European

VOLUME 23, 1961/62, No. 1 pp. 1-28 Published 16th October 1961

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

THE "PASCAL", A FAST DIGITAL ELECTRONIC COMPUTER FOR THE PHILIPS COMPUTING CENTRE

by W. NIJENHUIS *).

681.14-523.8

One of the aims in the design of digital electronic computers is the attainment of higher and higher computing speeds. In view of the fact that Philips supply many components, such as valves, transistors, diodes and magnetic memory cores, to computer manufacturers, it was important to have first-hand experience as to the demands made on components for computers in this and other respects. Investigations on this subject go back to about 1954 when an experimental computer was built in the Philips Research Laboratories. This machine has also been used for mathematical work of the laboratories. On the basis of the experience thus obtained, two larger and faster machines have now been built, the PASCAL and the STEVIN. Both machines will be put to use in the Philips Computing Centre to test their utility in scientific and administrative problems taken from actual practice.

*This article gives a short description of the design and properties of the PASCAL, without going into the detailed circuitry. The article begins with an introduction to the field, outlining the general design of digital electronic computers and defining a number of concepts. For a fuller treatment the reader is referred to various textbooks on the subject **).*

A digital electronic computer can be used to solve problems of widely different kinds with great speed. In principle such machines can only perform the more elementary mathematical operations, such as addition, subtraction, multiplication, and division, and a number of simple organizational operations including, for example, discrimination (the act of distinguishing whether a number is larger or smaller than another). The problems to be solved can, however, frequently be reduced to a series of such elementary operations. These machines can thus be used to solve scientific problems involving integrals, higher-degree equations, ordinary and partial differential equations, matrices, etc., They have proved equally usefull in dealing with administrative problems in the field of pay-billing, stock administration, order and invoice processing. The possibilities have never been expressed more clearly than in the pronouncement of Lady Lovelace: "The machine can do whatever we know how to order it to perform" [1]).

An experimental machine of this kind, which is suitable for the solution of scientific problems, has been built in the Philips Research Laboratories in Eindhoven. The machine has been given the name PASCAL, after Blaise Pascal who in 1642 constructed one of the first computers (which was of course mechanical). A second machine built in the laboratory and named after Simon Stevin, one of the fathers of accountancy, is intended more for administrative purposes. The STEVIN is basically similar to the PASCAL, but the auxiliary equipment has been extended in several respects so that it can

*) Research Laboratories, Eindhoven.

**) See, for example, C. B. Tompkins, J. H. Wakelin and W. W. Stifler Jr., High-speed digital computing devices, McGraw-Hill, New York 1950; R. K. Richards, Digital computer components and circuits, Van Nostrand, New York 1957; G. Haas, Grundlagen und Bauelemente elektronischer Ziffernrechenmaschinen, Philips Technical Library, Centrex, Eindhoven 1961.
An elementary introduction to computers (including their historical development) is to be found in the book referred to in footnote [1]).

[1]) B. V. Bowden, Faster than thought, Pitman, London 1953, pp. 30 and 398.

Table I. Principal data of the PASCAL, a binary parallel machine of the single-address type.

Word length	42 bits + 2 parity bits
Numbers with fixed point	41 bits + sign bit
Numbers with floating point	exponent (powers of 2): 7 bits + sign bit, fractional part: 33 bits + sign bit
Instructions	2 per word: 6 bits for the operation, 3 modification bits, 1 address interpretation bit, 11 address bits
Clock-pulse frequency	660 kc/s (basic period 1.5 μsec)
Some processing times *)	
addition and subtraction	with fixed point about 10 μsec with floating point 14-64 μsec
multiplication	with fixed point about 71 μsec with floating point about 59 μsec
division	with fixed point about 73 μsec with floating point about 61 μsec
number of instructions carried out per second in a typical programme	60 000
Memories (kinds and capacities)	
plug-board memory	16 words of 44 bits
modification memory	8 half-words of 22 bits
magnetic-core memory	2016 words of 44 bits
drum memory	16 384 words of 44 bits
magnetic tape	about 10^6 words of 44 bits with bits for longitudinal and transverse parity checks
Input (types and speeds)	
punched tape (5 or 8 channels)	120 characters per second with the possibility of stopping at each character; or 1200 characters per second, with continuous reading and using a special character for stopping
punched cards	2½ or 12 cards per second (both "Bull" and "IBM" code)
magnetic tape	on average 8000 words per second ("Ampex")
Output (types and speeds)	
punched tape	60 characters per second ("Teletype")
punched cards	1¼ cards per second ("Bull" tabulating machine)
magnetic tape	8000 words per second ("Ampex")
electric typewriter	10 characters per second ("IBM")
line printer	2½ lines of 90 characters per second ("Bull" tabulating machine)

*) Including waiting time for fetching the number and the instruction from the memory.

be used for processing the large number of data that accompany administrative problems. Both machines, which belong to the class of "fast" machines, have been installed in the Philips Computing Centre (*fig. 1*), where the most important electronic computers used in the Dutch Philips factories have recently been brought together.

Fig. 2 gives a general view of the PASCAL. *Table I* contains a summary of the properties of this machine, in the nomenclature usual for electronic computers. In the course of this article these properties will be explained. On the assumption that this nomenclature is not familiar to most readers, a general survey of the components of a digital electronic computer will first be given, in which all the concepts and terms mentioned in Table I will be explained.

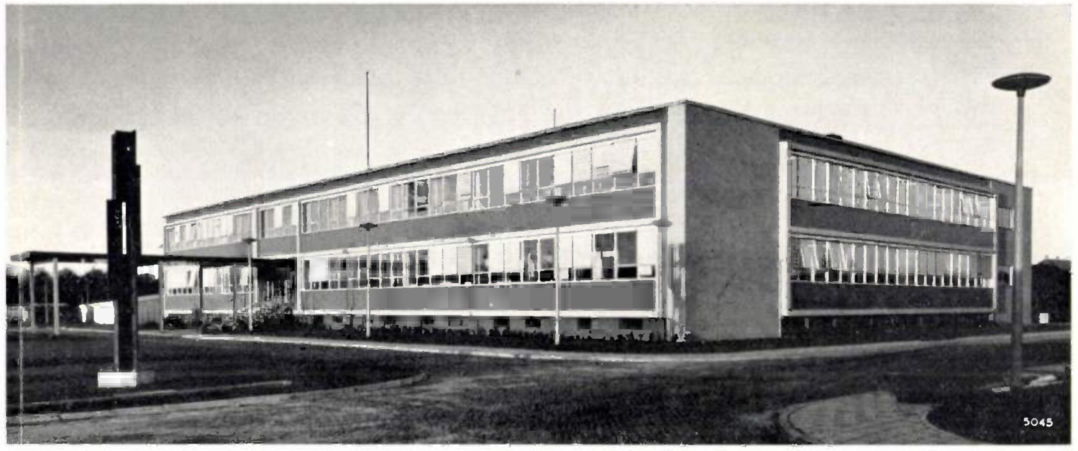


Fig. 1. The Philips Computing Centre in Eindhoven. In this building, which was officially opened on 21 March 1961, the principal electronic computers used by Philips in Eindhoven for administrative and scientific purposes have been collected. Apart from the PASCAL and the STEVIN it houses at present two IBM 650 digital computers, a Bull Gamma B digital machine and an analogue computer, the PACE.

The binary system

The description of the PASCAL in Table I begins with the word "binary", and therefore we start with a few remarks concerning the binary or two-digit system.

A digital machine calculates with discrete numbers, each number being expressed according to a certain code by a series of ciphers (digits). These digits are represented by elements or circuits that can take up several stable states. This is different from what happens in an analogue computer, in which a number is represented by a physical quantity that is continuously variable.

Most digital machines do not work with the decimal system but with the binary system. This system comprises only the digits 0 and 1 (called binary digits or "bits"), so that the numbers can be represented by means of a series of *bi*stable elements, such as flip-flops (circuits with two valves or transistors of which either the one or the other may be conducting) or magnetic cores (small ferromagnetic rings that can be magnetized in one of two opposite directions). Such a series of bistable elements, one for each bit, is called a *register*. It is of course quite true that there are also elements that can exist in ten stable states (e.g. the decade counter tube E 1 T [2]). Using such elements one would there, fore also be able to work with the decimal system, but they require a much more complicated circuit, so that this system is hardly ever used in computers If some modern computers are yet said to work with the decimal system, this generally means that they make use of a system in which each decimal digit is itself separately coded by some means or other in binary form.

A disadvantage of the binary system is that the numbers are at least three times as long as in the decimal system. The advantages, however, amply counterbalance this disadvantage.

In the binary system the arithmetical rules are very simple:

$$0+0=0 \qquad 0+1=1 \qquad 1+0=1 \qquad 1+1=10$$

$$0\times0=0 \qquad 0\times1=0 \qquad 1\times0=0 \qquad 1\times1=1$$

[2] A. J. W. M. van Overbeek, J. L. H. Jonker and K. Rodenhuis, A decade counter tube for high counting rates, Philips tech. Rev. 14, 313-326, 1952/53.

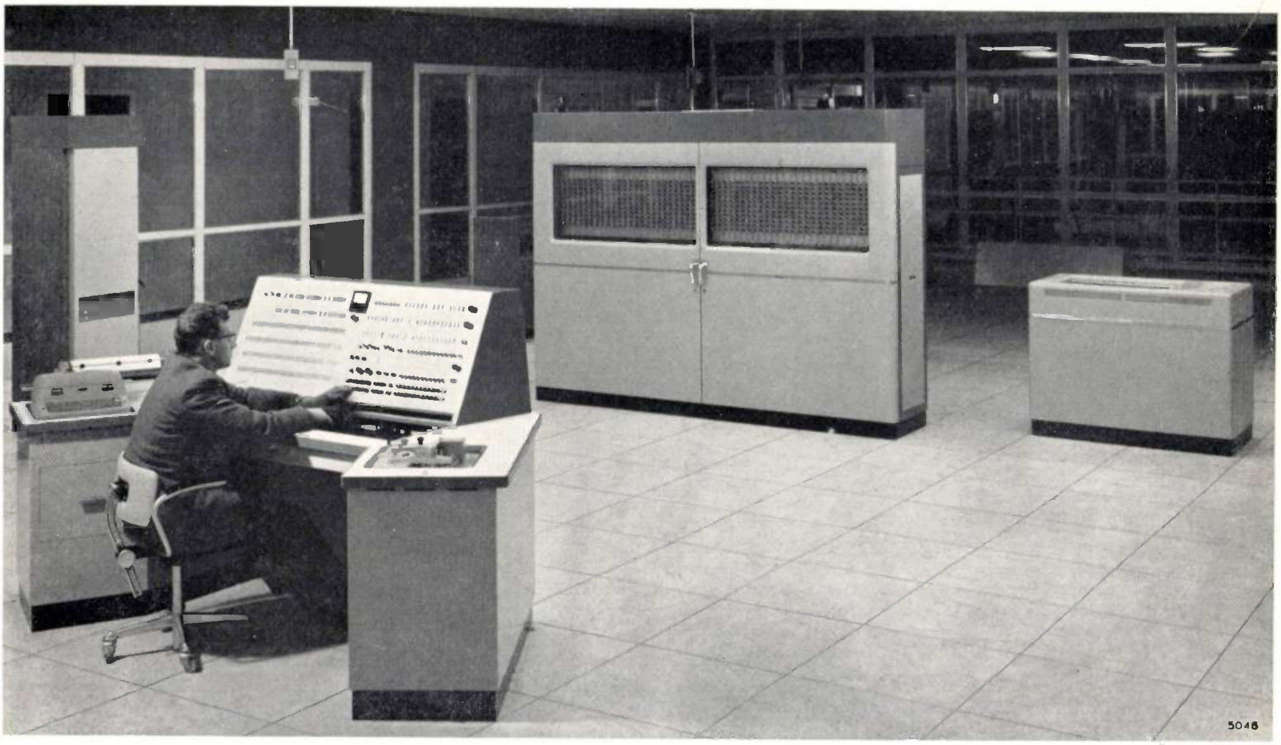


Fig. 2. General view of part of the PASCAL equipment. The large cabinet contains the principal components of the computer, viz. the arithmetical unit, the control unit and the magnetic-core memory. The low cabinet on the right contains the magnetic-drum memory. The PASCAL is operated from the desk in the foreground. The following are not visible: the magnetic tape recorders of the memory, the punch-card equipment for input and output, and the line printer.

We will give a simple example of the use of the binary system. In the decimal system the number 13.625 has the value of $1\times 10^1 + 3\times 10^0 + 6\times 10^{-1} + 2\times 10^{-2} + 5\times 10^{-3}$. In the binary system this number becomes 1101.101, as the value of this is $1\times 2^3 + 1\times 2^2 + 0\times 2^1 + 1\times 2^0 + 1\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3}$ which, expressed in the decimal system, is equal to 13.625.

Here are a few calculations in the binary system:

$$\begin{array}{rcr} 110110011 & = & 435 \\ +111001 & = & +57 \\ \hline 111101100 & = & 492 \end{array} \qquad \begin{array}{rcr} 1110010 & = & 114 \\ -1011 & = & -11 \\ \hline 1100111 & = & 103 \end{array} \qquad \begin{array}{rcr} 1101 & = & 13 \\ \times 1011 & = & \times 11 \\ \cline{1-1} 1101 & & \\ 11010 & & \\ 1101000 & & \\ \hline 10001111 & = & 143 \end{array}$$

Survey of the components of a digital computer

Practically every digital computer consists of the following components:

1) An arithmetical unit, in which the operations of addition, subtraction, multiplication and division, and a few others, can be carried out.
2) A memory, in which the data that have to be used or stored during the computation are kept. The memory consists of registers, each of which is indicated by a number, the *address*. The contents of a memory register are called a *word*. This may be a *number*, or an *instruction* expressed (coded) in a certain way as a binary number.

   An instruction indicates what operation is to be carried out. There must also be an indication concerning the number or numbers on which the operation is to be performed. In this article we shall only discuss computers having instructions of the single-address type, such as those used in the PASCAL. Here each instruction gives the address of *one* number. Provision has to be made for the second number that is normally concerned in a process to be already available in the arithmetical unit, either as a result of the previous process or as the result of a special instruction.
3) A control unit, whose function it is to ensure that the separate instructions are carried out in the correct sequence. The complete list of instructions to be carried out during the desired computations, the *programme*, is stored in the memory.
4) Input and output devices. The programme, and the numbers belonging to it, are supplied to the machine by means of an input device, e.g. a

punched-tape or a punched-card reader. To extract the answers from the machine, there is an output device, e.g. an electric printer or a card punch.

The arithmetical unit

The arithmetical unit of most machines contains three registers (see *fig. 3*), which may each consist of a row of flip-flops. The register M forms the connection with the memory. If a number is brought from the memory to the arithmetical unit, the memory supplies the requisite voltage pulses to set the flip-flops of M in the corresponding states. The successive digits of the number can be transported either one by one (serial transport) or simultaneously (parallel transport). The latter requires more equipment but less time and is therefore used in the PASCAL. All numbers must also pass through M on their way *to* the memory.

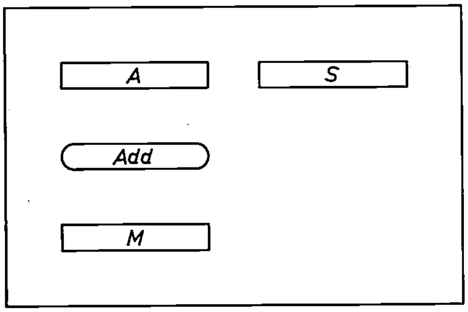


Fig. 3. Diagram of the arithmetical unit of a digital computer. The register M contains the number brought from the memory. The adder *Add* gives the sum of the number in M and the number in the register A (the accumulator). (This may, for example, be the result of a previous operation.) The register S (shift register) is used in multiplication and division. Each of the registers A, M and S consists of a series of flip-flops (42 in the PASCAL).

The second register of the arithmetical unit is the *accumulator A*, so called because the sum can be accumulated therein during addition. The third register S (*shift register*) is used in multiplication and division, among other things.

The central part of the arithmetical unit is the *adder Add*. The adder forms the sum of the numbers standing in A and M. This is a fundamental operation, and the circuits by which it is carried out (and which are built up from "logical" circuits) will presently be discussed in more detail. During addition the sum thus formed is generally taken up in A, and the previous contents of this register are lost.

The adder *Add* may be designed to add the numbers in A and M digit by digit, or it may contain a specific circuit for each digit, so that all digits of the numbers are processed at the same time. In the former case one speaks of a serial adder, in the latter of a parallel adder. Analogous to what was said about transport, the parallel adder requires a more complicated circuit but works much more rapidly than the serial adder.

Subtraction does not differ much from addition: the number arriving at M from the memory is there first given the opposite sign (most digital computers are designed so that this can be done quite simply by changing ones into noughts and vice versa; this is known as inversion) and then added to the number standing in A.

In a multiplication the multiplier is first placed in S by a separate instruction. The register A is "cleared" by the multiplication instruction proper (if there is a number in A that has to be kept, this number must first be transferred to the memory by a suitable instruction) and the multiplicand is brought to M. The multiplication process now takes place as follows:

a) If the last (least significant) figure in S is a 1, the contents of M are added to A, after which A and S are both shifted one place to the right. During this shift the last figure of the multiplier, which has now performed its task, is lost. The last figure from A, which must not become lost as it is part of the product, is inserted in the now empty first position of S. The first position in A, being now free, is filled with a zero (it being assumed that in this and the following example the numbers are positive; cf. the example in *fig. 4* which represents the various stages in the multiplication of $11 \times 13 = 143$, for registers of 5 bits).

b) If the last figure of S is 0, nothing is added, but the contents of A and S are still shifted one place to the right.

Because of the shift, both in case *a*) and in case *b*), the multiplier digit to be considered always stands in the last position of S. The adding and/or shifting is repeated until the multiplication is finished; the whole multiplier will then have been shifted out of S and the product (which will now have twice the length of a register) will occupy the registers A and S together.

Division proceeds analogously but in the opposite direction to multiplication. Here the dividend must first be placed in the A register by means of a separate instruction. (S may also contain a part of the dividend if the latter happens to consist of more digits than normal.) The division instruction proper then places the divisor in M. The arithmetical unit contains a circuit that first "compares" the contents of A and M. The contents of M are now subtracted from the content of A if possible, and a 1 is placed at the right-hand end of S if subtraction takes place. The contents of A and S are then shifted together one place to the left. If it appears that the subtraction is not possible, only a shift takes place, and a 0 is placed at the right-hand end of the S register. At the end of the process the quotient will be in S and the remainder in A.

The procedure followed in carrying out such a composite instruction is called a microprogramme.

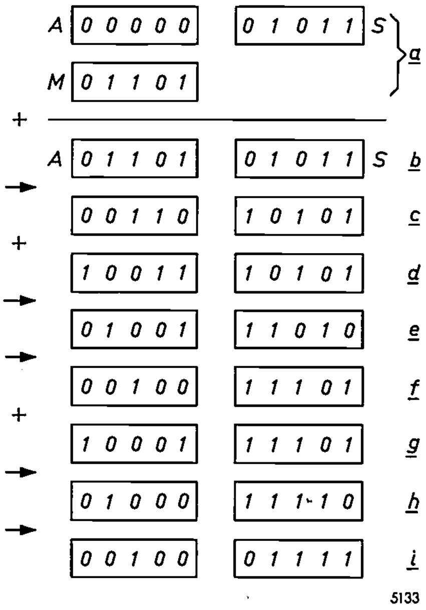


Fig. 4. Example of a multiplication in the arithmetical unit: the multiplication 1011×1101. *a*) Initial state of the registers A, S and M. *b*) - *i*) State of the registers A and S during the successive steps of the multiplication. The last figure in S always indicates what the following step has to be. If there is a 1, the contents of M are added to those of A (see *b*, *d* and *g*), after which the contents of A and S are together shifted one place to the right (see *c*, *e* and *h*). If there is a 0, only shifting takes place (see *f* and *i*). At the end of the operation the product (twice the register length) is found in A and S together.

The memory

The memory of an electronic computer must comply with two conditions: it must be large, i.e. it must be able to contain many words, and it must be rapid, i.e. the selection and reading of a word must not take long. As these conditions are more or less incompatible, most machines have several kinds of memories, which comply with either the one or the other demand.

In practice one frequently meets a magnetic-core memory as a rapid memory (*fig. 5*) and a magnetic drum memory (*fig. 6*) as a large (but slow) memory. The magnetic core memory then works in direct combination with the arithmetical unit and is called the working memory, while the drum memory contains the information that is not directly concerned in the computing process. Both kinds of memory have been described previously in this journal [3].

The instructions

As already mentioned, an instruction indicates which operation is to be carried out and where in the memory the relevant number is to be found. One might imagine a list of all admissible operations, so that each one could be indicated by its number in the list. In this way each instruction takes the form of:

number of the operation	address of the number

Some find it easier to indicate an operation with a readily remembered group of letters. This can, if necessary, be translated by the machine itself into the numerical code during the input of the problem. For example, *TOA* 513 might mean: add to A the number found at address 513. The number enters the arithmetical organ via register M and is added to whatever there was already in A. This sum can, for example, then be stored at address 127 by the instruction: *SAP* 127 (store A positive at address 127). In a multiplication it is first necessary to set up the multiplier in S by an instruction, e.g. *LDS* 500 (load S with the number at address 500). The microprogramme of the multiplication is then started by means of the instruction *MPY* 501 (multiply the number at address 501 by the number in S).

The instructions are stored in a row of successive addresses in the memory. When an instruction has been carried out, the control unit (cf. the following section) normally fetches the new instruction from the next address. If one wishes to go over to another row of instructions, "jump" instructions are used. On the instruction, for example, *JUN* 203 (jump unconditionally to address 203), no processing takes place but the control fetches the next instruction from address 203 and then proceeds to address 204 and subsequent ones. Such a jump may also be "conditional", i.e. it is carried out only if a certain condition is fulfilled. Thus, for example, the instruction *JAP* 200 means: jump, if

[3] H. J. Heijn and N. C. de Troye, A fast method of reading magnetic-core memories, Philips tech. Rev. 20, 193-207, 1958/59.

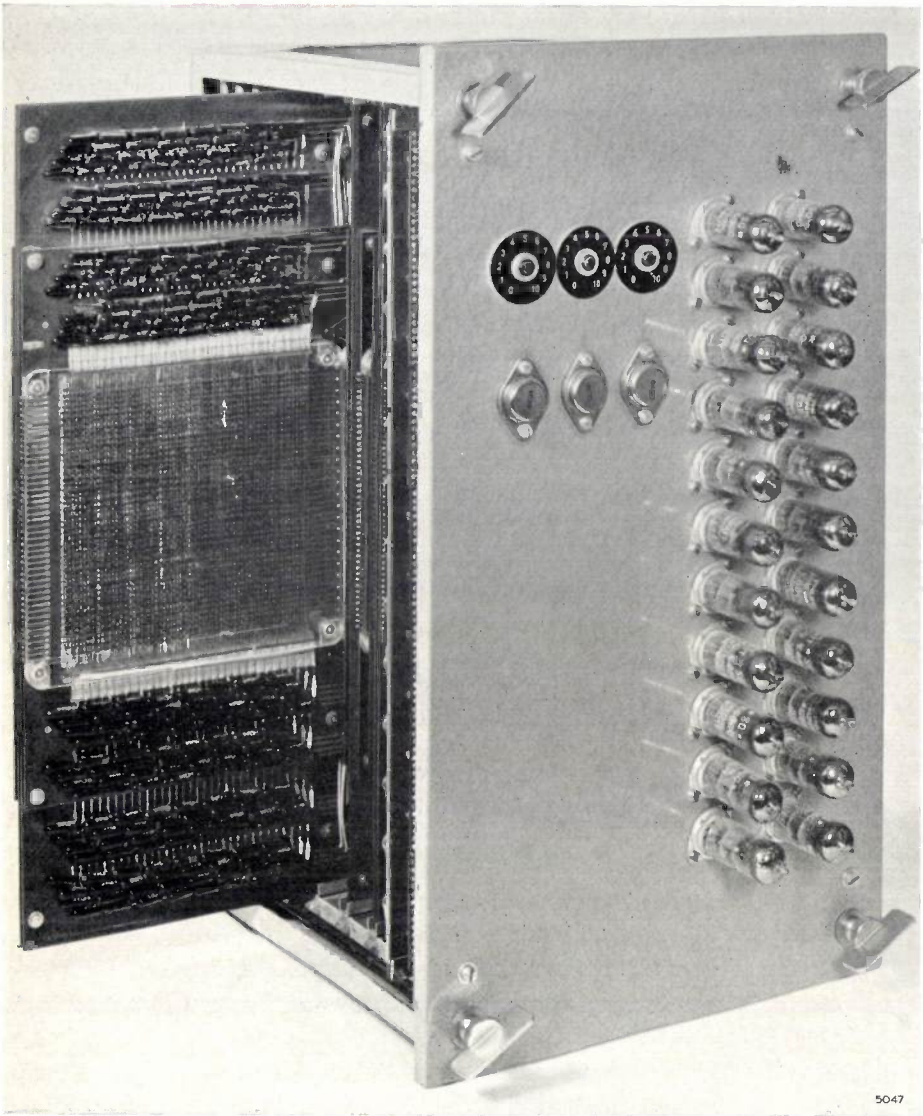


Fig. 5. A magnetic-core memory. This memory consists of four panels, each containing 256×44 cores for 256 words. One panel has been pulled out from the chassis. Printed wiring is used; the connections between the panels take the form of connection strips that engage in contact clamps. The chassis also contains circuits that serve for writing data into the memory and for reading them out of it. The PASCAL contains two of these memories.

the accumulator A is positive, to address 200. If the contents of A are negative, this jump is not carried out and the programme continues normally.

A number of instructions in the PASCAL which differ from the conventional ones will presently be discussed in more detail.

The control unit

The overall working of a digital computer proceeds in two phases. In the "control phase" an instruction, consisting of an operation part and an address part, is brought from the memory to the control unit and there placed in a register, the C register. A selection circuit connected to the C register selects the microprogramme indicated by the operation part of the instruction, while the address part of the instruction is transported to a second register of the control unit, the address-selection register T. There is a further selection circuit connected to this register; this selects the word to be processed from the appropriate address in the memory. At the instant that the address part of the instruction is transferred to T, the latter still contains the address of the instruction in the process of being carried out, and this address must not be lost. (It should be remembered that each instruction *has* an address where it is to be found, and *contains* the address of the number to be processed.) It is therefore temporarily placed in a suitable place of safety. After this the number specified by the instruction may be taken from the memory and brought to the arithmetical unit.

Then the control changes over to the "operation phase", and the microprogramme for the corresponding operation is started. Once this is completed, the machine switches back to the control phase. The address of the instruction just completed is brought out from its place of safety and, increased by one, brought back to T, so that this register now contains the address whence the new instruction has to be fetched. The cycle can then start again.

When a jump instruction (in which as already mentioned no processing takes place) occurs, the number in T is replaced by the address mentioned in the instruction, so that the following instruction is fetched from this place, the operation phase being simply omitted.

Input and output devices

For the input of programmes and numbers, use can be made of punched tape, punched cards or magnetic tape. On these the information need not be placed in the binary form required by the machine but may, for the convenience of the user, take a decimal form or the above-mentioned letter code. For punched tape one can, for example, make use of the teleprinter code.

The machine is fed with a so-called input programme with which this code is translated into the binary form. The mechanical input devices work much more slowly than the machine can perform the translation, so there is plenty of time for this translation work.

Similarly, an output programme arranges for the answers to be given to the output devices, which may take the form of, for example, an electric typewriter, a tape punch, a line printer, a card punch or a magnetic-tape apparatus.

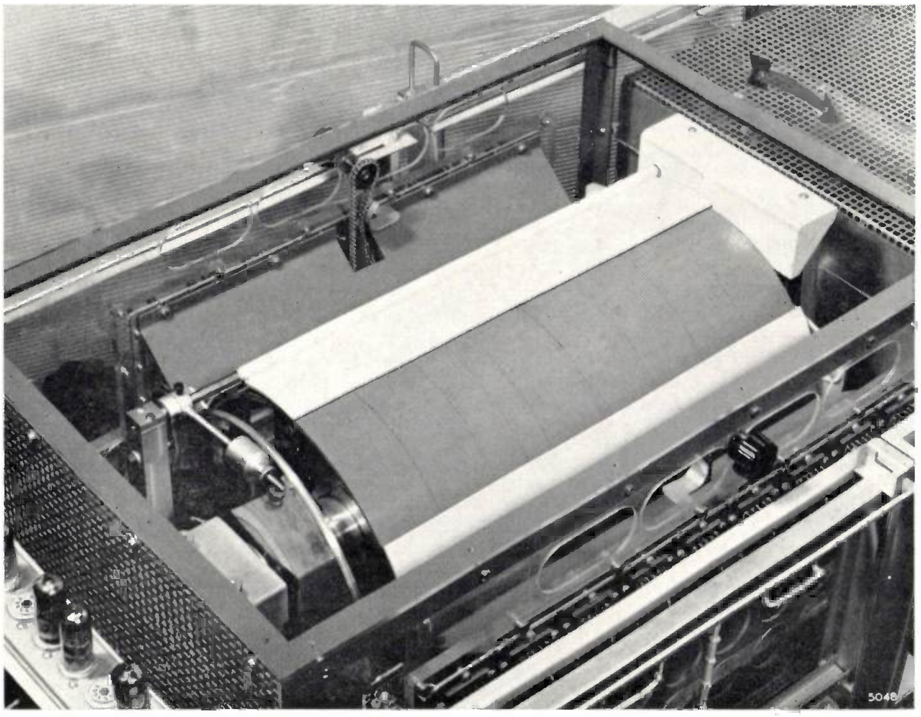


Fig. 6. A drum memory. The continuously rotating drum, which is coated with a thin layer of magnetizable material, is driven by an electric motor. During rotation, the drum passes under a series of heads that magnetize small areas of the surface in one of two opposite directions and that are also used for reading the magnetization. The drum here illustrated is used for the PASCAL; it rotates with a speed of 6000 rev/min. It has 256 heads which are arranged in two horizontal rows along the drum. Some connections to those of the front row are visible. The white metal plates ensure that the heads do not touch the drum when it expands. For this purpose the plates are pivoted and float on the air stream produced by the rotating drum. As the drum expands, the plates are raised a little further and the angular displacement thus produced increases the distance between the heads and the shaft of the drum.

Logical circuits

To make clear the character of logical circuits, we might perhaps best consider the addition circuit, *Add* in fig. 3. Let us consider the situation somewhere in the middle of the number. We have to add the digit a from A to the digit m from M, and we may have to carry one from the previous digit, which may be regarded as the addition of a further digit c_{i}. The result is the digit s, and c_{u} to be carried to the next digit.

Table II gives the possible combinations of a, m and c_i that may occur, and the corresponding values of s and c_u. It can be seen that, for example, $c_u = 1$ if:

$$\begin{aligned} & a = 1 \text{ and } m = 1, \\ \text{or: } & a = 1 \text{ and } c_i = 1, \\ \text{or: } & m = 1 \text{ and } c_i = 1, \\ \text{or: } & a = 1,\ m = 1 \text{ and } c_i = 1. \end{aligned}$$

Table II. The sum s and the outgoing carry c_u at a certain position in the adder, for different values of the digits a and m to be added and the incoming carry c_i.

a	m	c_i	s	c_u
0	0	0	0	0
1	0	0	1	0
0	1	0	1	0
0	0	1	1	0
1	1	0	0	1
1	0	1	0	1
0	1	1	0	1
1	1	1	1	1

Now we may represent a 0 by e.g. a low voltage and a 1 by a high voltage (for example 10 volts and 15 volts). A circuit can then be constructed with diodes and resistors which, when supplied with voltages corresponding to the values of a, m and c_i, only supplies a high output voltage if the above conditions are complied with. This circuit thus gives the required result c_u.

One may also say that the circuit carries out the operations *and* and *or* occurring in the conditions. These are operations from a branch of mathematics called propositional logic, hence the name "logical circuits".

Logical circuits are built up from two basic types corresponding to the two fundamental operations, viz. the *and* and the *or* circuit. *Fig. 7a* represents the *and* circuit. Point P has the high voltage. The output, point U, will now have the high voltage only if both the points x *and* y have the high voltage. This circuit will thus give 1 only if x *and* y are each equal to 1. In the *or* circuit, fig. 7*b*, point Q has the low voltage. The point U will have the high voltage if either x *or* y *or* both have the high voltage. This circuit thus gives a 1 if x *or* y *or* both are equal to 1.

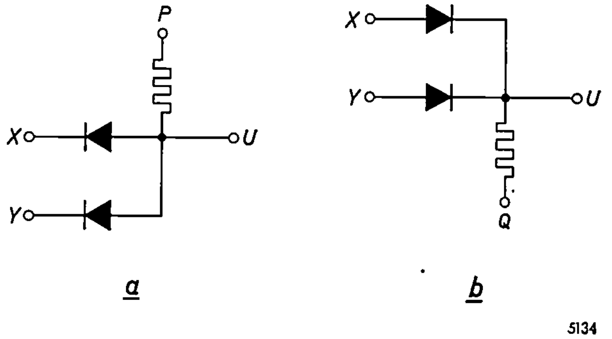


Fig. 7. The two basic types of logical circuits. *a*) *and* circuit. Point P here has a "high" voltage. Point U only has a high voltage if both x and y have a high voltage. *b*) *or* circuit. Q here has a "low" voltage. Point U has a high voltage if x or y or both have a high voltage.

The circuit for obtaining c_u from a, m and c_i might be as shown in *fig. 8*. To reduce the delaying effect of parasitic capacitances in the circuit, the voltage V is chosen larger than the voltages a, m and c_i.

One of the functions of logical circuits in computers is to activate a certain part of the machine on the receipt of certain combinations of input signals. This is the case for example in the above-mentioned selection circuit connected to the operation part of the C register in the control unit which switches on the correct microprogramme, and in the selection circuit of the T register which brings about the reading of the desired address in the memory.

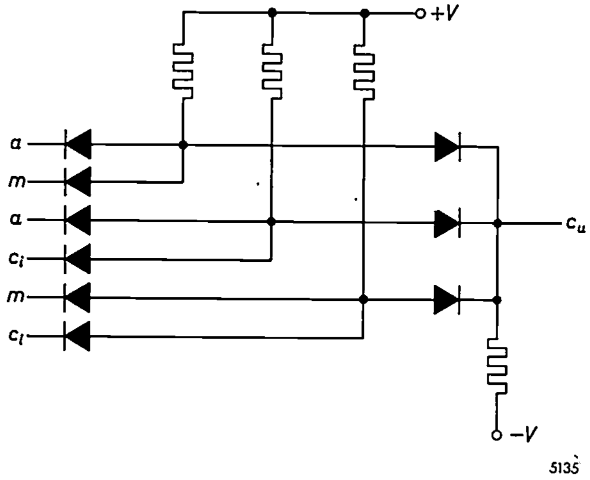


Fig. 8. Example of a logical circuit which determines the carry c_u to the following digit from the incoming carry c_i and the digits a and m at a given position in the registers A and M (for example: 0 is represented by a low voltage of 10 V and 1 by a high voltage of 15 V). This circuit is composed of three *and* circuits and one *or* circuit. (The adder of the PASCAL uses a different circuit for a more rapid determination of the carry.)

Special features of the PASCAL

So much for the general description of the working of digital electronic computers. In the following section of this article we shall discuss a few special features in the equipment of the PASCAL, and we shall follow more or less the same arrangement as used in the first section. We shall here deal with some more new concepts, such as the subroutine, the number coding with floating point, and the modification of an instruction.

The PASCAL shares many of its special features with other machines of recent years; this will not be mentioned every time. As far as we know, the following have not yet been used in other machines: the method for rapid transfer of instructions in the addition circuit; the Σ switch; the plug-board memory; the method of modifying instructions; the link instructions for going over to a subroutine, and a few more special instructions.

At the end of the article we shall discuss a few constructional problems and make some remarks concerning the PASCAL in comparison with other machines.

The coding of the numbers

The binary notation for numbers was discussed in the beginning of this article. It is obvious that when numbers expressed in this notation are added or subtracted, all the numbers must have the decimal point, or rather the "binary point", in the same place. This is referred to as *fixed-point notation.* The possible magnitude of numbers expressed in this notation is limited. It was mentioned in Table I that 41 bits are available for fixed-point numbers in the PASCAL. In order to distinguish positive and negative, one bit (sign bit) is added, 0 being used to indicate a positive and 1 a negative number. Now if the point is, for example, placed at the end of the number, the range covered by 41 bits extends from 0 to about 10^{12} (12 decimals). This range can be increased considerably by using the *floating-point* notation, in which numbers are indicated by a fractional part p and an exponent q. The value of such a number in the binary system is then $p \times 2^q$, the absolute value of p always being chosen to lie between a half and one, so that q really indicates the position of the point. In the PASCAL the $41 + 1$ available bits are divided into a fractional part of 34 bits, one of which is a sign bit, and an exponent of 8 bits, one of which is a sign bit. The absolute value of the numbers can now lie between $\frac{1}{2}\times 2^{-127}$ and $1\times 2^{+127}$, or in other words between about 10^{-38} and 10^{+38}. The accuracy (33 bits in place of 41) is of course smaller than in the fixed-point notation and will cover about 10 decimals.

Addition and subtraction of numbers expressed in the floating-point notation is more complicated and takes longer than for numbers in the fixed-point notation, because the exponents of both numbers must first be compared and made equal, while an extra step may be necessary after addition in order to bring the absolute value of the fractional parts back to between $\frac{1}{2}$ and 1. Multiplications and divisions are a little faster, mainly because the length of number (fractional part) to be processed is shorter (see Table I). There are special microprogrammes in the control unit for working with numbers in the floating-point notation.

The arithmetical unit of the PASCAL

The arithmetical unit of the PASCAL has the three registers M, A and S discussed above. The registers M and A serve as inputs for the adder, which is a parallel one. Only 0.8 μsec after the arrival of the numbers in the registers, the answer will be available at the output of the adder. Special measures were obviously required to attain such a short addition time. Thus the carries were not only transferred digit by digit for all 41 digits, but also determined for successive blocks of 7 digits at once and passed on to the following block. In this way it was found possible to gain time without having to use too many extra components.

The determination of a carry takes 40 nanoseconds for each digit. A separate logical circuit determines the carry for a block of 7 digits to the first digit of the next block, from the incoming carry and the state of the registers A and M for these digits. This takes 100 nsec. In the second block the passing on of the carries and the determination of the carry of the third block can already start while the carry from the third to the fourth digit in the first block has still to be determined. The last carry of the addition is obtained after 780 nsec in this way, while without these block carries it would take $42\times 40 = 1680$ nsec.

This method of working also affords a check on the operation of the computer, because at the end of each block the carry is produced in two ways: first with the separate logical circuit, and immediately afterwards by virtue of the fact that the carries in the block are determined digit by digit. If these two carries do not agree, an error has occurred and the machine is automatically stopped.

A complete addition, including the fetching of the instruction as well as of the number, admittedly takes much longer than the above-mentioned addition time of 0.8 μsec (in the PASCAL it takes on the average 10 μsec), but the very short intrinsic addition (or subtraction) time is still of prime importance, particularly for division. Division is after all made up of a number of subtractions, and furthermore the course of the process differs according as subtraction is or is not possible, so that the microprogramme of the division must still decide for each subtraction how to continue. Thanks to the very short addition time each division step in the PASCAL still does not occupy more than 1½ μsec.

The clock pulses

The above-mentioned time of 1½ μsec is called the basic period of the machine, and all separate steps in the working of the machine occur at intervals of 1½ μsec or multiples of this. These intervals are determined by a periodic series of voltage pulses, the clock pulses, which are formed by means of a disc fixed to the shaft of the drum memory. The circumference of the disc is provided with small teeth of a permanently magnetic material. As the disc rotates, the teeth generate an AC voltage in a reading head, from which the pulses are taken. They have a frequency of 660 kc/s and thus appear at intervals of about 1½ μsec.

To illustrate the function of the clock pulses in the working of the PASCAL, let us consider a logical circuit. This provides, as mentioned above, a high or a low output voltage, depending on whether or not a certain combination of input signals is present. This output voltage nearly always serves for setting a flip-flop to a corresponding state. This must often be done for several flip-flops simultaneously. The output voltages of the different logical circuits are not, however, usually available at precisely the same instant (this depends on, for example, the size of the circuits). The instants at which the flip-flops are activated must not therefore be determined by these output voltages. The moment of activation of each flip-flop is therefore determined by the first clock pulse appearing after the output voltage of the logical circuit concerned has been formed. This happens as follows.

The flip-flop F (*fig. 9*) is not activated by the output voltage of the logical circuit L but by the clock pulse k; the output voltage of L does however determine which half of the flip-flop is to be supplied with the clock pulse, by means of the two-way switch P. Two voltages of opposite sign are needed to operate the two-way switch, i.e. the polarity of these voltages determines the state of P. The voltages are formed by the circuit N and their polarity is determined by whether the output voltage of L is high or low. Two cathode followers K connected to the flip-flop supply the power for activating other parts of the machine. The circuit of fig. 9 is a basic circuit which occurs in many different parts of the machine.

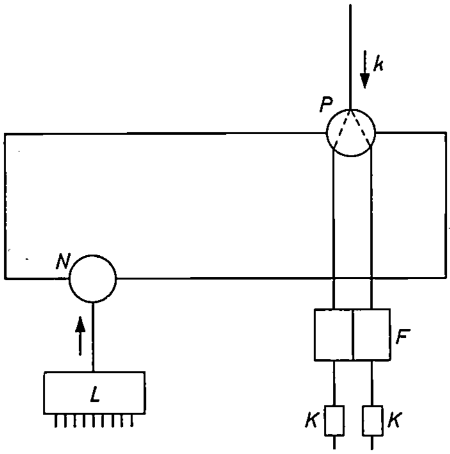


Fig. 9. Basic circuit used in many parts of the PASCAL. This kind of circuit is used to set a flip-flop to a state determined by the output voltage of a logical circuit at the instant when a clock pulse appears. The logical circuit L supplies either a high or a low voltage. On receipt of this, the circuit N produces a positive and a negative voltage, the polarity of which is determined by the magnitude of the signal from L. The state of the two-way switch P is in its turn determined by the polarity of the voltages from N. The first subsequent clock pulse k is now passed by this switch to one of the two halves of the flip-flop F. Depending on the magnitude of the voltage supplied by L, this flip-flop will thus take up a certain position. Cathode followers K connected to F supply the power for activating other parts of the machine.

The basic period of 1½ μsec has been chosen so that all logical circuits will have supplied their output signal within this period.

The Σ switch

The outputs of the adder are connected to a six-way electronic switch Σ. This makes it possible to transfer the sum obtained in the adder to one of a number of other registers (see *fig. 10*). The Σ switch actually consists of 42 identical switches, because it must be possible to effect the six connections for each digit simultaneously. First of all the sum can be placed in A; this 'is necessary in normal addition instructions. The sum can also be placed in A shifted one place to the left or right, so that shifting during division or multiplication does not cost any additional time. Because of this little trick the PASCAL requires an average of only 71 μsec for a multiplication (including fetching the instruction and the multiplicand), and 73 μsec on average for a division.

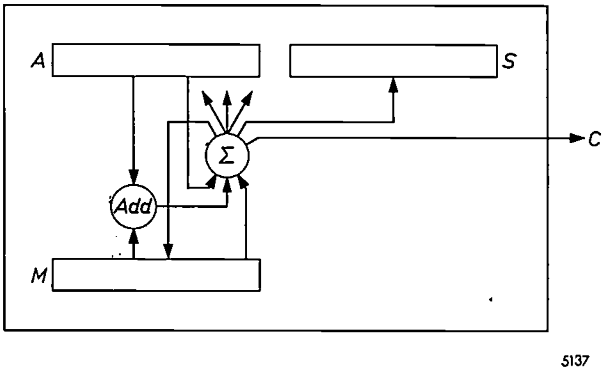


Fig. 10. Diagram of the arithmetical unit of the PASCAL with the 6-way Σ switch. The latter is able to make the requisite connections (42 in parallel, one for each digit) between the adder *Add*, the registers A, M and S of the arithmetical unit and register C of the control unit.

The Σ switch also makes it possible to transport the sum obtained in the adder to the S register, which can thus be used also as an accumulator, and to the M register, so that the sum can be written directly in the memory. An arbitrary memory address can therefore also serve as an accumulator.

Finally, Σ can transport the sum to the control unit. This is used in modifying the instructions. (This important facility of the PASCAL will be discussed below.)

It is also possible to transfer the contents of register A or of register M to the six above-mentioned locations. These possibilities are used in a number of instructions in certain phases of the control.

To give an idea of the speed of the PASCAL, we may mention that it processes about 60 000 instructions per second for a programme containing one long operation (multiplication or division) for every ten short ones (addition, subtraction, etc.).

The memories of the PASCAL

As already mentioned, most computers have a rapid working memory and a large, slow, secondary memory.

The PASCAL has in all five kinds of memory. The working memory is of the magnetic-core type. The most important slow memory is a magnetic drum. There is also a very large slow memory with magnetic tape and two small special memories: a "plug-board memory" and a "modification memory".

a) The magnetic-core memory can accommodate 2016 words. The address part of an instruction consists of 11 bits, with which a selection can be made from $2^{11} = 2048$ addresses. The 32 addresses that do not occur in the magnetic-core memory are used for the special memories and for a few registers, among others A and S of the arithmetical unit. Selection of a certain address and reading of the contents requires 3 μsec in this memory, re-writing after reading a further 3 μsec.

b) The drum memory can contain 16 384 words, divided into "blocks" of 128 words. A block is written on two tracks that can each contain 128 half-words and which are written and read simultaneously, each with its own head. Writing erases whatever registration was already present. The access time with a drum memory (i.e. the time required for the selection of a certain word and the reading of it) is in principle variable, as a word can only be read when the relevant part of the rotating drum passes under the head of this track. In the PASCAL this variable access time causes no difficulties, as a block of words is always transferred in its entirety from the drum to the magnetic-core memory (which is divided into blocks of 128 words for this reason). As soon as the control receives a transport instruction, transport starts immediately and is maintained for a complete revolution of the drum. Provision is made for placing the words in the positions in the magnetic-core memory corresponding to their correct positions in the block, even if reading is started in the middle of a block. During such a transport of a block from the drum to the magnetic-core memory, transport of another block may take place in the reverse direction. One revolution of the drum takes 10 msec, so the transport speed is 12.8 words per msec in each direction.

During such transport the PASCAL continues to compute, albeit sometimes at slightly reduced speed (the reduction amounts at the most to 15%; it would take us too far afield to go into this in more detail).

c) The magnetic tape serves as a much bigger but much slower memory. It is possible to store some 10^6 words on each roll of magnetic tape. Here again exchange of data with the magnetic-core memory takes place in blocks of 128 words and, if more than one magnetic-tape apparatus is used, is possible in two directions simultaneously. A number of magnetic-tape units can be seen in *fig. 11*.

d) The plug-board memory can contain 16 words, which are inserted by placing plugs in appropriate holes; these 16 registers bear the addresses 0-15, which do not occur in the magnetic-core memory (see above). The plug-board memory serves for the storing of parameters that should be easy to change during computation, for the insertion of small test programmes, and for the starting of a new programme. To make this last even easier, the register with address 0 is equipped with push-buttons instead of plugs. The plug-board memory and the push-buttons for address 0 are to be found on the control desk (see fig. 11).

e) The modification memory contains eight half-words (i.e. space for eight instructions, see below). Transistorized flip-flop registers are used for the storage. This makes for a very short access time, just as in the plug-board memory. The registers can be filled in the same way as those of the magnetic-core memory by using normal instructions. They are given the addresses 16-23. The use of the modification memory will shortly be described in more detail.

After the discussion of the arithmetical unit and the memory we ought now to look more closely at the control of the PASCAL. It is, however, necessary first to consider anew the coding of the instructions.

The coding of instructions in the PASCAL

The chosen word length of 42 bits is determined mainly by the desired computing accuracy. The instructions are, however, very much shorter. Each word can then contain two instructions that are processed by the control successively. They are indicated as I_1 and I_2.

For each instruction 21 bits will now be available. 6 of these serve to indicate the kind of operation,

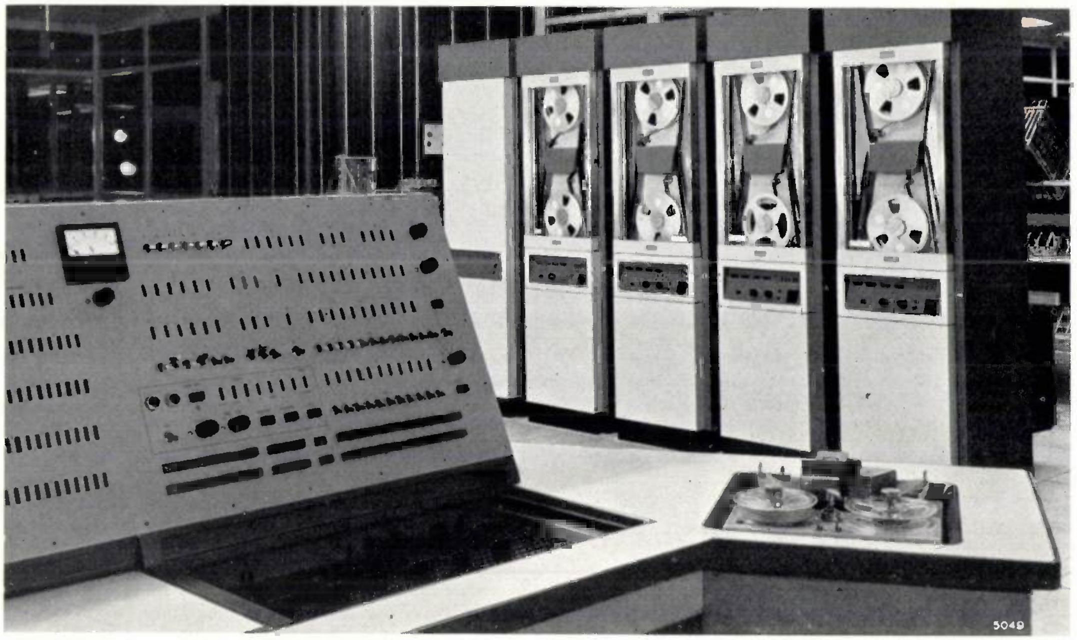


Fig. 11. Part of the control desk. At the back are some tape units which are used as a large, slow memory. The plug-board memory can be vaguely seen on top of the desk while the push-button register for address 0 is at the bottom right-hand side of the panel. To the right on the desk there is the reader for the input of programmes recorded on punched tape.

so that $2^6 = 64$ different operations can be carried out. As we have already seen, 11 bits are required for the address of the number to be processed (for a total of $2^{11} = 2048$ addresses in the various memories). Of the 4 remaining bits of an instruction, 3 are used for selecting one of the 8 modification registers; for an instruction can still be modified even after having been fetched from the memory but before being executed. By making use of the possibility of modifying instructions it is sometimes possible to achieve great economies in time and in memory space in the case of programmes in which a certain series of instructions has to be repeated a number of times. This will be made clearer below in the section in small print.

In the PASCAL a distinction is made between modifications in a register with an even number, and those in a register with an odd number. Since a magnetic-core memory has the property that a word is erased from the memory when it is read, such a word must be re-written in the memory. If the word that is read and brought to the arithmetical unit is an instruction pair, such re-writing is postponed till after the modification of the instruction. While the modification is being added to the instruction in the arithmetical unit, both the original and the modified form of the instruction are present in the arithmetical unit, so that it is possible to choose which one is to be re-written in the memory. When an even modification register is involved, the unmodified instruction is re-written, while with the odd registers the modified one is taken.

Of the 21 bits there now still remains 1 bit, the so-called address-interpretation bit. This determines how the address part of the instruction has to be interpreted. In normal cases this bit is a 0. If, however, it is a 1, then the address part of the instruction is not considered as an address of a number but as the number itself. This possibility can be utilized if some operation has to be carried out on a positive number < 2048. No separate memory position will then be required, for the operand is already given as part of the instruction.

The form of a complete instruction is represented in *fig. 12*.

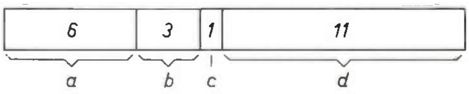


Fig. 12. Distribution of the 21 bits of an instruction: *a*) 6 bits indicate the operation to be carried out; *b*) 3 bits indicate the required register in the modification memory; *c*) one bit serves for the address interpretation; *d*) 11 bits indicate, if there is a 0 in (*c*), the address in the memory of the number to be processed. If in (*c*) there is a 1, the contents of (*d*) are to be taken as the number to be processed.

As an example of the application of a modified instruction, let us consider the calculation of the scalar product of two vectors **a** and **b** having ten components: $\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + \ldots + a_{10}b_{10}$. The components $a_1 \ldots a_{10}$ and $b_1 \ldots b_{10}$ are placed in the memory at addresses $n \ldots n+9$ and $n+10 \ldots n+19$. Now here is a process recurring ten times: multiply

a_i by b_i, add this to the previous result that might, for example, be located at address m, and store the new result again at address m. This requires the following instructions, whose meanings have already been indicated on p. 5 (they are placed at addresses k, $k+1$, etc., assuming for the sake of simplicity that there is only one instruction per address):

k	*LDS*	n,
$k+1$	*MPY*	$n+10$,
$k+2$	*TOA*	m *),
$k+3$	*SAP*	m.

*) The product now contains twice as many figures, but the part in *S* can be neglected.

Next the same cycle has to be carried out with the other components of the vector. This can of course be done by repeating these four instructions a further nine times with other values of n. Forty instructions will then have to be stored in the memory and carried out.

We could alternatively increase by one the addresses n and $n+10$ of the instructions in k and $k+1$ and then begin again at k by means of the following seven instructions:

$k+4$	*LDA*	k,
$k+5$	*TOA*	"1" *),
$k+6$	*SAP*	k,
$k+7$	*LDA*	$k+1$,
$k+8$	*TOA*	"1",
$k+9$	*SAP*	$k+1$,
$k+10$	*JUN*	k.

*) Here the number 1 is placed in the instruction itself: the address-interpretation bit is thus 1.

This time 11 instead of 40 instructions must be stored in the memory, but 110 instructions have to be carried out, so that the whole process takes much longer. (Counting is furthermore required to determine when the cycle has been carried out ten times.)

The second series of instructions will, however, become unnecessary if the instructions at addresses k and $k+1$ are modified "odd" by means of the number 1. Before they are carried out, the address part of these instructions is then increased by 1 and in this form the instructions are replaced in the memory. (The first time the addresses must in this case be $n-1$ and $n+9$, as modification also takes place when the instructions are carried out for the first time.) The programme is then reduced to the original four instructions plus a jump instruction, and these have to be carried out 10 times (the requisite counting will again be left out of account). Therefore not more than 5 instructions have to be stored in the memory, and 50 instructions have to be carried out. Instructions with modification each require only $1\frac{1}{2}$ μsec more than the non-modified ones. Compared to the first method there is therefore only a small loss of time, accompanied by great saving of space in the memory. Compared to the second method there is a gain in both respects.

The control of the PASCAL

Let us now consider the control of the PASCAL in greater detail with reference to the block diagram of *fig. 13*. This shows the arithmetical unit *Cal*, the control unit *Con* and the memory *Mem* with their registers. The diagram also shows the paths that can be followed by numbers and instructions (full lines) and by selection signals (broken lines).

We shall follow the working of the machine during the fetching and execution of an instruction. Suppose that it is the turn of the instruction pair at address 51. The number 51 will then be found in the *T* register, which performs the selection of the address in the memory. Let us further consider the simplest possible situation, viz. that the result of the previous operation is still left in the *A* register. The selected word with the two instructions I_1 and I_2 comes from the memory and first enters the

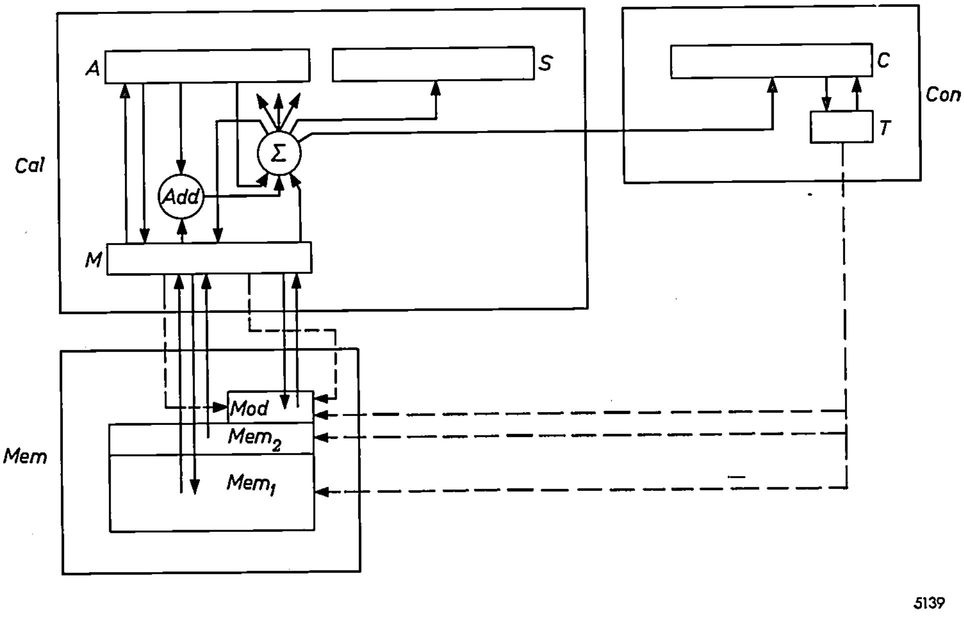


Fig. 13. Block diagram of the PASCAL with the arithmetical unit *Cal*, the control unit *Con* and the memories *Mem*. In the arithmetical unit the registers *A*, *M* and *S* and the adder *Add* are shown. Furthermore the arithmetical unit contains the six-way switch *Σ*, which at the command of the control unit can make the requisite connection between the various registers and the adder. The transport of numbers and instructions necessary for carrying out the processes is shown by full lines. The broken lines indicate the path of selection signals from the *T* register to the memory for the fetching of instructions and numbers (Mem_1 is the magnetic-core memory, Mem_2 is the plug-board memory) and from the *M* register to the modification register *Mod* for fetching the modifications of an instruction.

M register. If the instructions have to be modified, the correct registers must be selected from the 8 modification registers, using the selection circuits connected to the part of the *M* register where the three modification selection bits are to be found. The two modifications (one for each instruction) must now be added to the instructions. For this the *A* register is required, so it must first be cleared. Its contents must however be preserved. These can for the time being be placed in the control register *C*, for the contents of this (the old pair of instructions) may be destroyed. The instruction pair is itself brought from *M* to *A*, while the selected modifications, different for each half-word, go to *M*.

The adder now supplies at its output the instructions in modified form, while in *A* at this instant the instructions are still present in their unmodified form. Depending on the kind of modification (even or odd) one of these two forms is replaced in the memory via *M*. The modified instruction pair is brought to *C* and the original contents of *A* are simultaneously replaced in *A*.

Next the first instruction (I_1) must be performed. Its address part is transferred from *C* to *T* in order to select from the memory the number with which the calculation is to be carried out. The number 51 that stood in *T* must not be lost and is therefore stored for the time being in the corresponding part of *C*. (This exchange is necessary as only *T* can select a word from the memory.) When the desired number has arrived from the memory in *M*, the operation to be performed can start. If this is an addition, it is only necessary for the sum formed in *Add* to be transferred to *A*. This has already been explained above, as has the procedure for other operations.

Once the operation has been completed, a "ready" signal indicates that I_1 and I_2 in the *C* register must be changed, after which the second instruction is carried out. When the "ready" signal of the second instruction is also given, the next instruction pair must be selected from the memory. The address of the old instruction pair, increased by one, is therefore placed in *T* and the whole cycle is repeated.

The PASCAL's special instructions

The designer of computers is still given a chance to show his inventiveness in providing facilities for special instructions, and this is the main reason why the many machines built at the present time get an individual character. In the present section we shall review some of the special instructions that can be handled by the PASCAL. We have already mentioned the *conditional jump* instructions. A special form of these is the jump instruction that depends on the result of a previous *comparison* instruction. By means of a comparison instruction, *COM n*, the number at address *n* is compared with the number in *A*. The three possibilities of "bigger" "equal" and "smaller" determine the positions of two flip-flops. The particular jump instruction might now be, for example: Jump to address *m* if the result of the last comparison was "bigger". The control examines the contents of the two comparison flip-flops and determines whether or not the jump must be carried out. Three jumps of this kind are thus possible.

In view of the fact that the PASCAL contains two instructions per word and that it is necessary to be able to jump both to the I_1 half and to the I_2 half, all the various jumps occur in duplicate; jumps to I_1 have an even instruction number, while those to I_2 have an odd number.

For the *transport* of data from the drum memory to the magnetic-core memory and vice vèrsa the *transport* instructions *DTC* (drum to core) and *CTD* (core to drum) are used. It has already been mentioned in the discussion of the memories that a great saving of time is obtained with the PASCAL because 1) transport may take place in both directions simultaneously, 2) transport may commence at any instant and 3) the machine can in the meanwhile continue computing.

In thc PASCAL there is also a *count* and a *repeat* instruction. Both these instructions may only occur in the I_2 half of a word. In the *count* instruction a count is maintained in the address part of the instruction of how many times this instruction has been carried out in the programme; after a number of times that can be set in advance the next instruction is skipped. This is a means of getting out of a cycle. One might imaginc, for example, that the count instruction is followed by an instruction: "jump back to the beginning of the cycle"; then this jump backwards will initially be carried out, until after a preset number of times it is omitted.

In the *repeat* instruction the I_1 half of the instruction pair is carried out a preset number of times before the programme continues. In this way one can, for example, carry out a multiplication by 3 by combining a repeat instruction, set to 3 times, with an add instruction in the same instruction-pair. It costs half a word in the memory space, but in fixed-point notation it produces a gain in speed, in view of the fact that three such additions take less time than one multiplication.

Finally let us here discuss the special instructions *link-I* and *link-II* that serve for introducing a subroutine.

A subroutine is a programme that on its own forms a more or less complete entity and that because of its utility in many calculations is cast in a standard form. Such programmes are generally used for the calculation of, for example, sin x, $\sqrt{x}$, log x, and e^x. Usually the subroutines are in stock on a magnetic tape or a punched tape, and it must be possible to insert them into a programme at the place where they are needed. There are, however, advantages in placing a subroutine in an arbitrary row of addresses in the memory, and to include in the (main) programme at whatever position is necessary a jump instruction for jumping to the initial address of the row. After the subroutine has been carried out the machine must then jump back to the place where the main programme was interrupted. It must be possible to use the subroutine in different places of the same main programme, and the instruction which controls the jump back to the main programme must take this into account. Now this is automatically ensured by the link-I instruction. The address part of this instruction contains the address k, say. The subroutine proper begins at $k+1$; the link instruction now puts the subroutine into operation by jumping to $k+1$ and simultaneously places a jump instruction at k, indicating the return jump to the correct part of the main programme. It is then only necessary to make sure that each subroutine will finally jump back to its first address (k); there the machine finds the jump instruction bringing it back to the correct part of the main programme.

Even more flexible in use is the *link-II* instruction. This instruction refers to a subroutine via an "address book". At a fixed position in the memory a small list is made of places where subroutines are stored. This is done in the form of jump instructions referring to their respective initial addresses. In all, 30 subroutines may occur. The link-II instruction now puts into action a subroutine by directing the machine to the relevant number in the small list, after which the transition to the subroutine and the return to the main programme occur completely automatically.

In order to illustrate how a link-II works, we may mention that among the 64 possible instructions, with the operation numbers 0 to 63 given by 6 bits, there are 32 that do not use their 11 address bits in their proper meaning of "position of a number to be fetched or stored". These are instructions in which no number is processed, e.g. the jump instructions, the *DTC* and *CTD* instructions, and the count and repeat instructions. For such instructions an address-interpretation bit equal to 1 would have no sense. If they do however have an address-interpretation bit equal to 1, these instructions all (except the count and repeat instructions) represent link-II instructions in which the operation number indicates directly the number of the subroutine in the list. The advantage of the link-II instructions is that the programmer has at his disposal another 30 operation numbers with which he can indicate computing procedures of his own choice while, just as with normal instructions, he still has room for indicating the number to be processed. Thus one programmer might use these instructions to indicate, for example, an addition, a subtraction, or a multiplication of complex numbers with floating point, while another might perhaps use them for addition, multiplication etc. of matrices.

The parity check

The reader will probably have noticed that in what was mentioned above we have spoken of a word length of 42 bits, but that in Table I mention was made of a total word length of 44 bits. This difference is due to the parity bits. These digits, one per half-word, provide a check on the operation of the machine. They have no significance in either the number or in the instruction, but they ensure that the total number of ones in the relevant half-word is odd.

By means of a special device in the M register the number of ones is investigated and a parity bit (1 or 0) added: each word going from the M register to the memory is thus provided with two appropriate parity bits. If a word from the memory enters the M register, the number of ones is checked again. If this is now found to be even, an error is present and the machine stops. If there are two (or an even number of) errors per half-word, they remain undiscovered; the chance of more than one fault occurring is, however, very small.

Recording on magnetic tape — in which the chance of errors is generally greater — is checked by a number of extra bits in addition to these two parity bits. The 44 bits of a word are recorded in four rows on 14 tracks of the tape. The remaining 12 positions are taken up by check bits.

Input and output facilities in the PASCAL

Punched tape, punched cards, or magnetic tape may be used for the input of the PASCAL. For the output, use can be made of an electric typewriter, a tape punch, a card punch, a line printer and magnetic tape equipment.

The typewriter is much slower than the other output devices. It is therefore used mainly to write out instructions for the operating staff (who need not know what the computing problem was) as indicated by the programme, e.g. to tell them that a new roll of tape must be inserted in one of the magnetic-tape units of the memory.

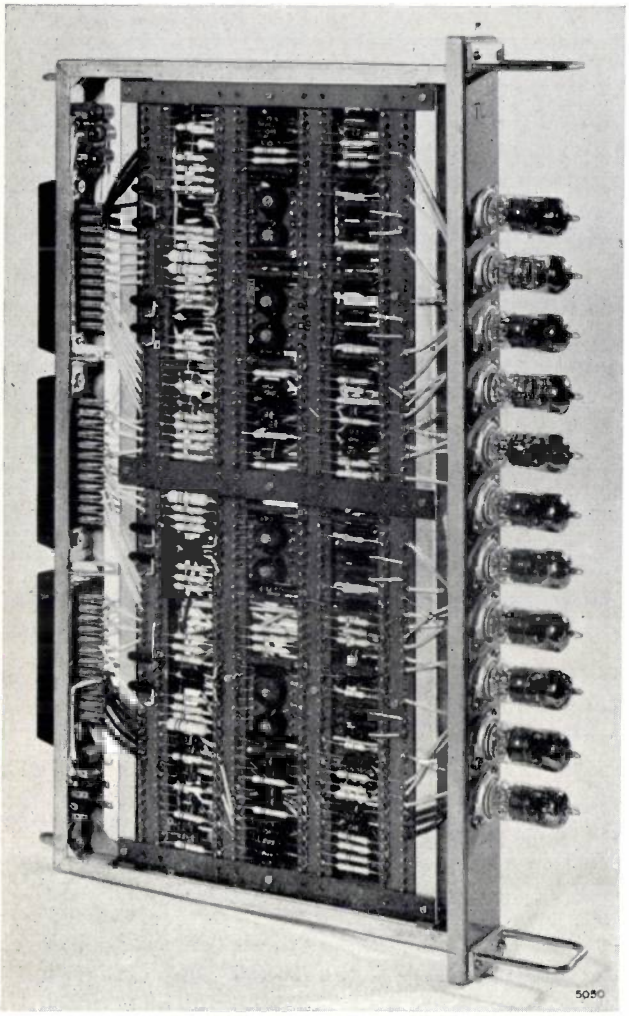


Fig. 14. A circuit panel of the arithmetical unit. This interchangeable element contains all the circuits corresponding to one digit in the arithmetical unit. It embraces the flip-flops of the registers A, M and S, the addition circuit *Add* and the six-way switch Σ. As the numbers to be used have 42 binary digits, the whole arithmetical unit contains 42 of these circuit panels.

The tape punch, the card punch and the magnetic tape are used if the results of computation have to be processed again later on by the machine. If the results must be made available in readable form, use is made of a line printer, which prints a whole line of 90 characters simultaneously at a rate of $2\frac{1}{2}$ lines per second. This speed is frequently still too low for printing the results during computation. It that case a magnetic-tape unit can be used for recording the information very quickly. A separate apparatus can then read these data in groups from the tape and supply them to a fast line printer that prints 10 lines of 120 symbols per second.

Special constructional features of the PASCAL

The PASCAL is built up of interchangeable circuit panels. The size of these elements is determined by the arithmetical unit, since the components of the registers A, M and S, the adder and the six-way switch Σ referring to a single digit are assembled together (*fig. 14*). The arithmetical unit thus has 42 identical elements. An advantage of this system is that not many connections are required among the various circuit panels, so that the wiring of the machine can remain fairly simple for such a complicated apparatus. The size of the circuit panels was also found to suit the rest of the machine extremely well, so that the PASCAL has in all 125 elements, all of the same dimensions and all accommodated in a cabinet 50×190×270 cm. This cabinet also houses the stabilized voltage-supply unit, the

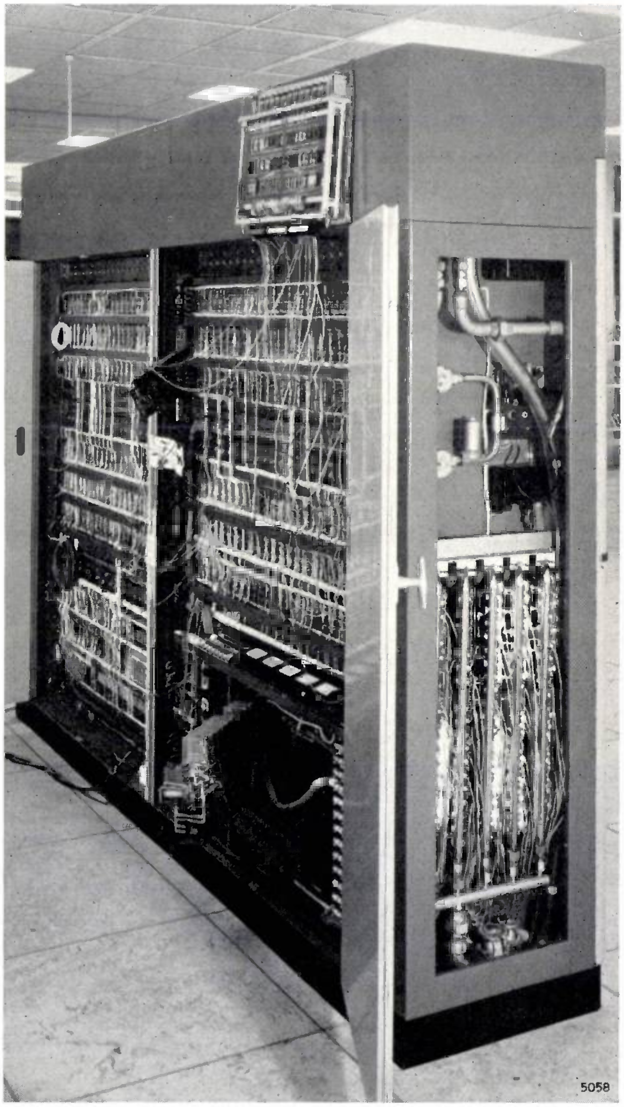


Fig. 15. Rear view of the cabinet in which the principal components of the PASCAL are housed (see fig. 2), during testing of the machine. The doors are open, so that the wiring can be seen. A number of power transistors are visible at the side of the cabinet; these serve for the stabilization of the supply voltages. For reasons of cooling they are mounted on metal strips fixed to tubes in which water flows. The pilot lights at the end of the cabinet are each in series with a transistor and assure an equal distribution of the current over the transistors. They also serve as fuses.

magnetic-core memory and the modification memory. *Fig. 15* shows the cabinet from the rear with the doors opened, so that the wiring can be seen; *fig. 16* shows the front of the cabinet, likewise with the doors opened. In view of the fact that, when the machine was designed, suitably fast transistors were not yet available for a number of functions, the PASCAL contains some 1200 valves. The number of transistors amounts to about 10 000, and the number of germanium diodes to about 15 000.

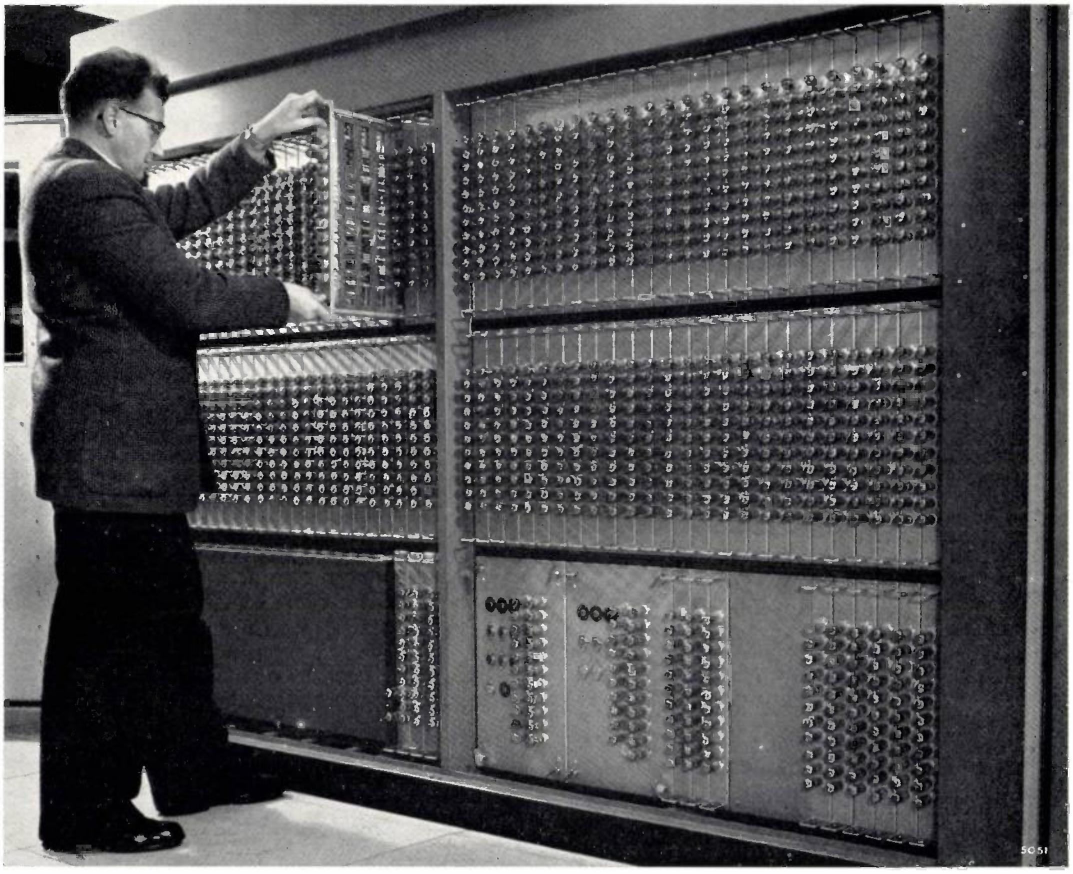


Fig. 16. Front of the cabinet in which the principal components of the PASCAL are housed (cf. fig. 2), with the doors opened. The uppermost row of circuit panels (cf. fig. 14) forms the arithmetical unit. The centre row forms the control unit. Below, from left to right: the power supply unit, the magnetic-core memory and the modification memory. When the cabinet is closed, the heat developed therein is removed by circulating air which is cooled by tap water.

The heat developed in the cabinet amounts to some 10 kW. When the front doors are closed, the cabinet forms a sealed unit. Air circulating inside takes up the heat and is cooled by tap water in a heat exchanger. The opening of the doors at the back does not interfere with the air circulation, so that the wiring and all points that are important for maintenance remain accessible while the machine is working. In fig. 15 can be seen how the power transistors for the stabilization of the supply voltages (at the side of the cabinet) are cooled. They are mounted on metal strips that form a unit with the tubes through which the tap water flows.

The operation of the PASCAL is controlled entirely from the control desk (see figs. 2 and 11). An upright panel contains the switches for starting and stopping the machine as well as the switches for slow operation, sometimes required when it is necessary to follow the various operations visually during maintenance. Pilot lights on the left-hand half of the panel indicate the contents of the *S*, *A* and *M* registers, and ones on the right-hand half indicate the contents of the control register *C* and of *T*. The lowest row of pilot lights on the left shows continuously the contents of an arbitrary memory address that can be selected by means of switches on the right of the panel. Further pilot lights show the state of the two flip-flops for the comparison instruction (see page 14), the state of affairs of transports in the course of being carried out, and the occurrence of any parity errors. A meter in the centre of the panel indicates during what percentage of the time the machine is actually "computing"; the remaining time is used in waiting for transport or for input and output. The plugboard memory is situated in a recess sunk into the top of the control desk proper, while the push-buttons for the address 0 are at the bottom right-hand side of the control panel.

On the far right of the desk there is the punched-tape reader for the input of programmes, and on the left there is the electric typewriter. The tape punch is accommodated in a drawer of the desk.

As already mentioned, the PASCAL and the STEVIN with their addition time of 10 μsec belong to the class of "fast" machines; in various parts of the world there are however already designs ready for machines with a computing speed ten times as great [4]). On the other hand, most machines now in general use have a computing speed about ten times smaller and also a smaller memory.

Modern commercial computers that are comparable with the PASCAL as regards computing speed generally have a more extensive magnetic-core memory. The disadvantage of the PASCAL in this respect is however more or less counterbalanced by the rapid transport between the drum memory and the magnetic core memory. Only when there are less than 8 short operations (such as additions) per word transported will such transport produce a serious lowering of the computing speed.

The STEVIN — identical with the PASCAL — works more rapidly than is required for the solution of most administrative problems. The high speed compensates for the lack of a few features that are usually found in machines designed specifically for administrative work.

Summary. The digital electronic computer PASCAL developed in the Philips laboratories in Eindhoven is a binary machine of the single-address type with a word length of 42 bits (either a number or two instructions are given per word). The machine has a magnetic-core memory as a fast working memory, containing 2016 words, and a drum memory as a slow memory with 16 384 words, as well as several magnetic-tape units on which 10^6 words can be stored per tape. In addition there is a small memory, in which 15 words can be inserted manually by plugs and one word by push-buttons, and a modification memory of 8 half-words.

Computation can be performed in either the fixed-point or the floating-point notation; in the latter the fractional part has a length of 33 bits plus a sign bit, and the exponent has a length of 7 bits plus a sign bit (accuracy of 10 decimal places). In fixed-point notation (accuracy 12 decimal places) an addition requires about 10 μsec and a multiplication about 71 μsec. This great speed is attained by means of circuits which determine the carries in the adder very rapidly and by a very efficient organization of the traffic in the arithmetical unit with the help of a special circuit (the 6-way Σ switch). On the average 60 000 instructions are processed per second.

The input of the programmes and the output of the results can take place by means of punched tape, punched cards or magnetic tape. A line printer can also be used for the output.

Before an instruction, coded in the form of a number, is carried out, it can be modified by the addition of a word from one of eight modification registers. Thereafter the instruction may be replaced in the memory either in its original or its modified form. The PASCAL has several special instructions, e.g. the count and repeat instructions, and the link instructions, which assist in the transition to subroutines stored at arbitrary sites in the memory. There are also special instructions for the transport of data from the drum memory to the magnetic-core memory and vice versa. Such transport can take place simultaneously in both directions while the machine continues computing, albeit with a slightly reduced speed (this reduction amounts at the most to 15%).

After a short discussion of calculations with the binary system, this article gives a survey of the principal components contained in all digital electronic computers, leading on to a general survey of the working of such computers. The abovementioned special features of the PASCAL are discussed in the second part of the article. After a description of certain constructional details of the PASCAL, such as the cooling (it contains 1200 valves, 10 000 transistors and 15 000 diodes), some concluding remarks are made about the PASCAL in comparison with other machines.

[4]) A summary of digital electronic computers that are at present available commercially or in a state of development will be found, for example, in: Martin H. Weik, A survey of domestic electronic digital computing systems, United States Department of Commerce, Office of Technical Services, 1955; Isaac L. Auerbach, European electronic data processing — A report on the industry and the state-of-the-art, Proc. Inst. Radio Engrs. 49, 330-348, 1961 (No. 1).

NEW DEVELOPMENTS IN OXIDE-COATED CATHODES

I. OXIDE-COATED CATHODES FOR LOADS OF 1 TO 2 A/cm^2

II. AN OXIDE-COATED CATHODE WITH A HALF-WATT HEATER FOR CATHODE-RAY TUBES

The first of these articles deals with oxide-coated cathodes designed to be loaded with current densities several times higher than ordinary oxide-coated cathodes can withstand, and which yet have a life of a few thousand hours. Although these new cathodes cannot rival the L type of dispenser cathode as regards their permissible current density, they are cheaper and have a lower operating temperature, which removes the problem of insulation between heater and cathode.

Where cathode-ray tubes are used in fully transistorized, battery-fed equipment, such as television receivers, television cameras, oscilloscopes, etc., low power consumption by the heater is an important consideration. The second article concerns an oxide-coated cathode developed for applications of this kind, whose heater consumes only 0.5 W.

I. OXIDE-COATED CATHODES FOR LOADS OF 1 TO 2 A/cm^2

by H. J. LEMMENS *) and P. ZALM *).

621.3.032.213.23:621.3.032.213.63/.64

In specifications of the maximum permissible loading of oxide-coated cathodes in vacuum tubes it is normally stated that the average current density at the cathode surface should not exceed 0.5 to 1 A/cm^2 [1]). This limitation is due to the resistance of the oxide coating and in some cases also to the interface resistance in the cathode. With increasing load the dissipation I^2R in the oxide coating rises in proportion to the square of the thermionic current I, whereas the power which the emitted electrons drain from the cathode rises only in direct proportion to I, being equal to

$$I(\varphi + V_{\mathrm{m}} + \frac{2kT}{e}),$$

where φ is the work function, V_{m} the potential barrier (a few tenths of a volt) due to the space charge, k is Boltzmann's constant, T the absolute temperature of the cathode, and e the charge of the electron. Above a certain current I, therefore, I^2R is greater than the power drained off by the electrons. The emissive layer can become overheated under these conditions; the thermionic current will then no longer be constant and the oxide coating will be damaged [2]).

For loads greater than 1 A/cm^2, dispenser cathodes [3]) are generally preferred; the emissive surface of these cathodes consists of porous tungsten, activated by adsorption of Ba and BaO. The latter are either supplied from a chamber behind the porous tungsten body (L cathode [4])) or are incorporated in the body itself (impregnated cathode [5])). In these dispenser cathodes the *metallic* character of the emitter predominates. The emissive layer is less than 1 μ thick, and consequently has a very low resistance R. Dispenser cathodes have two drawbacks, however. The first is their relatively high price (due to their rather complicated construction), which limits their use to special tubes. The second is that they have to operate at higher temperatures than oxide cathodes (at 1050-1100 °C as against 740-830 °C), which is a point of importance where low heater consumption is required. Moreover, the higher temperature causes difficulties with the insulation between heater and cathode.

*) Research Laboratories, Eindhoven.

[1]) D. A. Wright, A survey of present knowledge of thermionic emitters, Proc. Instn. Electr. Engrs., Part III, 100, 125-142, 1953.
A. H. W. Beck, High-current-density thermionic emitters: a survey, Proc. Instn. Electr. Engrs., Part B, 106, 372-390, 1959.

[2]) In considering the maximum permissible load we are concerned only with its influence on the emission of the cathode, and not with the limits set to this load by the effect which the resistance of the cathode has on the electron-tube characteristic.

[3]) A. Venema, Dispenser cathodes, I. Introduction, Philips tech. Rev. 19, 177-179, 1957/58.

[4]) H. J. Lemmens, M. J. Jansen and R. Loosjes, A new thermionic cathode for heavy loads, Philips tech. Rev. 11, 341-350, 1949/50.

[5]) R. Levi, Dispenser cathodes, III. The impregnated cathode, Philips tech. Rev. 19, 186-190, 1957/58.

Functions of the oxide coating

On the face of it, the obvious way of solving the resistance problem of oxide-coated cathodes would be to make their emissive layer very thin too (e.g. between 5 and 20 μ), particularly since it is known that the work function of cathodes with such a thin oxide coating is the same as that of cathodes with the usual layer thickness of about 80 μ. The solution fails, however, if a reasonably long life is required of the cathode. An examination of the functions of the oxide coating will make this immediately evident. Its primary function, of course, is to emit electrons. Scarcely less important, however, is its function as the "storehouse" for replenishing the emissive surface [6]). As various investigations have shown [7]), in conventional cathodes coated with barium-strontium oxide the emissive layer, formed by minute crystals of (Ba,Sr)O, consists of virtually pure SrO at the boundary between oxide and vacuum. At the surface — and perhaps inside the crystals too — this SrO layer is activated by Ba and governs the work function of the cathode. As we shall presently see, Ba is used up during emission (some reacting to give the silicate and some being released as metal and evaporating); continuous replenishment of Ba is therefore necessary, and this is ensured by the reaction of BaO with the activators in the nickel of the cathode (Al, Mg, Si, etc.) [8]), and probably also by electrolysis of the oxide coating [6]).

The separation of these two functions — thermionic emission and barium replenishment — has led to the development of two types of oxide-coated cathode for heavy loads:

a) a "storage" cathode, for current densities from 1.5 to 2 A/cm^2, and

b) a double-coated cathode, for current densities from 1 to 1.5 A/cm^2.

The storage oxide cathode

One possible construction of a storage type of oxide-coated cathode is shown in *fig. 1*. A nickel cylinder *1* is pressed in at the top so as to form a chamber *2*, which is filled with a reserve supply of, say, (Ba,Sr)O. The chamber is capped by a piece of nickel gauze *3*, to which the emissive oxide coating *4* is applied.

The object of this design is to eliminate the above-mentioned drawbacks of normal oxide-coated cathodes: the formation of an interface resistance, the accompanying change in the composition of the emissive layer during operation, and excessive resistance in the oxide coating.

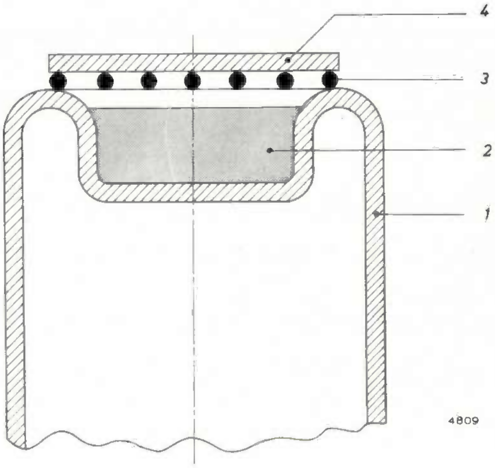


Fig. 1. Cross-section of a storage oxide cathode. *1* nickel cylinder. *2* storage chamber, filled with e.g. (Ba,Sr)O. *3* nickel gauze (exaggerated for clarity). *4* emissive oxide coating.

Avoidance of interface resistance

In normal oxide-coated cathodes the nickel serves a dual purpose: it is the base for the emissive layer and the source of the activating elements, Al, Mg, Si, etc. A consideration affecting the choice of the composition of the cathode nickel is that the products of the reaction between the activators and Ba should not give rise to excessive resistance between the nickel and the oxide. On the other hand, sufficient Ba must be produced for the emissive layer. With the design illustrated in fig. 1, interface resistance between gauze and oxide can be avoided if the composition of the nickel used for the gauze is suitably chosen, the barium for activating the emissive layer being drawn from the chamber through the reaction of its contents with activators in the nickel cylinder and with any metallic powders that may have been added. If the nickel contains Si as activator, the fillings found suitable are:
(Ba,Sr)O, possibly mixed with 3% by weight of Ti (added as titanium hydride), and
$BaAl_2O_4$ with 25% by weight of Ti.

The reaction between BaO and Ti is

$$BaO + \tfrac{1}{2}Ti \rightleftarrows Ba\,(g) + \tfrac{1}{2}TiO_2\,, \qquad (1)$$

where (g) indicates that the barium is produced in

[6]) L. S. Nergaard, The physics of the cathode, R.C.A. Rev. **18**, 486-511, 1957.

[7]) H. Gaertner, Electron diffraction on oxide-coated filaments, Phil. Mag. **19**, 82-103, 1935.
J. A. Darbyshire, Diffraction of electrons by oxide-coated cathodes, Proc. Phys. Soc. London **50**, 635-641, 1948.
H. Huber and S. Wagener, Die kristallographische Struktur von Erdalkaligemischen. Untersuchungen mit Hilfe von Röntgen- und Elektronenstrahlen an Oxydkathoden, Z. techn. Phys. **23**, 1-12, 1942.

[8]) E. S. Rittner, A theoretical study of the chemistry of the oxide cathode, Philips Res. Repts **8**, 184-238, 1953.

gaseous form. The equilibrium vapour pressure p_{Ba} of the barium (in mm Hg) is given for this reaction by

$$\log p_{\mathrm{Ba}} = -\frac{14\,000}{T} + 8.72$$

(with T in °K), and for the reaction

$$\mathrm{BaAl_2O_4} + \tfrac{1}{2}\mathrm{Ti} \rightleftarrows \mathrm{Ba\,(g)} + \mathrm{Al_2O_3} + \tfrac{1}{2}\mathrm{TiO_2} \quad (2)$$

it is given by

$$\log p_{\mathrm{Ba}} = -\frac{19\,400}{T} + 7.6.$$

This means that, at an operating temperature of $T = 1050$ °K, the maximum Ba vapour pressure in reaction (1) is 2.5×10^{-5} mm Hg, and in reaction (2) only 1.6×10^{-11} mm Hg. These are important data for the application of storage cathodes. Where electron tubes are concerned whose vacuum conditions are not all too good, e.g. because thorough degassing of the metal parts is not readily possible, some generation of barium vapour is an advantage, in that the vapour counteracts poisoning of the cathode by the residual gases. In high-vacuum tubes, on the other hand, the lowest possible Ba vapour pressure may be desirable to avoid grid emission caused by barium deposits on the grid.

The formation of interface resistance is not the only adverse consequence of the reaction of the activator in the cathode nickel with the BaO; another is a change in the composition of the oxide coating. If the cathode nickel contains, for example, 0.1% by weight of Si as activator, more than 10% of the BaO present may react with the Si to form $\mathrm{Ba_2SiO_4}$[9]) at the interface of the nickel and the oxide coating. (This estimate relates to a coating of 10 mg of (Ba,Sr)O per cm^2 on nickel 100 μ thick.) Moreover, in this case just as much Ba will evaporate as will form silicate by the reaction:

$$4\,\mathrm{BaO} + \mathrm{Si} \rightleftarrows \mathrm{Ba_2SiO_4} + 2\,\mathrm{Ba\,(g)}\,.$$

Minimizing the resistance of the oxide coating

The resistance of the oxide layer has already been mentioned as one of the drawbacks of normal oxide-coated cathodes. The emissive coating of a storage cathode, however, can be relatively thin, and thus its resistance low. Higher permissible loading makes it possible to operate this cathode at a lower temperature than ordinary (Ba,Sr)O-coated types. This reduces the tendency of the oxide coating to lose its porosity through sintering. Since the conduction of the emissive layer is largely governed by electron transport through the pores between the crystals [10]), this again has the effect of keeping down the resistance of the coating. In addition, a more open structure will ensure better dispensation of Ba to the boundary face between oxide and vacuum.

A further way of minimizing the sintering process is to use another oxide instead of (Ba,Sr)O. Substances which have proved suitable are SrO, (Ca,Sr)O and CaO. During operation, coatings of these substances do not lose their porosity to the same extent as (Ba,Sr)O, and consequently the increase in their resistance is less marked.

Results with storage oxide cathodes

It will be clear from the foregoing that there is no essential difference between conventional oxide-coated cathodes and the storage types described. The latter, however, have more favourable characteristics. Before demonstrating these, we shall recall a dictum of Nergaard: "Every cathode is in equilibrium with its environment" [6]). It is possible, for example, to subject a normal (Ba,Sr)O cathode for a long period of time to a continuous load of 1 A/cm^2, provided the residual gases have been sufficiently removed by careful degassing. The nickel used for such a cathode cylinder can be given a low activator content, as a result of which the composition of the oxide coating remains fairly constant during the life of the cathode and no appreciable interface resistance is found. A storage cathode, however, where the cylinder consists of Si-activated nickel, the filling of (Ba,Sr)O, and the emissive coating on the gauze is also (Ba,Sr)O, can have a life of more than 5000 hours under a constant load of 1.5 A/cm^2 (curves *a* and *b* in *fig. 2*), in conditions which a normal oxide-coated cathode is unable to withstand for as much as 200 hours (curve *c* in fig. 2).

Cathodes whose emissive layer consists of SrO, CaO or mixed crystals of (Ca,Sr)O differ little in their characteristics from types coated with (Ba,Sr)O; the work function of cathodes coated with SrO is almost the same, with CaO slightly higher and with (Ca,Sr)O perhaps somewhat lower. The characteristics of the latter type are optimum when the mole fraction of CaO is 60%.

[9]) X-ray and chemical analyses have shown the composition of the interface to be $(\mathrm{Ba_{0.8},Sr_{0.2}})_2\mathrm{SiO_4}$.

[10]) R. Loosjes and H. J. Vink, The conduction mechanism in oxide-coated cathodes, Philips Res. Repts 4, 449-475, 1949.
R. Loosjes and H. J. Vink, Conduction processes in the oxide-coated cathode, Philips tech. Rev. 11, 271-278, 1949/50.

The work function of cathodes can in principle be found from Richardson's equation:

$$J_s = AT^2 \exp\left(-\frac{e\varphi}{kT}\right),$$

where J_s is the saturation current density. The emissive surface of an oxide-coated cathode, however, is not homogeneous. The work function, therefore, is not everywhere identical and in practice is expressed by a suitably weighted mean. The values given below were found either from the slope of the line giving ln J_s/T^2 as a function of $1/T$, or by inserting the value 120 A/cm^2(°K)^2 for A. By the first method the work function of cathodes coated with (Ca,Sr)O is found to be 1 eV; the second method gives 1.37 eV at 900 °K and 1.42 eV at 1020 °K.

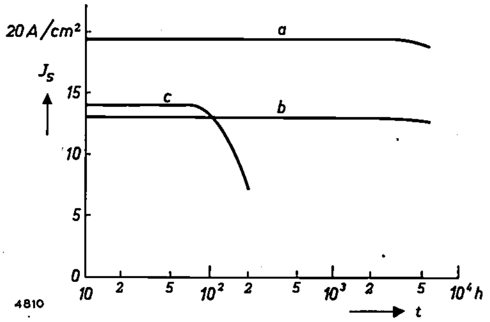


Fig. 2. Saturation current density J_s of various oxide-coated cathodes, as a function of the time t during which they were continuously loaded at 1.5 A/cm^2. Curves a and b relate to storage cathodes, as illustrated in fig. 1, operated at 780 °C and 730 °C, respectively. Diameter of emitting surface 1.5 mm, chamber contains (Ba,Sr)O mixed with 3% by weight of Ti, coating is pure (Ba,Sr)O. Curve c relates to a normal oxide-coated cathode, operated at 820 °C.

The double-coated cathode

After what has been said about the storage type of oxide-coated cathode, the principle of the double-coated type can be stated quite simply. It differs from the conventional type in only one respect. Ordinarily the Ba activation is due to the BaO reacting with an activator originating from the cathode nickel; in the double-coated type the activator is entirely or partly incorporated in the oxide layer, Ti or Zr being added as hydrides to the (Ba,Sr)CO$_3$ sprayed on to the nickel base. Since the presence of the activator in the entire emissive layer proves to have an adverse effect on the emission characteristics, a second coating containing no activator is sprayed on to the first. This second coating may consist of (Ba,Sr)CO$_3$, (Ca,Sr)CO$_3$, (Ba,Ca,Sr)CO$_3$, etc. After decomposition of the carbonates and activation, the cathode obtained is again essentially a storage type, but in this case the barium-producing layer and the emissive layer are in direct contact with one another.

The characteristics of the double-coated cathode lie between those of the ordinary oxide-coated type and those of the storage type. At a current density of 1.5 A/cm^2 it has a life of between about 1200 and 1500 hours, and at 1 A/cm^2 between about 2000 and 2500 hours. Its performance is thus inferior to that of the storage cathode, but it has the compensatory advantages of simpler construction and a lower price.

II. AN OXIDE-COATED CATHODE WITH A HALF-WATT HEATER FOR CATHODE-RAY TUBES

by F. H. R. ALMER *) and A. KUIPER *).

621.3.032.213.6

The advent of the transistor and the crystal diode has led, or is leading, to the development of portable battery-fed electronic equipment containing cathode-ray tubes, such as television receivers, television cameras and oscilloscopes. The power consumed by the C.R.T. heater in such equipment constitutes a heavy load on the battery. The cathodes in standard, mains-operated picture tubes in America, for example, use a 4-W heater. Philips C.R.T. cathodes consume only about 2 W (6.3 V × 0.3 A), but even that is a severe drain on a battery.

There is thus a need for a cathode for cathode-ray tubes of equivalent performance that will consume considerably less than 2 W. In the following a cathode is discussed for which the heater power is only 0.54 W at a heater voltage of 6.3 V. This cathode is now fitted in oscilloscope tubes (type DH 7-11) and in vidicons (type 55 850). The design described offers prospects of using higher heater voltages, e.g. 12 V, which in many cases is the most favourable supply voltage for transistors.

*) Electron Tubes Division, Eindhoven.

The work function of cathodes can in principle be found from Richardson's equation:

$$J_s = AT^2 \exp\left(-\frac{e\varphi}{kT}\right),$$

where J_s is the saturation current density. The emissive surface of an oxide-coated cathode, however, is not homogeneous. The work function, therefore, is not everywhere identical and in practice is expressed by a suitably weighted mean. The values given below were found either from the slope of the line giving ln J_s/T^2 as a function of $1/T$, or by inserting the value 120 A/cm²(°K)² for A. By the first method the work function of cathodes coated with (Ca,Sr)O is found to be 1 eV; the second method gives 1.37 eV at 900 °K and 1.42 eV at 1020 °K.

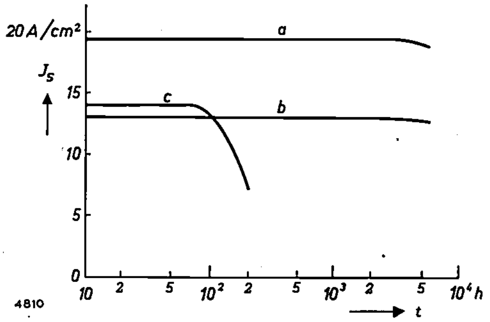


Fig. 2. Saturation current density J_s of various oxide-coated cathodes, as a function of the time t during which they were continuously loaded at 1.5 A/cm². Curves a and b relate to storage cathodes, as illustrated in fig. 1, operated at 780 °C and 730 °C, respectively. Diameter of emitting surface 1.5 mm, chamber contains (Ba,Sr)O mixed with 3% by weight of Ti, coating is pure (Ba,Sr)O. Curve c relates to a normal oxide-coated cathode, operated at 820 °C.

The double-coated cathode

After what has been said about the storage type of oxide-coated cathode, the principle of the double-coated type can be stated quite simply. It differs from the conventional type in only one respect. Ordinarily the Ba activation is due to the BaO reacting with an activator originating from the cathode nickel; in the double-coated type the activator is entirely or partly incorporated in the oxide layer, Ti or Zr being added as hydrides to the (Ba,Sr)CO_3 sprayed on to the nickel base. Since the presence of the activator in the entire emissive layer proves to have an adverse effect on the emission characteristics, a second coating containing no activator is sprayed on to the first. This second coating may consist of (Ba,Sr)CO_3, (Ca,Sr)CO_3, (Ba,Ca,Sr)CO_3, etc. After decomposition of the carbonates and activation, the cathode obtained is again essentially a storage type, but in this case the barium-producing layer and the emissive layer are in direct contact with one another.

The characteristics of the double-coated cathode lie between those of the ordinary oxide-coated type and those of the storage type. At a current density of 1.5 A/cm² it has a life of between about 1200 and 1500 hours, and at 1 A/cm² between about 2000 and 2500 hours. Its performance is thus inferior to that of the storage cathode, but it has the compensatory advantages of simpler construction and a lower price.

II. AN OXIDE-COATED CATHODE WITH A HALF-WATT HEATER FOR CATHODE-RAY TUBES

by F. H. R. ALMER *) and A. KUIPER *).

621.3.032.213.6

The advent of the transistor and the crystal diode has led, or is leading, to the development of portable battery-fed electronic equipment containing cathode-ray tubes, such as television receivers, television cameras and oscilloscopes. The power consumed by the C.R.T. heater in such equipment constitutes a heavy load on the battery. The cathodes in standard, mains-operated picture tubes in America, for example, use a 4-W heater. Philips C.R.T. cathodes consume only about 2 W (6.3 V × 0.3 A), but even that is a severe drain on a battery.

There is thus a need for a cathode for cathode-ray tubes of equivalent performance that will consume considerably less than 2 W. In the following a cathode is discussed for which the heater power is only 0.54 W at a heater voltage of 6.3 V. This cathode is now fitted in oscilloscope tubes (type DH 7-11) and in vidicons (type 55 850). The design described offers prospects of using higher heater voltages, e.g. 12 V, which in many cases is the most favourable supply voltage for transistors.

*) Electron Tubes Division, Eindhoven.

The power taken up by the heater is dissipated partly by radiation and partly by thermal conduction through the heater terminals and the cathode supports. Both the radiation and the conduction must be reduced to achieve any appreciable reduction in heater power. Which of the two is reduced more than the other is a matter of some importance, as will be shown below.

One of the main points to be observed in the design of a cathode is the possible spread in its working temperature. Below a certain minimum temperature the thermionic emission is too small; above a certain maximum temperature the life of the cathode is shortened and the emissive material will evaporate to such an extent as to endanger the insulation between the electrodes. The cathode temperature may deviate from the desired value for the following reasons:

1) The heater power is incorrect,
   a) because the supply voltage differs from the rated value,
   b) because of unavoidable differences in the heaters themselves.
2) The heat balance of the heater + cathode assemblies shows a statistical spread for a given heater power.

The reason mentioned under 1(*a*) — different supply voltage — applies especially to cathodes fed from batteries, whether accumulators or dry cells, the voltage of which shows relatively much greater variations than most AC mains.

Other requirements to be met by such cathodes are:

1) the ability to withstand shocks and vibrations,
2) good insulation between heater, cathode and Wehnelt cylinder,
3) the ability to heat up quickly, so that the television picture becomes visible very soon after switching on.

These questions are discussed below.

Reasons for spread in cathode temperature

Variations of the heater voltage

The influence of variations in the heater voltage V_f on the temperature T_k of the cathode can be subdivided into

a) the effect of variations in V_f on the power P_f taken up by the heater [1]), and

b) the effect of variations in P_f on the temperature T_k.

If we call the heater current I_f, then $P_f = V_f I_f$, and so

$$\frac{\mathrm{d}P_f}{P_f} = \frac{\mathrm{d}V_f}{V_f} + \frac{\mathrm{d}I_f}{I_f}.$$

Let c be the ratio of the differential resistance $\mathrm{d}V_f/\mathrm{d}I_f$ of the heater to the resistance V_f/I_f, then

$$\frac{\mathrm{d}I_f}{I_f} = \frac{1}{c}\frac{\mathrm{d}V_f}{V_f},$$

and therefore

$$\frac{\mathrm{d}P_f}{P_f} = \left(1 + \frac{1}{c}\right)\frac{\mathrm{d}V_f}{V_f}. \quad \dots . \quad (1)$$

The relative variation of the heater power for a given relative variation of the heater voltage thus decreases as the factor c increases. In well-designed indirectly heated cathodes the value of c is between 1.6 and 2.0, depending on the heater temperature and the construction, and the factor $(1 + c^{-1})$ is therefore between 1.63 and 1.50. Here, then, there is not much chance of improvement.

The effect of variations in P_f on the cathode temperature T_k depends to a great extent on how the power P_f is removed from the cathode. The two processes responsible for its removal are radiation and thermal conduction (via the cathode supports and the pins which make the electrical connections). Let us first consider a cathode whose power is almost entirely dissipated by conduction. As an approximation we may then write:

$$T_k - T_0 \propto P_f,$$

where T_0 is the ambient temperature. Turning to the other extreme, we may write for a cathode which dissipates its heat almost entirely by radiation:

$$P_f = C(T_k)\,(T_k^4 - T_0^4),$$

where C is a factor dependent on the temperature. In an oxide-coated cathode T_k is roughly 1070 °K and T_0 about 350 °K, which means that T_0^4 is barely 1% of T_k^4. In our following considerations we shall therefore disregard T_0^4 in relation to T_k^4. Furthermore, in a limited temperature range around the nominal cathode temperature the factor C can be replaced by a constant factor and the exponent 4 by a larger exponent n; we then have $T_k^n \propto P_f$, so that

$$T_k \propto P_f^{1/n}.$$

Evidently, then, in the case of a cathode which is a pure radiator T_k depends much less on P_f than in the other case, where all heat is dissipated by conduction. We therefore conclude that, in order to

[1]) We consider here only the most unfavourable case where no other heaters or resistors are connected in series with the heater, i.e. where V_f is equal to the supply voltage.

minimize the dependence of T_k on V_f, *the dissipation of the heat by conduction should be limited as much as possible in favour of its removal by radiation.*

Let the power removed from the cathode by radiation be βP_f; that removed by conduction is then $(1-\beta)P_f$. The heat balance can therefore be written:

$$\begin{aligned} P_f &= \beta P_f \;\; + (1-\beta)P_f \\ &= k_1 T_k^{\,n} + k_2(T_k - T_0)\,, \quad . \; . \quad (2) \end{aligned}$$

where k_1 and k_2 are proportionality factors.

Again in a limited temperature range (e.g. from 1000 to 1200 °K) we may write:

$$P_f \propto T_k^{\,a}\,. \quad . \; . \; . \; . \; . \; . \quad (3)$$

The exponent a can be determined by measuring T_k on the cathode as a function of P_f. We can then calculate the magnitude of the fractions β and $1-\beta$ for the cathode in question, i.e. the fractions of P_f that are dissipated by radiation and conduction respectively, using the following relationship between a and β:

$$\beta = \frac{a - \dfrac{T_k}{T_k - T_0}}{n - \dfrac{T_k}{T_k - T_0}}\,. \quad . \; . \; . \; . \; . \quad (4)$$

The exponent n is only slightly temperature-dependent; e.g. for nickel, $n \approx 4.6$ in the entire temperature range from 600 to 1200 °K. Disregarding $\mathrm{d}n/\mathrm{d}T_k$ therefore, we find from (2):

$$\frac{\mathrm{d}P_f}{\mathrm{d}T_k} = n\,k_1\,T_k^{\,n-1} + k_2 = \frac{n}{T_k}\,\beta P_f + \frac{1-\beta}{T_k - T_0}\,P_f,$$

or

$$\frac{\mathrm{d}P_f}{P_f} = \left[n\beta + (1-\beta)\frac{T_k}{T_k - T_0}\right]\frac{\mathrm{d}T_k}{T_k}\,. \quad . \; . \; . \quad (5)$$

From (3) we find:

$$\frac{\mathrm{d}P_f}{P_f} = a\,\frac{\mathrm{d}T_k}{T_k}\,. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (6)$$

We then obtain from (5) and (6):

$$a = n\,\beta + (1-\beta)\frac{T_k}{T_k - T_0}\,.$$

Solving this expression for β, we arrive at (4).

Fig. 1 shows a plot of T_k versus P_f for $a = 1$, $a = 3$ and $a = 4$. It can be seen that the dependence of T_k on P_f decreases as a increases, and, as appears from (4), a larger a is associated with a larger β, that is with a greater heat loss by radiation and a smaller loss by conduction. In agreement with our foregoing calculations, then, the effect of reducing the loss by thermal conduction is indeed favourable.

The marked dependence on a shown by the ratio $\beta/(1-\beta)$ between radiation and conduction losses appears from the figures in *Table I*, which refer to two different designs ($T_k = 1080$ °K, $T_0 = 350$ °K, $n = 4.65$, $a = 3.0$ and 3.8).

Table I.

	a=3.0	a=3.8
β	0.48	0.80
$1-\beta$	0.52	0.20

Fig. 2 shows T_k as a function of the heater voltage V_f for $a = 3$ and 4.

Variations of the heat balance of heater + cathode assemblies

Various specimens of the same type of cathode, operated at identical heater power, show differences in temperature owing to the unavoidable statistical spread in the heat balance of heater + cathode assemblies. An extensive investigation has shown that the main cause of this spread is an undefined heat loss at places where there is fairly loose thermal contact between components at different temperatures, e.g. the cathode body and the ceramic or mica disc to which the cathode body is secured. As regards

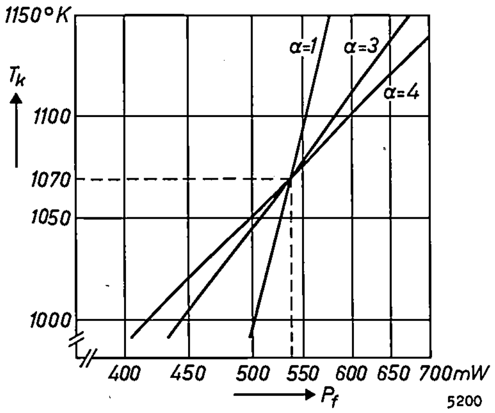


Fig. 1. Temperature T_k of a cathode as a function of heater power P_f, for various values of the exponent a in (3).

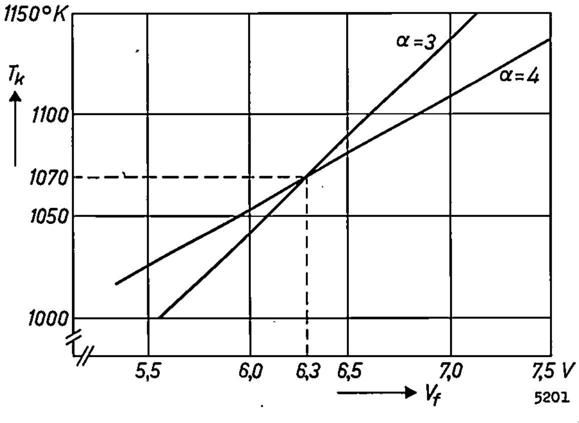


Fig. 2. As fig. 1, but now with the heater voltage V_f as abscissa.

radio tubes of conventional construction and comparable P_f, the standard deviation σ from the nominal cathode temperature is generally 20 °C[2]. Similar and even greater standard deviations are found where the cathodes used in cathode-ray tubes are mounted in ceramic discs, as is the case in America. The standard deviation of the 2-W cathode now fitted in Philips television picture tubes has been reduced to 10 °C by ensuring a defined heat loss through welded metal supports (of poor thermal conductivity). A similar method has been used (see below) to give the half-watt cathode the same low standard deviation of 10 °C.

Design of the new cathode

To draw up the heat balance of the new cathode we must start from various data that are more or less fixed beforehand.

1) Situated opposite the cathode, at a distance of 50 to 150 μ, is the Wehnelt cylinder, which has a bore of 1 mm or less. Since the size of the bore and its positioning in relation to the emissive coating show slight differences for different electron guns, the smallest linear dimension of the emissive surface must not be less than 1.4 mm. The minimum surface area must therefore be 2.0 mm^2.
2) At an operating temperature of 800 °C the emissive surface of 2.0 mm^2 radiates 60 mW (or less, depending on the reflexion coefficient of the Wehnelt cylinder).
3) The cathode body must accommodate the heater together with its insulation. Taking the maximum heater voltage of 12 V mentioned in the introduction, and having regard to the minimum diameters in which tungsten wire is readily workable in mass production, we see that the minimum volume of the cathode body must be about 2.5 mm^3.
4) For practical reasons the nickel body of the cathode (volume $\approx$ 2.5 mm^3) is made rectangular: length 3.5 mm, cross-section 0.5×1.4 mm.

The outside surface area of the cathode body is 14.5 mm^2, only 2 mm^2 of which is occupied by the emissive layer. The nickel surface of 12.5 mm^2 radiates 200 mW at an operating temperature of 800 °C[3]. One end of the body is closed, the other can be partly closed by a lug after insertion of the heater. Some heat is still lost through the opening which is left, however. The heater further loses heat by radiation from the leads outside the cathode and by conduction via the heater terminals. For these three items a total of 180 mW was measured. At a rough estimate, the open end accounts for 100 mW of this total and the heater ends for 80 mW; half of the latter may safely be regarded as radiated. The total radiated power is thus about 60 + 200 + 100 + 40 = 400 mW. It being technically feasible to limit the loss through the cathode supports to 100 mW, the total heat removed by conduction is about 40 + 100 = 140 mW. The heater power P_f is consequently 400 + 140 = 540 mW, a value which is low enough for the applications mentioned at the outset. The radiated fraction β and the conducted fraction $1 - \beta$ of P_f are thus 74% and 26%. This roughly corresponds to $a = 3.8$ in Table I, i.e. the cathode temperature depends relatively little on the heater voltage (fig. 2), just about as little as in the case of the 2-W cathode.

The detailed heat balance is set forth in *Table II*; the figures relating to the present 2-W cathode are given alongside for comparison.

Table II. Detailed heat balance of the present 2-W cathode for cathode-ray tubes and of the new half-watt cathode.

	Present 2-W cathode	New half-watt cathode
Radiation from oxide coating . .	90 mW	60 mW
Radiation from nickel surface . .	760 mW	200 mW
Radiation from open end	400 mW	100 mW
Radiation from heater leads . .	100 mW	40 mW
Conduction via heater leads . . .	100 mW	40 mW
Conduction via cathode supports	440 mW	100 mW
	1890 mW	540 mW
$P_\mathrm{f} = V_\mathrm{f}I$	6.3 V × 300 mA	6.3 V × 86 mA

Construction of the half-watt cathode

Fig. 3 shows a magnified photograph of the new cathode. The details are to be seen more clearly in *fig. 4a* and *b*, where a model 15 times actual size is shown.

The cathode body *1* has the dimensions stated above, so that it radiates about 360 mW. It is supported by two nickel-iron strips *6* of low thermal conductivity. Being long and thin they account for a heat loss by conduction of no more than 100 mW. To make the cathode capable of withstanding the shocks to which portable apparatus are exposed, the strips are fitted with a reinforcement rib *7* and "wings" *8*. These wings rest on the upper of the three

[2] The "standard deviation" is a convenient and widely used measure of spread; see for example P. J. McCarthy, Introduction to statistical reasoning, McGraw-Hill, New York 1957, p. 111 et seq.
In a Gaussian distribution of an infinitely large batch, 32% deviates by more than σ from the nominal value, 4.5% by more than 2σ, and 0.3% by more than 3σ.

[3] This does not contradict the statement under (2) that the emitting surface of 2 mm^2 radiates 60 mW. The radiation coefficient of the total radiation (of all wavelengths) from nickel is only about half that of the emissive material.

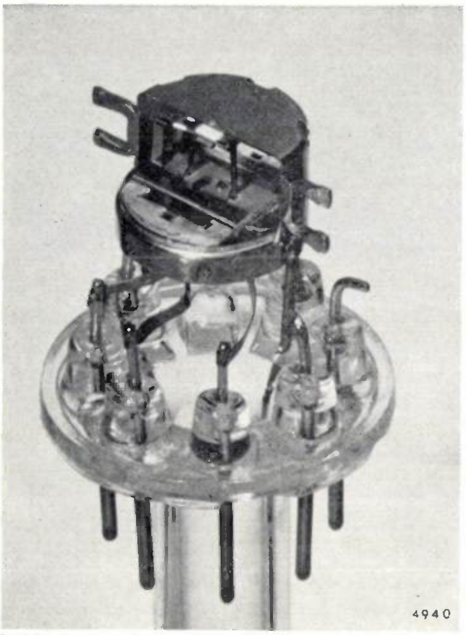


Fig. 3. Half-watt cathode for cathode-ray tubes. V_f = 6.3 V, I_f = 86 mA.

mica discs (*10*, *11*, *12*) in which the strips are secured, and prevent lateral movement. Shocks and vibrations in the longitudinal direction can do little harm, as the two supports form a parallelogram with the cathode body and the mica discs; any lengthwise displacement of the cathode is therefore accompanied by a much smaller vertical movement towards the neighbouring electrode.

Mounting the cathode supports in mica discs might not seem conducive to a small spread in temperature. The conditions are chosen in such a way, however, that any variable thermal contact that may arise between the supports and the mica will not significantly affect the temperature of the cathode. We can explain how this is done with the aid of *fig. 5*. The thermal resistances are drawn as electrical resistances; k represents the cathode, R_1 and R_2 the thermal resistance of the part of the supports above and below the mica, R_3 the variable contact resistance, and R_4 the thermal resistance of the mica, which is surrounded by the Wehnelt cylinder g_1. The ratio $R_1 : R_2$ and the spacing between cathode and mica are now chosen so that the temperature of the contact point P differs only slightly from the temperature to which the mica is raised, mainly by the heat radiated from the cathode. (T_k is 800 °C, at P the temperature is between 400 and 500 °C, and the mica is at about 300 °C.) The variable contact resistance R_3 thus joins points at only slightly different "potentials", so that the "potential" T_k of the cathode is virtually independent of the magnitude of R_3.

The heater consists of the conventional coiled tungsten wire. To make proper use of the space inside the cathode body, the windings are made oblong on a double mandrel (*fig. 6a*). Tungsten

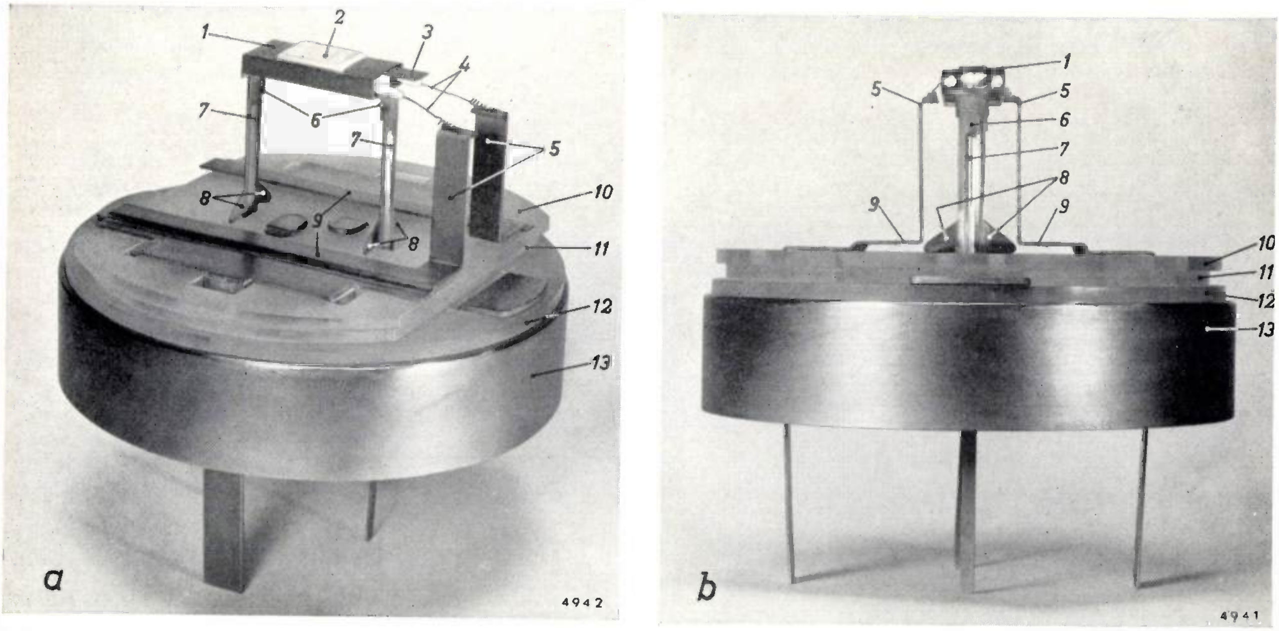


Fig. 4*a* and *b*. Model of the half-watt cathode (linear dimensions 15 times actual size). *1* nickel cathode body, with emissive portion *2*. The lug *3* can be folded to reduce radiation through the open end. *4* straight ends of heater wires. *5* heater terminals. *6* cathode supports of nickel-iron, with reinforcement rib *7* and wings *8*. *9* screen to prevent the formation of metal deposits on mica disc *10*. *11* and *12* mica discs. The diameter of *11* is smaller than that of *10* and *12*, to produce a long creep path (labyrinth) between *5* and *6* on the one part and the metal ring *13* on the other. The Wehnelt cylinder (not shown) is fixed to *13*.

wire of 30 μ diameter can be reproducibly wound in this way with a pitch of 45 μ. The wound mandrel is bent into an M shape and given the usual insulating coating of aluminium oxide (fig. 6b and c). The wires projecting outside the cathode body are left straight so as to reduce their radiating surface, but the ends are coiled to make them better visible for welding.

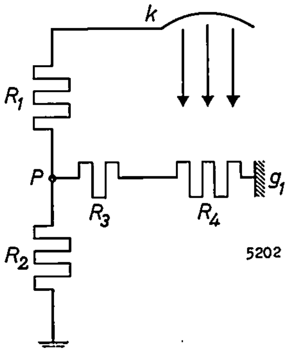


Fig. 5. Thermal resistances are shown here as electrical resistances and temperatures are treated as electrical potentials. k cathode, with temperature $T_k = 800$ °C. R_1, R_2 thermal resistance of cathode supports above and below the mica discs. P point where the supports touch the mica. R_3 variable contact resistance. R_4 thermal resistance of mica. g_1 Wehnelt cylinder. The ratio $R_1 : R_2$ and the distance between cathode and mica are chosen so that the temperature at P (400-500 °C) differs little from the temperature acquired by the mica as a result of cathode radiation (about 300 °C). The magnitude of R_3 therefore has little influence on T_k.

Metal deposits on the mica might cause bad electrical insulation between heater, cathode and Wehnelt cylinder. The heater leads are therefore provided with a screen (*9* in fig. 4*b*) which shields the underlying part of the mica *10*. A further safeguard is the smaller diameter of the mica *11* sandwiched between *10* and *12*, which creates a long creep path (labyrinth) between the Wehnelt cylinder (*13*) on the one part and the cathode and heater on the other (see fig. 4*b*).

As mentioned in the introduction, the new cathode is primarily intended for tubes in transistorized apparatus. Now one advantage of transistors is that they start operating as soon as the apparatus is switched on. That can never be expected of an indirectly heated cathode, of course, but an effort can certainly be made to shorten the heating-up time. The half-watt cathode, used in a television picture tube, takes 8 seconds to reach 10% emission — sufficient for the picture to appear on the screen — and 13 seconds to reach 80% emission. The present 2-W cathode takes more than three times as long to reach these emission values.

The permissible voltage V_{kf} between the heater (f) and the cathode (k) is 50 V (k negative) or 100 V (k positive)[4]. If a higher V_{kf} is required for other applications, thicker insulation will of course be needed between cathode and heater, which will entail a larger emissive surface. The heater power will then have to be somewhat higher, though it will still be less than 1 W.

[4] For the influence of polarity on the insulation between cathode and heater, see Philips tech. Rev. **18**, 188, 1956/57.

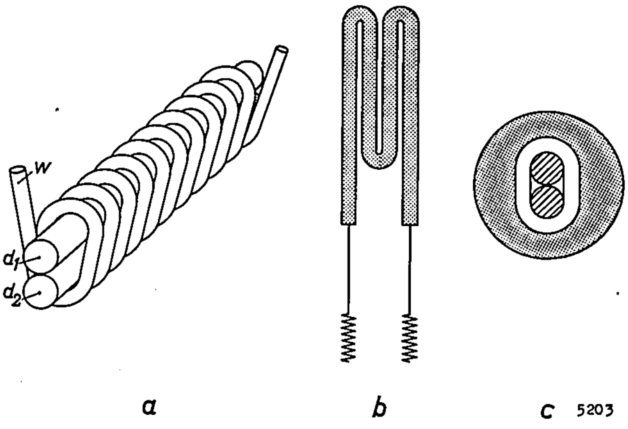


Fig. 6. Heater of half-watt cathode ($V_f = 6.3$ V).
a) Tungsten wire w of 30 μ is wound on a double mandrel d_1-d_2 of molybdenum wire.
b) The coiled mandrel is then bent into the shape of an M and the tungsten coil is coated with aluminium oxide.
c) Cross-section of mandrel with coil and insulation.
The mandrel is dissolved in strong acid, leaving behind the insulated heater. The free ends (see *b*) are left straight so as to reduce the radiating surface; their tips are coiled, however, to make them better visible for welding.

Summary of I and II. The permissible current density of conventional oxide-coated cathodes is limited by the resistance of the oxide coating and by interface resistance to 0.5-1 A/cm^2. Recognition of the fact that the oxide coating has two distinct functions, i.e. to emit electrons and to keep the emissive surface supplied with barium, has led to designs in which these two functions are separated. On these lines two new types of oxide-coated cathodes have been evolved: the storage cathode and the double-coated cathode, for loads of 1.5-2 A/cm^2 and 1-1.5 A/cm^2 respectively. The first type can have a life of over 5000 hours at a load of 1.5 A/cm^2, the second 1200-1500 hours at 1.5 A/cm^2 and 2000-2500 hours at 1 A/cm.

In battery-fed transistorized equipment containing a cathode-ray tube (portable television receivers, television cameras, oscilloscopes) there is a need for a C.R.T. cathode that will consume very little power, e.g. 0.5 W. Factors of importance in the design of such cathodes are the effect of heater-voltage variations on the cathode temperature, the ability to withstand shocks and vibrations, the insulation between heater, cathode and Wehnelt cylinder, and the heating-up time. The effect of heater-voltage variations can be minimized by reducing the heat loss of the cathode by conduction in favour of that by radiation. An indirectly heated oxide-coated cathode is described for 6.3 V and 86 mA (0.54 W). The standard deviation of the cathode temperature is only 10 °C. A picture tube fitted with this cathode gives a visible picture only 8 seconds after switching on. The cathode is used in oscilloscope tubes of type DH 7-11 and in vidicons of type 55 850.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2830: A. Verloop, G. J. B. Corts and E. Havinga: Studies on vitamin D and related compounds, X. Preparation and properties of cis-isotachysterol$_2$ (Rec. Trav. chim. Pays-Bas **79**, 164-178, 1960, No. 2).

The photochemical conversion of isotachysterol$_2$ (in ether) has been studied; the effects of wavelength and time of irradiation on the composition of the irradiation mixture are discussed. On chromatography of the crude irradiation product a compound was obtained in a pure, non-crystalline form. The compound did not show antirachitic activity. The ultra-violet and infra-red absorption spectra, the iodine-catalysed quantitative conversion to isotachysterol$_2$, the behaviour on heating in solution and the very low reactivity towards maleic anhydride led to the conclusion that the compound is the Δ6,7-cis isomer of isotachysterol$_2$, i.e. 9,10-secoergosta-5(10),6-cis,8(14),22-tetraen-3β-ol.

2831: B. Bölger, B. J. Robinson and J. Ubbink: Some characteristics of a maser at 1420 MHz (Physica **26**, 1-18, 1960, No. 1).

Description of the construction of a solid-state maser (a microwave amplifier that operates by stimulated emission) for a frequency of 1420 Mc/s, i.e. the frequency of the line radiation from interstellar hydrogen. The operative ions are Cr^{3+} ions on Co sites in $K_3Co(CN)_6$, which possess four energy levels. The Cr^{3+} concentration is 0.05%. Using the pump frequency f_{13} of 3850 Mc/s, the populations n_1 and n_3 of levels *1* and *3* are equalized. When, in consequence of this, the population n_2 of the intermediate level *2* becomes greater than n_1, radiation of frequency f_{12} (the signal frequency) will give rise to stimulated emission and be amplified. The energy-level differences are adjusted to the required values (f_{12} must be 1420 Mc/s) with a variable external magnetic field. The article also describes resonance experiments designed to determine the possible adjustments of the magnetic field, the amplification and bandwidth obtained, and the behaviour of the maser at saturation. Some results achieved with ruby (Cr^{3+} in Al_2O_3) are mentioned. This particular maser is not very satisfactory for radio-astronomical measurements because of its narrow bandwidth and poor stability. Better results could be obtained by placing the emissive crystal in a slow-wave structure.

2832: N. W. H. Addink: The determination of trace elements (J. Iron and Steel Inst. **194**, 199-211, 1960, No. 2).

Description of a method of spectrochemical analysis, developed by the author and used in Philips Research Laboratories at Eindhoven, which is based on complete evaporation of the material to be analysed by a DC carbon arc. The method is suitable, among other things, for investigating very small quantities of material (as little as 0.01 microgram). The author deals with various significant factors involved, in particular with the speed of evaporation and its influence on the percentage of material lost. This factor has tended to be somewhat neglected. Results are mentioned of the analysis of solutions with and without chemical pre-concentration. The method of X-ray fluorescence analysis is then discussed and compared with spectrochemical analysis, using results obtained by both methods on a number of standard steels and determinations of zinc and iron in blood. The spectrochemical method gives reasonably accurate results and requires only very little material. The method of X-ray fluorescence gives high accuracy, but calls for the use of somewhat more material.

2833: M. Avinor: Visible emission of nickel-activated CaS (J. chem. Phys. **32**, 621-622, 1960, No. 2).

It is shown that red visible emission in nickel-activated CaS, as reported by Lenard *et al.*, is not caused by the presence of nickel alone. The red emission is presumably due to associated nickel and copper centres. By itself, nickel activation causes weak green fluorescence and copper activation blue-green fluorescence.

2834: H. Koopman and J. Daams: Investigations on herbicides, III. 2,4-disubstituted 6-chloro-1,3,5-triazines (Rec. Trav. chim. Pays-Bas **79**, 83-89, 1960, No. 1).

In view of their herbicidal properties a number of 2,4-substituted 6-chloro-1,3,5-triazines was synthesized. Their herbicidal properties are briefly discussed.

VOLUME 23, 1961/62, No. 2 pp. 29-68 Published 7th December 1961

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

Photo S. T. Karlsson

ELECTRONICS IN THE WOOD INDUSTRY

by M. HAMMAR *), A. RYDAHL *) and B. WESTERLUND *).

621.38:674

Production methods in many industries are changing rapidly with the wide introduction of electronics. Examples include photoelectrical devices, high-frequency heating, telemetering, and control and automation systems. It cannot be the aim of a single article to attempt a survey of the industrial applications of electronics, even for the case of a single industry — the wood industry. Neither does the present article deal extensively with specific electronic equipment, although here and there some details are discussed. The purpose of the article is rather to illustrate the various ways in which industrial processes can be improved by the application of electronics. The nature of the material handled in the wood industry — living matter, with all its freakishness in composition, shape and properties — lends a special flavour to the problems encountered here.

The topics dealt with in this article represent part of the

*) Svenska Aktiebolaget Philips, Stockholm.

activity of Svenska AB Philips, Stockholm in developing "tailor-made" industrial electronic equipment. One reason for focusing attention on Sweden is the important world role of the Swedish wood industry: in 1957 Sweden ranked third of all countries in the production of pulp and fifth in the production of semi-finished and finished coniferous wood products, and in exports of these goods it ranked first and second respectively. Another reason is the fact that wages in this country, being the highest in the world next to U.S.A. and Canada, are fostering investments in equipment for saving labour and increasing productivity. Sweden may therefore be regarded as a "pilot land", foreshadowing the developments in other countries.

Introduction

The three main Swedish industries are mining, steel and wood. In each of these industries, electronic equipment is being used more and more. Conditions in Sweden are favourable for such a development: there is a demand for high output and high quality, which makes new means to these ends welcome, and the introduction of new equipment is stimulated by the pressure of high wages and by the readiness of the manufacturers to pool their knowledge. (Technical collaboration has a long history in Sweden: in the iron industry it began in 1747, and the wood industry has had common research establishments for many years.)

The wood industry is particularly interesting because of the variety of its products, semi-finished or finished — ranging from planks, mining props and telephone poles to pulp, board and paper, and from matches, toys and shoe lasts to parquet blocks, furniture and prefabricated houses. These products constitute about 25% of the total value of Swedish exports. Many of the processes in the wood industry have already been mechanized to a high degree. The help that electronics can nowadays offer in the wood industry is chiefly concerned with two types of operations: *a*) physical ones, including *drying* and *gluing*, for which the required heat can very effectively be produced in a high-frequency furnace; and *b*) organizational ones, including *inspection*, *sorting*, and *control*. The latter group of operation are usually preceded by *measurements* which, again, may involve electronics.

We shall not dwell here on the process of *gluing*, since this important application of high-frequency (dielectric) heating was introduced many years ago and has already been described in this journal [1]). We may however mention that in the past few years installations have been designed in which high-frequency gluing is performed as a *continuous* production process, the objects to be glued being transported through the high-frequency oven on a conveyor belt [2]). A similar procedure, which greatly improves the economy of the process, is adopted in one case of high-frequency *drying*, a comparatively new development which will be described in this article. After this, several examples of electronic *inspection* and *sorting* will be presented. Finally, the complete production process in a *sawmill*, where electronics is put to work at a number of stages, will be considered.

High-frequency drying of shoe lasts

Lasts are the basic tool for the mass manufacture of shoes. When the model for a certain type of shoe in a given size has been designed, a steel prototype of the required last, on which the upper leather will be moulded, is made on a precision milling machine. The requisite number of wooden lasts is then manufactured from this prototype by means of a copying machine. The very strong, hard wood of the beech is generally used for the lasts, since they must remain smooth and retain the designed shape very accurately during the moulding of the leather, the nailing on of the soles, etc. Two or three hundred shoes can be made on a properly manufactured last of this material.

The wood to be machined must have a very low moisture content (7-9%), in order to prevent subsequent shrinking or warping of the last. This would of course affect the correct shape and size, and it would endanger the precise fit of those parts of the last which must be detachable or movable in order to enable it to be withdrawn from the completed shoe.

The rough cut blocks, which may originally contain as much as 65-75% of moisture, must be dried very cautiously, because of the danger of cracking. This danger is enhanced by the very hardness and density of the beech wood which make it so suitable for lasts. Natural drying is a very slow process and cannot reduce the moisture content to less than 30 or 20%. Moreover, when drying proceeds very slowly, there is a risk that the wood will be attacked by fungi ("blue stain" or mould). Artificial drying is therefore the usual practice in many branches of the wood industry. A great number of procedures have been developed for this purpose [3]), the most common

[1]) Philips tech. Rev. **11**, 239, 1949/50.

[2]) See e.g. G. Wästberg, Die Hochfrequenz-Holzverleimung in Schweden, Holz als Roh- und Werkstoff **16**, 177-183, 1958.

[3]) See e.g. F. Kollmann, Technologie des Holzes und der Holzwerkstoffe, Springer, Berlin 1955, Vol. 2, pp. 255-380.

method consisting in heating the wood by a mixture of steam and air in such a way that the outer layers of the wood do not start giving off water before the inner layers have also warmed up and can participate in the process of evaporation (and shrinkage!). Dielectric heating in a high-frequency field is clearly much better suited for the purpose: the heat is not applied from the outside but is developed in the material itself, and the innermost parts of the wood acquire an even higher temperature than the outer layers, which are cooled by the surrounding atmosphere. The water from the middle is thus the first to evaporate and it can find its way out relatively unimpeded through the pores of the outer layers, which have not yet shrunk.

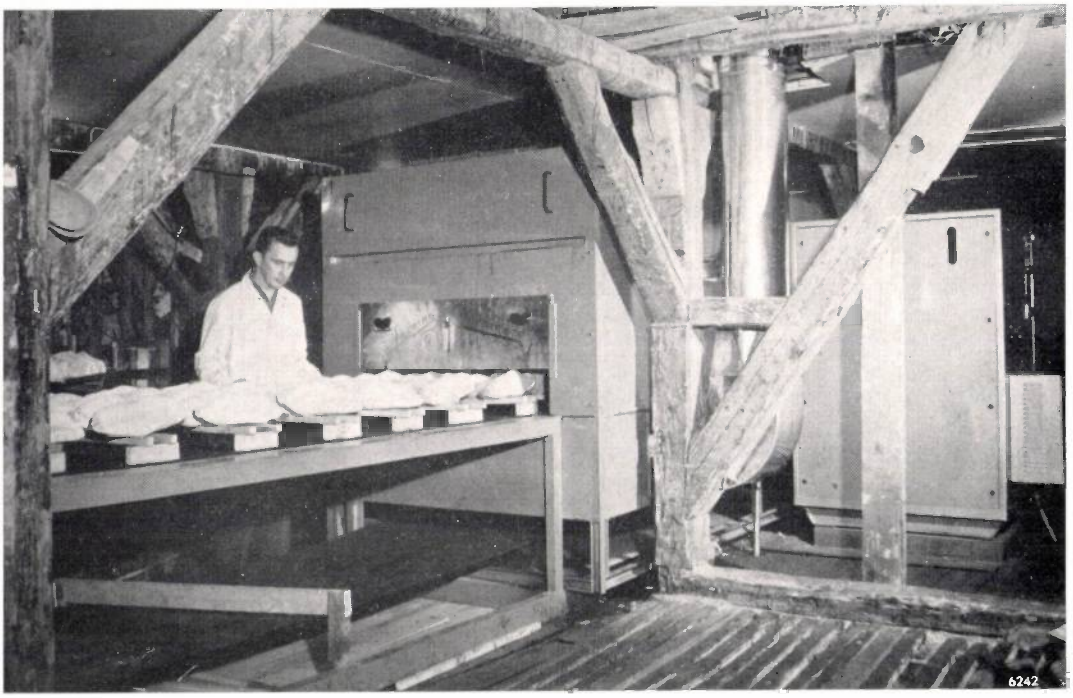


Fig. 1. Installation for high-frequency drying of shoe lasts at Järrestad, designed by Svenska AB Philips, Stockholm in collaboration with Svenska Skolästfabriken. The belt loaded with lasts travels through the HF oven at about 3 m/hour. To the right, the 20-kW HF generator.

The high-frequency drying of wood, first conceived in 1928 in the U.S.A. and extensively investigated since 1934 in Russia, Germany and other countries [3]), is relatively expensive, but so are the losses that may be inflicted by other drying procedures. Moreover, high-frequency drying is much faster than other methods. It will therefore be an economic proposition in many cases, and especially for small, carefully machined objects such as shoe lasts. Moreover, recent developments have considerably improved the economy of high-frequency drying by making the process *continuous*: the material is then transported on a conveyor belt through the high-frequency field of an open capacitor. Svenska Skolästfabriken at Järrestad in the south of Sweden have used such a method for the last two years with considerable success. A few details of their installation [4]) (*fig. 1*) are given here.

The high-frequency generator operates at 12 Mc/s and delivers 20 kW of HF power. It comprises a power-supply unit with a high-tension transformer and a 3-phase full-wave rectifier, and a HF unit with an oscillator tube and an oscillating circuit in which the oven serves as the capacitor (*fig. 2*). The oven contains one lower electrode about 3 metres long (perforated so that the water can drain off), over which the conveyor belt (of fine-meshed stainless-steel gauze) carrying the wooden blocks is passed, and three upper electrodes each about 1 metre long. The three upper electrodes are interconnected by flexible sheets and are separately adjustable in height. This subdivision of the capacitor into three independent sections is necessary in order to control the HF field strength and thus the degree of heating the blocks will undergo at different

[4]) Designed in the HF-heating laboratory of Svenska AB Philips, Stockholm in conjunction with Svenska Skolästfabriken. A series of identical installations are being built for other plants in collaboration with this company.

points when travelling through the oven; they should be heated strongly in the first section of the oven, then given a "rest period" with moderate heating, while in the last section the heating should increase again. The design of the HF circuit had to allow for an unusually large spacing between the upper and lower electrodes of the capacitor (*fig. 3*): the water dripping from both ends of the wooden blocks must be prevented from short-circuiting these with the lower electrode, as this might cause the wood to be overheated in places, or even to burst into flame. The upper electrode must also be a considerable distance above the blocks in order to avoid excessive fields on protruding ribs and points.

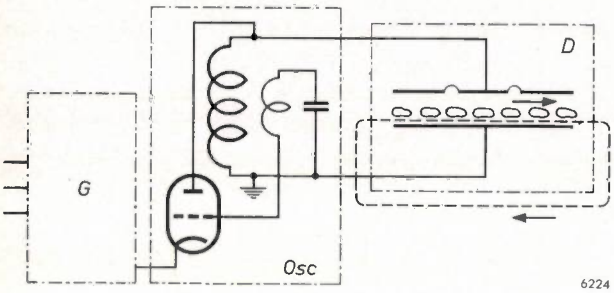


Fig. 2. Circuit diagram of HF generator and oven. *G* power-supply unit. *Osc* HF unit. The large capacitor which forms part of the oscillating circuit is situated in the oven *D*.

Since between 500 and 800 grams of water normally have to be removed from each block, while the 20-kW oven can remove a total of 35 kg of water per hour, an average of 30 pairs of lasts can be dried per hour. About 10 pairs can usually be accommodated per meter of the conveyor belt, so that the belt can move at a rate of about 3 m per hour. During the passage through the oven each block shrinks by about 12% in width and 6% in height (*fig. 4*). Thermo-couple measurements have shown that the HF field heats the middle of the block to 85 °C, and its surface to 50 °C.

The complete drying process established at Järrestad after a certain period of experimentation is represented in *fig. 5a*. The blocks of wood are taken from stock after 2 months of cold storage, when the initial moisture content of 62% has dropped to 35-50%. They are then stored at room temperature for about 2 days (pre-drying), HF-dried in less than 2 hours, and again stored at room temperature for 1 or 2 days before machining, in order to allow the wood to cool down and to allow the remaining moisture content of about 7-9% to distribute itself uniformly. This procedure should be contrasted with that in use before the intrcduction of HF drying (see fig. 5*b*): the blocks had to be kept in cold storage for 6 months, were then dried by hot air for 3 months, and finally stored at room temperature for 1 month.

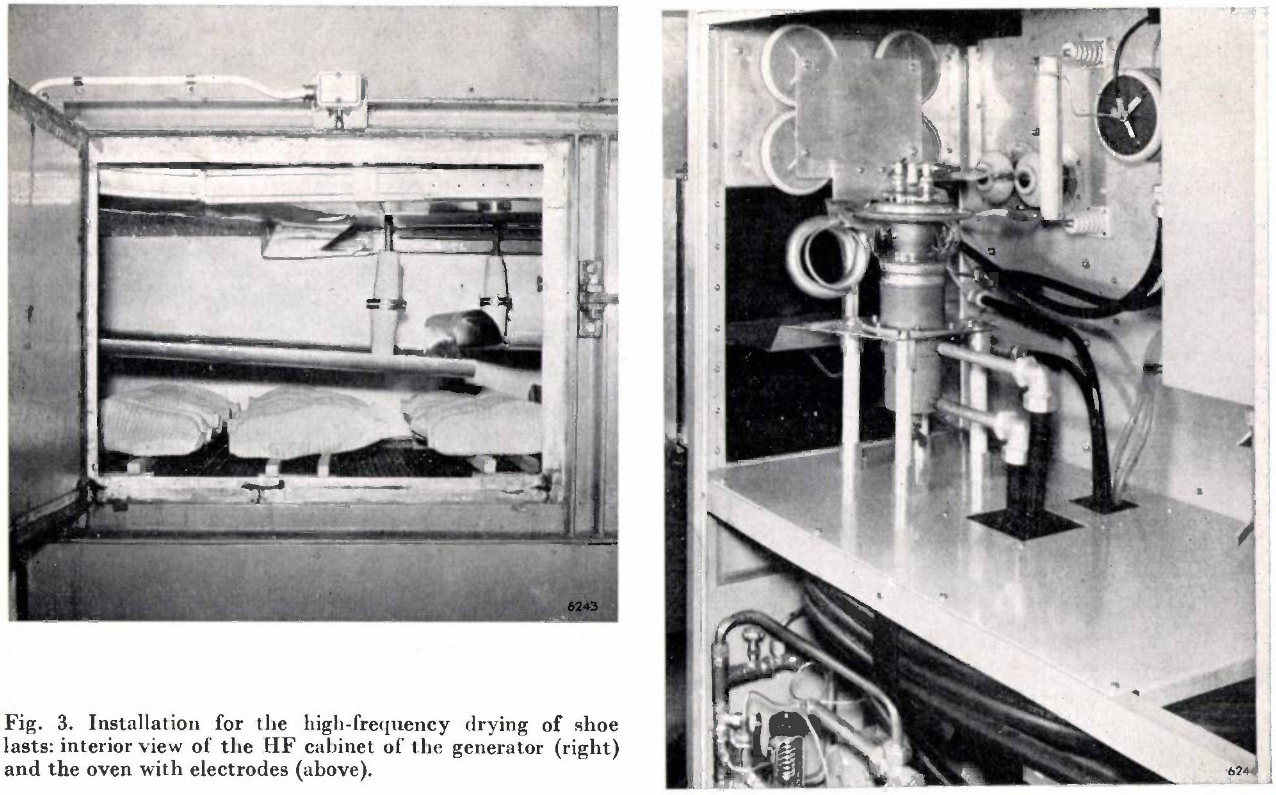


Fig. 3. Installation for the high-frequency drying of shoe lasts: interior view of the HF cabinet of the generator (right) and the oven with electrodes (above).

A comparison between the times involved bespeaks the advantage of HF drying quite eloquently: the wood stocks and storage space required for a given output are enormously reduced. Storage costs at Järrestad have been brought down to 15% of their former levels. (The saving of the energy formerly used for heating the hot-air drying chambers is not a real advantage, since this energy was provided extremely cheaply by burning scrap wood.)

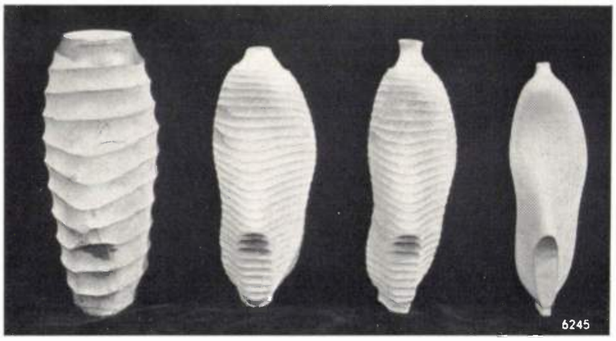


Fig. 4. Four lasts in different stages of manufacturing. From left to right: rough cut block; a similar block roughly machined to the shape of the last, before drying; the same after drying; machined to final shape.

This large saving is obtained at a relatively low investment cost, thanks to a change in the manufacturing process made possible by the new drying method. The blocks of wood cut from the trees must be large enough to accommodate every type and size of last. With the former method, where the drying period was much longer than delivery times scheduled for batches of lasts, it was unavoidable that a large quantity of wood in each block was kept in the drying process, taking up space in the store and in the hot-air chambers, only to be removed in the preliminary and final machining of the block. With HF drying, blocks selected for a batch of lasts (after 2 months' storage) can be roughly machined *before* passing through the HF oven (see fig. 4), thus considerably reducing the quantity of wood to be dried and therefore the size and power of the HF generator to be installed for a given production.

Other advantages arise in addition to the savings in storage costs. The long drying periods previously used could not prevent the rejection rate because of cracks from amounting to 8-10%; the rejection rate because of fungi was sometimes even higher. With HF drying, only 1-2% of the blocks have to be rejected for cracks, and none for fungi. Moreover, the HF-dried wood is found to be definitely of superior quality: it is brighter, more uniform in colour and in strength, and easier to machine and to polish. The latter advantages could hardly have been predicted, although wood experts state that they can be explained by considering the micro-processes occurring in the wood during drying. This, however, is taking us beyond the scope of this article.

Inspection of veneer for match boxes

The manufacturing of matches was one of the first industrial processes to be virtually completely mechanized. A large number of single steps are involved and a variety of machines are used for performing them. The logs (aspen wood is generally used) are peeled to produce veneer of two thicknesses, one for the match sticks, the other for the

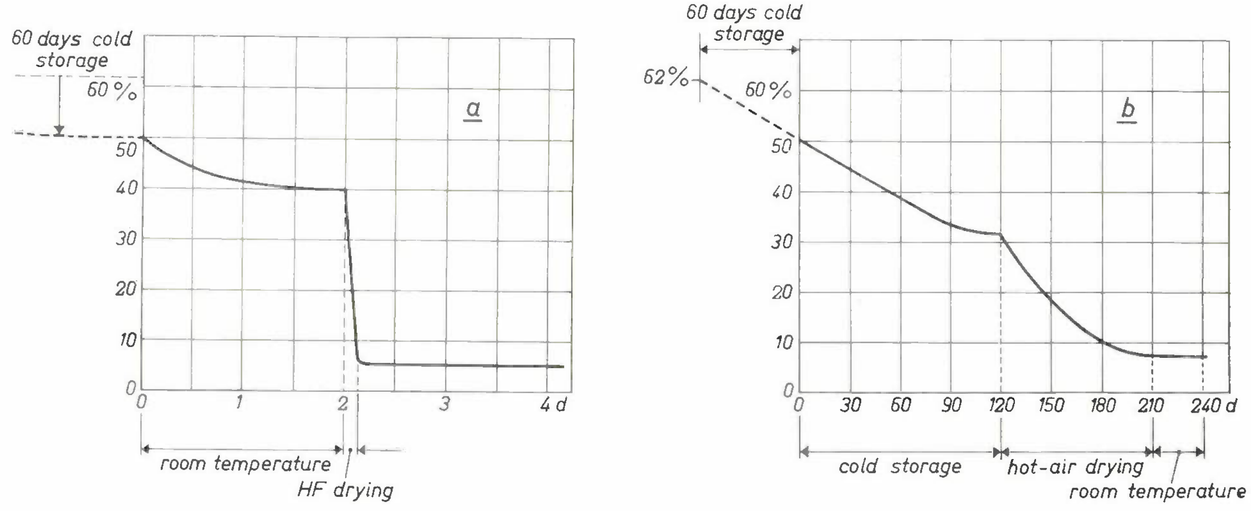


Fig. 5. *a*) Time diagram of new drying process for lasts at Järrestad: water content of the wood in wt.% as a function of time in days, beginning after two months' cold storage. *b*) Time diagram of old drying process. Cold storage actually lasted for 6 months, but the first two months are put before the zero of the time scale, to make the two graphs comparable.

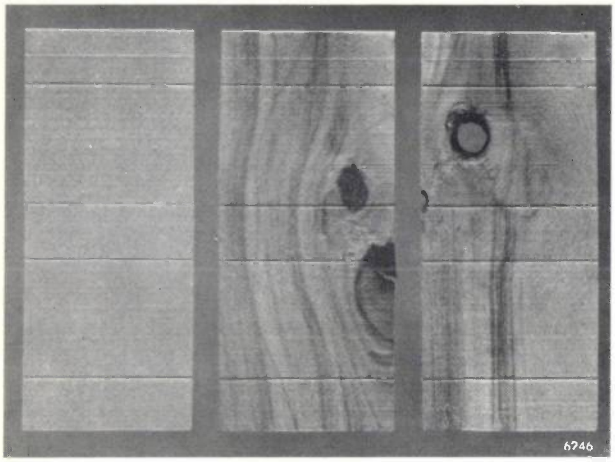


Fig. 6. Strips of veneer for the sleeves of match boxes. Each strip is folded along the grooves. From left to right: normal strip; strip with knot; strip with hole.

match boxes. After the veneer has been cut to the required lengths, it follows two separate lines. The first line includes: cutting the veneer strips to obtain the match sticks (one machine will make as many as 70 million a day); collecting, drying and chemically polishing and impregnating the sticks; eliminating short and faulty ones; aligning the sticks parallel and passing them through an enormous dipping machine for providing them with the heads, hardening these in an adjacent drying installation and collecting the finished sticks for filling the boxes. The second line makes the boxes: the veneer strips, coming in two widths, narrow for the drawer, broad for the sleeve, both already grooved during cutting, are fed to two separate machines. These will fold them (at a rate of 200 per minute), add the cardboard bottom for the drawer and paste paper strips on to them, whereafter the sleeve is completed by pasting on the colourful emblem and painting the striking surfaces on both sides. Box and sleeve are finally fed to the filling machine to meet the sticks.

Very little manual labour is involved in this manufacturing process: 3 man-seconds are required for one match box, completed, filled with its 50 sticks and packed.

About 10% of this labour is spent on one seemingly very simple step: inspecting the veneer for the boxes and removing those strips containing knots, holes or other visible defects (*fig. 6*). The girls in charge of the folding and pasting machines let the piled-up strips of veneer pass through their fingers like a fan and drop every defective one with a quick movement. One girl can look after and feed two machines, about half her time being occupied by inspecting and sorting the veneer strips, at a rate of 400 per minute.

However elegantly the girls perform their task, it is evident that here is a point where electronics comes into its own: the inspection can be done photoelectrically and the removal of the defective strips can be effected with the aid of a relay. Equipment which performs this task for the broad veneer

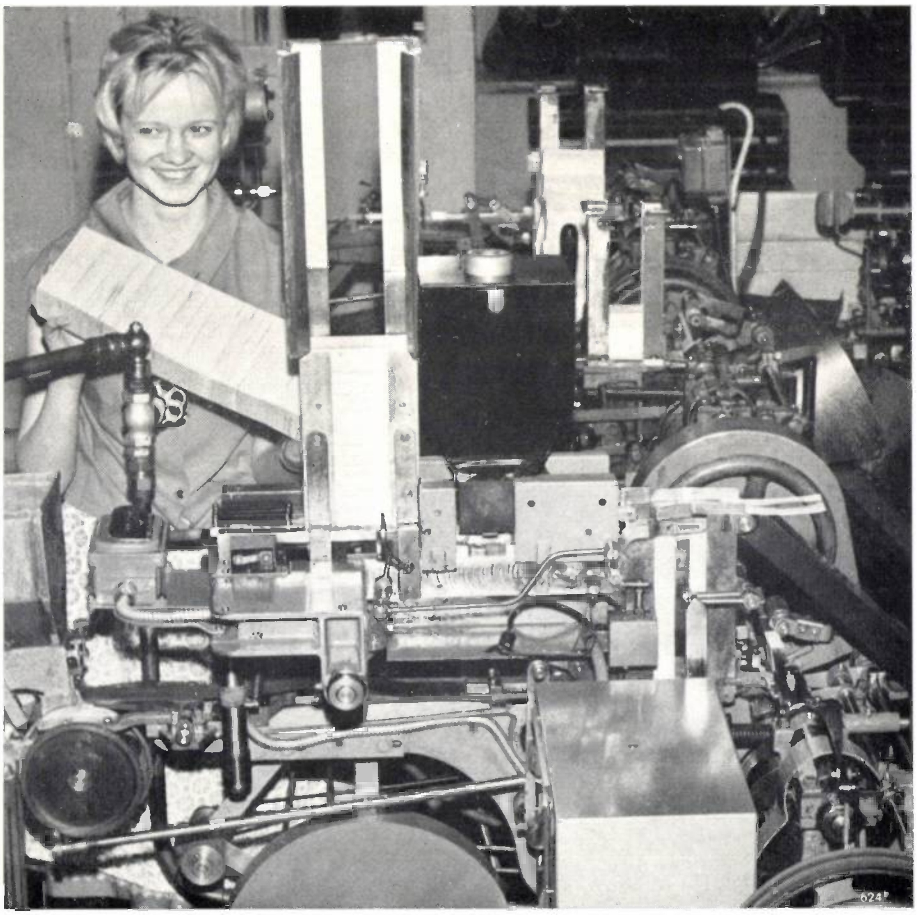


Fig. 7. Folding machine for match-box sleeves at the Vetlanda factory of Svenska Tändsticks AB. Top centre, the magazine for the strips of veneer (the girl is holding another stack of strips to fill the magazine) and the box containing the scanning mechanism for inspecting the strips. To the right a defective strip is just leaving the slide and entering the chute to the reject box.

strips has been designed and installed for Vetlanda Tändstickfabriken, a member of the group Svenska Tändsticks AB, Jönköping. A considerable number of instruments of the same type will in due course be installed in the factories of this group in consequence of the very satisfactory results obtained with the prototype. The apparatus is shown in action in *fig.* 7 and will now be described briefly.

The strips of veneer contained in the magazine of a folding machine are passed in rapid succession over a smooth metal surface (the "slide") by several pairs of rollers. A photoelectric cell above the slide observes the light reflected from the moving strip when this is scanned in the transverse direction by a flying spot. (In reality, there are two scanning spots, see below.) If the strip of veneer is faultless, the photoelectric current will be approximately constant and the strip passes on to the folding mechanism. If, however, the scanning spot passes a knot, which is darker than the normal wood (or a hole, whose edges likewise are darker), the photoelectric current diminishes suddenly, and by this negative pulse actuates a relay, energizing a solenoid which raises a hinged steel plate at the end of the slide, causing the defective strip of veneer to enter a chute leading to a reject box. 700 scans, covering six pieces of veneer, can be carried out per second.

An interesting point in this rather commonplace set-up is the solution of the well-known problem of the spurious pulse produced at the edges of the scanned object: every time the flying spot crosses one of the edges, it gives rise to a signal similar to that indicating a knot or hole, so that every strip would be discarded unless special precautions were taken. Keeping the spot exactly within the width of the strips is hardly possible. The difficulty is eliminated in a simple way by using *two* alternating scanning spots, which move in opposite directions. The photocell current caused by each spot is initially suppressed until the spot is well on the strip, but is allowed to flow until the spot has pursued its course well beyond the far edge. The slide on which the strips move has a *brighter* surface than the veneer. Every time a spot leaves the strip a *positive* current pulse is now delivered by the photoelectric cell, but this is made ineffective by a simple circuit, which causes the above-mentioned relay to react only to negative pulses.

The two flying spots are produced by means of a small light bulb surrounded by a rotating perforated drum and flanked by two stationary mirrors; the arrangement is shown in *fig. 8* and a few details are explained in the caption. *Fig. 9* shows the device in its cover and the cabinet containing the electronic circuits.

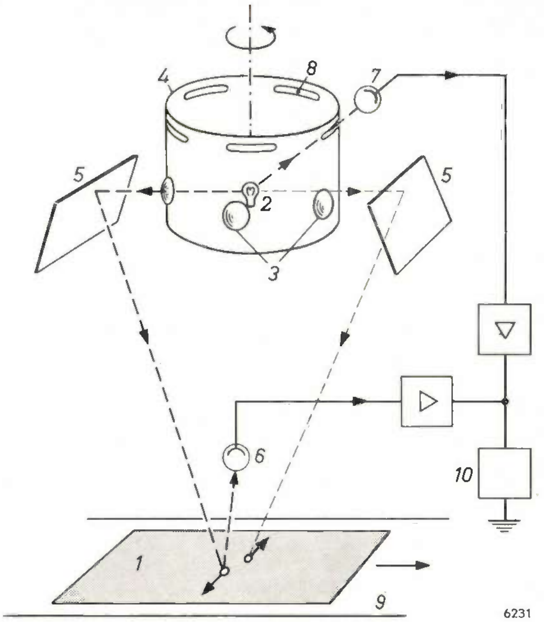


Fig. 8. Scanning system for the inspection of veneer strips. Two flying spots appearing alternately on the strip *1* and moving in opposite directions are produced by the lamp *2* by means of lenses *3* in the rotating drum *4* and fixed mirrors *5*. The photocell *6* receives the reflected light of the spot and produces a negative current pulse firing a thyratron and thereby actuating the reject mechanism as soon as a dark part is encountered in the strip. Until each spot has crossed the first edge of the strip, the photocell current is suppressed by the signal of the auxiliary cell *7* illuminated by suitably phased slits *8* in the rotating drum. The positive current pulse produced by *6* at the moment when each spot crosses the second edge of the strip and arrives on the brighter surface *9* underneath is made ineffective by a diode in the thyratron circuit *10*.

Owing to the introduction of this automatic inspection equipment, one girl can now run 4 machines instead of 2. For the time being, only the *broad* veneer strips used for the sleeves of match boxes are sorted in this way. Similar equipment for inspection of the narrow strips for the drawers is being developed.

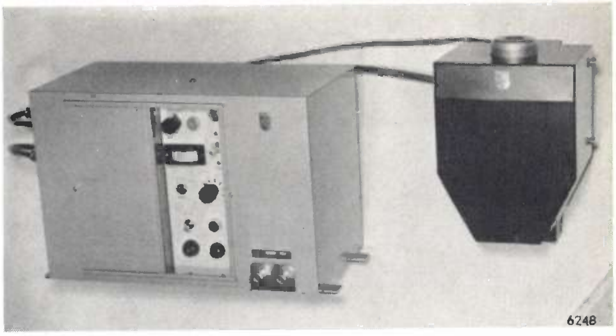


Fig. 9. Scanning device in its cover (right) and cabinet with electronic circuits.

Sorting of parquet blocks

The following example is not very different in principle from the one just discussed, but it illustrates the possibility of more sophisticated scanning, adapted to the variety encountered in the appearance of wood. This example concerns equipment developed and installed for Limhamns Träindustri at Ronneby, where it has been working for more than a year. The equipment automatically sorts oak parquet blocks into four different classes, viz., bright, medium, dark, and patterned (*fig. 10*).

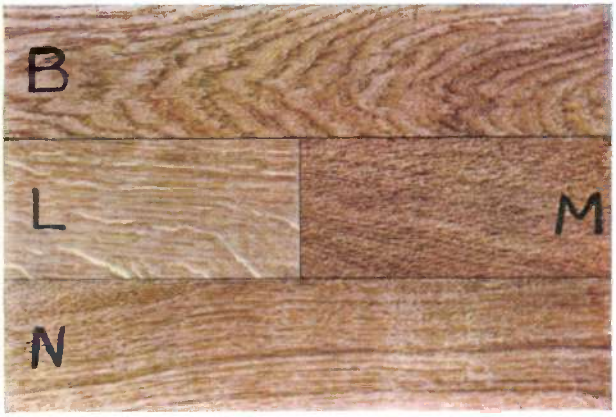


Fig. 10. Oak parquet blocks of four different classes: *L* bright, *N* medium, *M* dark, *B* patterned.

Blocks of fixed dimensions (length between 20 and 45 cm, width between 5 and 8 cm) are fed into a magazine, from which a conveyor belt takes them, one after another, to a photoelectric scanner. The scanner contains two photocells (*fig. 11*). The first cell measures the average brightness (reflectivity) of the block as it passes under a uniformly illuminated window of length 10 cm and width equal to that of the block. The reflectivity limits for the medium class can be preset within a wide range. The blocks which are darker or brighter than these limits are assigned to the dark class and the bright class respectively. This information, conveyed by the photocell current, is stored for a short time in an *RC*-type "memory": it will actuate a mechanism depositing the block in one of three boxes which it will pass when it is carried further by the continuously moving belt. Before this happens, however, the block passes under the second photocell, where it is scanned by a light spot with an area of 3 × 5 mm, reflected from a rotating mirror and thereby describing a circular path on the block. If the wood has a pattern of darker and brighter lines, the photocell observing the light reflected from the block will deliver an alternating current, whose amplitude will be larger the more pronounced the pattern. By amplifying this current and passing it through a suitable circuit (similar to the clamping circuit well known from television techniques), a current of the shape shown in *fig. 12* is obtained and fed into an integrating *RC* circuit. If the voltage on the capacitor reaches a certain (adjustable) value, the information concerning the average brightness, previously stored in the memory, is cancelled and replaced by a signal which will deposit the block in a fourth box, for the "patterned" class.

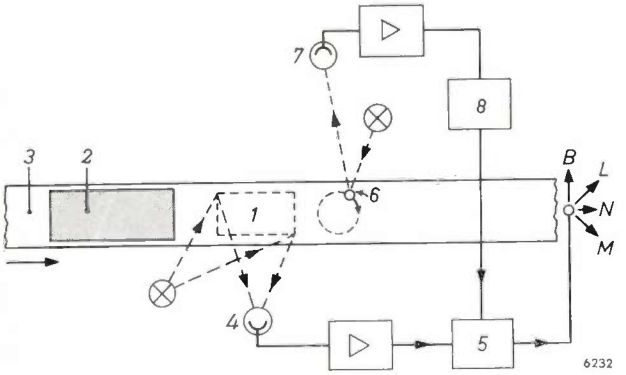


Fig. 11. Scanning system for sorting parquet blocks. *1* uniformly lit window, underneath which the blocks *2* are transported by belt *3* and which is observed by photocell *4* for measuring the average reflectivity of each block. The signal produced is stored in the memory *5*. The scanning spot *6* produces a varying signal in photocell *7*, which is integrated in *8*. If the integrated signal exceeds a certain limit it replaces the signal stored in *5*. The output of *5* controls the position of "points" directing the blocks to boxes *L*, *N*, *M* or *B*.

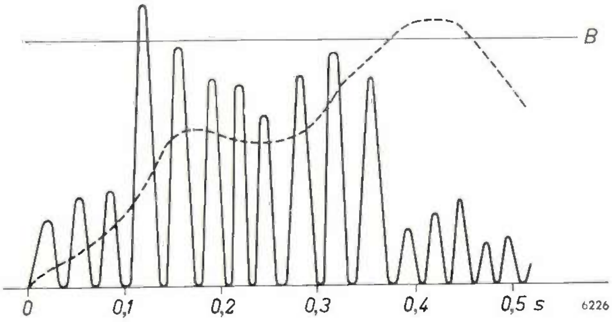


Fig. 12. Signal obtained when the scanning spot moves on a block. The broken line is the integrated signal, which must exceed the limit *B* for the block to be classified as "patterned".

The apparatus is illustrated in *fig. 13*. The sorting speed is about 1 block per second.

The obvious advantage offered by electronic sorting is again the saving of labour, but it should be emphasized that the improvement of quality is of equal or even greater importance: substituting physical criteria for the subjective comparison in the evaluation of brightness and variegation contributes greatly towards obtaining a more uniform product.

Electronic equipment in a sawmill

In order to illustrate some of the roles electronics can play in a sawmill, the complete production

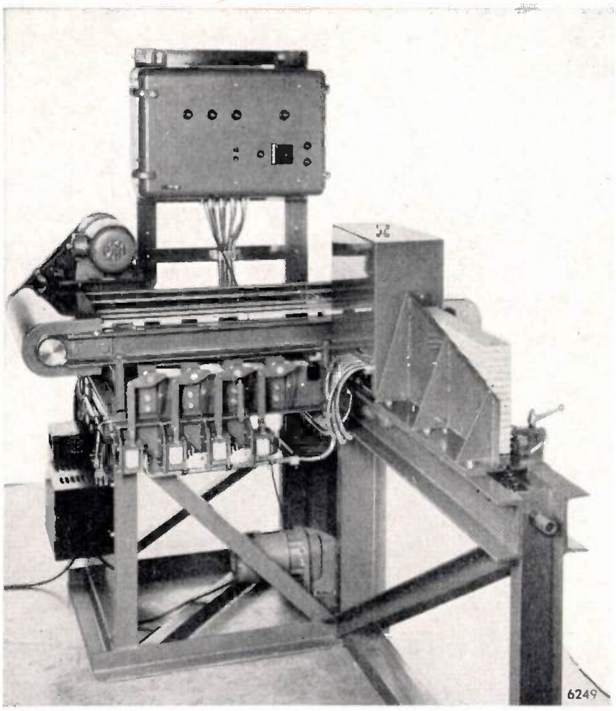


Fig. 13. Equipment for sorting parquet blocks installed for Limhamns Träindustri at Ronneby.

process in such a plant will be described, with special reference to the sawmill of Domsjö Såg at Örnsköldsvik in the northern part of Sweden, which is part of the group Mo och Domsjö AB. This mill, one of the largest of Sweden, with an annual output of 17 000 standards [5]), has been using two large electronic installations for sorting logs and boards for the past few years.

The logs arriving at the plant from the forests have to undergo three main operations: debarking, sawing and drying. The initial stages of this sequence at Domsjö Såg (and also at other large sawmills) are largely based on the availability of a large harbour basin, since the logs are more easily manoeuvred in water than on land. A large part of the logs to be sawn actually arrive at the mill in enormous floats (title photograph). The logs arriving by truck — about half of the total number at Domsjö Såg — are also usually unloaded into the harbour basin. Freezing of the harbour basin during the winter is therefore a serious though unavoidable drawback of the geographical situation of most Swedish sawmills. In order to postpone freezing for as long as possible, the water is stirred in some cases.

Let us now follow the logs on their way through the plant (see the "flow diagram", *fig. 14*). When the logs have been taken from the truck or the float, they will be put on a conveyor belt carrying them to the debarking machine. Before this is done, half of them must be turned round: they must enter the debarking machine narrow end first, but during transport half of the logs point one way and half the

[5]) 1 standard is equal to 4.672 m³ and corresponds to 40-50 logs.

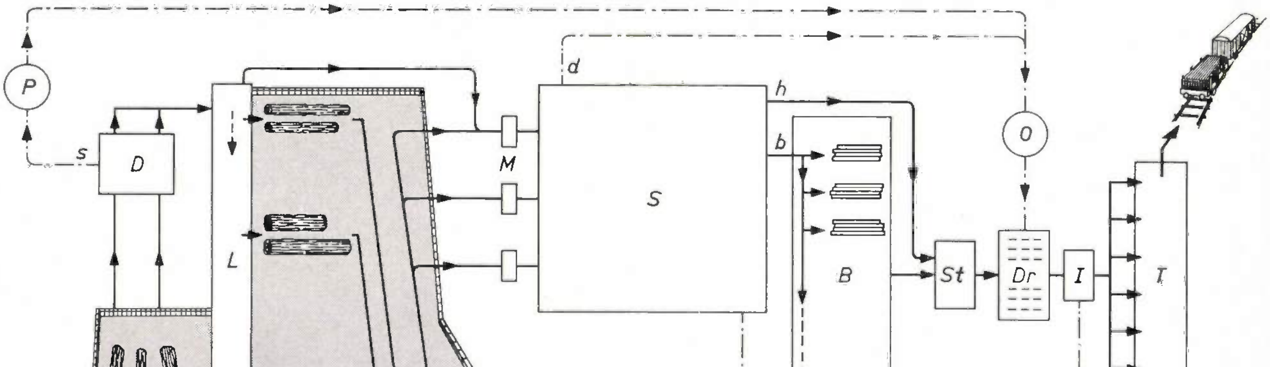


Fig. 14. Flow diagram of the sawmill of Mo och Domsjö AB at Örnsköldsvik. *D* debarking machines. *L* log-sorting installation. *Mag* magazine for sorted logs. *M* metal detectors. *S* sawing lines. *B* board-sorting installation. *St* stacks of boards. *Dr* drying plant. *I* final inspection. *T* storehouse for boards. *P* press for bark *s*. *O* boilers using dry bark and sawdust *d* as fuel. *h* heart planks, *b* boards, *p* scrap wood for pulp. The process in the block marked *S* is explained in the text with reference to fig. 16.

other. The bark removed during debarking is compressed and used for fuel. The logs coming from the debarking units (two at Domsjö Såg) are put on a common platform where their diameter at the narrow end is measured. This measurement places each log in one of 23 thickness classes, corresponding to 23 bays arranged on a line 200 m long in the harbour basin. The logs are then put on a conveyor running along these bays and are automatically unloaded into the right bay by means of electronic sorting equipment. This equipment will be described below in some detail. About 3400 logs are sorted in an 8-hour shift in this way at Domsjö Såg. When a bay has been filled with logs, they are tied together with chains and taken away by a tug-boat for storage in a magazine occupying part of the harbour basin until their class is selected for sawing.

The logs to be sawn are towed or floated to the conveyor belt for the frame saws. Domsjö Såg is equipped with three complete sawing lines, each fed by a separate conveyor. Before entering the sawing line each log is passed through an *electronic metal detector* for indicating logs containing large nails or other metallic objects which might damage the saw blades (*fig. 15*). (In some sawmills, the metal detector is situated before the debarking machine.)

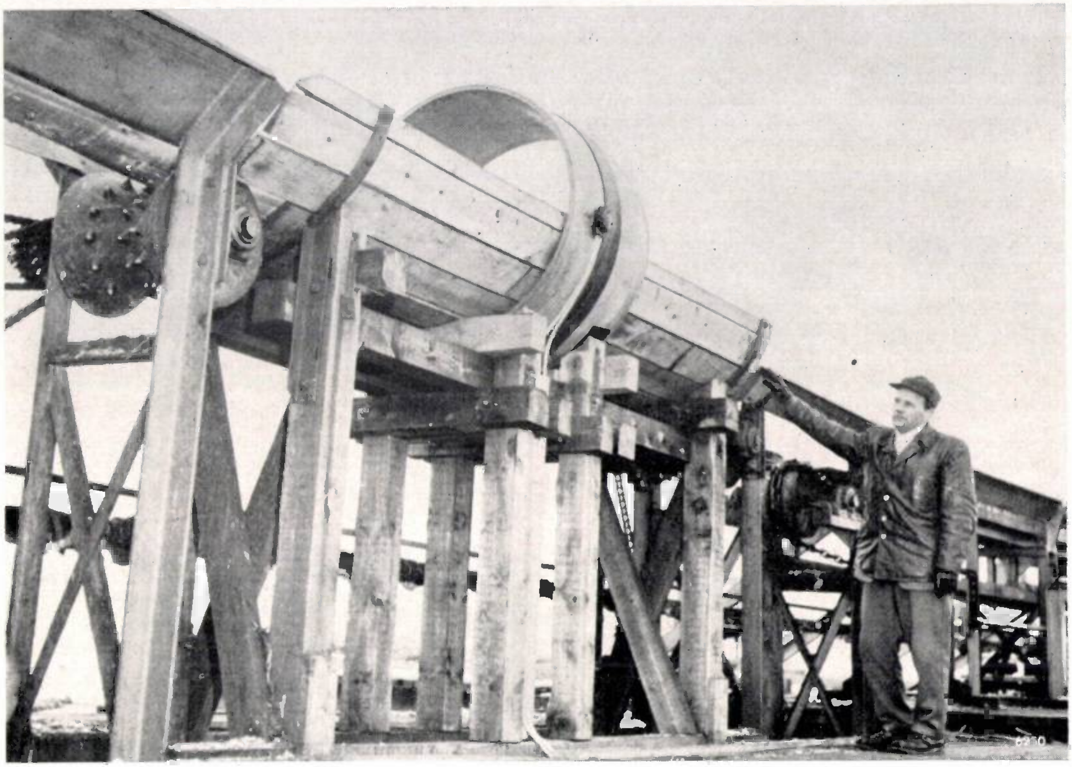


Fig. 15. Metal detector, in use in the sawmill of Tunabergs Trävaru AB at Koppartorp. The logs on their way to the saws are transported on a non-metallic conveyor through the coil 1 metre in diameter seen in the centre of the picture. Logs containing big nails or other metal objects that might damage the saws change the inductance of the coil during their passage and thereby actuate a relay giving an alarm signal. (The relay in some installations causes a paint gun to produce a spot on the log at the location of the metal, or it causes the log to be put off the conveyor.)

Although the principle is well known and has been applied in many other fields (see e.g. E. Blasberg and A. de Groot, Metal detectors, Philips tech. Rev. **15**, 97-104, 1953/54), the design and the installation for the present application is relatively difficult: because of the large coil diameter required and the small dimensions of the objects to be detected, the apparatus must be made very stable electrically and mechanically. The latter requirement is quite a problem in a sawmill.

After sawing, the planks or "boards" are sorted according to their thickness and width. This is done in another large sorting plant, using an electronic control system, which will also be described below. Next, the boards are loaded on to special trucks carrying 3 standards each and transported to the drying plant. The importance of artificial drying in fighting blue stain and mould has already been stressed in the first section of this article. The drying plant of Domsjö Såg operates on hot air and contains 8 channels, each taking 15 stacks of boards. The plant has a capacity of 70 standards per 24 hours.

After cooling down, the dried boards are transported to a platform, where a final inspection takes place, poor ends are cut off and the boards are marked according to quality and size. They are then taken to the storehouse on the waterfront, from where they will be shipped in due time.

We shall presently describe the log-sorting plant. Before doing so, however, we must explain what purpose is served by such a gigantic sorting operation. To understand this, we must first consider the sawing strategy. This strategy most visibly influences the course — and the economics — of the whole production process.

The usual method for most large sawmills is "block sawing"; see *fig. 16.* A log, of length varying between say 9 and 21 feet, is put on a carriage and is passed through a reciprocating frame saw containing a number of parallel saw blades, which simultaneously cut a number of boards from two opposite sides of the log. These boards originating from the rounded parts must be severely cut down in subsequent "edging" machines, and they will usually finish up shorter than the log because of the taper of the tree. The remaining main part of the log is turned through 90° and passed through another frame saw. Again, only thin and relatively short and narrow boards can be obtained from the rounded sides cut off by the outermost saw blades. The spacings of the saw blades in both frame saws should be chosen such that the heavy "heart" planks (fig. 16*a*), which do not need subsequent edging and represent the most valuable part of the log, will be as wide and thick as possible, while at the same time the sides of the log should also be used to the greatest advantage, with the least possible amount of scrap wood. The best compromise for a given taper and average "crookedness" of the relevant type of tree is a matter of experience — although ideas of putting electronics to work for this task too are being considered.

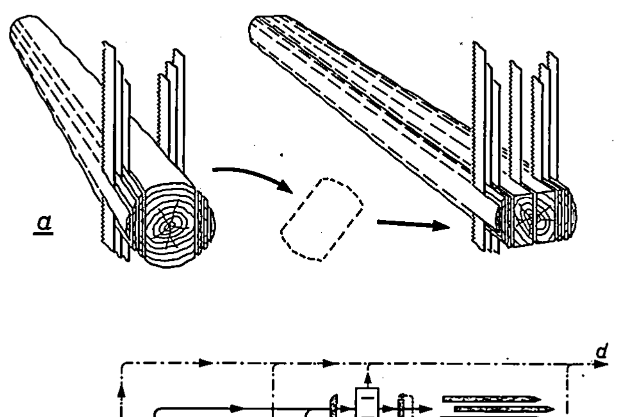


Fig. 16. *a*) Block sawing. A log is first passed through a frame saw which simultaneously cuts a number of thin boards from both sides (left); then it is turned through 90° and passed through a second frame saw (right). The thinner boards must be "edged" and "adjusted". The whole process is diagrammatically represented in figure (*b*), which represents the contents of block *S* in the flow diagram of fig. 14. *M* metal detector. S_1 and S_2 frame saws. *E* "edging" saws. *A* "adjusting" saws. *h* heart planks, *b* boards, *p* scrap wood for pulp, *d* sawdust.

It follows from the above that the spacings of the saw blades in each frame saw must be carefully adjusted in accordance with the *thickness of the log.* The sawmill must of course deal with logs of widely varying diameter, ranging e.g. from 5 inches to 18 inches. In order to avoid frequent changing of the adjustment of the saw blades, the logs are sorted into a number of thickness classes and one sawing line is fed logs of one thickness class for at least one shift, in some cases even for several days on end.

The sorting of the logs

A sketch of the lay-out of the log-sorting installation is given in *fig. 17. L* is the sorting platform where the logs are measured. The operator is seated at a desk carrying 23 buttons which correspond to the 23 thickness classes [6]) into which the logs must be sorted (*fig. 18*). In an adjacent cabinet a paper strip about 10 cm wide, on which 23 parallel "tracks" are available, moves beneath a row of 23 punches at a very low speed, about 2.5 mm/sec (*fig. 19*). The movement of the paper strip is geared directly to that of the conveyor belt running along the platform (the speed of the conveyor belt is about 1.2 m/sec). As soon as the thickness class of a log has been determined — let it be class number *k* —, the log is pushed off the platform on to the moving conveyor, and the operator presses the button *k.* This causes a hole to be punched in the paper strip by punch number *k.* A photocell is placed above each of the 23 tracks of the paper at such a distance from

[6]) A few of these are in reality special categories, such as "crooked" or the like, for logs which have to be diverted from the normal course of sawing.

the corresponding punch that the punched hole will reach the cell and let light fall on to it at the exact moment that the log on the conveyor will have reached the corresponding bay (k). For the most distant bay this will only happen as much as 3 minutes after the log was put on to the conveyor. The current of photocell k then energizes a relay causing a pneumatically or electrically driven lever at bay k to push the log off the conveyor belt into the bay.

The paper strip obviously acts as a delay system for bridging the varying time intervals which pass before each log arrives at the right bay. This mechanical delay system, unusual though it may seem in electronic equipment, is very simple and offers the important advantage that it does not involve any real time relationships but only place relationships: if the conveyor belt is stopped for some reason, the logs resting on it will still be delivered to the right bays when it starts up again.

An interesting detail of the equipment is the fixing of the exact moment when the hole for a log

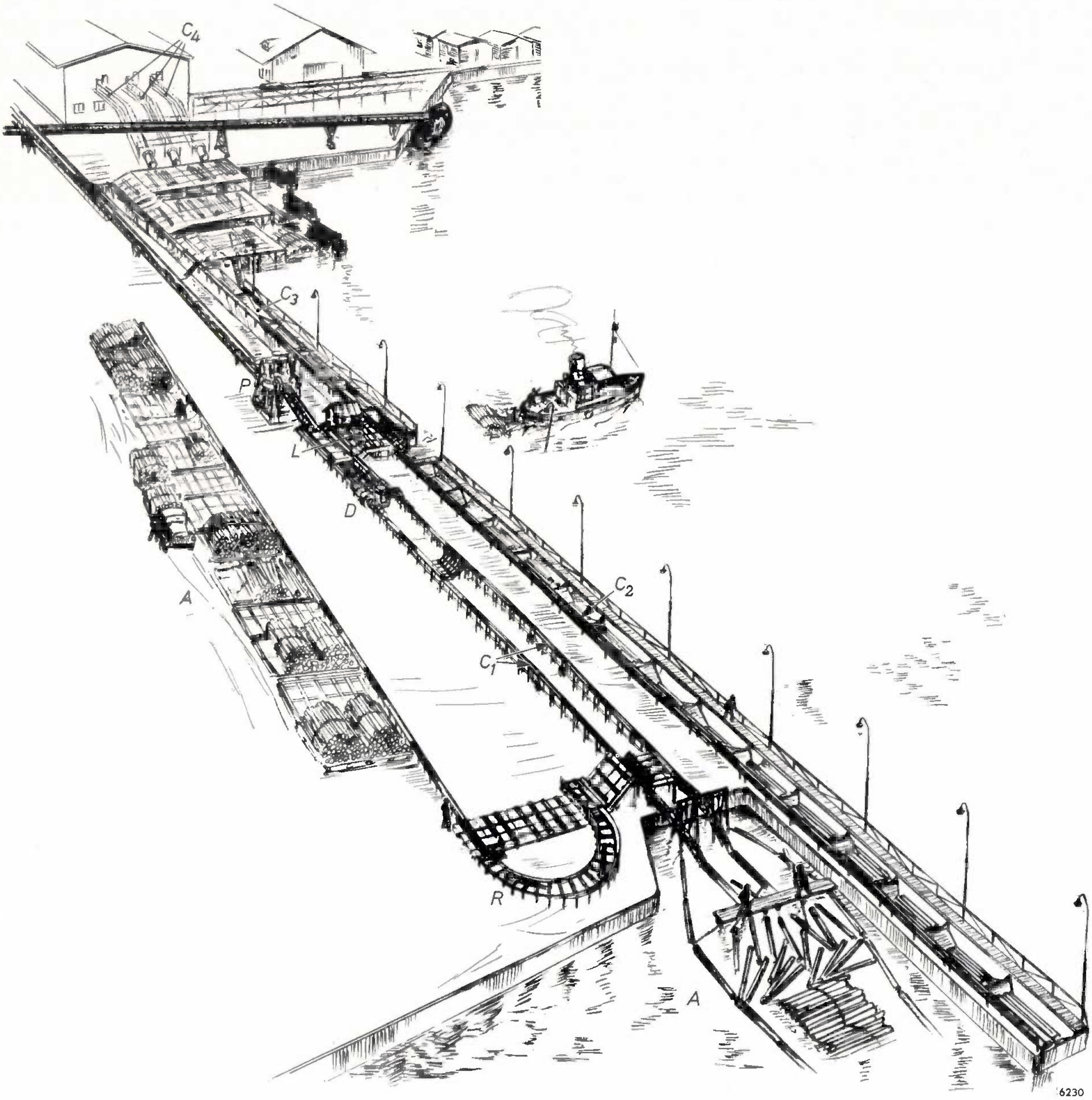


Fig. 17. Log-sorting area of Domsjö Såg. A arrival of the logs by truck (especially in wintertime when the harbour is frozen) or by float. R reversal of half of the logs (shown here as being done on land; more usually done in the water). C_1 conveyors to the debarking machines D. P machine for compressing the bark. L platform where the logs are measured. C_2 conveyor running along the bays for the different classes of logs. (Some of the bays are situated on the other side of L. The logs for these bays are placed on the conveyor C_3, which also carries logs directly to the sawing lines; this trivial complication has been neglected in the further description.) C_3 conveyor running to the sawing plant. C_4 three separate conveyors, each feeding logs to one of the sawing lines. — The conveyors at Domsjö Såg were made by Bahco Erenco AB.

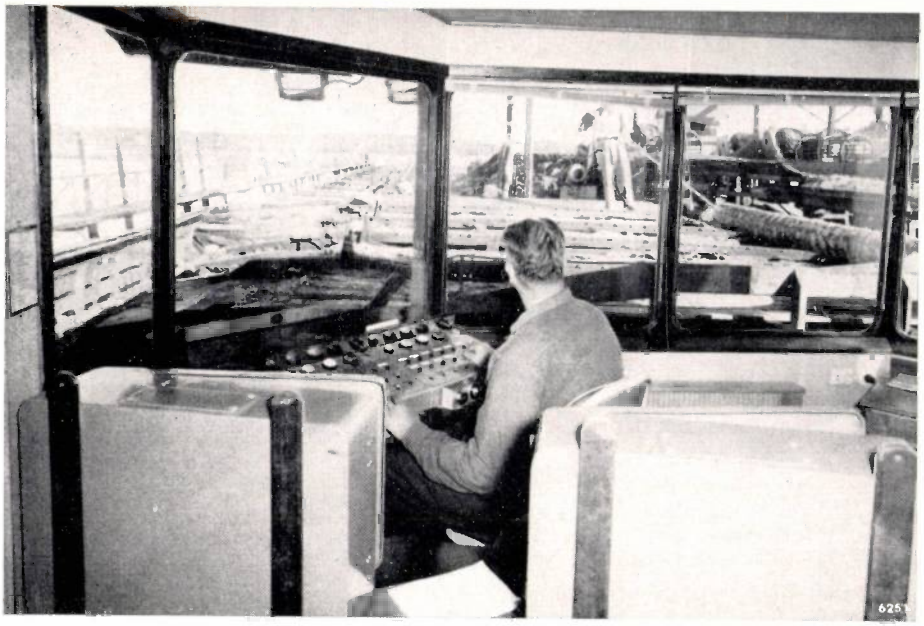


Fig. 18. Operator at desk containing the 23 buttons for directing the logs to their respective bays. The logs arrive from the right-hand side on the platform visible through the window. To the left the conveyor with the 23 bays.

has to be punched into the paper. This moment must obviously be governed not by the operator's action in pressing the button but by the actual arrival of the log at a certain point of the conveyor, say point *C* in *fig. 20*. Moreover, the *length* of the log, which may vary from 9 feet to a maximum of 21 feet, must be taken into account. The bars which push the logs into the water must not be more than about 9 feet long, as otherwise there would be a risk of one bar pushing two logs at the same time. If, on the other hand, a bar of length 9 feet should push a 21-foot log at a one end, this log would fall into the bay at an angle, thus seriously hampering the tying together of the logs in neat bundles. The electronic sorting equipment therefore contains a device which ensures that the hole for each log is punched into the paper at the moment when the mid-point of the log passes the fixed point *C*, so that the pre-determined lever will be actuated at the moment when the mid-point of the log will pass it.

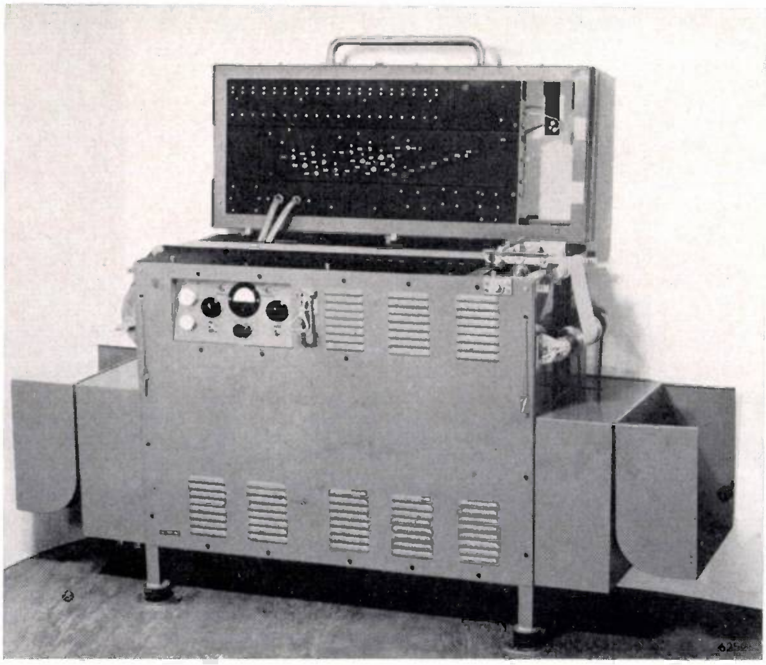


Fig. 19. Cabinet with electronic circuits and the moving paper strip acting as a delay system. A series of punches and the corresponding photocells may be seen arranged on oblique lines in the middle of the lid, which is opened to show the interior. The seemingly haphazard arrangement is in fact chosen with regard to the available space; only the distance between each punch and the photocell on the same "track" matters. (Even with a judicious arrangement, 23 punches cannot be accommodated by one strip. The actual installation for 23 bays therefore comprises two complete delay cabinets. This trivial complication has been disregarded in our description.)

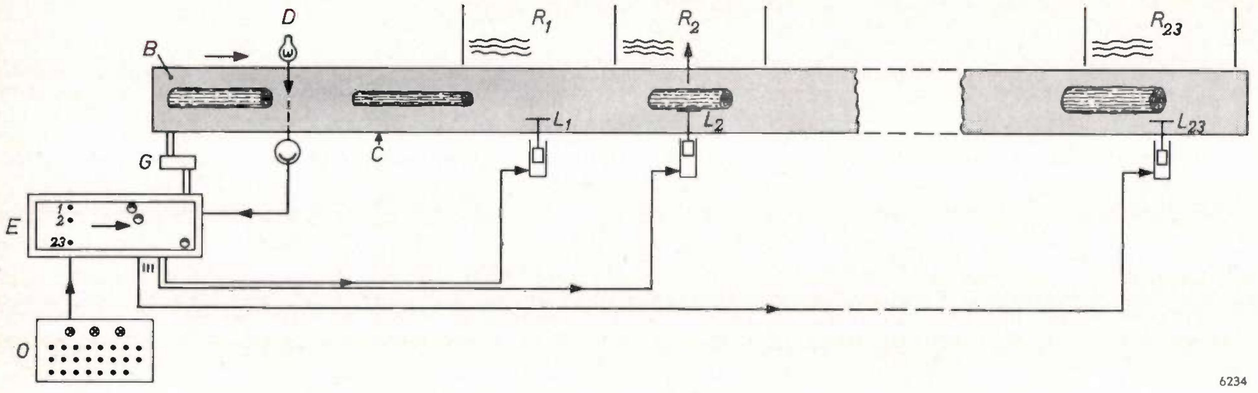


Fig. 20. Log-sorting installation. *O* operator's desk with 23 buttons. *E* cabinet with electronic circuits and paper strip. *G* gearing by which the movement of the strip is coupled to that of the conveyor belt *B*. R_1-R_2- . . . -R_{23} bays, with pneumatically driven levers L_1-L_2- . . . -L_{23} for dropping the logs. *D* lamp and photocell of mid-point-seeking device.

This device will be described with reference to *fig. 21*. A beam of light falling on to a photocell is directed across the path of the logs at point *D*, situated a distance equal to half the maximum length of a log (i.e. $\frac{1}{2}\times 21' = 10\frac{1}{2}'$) before *C*. The beam and hence the photocell current will be interrupted by a passing log, and will reappear when the tail of the log arrives at *D*. At this moment the head of a log of the maximum length $l_{\max}$ (21′) would be at *E* and its mid-point at *C*, but the head of a shorter log would only have got to *E′*, say, and its mid-point to *C′*. The reappearance of the current cannot thus be used for controlling the actual punching after the operator has pressed the button. Instead, a switching disk is incorporated in the delay-system cabinet, which is started rotating at an angular velocity ω by the disappearance of the photocell current (arrival of the head of the log at *D*) and which switches on the punch after rotating through a certain angle Θ. At the speed ω, the rotation through Θ would be completed when a 21-foot log had the (correct) position *DE*, but the shorter log considered above would then have travelled too far and reached the position *D″E*. The speed of the disk is however *doubled* by the reappearance of the photocell current (tail of shorter log at *D*, head at *E′*). The remaining part of the rotation through Θ is therefore completed in a shorter time, namely such that the shorter log will have arrived at the position *D‴E‴*, exactly half-way between the positions *DE′* and *D″E*, and its mid-point will now be at *C*, where it should be at the moment of punching.

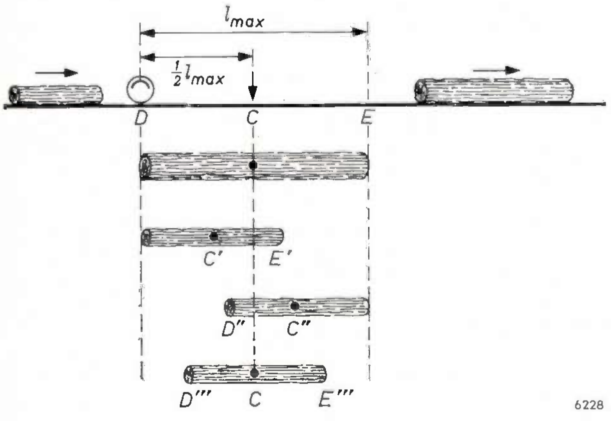


Fig. 21. Diagram for explaining the mid-point-seeking device.

When two relatively short logs are put on to the conveyor in quick succession, it may happen that the head of the second log will arrive at *D* before the switching disk has completed its full rotation for the first log. Another switching disk is therefore provided, the two disks being used alternately by means of an interlocking circuit.

A few more details of the log-sorting installation may be mentioned briefly. The operator has a special button at his disposal for cancelling his last decision, so that — within a certain lapse of time — he can correct an error. Pilot lights indicate the state of affairs. Another special button takes care of the logs in the thickness class being sawn on that day: this button actuates a set of "points", throwing the log on to a subsidiary conveyor belt as soon as it comes off the platform. It is then carried not to the sorting bays but in the opposite direction, to the sawing lines (fig. 17). All the logs sent in either direction are automatically counted. The paper strip of the delay system is stored on a 250-m spool, sufficient for approximately three 8-hour shifts. A sensing lever pressing against the paper on this spool will sound an alarm when a length of paper sufficient for only 15 minutes is left.

The measuring of the diameter of the logs used to be done by means of calipers, or simply by visual estimation after the operator had gained some experience, but an *electronic gauge* has recently been developed for the purpose and installed at one sawmill (*fig. 22*). This instrument quite considerably improves the speed and reliability of the measurement.

Let us now turn to the sorting plant for the boards.

The sorting of the boards

We have already mentioned (page 39) that in block sawing, the spacings of the saw blades in each frame saw must be carefully adjusted in accordance with the thickness class of the logs being sawn, to make the best use of the wood (cf. fig. 16a). A given setting of the frame saw will give boards of one, two or three different thicknesses in addition to the heavy heart planks. Moreover, the boards of a given thickness will differ in width, depending on how near the edge of the log they were cut, and on the varying degree of edging necessary to remove the curved sides. For one setting of the saw as many as 24 different types of board — different combinations of thickness and width — may be obtained (disregarding the heart planks) [7]. These boards, emerging from the saws in a random sequence (at a rate of about 1 per second at Domsjö Såg), must be sorted in types, not only with a view to their final delivery to the customer, but primarily in order to enable them to be *stacked* in regular fashion for the process of *drying*: only boards of one type can be combined to form a regular lattice, which is desirable for easy handling and controlled drying conditions in the hot-air chambers. Fortunately, the heavy heart planks need not be sorted since their width and thickness are fixed as long as logs of one diameter range are being sawn. All the thinner boards, after having been checked for visible defects, are transported by a single rapid belt conveyor (speed 210 m/min) to the sorting plant (*B* in the flow diagram of fig. 14).

[7] Differences in length, caused by the different length of the logs and by the variable edging necessary for removing faulty and tapered ends, are not essential to these considerations.

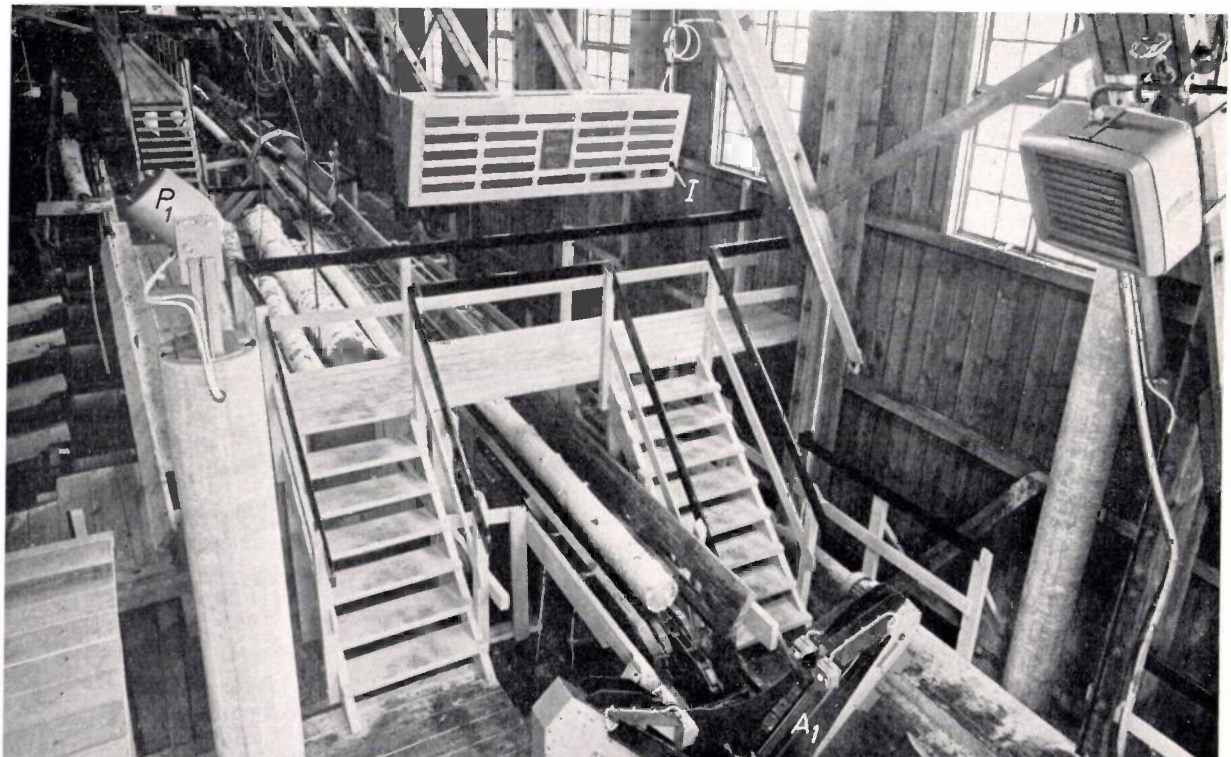


Fig. 22. Electronic minimum-diameter gauge, installed at Skutskärsverken, Skutskär. In each of two arms A_1 and A_2, which form a letter V, a belt with a transverse slit runs along a series of 160 small lamps. The light of these lamps successively transmitted by the slit falls on a photocell mounted opposite each arm at a distance of 3.5 m (P_1; the other photocell cannot be seen) and produces a series of pulses. Fifteen pulse trains per second are produced by each arm and are fed to a counter with two registers, where they arrive alternately. A log which is passed between the two arms will screen a number of lamps from each photocell and thus decrease the number of pulses in each train. The number of pulses missing from these two trains is a measure of the thickness of the log in two directions at right angles to each other. Two consecutive pulse trains arriving at the counter are compared, the larger of them is stored in one of the registers and compared with the next pulse train appearing in the other register, and so on. The maximum thus found after the log has passed gives the minimum diameter. The accuracy is better than $\pm \frac{1}{4}$ inch. The gauge is provided with 22 relays giving output signals used for automatic sorting of the logs and for operating a display unit (*I*) which was used in the initial period after the installation of the equipment to provide a check on its operation.

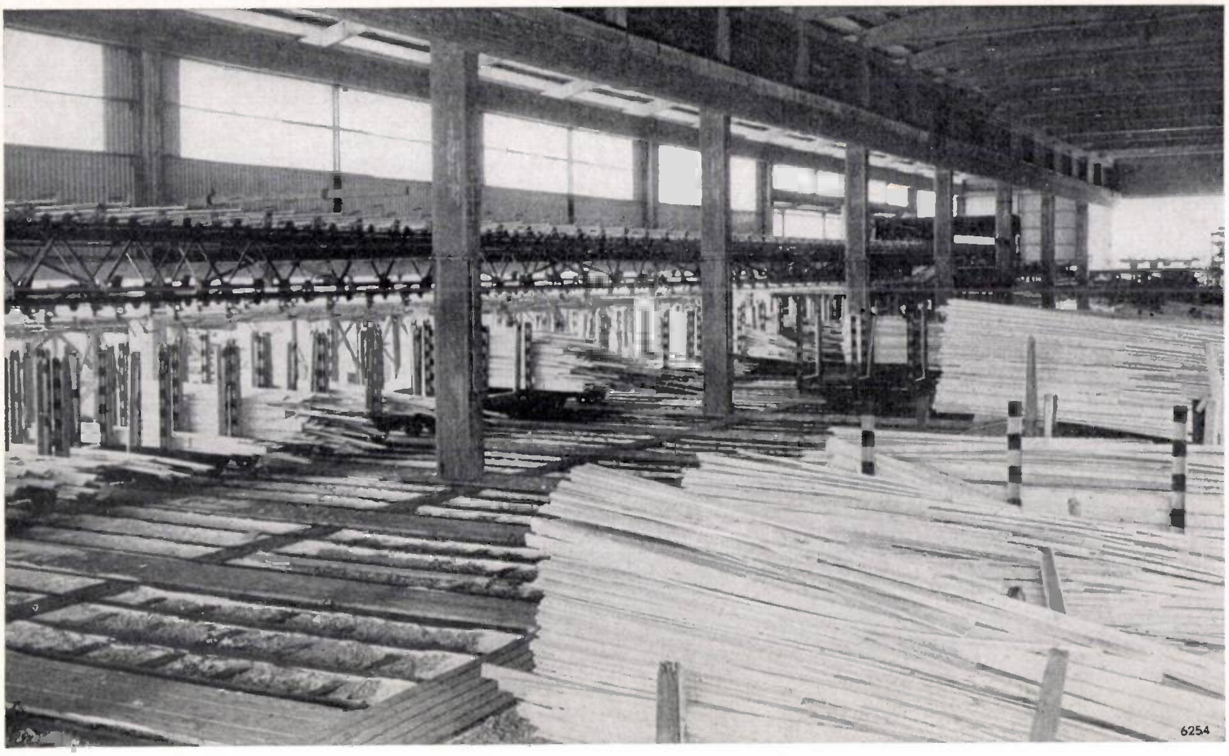


Fig. 23. Board-sorting installation of Domsjö Såg. The planks are dropped from the large conveyor system on the left into the respective bays, where they are collected on trucks. A number of trucks which have already been filled and removed are seen at the right-hand side.

The most conspicuous part of the board-sorting plant at Domsjö Såg is a huge conveyor system 65 m long and 5 m wide, by which the boards are transported, in a direction perpendicular to their length, over a series of 30 bays; see *fig. 23* [8]). In this system each board is held by a suspension unit containing several hooks fixed on a spring-loaded shaft (*fig. 24*). The shaft carries three disks which can be slid along it to various positions. Above each bay is a release mechanism containing a lever, which is at a different position for each bay. When one of the disks of a suspension unit passing overhead hits this lever, the shaft of the suspension unit is released from its spring, so that the hooks will drop the board into a truck in the bay.

Now, the actual sorting is effected by equipment which automatically measures the width and thickness of each board arriving at the conveyor

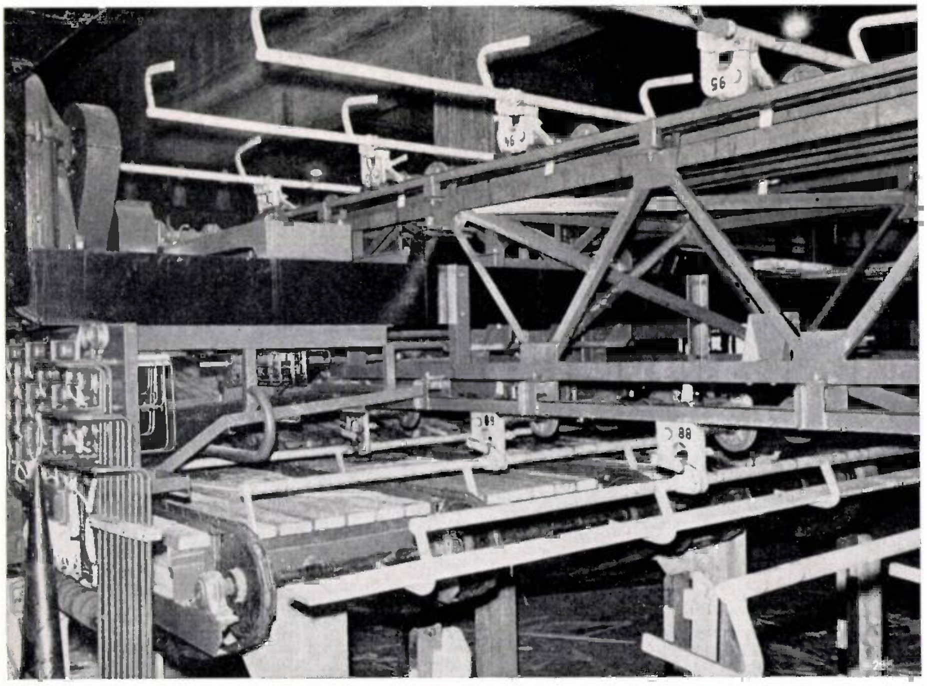


Fig. 24. First part of the conveyor system with suspension units. Unit 88 is carrying a board. On the upper side the units are returning to the starting point to the left in the photograph. The disks carried by each suspension unit and actuating the release mechanism of the respective bays are automatically shifted to the desired position by one of 30 relay units visible to the extreme left.

[8]) The mechanical system was constructed by AB Nordströms Linbanor, the chief contractor for the board-sorting installation.

system and which, depending on these measurements, energizes one of 30 relay units. Each of the relay units controls a solenoid valve with a hydraulic ram in such a way that one of the disks of the simultaneously arriving suspension unit, which is going to take over the arriving board, is pushed to a coded active position [9]. (Just before this happens, the disks of this suspension unit are automatically reset to a zero position.) The functioning of this equipment can be explained in more detail with reference to *fig. 25.*

The *width* of the moving board is measured by means of a photocell arrangement which delivers a number of pulses proportional to the width of the board (20 pulses per inch). The pulses are fed to an active cold-cathode counter [10]) which, dependent on the number of pulses received, will give a positive pulse on one of 14 lines corresponding to 14 different widths varying from say 2 to 11″. At the same time the *thickness* of the board is measured by a simple mechanical gauge, which controls a stack of 14 three-way switches connected to the above-mentioned lines. The positive pulse is thus directed to one of 42

[9]) It would be difficult to make a ram long enough to shift a disk to any one of 30 different positions on the shaft. This is the reason why three disks are used, two of which remain in their zero position while the third is shifted to one of 10 active positions.

[10]) See e.g. F. Einramhof and P. Havas, Philips tech. Rev. 21, 309, 1959/60.

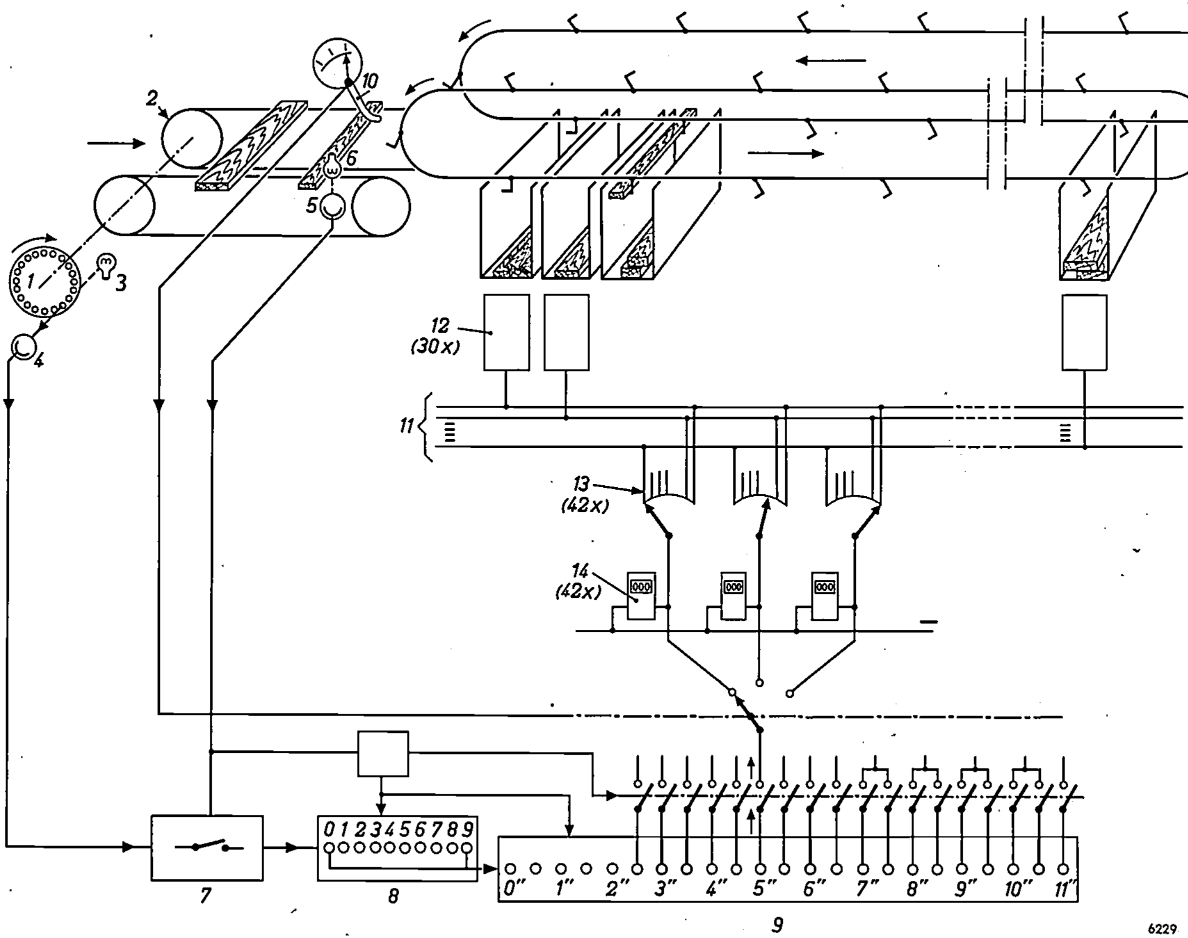


Fig. 25. Sketch of the board-sorting system. The perforated disk *1* is coupled to the conveyor *2* and produces a continuous series of pulses by interrupting the light of lamp *3* falling on photocell *4*. The number of pulses is directly proportional to the distance covered by the conveyor. A board arriving at photocell *5* will interrupt the light of lamp *6* and thereby open the gate circuit *7* where the aforementioned pulses arrive. The number of pulses transmitted by the gate is a measure of the width of the board — irrespective of the speed of the conveyor. After every ten pulses (corresponding to 0.5″), the active cold-cathode counter *8* will deliver a pulse to the shift register *9*, which has fourteen output lines. The reappearance of the light on photocell *5* will produce a signal closing the gate *7* and causing a positive pulse to appear on the last attained output line of *9*. This output line (which thus corresponds to the width range of the board being measured) is connected to one of three sub-outputs, selected by the mechanical thickness gauge *10*. Each of the 42 outputs thus available can be connected by means of a jack-and-cord switchboard *11* to one of the 30 relay units *12* which control the hydraulic setting of the disks on each suspension unit. Each output is also provided with a selector *13* bypassing the switchboard. The selector is actuated by a negative pulse from the corresponding preset counter *14* for selecting a free relay unit when a bay has been filled.

outputs, arranged on a switchboard. Each of these 42 outputs can be connected to each of the 30 relay units by a jack and cord (*fig. 26*).

In reality, not more than 8 width classes (24 types of boards) are to be expected on a given day, as explained before. Only 24 outputs will therefore actually be used at one time.

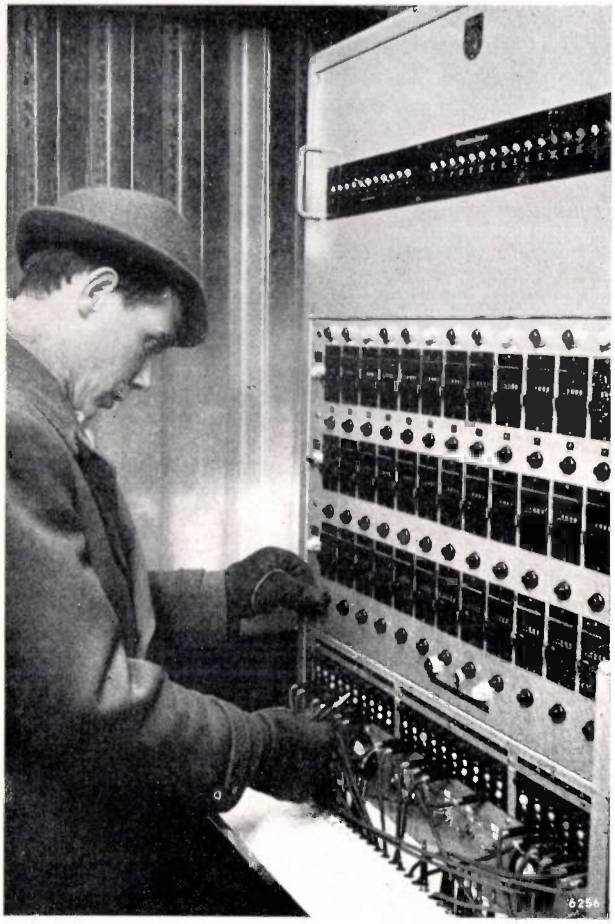


Fig. 26. Cabinet with 42 outputs (below) of the device for measuring the width and thickness of the boards. Each output can be connected by jack-and-cord with each of the 30 relay units for the 30 bays (no more than 24 types of boards will occur during one shift owing to the established sawing strategy). Above these inputs with their signalling lamps are the 42 preset counters for the automatic selection of a free relay unit when one of the bays has been filled.

When the truck of a bay has been filled to capacity, the bay must be put out of use until the personnel have emptied the bay and put another truck in it. To prevent difficulties arising from bays being filled too high, an automatic changing system has been provided. Each of the 42 outputs on the switchboard is provided with an electro-mechanical counter, which can be preset to the number of boards of the pertaining type that can be collected in one bay. When this preset number is reached the counter will deliver a negative pulse which starts a rotary 30-step switch (cf. fig. 25) automatically selecting a relay unit (i.e. a bay) not being used at the moment. The following boards of this type thus will be discharged into the new bay. At the same time the counter is automatically reset to zero, a lamp is lit on the switchboard to indicate the relay unit now occupied and another lamp is lit at the bay which has been filled, warning the personnel to empty the bay.

Emptying a bay will take some time and other bays may become filled in the meantime. It is, however, unlikely that more than 4 or 5 bays will be waiting to be emptied at any given moment. Thus, 30 bays (30 relay units) are sufficient for the 24 different types of boards.

The board-sorting installation can be operated manually, in the eventuality that the electronic equipment should fail. This precaution was necessary in order to ensure the continuous flow of logs and boards through the plant under all circumstances. No use has had to be made of this facility at Domsjö Såg so far.

Let us now have a closer look at the changes brought about in the production process by the application of electronics. Log sorting in accordance with the sawing strategy used to be done, in a cruder form and manually, before electronic means were available. Electronics is saving labour and improving accuracy here too, but as in the other cases other very important advantages are also obtained. In the manual process of sorting, the floating logs were grouped by simply pushing them to the desired bay, and the available water surface was exhausted when it was covered by a layer one log thick. In the modern process, the logs are collected in bundles of 100-300 which are tied together, covering the water with a layer several logs thick, on the average. Many more logs of each class can thus be kept in the storage space of a given harbour basin. This saving of storage costs is similar to that found in the shoe-last factory, but that is not all. The possibility of feeding one single class of logs to all the sawing lines during at least one shift ensures that not more than 24 types of boards will come out of the saws during this period. Had this not been so, the very large and expensive board-sorting plant would have needed more than 50 bays instead of 30, in order to cope with all the 42 possible types of boards at the same time. The enormous hall and mechanical conveyor system necessary for this number of bays would have more than doubled the investment costs for this operation.

Tying the logs together in bundles, rendered possible by the sorting process described above, can bring another unexpected advantage. Logs lose a large part of their buoyancy when debarked, so much so that with some types of trees 20% of the logs will sink after some time and disappear to the bottom of the harbour basin. These logs ("sleepers") have to be regained by regular dredging operations and fitted with floating planks of light wood in order to bring them to the conveyors for the saws. When tied together, however, the logs still capable of floating — 80% of the total — will keep the other ones afloat too, and no difficulty is experienced in bringing the whole bundles to the saws. Logs that sink when a bundle is untied can easily be spotted and put directly on the conveyor.

The applications of electronics described in this article will have made it clear that the potentialities of electronics are well appreciated and widely used in the wood industry. There is no doubt that more and wider applications in this field will be found, especially in Sweden, where close collaboration between the' wood industry and the electronic industries is now firmly established.

Summary. Description of a series of relatively new applications of electronics in the Swedish wood industry, with emphasis on the character of the changes effected in manufacturing processes. High-frequency gluing is mentioned only in passing, since this has been common practice for a number of years. High-frequency drying introduced in a factory making shoe lasts at Järrestad has reduced storage costs to 15% of their former value and at the same time diminished the rejection rate because of cracks from 8-10% to 1-2%. In a match factory at Vetlanda, strips of veneer for match boxes are inspected for defects at a rate of 400 strips per minute by a photoelectric scanning device, reducing the manual labour involved in the total manufacturing process by 10%. A similar photoelectric device developed for sorting oak parquet blocks into four categories in a factory at Ronneby has resulted in a notable improvement of the quality of the product, in addition to saving labour. The important role electronics is beginning to assume in large sawmills is discussed at some length. A log-sorting plant for 3400 logs daily and a board-sorting plant for 10 times this number of boards, both in use in a sawmill at Örnsköldsvik for the past few years, are described and brief mention is made of a metal detector and an automatic minimum-diameter gauge for logs. Saving of labour and storage costs are again the chief advantages obtained, in addition to simplifications in the production process.

A SNOW SEPARATOR FOR LIQUID-AIR INSTALLATIONS

by C. J. M. van der LAAN *) and K. ROOZENDAAL *).

533.24:66.078

The time during which a gas refrigerating machine can continuously produce liquid air is limited by the gradual blockage of the separator in which the gaseous impurities removed from the air feed settle in the form of ice or snow. The old form of separator had to be defrosted after every 120 litres of liquid air produced, i.e. every twenty-four hours when the machine was in continuous operation. With the new design described below, which is quite different from all known designs, defrosting is necessary only after 600 litres of liquid air have been produced.

Before air is liquefied, it must be freed from impurities such as water vapour and carbon dioxide, otherwise the equipment would soon be blocked up by ice and solid carbon dioxide.

Water vapour and carbon dioxide can be removed by the use of chemicals such as silica gel or potassium hydroxide. This method has drawbacks, however: potassium hydroxide is a corrosive substance and very unpleasant to use; silica gel has to be regenerated from time to time, which complicates the installation, and moreover under tropical conditions a large drying plant is needed. A smaller drying plant might be sufficient if the air were compressed, but the use of a compressor is in itself an added complication. It consumes electrical energy, calls for maintenance, and must be of a special type in which the air is not contaminated by lubricants.

The Philips gas refrigerating machine [1]) was therefore designed right from the start to freeze out the impurities before the air is liquefied. This was formerly done in an "ice separator" (*fig. 1*), consisting of horizontal perforated copper plates kept cold by the

*) Industrial Equipment Division, Eindhoven.

[1]) J. W. L. Köhler and C. O. Jonkers, Fundamentals of the gas refrigerating machine, and Construction of a gas refrigerating machine, Philips tech. Rev. **16**, 69-78 and 105-115, 1954/55.

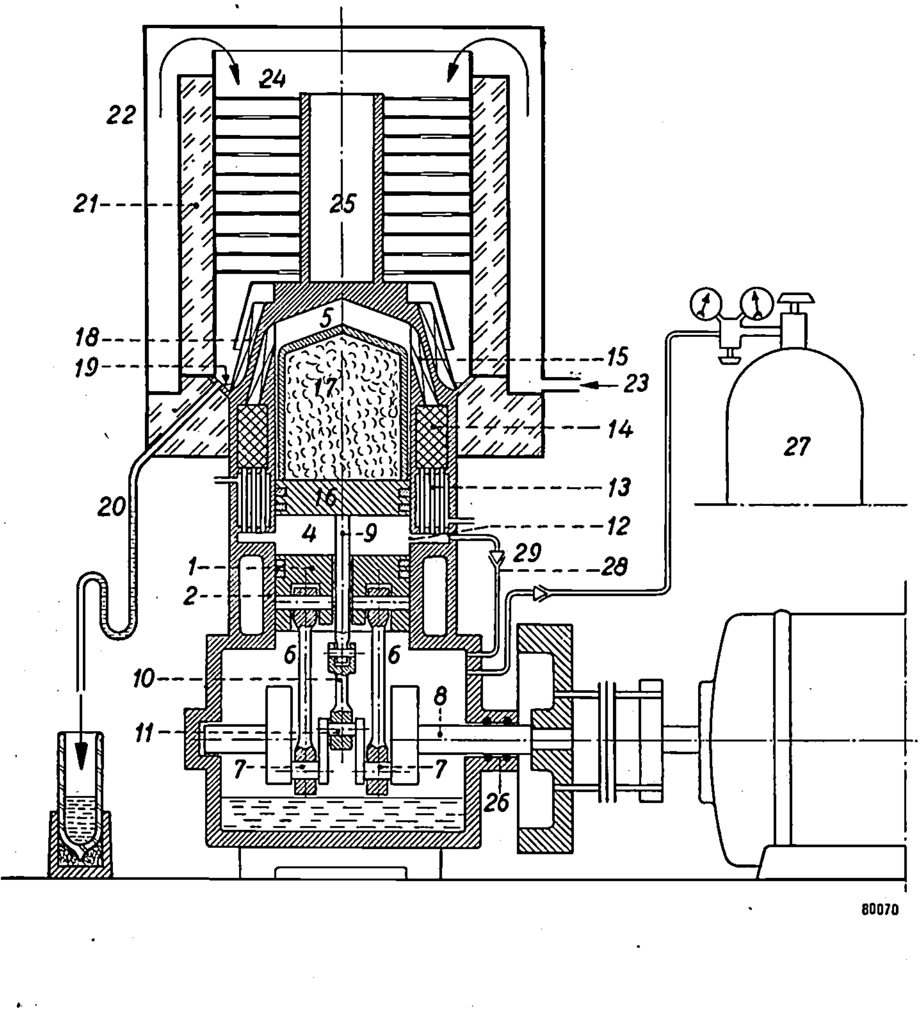


Fig. 1. Simplified cross-section of a gas refrigerating machine for producing liquid air [1]), illustrating the ice separator consisting of perforated copper plates *24*. The plates are in contact with the head ("freezer") *15* of the refrigerating machine via the tubular structure *25*. The air enters at *23* under atmospheric pressure, flows along the plates *24* where moisture and carbon dioxide are removed, is liquefied on the condenser *18*, and is tapped off via the annular channel *19* and the delivery pipe *20*. *21* is an insulating wall, *22* the jacket around the ice separator.

The significance of the other figures is: *1* main piston, *2* cylinder, *4* and *5* spaces between which the gas flows to and fro, *6* connecting rods, *7* cranks, *8* crankshaft, *9* displacer rod with connecting rod *10* and crank *11*, *12* ports, *13* cooler, *14* regenerator, *16* piston and *17* cap of the displacer, *26* gas-tight shaft seal, *27* gas cylinder supplying refrigerant, *28* supply pipe, *29* one-way valve.

machine itself. The moisture and carbon dioxide in the air formed deposits of ice or snow on these plates, and the purified air passed through the holes.

The great drawback of this type of separator is that it soon becomes clogged up, owing to the ice and snow settling principally on only one or two of the plates. This is due to the fact that the moisture content of saturated air decreases very rapidly as the temperature drops (*fig. 2*). Nearly all the moisture therefore settles on those plates whose temperature is only slightly below 0 °C. The already dry air is then further cooled by the following, colder plates, but the spaces between them are no longer filled with snow, except at the last, very cold plates, where carbon-dioxide snow is formed. In spite of the 14-litre capacity of this ice separator, about 1 kg of snow and ice is enough to cause a blockage. A machine fitted with such a separator is therefore unable to produce more than 120 litres of liquid air before it has to be stopped for defrosting.

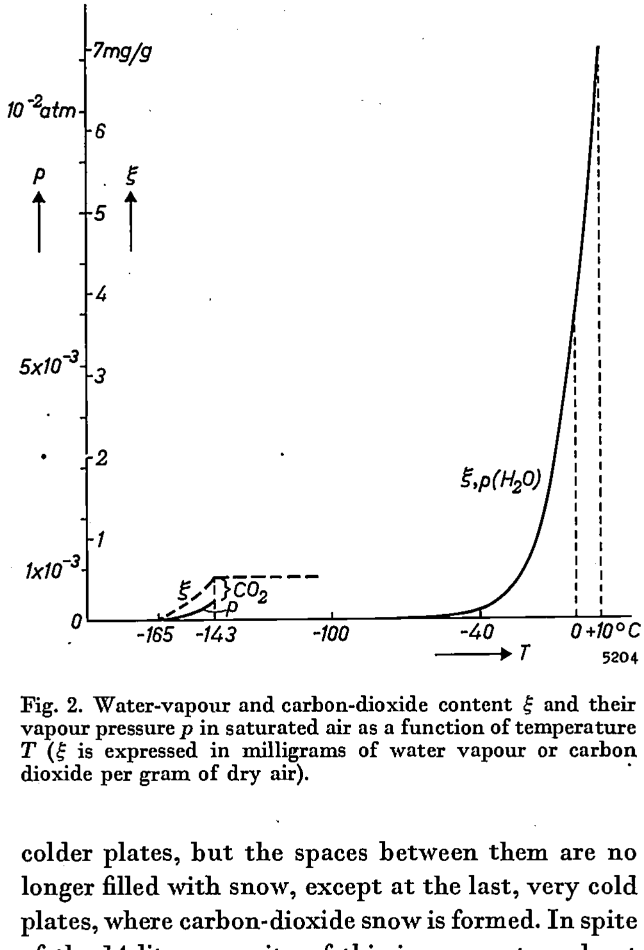


Fig. 2. Water-vapour and carbon-dioxide content ξ and their vapour pressure p in saturated air as a function of temperature T (ξ is expressed in milligrams of water vapour or carbon dioxide per gram of dry air).

Efforts to improve on this performance have resulted in a new design, which will be discussed below. The new separator traps the water vapour of the air only in the form of snow, hence its name.

The snow separator

The design of the new separator is based on a surprising fact which was discovered by chance by the authors in an investigation directed towards another end [2]). It was found that when a stream of air is passed through a gauze kept at a very low temperature the layer of snow formcd on the gauze remains porous for a considerable time. At first sight one would expect that thc layer of snow would quickly grow into a compact mass, but in fact the crystals form in such a way that this is not so.

In the new separator, whose size is no greater than that of the former type (14 litres), this phenomenon has been turned to practical use so that about 5 kg of snow can be stored without causing a blockage. This means that the machine can produce five times as much liquid air in continuous operation. Whereas the old type of separator had to be defrosted every day, once or twice a week is now sufficient.

The principal component of the snow separator is a cylinder of fine copper gauze (*fig. 3*) kept at low temperature in a manner presently to be discussed. This "snow gauze" is surrounded by a double jacket (the snow forms in the space between the jacket and the gauze). The air feed flows via the double jacket, in which it is precooled, to the snow space, passes through the gauze and is then condensed on the head of the refrigerating machine. As the air flows through the gauze, a layer of snow forms on its surface. The layer progressively thickens into a snow cap, through which the air still has to pass. A "snow cake" from which a piece has been removed is shown in *fig. 4*. This cake took a week to grow to a thickness of about 10 cm; the innermost layer, about 0.5 cm thick, consists of carbon-dioxide snow. The density of the snow is fairly high, the values measured being 0.4 g/cm^3 for the water snow and 0.9 g/cm^3 for the carbon-dioxide snow. Nevertheless, the layer is still porous, and the pressure drop across it when the machine is working at full capacity is no more than 20 cm water column.

What are the conditions to be fulfilled in order to keep the snow layer porous? To answer this question we must have a clear picture of the way in which the layer of snow is formed.

Formation of the snow layer

As long as no snow has yet formed on its surface, relatively warm air enters into contact with the gauze. To prevent the transport of the fairly large amount of moisture still present in the air feed, the gauze must be capable of cooling the air to a very low temperature. First of all, then, the gauze itself must be kept at the very low temperature

[2]) Dutch Patent No. 98 130.

of —165 °C (the temperature of liquid air is —194 °C). Secondly, to provide good thermal contact with the air, the gauze must have a fine mesh.

Once the machine is started, part of the impurities will initially settle on the gauze by diffusion, i.e. through the movement of water vapour and carbon dioxide caused by local differences in partial vapour pressure. At the surface of the cold gauze the partial vapour pressures are much lower than in the warmer air feed (fig. 2). As a result, the layer of snow grows counter to the direction of air flow. In the beginning a small proportion of the impurities forms snow crystals in the air stream itself and passes through the gauze. This is unavoidable. The rest, however, settles on the gauze and increases the heat-transfer surface area, forming a thin layer of snow pierced by numerous narrow channels. Both the heat transfer and the deposition of impurities are virtually 100% effective as soon as this thin layer is formed.

The snow first settles in the form of relatively long needles. The space between the needles is then gradually filled up with smaller snow crystals. As long as the temperature of the snow layer remains far enough below the melting point, no ice forms. Although the accumulation of smaller crystals decreases the porosity of the first layer, a fresh and very porous layer, with long needles, has meanwhile formed on the outside; this absorbs the major part of the following small crystals, thus preventing any stoppage of the first layer for a considerable time, and so the process goes on. Conditions must therefore be chosen in such a way that the snow settles mainly on the outside of the layer, thus minimizing the quantity of impurities that can settle inside.

Since the mass of an impurity transported per second is proportional to the partial pressure of that impurity in the air, steps must be taken to ensure that the temperatures in the snow layer are such that the partial pressure therein (i.e. the vapour pressure corresponding to the local temperature) is much lower than the partial pressure in the air feed. To meet this requirement, the temperature everywhere in the snow layer must be kept below a specific value, since the vapour pressure drops sharply with decreasing temperature (see fig. 2). Experiments have shown that the temperature of the outside of the water-snow layer must remain below —40 °C if the dew point of the incoming air is 10 °C. Since the outside layer is heated by the incoming air, the layer must be kept at this low temperature by conduction via the snow crystals to the cooled gauze.

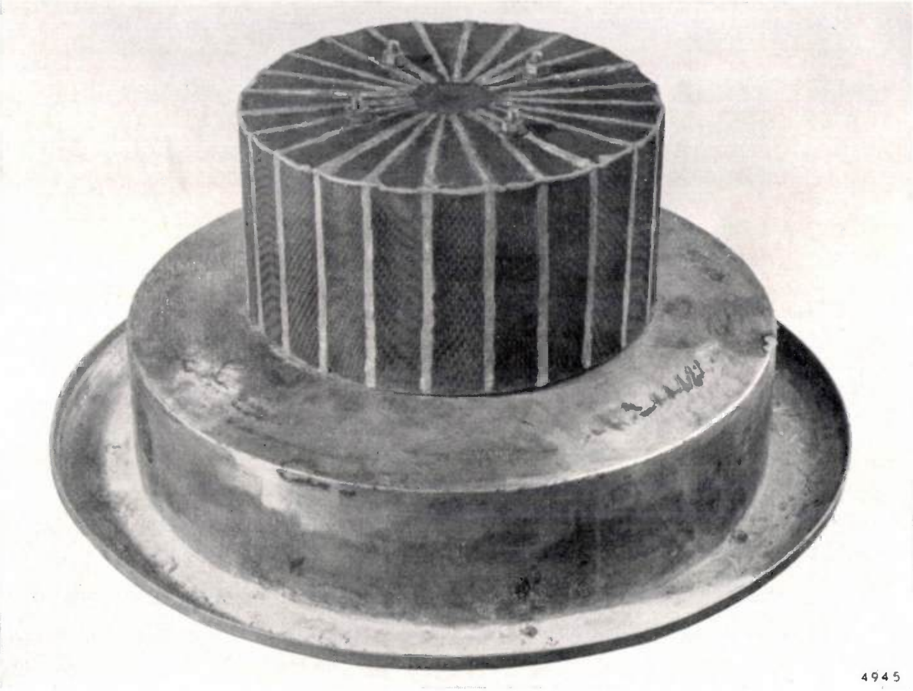


Fig. 3. Cylinder of fine copper gauze, on which water vapour and carbon dioxide from the air are deposited in the form of porous snow.

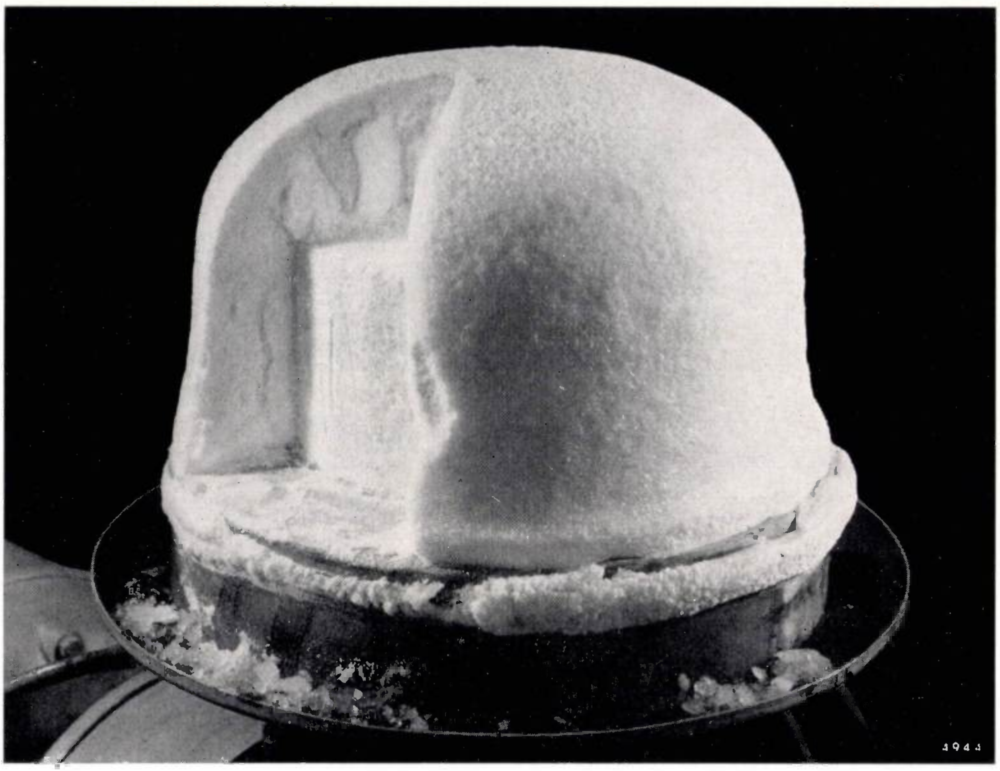


Fig. 4. A porous snow cake, about 10 cm thick, grown on the gauze in about a week.

Where the surface temperature of the snow layer is —40 °C, 99% of the moisture in the air feed will settle on the outside, and only 1% will enter the layer as vapour and form snow further inside where the temperature is lower (see fig. 2). Where the temperature inside the layer is lower than —143 °C, the carbon dioxide also forms snow. In order to ensure adequate trapping of the carbon dioxide, the temperature of the gauze should not exceed —165 °C. At this temperature 1% of the carbon dioxide in the air feed is still transmitted; this has been found to be a tolerable percentage.

With the first designs of the snow separator, the moment at which the machine had to be stopped for defrosting because of excessive resistance to air flow was determined by the amount of carbon dioxide in the layer of water snow. An important improvement was later introduced by surrounding the cylindrical snow gauze (fig. 1) with a layer of metal gauze folded in zigzag form roughly 12 mm thick (*fig.* 5). The water snow then forms on the outside of this "concertina gauze", whilst the carbon-dioxide snow has ample space to settle inside it. This substantially postpones the moment at which the resistance to flow becomes excessive.

In the foregoing we have referred to *the* temperature of the snow gauze, as if this temperature were everywhere the same. This is not so, however. Heat is supplied to the gauze over its entire surface, and must flow by conduction through the wires to the places where the gauze is brazed to cooled strips or pipes (these structural details are indicated in the figures at the end of the article). Midway between the brazed joints the gauze is therefore not so cold. Evidently, the temperature differences in the gauze must be small if the temperature at all points is to be kept below —165 °C, in other words the gauze must be a good heat conductor. Simple calculations and experiments have shown that the only suitable material for the gauze is copper wire, which must not be too fine.

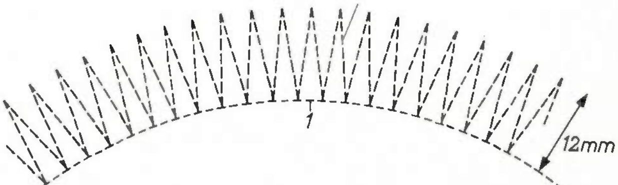


Fig. 5. A layer of "concertina gauze" *2*, about 12 mm thick, is fitted around the snow gauze *1*. This promotes the deposition of water snow on the outside of the layer, and offers more space for carbon-dioxide snow to settle.

Another initial difficulty was that if the air feed entered the snow space through a small aperture, and thus at such a high velocity that turbulences arose, the snow layer very soon became impenetrably dense. We can explain this as follows. The eddying air is strongly cooled by the surface of the snow, giving rise to hard snow crystals that show no tendency to settle to form the porous layer required. Snow of this structure, called "polar snow", evidently fills up the fine channels completely, causing a sharp increase in the air-flow resistance. If the air is allowed to enter the snow space through a large aperture, and thus at lower velocity, no difficulties are experienced. Although the large temperature differences existing might be expected to cause convection currents in the air, these currents (if they exist) are not troublesome.

The thermodynamics of the snow separator

We shall now consider the thermodynamics of the snow separator, with special reference to the magnitude of the air flow which can be handled. For this purpose we make the following simplifying assumptions.

1) The temperature of the air and that of the snow are everywhere equal, i.e. there is infinitely good thermal contact.
2) The entire latent heat of sublimation of the impurities and the heat absorbed while cooling the air feed to the temperature of the outer surface of the snow layer are released at the surface (in reality this heat is released in a layer a few millimetres thick).
3) Since, in accordance with assumption (2), the latent heat of sublimation of the small quantity of impurities deposited *in* the snow layer is disregarded, the enthalpy H of the air inside the layer must be a linear function of the temperature T. The rate of change of the enthalpy with temperature is equal to the specific heat c_p of air at constant pressure:

$$\frac{\mathrm{d}H}{\mathrm{d}T} = c_p \,. \qquad \qquad (1)$$

The variation of the enthalpy H of the air with temperature T is shown in *fig. 6*. (Since the enthalpy of a gas is determined but for an additive constant, which we can choose at our convenience, we are at liberty to assume the enthalpy of the air feed (at 10 °C, 1 atm) to be zero. This simplifies the calculations given below.)
4) The thermal conductivity λ of the snow is assumed to be identical and constant at all points (in reality λ depends on the temperature and density of the snow, and also on its structure).
5) Although the snow gauze and the outside of the cake of snow have the form of coaxial cylinders, we shall regard them as parallel flat surfaces. A cylindrical section through the snow layer at a distance x from the gauze then becomes a flat surface whose area A in fact depends on x. We will assume however that this area is constant and equal to the average of the surface area of the gauze and that of the cake of snow.

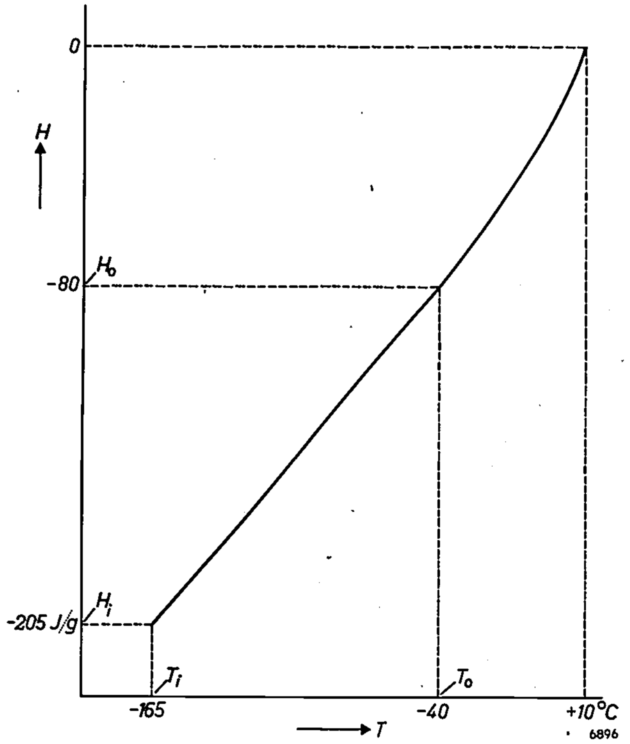


Fig. 6. Enthalpy H of air as a function of temperature T. The enthalpy of the ambient air is assumed to be zero.

In the steady state, the total heat flow through every cross-section of the layer is constant. It may easily be shown that it follows from assumption (3) that the magnitude of this heat flow is in fact zero. (This brings out the rather artificial nature of assumption (3), since one normally regards the heat content of a body as being positive, but this in no way detracts from the validity and utility of the assumption.) The total heat flow consists of an enthalpy flow in the air and a heat flow conducted by the snow. The enthalpy flow is equal to mH, where m is the mass of air displaced per second, and the heat flow through the snow is equal to $-\lambda A \mathrm{d}T/\mathrm{d}x$. We may therefore write:

$$mH - \lambda A \frac{\mathrm{d}T}{\mathrm{d}x} = 0\,.$$

Subject to the assumptions mentioned above, this equation defines the variation of the temperature of the air with the distance x from the snow gauze as shown in *fig.* 7. Making use of eq. (1), and integrating over the layer thickness x_0, we find:

$$-\frac{m\,c_p x_0}{\lambda A} = \ln\frac{H_i}{H_0}, \qquad \dots . \quad (2)$$

where the subscripts i and o relate to the inside and outside of the layer respectively.

For the reasons already discussed, we put the temperature T_i of the gauze at —165 °C and the maximum permissible temperature T_0 of the surface at —40 °C; it may be seen from fig. 6 that the corresponding enthalpy values are then $H_0 = —205$ joules/gram and H_0 = —80 joules/gram. In the design adopted the maximum thickness x_0 of the snow was 7 cm and the average cross-sectional area A was 0.2 m². Given c_p = 1 joule/gram°C and $\lambda = 0.6$ W/m°C, we find from (2) that the maximum permissible air flow is $m = -1.6$ grams per second. (The minus sign indicates that the air flow is in the negative x direction, as shown in fig 7.) This is in reasonable agreement with the value of 1.8 grams per second found in practical operation.

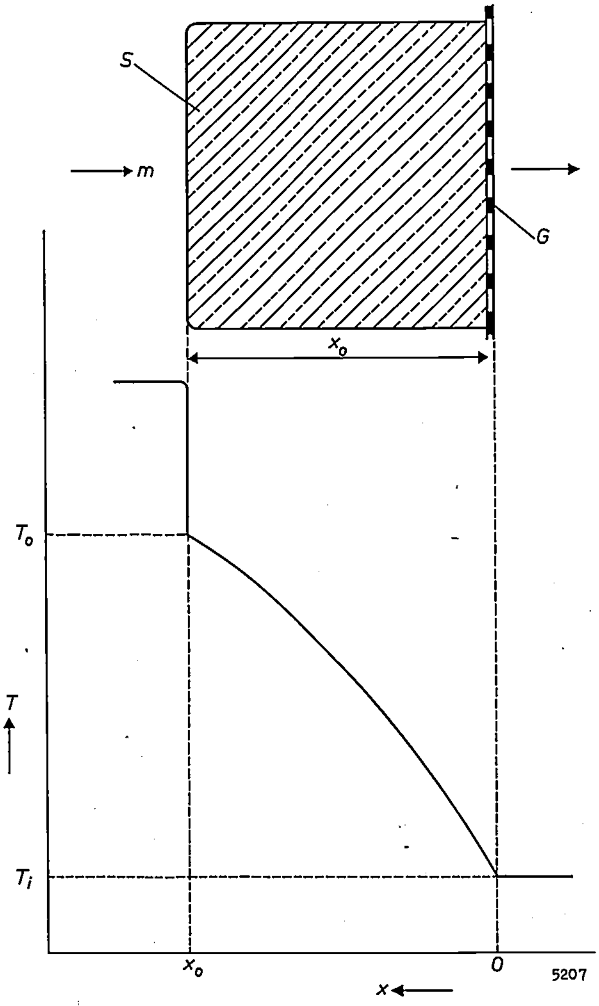


Fig. 7. Above: axial cross-section through the snow cake S, of thickness x_0. The snow gauze is denoted by G, the air flow by m. (Since the air flow is in the negative x direction, the value of m will also be negative.)
Below: variation of temperature T with distance x from the snow gauze.

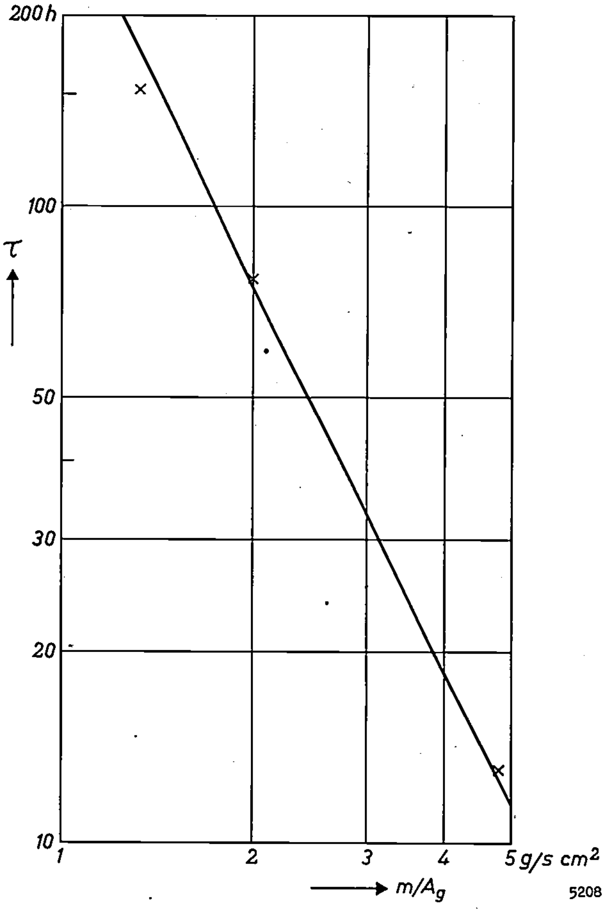


Fig. 8. "Operating time" τ as a function of air flow m per unit surface area A_g of the gauze. The straight line represents the relation $\tau \propto m^{-2}$; the crosses relate to measured values.

If the moisture content is 0.75% by weight (dew point 10 °C) and the density of thc snow is 0.4 g/cm³, it follows that the "operating time" is 4½ days (i.e. the machine can run continuously for 4½ days before defrosting). If the air flow is twice as large, so that twice the amount of impurities is supplied per second, we see from (2) that the thickness x_0 at which the temperature has risen to the point where the separator no longer works efficiently (—40 °C) is then halved, and so too therefore is the snow storage capacity. The operating time τ is consequently four times shorter. This quadratic effect ($\tau \propto m^{-2}$) is in fact found in practice (*fig. 8*).

It also appears from (2) that it must be advantageous to increase the thermal conductivity λ, for example by introducing copper pins or strips in the

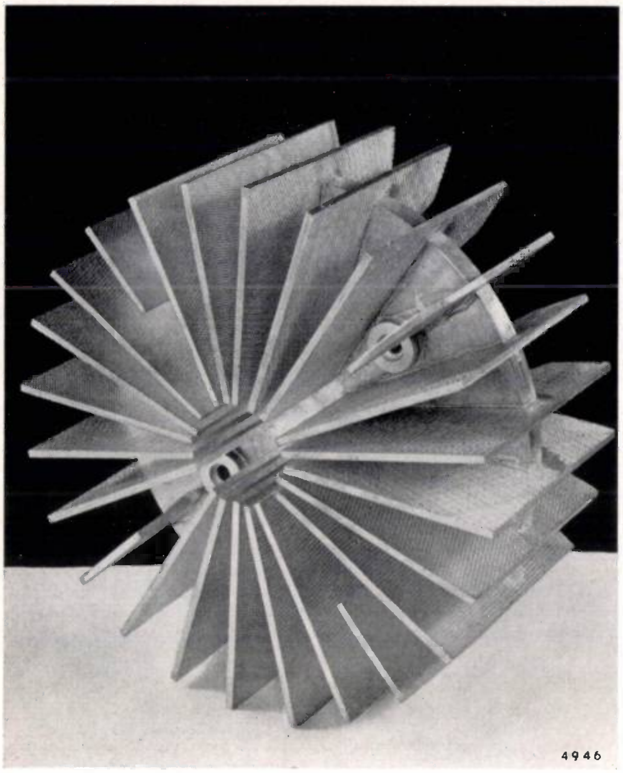


Fig. 9

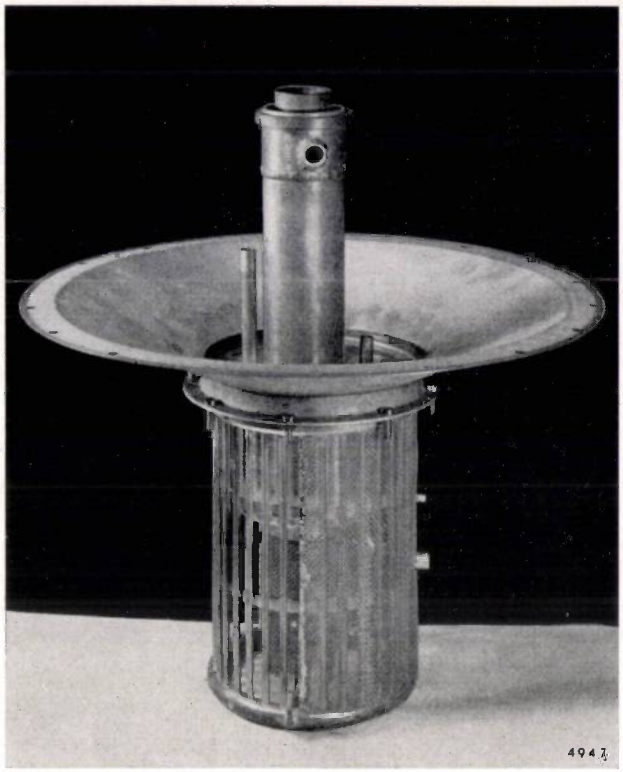


Fig. 10

Fig. 9. In the gas refrigerating machine the snow gauze is brazed to the radial copper grid illustrated here, which is bolted to the cold head of the machine and thence dissipates the heat directly.

Fig. 10. In air-fractionating installations for producing liquid nitrogen, the snow gauze (here partly removed) is brazed to a crown of copper pipes, through which liquid oxygen of about −180 °C flows.

Fig. 11. Air-fractionating column (without refrigerating machine) for supplying liquid nitrogen [3]). The bottom, dark section is an insulating jacket. in which the structure shown in fig. 10 is located.

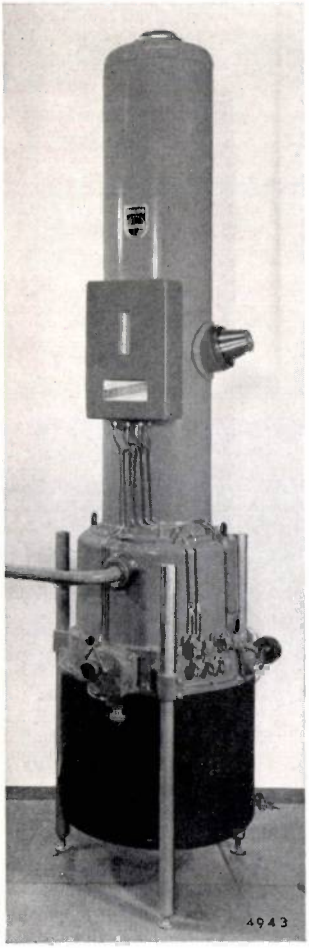


Fig. 11

snow. The concertina gauze mentioned above also works along these lines.

Finally, the snow separator as used in a liquid-air machine and in an air-fractionating column [3]) is illustrated in *fig. 9, 10* and *11*. Various structural details are mentioned in the captions.

[3]) J. van der Ster and J. W. L. Köhler, A small air fractionating column used with a gas refrigerating machine for producing liquid nitrogen, Philips tech. Rev. **20**, 177-187, 1958/59.
J. van der Ster, The production of liquid nitrogen from atmospheric air using a gas refrigerating machine, thesis, Delft 1960.

Summary. Before liquefying air, it is necessary to remove the moisture and carbon dioxide it contains. Until recently, the Philips gas refrigerating machine used a separator on which these impurities settled in the form of ice and snow. A drawback of this separator was that it had to be defrosted once a day. An entirely new design of separator is discussed here, consisting of a cylindrical jacket of copper gauze, which is kept at a temperature of −165 °C and through which the air feed is passed. Water vapour and carbon dioxide settle on the gauze as a layer of porous snow (hence the name "snow separator"). The layer remains porous for days and causes no appreciable pressure drop. Defrosting is necessary only after a continuous production of 600 litres of liquid air (five times as much as before), i.e. once or twice a week. The snow separator is also being used with advantage in air-fractionating systems.

DOSEMETERS FOR X-RADIATION

by J. HESSELINK *) and K. REINSMA *).

621.386.82:615.849.5-015.3

In recent years the conviction has gained ground that dosimetry should no longer be limited to X-ray therapy, but should be extended to the small doses administered in X-ray examinations. In the article below, after a discussion of dosemeters for use in X-ray therapy, an easily operated instrument for determining diagnostic doses is described. This instrument, unlike therapy dosemeters, measures the total energy incident on the patient. It can help radiologists in their efforts to use X-radiation sparingly.

Introduction

If proper use is to be made of X-rays in medicine, it is necessary to know with reasonable accuracy the amount of radiation administered to the patient. In therapeutic treatments this applies first and foremost to the doses received by patients individually; the optimum effective dose here is sometimes not much smaller than the maximum permissible dose, and in such cases the margin of uncertainty must be no more than 3 to 5%. In diagnostics we are mainly concerned with the amount of radiation received by the population as a whole. Although the diagnostic dose given to each patient at a time is generally small, X-ray examinations as a method of diagnosis are used on a very wide scale in the more advanced countries. No less than 75 to 90% of the total quantity of medically administered radiation reaching the sex glands of the inhabitants of these countries (the gonad dose) is due to X-ray examinations. In some of these countries this amount is roughly equal to that from natural sources (cosmic radiation, radioactivity of the soil and building materials, K^{40} and C^{14} in the human body, etc.) so that caution is obviously called for. The conviction is growing that even the small diagnostic doses received by the sex glands may, in their cumulative effect, be genetically harmful to the population. In view of the increasing use of X-radiation for diagnostic purposes it therefore seems desirable not to limit dose measurements to therapeutic treatments. The regular measurement of diagnostic doses will enable the radiologist to make more sparing use of X-rays, and may lead to the development of new equipment with which the radiologist can obtain more information with a smaller dose [1].

The physical quantity with which the biological effect of radiation can best be correlated is now generally held to be the *energy absorbed by the patient* [2]. Sincc there is no way of measuring that energy directly, an attempt has to be made to derive its magnitude from the measurement of some physical or chemical effect brought about by the radiation.

In the course of the years many and various dosimetric methods and instruments have been developed and put to use. We may mention the photographic plate, thc ionization chamber, the Geiger-Müller counter, the proportional counter and the scintillation counter; new prospects are offered by the conversion of ferrous ions into ferric ions in a suitable iron compound, the colouring of certain types of glass and the change in the resistance of cadmium sulphide upon irradiation. In this article, however, we shall be solely concerned with instruments whose operation is based on the ionization produced by X-rays in passing through a gas — in our case air — and which measure the charge carried by the ions thus formed. Such an instrument consists essentially of a gas-filled space containing two electrodes connected to a measuring device.

To measure the charge produced by ionization per unit time, a potential difference is applied between the electrodes and the current flowing in the circuit is measured. This allows the determination of the dose rate at the site of thc chamber. The total dose is found from this by integrating the current with respect to the time of exposure. The integration may be carried out by the circuit itself, simply by incorporating a capacitor in it. After the irradiation it is then simply a matter of measuring the potential to which the condenser has been charged by the ion current. The gas-filled space referred to, which functions as a radiation detector, is in both cases called an *ionization chamber.*

For measurements at many places at the same time, without the need for separate measuring

*) X-ray and Medical Apparatus Division, Eindhoven.

[1] Report of the United Nations Scientific Committee: "On the effects of atomic radiation", Ch. 3, New York 1958.

[2] Recommendations of the International Commission on Radiological Units (I.C.R.U.) 1950. See Brit. J. Radiol. 24, 54, 1951.

equipment for each chamber, use is made of *condenser chambers*. These work on the same principle as the ionization chamber, but they are not attached to the measuring equipment during irradiation. The chamber is first charged to a certain voltage. It is then removed for exposure to the radiation, and returned to the measuring equipment for reading the decrease in voltage. A condenser chamber is therefore in fact a charged capacitor with a gaseous dielectric, which is partly discharged during irradiation. Obviously, the condenser chamber only measures doses and not dose rates, although the average dose rate can simply be found by dividing the dose reading by the exposure time.

In this article we shall discuss examples of both types of instrument: ionization chambers for use in radiotherapy, condenser chambers for various purposes, and finally a special type of ionization chamber for diagnostic use.

The question now arises as to what relation exists between the energy which an irradiated patient absorbs — either totally or in a specific part of the body — and the ionization which the same radiation causes in the same period of time at the same place in free air. Unfortunately the relation is not always a simple one. We shall therefore consider briefly the physical factors that govern this relation, after first giving the definitions of two concepts of dose in current use and the units in which these doses are expressed.

Dose definitions and units; physical principles of dosimetry

The oldest dose specification still in common use, and now called the exposure dose, is based on the ionization which the radiation produces in air. The unit of exposure dose is the roentgen, defined as "an exposure dose of X- or γ-radiation such that the associated corpuscular radiation per 0.001293 grams of air produces, in air, ions carrying 1 electrostatic unit of quantity of electricity of either sign"[3]).

Two points should be noted in this connection. As mentioned, the relation between the exposure dose just defined and the energy absorbed by the patient is not always simple. The continued use of ionization as a direct measure of dose is due to the fact that the ionization can readily be measured and conveniently used for charting a radiation field (for determining isodose curves, etc.). Further, it was discovered in the early days of radiology that a close relation existed between the exposure dose and the biological effect of the rays on muscular tissue, for the types of X-rays then most commonly used.

Secondly, some remarks on the words: "the associated corpuscular radiation per 0.001293 grams of air". The ionization of the air produced directly by the X-ray quanta is negligible compared to that produced by the electrons which the quanta eject from the atoms in their path. These (secondary) electrons acquire the whole of the energy of the quanta in question (photoelectric effect), or a part of it (Compton effect), as kinetic energy, which they in turn lose by collision with other atoms or molecules, giving rise to excitation or ionization. Such electrons produce in air an average of one ion pair for every 34 eV of kinetic energy they possess, so that an X-ray quantum of energy 50 keV would produce about 1500 ion pairs, only one or two of them directly. The exposure dose as defined above is therefore not identical with the ionization produced in the volume occupied by 0.001293 g of air (which is exactly one cubic centimetre at 0 °C and 76 cm Hg), but with the ionization produced by the secondary electrons released in that volume. This makes no difference in a large space uniformly irradiated with X-rays, but it does in a small enclosed space. We shall return to this point when discussing the properties of an ionization chamber.

The other dose specification now in use is the "absorbed dose"[3]), defined as "the amount of energy imparted to matter by ionizing particles, per unit mass of irradiated material, at the place of interest". It is expressed in "rads". One rad is defined as 100 erg/g, or 10^{-2} joule/kg. The total energy absorbed by a patient during exposure to radiation is called the "integral absorbed dose", and is usually expressed in kilogram-rads (not to be confused with kilorads); one kg rad is 10^5 erg or 10^{-2} joule.

The usual practice in X-ray therapy has been to measure the exposure dose, in roentgens. The absorbed dose in the skin and in deeper areas of the body can then be found with sufficient accuracy from tables and charts. Unfortunately this is scarcely practicable in the diagnostic use of X-rays. In the first place the tube voltage and amperage, filtration, and the size and location of the field are often varied several times in the course of each examination, making it virtually impossible to measure the total dose in roentgens. Another difficulty is the fact that the radiation used for diagnosis is not hard, as it generally is for X-ray therapy, but soft.

To explain this difficulty, and by way of introduction to the discussion of our dosemeter for

[3]) Report of the I.C.R.U. 1956, Handbook 62, Nat. Bur. Standards, Washington.

diagnostic use, we shall touch briefly on a few points concerning absorption. The absorption of X-rays is a function of their quantum energy. At a given quantum energy, however, a given substance always has the same absorption per gram/cm^2; water in the liquid state, for example, absorbs just as much per gram/cm^2 as in the vapour state. (We are concerned with absorption in a thin laycr whosc thickness is expressed in mass per unit area.) The absorption can differ considerably from one substance to another, depending on the atomic number — the higher the atomic number the greater the absorption of radiation of given quantum energy — and on the relative quantities of the elements contained in the substance. The chemical binding of the elements plays no significant part. Air, water and wet tissue, for example, show roughly the same absorption per gram/cm^2. The absorption in bony tissue, which contains the heavier elements calcium ($Z = 20$) and phosphorus ($Z = 15$), is much greater. The absorption per gram/cm^2 is expressed by the *mass energy-absorption coefficient*, which is equal to the linear energy-absorption coefficient μ divided by the density ϱ. (Since there is no possibility of confuzion, we write simply μ in place of the approved symbol μ_{en}.)

The dependence of the mass absorption coefficient on the quantum energy of incident monochromatic X-radiation is shown in *fig. 1*. It can be seen that the curve is almost horizontal for hard radiation (high-energy radiation), but the variation of μ/ϱ with quantum energy is very marked for soft radiation. A similar curve is found for all substances.

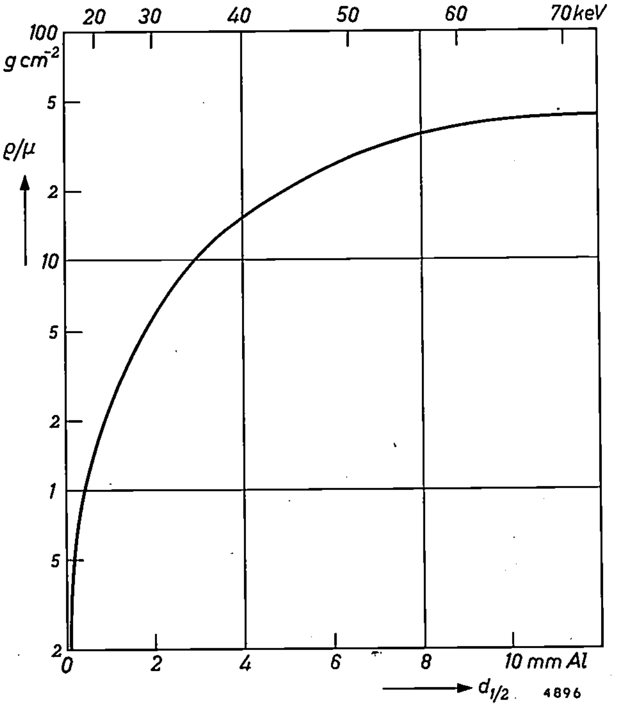


Fig. 1. The reciprocal of the mass-energy absorption coefficient μ/ϱ of air for monochromatic X-rays as a function of radiation-quality. The latter is expressed as the half-value layer $d_{\frac{1}{2}}$ and also as the quantum energy (in keV). For hard radiation (as used in therapy) μ/ϱ is practically constant; for soft radiation (diagnostic) it varies considerably.

As the radiation from an X-ray tube is not monochromatic, it cannot be characterized by a single quantum energy. In practice the spectral intensity distribution — called the *quality* of the radiation — is specified by the thickness which a filter of aluminium or copper must have in order to reduce the dose rate (roentgens per unit time) to one half, one quarter, one eighth, etc., of the initial value. These thicknesses are termed the first, second, third, etc., *half-value layers* (HVL). In medical practice it is usual to determine only the first half-value layer (HVL$_1$).

The intensity of monochromatic (monoenergetic) radiation decreases exponentially with the mean free path in the absorber, and the quality of the radiation is therefore satisfactorily defined by HVL$_1$. The filter that reduces the intensity from $\frac{1}{2}$ to $\frac{1}{4}$ is then just as thick as the first, and so on. Where non-monochromatic radiation is concerned, the layers are successively thicker, the soft rays being most strongly absorbed and the remainder becoming progressively harder. The difference in thickness is greater the broader the radiation spectrum. The quotient HVL$_1$/HVL$_2$ is termed the homogeneity factor, and is thus unity for monochromatic radiation.

In fig. 1 the quality of the radiation is given along the abscissa in terms of both quantum energy and half-value layer. Although strictly applicable only to monochromatic radiation, each point of the curve is roughly valid for non-monochromatic radiation of the same HVL.

To indicate the specific difficulties encountered when integral doses of soft radiation are to be measured, we shall now consider the way in which the energy flux density, i.e. the energy per unit time passing through unit area of surface normal to the X-ray beam, is related to the ionization produced in air per unit time in that area. We note that the ionization is proportional to the *absorbed* energy: the above-mentioned ionizing energy of 34 eV per ion pair holds for all radiation qualities concerned. Since the energy absorbed per unit time in a small volume is equal to the product of the energy flux density and the absorption coefficient μ, we may infer that the relation between the dose rate in roentgens per unit time and the energy flux density contains the factor μ. This factor will of course vary with the quality of the radiation in the same way as μ/ϱ. We have seen from fig. 1 that μ/ϱ is practically independent of the radiation quality where the rays are hard (for therapy) but by no means so where the rays are soft (for diagnostics). It is therefore understandable that the absorbed energy in an X-ray examination cannot easily be derived from the dose expressed in roentgens. For that to be possible the spectral intensity distribution of the X-rays would have to be known exactly.

With the diagnostic dosemeter to be described in this article the total energy incident on the patient can be accurately measured without its being necessary to know precisely the quality of the radiation used. Since the precision required in a measurement of the integral absorbed dose in diagnostics is not high — an uncertainty of 25% is permissible [1]) — this dose can be determined near enough by assuming that 20 or 30% of the incident energy is transmitted or scattered and the remainder absorbed [4]).

Although the gonad dose is obviously not determined in this way, it can be estimated if we know the total energy absorbed by the body, and the part of the body exposed to the rays (chest, stomach, etc.). Measurement of the integral absorbed dose is therefore a useful means of arriving at the gonad dose, which is itself very difficult to measure [5]).

The method by which the new dosemeter for diagnostics is made independent of the radiation quality will be discussed at the end of this article. First we shall consider a number of instruments for measuring therapeutic doses [6]).

Dosemeters for X-ray therapy

Ionization chambers

An ionization chamber for X-ray dosimetry is sketched in *fig. 2*. The cylinder wall *1* and the central pin *2*, both of which are conductive, form the electrodes mentioned in the introduction. Between them, fitted in an insulating bush *3*, is a third electrode *4*, in the form of a ring. This is given roughly the same potential as the central electrode and ensures that any leakage current across the insulation is not included in the measurement. In a calibration procedure, the chamber is given a carefully defined volume such that its sensitivity (in coulombs/roentgen) is conveniently adapted to that of the measuring equipment. In other words, full deflection of the meter is arranged to correspond with a round number of roentgens, say 3 or 10.

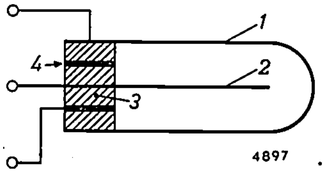


Fig. 2. Ionization chamber for dosimetry in therapeutic radiology (schematic). *1* wall. *2* central electrode. *3* insulation. *4* guard ring.

The current *I* that flows in an ionization chamber upon exposure to X-rays and the charge *Q* of the ions produced per unit time are proportional to one another below a certain critical value of *Q*. Above that value, *I* increases less than one would expect. The smaller the potential difference *V* between the electrodes *1* and *2*, the lower are the *Q* values at which this effect occurs (*fig. 3*). The reason for the effect is that the ions have a certain chance of recombining, thereby losing their charge. *Fig. 4* will help to make this clear.

This figure shows the relation between *I* and *V* for three different values of *Q*. At small values of *V* the current is seen to rise fairly steeply with the voltage (region *A*), after which it flattens out to an almost constant level (region *B*), and finally curves upwards again (region *C*). The explanation is that the ions travel relatively slowly at lower voltages (region *A*), so that many of them are able to recombine before reaching one of the electrodes. When the voltage is raised, the ions travel faster and thus have less opportunity to recombine; the current then rises until it becomes virtually independent of the applied voltage (region *B*). However, if *V* is raised to a value where the electrons in the electrical field inside the chamber absorb sufficient energy to produce ionization themselves, the newly formed ions add to the current, which thereupon rises again with *V* (region *C*). (This is the gas-amplification effect which underlies the operation of the proportional counter.)

As can be seen, the boundary between region *B* and region *C* lies at the same value of *V* at all three dose rates. This is not so as regards the boundary between *A* and *B*. With increasing *V*, the recombination effect persists at higher *Q* values, i.e. higher dose rates; there is then a greater concentration of ions and electrons and thus the ions individually have a greater chance of recombining.

It is now clear why the *I*-*Q* curves in fig. 3 begin to flatten out at high *Q* values. If *Q* is increased whilst *V* is kept constant, at a certain moment the boundary between the regions *B* and *A* (fig. 4) will

[4]) Cf. E. Zieler, Fortschr. Röntgenstr. **92**, 211, 1960. The fraction absorbed in various methods of chest examination is already known. See E. Zieler, *ibid.* **94**, 248, 1961.

[5]) See chapter IV (by J. Feddema and W. J. Oosterkamp) of the book "Modern trends in diagnostic radiology". (2nd series), Butterworth, London 1953, and also A. Neboschew and O. Schaft, Zur Überwachung der Patientenbelastung während der Röntgendurchleuchtung, Röntgen-Bl. **12**, 244, 1959.

[6]) For a thorough treatment of dosimetric methods in X-ray therapy and their physical background, see G. J. Hine and G. L. Brownell, Radiation dosimetry, Acad. Press, New York 1956.

be passed. This will happen less quickly the higher V is chosen (as long as V remains within region B, of course).

Besides the potential applied between the electrodes, another factor of importance in an ionization chamber of the type sketched in fig. 2 is the wall.

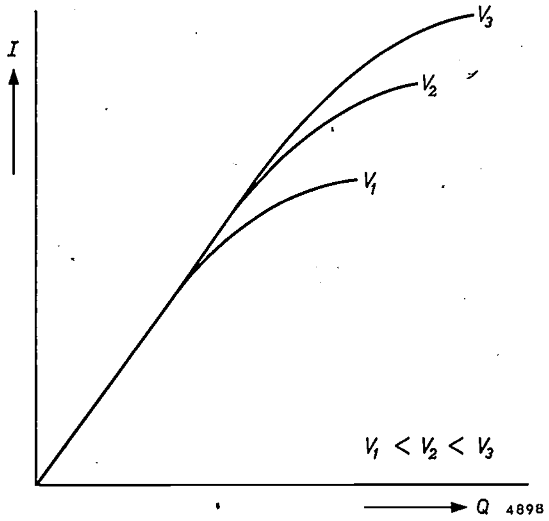


Fig. 3. Relation between the current I flowing in an irradiated ionization chamber and the charge Q produced by ionization per unit time, for various values of the applied voltage V.

It is for this reason that we have compared the current I with the charge Q produced per unit time in the chamber and not with the dose rate q. To explain this we recall the remark we made on the fact that the definition of the roentgen is based on the ionization caused by the secondary electrons generated in 0.001293 gram of air. Since these electrons have high kinetic energies — as we have already mentioned, the fastest acquire almost the total energy of the X-ray quantum — some of the secondary electrons liberated in the wall also enter the ion-collecting volume of the chamber and there contribute to the ionization of the air. On the other hand, some of the electrons formed in the air travel to the wall and contribute less to the ionization than would otherwise be the case. On the whole, these two effects are not entirely compensatory. In itself this would present no difficulties — the chamber must in any case be calibrated — if it were not for the fact that the resultant effect depends on the quality of the radiation. At constant V and constant true dose rate the current I of an ionization chamber is therefore dependent as a general rule on the quality of the radiation.

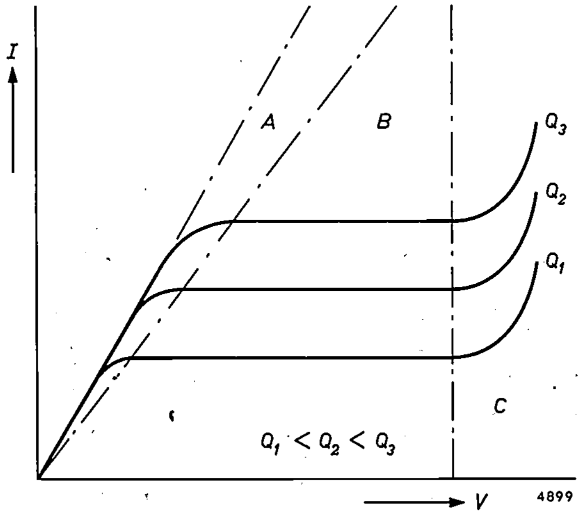


Fig. 4. Relation between the current I and the applied voltage V for various values of the charge Q formed per unit time in the chamber. The operating point should be in region B.

The wall also has another influence on the ion current: it absorbs part of the radiation, so that the intensity of the radiation inside the chamber is lower than the intensity at that spot in the absence of the chamber. This effect is again dependent on the quality of the radiation. By suitable choice of the wall material, the two wall effects can be made to compensate one another substantially in a particular range of radiation qualities [7]). In this way it is possible to build a chamber which has a reasonably constant sensitivity in a limited range of radiation qualities. This is sufficient for practical purposes, although of course more than one ionization chamber is needed to cover the whole range of therapeutically employed radiations. *Fig. 5* shows three of the ionization chambers of the Philips Universal Dosemeter, each of which is suitable for a specific range of radiation qualities. Particulars of their construction are given in *fig. 6*. Finally, *fig. 7* shows an ionization chamber fitted to the end of a 70-cm-long rubber tube for introduction into a body cavity, e.g. the oesophagus.

The increased ionization due to the secondary electrons liberated in the wall, and the decreased ionization due to secondary electrons penetrating the wall, compensate one another if the wall material apd the gas filling have the same chemical composition. The difference in density is immaterial. If the wall and gas are not identical in chemical composition, then they must have the same mass absorption coefficient for X-radiation and the same atomic stopping power for electrons. In that case the wall is said to be equivalent to the gas filling. A wall that is equivalent to air for all qualities of radiation does not exist. As we have seen, the wall of the ionization chambers used in the Philips Universal Dosemeter is not meant to be entirely air-equivalent, the object being to compensate as far as possible for the absorption of X-rays in the wall.

[7]) See e.g. W. J. Oosterkamp and J. Proper, Acta Radiol. 37, 33, 1952.

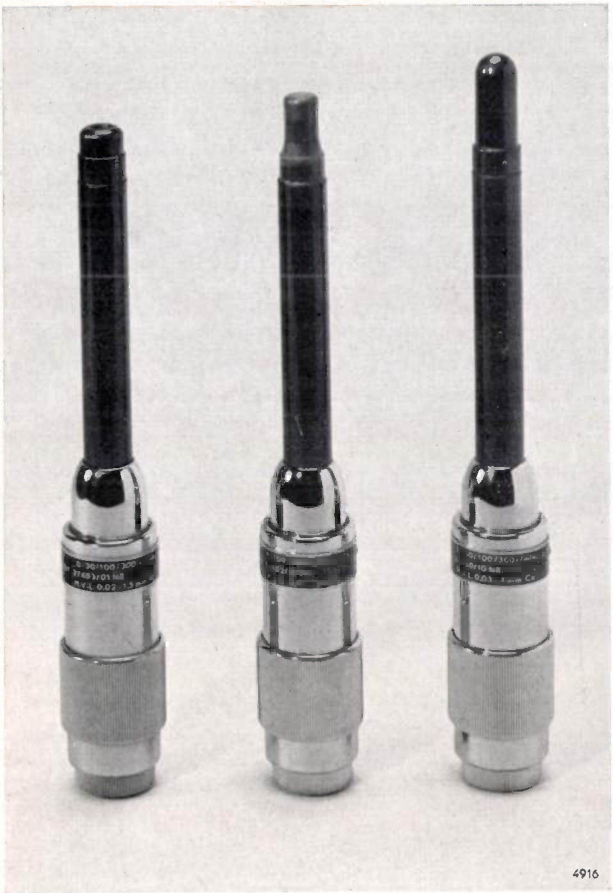


Fig. 5. Three of the ionization chambers of the Philips Universal Dosemeter. The one on the left is for dose rates up to 300 roentgens/second of radiation having a half-value layer (HVL) of 0.02 to 1.5 mm Al. The corresponding figures for the middle one are 1000 roentgen/min and 0.4 to 7 mm Al, and for the right-hand one 300 roentgen/min and 0.03 to 4 mm Cu. The ion-collecting space is at the top (cf. fig. 6).

The curve in *fig. 8* gives an idea of the extent to which the ionization current, at a given dose rate, is independent of the radiation quality. This is the correction curve applied to the reading given by the Philips Universal Dosemeter when fitted with the ionization chamber shown on the right in fig. 5 — a chamber used in deep therapy. The response for qualities between 0.2 and 1.5 mm Cu (i.e. roughly 5 and 20 mm Al, respectively; cf. fig. 1) is seen to vary by less than 1%. Even at 0.1 and 2.5 mm Cu, the variation is only 3%.

The ionization chambers which form a part of the Philips Universal Dosemeter are all calibrated before they leave the factory. This is done with an extremely stable (substandard) ionization chamber, the calibration curve of which is obtained by comparison with the standard ionization chambers [8]) at the National Physical Laboratory, Teddington (England), the Physikalisch-technische Bundesanstalt, Brunswick (Germany) and the Rotterdams Radiotherapeutisch Instituut (the Netherlands) [9]).

The calibration curve is constructed so as to ensure that the sensitivity of the Philips standard is equal to the average of the sensitivities of the three standard chambers mentioned and that of the National Bureau of Standards at Washington (U.S.A.). (The sensitivity of the latter standard chamber as compared with that of the National Physical Laboratory is given in the literature.) The maximum differences in sensitivity between the Philips substandard and these chambers (which occur at different radiation qualities in each case) are:

Nat. Phys. Lab.	0.9% higher
Nat. Bur. Stand.	0.5% higher
Rott. Radiother. Inst.	0.3% higher
Phys.-Techn. Bundesanst.	1.4% lower.

Only those ionization chambers are passed after calibration whose sensitivity for any given radiation quality differs by no more than 3% from the value found from the curve provided with the instrument (cf. fig. 8).

Condenser chambers

As we have seen, a condenser chamber is essentially a charged capacitor with a gaseous dielectric, which is gradually discharged upon exposure to

[8]) In a standard ionization chamber the wall plays no part whatsoever. The principle is described e.g. in the book by Hine and Brownell, quoted in note [6]).

[9]) Carried out in 1957 by W. J. Oosterkamp and J. Proper of Philips research laboratories at Eindhoven.

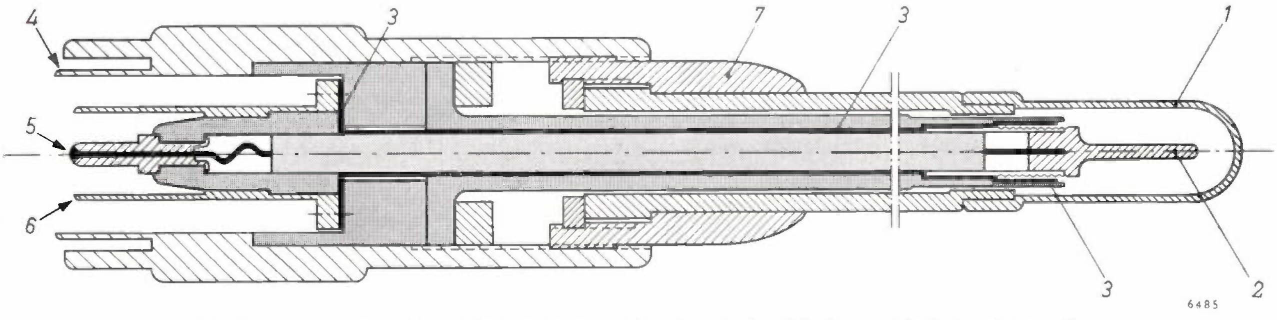


Fig. 6. Cross-section of one of the ionization chambers in fig. 5 (schematic). *1* chamber wall. *2* central electrode mounted on an insulating bushing with a ribbed surface to lengthen the creep path. *3* (thick black line) guard ring. *4*, *5* and *6* cable terminals, connected to electrodes *1*, *2* and *3* respectively. Except for the above-mentioned bushing, all insulating parts are shaded. The volume of the chamber can be adjusted to give the desired sensitivity by screwing the part *7*, fixed to the chamber wall *1*, in or out.

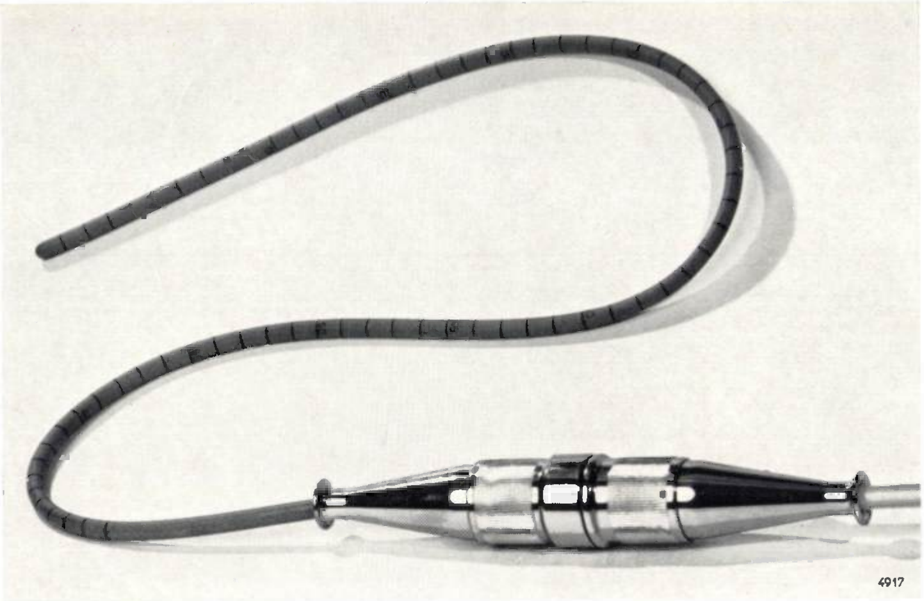


Fig. 7. For dose measurements in a body cavity, e.g. the oesophagus, a small ionization chamber is fitted to the end of a rubber tube 70 cm long.

radiation. Since a condenser chamber is not connected to the measuring circuit during exposure, it possesses no guard ring, and therefore special care must be paid to the insulation. It is necessary, for example, to ensure that the surface of the insulator is always kept dry.

From what has been said above about the effect of ion recombination on the measurement, it may be inferred that all values which the potential difference V of the electrodes acquires during a measurement must fall in the region within which the effect of recombination is not troublesome — region *B* in fig. 4. A condenser chamber must therefore not be discharged too far.

Like the chamber shown in fig. 7, condenser chambers can be used during therapeutic irradiations for dose measurements in body cavities. Very small ones are used for this purpose. Furthermore, as mentioned, they make it possible to measure doses at numerous places at the same time without having to use an equal number of measuring circuits.

Condenser chambers are also used as radiation monitors for personnel protection. They may then be carried around in the pocket, and serve as a means of ascertaining whether the measures of radiation protection adopted are in fact effective. Such chambers are usually given the convenient form of a fountain pen. Large condenser chambers, which are of course much more sensitive, are used for spot checks of the radiation level in places of work and also for detecting and measuring leakage radiation. Three types of condenser chamber are shown in

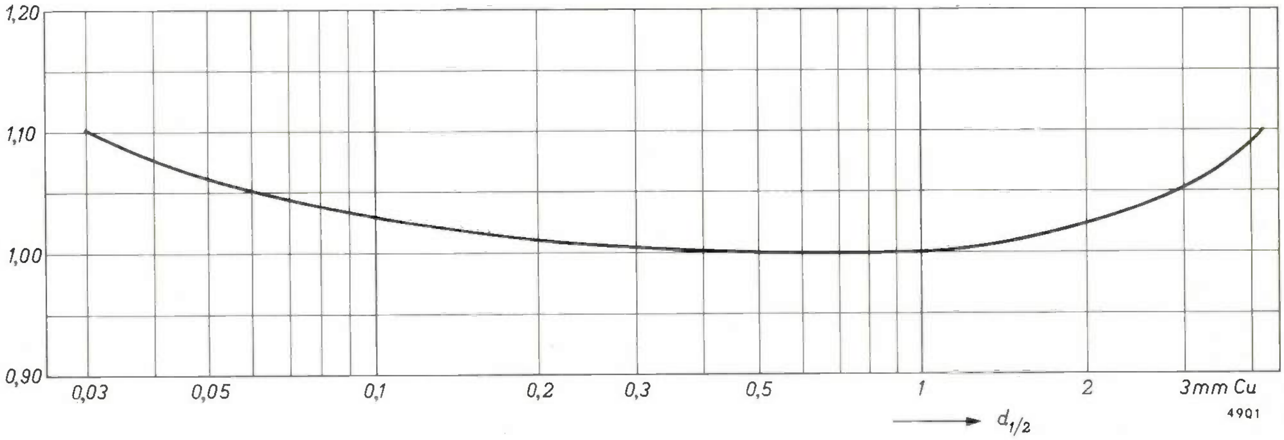


Fig. 8. The correction factor to be applied to the reading on the Philips Universal Dosemeter to find the actual exposure dose differs only very little from 1 in a fairly wide range of radiation qualities. The curve drawn here applies to the ionization chamber on the right in fig. 5.

fig. 9. The one on the right (see also *fig. 10*) has a very thin window (1 mg/cm^2) which makes it suitable for the measurement of very soft X-rays and β-rays.

Measuring circuit of the Philips Universal Dosemeter

The function of the measuring circuit associated with an ionization chamber is to give the electrodes — including the guard ring — the correct potential, and to give a direct reading of the ion current or of the charge. As an example, we shall discuss the circuit used in the Philips Universal Dosemeter. A diagram of the circuit, partly in block form, is shown in *fig. 11*. To measure a dose *rate*, switch S is turned to position *1*. The ion current then flows through resistance R, and the potential difference produced between the ends p and q of this resistance is measured with the circuit shown to the right of the broken line. Block I represents an electronic voltmeter. Since the current which this draws from the input circuit must be small compared with the ion current (which is itself of the order of 10^{-10} A) I must have a very high input resistance. This is achieved by making use of a vibrating-reed type of electrometer [10]).

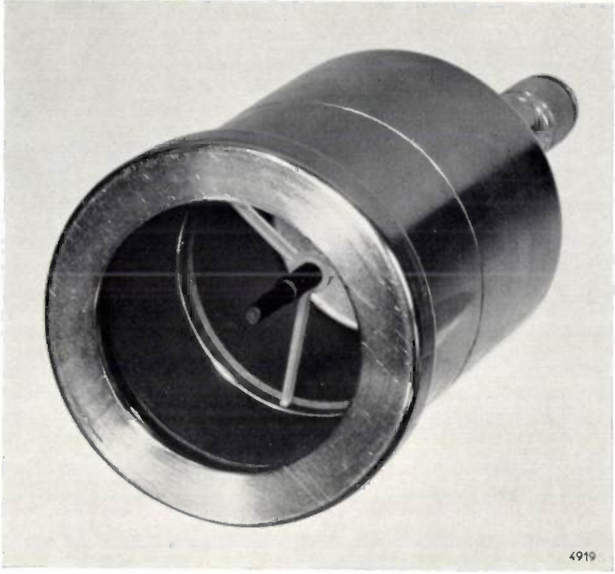


Fig. 10. Condenser chamber for very soft radiation with protective cap removed, showing the central electrode.

The output signal from I — an alternating voltage whose amplitude is proportional to the input voltage and whose frequency is equal to that of the vibrating-reed capacitor — is amplified in II, rectified in III, and applied to the moving-coil instrument M.

The sensitivity of the entire apparatus is stabilized by means of negative feedback, a variable portion of the output signal from III being fed back to I in antiphase by circuit IV. The sensitivity can also be varied in this way, which is desirable for two reasons. In the first place, it is then easy to switch over to another measuring range, and secondly there is no need to correct the reading of the meter if the temperature and pressure of the air differ from the values for which the chamber is calibrated. The correction required is made in advance by correspondingly altering the sensitivity of the measuring circuit.

The rectifying circuit III is both phase-sensitive and selective: it does not pass signals whose frequency differs from that of the vibrating capacitor, except for the

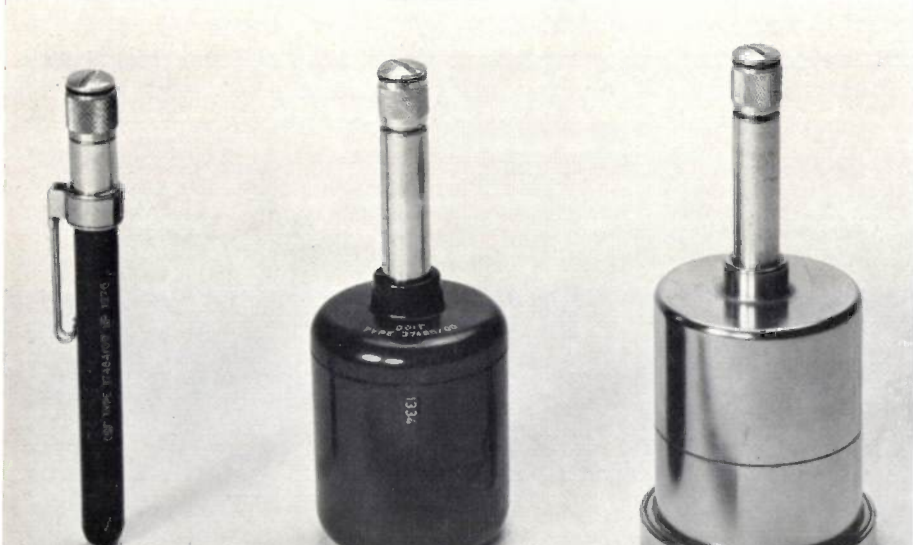


Fig. 9. Three condenser chambers used to monitor radiation for personnel protection. The smallest can be carried in the pocket like a fountain pen; the larger one in the middle, which is highly sensitive, can serve for measuring the radiation level in work places. The one on the right is for the detection of very soft X-rays and β-rays (cf. fig. 10).

[10]) See J. van Hengel and W. J. Oosterkamp, A direct-reading dynamic electrometer, Philips tech. Rev. **10**, 338-346, 1948/49. The circuit of the Philips Universal Dosemeter, largely developed by J. Fransen, differs in some points from the circuit described there, but is the same in principle.

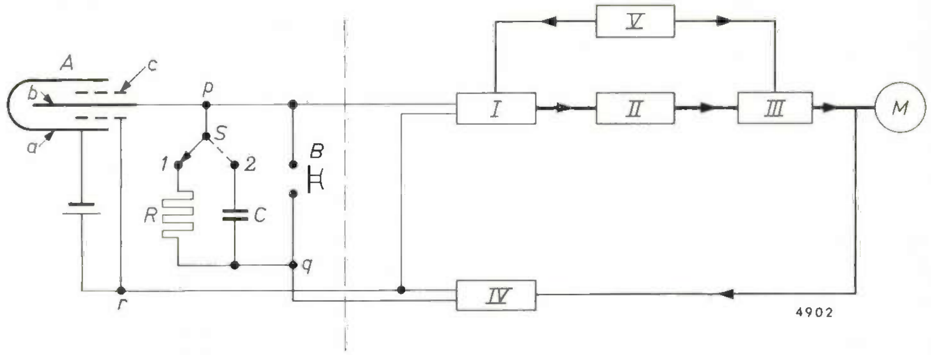


Fig. 11. Measuring circuit, partly in block form, of the Philips Universal Dosemeter. When the ionization chamber *A* (*a* is the wall, *b* is the central electrode and *c* the guard ring) is exposed to radiation, the ion current flows through the resistance *R* if switch *S* is turned to position *1*. The potential difference then appearing between points *p* and *q* is proportional to the ion current, and a measure of the dose rate. The potential is measured by a vibrating-reed electrometer *I*, amplified in *II* and rectified in *III*. The latter is a phase-sensitive rectifier which only passes signals having the frequency of the vibrating capacitor (and its odd harmonics). *M* meter. *IV* feeds a variable proportion of the output signal back to *I* in antiphase. This negative feedback stabilizes the amplification against mains fluctuations and makes it a simple matter to vary the sensitivity. *V* oscillator which drives the vibrating capacitor in *I* and supplies an auxiliary signal of the same frequency to *III*. When switch *S* is in position *2*, the ionization current is integrated and the dose in roentgens is measured. Pushbutton *B* serves to discharge the capacitor *C* after each measurement.

If it is desired to connect a condenser chamber between *p* and *r*, switch *S* must be set in a third position similar to position *2* but in which *C* has a much lower value.

odd harmonics. To achieve this, the output signal from the oscillator *V*, which drives the vibrating capacitor in *I*, is also fed to *III*. The object of making the circuit phase-sensitive is to prevent the negative feedback changing into positive feedback if the strength of the input signal should drop sharply, in which case the signal supplied by *IV*, which follows the change relatively slowly, may have a greater amplitude than the input signal itself. Owing to the fact that the circuit, for all practical purposes, only passes signals of the measuring frequency, interfering signals cannot affect the meter reading.

A detailed description of the operation and circuitry of blocks *I*, *III* and *IV* will be found in the article quoted under [10]).

If the *dose* and not the *dose rate* is to be measured, switch *S* is turned to position *2*. After the exposure, the voltage to which the ion current has charged the capacitor *C* is then measured and read off in roentgens.

For checking the operation of the circuit, a calibration device is incorporated in the Philips Universal Dosemeter. This is an ionization chamber, sealed off from the outside air and filled with nitrogen, which is connected to the electrometer circuit in the same way as the chamber used in the measurement. A small quantity of radium is applied to the central electrode; the radiation from this produces ionization in the nitrogen and causes a current of about 4×10^{-10} A to flow in the resistance *R* (fig. 11).

The sensitivity can be corrected for differences between the actual temperature and barometric pressure and those at which the chamber was calibrated by connecting to the circuit an ionization chamber which is in open communication with the atmosphere and in which the central electrode is fitted with a hollow glass bead containing radium (*fig. 12*). The α radiation emitted by the radium is absorbed in the bead wall; the β- and γ-radiation produce the ionization.

The correction can also be derived from a table if the pressure and temperature are accurately known.

A photograph of the Philips Universal Dosemeter is shown in *fig. 13*.

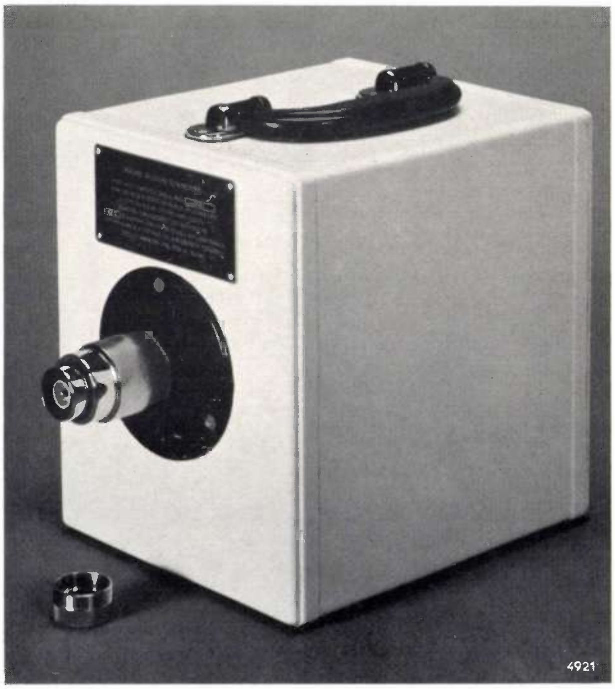


Fig. 12. Ionization chamber in open communication with the atmosphere and containing an internal radiation source for correcting the sensitivity of the measuring circuit (figs. 11 and 13) to allow for different values of temperature and air pressure.

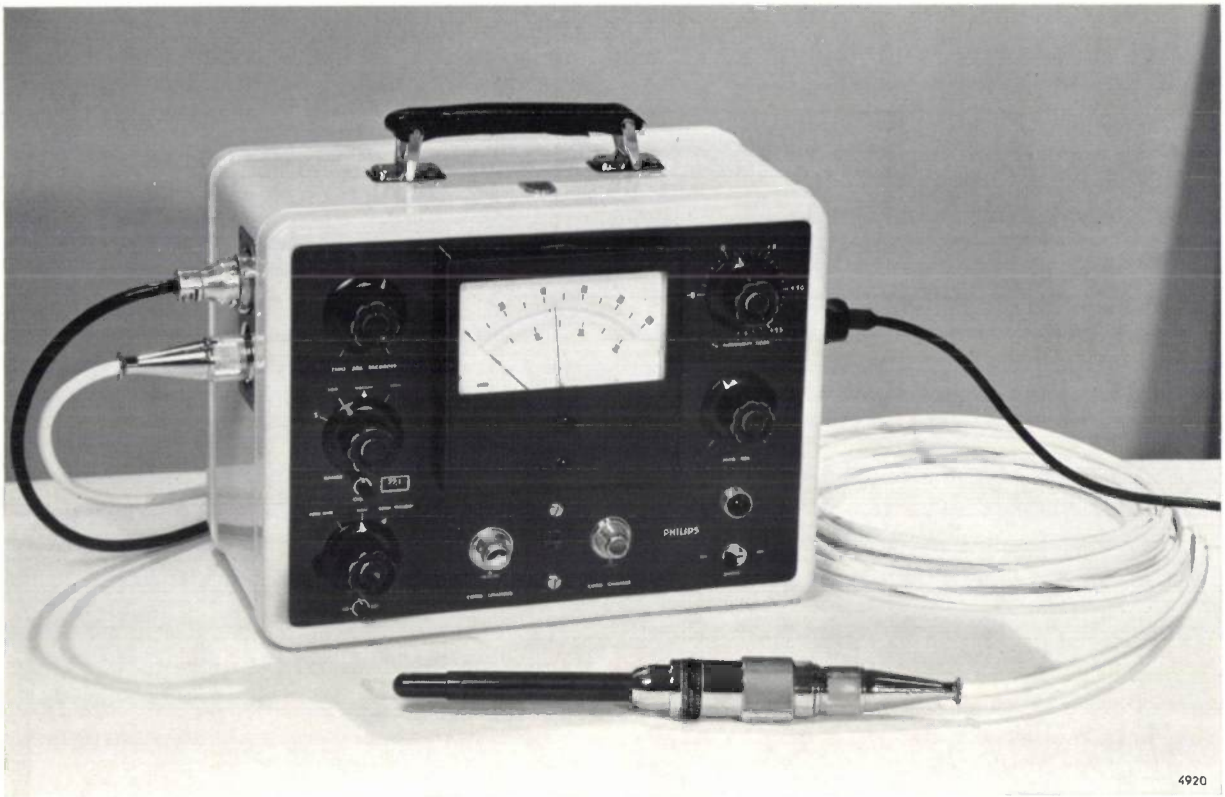


Fig. 13. The power supply and measuring unit of the Philips Universal Dosemeter, with one of the ionization chambers connected to it. On the right of the front panel can be seen, from the bottom upwards: mains switch, signal lamp, control for zero adjustment, control for sensitivity correction. On the left: push-button *B* and switch *S* (fig. 11), push-button for switching on the calibration device, range selector switch, and control for zero adjustment of any recording apparatus that may be connected. A condenser chamber can be connected to the terminals underneath the meter, after removing the protective cap. The terminal on the right is for charging the chamber, the one on the left for measuring the residual charge. The black cable on the right is the mains lead. The plug top-left carries the lead to a recording instrument.

A dosemeter for diagnostic use

We now come to the dosemeter for diagnostic use. As mentioned, this instrument measures the total radiant energy incident on the patient. For this purpose the detector is given the form of a large flat box, which is placed in front of the patient. The box is large enough to transmit the entire X-ray beam. The instrument is made independent of the quality of the radiation by using two ionization chambers separated by a filter. Provided certain conditions are fulfilled, the difference between the ion currents of the two chambers is proportional, over a fairly wide range of radiation qualities, to the energy flux incident on the patient [11]). The conditions to be fulfilled will be examined with reference to *fig. 14*.

In this figure, electrodes *1* and *2* together form the first ionization chamber (*P*), and the second (*S*) is formed by electrodes *3* and *4*. Both chambers are fitted with the usual guard ring (*5* and *6*). The respective distances between the electrodes are d_P and d_S. The X-ray beam (shaded) passes from the focus *A* successively through chamber *P*, filter *B* and chamber *S*. Electrodes *2* and *3* in the two chambers are interconnected, and the voltages are applied with opposite polarities. The output terminals *p* and *r* are connected to a measuring circuit using negative feedback, in the same way as the single ionization chamber in fig. 11 is connected to the circuit in the Philips Universal Dosemeter. (The corresponding points in this figure are also denoted by *p* and *r*.)

When a suitable voltage is applied between electrodes *1* and *2* (fig. 4), the charge produced per unit time in chamber *P* is equal to the current I_P and depends on the energy flux $\dot{E}_P$ at the place of interest according to the equation:

$$I_P = e\times\mu\times d_P\times \dot{E}_P/W = C\,\mu\,d_P\,\dot{E}_P \; . \quad (1)$$

[11]) This idea was the fruit of a discussion between one of the authors and K. Bronsema, X-ray and Medical Apparatus Division, Eindhoven.

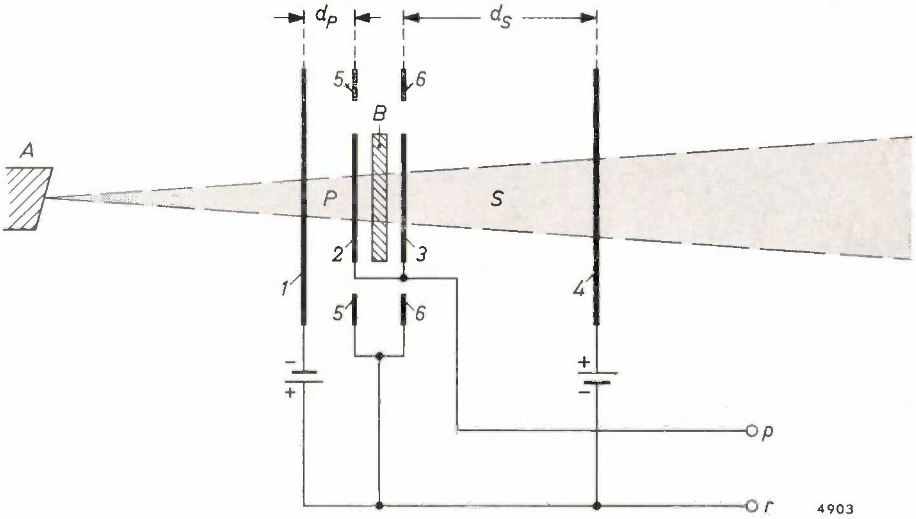


Fig. 14. Twin ionization chamber for measuring integral doses in X-ray examinations. The X-ray beam (shaded) passes from the anode *A* through the ionization chamber *P* (electrodes *1* and *2*), the filter *B* and the ionization chamber *S* (electrodes *3* and *4*). The chambers have the form of a flat box and are large enough to transmit the whole of the X-ray beam incident on the patient. Electrodes *5* and *6* are the guard rings.

Here *e* is the elementary charge, *W* the energy required to produce one ion pair — 34 eV in air — and *C* is the ratio e/W. Similarly, for chamber *S* we have:

$$I_S = C\,\mu\,d_S\,\dot{E}_S. \qquad (2)$$

The difference between the two currents is:

$$I = I_S - I_P = C\,\mu\,(\dot{E}_S\,d_S - \dot{E}_P\,d_P)\,. \qquad (3)$$

Disregarding the absorption in the walls, $\dot{E}_S/\dot{E}_P$ is equal to the transmission factor a of filter *B*. Equation (3) may therefore be written

$$I = C\,\mu d_S \dot{E}_S\left(1 - \frac{1}{a\beta}\right), \qquad (4)$$

where $d_S/d_P = \beta$. In these equations, μ and a both depend on the quality of the radiation; a is greater and μ smaller the harder are the rays.

If the factor $\mu(1 - 1/a\beta)$ is independent of the radiation quality, so too is the relation between I and $\dot{E}_S$. Since we are perfectly free to choose the value of β, and can also control the form of the variation of a within certain limits by a suitable choice of the thickness and material of the filter, this requirement can be exactly fulfilled for two qualities of radiation, and met to a fair approximation for the radiation in the neighbouring range.

Fig. 15 illustrates the extent to which the response of the apparatus so designed is independent of the radiation quality. The ratio between the response and the energy flux leaving the chamber is plotted as a function of the anode voltage of the X-ray tube [12]). At voltages of 55 and 130 kV (HVL respectively 2.3 and 6.3 mm Al) the reading is seen to be exact, and between these values it deviates by no more than about 15%. The values of certain quantities governing the performance are given in the caption to fig. 15.

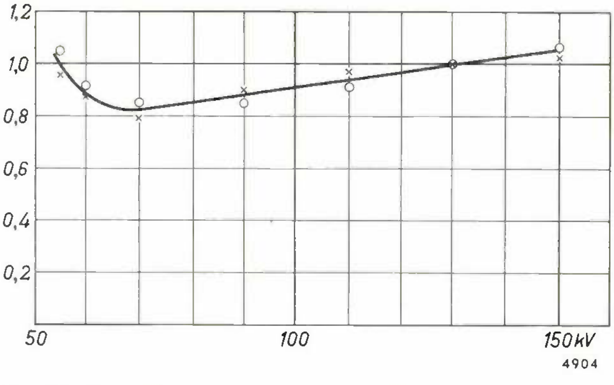


Fig. 15. Correction factor to be applied to the reading of a diagnostic dosemeter using a twin chamber. In a wide range of radiation qualities this factor differs only slightly from unity. (The half-value layers corresponding to anode voltages 50, 130 and 150 kV are 2.3, 5.6 and 6.3 mm Al, respectively.) The points × were obtained with a scintillation detector, the points ○ were calculated from the dose in roentgens measured with an ionization chamber and from the relevant value of μ. (The experimental instrument was fitted with an aluminium filter $1\frac{1}{2}$ mm thick. At 50 kV its transmission factor a is 0.50, at 130 kV it is 0.74. Distances between electrodes: d_P = 2.0 mm and d_S = 4.5 mm).

[12]) See K. Reinsma, Dosismeters voor de röntgendiagnostiek, Centrex Publishing Co. Eindhoven 1960, where further particulars will be found. (To be published shortly in English.)

The experimental points in fig. 15 were obtained by two entirely different methods. The first method [13]) used a scintillation counter and count-rate meter. Before the actual measurements were made, the response of the measuring equipment was determined by exposing the detector to monochromatic X-radiation — obtained by diffraction in a crystal platelet — and counting the individual quanta. The energy flux corresponding to a given reading on the meter could then be directly computed from the number of quanta counted and the known quantum energy. This was done with tungsten and molybdenum $K\alpha_1$ radiation.

The sensitivity of the twin chamber to monochromatic tungsten and molybdenum radiation was also determined. By then repeating the measurement with non-monochromatic radiation it was found that it made no difference to the response of the twin chamber whether it was exposed to monochromatic or non-monochromatic radiation, provided the half-value layers of both were identical and provided, of course, that the energy fluxes determined with the aid of the scintillation instrument were also identical. It was accordingly assumed that the same would apply to all quantum energies involved, and the measurements from which the curve in fig. 15 was plotted were therefore done with non-monochromatic radiation.

The measurements were also carried out with an ionization chamber instead of a scintillation counter. The energy flux (see equation (1)) was then determined from the measured dose rate and the value of μ found for the radiation used. This value was found by again considering the radiation as monochromatic radiation of the same half value layer [14]). As appears from fig. 15, the results obtained by both measurements are quite close.

Although the negative-feedback electrometer circuit for the twin chamber broadly resembles that in the Philips Universal Dosemeter, there are several important differences. Firstly, an electrometer pentode, type 4068, is used instead of a vibrating-reed capacitor. This tube is biased to give a grid current of about 3×10^{-15} A, corresponding to only 0.1% of the smallest difference current measured in practice. Secondly, the current is always integrated. The instrument thus indicates the total energy incident on the patient since the beginning of the examination, and not the energy flux. Further particulars will be found in the publication quoted under [12]).

Several experimental dosemeters for diagnostic use have been built at Eindhoven on the principle described, and have already been used in various hospitals. An instrument of this kind will be marketed in the not too distant future.

Summary. In recent years it has proved desirable to measure the X-ray dose administrated to a patient not only in therapeutic irradiations but also in diagnostic examinations. In therapeutic practice the dose received by a given part of the body is found from the exposure dose measured in roentgens, which is directly related to the ionization produced by the rays in air. A description is given of the ionization chambers and electrometer circuit of the Philips Universal Dosemeter, which can be used for this purpose. In X-ray examinations, it is impracticable to measure the dose in this way. A dosemeter for diagnostic use is described which measures the total radiant energy incident on the patient during an examination. From this it is possible to determine with sufficient accuracy the integral absorbed dose, i.e. the energy absorbed by the patient (unit: kg rad = 10^{-2} joule).

[13]) Developed in cooperation with C. Albrecht of Philips Research Laboratories and A. F. J. van Himbergen of the X-ray and Medical Apparatus Division, Eindhoven.

[14]) Use was made of the data presented in the I.C.R.U. report mentioned in note [3]).

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, the Netherlands, where a limited number of reprints are available for distribution.

2835: J. G. van Wijngaarden: A travelling-wave tube for the frequency band of 3800 to 5000 Mc/s (Le Vide **15**, 36-40, 1960, No. 85; in French and in English).

Details are given of the construction of a travelling-wave tube for the frequency band mentioned. The theory is assumed to be known. Particular mention is made of glass-to-metal seals, alignment and degassing, and the design of the permanent magnet. The object of the design, which is to produce a tube that can be replaced without requiring adjustment, whose life should exceed ten thousand hours, which should possess constant characteristics and have a well-screened magnetic field, is achieved through the use, among other things, of a dispenser cathode (L cathode).

2836: B. B. van Iperen: Reflex klystrons for millimeter waves (Proc. Symp. on millimeter waves, New York, March 31, April 1 and 2, 1959, pp. 249-259, Polytech. Inst. Brooklyn, 1960).

See Philips tech. Rev. **21**, 221-228, 1959/60.

2837: H. J. Prins: Transforms for finding probabilities and variate values of a distribution function in tables of a related distribution function (Statistica neerl. **14**, 1-17, 1960, No. 1).

In statistics, distribution functions are used which can often be transformed one into the other by means of suitable substitutions. This article gives the transforms for the most familiar distributions. With their aid, statisticians are enabled to add to existing tables. Knowledge of the relations between the distributions can also provide more insight into certain methods of statistical testing.

2838: J. S. C. Wessels and H. Baltscheffsky: Adenosine triphosphatase activity in chloroplasts (Acta chem. scand. **14**, 233-246, 1960, No. 2).

Adenosine triphosphate (ATP) was added to a suspension of spinach chloroplasts, and also $MgCl_2$ in some experiments. Inorganic phosphate was then split from the ATP at a temperature of 30 °C, and the amount of phosphate produced after a certain time was determined colorimetrically. The experiments were done with several chloroplast preparations, and the influence of pH was investigated in each case. The reaction is attributed to an enzyme present in the chloroplasts, here referred to as ATP-ase. At $p\mathrm{H} = 7.5$ the activity of the preparations tested was more or less identical. It was greatly stimulated by $MgCl_2$ and inhibited by chlorpromazine. The reaction in the presence of $MgCl_2$ was also investigated kinetically. Experiments in which ATP was replaced by Na pyrophosphate exclude the possibility that the ATP-ase activity at $p\mathrm{H} = 7.5$ involves a liberation and a subsequent hydrolysis of inorganic pyrophosphate. At $p\mathrm{H} =5.5$ the activity differed considerably from one preparation to another, and was not sensitive to $MgCl_2$ and chlorpromazine.

The question is also considered whether a functional relation exists between the investigated reactions and the enzymatic mechanism by which phosphorylation of adenosine diphosphate (ADP) is coupled to electron transport during light-induced phosphorylation.

2839: J. Hornstra: Models of grain boundaries in the diamond lattice, II. Tilt about <001> and theory (Physica **26**, 198-208, 1960, No. 3).

Models are presented for large-angle grain boundaries with a <001> tilt axis. For all angles of tilt, dislocations without dangling bonds can be used in the construction of these models. One model may be considered either as an array of edge dislocations or as an array of 45° dislocations. In the last section the crystallography of regular grain boundaries is discussed, the grain-boundary index is introduced and the application of the transformation matrix to grain-boundary problems is illustrated.

2840: M. Koedam: Sputtering of single crystal metals bombarded with rare gas ions of low energy (50-350 eV) (Proc. 4th int. conf. on ionization phenomena in gases, Uppsala 17-21 August 1959, edited by N. Robert Nilsson, pages ID 252-ID 254, North-Holland Publ. Co., Amsterdam 1960).

Single crystals of Cu and Ni have been bombarded with rare-gas ions of low energy. The directional distribution of the atoms sputtered from a (111) and a (110) surface bombarded with normally incident ions have been determined. Certain preferential sputtering directions are found. The sputtering yield in the preferential (110) direction has been determined as a function of the ion energy (50-350 eV) for Kr^+, Ar^+ and Ne^+ ions.

2841: G. J. M. Ahsmann and Z. van Gelder: The normal cathode fall on single crystal cathodes (as **2840**, pp. ID 266-ID 268).

A glow discharge has been applied to a number of single-crystal cathodes. For most crystals the burning voltage depends on the orientation of the crystal face. Some crystal faces, such as the (100) face of germanium, are not stable but become somewhat roughened by the discharge; it appears that there is a tendency for another crystal face to be exposed by sputtering, probably the (111) face.

It is shown that it is possible to obtain very reliable and stable values of the normal cathode fall by using single-crystal cathodes.

2842: M. Klerk: A contribution to the investigation of the striated positive column (as **2840**, pp. IIA 283-IIA 285).

An experimental investigation of the striated column of low-pressure DC discharges and HF discharges in the range 10-30 Mc/s has led to the conclusion that the origin of the striations is principally seated in the plasma of the column itself. In some cases, however, the boundary conditions at the ends of the column cause a disturbance of the plasma and so give rise to the occurrence of striations.

2843: G. J. M. Ahsmann: The impedance and recovery time of glow discharges in mixtures of rare gases (as **2840**, pp. IIA 309-IIA 313).

It is shown that the complex impedance of a glow discharge in a rare gas is considerably reduced by the addition of another rare gas with a lower ionization potential than the main gas. Both the real and the imaginary part decrease. It is shown that the self-inductance of the discharge decreases in inverse proportion to the mobility of the ions in the Crookes dark space.

The addition of a rare gas with a lower ionization potential than the main gas also has a large effect on the recovery time of the discharge.

2844: Th. P. J. Botden: A decade indicator glow-discharge tube operating on signals of low current and voltage (as **2840**, pp. IID 443-IID 447).

See Philips tech. Rev. **21**, 267-275, 1959/60.

2845: O. Reifenschweiler: Massenspektrometrische Untersuchungen der Ionenemission und Ionenverteilung von Gasentladungsplasmen (as **2840**, pp. IIE 541-IIE 548). (Research on ion emission and ion distribution of gas-discharge plasmas by means of mass spectrometry; in German.)

A method is discussed for investigating the ion emission from gas-discharge plasmas, enabling conclusions to be drawn on the ion-density distribution inside the plasma. Both radio-frequency and arc discharges in hydrogen are used. An ion-extraction system and an electrostatic-lens system produce an image of the plasma boundary, and the current-density distribution of the various types of ions is measured along the diameter of this image. The relative contributions of different ions are found to differ strongly. A relatively weak magnetic field perpendicular to the discharge axis causes the emission maxima of the plasma boundary to shift, the extent of the shift varying with the type of ion. Finally a movable extraction probe is described, suitable for the determination of ion densities at arbitrary points inside the plasma.

2846: C. Z. van Doorn: Thermal equilibrium between F and M centers in potassium chloride (Phys. Rev. Letters 4, 236-237, 1960, No. 5).

Single crystals of KCl (measuring $4 \times 5 \times 1.3$ mm) were heated to 697 °C in potassium vapour at various pressures. A heating time of ten minutes was sufficient to obtain equilibrium, after which the high-temperature equilibrium was frozen in by rapid quenching in CCl_4. The absorption of the F and M bands was measured at 77 °K. The M concentration was found to vary quadratically with the F concentration, which is in agreement with the $2F \rightleftarrows M$ equilibrium (Van Doorn and Haven) and inconsistent with the assumed equilibrium $F +$ vacancy pair $\rightleftarrows M$ (Seitz-Knox model). A rough determination of the equilibrium constant $K' = [M]/[F]^2$ at different temperatures showed the temperature dependence to be small, corresponding to a heat of formation of the M centres of about 0.01 eV.

VOLUME 23, 1961/62, No. 3 pp. 69-100 Published 25th January 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

NEUTRON DIFFRACTION

by J. A. GOEDKOOP *).

539.125.5:620.179.1

Now that more and more nuclear reactors are being put into operation, neutron diffraction is steadily gaining in importance as a method of analysing crystal structures. One principal reason for this is that neutron diffraction can bring to light certain interesting structural details that are difficult or impossible to determine by X-ray diffraction. Cases in point are the sites of hydrogen atoms and the orientation of magnetic moments.

Since 1952 neutron-diffraction investigations have been carried out with the aid of the nuclear reactor at Kjeller, Norway, as part of the research programme undertaken jointly by the (Norwegian) Institutt for Atomenergi and the (Dutch) Reactor Centrum Nederland. In several cases the structural problems tackled in this programme originated in the Philips Laboratories at Eindhoven, and have been solved with their close cooperation. The present article first describes the technique of neutron diffraction, and goes on to show how this technique has provided important additional data on the structure of three compounds, already investigated by X-ray diffraction at Eindhoven [1]).

Introduction

The structure of a crystalline solid can be investigated with the aid of the diffraction pattern produced when the crystal is irradiated with X-rays. The various methods in use werc recently discussed at some length in this journal [1]).

X-radiation is suitable for this purpose because its wavelength is of the same order of magnitude as the interatomic distances in a crystal, that is roughly 1 Å. This being so, a diffraction pattern is produced in which the deflection angles can be measured accurately. This is easily seen from Bragg's equation:

$$2d \sin \Theta = \lambda, \quad . \; . \; . \; . \; . \quad (1)$$

where d is the distance between lattice planes, Θ half the angle of deflection and λ the wavelength of the radiation [2]).

Now it is known from wave mechanics that a parallel beam of particles of mass M and velocity v behaves in many respects like a plane wave of wavelength

$$\lambda = h/Mv, \quad . \; . \; . \; . \; . \; . \quad (2)$$

where h is Planck's-constant. If the mass and velocity of the particles are such that the wavelength is of the order of 1 Å, the phenomena resulting from the passage of these particles through a crystal may be expected to resemble X-ray diffraction phenomena. Shortly after De Broglie put forward equation (2), this was in fact observed from the diffraction of electrons in a crystal. Electron diffraction has since found practical application in the structural analysis of crystals, although it has proved more valuable for investigating the molecular structure of gases. The reason for this is the very high intensity of electron

*) Attached to the Netherlands Reactor Centre at Petten, also Professor Extraordinary of the study of matter with the aid of neutrons at Leyden University.

[1]) P. B. Braun and A. J. van Bommel, X-ray determination of crystal structures, Philips tech. Rev. 22, 126-138, 1960/61 (No. 4).

[2]) The concept "lattice plane" used here differs somewhat from that used in the original formulation of Bragg's law. What was originally the nth-order reflection from a group of lattice planes spaced a distance d' apart, is now a (1st order) reflection from a group of lattice planes whose distance apart is $d = d'/n$. This of course dispenses with the factor n, customarily found in the formulation of Bragg's law.

scattering, as a result of which even a small quantity of gas still gives a reasonably measurable diffraction pattern.

The development of nuclear reactors has in recent years made another elementary particle available in large quantities for structural analysis, namely the neutron, an uncharged particle having a mass of 1.6×10^{-24} grammes (i.e. roughly the mass of a proton). It is easy to calculate that a wavelength of 1 Å corresponds to neutrons with a velocity of 4 km/sec. It is a fortunate circumstance that a nuclear reactor contains very large amounts of neutrons having velocities near this. These neutrons can be extracted through a channel in the reactor shield.

Since the scattering of neutrons differs in various respects from that of X-radiation, its investigation can serve to supplement structural analysis with X-rays. For this reason neutron diffraction as a method of determining crystal structures has made great progress in recent years. In this article we shall deal first with the experimental and theoretical basis of this method [3]). We shall then turn to the three examples described in the above-mentioned article [1]) on X-ray analysis, and show how neutron diffraction has provided additional information in each of these cases.

All experiments were done with the aid of the Kjeller nuclear reactor, Norway, as part of the research programme undertaken jointly by the (Norwegian) Institutt for Atomenergi and thc (Dutch) Reactor Centrum Nederland.

The nuclear reactor as a source of neutrons

In a nuclear reactor, neutrons play an essential part as the carriers of the chain reaction [4]). Upon the splitting of a uranium nucleus that has captured a neutron, other neutrons are liberated, and matters are so arranged in a nuclear reactor that on the average one of these is captured by another uranium nucleus, giving rise to a further fission. The chance of the uranium nuclei capturing a neutron is greater the lower the velocity of the neutron. In most reactors, therefore, the neutrons, which are liberated at very high velocities upon the fission of a nucleus, are decelerated by causing them to collide with the atomic nuclei of a material specially added for that purpose, called the moderator. To slow down the neutrons effectively, these nuclei must be light; moreover they must have no tendency to capture neutrons. Materials meeting these requirements and therefore used in practice as moderators are ordinary water, heavy water (i.e. water containing deuterium instead of hydrogen) and graphite.

As long as the neutrons have a high velocity, they will lose energy upon every collision with an atomic nucleus in the moderator and will thus be slowed down. The nuclei of the moderator, however, are not stationary but are in thermal motion. When the kinetic energy of a neutron has dropped after successive collisions to a value comparable with that of the energy which the moderator atoms possess due to thermal agitation, acceleration as well as deceleration becomes possible. The result is that the neutrons tend to a state of thermal equilibrium with the moderator, and thus move through the latter with the velocities that might be expected for a rarefied gas of atomic weight 1 at the prevailing temperature. Of course, absolute equilibrium is never achieved, since high-energy neutrons are continually added to the system, and moreover low-energy neutrons are absorbed more rapidly than high-energy ones.

The equilibrium distribution of the velocities in such a gas can be derived from the kinetic theory of gases. It is found that, at a temperature of 25 °C, the velocity most frequently occurring is 2200 m/sec, corresponding to a wavelength of 1.8 Å.

Experimental technique of neutron diffraction

The shield around the reactor is fitted with narrow, straight channels, through which neutrons that happen to enter in the right direction can escape freely (*fig. 1*).

In one of the articles referred to above [1]) it is stated that X-ray diffraction patterns are as a rule obtained with *monochromatic* radiation. The neutrons extracted from the reactor, however, have widely differing velocities, corresponding to a wide range of wavelengths. For this reason the first requirement is to select neutrons of the appropriate wavelength from the beam. This is done in a manner familiar in X-ray practice, namely by diffraction: a single crystal (generally a metal crystal) placed in the beam scatters the neutrons in many directions. Each direction corresponds to a different wavelength, determined by the crystal used and its orientation in accordance with Bragg's equation (1). By using such a crystal, a monochromatic beam of the desired wavelength can thus be extracted through an aperture in the shield (fig. 1). This neutron beam, which is of course

[3]) This subject is dealt with at length by G. E. Bacon, Neutron diffraction, Clarendon Press, Oxford 1955; see also C. G. Shull and E. O. Wollan, Applications of neutron diffraction to solid state physics, Part 2, Academic Press, New York 1956; G. R. Ringo, Handbuch der Physik, Part 32, 552, Springer, Berlin 1957.

[4]) See J. J. Went, Philips tech. Rev. **21**, 109, 1959/60.

much weaker than the original one, can then be used in much the same way as a monochromatic beam of X-rays. One can thus, for example, place a single crystal or a powdered specimen in the beam and ascertain the intensity with which the neutrons are scattered in various directions.

Neutrons are not capable of blackening a photographic film. Although there are ways and means of recording a neutron diffraction pattern photographically, in practice neutrons are generally detected by means of a counter tube, a device which is also coming increasingly into use in X-ray diffraction [5]). Since they possess no charge, neutrons cannot cause ionization and thus cannot actuate a counter tube directly. This difficulty is overcome by using a counter tube filled with boron-trifluoride gas. The boron nucleus readily captures neutrons, disintegrating into two charged fragments. The ionizing action of these fragments produces a discharge in the counter tube, so that the number of neutrons entering the tube can be counted.

To determine the intensity distribution in various directions, the neutron counter is mounted on the moving arm of a goniometer, so that it describes an arc around the investigated specimen, on the same principle as in an X-ray diffractometer [6]). Details of the arrangement will be found in fig. 1, which shows the main diffraction set-up at Kjeller [7]). *Fig. 2* shows the equipment in use.

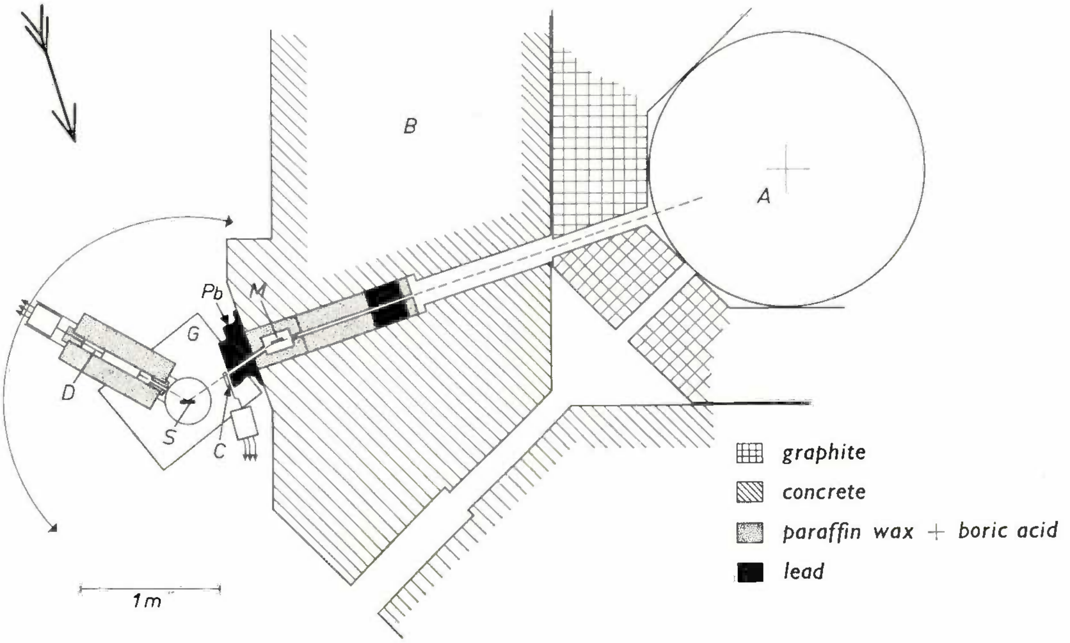


Fig. 1. Diagram of the neutron-diffraction system at Kjeller. The neutrons are extracted from the reactor *A* through a channel in the reactor shield *B*; they encounter in their path a large single crystal of lead *M*, which — still within the shield — scatters them in a number of directions. Neutrons travelling in one specific direction pass through an aperture in the lead shield *Pb*: the whole system thus works as a monochromator. In the arrangement at Kjeller the crystal is so situated that only those neutrons that are reflected by the (111) planes of the crystal pass through the aperture, resulting in a monochromatic beam of wavelength 1.03 Å. The sample *S* to be analysed (single crystal or powder) is mounted on the shaft of a goniometer *G*, around which the BF_3-filled neutron counter *D* can be rotated. The counter is surrounded by a thick shield to protect it from stray neutrons in the reactor building. The apparatus works automatically in steps: when the counter is stationary the number of scattered neutrons is counted during a period in which a specific number of neutrons (e.g. 100 000) are recorded in a small monitor counter *C*, which intercepts part of the monochromatic beam. At the end of this period the number counted is printed on a paper strip and the counter arm is swung through an angle of 0.2°, the slide carrying the sample being shifted through an angle of 0.1°. A new counting period then starts. — Fig. 2 shows a photograph of this set-up taken in the direction of the arrow shown here top left.

[5]) See W. Parrish, X-ray intensity measurements with counter tubes, Philips tech. Rev. **17**, 206-221, 1955/56.

[6]) See W. Parrish, E. A. Hamacher and K. Lowitzsch, Philips tech. Rev. **16**, 123, 1954/55.

[7]) For further particulars see J. A. Goedkoop, Ned. T. Natuurk. **23**, 140, 1957.

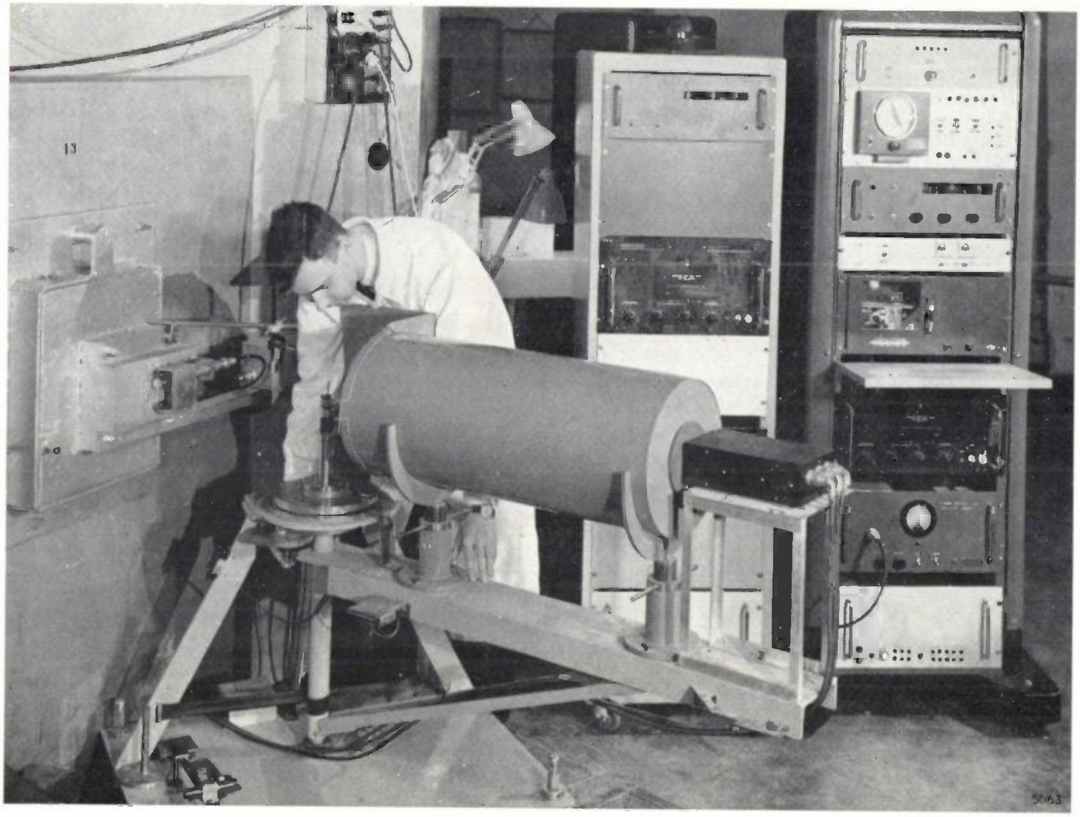


Fig. 2. View of the neutron-diffraction equipment at Kjeller. The direction of observation is indicated by the arrow in fig. 1. The neutron aperture, covered by a plate, is situated in the middle of the lead block at the left-hand side of the picture. The monitor counter can be seen to the right of this aperture. The sample is visible in front of the experimentor's sleeve. The large cylinder is the shield around the neutron counter, which is mounted on a goniometer arm that turns around the same axis as the sample. On the extreme right of the arm can be seen the pre-amplifier for the neutron counter. Under the sample holder can be seen some of the gear wheels of the driving mechanism. The control and recording system is accommodated in two racks in the background. The results are printed on a paper strip in the middle of the right-hand rack. (From: Atoomenergie en haar toepassingen **2**, 82, 1960.)

Neutron scattering by the atomic nucleus

Electromagnetic radiation is scattered by charged particles with an amplitude inversely proportional to the mass of the particle. For this reason X-ray scattering only provides information on the distribution of the *electrons* in a crystal: the atomic nuclei are too heavy to make any perceptible contribution.

In the case of neutron scattering it is the *nuclei* that cause the scatter, whilst the interaction between neutron and electron is very slight (with one important exception, to which we shall return presently).

The neutron, as will be known, is a normal constituent of all atomic nuclei, apart from ^{1}H. The forces acting between a neutron and an atomic nucleus in a collision are of the same nature as those that hold the neutrons and protons together inside the nucleus. A characteristic of these forces is their short range, roughly 10^{-12} cm, which is of the same order of magnitude as the diameter of the atomic nucleus. Since the wavelength at which we are working is about 10^{-8} cm, we can regard the atomic nucleus as a point scatterer, which means that the scattering is isotropic. The scattering action of a given atomic nucleus can thus be described by a single quantity, called the *scattering length* and denoted by the letter b.

In order to explain the significance of b, we shall consider a parallel beam of neutrons, containing n neutrons per cm^3, all having the velocity v cm/sec. The neutron flux is then nv, in other words nv neutrons pass per second through unit area perpendicular to the beam.

In terms of wave mechanics the beam is represented by a plane wave of amplitude $\sqrt{n}$ (the square of the absolute value of the wave function must everywhere be equal to the local density). A single atomic nucleus placed in the beam, being a point scatterer, will be the centre of a spherically scattered wave whose amplitude is proportional to $\sqrt{n}$ and inversely proportional to the distance r from the centre of the nucleus. The amplitude may therefore be represented by $(b/r)\sqrt{n}$. We have thus defined the quantity b, which evidently has the dimension of a length.

We may now introduce in this connection another useful concept. It follows from the above that the density of the scattered neutrons at a distance r is equal to $(b/r)^2n$. Supposing the scattering nucleus to be surrounded by a sphere of radius r, the number of scattered neutrons passing through this sphere per second will be

$$4\pi r^2 \times v \times (b/r)^2 n = 4\pi b^2 nv\,. \;.\;.\;.\;.\;. \quad (3)$$

Of the nv neutrons striking an area of 1 cm^2 the nucleus will thus, as it were, scatter those that are incident on a part $\sigma = 4\pi b^2$ of that area, and the others not at all. The quantity σ is therefore called the *effective scattering cross-section*, and is generally of the same order as the cross-section of the atomic nucleus.

Table I gives the scattering lengths [8]) for the elements with which we shall be concerned in the examples presented in this article. It would be beyond our present scope to deal with the physical factors governing the scattering length. In any case, a quantitative treatment of these factors is not as a rule possible, and in most cases the scattering lengths are determined empirically. We will not discuss this either, and shall thus simply consider the scattering lengths as given quantities in what follows.

Table I. Scattering lengths of some atoms for neutrons and for X-radiation.

Element	Atomic number Z	Neutron scattering length *) b in 10^{-12} cm	X-ray scattering length c in 10^{-12} cm	
			for sin $\Theta/\lambda = 0$	for sin $\Theta/\lambda = 0.5\times10^8$ cm^{-1}
Hydrogen	1	H: − 0.38 D: 0.65	0.28	0.02
Oxygen	6	0.58	2.25	0.62
Aluminium	13	0.35	3.65	1.55
Manganese	25	−0.37	7.0	3.1
Iron	26	0.96	7.3	3.3
Zinc	30	0.59	8.5	3.9
Barium	56	0.53	15.8	8.3
Thorium	90	1.01	25.2	14.4

*) Hydrogen excepted, these are the scattering lengths for the natural isotopic mixture.

The tabulated values call for some comments, of importance to our subsequent considerations.

a) The various isotopes of an element have different scattering lengths, a fact which is clearly demonstrated by the striking difference between H and D. An average value is taken for the other elements, applicable to the natural isotopic ratio.

b) In two cases, H and Mn, (and in the case of some other elements not mentioned here) the scattering length is negative. The significance of this may be understood when it is realized that the scattered neutron wave can have different phases in relation to the incident wave. In practice, two principal cases arise. At the scattering nucleus, the scattered wave is either in phase with the incident wave or in anti-phase with it. These two cases may be indicated by giving b a positive or negative sign: the phase difference is nearly always 180°, which we indicate by a plus sign; a phase difference of 0° thus corresponds to b negative.

c) Although the scattering lengths in Table I show a definite tendency to increase with increasing Z (as can be seen more clearly from a table for all elements), this increase is not particularly pronounced compared with the very marked irregularities (compare, for example, the neighbours Mn and Fe).

Comparison with X-ray scattering

We have noted that X-rays are scattered by the electrons in a crystal. Consequently, the X-ray-scattering action of an atom increases more or less linearly as the number of electrons increases, that is to say with increasing atomic number Z in the periodic system. If we consider a compound containing e.g. thorium ($Z = 90$) and hydrogen ($Z = 1$), it is evident that the scattering due to the hydrogen will be entirely insignificant compared with that due to the thorium. It will therefore be difficult, if not impossible, to determine by means of X-ray diffraction exactly where the hydrogen is located.

A glance at Table I shows that neutron diffraction gives much more chance of determining the position of the hydrogen nuclei.

It may also happen that a crystal contains two elements that are close together in the periodic system. In that case the electron density will not change appreciably if the two atoms exchange sites, so that it is difficult to distinguish one atom from the other on the basis of the X-ray diffraction pattern. Here, too, neutron diffraction makes more reliable determination possible. A clear case in point is that of Fe and Mn, already mentioned in passing; another example discussed below also demonstrates a third very important difference between X-ray and neutron diffraction, which is that the latter is able to provide information about the magnetic structure of crystals.

Before going into this interesting aspect and discussing examples, we shall pursue for a moment the comparison of neutron and X-ray scattering. Electrons are point scatterers for X-radiation, but the observed X-ray diffraction pattern is necessarily an average over a time in which the electrons have undergone many revolutions. For that reason the picture we get of the electrons is a "smeared"

[8]) After Bacon, loc. cit. [3]).

continuous distribution of electrons in space. Something similar, though to a much lesser extent, applies to the nuclei as scatterers of neutrons. As a first approximation the nuclei are located at fixed sites in the crystal; the scattering of neutrons is thus everywhere zero except at the sites of the nuclei (integrating the neutron scattering over the cross-sectional area of the nuclei, we obtain a value proportional to the scattering length). The thermal agitation of the nuclei, however, is responsible for some spread in the mass distribution, although as a rule the resultant "smear" of the nuclei is small compared with that of the electron clouds.

The similarity between the mass distribution of the nuclei and that of the electrons, which obviously share the same periodicity in crystals, makes it clear that we can resolve the density of the nuclei into Fourier components, just as is known to be possible for the electrons. In the article cited above [1]) it was shown that these Fourier components can in fact be derived from the diffraction pattern. The structure can then in principle be determined simply by adding the components (Fourier synthesis). In *fig. 3a* and *b* an example is given of the application of this Fourier method with X-rays and neutrons, respectively.

Although this indicates that neutron diffraction and X-ray diffraction may in principle be identically treated, the quantitative difference mentioned nevertheless means that the comparison is no longer valid in some points. In the first place, where X-rays are concerned the relatively large extent of the electron cloud gives rise to considerable path differences between the rays scattered by one atom. Owing to the phase differences involved, the X-ray intensity resulting from the scattering by a single electron cloud decreases with increasing angle of deflection, and there is consequently a tendency for the reflections of increasing angle in the diffraction pattern to die out. In view of the much more concentrated distribution of the nuclear mass, no such decrease occurs with neutron diffraction. Secondly, this more concentrated distribution of the nuclear mass makes

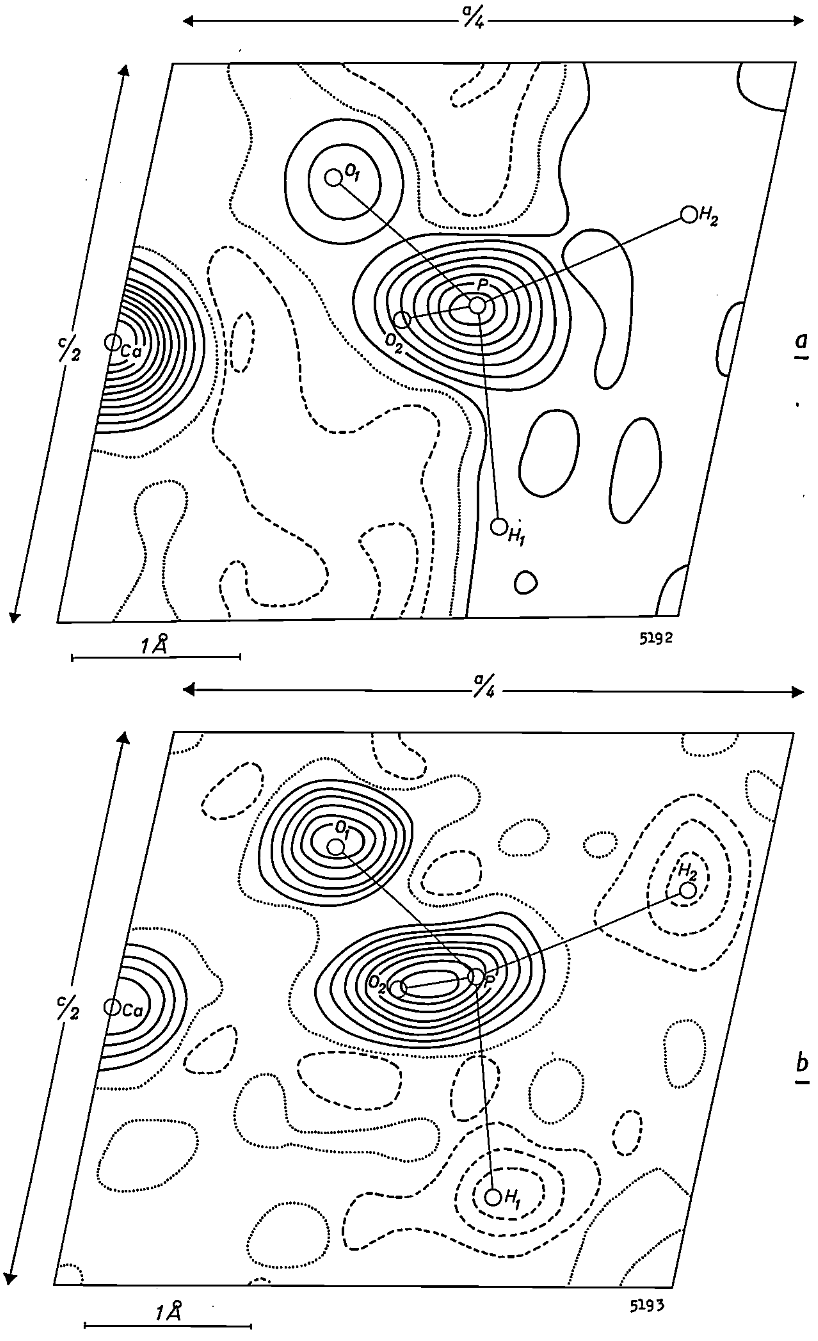


Fig. 3. Crystal structure of the compound $Ca(H_2PO_2)_2$, *a*) determined by X-ray diffraction, *b*) determined by neutron diffraction. Use was made in both cases of the Fourier method described in the article referred to under [1]). In *a*) the contours denote the distribution of the electrons, and in *b*) the distribution of the nuclei (the solid lines join places of equal positive density, the dotted lines places of zero density, and the dashed lines places of equal negative density). Note that the hydrogen atoms, of which there is no trace in *a*), are clearly visible in *b*), demonstrating one of the advantages of neutron diffraction.

The negative sign of the density at certain places in *a*) and in the "flat" parts of *b*) has no physical significance: it arises from the method of calculation adopted; for convenience a constant positive term has been omitted at all points. As regards the hydrogen atoms in *b*), however, the sign does have a physical reason, which is a phase effect involved in the scattering. For particulars see text. (After B. O. Loopstra, JENER Publications No. 15, 1958.)

the Fourier method rather unsuitable: the alternative method of "trial and error" is to be preferred.

The applicability of this much simpler method is due to the fact that neutron diffraction analysis is frequently meant to provide only supplementary information on the basis of a fairly detailed picture of the structure already provided by X-ray analysis. The object may be, for example, simply to determine the position of the hydrogen atoms, with a choice to be made between several possible positions, or to determine whether the magnetic moments present have this or that orientation. Questions such as these can be answered quite satisfactorily by the method of trial and error. This method consists in testing various crystal models by comparing the observed reflection intensities with those calculated on the basis of the models. In simple cases the method can also be used in X-ray diffraction. The calculation of the (relative) intensities of the reflections, i.e. the squares of the scattering amplitudes F, from an assumed crystal model is in itself no problem, and in the case of neutron diffraction is even somewhat simpler than in the case of X-ray diffraction. The scattering amplitudes F can be found quite simply by adding the scattering lengths of the atoms contributing to the scatter, taking into account the phase differences due to the various path differences.

Because the Fourier method is not much used in neutron diffraction analysis, we shall make no further use in this article of the representation of the above-mentioned Fourier components (density waves), on which the article cited under [1]) was based. This representation, where each reflection is regarded as produced from the fronts of a density wave, is particularly useful for elucidating the Fourier method. As regards the method of trial and error, we can base our considerations more simply on the alternative representation of individual scattering atoms, where each reflection is regarded as being produced by lattice planes.

We shall now briefly indicate how the phase differences are taken into account in the calculation of the diffraction intensities.

Suppose that atomic nuclei designated by j, having a scattering length $\mathbf{b}_j$, are situated at positions $\mathbf{r}_j$. Let S_0 and S be unit vectors in the incident and scattered directions, respectively; by adding the contributions, differing in phase, of the various atoms, we then find that the amplitude of the scattered wave must be proportional to

$$F = \sum_j \mathbf{b}_j \exp\{-2\pi\,\mathrm{i}\,\mathbf{r}_j \cdot (\mathrm{S}-\mathrm{S}_0)/\lambda\}. \quad . \; . \; . \quad (5)$$

A similar formula can be written for the scattering of X-radiation:

$$F = \sum_j \mathbf{c}_j \exp\{-2\pi\,\mathrm{i}\,\mathbf{r}_j \cdot (\mathrm{S}-\mathrm{S}_0)/\lambda\}. \quad . \; . \; . \quad (6)$$

The scattering lengths $\mathbf{c}_j$ in this equation, unlike the neutron scattering lengths, depend on the angle of deflection and the wavelength; in fact, $\mathbf{c}_j$ decreases with increasing value of $\sin\Theta/\lambda$. This naturally complicates the calculation somewhat, since for every value of $\sin\Theta/\lambda$ we must now know the corresponding value of the scattering length of each atom. The reason for this effect has already been mentioned, being the extensiveness of the electron cloud compared with the wavelength of the radiation. To give an idea of the magnitude of the decrease in question, columns 4 and 5 in Table I show the X-ray scattering lengths of the elements tabulated for $\sin\Theta/\lambda = 0$ and for $\sin\Theta/\lambda = 0.5\times10^8$ cm^{-1} respectively.

We shall now deal with the three examples of neutron analysis to which we have alluded, the supplementary structural data being obtained by the method of trial and error just discussed.

First example: determination of hydrogen atom sites in the alloy Th_2Al

The alloy Th_2Al can absorb gases. It may, for example, absorb hydrogen until the composition Th_2AlH_4 is reached. In the process the dimensions of the tetragonal unit cell are changed, but the relative positions of the Al and Th atoms remain unchanged. This is clearly revealed by X-ray diffraction. In Debye-Scherrer patterns of compounds having the compositions Th_2Al, Th_2AlH_2, Th_2AlH_3 and Th_2AlH_4, shown in *fig. 4*, it can be seen that the diffraction peaks are shifted, but their intensities remain virtually unchanged.

It is again quite evident from the latter fact that it is impossible to learn anything about the position of the hydrogen atoms from the X-ray scattering. There was therefore good reason to investigate these compounds with the aid of neutrons.

Since only polycrystalline samples of these compounds are available, their use in neutron diffraction involves a difficulty, especially where they contain hydrogen. In that case, in addition to *coherent* scattering, which obeys Bragg's law and with which we have been exclusively concerned in the foregoing, strong *incoherent* scattering occurs, which makes it very difficult to observe the diffraction peaks. (Incoherent scattering is discussed in the appendix to this article. Its salient feature is that, by contrast with the ideal case, there is not complete extinction of all the scattering contributions in other than the Bragg angles.)

In these investigations, therefore, deuterides were used instead of hydrides, because they give appreciably less incoherent scattering. It may safely be assumed that the deuterium atoms will occupy the same sites in the deuterides as the hydrogen atoms in the hydrides.

An incidental advantage of the use of deuterides instead of hydrides — although it scarcely makes up for the much greater difficulty of preparing the sample — is that deuterium has a greater scattering length than hydrogen, as can be seen from Table I.

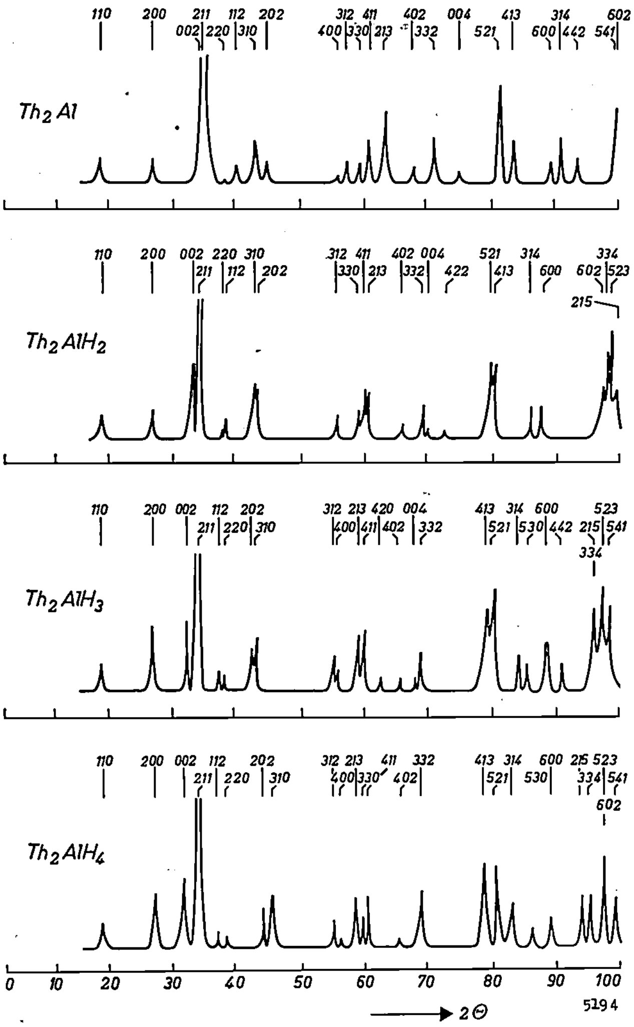


Fig. 4. Debye-Scherrer patterns of Th$_2$Al, Th$_2$AlH$_2$, Th$_2$ A /H$_3$ and Th$_2$AlH$_4$ obtained with X-rays of 1.79 Å. The intensit y $_{\text{is}}$ plotted versus the angle of deflection 2Θ. The Miller indices are given above the scattering curves. The diagrams differ from one another in the situation of the diffraction peaks (owing to a change in the size of the unit cell). There is little difference, however, in the intensities of the peaks, because of the small hydrogen contribution. The picture changes when neutrons are used instead of X-rays: compare these diagrams with those in fig. 5.

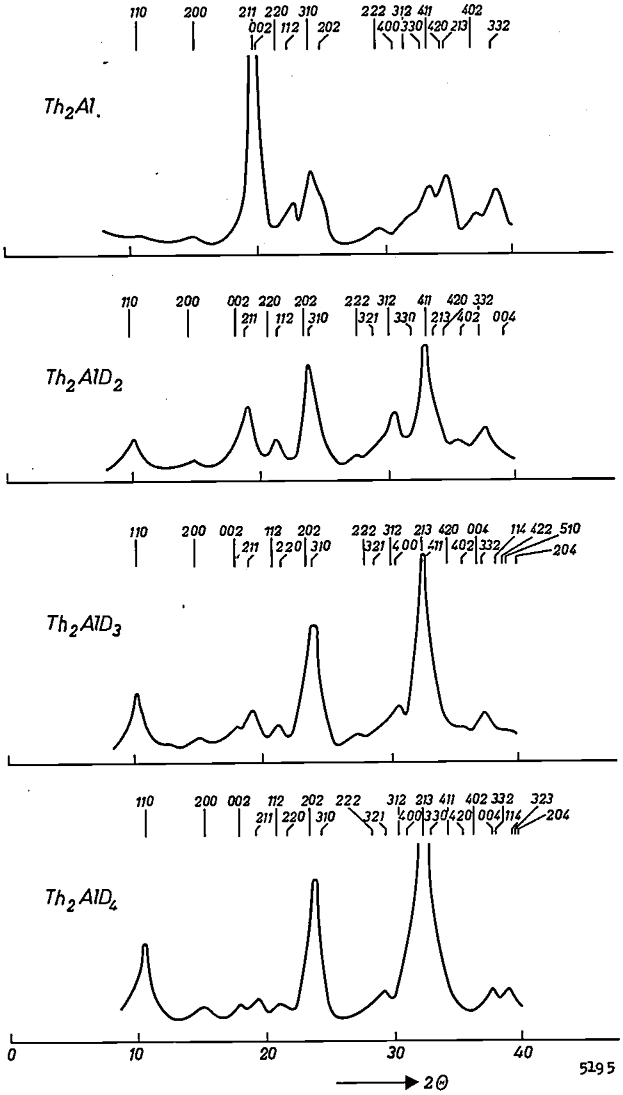


Fig. 5. Debye-Scherrer patterns of Th$_2$Al, Th$_2$AlD$_2$, Th$_2$AlD$_3$ and Th$_2$AlD$_4$ obtained with neutrons of 1.03 Å. The marked differences in the intensity of corresponding diffraction peaks demonstrate the very considerable scattering contribution made by the deuterium.

For neutron analysis [9]) samples of differing deuterium content were placed in a thin-walled tube of 1 cm diameter and mounted on the goniometer. The observed intensity, i.e. the pulse density (count rate) measured with the counter tube, was plotted as a function of the scattering angles for the various compositions in *fig. 5*.

Comparison with the X-ray diffraction patterns in fig. 4 shows practically the same succession of peaks, indicating that the substitution of deuterium for hydrogen causes no essential change in the crystal structure. The (relative) intensities of the maxima are now, however, entirely different, which is a clear illustration of the marked contribution which the deuterium makes to the scattering.

Comparison of the observed intensities with the values calculated for various models establishes the positions of the deuterium (and hence of the hydrogen) atoms in the unit cell. As assumed in [1]), they are situated in the tetrahedral interstices between the thorium atoms. There are four of these interstices for every aluminium atom, in agreement with the fact that the maximum hydrogen content achieved corresponds to the composition Th$_2$AlH$_4$. *Fig. 6* shows the unit cell of Th$_2$Al with four of these tetrahedra and the associated hydrogen atoms.

[9]) J. Bergsma, J. A. Goedkoop and J. H. N. van Vucht, Acta cryst. 14, 223, 1961 (No. 3).

Neutron diffraction by magnetic atoms and ions

Neutrons possess a spin, i.e. an angular momentum, associated with a magnetic moment. This means that, contrary to what has been said above, they *can* interact with an electron cloud when the latter is the carrier of a resultant magnetic moment. The scattering then produced is in general of the same order of magnitude as that caused by the atomic nucleus, so that if a crystal contains magnetic atoms or ions, scattering due to both causes may be observed.

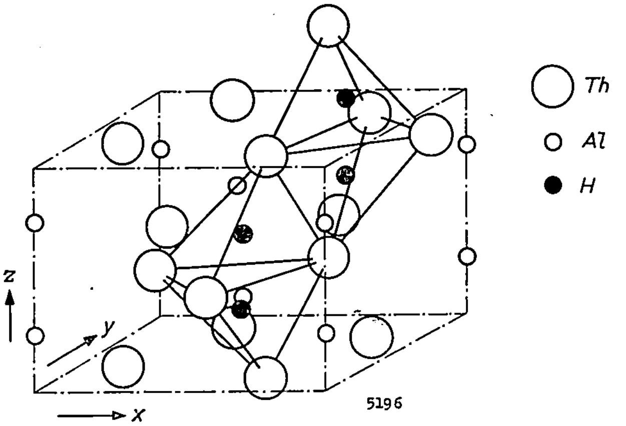


Fig. 6. The unit cell of Th_2Al. Neutron diffraction analysis shows that the gas atoms absorbed are situated in the tetrahedral interstices formed by the thorium atoms; four such tetrahedra, each with its hydrogen atom, are shown in the drawing. Where all these interstices are filled with hydrogen, the formula of the compound is Th_2AlH_4. This compound still has the same structure as Th_2Al, i.e. the thorium and aluminium atoms have the same relative positions, the change being a difference in the size of the cell.

The character of magnetic scattering depends strongly on the extent to which the magnetic moments of the atoms are in mutual alignment. If the crystal is paramagnetic, that is to say possesses no preferred orientation, the magnetic scattering is then not concentrated in diffraction peaks, as it is when due to the nuclei, but is diffuse.

Many substances are, however, only paramagnetic above a particular temperature; below this temperature the interaction between neighbouring atoms prevails over thermal agitation, so that over certain regions in the crystal (magnetized domains) all magnetic moments are oriented parallel or anti-parallel to a certain direction. Where the moments are all aligned in the same direction, the substance is said to be ferromagnetic; if some point in one direction and the others in the opposite direction, the substance is antiferromagnetic or ferrimagnetic, depending on whether the resultant moment per unit cell is zero or not. In the case of ferromagnetism and ferrimagnetism, the application of a sufficiently strong magnetic field influences the domains and brings about an externally measurable magnetization. The magnetic neutron scattering of such crystals possessing ordered magnetic moments does resemble to an appreciable extent the nuclear scattering so far considered, that is to say it is likewise concentrated in diffraction peaks according to Bragg's law. Their intensities are calculated in the same way as for nuclear scattering, except that instead of the nuclear scattering length b we must now use the quantity

$$p = 0.539 \times S\, \mathrm{f}(\sin \Theta/\lambda), \quad . \quad . \quad . \quad . \quad (7)$$

expressed in 10^{-12} cm.

S is the effective spin-quantum number, which is a measure of the magnetic moment of the scattering atom [10]) (the latter is equal to $2S$ Bohr magnetons). In antiferromagnetic and ferrimagnetic substances, S is given the positive or negative sign according to whether the moment is aligned in the one direction or the other. The other factor, $\mathrm{f}(\sin \Theta/\lambda)$, is the magnetic form factor, which is equal to unity where the reflection angle is zero, and becomes smaller as the angle decreases. The magnetic scattering of neutrons, unlike nuclear scattering, is indeed dependent on the scattering angle, and in this respect it shows more correspondence with X-ray scattering. The reason, of course, is that magnetic neutron scattering is also attributable to an electron cloud whose extent is of the order of 1 Å.

The dependence of this form factor on the scattering angle is not identical with that of the earlier introduced X-ray scattering length c, for the latter relates to the entire electron cloud around the nucleus. That cloud consists of individual shells, and the magnetic moment, if present, is attributable to the incomplete filling of one of those shells (the 3d shell, for example, in the case of Fe and Mn). It is therefore only the electrons in this shell that cause the magnetic scattering of neutrons, and consequently the magnetic form factor $\mathrm{f}(\sin \Theta/\lambda)$ is determined by the radial density distribution of those electrons alone.

In the same way as the nuclear scattering amplitude, which we shall now call F_{nucl}, was composed of the scattering lengths b of the atomic nuclei in the crystal, each multiplied by the appropriate phase factor, we can now combine the magnetic scattering lengths of the various magnetic atoms as given by equation (7) to form a magnetic amplitude, F_{magn}.

In most crystals there is no difference between the periodicity of the magnetic and the other atoms. The interference conditions are then identical for both magnetic and nuclear scattering, and therefore the observed diffraction intensities will consist of a

[10]) Our considerations apply to the case where the magnetic moment is entirely due to the electron spins, and not to the orbital motion of the electrons.

nuclear and a magnetic part. The theory shows that the total intensity is then proportional to

$$I = F^2_{\text{nucl}} + (\sin a)^2 \, F^2_{\text{magn}} \,, \quad . \; . \; . \quad (8)$$

where a is the angle between the preferred orientation of the magnetic moments in the crystal and the bisectrix of the angle between the incident and the scattered beams. If a is zero, the scattering is purely nuclear.

In some cases of antiferromagnetism and ferrimagnetism the magnetic moments of corresponding atoms in adjacent unit cells are oppositely oriented, and equally oriented moments are always twice as far apart. In that case there is thus indeed a difference in periodicity, the unit cell being larger in a magnetic respect than in a nuclear respect, and *extra reflections of purely magnetic origin will then occur.* The best known example of this is cubic MnO, where all edges of the unit cell are twice as long in the antiferromagnetic state [11]). It may be mentioned that neutron diffraction analysis has recently revealed cases of magnetic ordering (e.g. in metallic holmium) where neighbouring magnetic moments make angles with one another differing from 0° or 180°, giving rise to a spiral orientation [12]).

It might seem surprising that the intensities of the nuclear and magnetic scattering components are simply added together in equation (8). It might be thought that one would first have to add the amplitudes and then square their sum. This is certainly the right method if the neutrons are polarized, that is to say if the spins of the incident neutrons are artificially brought into identical alignment. The neutrons coming from the reactor are of course unpolarized, in which case the additive method gives the correct composition of the intensities. In the following we shall be concerned only with this case.

The magnetic scattering of neutrons makes it possible to determine the value of S for the various atoms in an ordered magnetic crystal, i.e. to establish the direction and magnitude of the individual atomic magnetic moments. We shall now consider two examples of this possibility, which obviously has no equivalent in X-ray diffraction.

Second example: the magnetic moments of Mn atoms in the alloy $Al_{0.89}Mn_{1.11}$ [13])

In the alloy $Al_{0.89}Mn_{1.11}$ the aluminium and manganese atoms are statistically distributed in a specific way over the two positions available for them in the tetragonal unit cell, i.e. the corner (A) and the middle (B) of the cell. It had been deduced from X-ray diffraction that the order parameter r, defined as the average fraction of Al atoms on the A sites, amounts to 0.03. This is explained in the article referred to under [1]). It is also mentioned there that the alloy can be made into a permanent magnet, so that there must be domains within which the magnetic moments of the manganese atoms are ordered. The following questions then arise:

a) In what crystallographic direction are the magnetic moments preferably oriented (in the absence of an aligning magnetic field, so that the moments are free to orient themselves)?

b) If a manganese atom occupies a B site, is its magnetic moment equally or oppositely oriented to that of the manganese atoms on most of the surrounding A sites? In other words, is the coupling between these neighbours ferromagnetic or ferrimagnetic? The results of X-ray diffraction analysis and measurement of the saturation magnetization of samples treated in different ways have prompted the supposition that this coupling is ferrimagnetic.

c) How large is the magnetic moment of each Mn atom?

The answers to these questions are to be found from the Debye-Scherrer pattern in *fig. 7*, obtained by neutron diffraction. The intensities derived from the peaks by planimetry are given in the last column of *Table II.*

Both in the figure and in the table the diffraction peaks are characterized by the Miller indices *h*, *k*

Table II. Observed and calculated intensities of neutron reflections from polycrystalline $Al_{0.89}Mn_{1.11}$.

Miller indices *hkl*	Magnetic angle factor $(\sin a)^2$	Magnetic form factor $f(\sin \Theta/\lambda)$	Calculated intensities				Observed intensities *)
			nuclear contribution	magnetic contribution	total		
001	0.00	—	6.56	0.00		6.56	6.70±0.12
100	1.00	0.68	8.00	3.59		11.59	11.66±0.11
101	0.62	0.54	0.22	0.98		1.20	1.32±0.08
110	1.00	0.45	0.09	0.44		0.53	0.73±0.07
002	0.00	—	0.03	0.00	0.03	6.68	6.47±0.12
111	0.77	0.34	6.12	0.53	6.65		
102	0.29	0.28	4.61	0.10		4.71	4.52±0.10
200	1.00	0.20	0.05	0.05		0.10	0.12±0.06
112	0.45	0.19	0.08	0.03	0.11	7.79	8.20±0.19
201	0.86	0.18	3.45	0.10	3.55		
210	1.00	0.15	3.18	0.07	3.25		
003	0.00	—	0.73	0.00	0.73		
211	0.89	0.11	0.12	0.03	0.15		
202	—	—	0.05	—	0.05	0.11	0.14±0.08
103	—	—	0.05	—	0.05		
212	—	—	4.33	—	4.33	6.48	6.26±0.12
113	—	—	2.15	—	2.15		

*) The limits of error indicated in this column relate only to the probable statistical error in counting the number of neutrons.

[11]) C. G. Shull and J. S. Smart, Phys. Rev. **76**, 1256, 1949.

[12]) Personal communication from W. C. Koehler.

[13]) P. B. Braun and J. A. Goedkoop, to be published in Acta cryst.

and l, whose significance we shall briefly recall. Let a, b and c be the edges of the unit cell (a parallelepiped), then two successive lattice planes with Miller indices h, k and l cut from the three edges respectively the portions a/h, b/k and c/l. The unit cell is here tetragonal; $c = 3.54$ Å is the tetragonal axis; the axes $a = b = 2.77$ Å are perpendicular thereto.

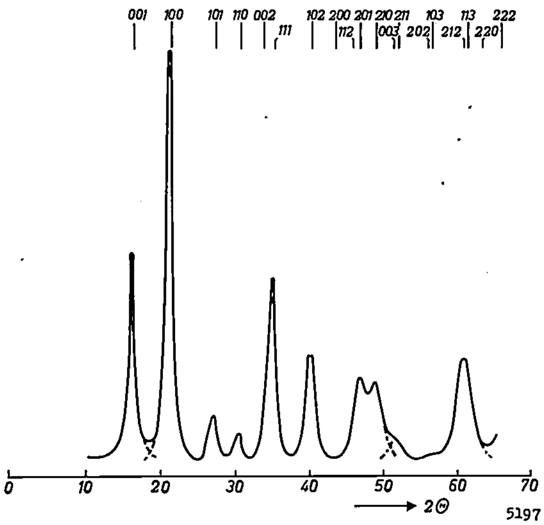


Fig. 7. Debye-Scherrer patterns of the alloy $Al_{0.89}Mn_{1.11}$ with neutrons of 1.03 Å. At large deflection angle 2Θ the intensities are almost entirely determined by nuclear scattering. They are weak where $h + k + l$ is even, and strong where $h + k + l$ is odd. This is due to the preference shown by the manganese and aluminium atoms for the corners and centre, respectively, of the tetragonal unit cell, and to the opposite sign of the scattering lengths of these atoms, see Table I. (In X-ray diffraction, where there is no such difference in sign, the reflections are strong where $h + k + l$ is even, and weak where $h + k + l$ is odd, i.e. the situation is the reverse of that in neutron diffraction.) The reflections 100, 101 and 110, but not 001, comprise an appreciable magnetic contribution in addition to that due to the nuclei.

Regarding the calculation of the intensity of the reflections, only two cases are to be distinguished here, namely reflections corresponding to lattice planes where $h + k + l$ is even and where $h + k + l$ is odd; for brevity we shall refer to them as even and odd reflections. It is easily seen that, if the A site lies in a lattice plane of a particular set, the B site for an even reflection will also lie in a lattice plane of the same set, but for an odd reflection it will lie midway between the lattice planes of that set. Thus, for even reflections the scattering contributions of A and B positions must be added, whereas for odd reflections it is necessary to take the difference.

In analysing the diffraction pattern we begin with large scattering angles, for it is here, owing to the decrease in magnetic scattering already discussed, that the intensities are almost entirely of nuclear origin. As with X-ray diffraction, the measured scattering intensities are compared with the intensities calculated for different values of the order parameter r. The best agreement was found for $r = 0.03$, which agrees within the accuracy of the measurement with the value found by X-ray diffraction. The corresponding distribution of the Al and Mn atoms over the corners and centres of the unit cell is represented in *fig. 8*.

We can now calculate the nuclear contributions of all intensities, including those at small scattering angles. The relevant values are given in the fourth column of Table II.

Comparison of these with the measured intensities makes it clear that there must still be appreciable magnetic contributions at small scattering angles. Let us assume for a moment that the magnetic moments of Mn atoms on A and B sites are in parallel orientation. At the value found for r the A sites will be occupied by an average of 0.97 Mn atoms, and therefore, in view of the total composition, there will be 0.14 on B sites. For even reflections we must add the scattering contributions; the scattering amplitude per unit cell is thus:

$$F_{\text{magn}} = 1.11\,p \quad (\text{with} \quad h + k + l \quad \text{even}),$$

where p is the magnetic scattering length per Mn atom, to be calculated from eq. (7). For odd reflections (still with the two magnetic moments parallel) we must subtract one scattering contribution from the other, giving

$$F_{\text{magn}} = 0.83\,p \quad (\text{with} \quad h + k + l \quad \text{odd}).$$

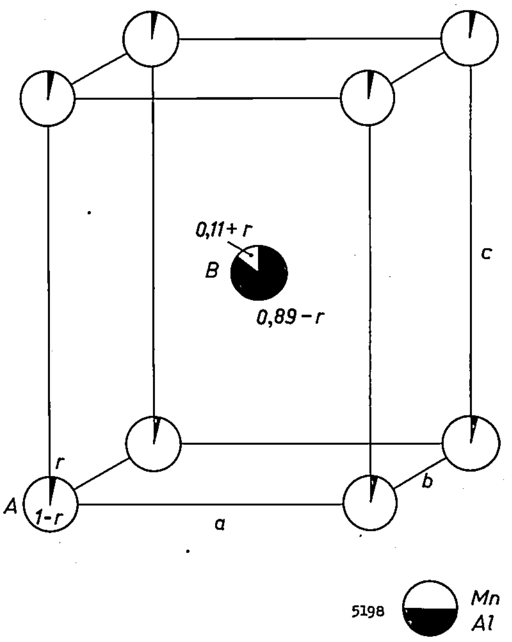


Fig. 8. The distribution of the aluminium and manganese atoms over the corners and centres of the unit cell of the compound $Al_{0.89}Mn_{1.11}$. The average fraction r of the aluminium atoms at the corners, here represented as a sector of a circle, is roughly 0.03, as found by both X-ray and neutron diffraction.

If, on the other hand, the moments on A and B sites are oppositely oriented, then according to the remarks concerning equation (7) the contribution of the B sites must be given the negative sign so that the two above equations must be interchanged.

As may be seen from Table II, the observed intensity for the first reflection (001) can be explained within the limits of experimental error by the nuclear contribution alone. It follows from the foregoing that F_{magn} cannot be zero, so we must assume that a in eq. (8) is zero for this reflection.

The Miller indices indicate that this reflection depends on lattice planes perpendicular to the tetragonal c axis. Since a is the angle between the perpendicular to the lattice planes and the preferred direction of the magnetic moments in the crystal, it may therefore be concluded that the magnetic moments must be oriented along the c axis, which answers question (a).

Armed with this knowledge, we can calculate $(\sin a)^2$ from equation (8) for the other reflections: the results are given in the second column of Table II. The third column contains the values of the magnetic form factor $f(\sin \Theta/\lambda)$, taken from the results of the neutron analysis of other manganese compounds.

Using these data in conjunction with equations (7) and (8), the value of the effective spin-quantum number S (see eq. (7)) giving the best agreement with the observed intensities was now calculated for the three reflections with the greatest magnetic contribution (100, 101, 111), in respect of the two possible mutual orientations of the spins on A and B sites. It was found that in the case of antiparallel orientation the three S values thus obtained were in better agreement with one another than in the case of parallel orientation. It was inferred from this that the spins are oppositely oriented, which disposes of question (b): as has already been assumed, the coupling is ferrimagnetic. The value of S thus found was 0.97, so the magnetic moment of the Mn atoms is 1.94 Bohr magnetons, which answers question (c).

The fifth column of Table II gives the magnetic contributions to the intensities, calculated for this model; the sum of nuclear and magnetic contributions in the sixth column is now seen to be in good agreement with the measured values. Further confirmation is that the saturation magnetization calculated for this structure is in quantitative agreement with the experimental value.

Third example: the magnetic structure of Y-$Ba_2Zn_2Fe^{III}_{12}O_{22}$

Our second example of magnetic structure analysis concerns Y-$Ba_2Zn_2Fe^{III}_{12}O_{22}$ (Y denotes the type of structure). In this case a single crystal was available, without which the neutron analysis of such an intricate compound would not have been easily possible. This structure was also discussed in the above-mentioned article [1]): the unit cell is hexagonal, with a very long c axis (43.6 Å), and contains 36 iron atoms, all of which have a magnetic moment. From magnetic measurements the substance is known to be ferrimagnetic and belongs to the class described as "ferroxplana", in which the magnetic moments are oriented perpendicular to the c axis. On theoretical grounds, Gorter [14]) had already drawn up a model of the magnetic structure, which is shown in *fig. 9*.

The primary object of the neutron analysis [15]) was to verify this model. It was done by measuring the intensities of 18 reflections from lattice planes perpendicular to the c axis (003, 006, . . . , 0054; the intermediate reflexions are "forbidden" on grounds of symmetry). Since the magnetic moments are also perpendicular to the c axis, the angle a of equation (8) is always 90° for such reflections, so that the magnetic scattering makes the maximum possible contribution.

Experimentally it is a very simple matter to measure such a series of intensities, because the reflections all lie in one plane. The crystal needs to be aligned only once, with the c axis in the appropriate (horizontal) direction, after which all reflections are successively recorded with one sweep of the counter tube.

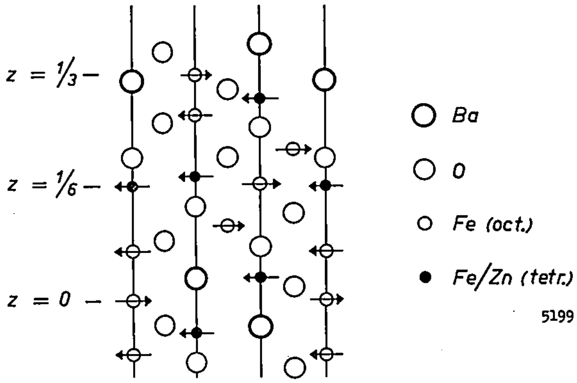


Fig. 9. The magnetic structure of Y-$Ba_2Zn_2Fe_{12}O_{22}$, as proposed by Gorter [14]) and confirmed by neutron diffraction. The figure represents a cross-section along the (110) plane; the z coordinate is expressed as a fraction of the lattice constant c. The directions of the magnetic moments of the iron atoms, which are all perpendicular to the hexagonal crystal axis, are indicated by arrows.

[14]) E. W. Gorter, Proc. Instn. Electr. Engrs. 104 B, suppl. No. 5, 255, 1957.

[15]) J. A. Goedkoop, J. Hvoslef and M. Živadinovič, Acta cryst. 12, 476, 1959.

The intensities found were then compared with the sum of the nuclear intensities (calculated from Braun's model[16]) and the magnetic intensities (from Gorter's model[14])). The results were in good agreement with the theory, so that this relatively simple measurement may be regarded as a confirmation of the postulated magnetic structure. Moreover, as regards the nuclear contribution, Braun's conclusion was again confirmed, namely that the zinc atoms together with an identical number of iron atoms are distributed at random over two different crystallographic positions. As appears from the neutron and X-ray scattering lengths of Zn and Fe, given in Table I, the neutron intensities are more sensitive in this particular point than the X-ray intensities.

As the intensity (in photons per cm^2 per second) of a given X-ray tube is still greater by several powers of ten than the neutron flux of the largest nuclear reactors, and in view of the comparative scarcity of nuclear reactors, neutron diffraction is at present used as a supplementary method for answering those questions that cannot readily, if at all, be answered by means of X-ray diffraction. Neutron analysis, then, is always preceded by X-ray analysis. In this connection, close cooperation is desirable between the research centres where nuclear reactors are available and the laboratories where the most diverse structural problems are frequently dealt with. The three examples presented in this article may serve as an illustration of such cooperation.

Appendix: Incoherent scattering

Coherent scattering, with which we have been primarily concerned, is found when the atoms at identical lattice sites in the crystal all contribute to the scattering with identical scattering lengths. *Incoherent scattering* occurs where they do not all contribute with the same scattering length. In that case it is readily seen that there is no longer complete extinction of all the scattering contributions in other than the Bragg angles. The phenomenon is observable in the scattering from an isotope mixture, for example. We shall take this example to illustrate the method of calculating the coherent and incoherent components of the scattering, after which we shall return to the case dealt with in the text (p. 75).

Suppose that on identical lattice sites there are n atoms of scattering length b_1 and m atoms of scattering length b_2. This is to a first approximation equivalent to *identical* scatterers on all the lattice sites in question, with a scattering length equal to the weighted mean:

$$b_{\rm H} = \frac{n}{n+m} b_1 + \frac{m}{n+m} b_2. \quad . \; . \; . \quad (9)$$

(The index H refers to the example on p. 75.)

These identical scatterers give rise to a diffraction pattern consisting of discrete peaks; their scattering length $b_{\rm H}$ is called the *coherent scattering length*. There still remains, however, a residual component of scattering power on each of the sites in question: on the n sites with a scattering length of:

$$b_1 - b_{\rm H} = \frac{m}{n+m}(b_1 - b_2)\,,$$

and on the m sites:

$$b_2 - b_{\rm H} = \frac{n}{n+m}(b_2 - b_1)\,.$$

These "residues", which are distributed at random over the available lattice sites, give rise to a continuous diffraction pattern. As a result of this random distribution, each "residue" can be considered independent of the other, and an effective scattering cross-section can be ascribed to each (see p. 73). We can now calculate the weighted mean of these; this is known as the *effective cross-section for incoherent scattering*:

$$\sigma_{\rm H,inc} = 4\pi \left[\frac{n}{n+m}(b_1 - b_{\rm H})^2 + \frac{m}{n+m}(b_2 - b_{\rm H})^2\right] = $$

$$= 4\pi \left[\frac{nm}{(n+m)^2}(b_1 - b_2)^2\right]. \quad . \; . \; . \quad (10)$$

We also introduce an *effective cross-section for coherent scattering*:

$$\sigma_{\rm H,coh} = 4\pi b_{\rm H}{}^2\,. \quad . \; . \; . \; . \; . \; . \; . \quad (11)$$

Finally we may define an *effective cross-section for the total scattering*, being the sum of (10) and (11):

$$\sigma_{\rm H,tot} = 4\pi \left[\frac{n}{n+m} b_1{}^2 + \frac{m}{n+m} b_2{}^2\right]. \quad . \quad (12)$$

Since only the coherent scattering can tell us anything about the structure of a substance, we can only determine the structure of those substances in which the effective cross-section for incoherent scattering is not unduly large compared with that for coherent scattering.

Let us now return to the hydrides of $\rm Th_2Al$, mentioned in the text. The incoherent neutron scattering here arises from the fact that both the neutron and the scattering proton have a spin of $\frac{1}{2}$. There are consequently two possible modes of scattering, i.e. with the spins of the particles parallel or antiparallel. The total spin of the system proton plus neutron is equal to 1 in parallel orientation and equal to 0 in antiparallel orientation, which means that the probability of the former is three times greater than of the latter (this has been explained elsewhere[3])). In the case of neutron scattering by a crystal containing hydrogen, this implies that in three-quarters of the encounters the spins of neutron and proton will be parallel and in one-quarter antiparallel.

The scattering lengths in the two cases differ considerably, and are denoted by b_1 and b_2 respectively:

$$b_1 = +1.04 \times 10^{-12} \text{ cm}; \quad b_2 = -4.7 \times 10^{-12} \text{ cm}.$$

Using equations (9), (10) and (11) we calculate that $\sigma_{\rm H,coh} = 1.8 \times 10^{-24}$ cm^2 and that $\sigma_{\rm H,inc} = 79 \times 10^{-24}$ cm^2.

The difficulty mentioned in the text is now clear: of the scattering caused by the hydrogen atoms, which is in itself very considerable, only a little more than 2% contributes to the interference peaks; the remainder produces isotropic scatter, which appears as a continuous background in the diffraction patterns. This is particularly serious where polycrystalline samples are concerned, since in their case only a small proportion of the crystallites takes part in the interference peaks, although the whole sample contributes to the

[16]) See P. B. Braun, Philips Res. Repts **12**, 491, 1957, and G. H. Jonker, H. P. J. Wijn and P. B. Braun, Philips tech. Rev. **18**, 145, 1956/57.

incoherent scattering. In a single crystal the situation is much more favourable, and hydrogen diffraction does appear in the patterns.

The effect discussed here in connection with hydrogen also occurs in cases of scattering by other nuclei that have a spin differing from zero. In vanadium, for example, coherent scattering is practically imperceptible for this reason. In deuterium, having a nuclear spin of 1, we find $\sigma_{\mathrm{D,inc}} = 2.0 \times 10^{-24}$ cm^2, $\sigma_{\mathrm{D,coh}} = 5.4 \times 10^{-24}$ cm^2, indicating that most of the scattering here is certainly capable of interference.

Summary. By collimating and selecting neutrons extracted from a nuclear reactor through a channel in the shield, a monochromatic beam is obtained (i.e. one containing neutrons of one specific velocity). A sample to be analysed is placed in this beam, and the scattering intensity in various directions is measured with a counter tube filled with BF$_3$. Methods analogous to those in X-ray diffraction are then used to draw conclusions regarding the crystal structure from the recorded diffraction pattern. Points where neutron diffraction corresponds to and differs from X-ray diffraction are discussed in this article. In general the scattering is caused by atomic nuclei, but in special cases it is due to the electrons, i.e. when the latter are carriers of a resultant magnetic moment. Neutron diffraction is particularly used for *a*) localizing hydrogen atoms; *b*) distinguishing between atoms having consecutive atomic numbers and which are therefore difficult to distinguish by X-ray diffraction; *c*) determining the orientation of magnetic moments. An example of each of these cases is discussed (the preparatory X-ray work for these examples has earlier been described in this journal). Among the results achieved with neutron diffraction has been the determination of the complicated magnetic structure of the compound Y-Ba$_2$Zn$_2$Fe$^{\mathrm{III}}_{12}$O$_{22}$, which has served to confirm the theoretical model proposed by Gorter.

PROJECTION TOPOGRAPHS OF DISLOCATIONS

by A. E. JENKINSON *).

548.4:778.33

Dislocations have become a tangible reality for the research worker since methods were devised for making them visible, either indirectly by "decoration" or directly with the aid of the electron microscope. X-ray diffraction is the basis of yet another elegant method for photographing and studying dislocations.

About three years ago, an elegant method was devised by A. R. Lang for making visible the dislocations present in thin slices of single crystals [1]). This method, which is based on X-ray diffraction, may be understood by reference to *fig. 1.*

A collimated beam of penetrating X-rays, e.g. MoKα or AgKα radiation, impinges on the crystal under study and is reflected according to the Bragg condition from an appropriate set of lattice planes, oriented preferably normal or nearly normal to the surface of the crystal slice. A suitable aperture in a screen behind the crystal allows the transmitted reflected beam to fall on a film, while the direct beam is intercepted. A diffraction image is produced on the film, representing a picture of the part of the crystal lattice traversed by the direct beam. The images of all points lying along one line in the direction of reflection coincide on the film. The topography within the crystal is thus pictured as seen in this particular projection.

If a sufficiently wide beam of almost parallel X-rays were available, a projection picture of the whole crystal could be produced with this simple set-up. Such a beam is however impracticable. Nevertheless a picture of the complete crystal can be obtained by mounting the crystal and the film on an accurately machined slide and moving them together in a certain direction, e.g. parallel to the crystal slice, while the primary X-ray beam and the screen are stationary. The Bragg condition is, of course, not affected by such a purely translational movement, and the topographical relation-

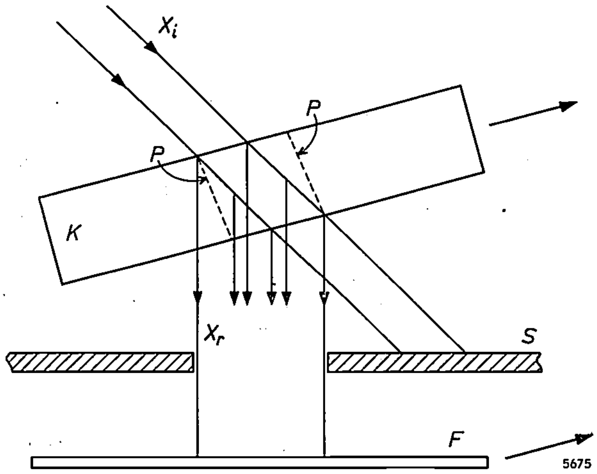


Fig. 1. Principle of projection topography as devised by A. R. Lang [1]). *K* slice of a single crystal. X_i incident X-ray beam. X_r diffracted X-rays. *P* lattice planes in Bragg reflection position. *F* film. *S* screen with aperture. Crystal and film are mounted on a slide and moved together during exposure in the direction of the heavy arrows.

*) Mullard Research Laboratories, Salfords, England.

[1]) A. R. Lang, The projection topograph: a new method in X-ray diffraction microradiography, Acta cryst. 12, 249-250, 1959.

incoherent scattering. In a single crystal the situation is much more favourable, and hydrogen diffraction does appear in the patterns.

The effect discussed here in connection with hydrogen also occurs in cases of scattering by other nuclei that have a spin differing from zero. In vanadium, for example, coherent scattering is practically imperceptible for this reason. In deuterium, having a nuclear spin of 1, we find $\sigma_{\mathrm{D,inc}} = 2.0 \times 10^{-24}$ cm², $\sigma_{\mathrm{D,coh}} = 5.4 \times 10^{-24}$ cm², indicating that most of the scattering here is certainly capable of interference.

Summary. By collimating and selecting neutrons extracted from a nuclear reactor through a channel in the shield, a monochromatic beam is obtained (i.e. one containing neutrons of one specific velocity). A sample to be analysed is placed in this beam, and the scattering intensity in various directions is measured with a counter tube filled with BF_3. Methods analogous to those in X-ray diffraction are then used to draw conclusions regarding the crystal structure from the recorded diffraction pattern. Points where neutron diffraction corresponds to and differs from X-ray diffraction are discussed in this article. In general the scattering is caused by atomic nuclei, but in special cases it is due to the electrons, i.e. when the latter are carriers of a resultant magnetic moment. Neutron diffraction is particularly used for *a*) localizing hydrogen atoms; *b*) distinguishing between atoms having consecutive atomic numbers and which are therefore difficult to distinguish by X-ray diffraction; *c*) determining the orientation of magnetic moments. An example of each of these cases is discussed (the preparatory X-ray work for these examples has earlier been described in this journal). Among the results achieved with neutron diffraction has been the determination of the complicated magnetic structure of the compound Y-$Ba_2Zn_2Fe^{III}{}_{12}O_{22}$, which has served to confirm the theoretical model proposed by Gorter.

PROJECTION TOPOGRAPHS OF DISLOCATIONS

by A. E. JENKINSON *).

548.4:778.33

Dislocations have become a tangible reality for the research worker since methods were devised for making them visible, either indirectly by "decoration" or directly with the aid of the electron microscope. X-ray diffraction is the basis of yet another elegant method for photographing and studying dislocations.

About three years ago, an elegant method was devised by A. R. Lang for making visible the dislocations present in thin slices of single crystals [1]). This method, which is based on X-ray diffraction, may be understood by reference to *fig. 1.*

A collimated beam of penetrating X-rays, e.g. MoKα or AgKα radiation, impinges on the crystal under study and is reflected according to the Bragg condition from an appropriate set of lattice planes, oriented preferably normal or nearly normal to the surface of the crystal slice. A suitable aperture in a screen behind the crystal allows the transmitted reflected beam to fall on a film, while the direct beam is intercepted. A diffraction image is produced on the film, representing a picture of the part of the crystal lattice traversed by the direct beam. The images of all points lying along one line in the direction of reflection coincide on the film. The topography within the crystal is thus pictured as seen in this particular projection.

If a sufficiently wide beam of almost parallel X-rays were available, a projection picture of the whole crystal could be produced with this simple set-up. Such a beam is however impracticable. Nevertheless a picture of the complete crystal can be obtained by mounting the crystal and the film on an accurately machined slide and moving them together in a certain direction, e.g. parallel to the crystal slice, while the primary X-ray beam and the screen are stationary. The Bragg condition is, of course, not affected by such a purely translational movement, and the topographical relation-

Fig. 1. Principle of projection topography as devised by A. R. Lang [1]). *K* slice of a single crystal. X_i incident X-ray beam. X_r diffracted X-rays. *P* lattice planes in Bragg reflection position. *F* film. *S* screen with aperture. Crystal and film are mounted on a slide and moved together during exposure in the direction of the heavy arrows.

*) Mullard Research Laboratories, Salfords, England.

[1]) A. R. Lang, The projection topograph: a new method in X-ray diffraction microradiography, Acta cryst. 12, 249-250, 1959.

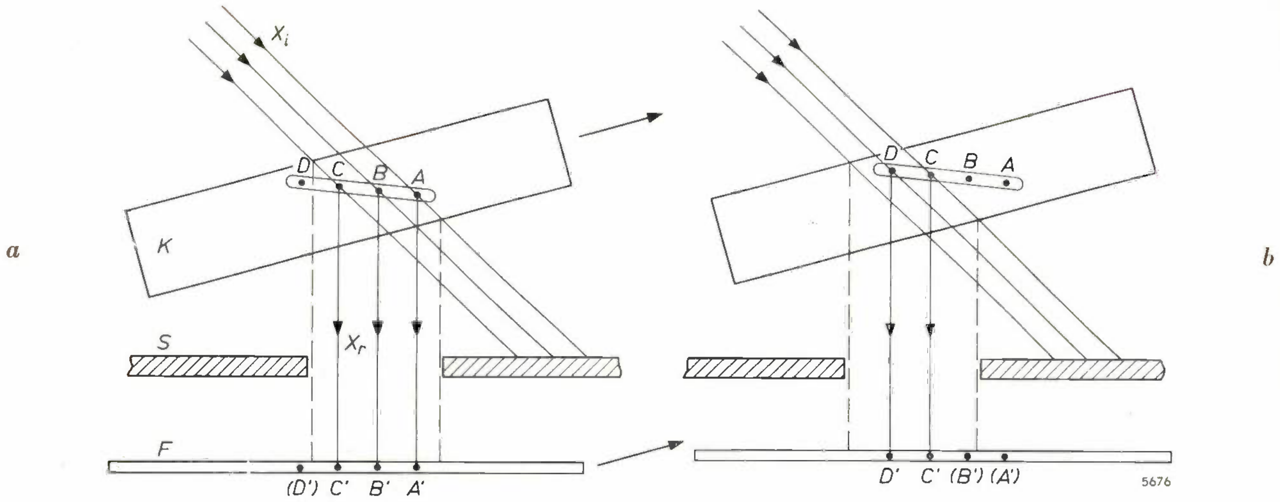


Fig. 2. Topographical details *ABCD* of an imperfection within the crystal are projected on to the film along the direction of X-ray reflection. In successive positions (*a*) and (*b*) of the crystal, each point within it, e.g. *C*, is always imaged in the same point of the film.

ship between crystal and film is obviously preserved in the overlapping diffraction images produced in successive positions (*fig. 2*). The entire crystal can be "scanned" in this way, much as a television picture is scanned (except that in the present case a ribbon-shaped scanning beam is used — see below — and a one-dimensional sweep is therefore sufficient).

An ideal crystal, having equal "reflectivity" in all parts, will give rise to a rather monotonous "picture", viz, a uniform blackening of the film. Any imperfections that disturb the lattice periodicity, however, such as dislocations, low-angle grain boundaries, etc., are found to reflect *more* than the perfect lattice surrounding them and therefore to produce an image of enhanced density on the film [2]). Thus a dislocation, which in this context can be regarded as a tubular region of relatively imperfect crystal set in a matrix of perfect crystal, will produce a line on the film representing the projection of the dislocation along the diffracted beam, as shown in fig. 2.

In *fig. 3* a "projection topograph" obtained in this way is reproduced. It shows part of a slice

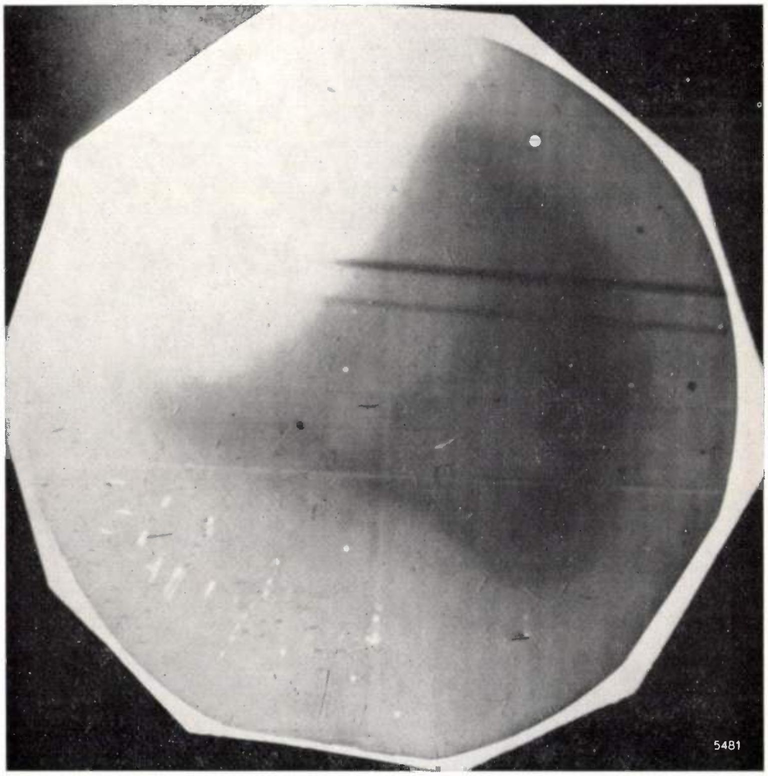


Fig. 3. Projection topograph of a germanium crystal slice about 0.1 mm thick. The total exposure time was 60 hours.

Part of the scanned crystal has produced no picture because the very thin slice had been bent in the process of mounting. The two dark bands crossing the picture are due to a spurious reflection. Other diffuse periodic lines in the picture occur because of some regular mechanical characteristic of the traverse unit. The central dark area represents a variation in the crystal thickness which is characteristic of an enhanced etching rate often found in dislocation-free Ge crystals (see A. G. Tweet, J. appl. Phys. **30**, 2002-2010, 1959). The "fringes" visible around this region and at the periphery of the crystal are due to a type of interference phenomena occurring at wedge-shaped specimens ("Pendellösung", see N. Kato and A. R. Lang, Acta cryst. **12**, 787-794, 1959).

[2]) This effect, which is not yet fully understood theoretically, is only shown by *thin* crystal slices, the criterion being $\mu d \leq 1$, where μ is the absorption coefficient and d the thickness of the slice. With a large crystal, e.g. $\mu d \approx 10$, the effect is replaced by the earlier known phenomenon of anomalous transmission (see e.g. L. P. Hunter, J. appl. Phys. **30**, 874, 1959), which entails a *decrease* in intensity of both the direct beam and the reflected beam as a consequence of imperfections.

about 0.1 mm thick of a germanium crystal rather free of dislocations [3]). The total exposure time was 60 hours.

Part of the scanned portion of the crystal slice has produced no picture. This is because the crystal has been bent in the mounting to such a degree that the relevant lattice plane was here no longer in the Bragg reflecting position. A bend with a radius of curvature of *1 mile* would be sufficient to produce this effect.

In actual practice, the crystal and film are made to perform many traverses during one exposure, instead of just one single scan. The reason is that the exposure is more easily controlled by varying the number of traverses than by changing the rate of scanning. Moreover, possible fluctuations of the X-ray intensity are rendered harmless in this way. For the very long exposure of 60 hours for fig. 3, the scanning beam made 80 traverses about 1″ long.

The apparatus for taking the projection topographs, together with the scanning mechanism, is shown in *fig. 4*.

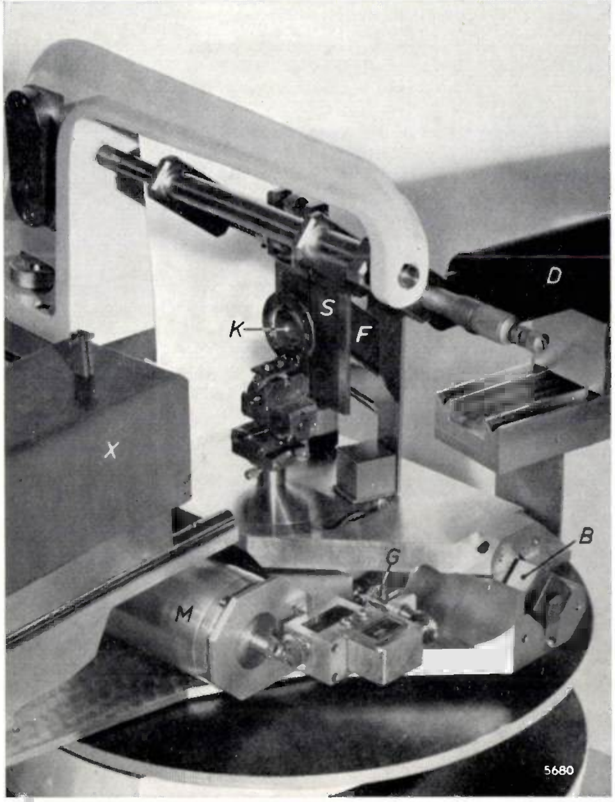


Fig. 4. Apparatus as devised by Lang [1]) for making X-ray-diffraction projection topographs. *M* electric motor for scanning. *G* gears driving micrometer screw for scanning. *B* accurately machined guide bar on which the slide carrying crystal slice and film moves. *X* collimating slit. *D* detector for setting apparatus. Other letters as in fig. 1.

The motor *M* is a synchro (magslip) driven from a control unit (not shown) in which the range of the to-and-fro movement can be pre-set.

A few more remarks on the scanning method should be made. The divergence of the X-ray beam in a plane perpendicular to that of fig. 1, i.e. to the plane of the incident and diffracted beams (transverse divergence) will not affect the resolution of the picture as long as the effective focus width is small enough [4]). A transverse divergence of 3° is used in practice, allowing a crystal 1″ wide to be scanned. The divergence *in* the plane of fig. 1 (longitudinal divergence), on the contrary, is strictly limited. The criterion is that the angular spread of the incident rays should be less than the difference in Bragg angles of the Kα_1 and Kα_2 components of the X-radiation used. In this way the α_1 component alone is selected to produce the topograph, and superimposition of images due to the α_1 and α_2 components is avoided. On the other hand, the longitudinal divergence should be large enough to cover completely the natural spread of reflection due to the imperfections under study. A suitable value is $1\frac{1}{2}$′.

The primary beam at the crystal is thus ribbon-shaped and has a cross-section of 1″ × 0.005″. It is obtained by suitable collimating slits [5]).

In regard to the scanning direction, it may be noted that a motion parallel to the film instead of parallel to the crystal slice would also be possible. In fact this would allow the smallest possible distance between crystal and film. Such a motion, however, would entail a displacement of the reflected beam as the crystal traverses the fixed primary beam, and it would therefore necessitate a relatively wide screen aperture, whereas in practice this aperture should be as small as possible in order to reduce the adverse effect of scattered radiation. The larger crystal-film distance necessary with the motion as shown in fig. 1 causes only a very slight increase of the geometric unsharpness.

With silicon the dislocation density is usually much higher than in the germanium crystal studied in fig. 3. See *fig. 5*, a topograph of a $\frac{1}{2}$-mm slice of a silicon crystal prepared by the "floating-zone" technique [6]). As may be seen from this figure, even in a thin slice of a silicon crystal the number of dislocations revealed by a projection topograph may be so large that it is difficult to visualize the actual situation of dislocations in the crowded environment and to interpret their distribution correctly. A further refinement of the method — also introduced by Lang — provides for this by producing *stereo* pictures. After one projection topograph has been completed, the crystal slice is rotated through twice

[3]) The crystal was provided by B. Okkerse of Philips Research Laboratories, Eindhoven. See: B. Okkerse, A method of growing dislocation-free germanium crystals, Philips tech. Rev. **21**, 340-345, 1959/60.

[4]) A microfocus X-ray generator such as has been developed for X-ray microradiography, with effective focus dimensions of about 100 × 100μ is very suitable (e.g., the Hilger microfocus unit with the line focus viewed almost end-on).

[5]) On the geometric considerations, see also A. R. Lang, Acta metall. **5**, 358, 1957, especially pp. 360/361.

[6]) See e.g. J. Goorissen, Philips tech. Rev. **21**, 185, 1959/60.

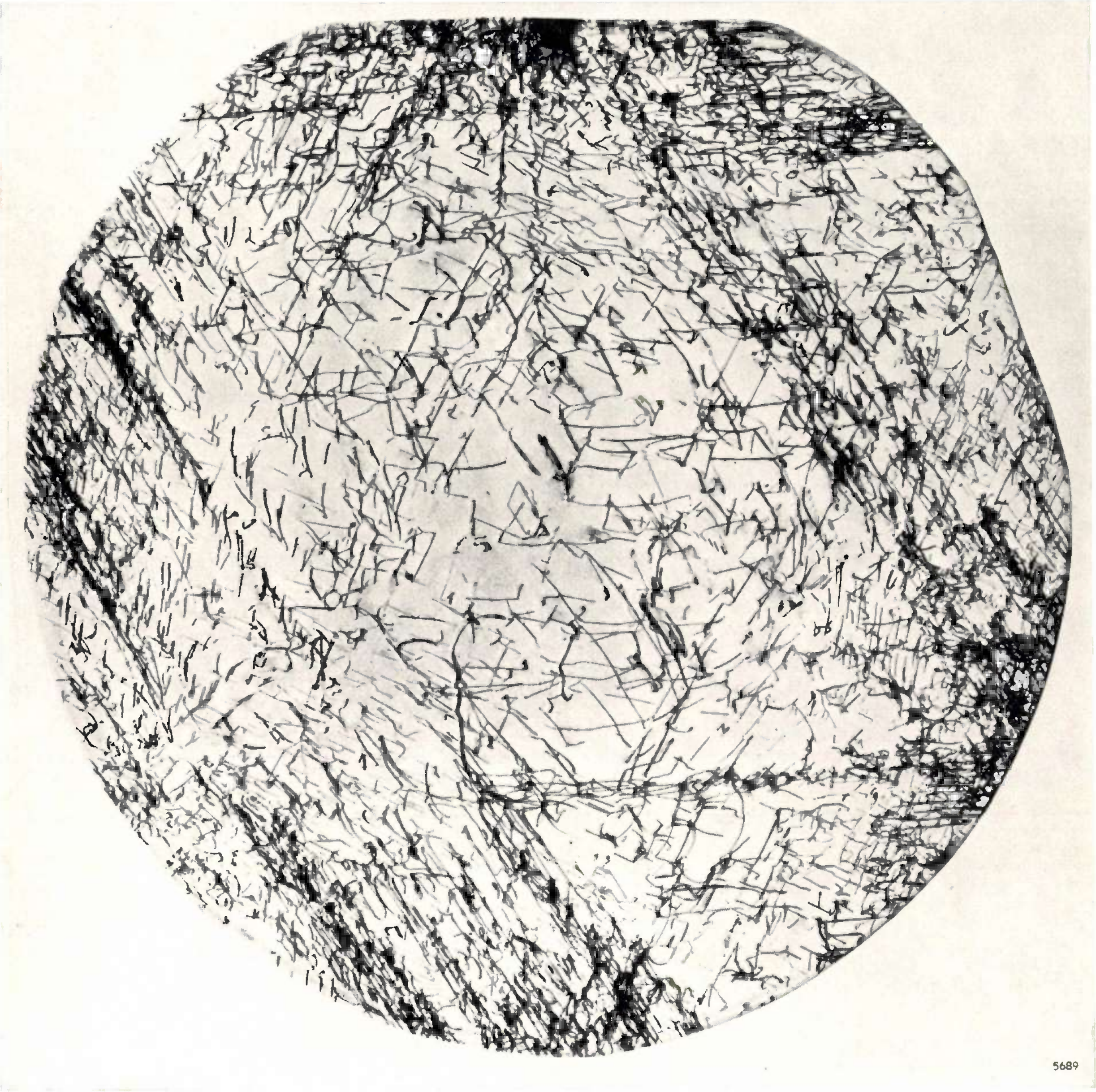


Fig. 5. Projection topograph of a ½-mm slice of a single crystal of silicon, cut from a single crystal prepared by the "floating-zone" technique. Total exposure time 20 hours. Reflection from the (02$\bar{2}$) plane was used, with Bragg angle $\Theta \approx 16°$.

the Bragg angle, as shown in *fig. 6*, and a second projection topograph is made in this position. (This amounts to using the Bragg reflection of index ($\bar{h}\,\bar{k}\,\bar{l}$) if in the first topograph the ($h\ k\ l$) reflection was used.) This will result in a different projection on

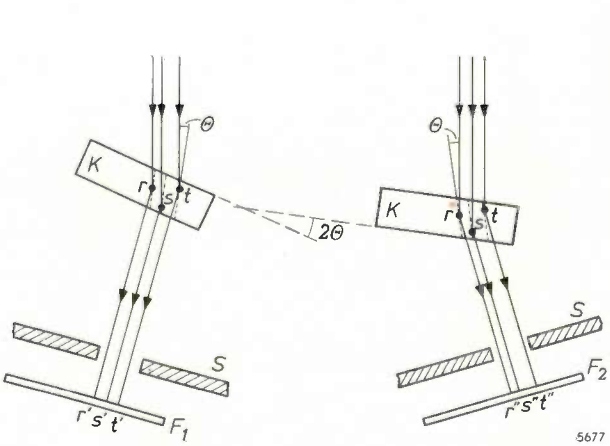


Fig. 6. A stereo pair of pictures F_1-F_2 is obtained by turning the crystal slice through twice the Bragg angle Θ as shown. Topographical details *rst* are pictured in points *r's't'* of F_1 and in points *r"s"t"* of F_2.

Other methods of rotation for obtaining stereo pairs are possible, but they would involve remounting of the crystal, which is a tedious procedure, especially because of the danger of bending (cf. fig. 3).

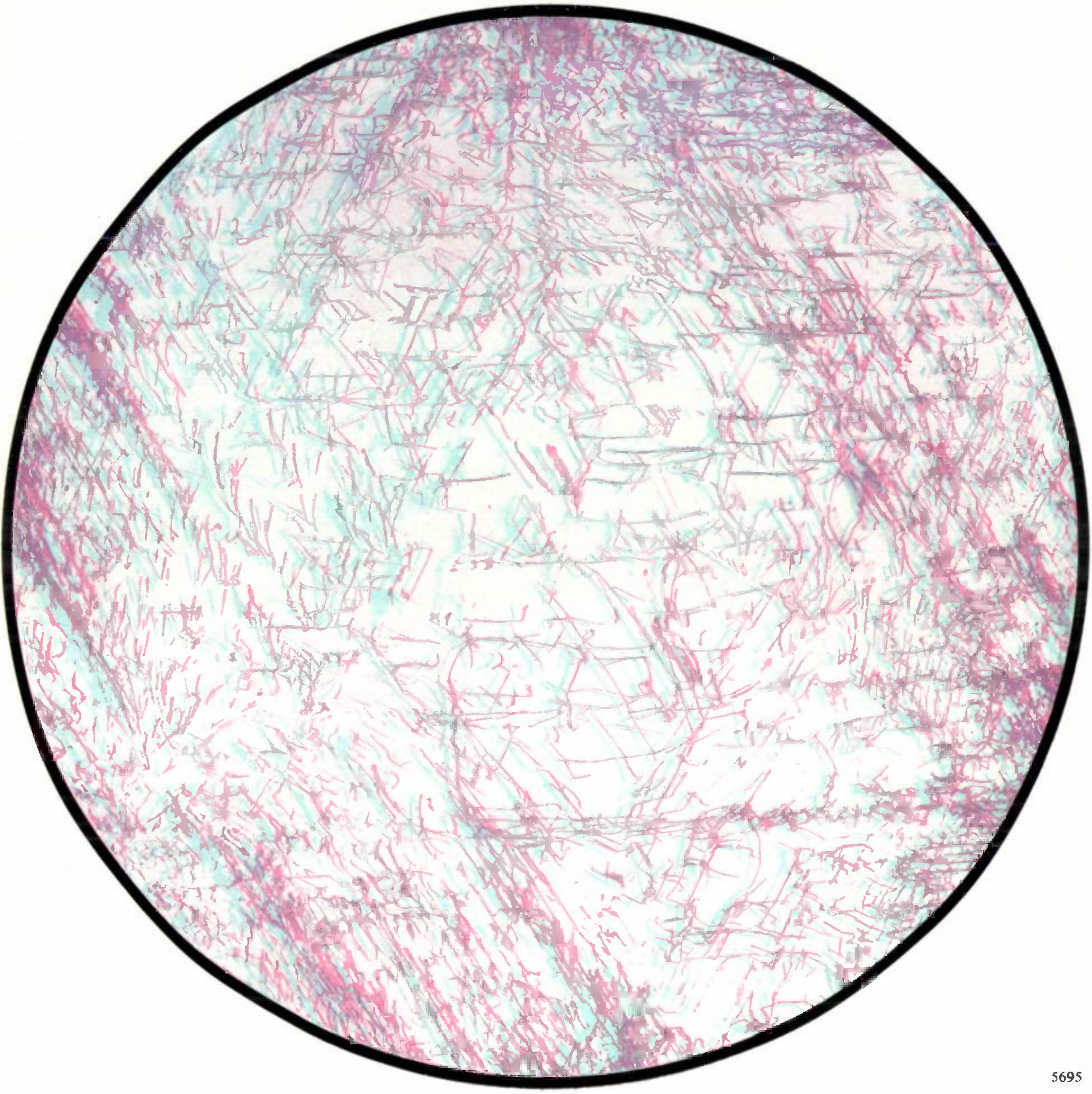
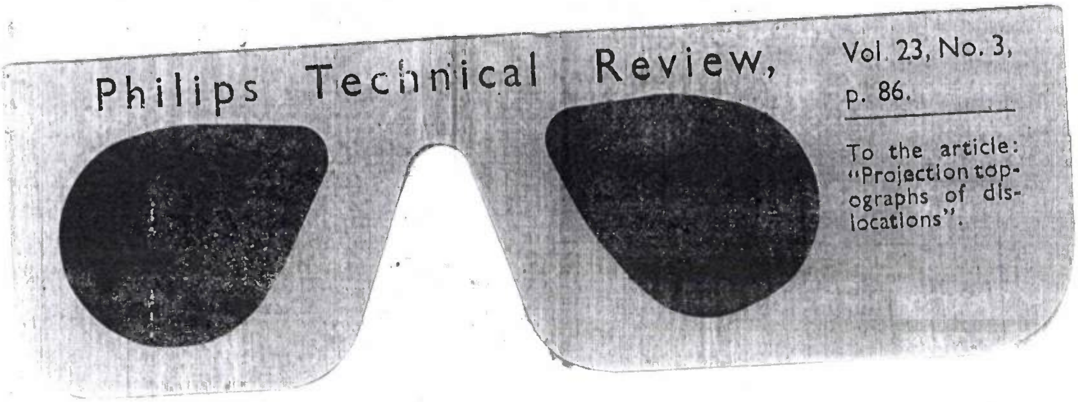


Fig. 7. Anaglyph of a stereo pair of projection topographs of silicon (cf. fig. 5). This figure should be viewed through the red-blue spectacles provided (red on the left). Most of the dislocations visible here as short pointed lines extending through the whole thickness of the slice are "60° dislocations" produced by plastic deformation. Many of the long lines lying in a plane parallel to the photograph, i.e. in what was at one time a solid-liquid interface, are screw dislocations. In this portion several examples of dislocation climb (spirals) and dislocation interaction (step-lines) are evident.



Philips Technical Review,

Vol. 23, No. 3, p. 86.

To the article: "Projection topographs of dislocations".

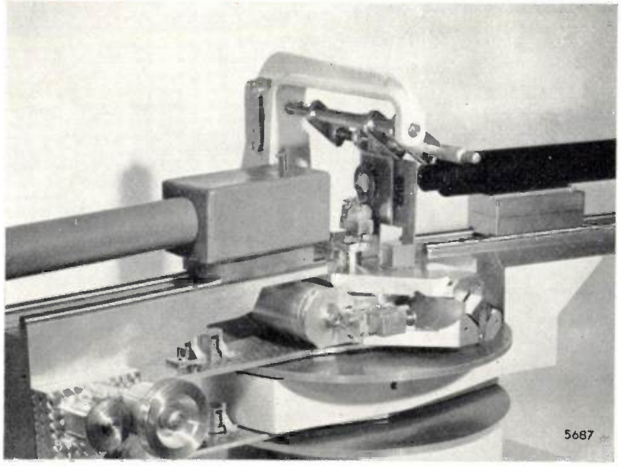


Fig. 4 *A*

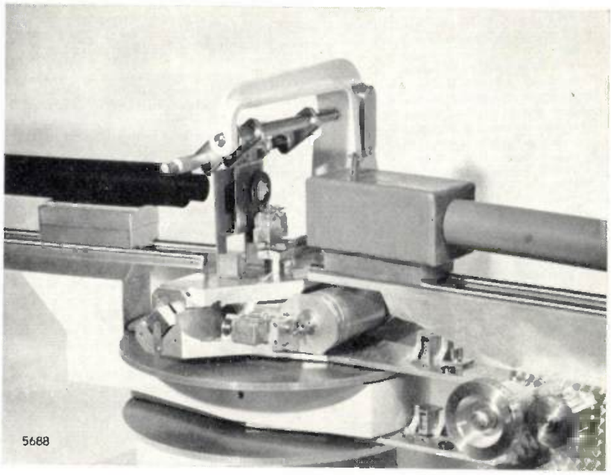


Fig. 4 *B*

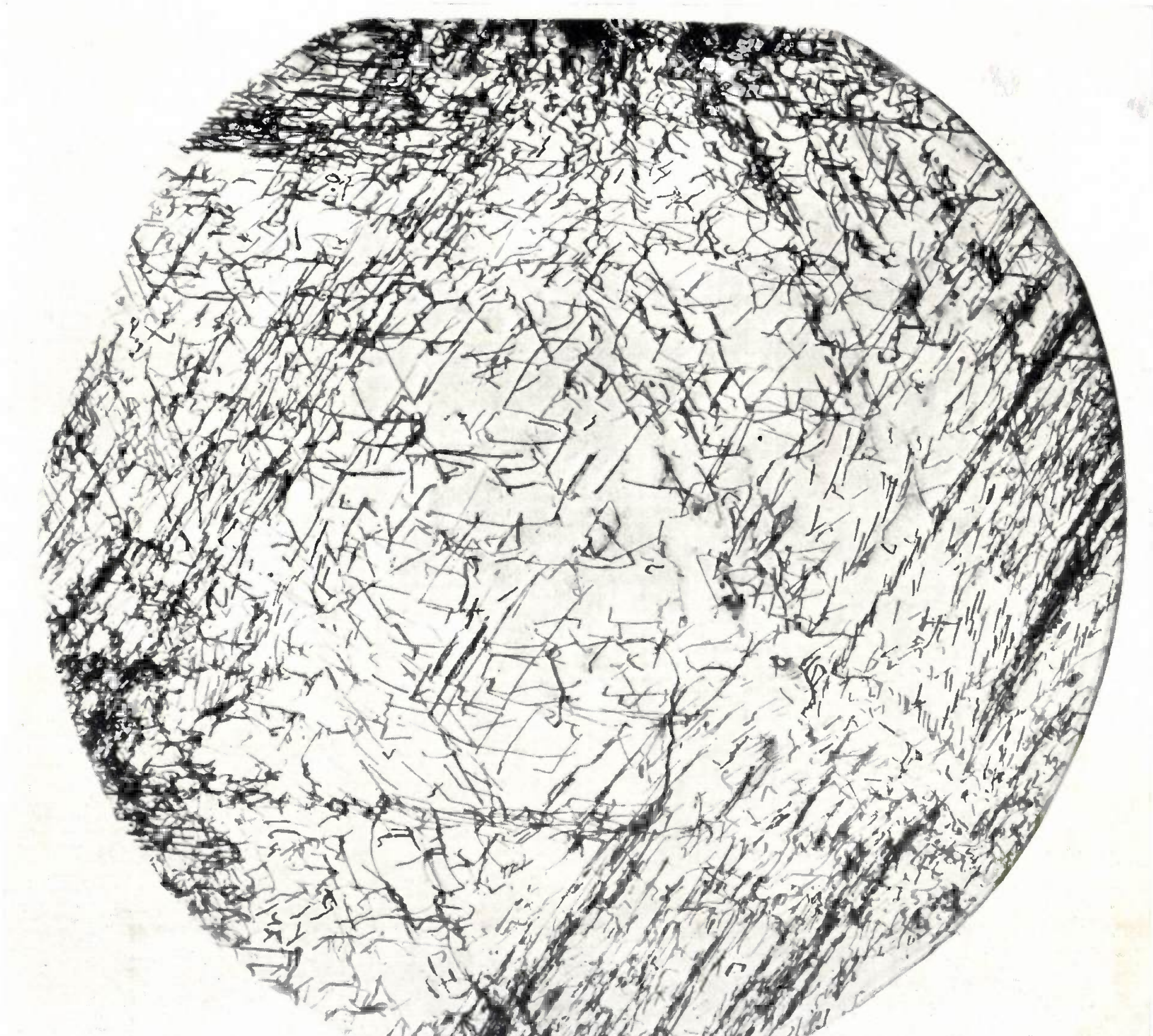


Fig. 5 *A*

to the film of the topographical details within the bulk of the crystal slice, as explained by fig. 6. The difference in projection is analogous to the difference obtained in visual observation when viewing one object from two slightly different angles. Hence, the two projection topographs form a stereo pair, which after suitable enlargement may be studied with the aid of a standard stereo viewer.

Fig. 7 shows an anaglyph of such a stereo pair, consisting of the topograph already shown in fig. 5, in which the Bragg reflection from the (02$\bar{2}$) plane was used, and the corresponding topograph with reflection from the (0$\bar{2}$2) plane. The dislocations within the crystal stand out clearly when this figure is viewed through the red-blue spectacles supplied.

A little known but quite useful method of obtaining the stereoscopic effect without a stereo viewer or an anaglyph makes use of a simple mirror [7]), as shown in *fig. 8*. One picture of the stereo pair must be reproduced as its mirror image for this method to be applicable. The second topograph corresponding to fig. 5 is reproduced in this way on the accompanying loose leaf (fig. 5*A*). This method, like the anaglyph, has the advantage over the stereo viewer that the viewing distance need not be approximately equal to the distance between the eyes, so that relatively large pictures can also be observed bit by bit.

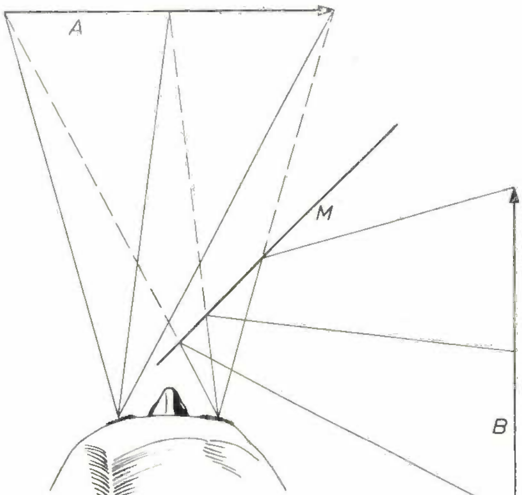


Fig. 8. Stereoscopic viewing with the aid of a simple mirror [7]). *A* and *B* are the stereo pair to be viewed; *B* must be reproduced as a mirror image. *M* is the mirror (preferably a metal one, to avoid double images). One "looks" with both eyes at one of the pictures (in this case *A*), but in fact the right eye receives the mirror image of *B*. The two images which are observed can be made to fuse into one by suitable adjustment of *B* and/or *M*.

The reader who wants to try this method of viewing will probably find it useful to practice with a photograph of a normal object first. Fig. 4 is therefore also reproduced as a stereo pair (fig. 4*A*, *B*) on the loose leaf.

Additional information on the dislocations in a crystal slice may be obtained by comparing several projection topographs of the same region, produced by reflection from *different lattice planes*. With anisotropic distortions of the lattice such as screw and edge dislocations, the mechanism giving rise to an enhanced reflection intensity and rendering the imperfection visible will not be equally effective for all directions of the primary and reflected beam.

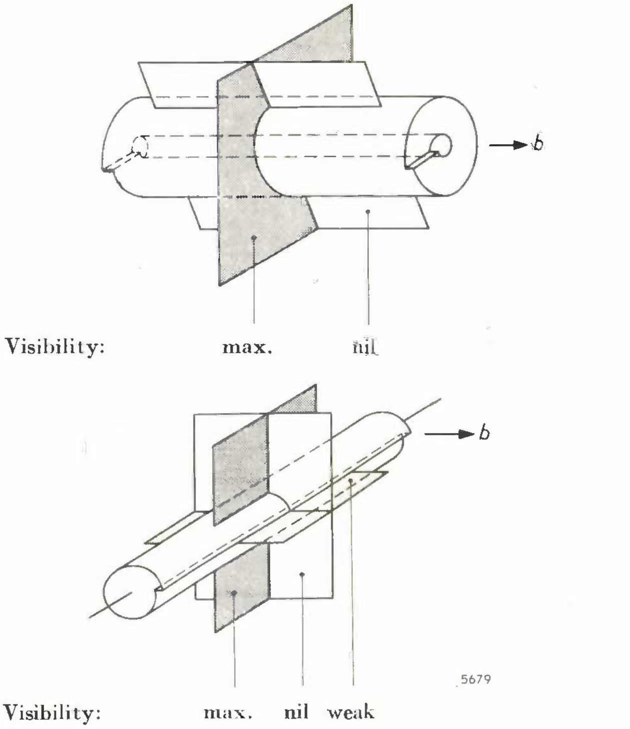


Fig. 9. Illustration of differing visibility of a dislocation in a projection topograph according to the orientation of its Burgers vector *b* with respect to the reflecting lattice plane.
Above: Screw dislocation. Reflection against the lattice plane indicated in grey will result in maximum visibility; reflection against any lattice plane oriented perpendicular to this grey plane will result in zero visibility of this dislocation.
Below: Edge dislocation. Reflection against the dark grey lattice plane yields maximum visibility; the dislocation will be weakly visible on reflection against the light grey lattice plane and it will remain invisible on reflection against the third lattice plane drawn here.

For example, the strain field around a screw dislocation is such that one would expect no X-ray image from it when its Burgers vector lies in the reflecting plane; maximum visibility would be expected when the Burgers vector is normal to the reflecting plane. Similar, although slightly more complicated, considerations apply to the edge dislocation. The result of these considerations is illustrated in *fig. 9*.

[7]) This method was devised by D. Brewster in 1849 (Trans. Roy. Scott. Soc. Arts 3, 247, 1851). See also M. von Rohr, Die binokularen Instrumente, Springer, Berlin 1920, 2nd Edn., p. 62.

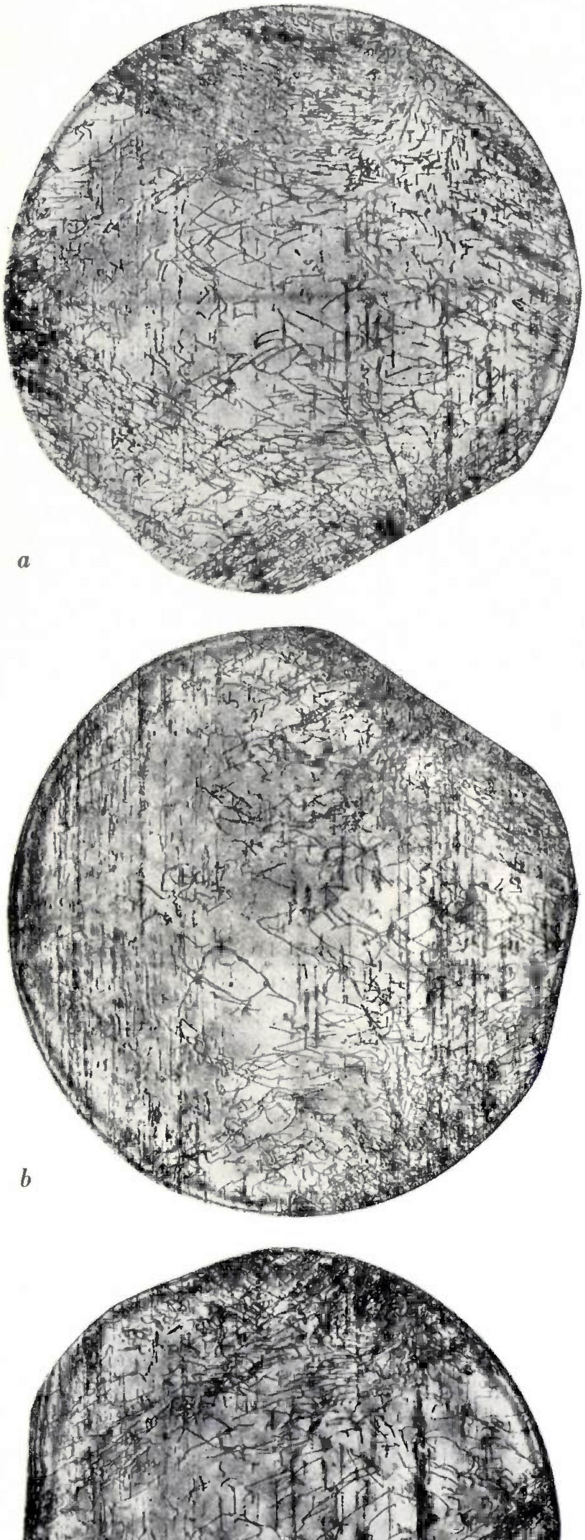


Fig. 10. Three projection topographs of the same slice as in fig. 5, made by reflection from *a*) the $(11\bar{1})$ plane, *b*) the $(1\bar{1}1)$ plane, *c*) the $(\bar{1}11)$ plane. The photographs illustrate the varying visibility of individual dislocations when different reflections are used.

The high density of dislocations at the boundary of the slice suggests the existence of dislocation sources in or near the crystal surface. Because of the relatively high density it is difficult to identify them and also to distinguish "grown-in" dislocations.

In *fig. 10* are shown three different projection topographs of the slice of silicon used for fig. 5: the lattice planes used were the $(11\bar{1})$, $(1\bar{1}1)$ and $(\bar{1}11)$ octahedral planes, the plane of the slice being (111). Now in silicon, the Burgers vectors lie in ⟨110⟩ directions. Hence, in a given {111} reflection, a dislocation with any one of the three possible ⟨110⟩ Burgers vectors (either positive or negative) *not* lying in the {111} plane is strongly visible, whereas a dislocation with any one of the three other ⟨110⟩ Burgers vectors which *do* lie in the {111} plane is invisible or almost so. It can therefore be concluded that a dislocation which is invisible in any *two* of the pictures of fig. 10 will have as its Burgers vector the direction of intersection of the two reflecting planes used. Thus, dislocations invisible in figs 10*a* and 10*b* but visible in fig. 10*c* have a Burgers vector in the direction [011]. Dislocations invisible in only *one* of the three topographs have a Burgers vector which is parallel to the direction of intersection of that particular reflecting plane and the (111) plane of the slice.

Summary. The method of projection topography of thin slices of single crystals by means of diffracted X-rays, as devised by Lang, is explained and a number of topographs are shown. The distribution of dislocations within a crystal is beautifully revealed by stereo pairs of projection topographs, and topographs of one slice made by diffraction from different lattice planes allow the identification of each dislocation according to the orientation of its Burgers vector.

OPTICAL MEASUREMENT OF RECORDED VELOCITIES ON STEREOPHONIC TEST RECORDS

by C. R. BASTIAANS *) and J. van der STEEN *).

53.082.531:681.854

Some thirty years ago it was shown that a simple relation exists between the peak lateral velocity of the cutting head in disc recording — the "peak stylus velocity" — and the width of the reflected bands of light which are observed when a beam of light is directed on to the walls of the groove in a test record. This relation offers a simple means of calibrating test records of which the peak stylus velocity must be accurately known. The results are not sufficiently reliable, however, for stereophonic disc recording. In the article below the authors derive the condition which must be fulfilled if accurate results are also to be obtained with stereodiscs, and describe an optical system of measurement incorporating the principles discussed.

In the gramophone industry, wide use is made of test records, i.e. gramophone discs on which tones of accurately known pitch have been recorded. They are also used in such electro-acoustical work as investigating the characteristics of cutting heads, pick-ups and loudspeakers.

For measurement purposes it is necessary to know precisely either the amplitude of the groove excursions on these records, or the "peak stylus velocity" (also known as the peak modulation velocity), i.e. the maximum lateral velocity of the cutter tip during the recording, which is proportional to the above-mentioned amplitude.

In 1930 Buchmann and Meyer described a simple method of measuring the stylus velocity by optical means [1]). They showed that the reflected bands of light observed when a parallel beam of light is directed on to the walls of the groove have a width b which is proportional to the peak stylus velocity $\hat{v}$, and is independent of the groove radius and the frequency of the recorded signal:

$$v = \pi n b , \quad . \; . \; . \; . \; . \; . \; . \quad (1)$$

where n is the rotational speed of the record in revolutions per second during the cutting process. If $\hat{v}$ is constant, which it frequently is on test records, the reflection is observed as two diametrically opposite bands of light having a constant width b (in *fig. 1* denoted by b_i and b_o). By measuring b, the value of $\hat{v}$ can be found from (1).

Using more rigorous proofs than were given by Buchmann and Meyer, we shall presently derive a more general formula, valid for both stereophonic and monophonic records [2]). First of all, we shall briefly recount the various methods of modulation used for disc recording.

Methods of modulation

The essential difference between an ordinary monophonic gramophone record and a stereophonic record is that two signals are recorded on the latter — the "left-hand" and "right-hand" signals. In early experiments attempts were made to provide each signal with its own sound track, two grooves being engraved either on one side of the disc [3]) or on two sides. This involved using two cutting heads and two pick-ups, and accurate synchronization proved to be an intractable problem. Moreover, it had the fundamental disadvantage of reducing the playing time of a stereodisc to half that of a monophonic disc of equal size.

In the modern stereodisc the two signals are impressed in the same groove. The cutting stylus is coupled to two driving systems, which in principle operate independently. To avoid interaction between the systems, the two directions of movement must be perpendicular to one another. The most obvious possibilities are then:

1) One signal is recorded *laterally* in the groove — as in a conventional monophonic disc — and the other *vertically*, i.e. on the "hill-and-dale" principle common half a century ago. This is known as the "0/90" method. The directions in which the wall of the groove moves in lateral and vertical modulation can be seen in *fig. 2a* and *b*.

*) Philips Phonographic Industries, Baarn, the Netherlands.

[1]) G. Buchmann and E. Meyer, Eine neue optische Messmethode für Grammophonplatten, Elektr. Nachr.-Techn. 7, 147-152, 1930. A shortened form of this article has appeared in English: J. Acoust. Soc. Amer. 12, 303-306, 1940.

[2]) We are indebted for this derivation to H. de Lang of Philips Research Laboratories, Eindhoven.

[3]) K. de Boer, Experiments with stereophonic records, Philips tech. Rev. 5, 182-186, 1940.

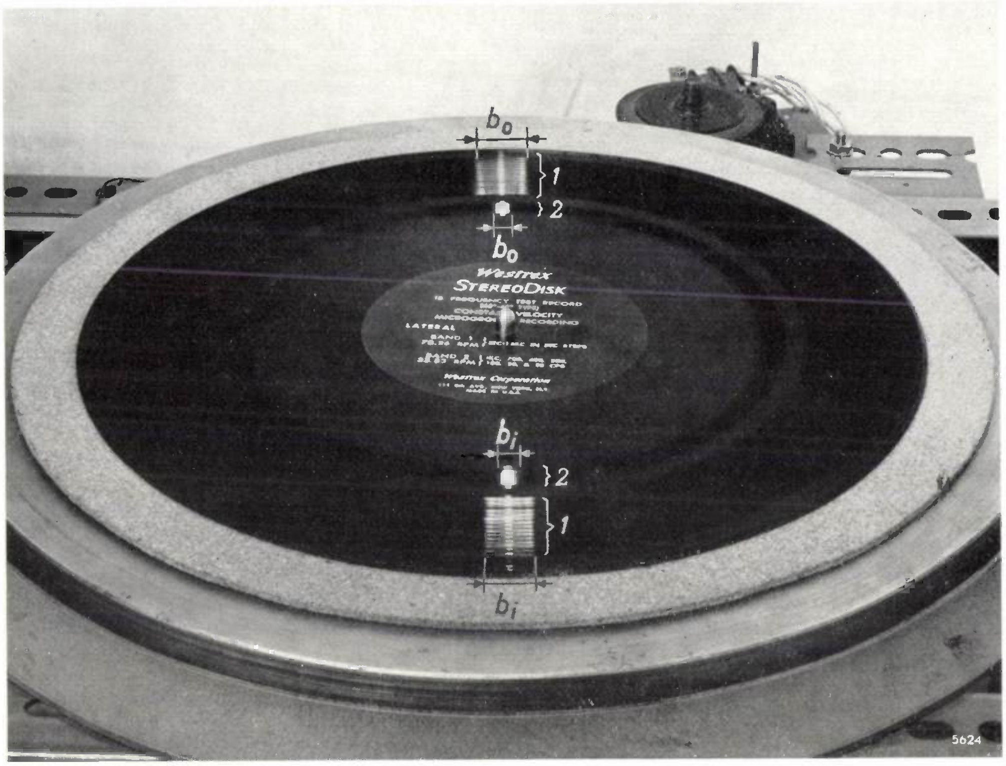


Fig. 1. On this test record a series of sine waves of different frequencies are recorded, in the turns *1* at 78.26 r.p.m. and in the turns *2* at 33.33 r.p.m. The peak stylus velocity is constant in each set of turns, but is lower in *2* than in *1*. Two diametrically opposite bands of light can be seen on each set of turns. In agreement with equation (1), the width of these bands is constant. (The difference between the width b_i on the front half and the width b_o on the rear half of the record is negligible to a first approximation.) For further particulars of this record see fig. 10.

2) Both signals are recorded symmetrically, the two cutting angles being 45° with respect to the surface of the disc. This is known as "45/45" or *45° modulation*. If only one of the two signals is present, then only one of the two groove walls is modulated (fig. 2*c*).

One advantage of the 45/45 method is that laterally it gives a summation of the two signals, so that the record can be played monophonically with a normal pick-up responsive to lateral movement only [4]). Nowadays, therefore, all stereophonic records are produced by the 45/45 method. Some test records, however, are modulated by the 0/90 system for the purpose of investigating certain characteristics of stereophonic pick-ups for 45°-modulated discs.

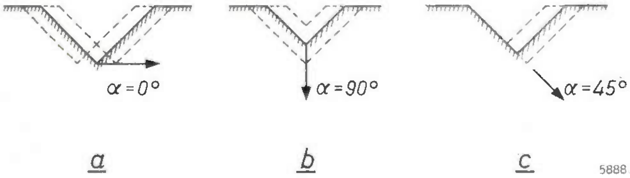


Fig. 2. Displacement of the walls of a groove (seen in axial cross-section) when *a*) laterally modulated, *b*) vertically modulated, and *c*) 45°-modulated. The arrow indicates the cutting direction of the stylus point. The angle made by the arrow with the plane of the disc is the cutting angle (0°, 90° and 45°, respectively).

In both methods it is possible in principle to keep the two signal channels entirely distinct. In practice, however, some degree of cross-talk is inevitable. Cross-talk causes distortion of the stereophonic sound picture and depends on the geometry of the cutting stylus and pick-up [5]). With the light-pattern system of measurement it is possible to determine not only the stylus velocity but also, in principle at least, the degree of cross-talk in so far as it is due to the cutting head and is thus "in the record", and not "in the pick-up".

Derivation of the relation between the light-band width, the peak stylus velocity and various other parameters

The relation between the observed width b of the band of light and the peak stylus velocity $\hat{v}$, which

[4]) The moving system of this (monophonic) pick-up must possess adequate compliance in the vertical direction to avoid damage to the vertical groove-modulation in the stereodisc.

[5]) J. L. Ooms and C. R. Bastiaans, Some thoughts on geometric conditions in the cutting and playing of stereodiscs and their influence on the final sound picture, J. Audio Engng. Soc. 7, 115-121, 1959.

is the maximum value of the sinusoidally varying stylus velocity v, depends in general on two angular parameters: the cutting angle α (see fig. 2) and the angle of incidence of the light β, which is at the same time the angle of observation. Both angles are measured with respect to tbe plane of the disc [6]).

For the sake of convenience we introduce a third parameter, i.e. the angle φ in *fig. 3*, which can be expressed directly in terms of b:

$$\sin \varphi = b/2R, \quad . \; . \; . \; . \; . \; . \quad (2)$$

where R is the radius of the modulated groove (we shall for the moment consider only one turn of the groove). 2φ is the angle subtended at the centre M by the outermost reflecting points A_1 and A_2 in this turn.

For simplicity we assume that geometrical optics are applicable to this case, which enables us to disregard diffraction effects. We can further simplify the problem by imagining the groove to be cut not in a rotating disc by a stylus moving only in a lateral direction, but in a stationary disc by a stylus which not only has a lateral motion but also rotates about the centre of the disc (at a rate of n revolutions per second). In the absence of modulation the stylus then has simply a "groove velocity" V, given by

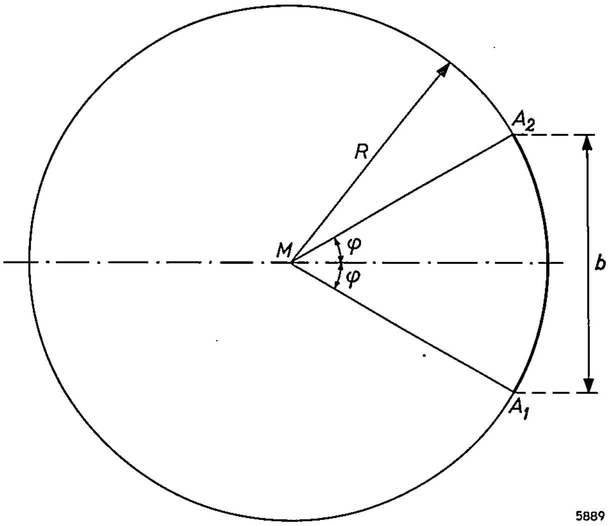


Fig. 3. In the turn of a groove of radius R only the thickly drawn portion A_1A_2 contains points that reflect light in the direction of the observer. A_1 and A_2 are the outermost points of reflection, b is the width of the band of light observed on successive turns.

$$V = 2\pi R\, n. \quad . \; . \; . \; . \; . \; . \quad (3)$$

The light incident at an angle β will be reflected in the same direction only by certain points in the groove, i.e. only by those points where the normal makes an angle β with the plane of the disc. *Fig. 4* shows a cross-section of the groove perpendicular to the resultant motion of the stylus point (the resultant motion being the vector sum of the velocities V and v). In this cross-section the bottom of the groove has a set of normals lying in the plane of the drawing with directions bounded by the normals n_1 and n_2. The condition for reflection in the direction β is not only that β should lie between n_1 and n_2 but also that it should be in the plane perpendicular to the direction of the resultant motion of the stylus.

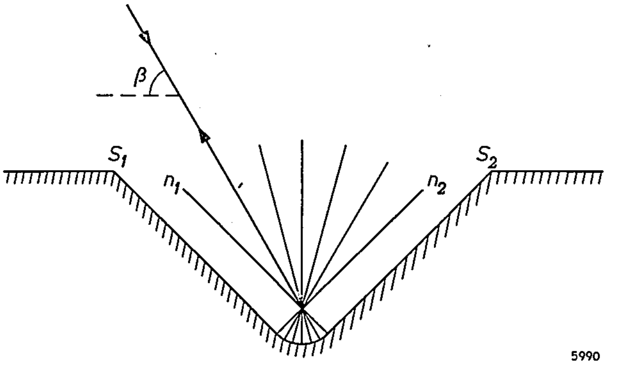


Fig. 4. Cross-section of a groove in a plane perpendicular to the resultant motion of the stylus point. The bottom of the groove has a set of normals limited by n_1 and n_2. Light will be reflected in the viewing direction β along one of these normals provided the plane of the cross-section is parallel to that direction.

The farther the points that reflect light in the required direction are removed from the vertical plane of symmetry of the observed band of light, the larger is the angle which the resultant stylus motion makes with the linear motion at the position of these points at the bottom of the groove (*fig. 5*). The outermost points that still reflect light in the direction β, A_1 and A_2, are therefore those where this angle is maximum, i.e. where the tangent of

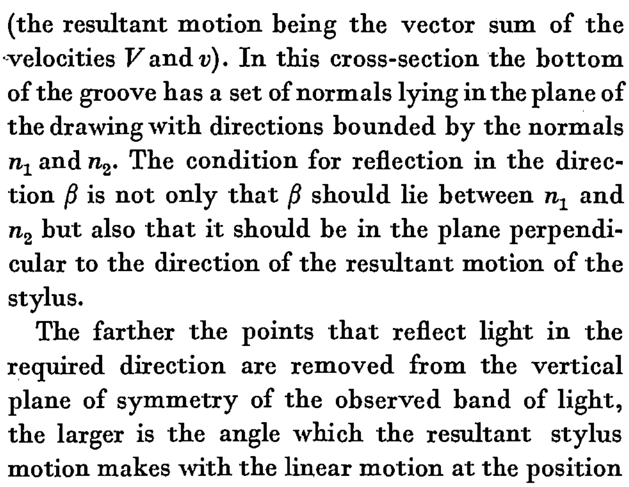


Fig. 5. The undulating line represents the bottom of the groove (seen from above), the dots are the points that reflect light in the direction of observation. The farther these points are from the plane of symmetry, the steeper is the portion of the undulating line on which they are located. At the outermost points of reflection, A_1 and A_2, the undulating line has its maximum slope in relation to the unmodulated groove. To the left of A_1 and the right of A_2 the observer sees no reflections.

[6]) The direction of incidence need not coincide with the direction of observation. Where they are different, the theory given here still holds if β is understood to be the angle which the bisectrix of the angle between the two directions makes with the plane of the disc. We shall confine ourselves here to the case where the two directions do coincide, as they do in the measuring equipment to be described below.

the angle is equal to $\hat{v}/V$. From this we can calculate $\hat{v}$ if we know the maximum angle as a function of the parameters α and β.

If conditions are such that the bottom of the groove is not visible, a band of light is still observable, of lower intensity but of the same breadth. This reflection is attributable to the fine tracks left in the walls of the groove by unavoidable irregularities in the cutting edges of the stylus. For such very fine tracks we cannot of course apply considerations of geometrical optics as we have done for the bottom of the groove. However, further analysis, taking diffraction into account, yields exactly the same result.

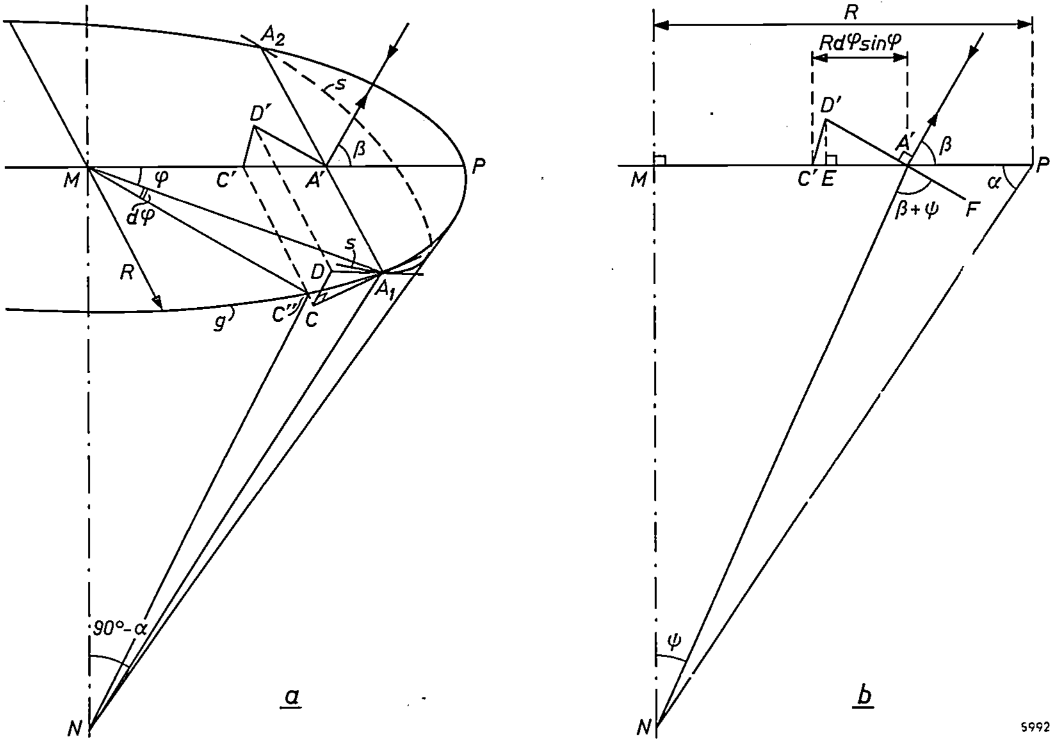


Fig. 6. *a*) The cone along which each point on the stylus cutting edges moves when cutting a groove. M centre of the disc. g turn of groove, of radius $R = MP = MA_1$. N apex of the cone (half angle $90°—\alpha$). A_1 and A_2 outermost points that reflect light in the direction of observation β. The plane through A_1A_2 perpendicular to the direction β cuts the cone in the curve s. The tangent of the maximum angle between the resultant stylus velocity $V + \hat{v}$ and the "groove velocity" V is CD/A_1C.
b) Triangle $A'C'D'$ is the projection of the triangle A_1CD in (*a*) on to the plane MNP; ψ is the projection of $\angle MNA_1$ ($= 90°—\alpha$) on to the same plane. $\angle NA'F = \beta + \psi$.

Another limitation of the application of geometrical optics is that the "wavelength" and the amplitude of the modulation in the disc are not always sufficiently large compared with the wavelengths of the light. At a frequency of 1000 c/s, and at the usual amplitudes on test records, this gives no trouble, but at higher audio frequencies it calls for more detailed analysis [7]).

An entirely different role is played in principle by the upper edges of the groove (the lines where the groove walls intersect the plane of the disc; S_1 and S_2 in fig. 4). The upper edges do not follow the motion of the point of the stylus, but describe a curved path in the plane of the disc with an amplitude that depends on the directions of the cutting edges of the stylus as well as on the motion of its point. It is not yet known how far reflections from the upper edges are of practical significance.

We shall now calculate the tangent of the maximum angle between the resultant motion of the stylus point and the unmodulated motion (this tangent, multiplied by V, gives the peak stylus velocity $\hat{v}$).

During a recording, every point on the cutting edges of the stylus moves along the surface of a cone of half angle $90°—\alpha$ (see *fig. 6a*). According to the considerations given above, at the outermost points of reflection, A_1 and A_2, the resultant motion (the vector sum of V and $\hat{v}$) is in the plane perpendicular to the direction β, and the stylus velocity $\hat{v}$ is directed along the generator of the cone. It follows that the resultant motion of the point of the stylus is tangential to the curve s in which the plane perpendicular to the direction β cuts the cone. The angle A_1MP is equal to the angle φ in fig. 3.

Our object is to find the tangent of the angle between $V + \hat{v}$ and V, i.e. the tangent of $\angle CA_1D$. This is equal to CD/A_1C, where the angle at C is a right angle. As C tends to A_1, CA_1 tends to arc $C''A_1 = R\,d\varphi$, whilst simple geometrical considera-

[7]) B. B. Bauer, Calibration of test records by interference patterns, J. Acoust. Soc. Amer. 27, 586-594, 1955.

tions (given below in small print) show the value of CD to be:

$$CD = R\,\mathrm{d}\varphi \sin\varphi \frac{\cos\beta\cos\psi}{\sin\alpha\sin(\beta+\psi)}, \qquad (4)$$

where the angle $\psi = \angle MNA'$ is the projection of $\angle MNA_1$ ($= 90°-\alpha$) on the plane MNP (fig. 6*b*).

We project the triangle A_1CD on to the plane MNP. This gives the triangle $A'C'D'$ (fig. 6*b*), where $A'C' = R\,\mathrm{d}\varphi\sin\varphi$. As C and D in fig. 6*a* approach A_1, the side $A'D'$ tends to:

$$A'D' = R\,\mathrm{d}\varphi\sin\varphi\frac{\cos\psi}{\sin(\beta+\psi)}$$

and the perpendicular $D'E$ to:

$$D'E = R\,\mathrm{d}\varphi\sin\varphi\frac{\cos\psi\cos\beta}{\sin(\beta+\psi)}.$$

From this we arrive at (4), since $CD = D'E/\sin\alpha$.

The ratio CD/A_1C is thus equal to

$$\frac{\cos\beta\sin\varphi\cos\psi}{\sin\alpha(\sin\beta\cos\psi+\cos\beta\sin\psi)} = \frac{\cos\beta\sin\varphi}{\sin\alpha(\sin\beta+\cos\beta\tan\psi)}.$$

But $$\tan\psi = \frac{MA'}{MN} = \frac{\cos\varphi}{\tan\alpha},$$

so $$\frac{CD}{A_1C} = \frac{\cos\beta\sin\varphi}{\sin\alpha\sin\beta + \cos\alpha\cos\beta\cos\varphi}.$$

By equating this with $\hat{v}/V = \hat{v}/2\pi Rn$, we find:

$$\hat{v} = \pi n \times 2R\sin\varphi\frac{\cos\beta}{\sin\alpha\sin\beta + \cos\alpha\cos\beta\cos\varphi},$$

or, taking eq. (2) into consideration:

$$\hat{v} = \pi n\, b\frac{\cos\beta}{\sin\alpha\sin\beta + \cos\alpha\cos\beta\cos\varphi}. \qquad (5)$$

We could also eliminate the $\cos\varphi$ in the denominator by use of eq. (2), but this would lead to a complicated equation. The result is clearer if we make use of the fact that φ is always a relatively small angle, so that to the first approximation $\cos\varphi = 1$. We may therefore write:

$$\hat{v} \approx \pi n\, b\frac{\cos\beta}{\cos(\alpha-\beta)}, \quad \ldots \quad (5\mathrm{a})$$

whilst in the general case the following series expansion applies:

$$\hat{v} = \pi n b\frac{\cos\beta}{\cos(\alpha-\beta)}\left[1+\frac{1}{8(1+\tan\alpha\tan\beta)}\left(\frac{b}{R}\right)^2+\ldots\right], \quad \ldots \quad (5\mathrm{b})$$

or, conversely:

$$b = B\left|\frac{\cos(\alpha-\beta)}{\cos\beta}\right|\left[1-\frac{\cos\alpha\cos(\alpha-\beta)}{8\cos\beta}\left(\frac{B}{R}\right)^2+\ldots\right], \qquad (6)$$

where $B = \hat{v}/\pi n$.

It is seen from (6) that b does not depend on R to a first approximation. Thus, provided the peak stylus velocity $\hat{v}$ (and hence B) is everywhere the same, successive turns of the groove each give two reflecting arcs having the same width b, and these arcs join together to form the bands of light visible in fig. 1.

Correction for the finite distance of the observer

Unless they are converged by optical means, the reflected parallel rays of light presuppose an infinitely distant observer. To an observer at a finite distance from the disc the band of light on the one half of the disc appears to be broader than on the other half. Let these breadths be b_o and b_i respectively (see fig. 1), then, as Bauer has shown [8]), b must be replaced by $2b_ob_i/(b_o+b_i)$. In the following we shall tacitly assume that this correction, where necessary, has been applied.

Special cases

It is now a simple matter to assign special values to the cutting angle α, and to investigate the manner in which the relation between b and $\hat{v}$ depends on β.

1) *Lateral modulation.* For $\alpha = 0$, equations (5b) and (6) become:

$$\hat{v} = \pi n b\left[1+\frac{1}{8}\left(\frac{b}{R}\right)^2+\ldots\right]$$

and $$b = B\left[1-\frac{1}{8}\left(\frac{B}{R}\right)^2+\ldots\right]. \quad \ldots \quad (7)$$

In this case, then, there is no dependence on β whatsoever; where monophonic discs are concerned the angle of incidence therefore has no effect on the result. In practice the correction terms $b^2/8R^2$ and $B^2/8R^2$ are nearly always negligible compared with unity. Disregarding these terms, we arrive at the original equation (1) of Buchmann and Meyer.

2) *Vertical modulation.* For $\alpha = 90°$ it follows from (5b) and (6) that:

$$\hat{v} = \pi n b \cot\beta$$

and $$b = B|\tan\beta|. \quad \ldots\ldots \quad (8)$$

In this case, then, there is a marked dependence on β. The correction term is zero.

3) *45° modulation.* For $\alpha = 45°$, equations (5b) and (6) become:

$$\hat{v} = \pi n b\frac{\sqrt{2}}{1+\tan\beta}\left[1+\frac{1}{8(1+\tan\beta)}\left(\frac{b}{R}\right)^2+\ldots\right]$$

and $$b = B\frac{|1+\tan\beta|}{\sqrt{2}}\left[1-\frac{1+\tan\beta}{16}\left(\frac{B}{R}\right)^2+\ldots\right]. \qquad (9)$$

Here again, b and $\hat{v}$ depend on β, however for positive values of $\tan\beta$ to a lesser degree than in vertical modulation.

4) *Appropriate choice of β.* If we compare equations (7), (8) and (9), omitting the correction term, we

[8]) B. B. Bauer, Measurement of recording characteristics by means of light patterns, J. Acoust. Soc. Amer. 18, 387-395, 1946.

find that for $\beta = 45°$ the same relation holds for both vertical and lateral modulation, i.e.:

$$b = \hat{v}/\pi n, \quad \ldots\ldots \quad (1)$$

and that the expression for 45° modulation differs from (1) by only a factor $\sqrt{2}$:

$$b = \hat{v}\sqrt{2}/\pi n.$$

For this reason the apparatus presently to be described has been designed for an angle of incidence and observation of 45°.

5) *Cross-talk.* Suppose that we wish to record a sinusoidal signal by 45° modulation of only one groove wall on a test record, leaving the other wall unmodulated. If the cutting angle α is exactly 45°, and assuming $\beta = 45°$, then (9) yields:

$$b_{45} = \hat{v}\sqrt{2}/\pi n$$

(disregarding the correction term). If on the other hand we make $\beta = 135°$, then with $\alpha = 45°$ we find from (9):

$$b_{135} = \text{exactly zero.}$$

However, if the cutting angle α is not exactly 45° but $45°+\varrho$, the results are:

$$b_{45} = \frac{\hat{v}}{\pi n}\sqrt{2}\cos\varrho\left[1 - \frac{1}{8}\cos\varrho(\cos\varrho - \sin\varrho)\left(\frac{\hat{v}}{\pi n R}\right)^2 + \ldots\right]$$

and

$$b_{135} = \frac{\hat{v}}{\pi n}\sqrt{2}|\sin\varrho|\left[1 + \frac{1}{8}\sin\varrho(\cos\varrho + \sin\varrho)\left(\frac{\hat{v}}{\pi n R}\right)^2 + \ldots\right].$$

A certain amount of cross-talk now exists on the second wall, the level of which is

$$10\log\left(\frac{b_{135}}{b_{45}}\right)^2 \text{dB} \quad \ldots . \quad (10)$$

below that of the signal on the first wall. The amount ϱ by which the cutting angle differed from 45° is given by:

$$\tan\varrho = \frac{b_{135}}{b_{45}} \quad \ldots\ldots\ldots \quad (11)$$

(assuming that we may neglect the correction terms).

It is here in particular that interference may be expected from the reflections from the upper edges of the groove, as mentioned above. These reflections give a light-band width which does not satisfy the basic formula (6) and which is superimposed on b_{45} and b_{135}. Equations (10) and (11) must therefore be used with some caution.

Experimental verification of the theory

To verify the theory, the light-band width b was measured as a function of the angle of incidence β for a Philips test record DV 140 205. On this record the first series of turns is 45°-modulated on the inner wall ($\alpha = 45°$), the second series is 45°-modulated on the outer wall ($\alpha = 135°$), the third series is laterally modulated ($\alpha = 0°$) and the fourth series vertically modulated ($\alpha = 90°$). The stylus velocity is constant for a given series of turns, but for the third and fourth series it is greater by a factor of $\sqrt{2}$ than for the first and second series. In *fig. 7* the full lines represent the results of the measurement, and the broken lines the values calculated from equations (9), (8) and (7), neglecting correction terms. The agreement is seen to be most satisfactory.

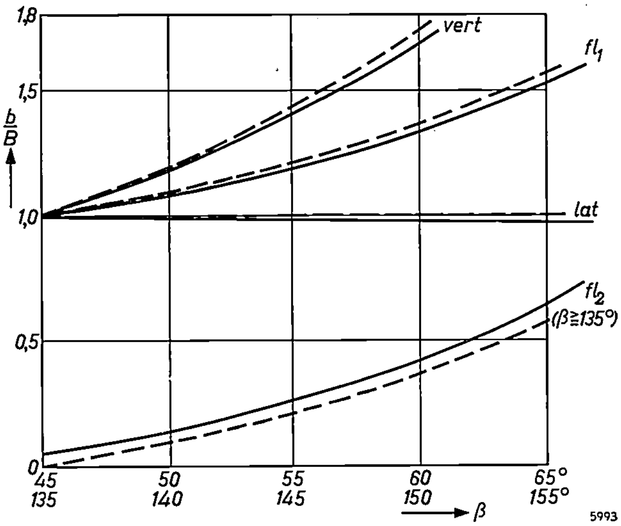


Fig. 7. Measured values (full lines) and calculated values (broken lines) of the light-band width b, divided by B, as a function of the observation angle β: *lat* for lateral modulation, *vert* for vertical modulation, fl_1 for 45° modulation (modulated groove wall), fl_2 idem (unmodulated groove wall; $\beta \geqq 135°$). The measurements were done on a Philips test record DV 140 205 (33⅓ r.p.m., recorded sinusoidal signal of frequency 1000 c/s, peak stylus velocity constant, but a factor of $\sqrt{2}$ lower for 45° modulation than for lateral and vertical modulation).

The measuring set-up

On the theoretical principles described, a measuring set-up for the calibration of test records has been constructed at Philips Phonographic Industries, Baarn. The light is incident on the record at an angle of 45°, and the reflection is observed from the same direction.

A diagram of the apparatus is shown in *fig. 8*, and a photograph in *fig. 9*. A collimator C delivers a parallel beam of light having a cross-section of 10 by 10 cm. The beam falls on the semi-reflective mirror M_1. The transmitted beam is incident on the left half of the test record G at an angle of 45° and there illuminates a rectangle measuring 10 by 14 cm which includes all the turns of the groove. The light is reflected back to the mirror M_1, which

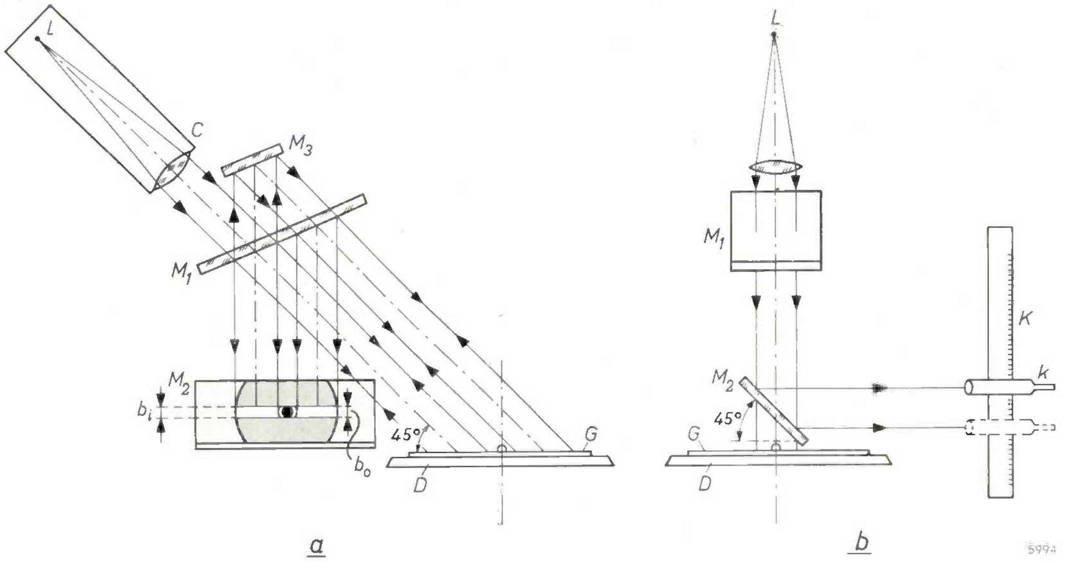


Fig. 8. Measuring set-up seen *a*) from the front, *b*) from the side. *G* record under investigation. *D* turntable. *C* collimator. M_1 semi-reflective mirror. M_2 and M_3 fully reflective mirrors. *K* cathetometer, with viewer *k*.

In the mirror M_2 two images appear side by side. The left image is from the left half of the record, the right from the right half (corresponding with the front and rear halves respectively in fig. 1). Their widths (in this arrangement their heights), b_i and b_o respectively, are measured with the cathetometer.

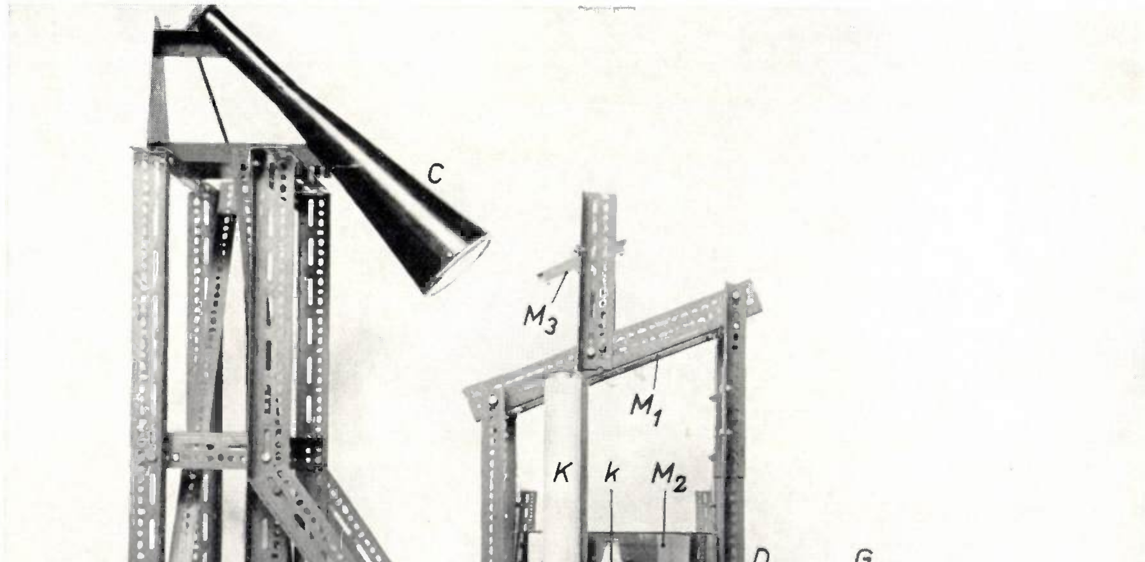
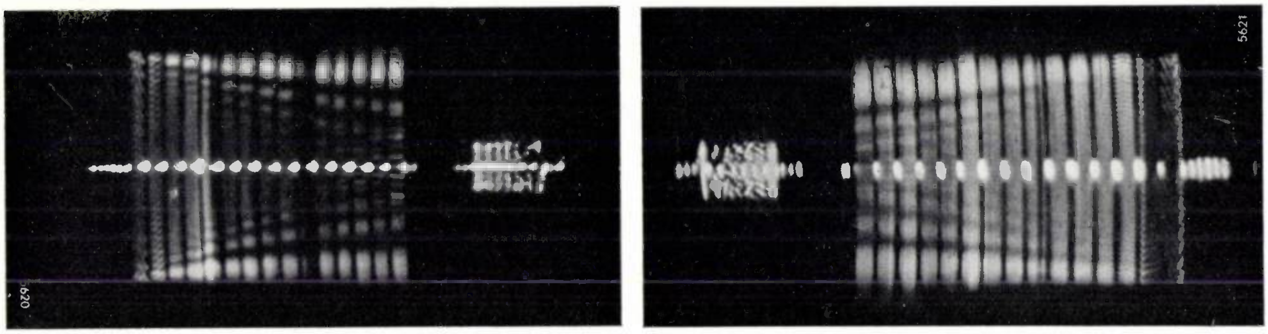
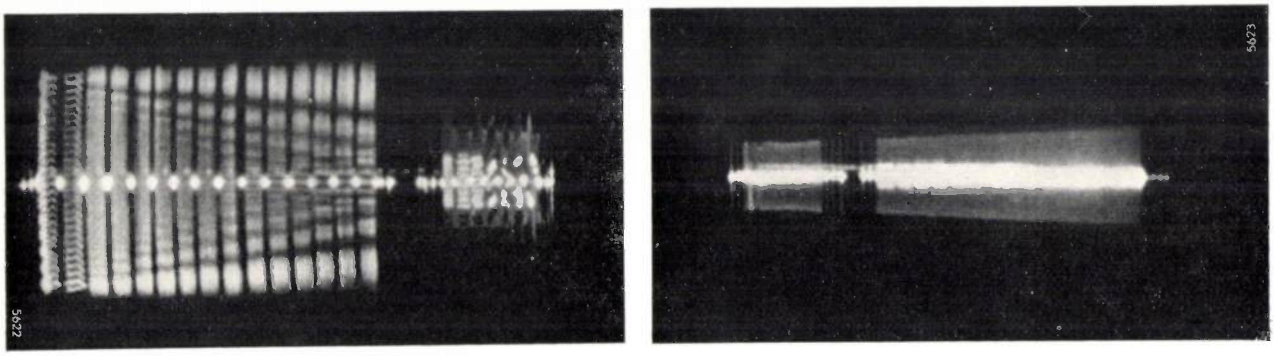


Fig. 9. Measuring set-up as in fig. 8, the letters having the same meaning. The cathetometer is 1.85 metres in front of the mirror M_2. The optical distance from the viewer to the left and right halves of the record is 2.77 and 2.85 metres respectively.



a *b*

Fig. 10. Photographs of the light patterns on the laterally modulated side of a "Westrex" test record observed in the apparatus described (the same side of this record is also to be seen in fig. 1); *a*) front (left) half, *b*) rear (right) half of record. A series of sinusoidal signals was recorded, first at 78.26 r.p.m., the pitch rising in steps of 1 kc/s from 1 to 15 kc/s, and then at 33.33 r.p.m., the pitch decreasing in steps from 1000 to 30 c/s. In both series the peak stylus velocity is seen to be reasonably constant.



a *b*

Fig. 11. As fig. 10, now with another "Westrex" test record on which the same programme is recorded, but only on the inner wall of the groove (45° modulation). Here too the peak stylus velocity is practically constant for both series of frequencies, see (*a*). It may be seen from (*b*) that there is some cross-talk on the outer wall of the groove.

reflects part of it downwards to the ordinary (fully reflective) mirror M_2 which makes an angle of 45° with the vertical. The emergent rays are thus horizontal.

The mirror M_1 directs roughly half the light from the collimator vertically upwards. This light is reflected by the ordinary mirror M_3 at an angle of 45° on to the right half of the record, where it again illuminates a rectangle of 10 by 14 cm; β must be considered here to be 135°. The light reflected within this rectangle passes out horizontally via M_1 and M_2.

Two images are thus seen side by side in the mirror M_2: on the left the image from the left half of the record, and on the right that from the right half of the record, with $\beta = 45°$ and 135° respectively. The heights of these images are the widths to be measured, b_i and b_o respectively, and they are determined (with an accuracy of up to 0.1 mm) by means of the cathetometer K, without it being necessary to shift the beam of light or the record. The cathetometer can also remain in position if a maximum error in the observation angle of $\pm 2°$ is acceptable (i.e. a maximum error in $\hat{v}$ of 0.16 dB at $\alpha = 45°$ or of 0.3 dB at $\alpha = 90°$).

Since the mirror M_1 loses roughly half its incident light upon every reflection or transmission, a fairly high luminous flux is required from the collimator and the measurement is done in a darkened room.

If the recorded tone is fairly low, the reflecting points are seen separately. It is then desirable to let the disc rotate at about 20 revolutions per minute during the measurement. For this purpose the disc lies on a turntable.

The maximum width that can be measured in this way is 10 cm. On a 78-r.p.m. record this corresponds to a peak stylus velocity of 40.8 cm/sec, and on a $16\frac{2}{3}$-r.p.m. record to 8.7 cm/sec.

Finally some photographs are shown of the patterns produced with this set-up on two "Westrex" test records. On each side of these records a series of sinusoidal signals rising in steps of 1 kc/s from 1 to 15 kc/s is recorded at 78.26 r.p.m., followed by a series of sinusoidal signals decreasing in steps from 1000 to 30 c/s, recorded at 33.33 r.p.m. On one record this programme was recorded by lateral cut on one side and by vertical cut on the other; on the second record 45° modulation was used, on one side on the inner wall of the groove and on the other on the outer wall.

Fig. 10 relates to the laterally modulated side of the first record, *fig. 11* to the side of the second record with the signal engraved on the inner wall of the groove. Figs 10*a* and 11*a* show the pattern on the left half of the record, figs 10*b* and 11*b* the pattern on the right half. It can be seen from figs 10a, 10b and 11a that the stylus velocity was reasonably constant. Fig. 11b shows a narrow band of light, indicating the presence of some cross-talk.

Summary. In 1930 Buchmann and Meyer published an optical method of measuring the peak stylus velocity with which a sinusoidal signal is recorded on a test record. They showed that this velocity is proportional to the width of the band of light observed on the record under specific conditions. This method was adequate for laterally modulated monophonic discs (cutting angle $\alpha = 0°$), but often unsatisfactory for stereodiscs ($\alpha = 45°$). The reason was found to be that the relationship between the stylus velocity $\hat{v}$ and the light-band width b is only independent of the angle β at which the record is observed (which is at the same time the angle of incidence of the light) in the case of lateral modulation. On stereodiscs and also on vertically modulated monophonic test records ($\alpha = 90°$) the relation between $\hat{v}$ and b depends on β. A general formula is derived giving $\hat{v}$ as a function of b, α and β. If β is made 45°, Buchmann and Meyer's simple relation also holds for stereodiscs (except for a factor $\sqrt{2}$) and for vertically modulated discs. The theory has been verified experimentally.

Finally a measuring set-up is described — made and used by Philips Phonographic Industries at Baarn — in which mirrors are employed to satisfy the condition $\beta = 45°$. Some photographs are shown of the patterns observed with this apparatus.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2847: R. van Strik: Uitschieters (Sigma **6**, 10-17, 1960, No. 1). (Outliers; in Dutch.)

Using the reasoning of mathematical statistics, the author deals with the problem of "outliers". The question is examined in how far it is permissible to omit from a sample of observations those values that are widely separated from the others. The conclusion is that it is permissible only when an irregularity in the method of observation can be demonstrated, or when sufficient information is available on the form of the frequency distribution and the spread of the variable. Special attention is paid to the possibility of non-normal, i.e. skew, distributions. Instead of rejecting an extreme value a transformation of the results is more appropriate in such cases, i.e. an examination of the values of e.g. log x or x^n instead of the measured values x. The argument is illustrated by examples.

2848: H. Koopman and J. Daams: 2,6-dichlorobenzonitrile: a new herbicide (Nature **186**, 89-90, 1960, No. 4718).

In the course of a screening programme, the herbicidal activity of 2,6-dichlorobenzonitrile has been investigated. This compound inhibits the germination of certain types of seeds in an agar solution in concentrations of 5×10^{-5} mg/ml, and also stunts the growth of young plants. The germination of wild oats was completely inhibited by spraying with amounts of 0.5-4 kg per hectare. Seeds of rice, ground nuts, maize and sunflower showed a distinct resistance. The substance is also active in vapour form. Its toxic effect on warm-blooded animals is low. The acute oral 50% lethal dose (LD50) for mice is greater than 6 g per kg body-weight, and the acute intraperitoneal LD50 is >3 g/kg. A full account of these investigations will be published elsewhere.

2849: N. W. H. Addink: Subnormal level of carbonic anhydrase in blood of carcinoma patients? (Nature **186**, 253, 1960, No. 4720).

In a recent paper (No. **2726**, these abstracts) the author showed that the zinc content of the blood of carcinoma patients was subnormal (a mean decrease of about 20 per cent being found). Since Zn atoms are a constituent of various metalloproteins, an investigation was made into the presence of Zn in the blood in the form of carbonic anhydrase. Analysis of blood fractions showed this to be probable. The possible effect of a subnormal level of carbonic anhydrase on the metabolism is mentioned, and the author urges an intensive study of the anti-carcinogenic activity of this substance.

2850: L. F. Defize and P. C. van der Willigen: Droplet transfer during arc welding in various shielding gases (Brit. Welding J. 7, 297-305, 1960, No. 5).

The paper reviews investigations on droplet transfer in the arc welding of steel with a consumable bare wire in a shielding gas atmosphere (CO_2 or argon). Droplet formation is studied with the aid of high-speed cinematography (3000 frames per second). In CO_2, with the wire connected to the positive pole, the wire melts obliquely and the droplets are transferred towards the side of the weld bead. This is attributed to asymmetric heating of the wire due to radiation from the pool of molten metal. The marked contraction of the plasma gives rise to noticeably coarse droplets unwanted in welding practice. This contraction is thought to be connected with the high thermal conductivity of CO_2 at arc temperature. In argon the plasma shows less contraction and no one-sided droplet transfer occurs at high currents. The fairly large anodic contact surface with argon produces finer droplets, which move axially through the arc. When the wire is connected to the negative pole, the droplets are coarse, which is attributed to the small diameter of the cathode spot. The addition of substances giving strong thermal electron emission to the surface of the wire can greatly increase the size of the cathode spot, and with argon a droplet transfer is obtained which is similar to that at the positive pole. With CO_2 it is not possible in this way to reduce the coarseness of the droplets.

2851: M. Avinor and G. Meijer: Vanadium-activated zinc and cadmium sulphide and selenide phosphors (Phys. Chem. Solids **12**, 211-215, 1960, No. 3/4).

It was known that addition of vanadium (V) kills the visible fluorescence of ZnS phosphors activated by Cu or Ag. This killing effect of V is now seen to consist of translating the emission into the infra-red, in the region of 2.0 μ, and not of dissipating the excitation energy completely into heat. The emission is also caused by V alone, but in most cases it is enhanced by Cu and Ag. It may be said that monovalent Cu and Ag behave here as co-activators of the trivalent V. The phosphors investigated were CdS, ZnS, CdSe and ZnSe, activated by V, V + Ag, V + Cu and V + Au. The spectral distributions of the emission bands were measured at 300 °K and 80 °K, a high-pressure mercury lamp with a water filter being used for excitation.

2852: G. Diemer: Power amplifiers using electro-optical effects (Electronics **33**, No. 9, 71-73, 1960).

Short review of 27 power amplifiers that can be designed to use various combinations of electric, radiative and thermal power (see also Nos. **2784** and **R 401**, these abstracts).

2853: H. Zijlstra: On magnetic annealing of a permanent magnet alloy (thesis Amsterdam, June 1, 1960).

"Ticonal" G is a magnet steel that derives its hard-magnetic properties from the presence of two finely dispersed phases. It is known that these phases occur in the form of elongated particles whose direction can be affected by a magnetic field during the heat treatment required for bringing the material into a double-phase condition.

The behaviour of this material during isothermal heat treatment in a magnetic field in the temperature range between 700 °C and 800 °C is described in this thesis.

From measurements of saturation magnetization after isothermal heat treatments it appears that the decomposition into two phases of the originally homogeneous alloy takes place very quickly and may be considered to be complete after the first few minutes.

From measurements of the magnetic anisotropy and of the mean distance between the precipitated particles while the heat treatment is going on, it appears that these quantities change their magnitude during a much longer period of time, so that they must be ascribed to changes of the shape and volume of the precipitated particles, the total quantity of each phase and its chemical composition remaining constant. The volume-diffusion process on which these changes are supposed to be based is analysed thermodynamically and related with the variations in the free energy of the fields of the magnetic dipoles about the particles and those in the free energy in the boundary plane between the phases. The diffusion equation thus obtained is applied to a model of prolate spheroids. From this, relationships are derived between the time of heat treatment and both the anisotropy and the distance between the particles, which appear to be in good agreement with the experimental results. In order to investigate the significance of this model, it is compared with a model of oblate spheroids. The behaviour of the model is found to have little sensitivity towards the type of the particles.

For comparison, surface diffusion along the boundary planes is investigated theoretically. The diffusion equations thus obtained are found not to correspond to the experimental data, so that it may be concluded that the latter process has little or no importance.

The good agreement between theory and experiments makes it possible to calculate the interfacial tension of the boundary between the phases; it appears to be very small (of the order of 1 erg cm^{-2}). The small interfacial tension is found to be of great importance for the response of a dispersion-hardening magnet steel to heat treatment in a magnetic field. It follows from the theory that the Mishima alloy, one of this group of magnet steels, which is generally believed not to become anisotropic during annealing in a magnetic field, ought certainly to exhibit this effect during heat treatment extending over a long period. This is proved by experiment. A description of the measuring methods and arrangements is given.

2854: J. van der Ster: The production of liquid nitrogen from atmospheric air using a gas refrigerating machine (thesis Delft, June 22, 1960).

The advent of the gas-refrigerating machine has made it possible to produce small amounts of liquid air in an economical way. The present thesis describes how liquid nitrogen can be produced from atmospheric air by combining the refrigerating machine with an air-rectifying column. The author discusses in particular the design and construction of an installation capable of producing more than 4 litres of liquid nitrogen an hour. The purity of the product is 99.8-99.9%; the power consumption is 1.36 kWh/l. The most conspicuous feature compared with other air-fractionating systems is that the air is not compressed. In consequence of this, the installation and its operation are extremely simple. Water vapour and carbon dioxide are removed from the feed by cooling the air in such a way that, in spite of the compact design, the installation can operate continuously for a week.

A detailed account is given of the automatic control of the installation, special attention being paid to the stability of the reflux. In the reflux-controlling system a vapour-bubble pump is used for returning the liquid nitrogen to the top of the column. An investigation of vapour-bubble pumps for liquid nitrogen and liquid oxygen is described in an appendix. Also included in an appendix are calculations and experiments on a hydrometer specially developed for simply and accurately measuring the purity of the nitrogen product.

The installation is safeguarded against explosion risks due to accumulations of hydrocarbons; the measures adopted to this end are discussed and various measurements are described.

2855: G. Meijer and R. van der Veen: Dual effect of nightbreak light (Acta bot. neerl. **9**, 220-223, 1960, No. 2).

Experiments with Salvia occidentalis, a short-day plant, showed that red nightbreak light which normally is effective in causing a long-day effect, can under certain conditions antagonize the long-day effect of a supplemental light period, thus causing a short-day effect. This short-day effect of nightbreak light disappears when the length of the nightbreak period is increased or when it is followed by an irradiation with far red (= near infra-red).

2856: E. H. Reerink, H. F. L. Schöler, P. Westerhof, A. Querido, A. A. H. Kassenaar, E. Diczfalusy and K. C. Tillinger: A new class of hormonally active steroids (Nature **186**, 168-169, 1960, No. 4719).

Preliminary communication concerning anima and clinical experiments with new synthetic compounds. Of these compounds the C-9 hydrogen atom occupies the β configuration and the methyl group at C-10 the α configuration. Therefore the compounds prepared represent the 9β, 10α analogues of the steroids of the natural series. Some of the compounds investigated showed marked progestational activity after parenteral and oral administration. No side-effects were observed. A full report will be published elsewhere.

2857: J. Verweel and B. J. M. Roovers: Magnetic properties and conduction phenomena in pure and substituted yttrium iron garnets (Solid state physics in electronics and telecommunications, Proc. int. Conf., Brussels, June 1958, edited by M. Désirant and J. L. Michiels, Vol. 3, pp. 475-487, Academic Press, London 1960).

In the substance $Y_3Fe_5O_{12}$, which has a garnet structure, ionic substitutions of divalent or tetravalent cations can be carried out which affect the resistivity. Replacement of a small amount δ of Y^{3+} by Ca^{2+} can cause some of the iron ions to become tetravalent. In the same way replacement of δFe^{3+} by δTi^{4+} or δSi^{4+} can cause δFe^{3+} to be reduced to δFe^{2+}. Samples were prepared and some physical properties, such as the resistivity as a function of temperature and the thermoelectric power, were investigated for different values of δ. The initial (complex) permeability was measured as a function of frequency and temperature, and the dispersion mechanisms are discussed.

2858: J. Smit, F. K. Lotgering and U. Enz: Anisotropy properties of hexagonal ferrimagnetic oxides (J. appl. Phys. **31**, suppl. to No. 5, 137 S-141 S, 1960).

In hexagonal crystals of the ferroxdure type the *c* axis is the preferred direction of magnetization. A magnetic field is needed to change the alignment of the magnetization, and in that case the crystal is subjected to a torque. The torque was measured as a function of the orientation of the crystal in relation to the field applied to polycrystalline material of composition $BaCo_xTi_xFe_{12-2x}O_{19}$, with aligned crystallites. It is shown that the anomalies found in the torque curves can be explained by the marked anisotropy of the cobalt ions.

2859: M. Koedam and A. Hoogendoorn: Sputtering of copper single crystals bombarded with A^+, Kr^+ and Ne^+ ions with energies ranging from 300-2000 eV (Physica **26**, 351-352, 1960, No. 5).

Earlier reported investigations into the sputtering of monocrystalline metal surfaces by ionic bombardment are being extended to higher ion energies (see also Nos. **2767** and **2840** of these abstracts).

2860: M. P. Rappoldt: Investigations on sterols, XIV. Studies on vitamin D and related compounds, XII. The photo-isomerization of lumisterol$_2$ (Rec. Trav. chim. Pays-Bas **79**, 392-400, 1960, No. 5).

Kinetic experiments on the isomerization of lumisterol$_2$ under the influence of ultraviolet light of various wavelengths are reported. The experiments indicate that neither tachysterol$_2$ nor ergosterol are primary reaction products of lumisterol$_2$. Lumisterol$_2$ is exclusively converted into pre-ergocalciferol. The quantum yield of this reaction is 0.41 at 2537 Å.

2861: M. P. Rappoldt and E. Havinga: Studies on vitamin D and related compounds, XI. Investigations on sterols, XIII. The photoisomerization of ergosterol (Rec. Trav. chim. Pays-Bas **79**, 369-381, 1960, No. 5).

Kinetic experiments on the photoisomerization of ergosterol were carried out using different sources of light. New methods of analysis for pre-ergocalciferol and tachysterol$_2$ had to be developed. The quantum yield of the photoisomerization of ergosterol with UV light of 2537 Å was found to be 0.31 at 20 °C. This value was obtained when ether, ethanol and petroleum ether were used as solvents. Tachysterol$_2$ is not formed from ergosterol during irradiation, but originates exclusively from pre-ergocalciferol. Excited ergosterol falls back to its ground state (69%) or transforms to pre-ergocalciferol (26%) and to unidentified products (5%).

2862: M. Avinor and G. Meijer: Emission of activated cadmium selenide phosphors (J. chem. Phys. **32**, 1456-1458, 1960, No. 5).

The fluorescence spectra at 80 °K and 300 °K of CdSe activated by Cu, Au and Ag, and co-activated by trivalent metals, were investigated. The phosphors were excited by radiation from a high-pressure mercury lamp passing through a $CuSO_4$ filter. For comparison, the emission spectra of CdS with the same activators and coactivators were also investigated. The emission bands of CdSe are shifted towards the infra-red in relation to those of CdS. The silver band of CdSe turned out to be completely quenched at room temperature. Besides the activator bands (Ag 0.92 μ, Cu 1.20 μ, Au 1.45 μ, at 80 °K) a near-edge emission of CdSe was found at 0.72 μ.

2863: A. A. Aldenkamp, C. P. Marks and H. Zijlstra: Frictionless recording torque magnetometer (Rev. sci. Instr. **31**, 544-546, 1960, No. 5).

An instrument for measuring magnetic anisotropy by recording magnetic torque curves is described. The special construction of the transducer which converts the torque exerted on the sample into an electrical signal makes it possible to avoid bearings, so that the instrument is essentially free of friction. The instrument is operated with commercially available electronic apparatus and is suitable for routine measurements on large numbers of samples. The maximum sensitivity is 1.5×10^{-5} Nm ($=150$ dyne cm) per centimetre deflection of the recorder stylus.

2864: F. J. Schijff: Safety and reliability in the instrumentation of nuclear reactors (Revue HF Tijdschrift 4, 223-229, 1960, No. 10).

The instrumentation of a nuclear reactor involves the use of instruments which are relatively unsafe and unreliable. Nevertheless, a reactor is required to possess a high degree of safety and great reliability, the first principally to protect life and property, the second mainly for reasons of economy. Both safety and reliability can be improved by means of coincidence circuits. Other measures of particular importance to safety are regular testing of the safety channels and constant comparison of the signals from identical channels.

VOLUME 23, 1961/62, No. 4 pp. 101-132 Published 25th January 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

THE BALANCE OF NATURE

by R. van der VEEN [*]).

632.95.024

The article below reproduces, with some minor changes, the text of the inaugural lecture delivered by the author on 6th February 1961 as professor extraordinary of botany at the University of Utrecht. The author sketches the "balance of nature" as a state of equilibrium in constant evolution, from primeval forest to single-crop farming, and from the attendant menace of agricultural pests to the present practice of combating these pests with chemical agents. He draws attention to the curious fact that plants themselves use chemical agents in an internecine struggle for existence, and develops the intriguing concept of a "negative biology" based on the action of pesticides, which promises to penetrate far into the secrets of living matter.

The text published here is supplemented, with the kind assistance of the author, by illustrations, a bibliography and an appendix of structural formulae.

Many publications addressed to the general public speak of the "balance of nature". They may often give the impression that this balance is more or less labile, so easily does it seem liable to disturbance. One might imagine that nature is comparable to a chemical balance, so that a slight shift in the balancing forces implied might be expected to "tip the scales" right over.

An equilibrium as achieved on a chemical balance is however never encountered in nature. Darwin [1]) and others have described at considerable length how the balance of nature in fact operates.

The primeval forest

Darwin was one of the first to see the tropical forest not as a community of plants in peaceful coexistence, but on the contrary as the scene of a silent but merciless struggle for food, space and, above all, light. The weapons employed in that strugglc arc extraordinarily varied. Some plants exploit their physiological aptitude for vigorous upward growth to break out rapidly above the gloom of the lower levels and so offer a broad crown to the light. Other plants are so physiologically adapted as to be able to populate these lower levels (*fig. 1*).

I mention the word physiological because, as a physiologist myself, I think in the first place of such properties as the highly economical photosynthesis required for the formation of enough assimilates at low light intensities; of growth hormones, like auxins and gibberellins, which induce exceptional elongation in many plants; and of factors which, in spite of that elongation, enable the plant to develop a sufficiently sturdy stem.

There are other plants, e.g. lianas, whose excessive elongation prevents them from acquiring the necessary sturdiness, and yet they still contrive to reach the higher levels by climbing up other plants, which may ultimately collapse under their burden.

Others again begin their lives in the light by germinating on an upper branch of a fully-grown tree. Such plants are called *epiphytes*. An epiphyte has to be extremely sparing with the mineral salts available to it, for it cannot in general draw them from the soil. If it is capable of sending out aerial roots, as some species of Ficus do, the roots develop downwards and may ultimately reach the ground and penetrate its surface; a tree is thus produced which, as it were, grows from top to bottom. The aerial roots multiply and become rapidly thicker. The epiphyte gradually proliferates around its "host", and at last destroys it. The "tree strangler", now firmly anchored by the lignification of its original aerial roots, spreads out its crown in the place once occupied by its host.

Although the innumerable seeds that fall to the

[*]) N.V. Philips-Duphar, Weesp, Netherlands.

floor of the forest may well germinate, their chance of survival is remote. The struggle is ceaseless and unrelenting, and only the best adapted and most fortunate can develop to maturity. But I need dwell no further on this fierce competition for survival in the primeval forest; it is a subject with which we have grown sufficiently familiar in the hundred years and more that have elapsed since the appearance of Darwin's brilliant work "On the origin of species".

The same fierce struggle as in the primeval forest is fought in practically all plant and animal communities. Here too the balance of nature is quite unlike a chemical succeeds in developing to maturity disturbs the existing pattern of relationships to some extent, for it intrudes itself among the other organisms, often destroying its neighbours, and in so doing changes the order of things. By the very fact of preserving itself, every living being, whether plant or animal, has a considerable influence on its environment, which would certainly have looked different without it. Thus, it can be said that the balance of nature is constantly being disturbed to some extent by each plant and each animal.

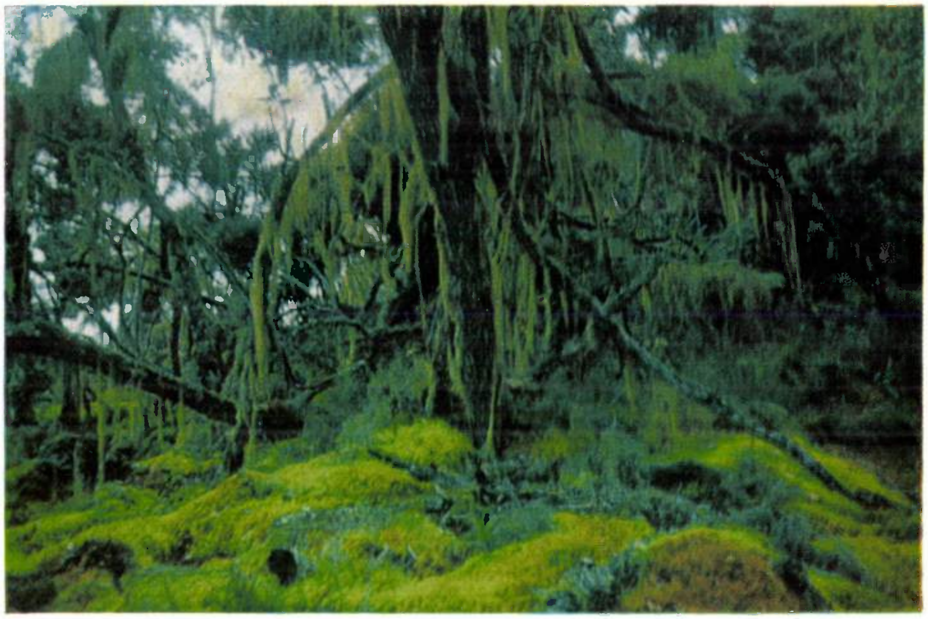


Fig. 1. Vegetation in the primeval forest: Ruwenzori massif in the Congo. The photograph was taken in a fairly open spot. The characteristic struggle in which the plants deprive each other of essential sunlight is naturally difficult to represent in a photograph.

Man and the balance of nature

Man too is a part of nature in this context, and to preserve himself is bound to shift the established balance, otherwise there would be no room for him. The early tillers of the soil must have been only balance. Every organism that too well aware of the difficulties which are involved in agriculture, in eradicating the existing plant growth and cultivating useful species. The surrounding vegetation will have constantly fought back to recover its lost ground. Only by dint of hard work, ceaseless weeding and the most intensive cultivation of the soil can they ever have succeeded in raising their plants to maturity. Not for nothing is it written in the old Dutch "Katechismus der Natuur" that God created weeds besides the useful herbs, for had He not done so, the farmers would have grown idle [2]).

Plainly, then, the agricultural population is interfering with the balance of nature in a singularly cruel fashion in felling trees and burning woodland, in ploughing and otherwise mechanically working the land. With the growth of world population this interference is assuming enormous proportions. In more densely populated regions little remains of the original flora and fauna. Even so, the best adapted species have lost nothing of their aggressiveness, and the farmer still has no chance to be idle.

The practice of agriculture requires that large acreages be devoted to thespecialized cultivation of individual crops (single-crop farming). As a logical result, adapted species amongst the other organisms, whether weeds, insects, eelworms or moulds, have been able to make large-scale invasions (*figs 2* and *3*).

This brings us to the diseases and pests of agriculture. Since almost every animal and plant has its own peculiar parasites, and every agricultural plague

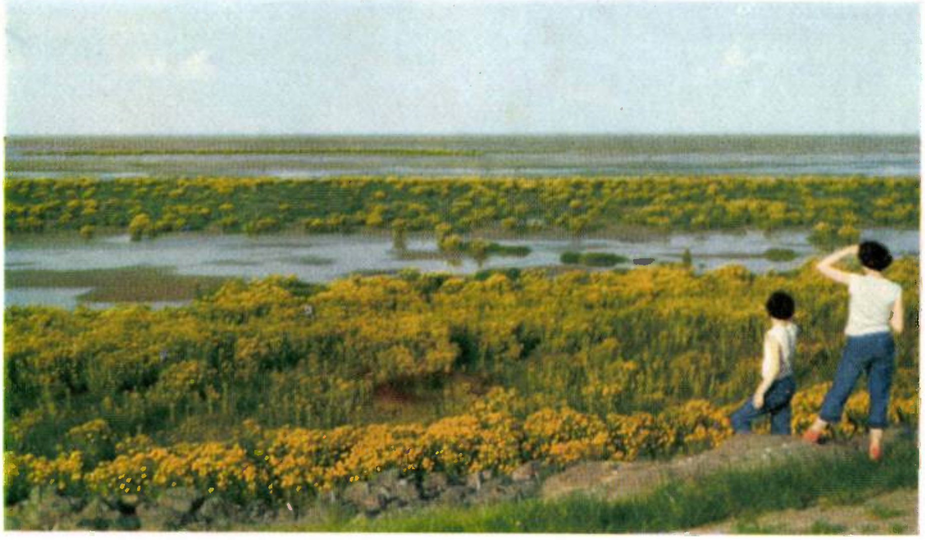


Fig. 2. |Land reclaimed from inland seas may often be invaded by adapted weeds. For example, during the drainage of the 140 000-acre Oost-Flevoland polder, formerly a part of the Zuijder Zee, broad expanses were conquered by "marsh endive" (Senecio paluster).

is caused by an animal or plant, pests which flourish in a given cultivated area are always followed by their own parasites.

The demands made on agriculture are steadily growing, and the damage caused by diseases and pests are felt more and more keenly. It is therefore not surprising that the agriculturist must take increasingly drastic measures to derive full benefit from his labour.

People have been so thoroughly accustomed to modern methods of farming, and to the periodic recurrence of pests and diseases, that they regard this state of things as a natural balance. If, for example, after having exterminated large numbers of insects

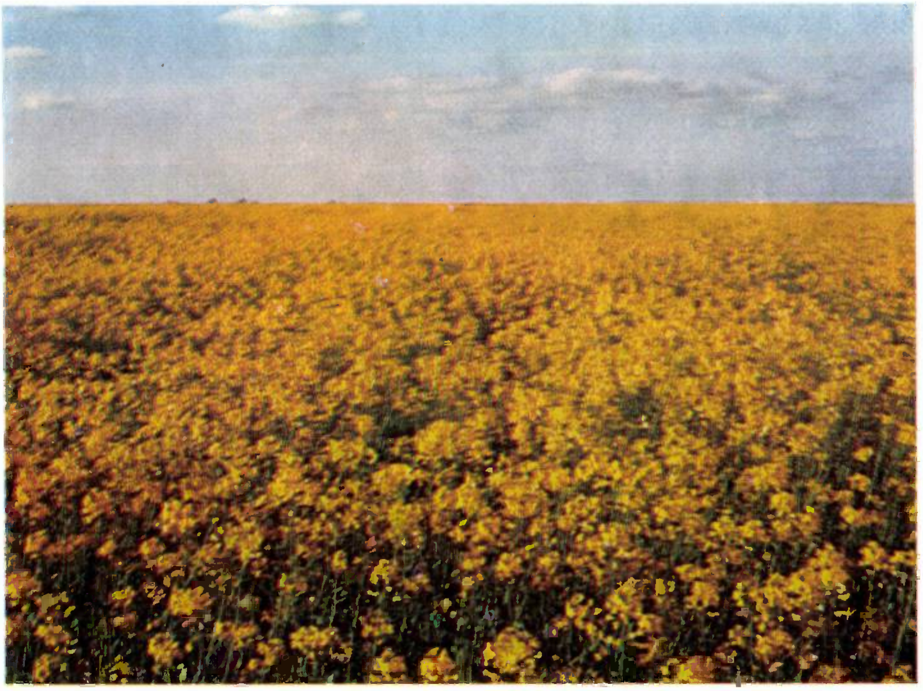


Fig. 3. Another part of the polder shown in fig. 2, in a more advanced stage. The weeds which were formerly present have been replaced by rapeseed. (The 86 000 acres under rapeseed in this polder form the largest area of its kind in Europe.) In the meantime, another pest has developed in this single-crop area: the fields of rapeseed have become thickly infested with weevils.

in his orchards by spraying with DDT [1] *), the fruit-grower finds that his action has enabled the mite population to multiply, he promptly blames himself for having disturbed the balance of nature. Nothing is farther from the truth; that balance was disturbed long ago, and now it has only been shifted a little further.

Usually it is the chemical control agents that are held responsible for such disturbances of the balance of nature. The physical methods, such as felling and burning, tilling, drainage, irrigation and the like, are generally accepted without demur.

It need hardly be said that there are no real grounds whatever for making this distinction between "inhumane chemical" and "humane physical" methods; it springs simply from the mistrust of innovations and the concomitant idealization of the old and familiar.

In fact, nature itself makes widespread use of chemical means of defence and aggression. Many plants, for instance, protect themselves from insects by toxic alkaloids. I need only mention the upas tree of South-East Asia (Antiaris toxicaria), whose milky sap is used as arrow poison.

Another mode of chemical warfare, this time between plants among themselves, takes place below ground. The roots of some plants excrete substances which, if absorbed by the roots of other plants, severely damage the latter or inhibit their growth.

This form of root competition is much more common than might be supposed. Many think of root competition as a struggle for mineral salts or water. That is by no means always the case. Bonner discovered that the rubber-producing guayule plant secretes cinnamic acid [2] from its roots, which acts as a growth-hormone antagonist when taken up by other plants and so inhibits their development[3]). In Java we were able to isolate from the soil in which Salvia occidentalis was growing a substance capable of almost completely inhibiting the growth of young coffee plants[4]) (*fig. 4*).

It was reported some years ago that when land badly infested by eelworms is planted with African marigolds (Tagetes), the eelworm population rapidly declines. This observation attracted considerabbe attention because of the difficulty of controlling this pest effectively. Investigations by Uhlenbroek and Bijlo showed that the roots of Tagetes contain a substance, terthienyl [3], which is extremely toxic to eelworms[5]). These are just a few of the many indications that the underground struggle in nature is fought, at least in part, by chemical means.

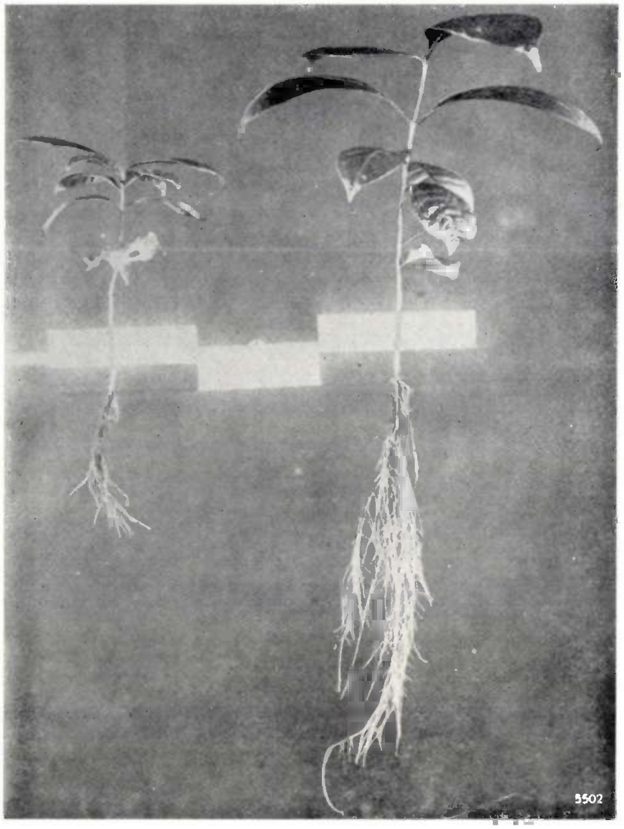


Fig. 4. Example of root competition. The coffee plant on the left was grown together with Salvia occidentalis in a water culture (after which the Salvia was removed). The coffee plant on the right was grown without Salvia. The growth of the left-hand specimen has been severely inhibited by substances excreted from the Salvia roots.

Above the ground, too, nature makes frequent use of chemical agents. One example is the shrub Encelia which has been studied by Went[6]) and Bonner[3]). This plant, a native of the Californian desert, exudes from its leaves a chemical substance that inhibits the germination of seeds. The sporadic rainfall in these arid regions washes the substance on to the ground under the shrubs. The seeds lying there would normally germinate in these humid intervals, but the inhibitor washed from the leaves prevents them from doing so. In this way the Encelia wards off the competition of other plants that might otherwise overwhelm it.

Chemical control of pests, a necessary measure

To control the spontaneous spread of weeds, the chemical industry has developed numerous herbicides which, administered in very small quantities, apparently block an essential biochemical reaction in plants.

Natural vegetation, already so impoverished by mechanical methods of agriculture, is menaced even more by these chemical agents. The botanist, of

*) The figures between square brackets refer to the structural formulae in the appendix.

course, views this development with regret. It is appalling to imagine what the Alpine meadows would look like if weed control of this kind were rigorously applied there. Our own native pastures have lost much of their former beauty now that they are no longer bright with buttercups, ladysmock, dandelions and sorrel, not to mention the wild orchid and other plants now rarely to be found.

We feel this as a sad deprivation and forget that the severest impoverishment was caused long ago by the growth of the population, resulting in the destruction of forests and the introduction of mechanical methods of cultivation. Chemical methods are simply a further step forward in a development that was initiated by our distant forebears, a development which is making the creation of nature reserves more and more urgently necessary.

Our most sensible policy is undoubtedly to take a realistic view of these chemical innovations in agriculture, and to set bounds to their application by establishing reserves where their use is forbidden. This whole course of events when all is said and done is the consequence of the enormous expansion of world population, which is a much more disquieting phenomenon.

Biochemical equilibria in cells

Before dealing with these pesticides any further, I should like to look for a moment at the structure of living matter, at the organism itself. There is a natural balance in an organism too, but of quite a different kind. Here it is a complex interplay of biochemical processes in which the chemical reactions are on the whole reversible. Each reaction thus represents a state of dynamic equilibrium. If the product of a reaction is removed by a subsequent reaction, the first reaction will continue to proceed in the same direction. If the reaction product accumulates, however, the reaction will come to a standstill or may even be reversed.

Nearly every compound formed in an organism should be seen as a link in a long chain which, with the uptake of energy, can lead to more intricate compounds, or which, upon the dissipation of energy, is normally degraded to simpler compounds.

The picture of a compound as a link in a long chain is really oversimplified. In fact we should think of it as a node in a network, indeed in a three-dimensional network. A system built up in this way can clearly be a stable one. Again, many vital processes are stabilized by ingenious feedback mechanisms. If a line is disturbed anywhere, the process usually continues by a roundabout route. Only the blockage of the most essential lines results in the death of the organism. Now what are these essential lines?

Indications in this direction may well be provided by the multifarious herbicides, fungicides and insecticides in use, for it is precisely these substances, applied in such extremely small concentrations, that yield such remarkable results, albeit in terms of the destruction of living matter. The study of these agents thus constitutes a sort of "negative biology". The following example will serve to illustrate the value of this negative biology in the fields of physiology and biochemistry.

A line of major importance is that along which energy is supplied in living matter. The energy required for synthesis is derived in most cases from the compound adenosine triphosphate [4], generally abbreviated to ATP. This compound is formed in processes of photosynthesis (photosynthetic phosphorylation), respiration (oxidative phosphorylation) and fermentation [7]).

In all three processes the energy made available by the oxidation of reduced compounds is stored in a phosphate bond of the ATP, which can then supply the energy for further processes that need it.

Any substance that impedes the formation of ATP must therefore be highly toxic to the organism. In agriculture substances of this kind are used which inhibit ATP formation in the process of oxidative phosphorylation. Known as "uncouplers", these agents do not stop respiration — on the contrary, they speed it up considerably — but no ATP is produced. The dinitrophenols are powerful uncouplers, and are used in large quantities as herbicides and insecticides (DNBP, DNOC [5]).

The dinitrophenols are also used as respiration uncouplers in many scientific investigations. Processes requiring energy are stopped by such a substance, as the necessary energy is normally supplied to the process via the ATP. In this way it is possible to distinguish energy-consuming processes from those that proceed freely. The dinitrophenols have been used for many years in physiological and biochemical research on e.g. the uptake and translocation of substances in plants. A good deal has thus come to be known about their action.

Another weed killer well known to agriculturists, 2,4-dichlorophenoxyacetic acid [6], belongs to the group of growth substances, its action on the plant being analogous to that of the normal growth hormone, indoleacetic acid [7]. Whereas, however, indoleacetic acid is a link in a chain of enzymatic reactions and is therefore dependent on other reactions both for its synthesis and for its breakdown, the dichlorophenoxyacetic acids do not fit into the biochemical scheme of the plant. They are assimilated and function as if they were growth hormones, but they cannot be broken down in the normal cycle and as a result the growth and development of the plant are completely disorganized (*fig.* 5).

There are some other chemical agents I should now like to mention because of their potential value to fundamental scientific research. Indeed, through their apparent ability to disturb the biological balance inside the cell, they may well provide us with the key to the essential processes in living matter. Unfortunately I must add at once that in this respect they have received far too little attention. What is at present known of their physiological and biochemical action stems mainly from the laboratories of the industries that produce them.

Years after the introduction of chlorophenyldimethyl urea [8], called "Monuron", as a herbicide in agriculture, nothing was known about the mechanism of its action. Wessels concluded on theoretical grounds that it was probably a potent inhibitor of photosynthesis [8]. This in fact proved to be the case, further research showing it to be one of the most active of such substances. So specific an inhibitor will undoubtedly be of considerable value to research into the process of photosynthesis.

Then there are the substituted triazines which, under the names of "Simazin" [9] and "Atrazin", are used in large quantities as weed-killers. It is known that these also inhibit photosynthesis, but it is quite likely that they arrest other vital processes too.

These are just some examples of substances about whose active mechanism we have at least some indications. But there are numerous other herbicides whose active mechanism is still completely unknown. One such substance is the germination inhibitor 2,6-dichlorobenzonitrile [10], minute doses of which make it impossible for most seeds to germinate, and moreover completely stop the growth of young plants (*fig. 6*). Others in this category are marketed under the proprietary names "IPC", "Avadex", "Dalapon", "Amitrol" and "Karsil". Further, there is "Reglone", a bromodipyridyl compound, extremely small concentrations of which (less than 1 pound per acre) are sufficient, in some way still unknown, to derange the physiological balance in plants so thoroughly that they die off in a very short time. It would be highly interesting to discover the particular system on which such minute quantities act.

Amongst the many fungicides in agricultural use, the only ones whose active mechanism is not completely unknown are the dithiocarbaminates [11]. Their physiological action has been elucidated, at least in part, by Kaars Sijpesteyn and co-workers [9], who also found the explanation for the double peaks in the curves showing the relation between effect and concentration (*fig.* 7).

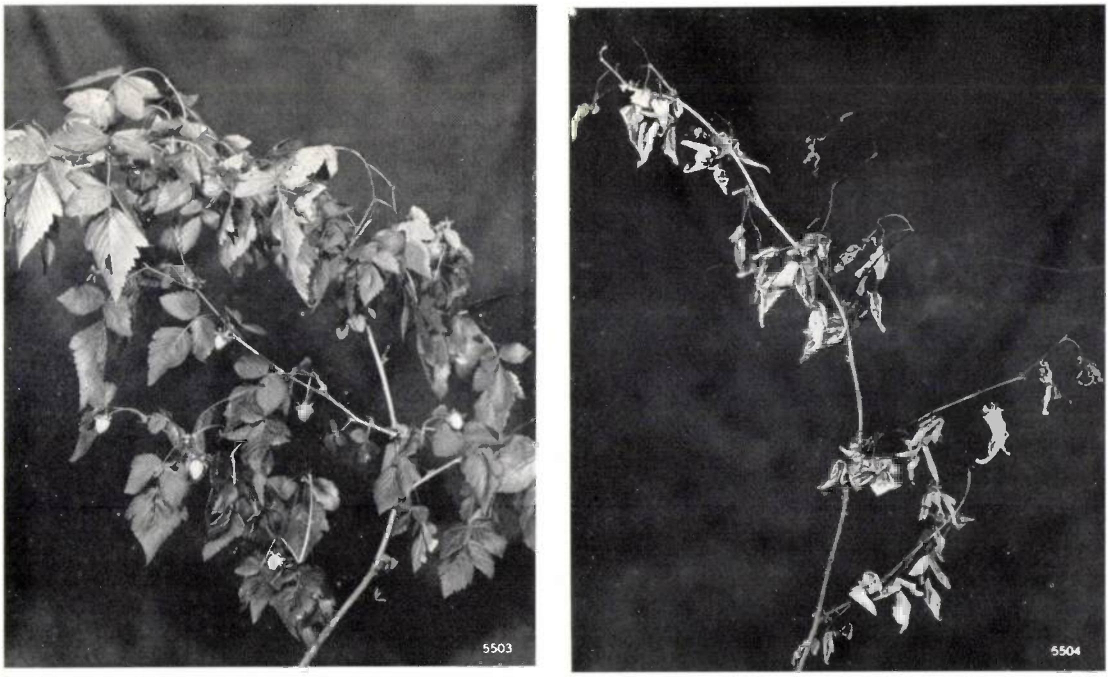


Fig. 5. The effect of the herbicide 2,4,5-trichlorophenoxyacetic acid on brambles is seen from the treated shrub on the right. (The shrub on the left is untreated.) This herbicide acts as a growth hormone, but does not fit into the biochemical scheme of the plant, and therefore completely disorganizes its development.

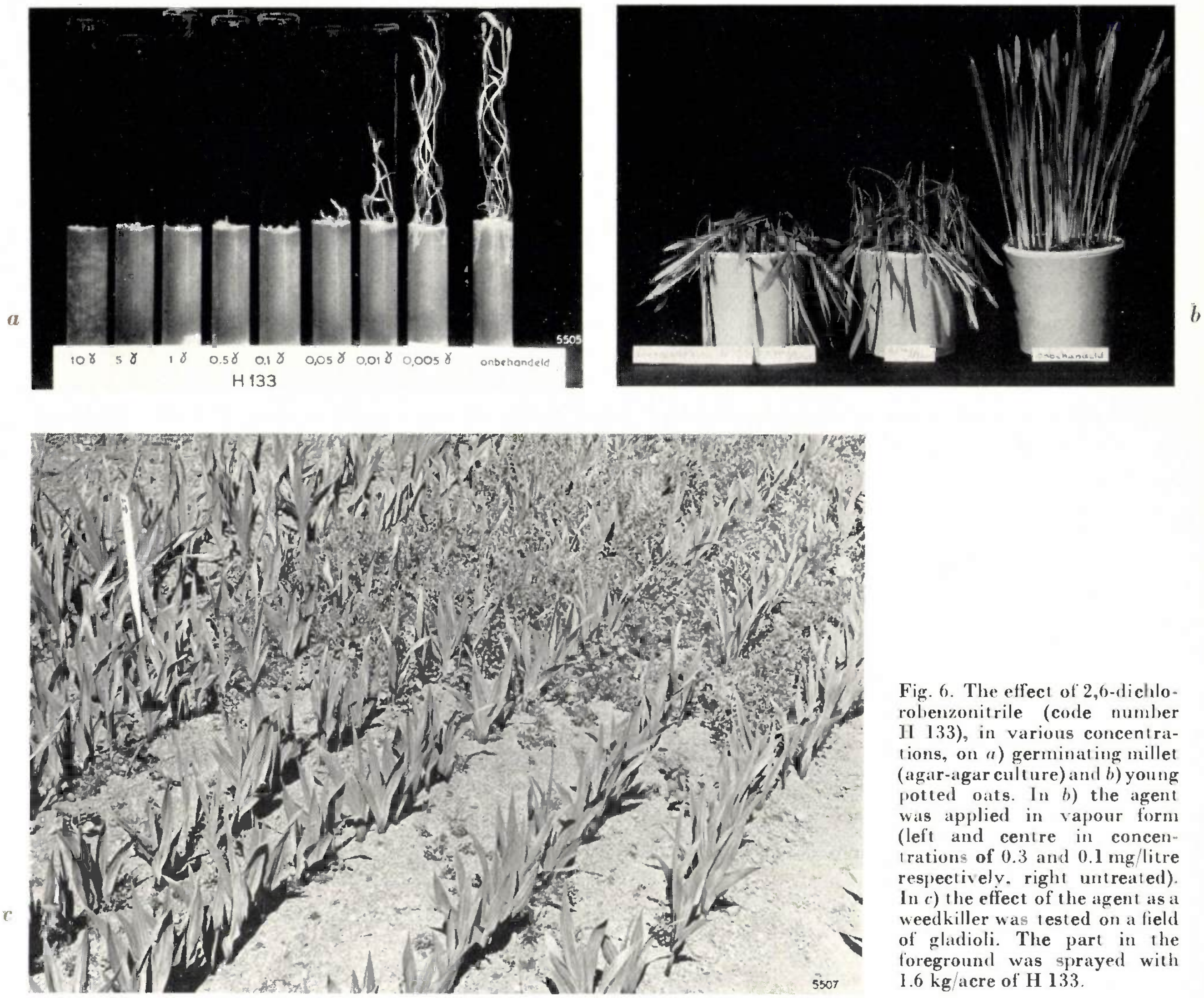


Fig. 6. The effect of 2,6-dichlorobenzonitrile (code number H 133), in various concentrations, on *a*) germinating millet (agar-agar culture) and *b*) young potted oats. In *b*) the agent was applied in vapour form (left and centre in concentrations of 0.3 and 0.1 mg/litre respectively, right untreated). In *c*) the effect of the agent as a weedkiller was tested on a field of gladioli. The part in the foreground was sprayed with 1.6 kg/acre of H 133.

Of the insecticides the group of phosphorus compounds is the subject of a great deal of biochemical research, owing to the fact that they inhibit the enzyme choline-esterase. Since the latter is essential to the proper functioning of the nervous system, the medical profession is keenly interested and various medicophysiological and biochemical laboratories are at present studying these substances.

Apart from this group, there are many more insecticides and acaricides which have so far only been studied as regards their practical application.

Conclusion

It is my opinion that there are many chemical agents used in agriculture that are certainly worth intensive and fundamental research to discover

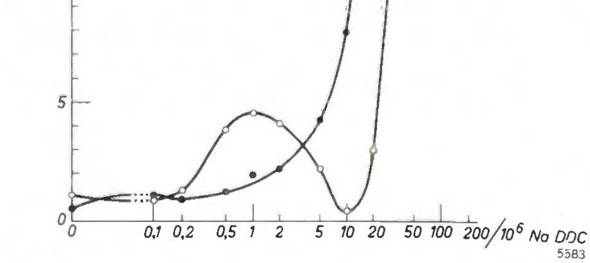
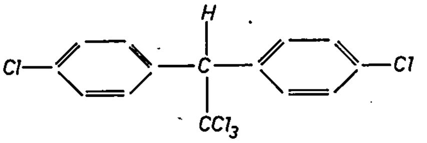


Fig. 7. Example of complications in the action of chemical control agents. The curves show the relation between the concentration of NaDDC [11] and its effect on Aspergillus niger in the presence and absence of copper (white and black circles, respectively). Where copper is present, the curves show two peaks. An explanation of this phenomenon will be found in reference [9]). The ordinate gives the content p of accumulated pyruvic acid, which is a measure of the biochemical disturbance on which the action of this pesticide depends.

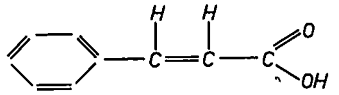
the mechanism of their action. Some of these substances are already subjects of such research. It is to be hoped that many of the others will soon be subjected to a similar investigation. The best way to make research along these lines yield rich fruit — apart from properly equipping the laboratories in question — is close cooperation between biologists on the one hand and biochemists and agriculturists on the other.

Appendix: structural formulae of some compounds mentioned in the text

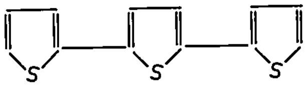
[1] DDT . . . .



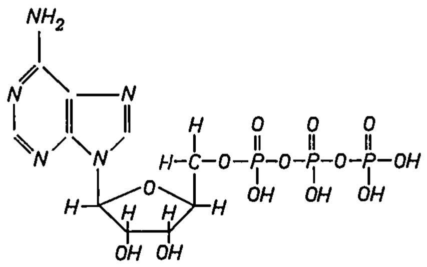
[2] Cinnamic acid . . . .



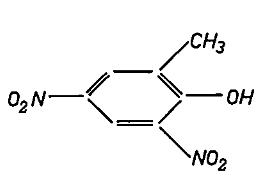
[3] Terthienyl . . . .



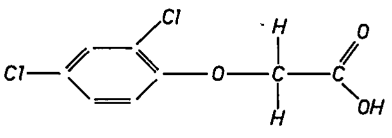
[4] Adenosine triphosphate (ATP). . . . .



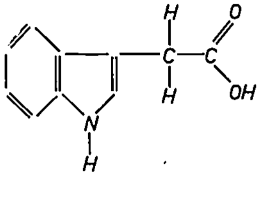
[5] DNOC. . . . .



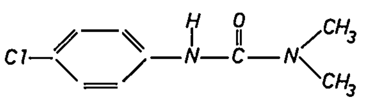
[6] 2,4-Dichlorophenoxyacetic acid



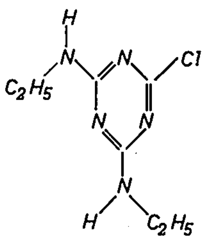
[7] 3-Indoleacetic acid . . . .



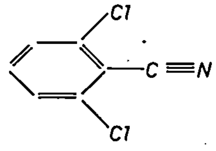
[8] Chlorophenyldimethyl urea . . .



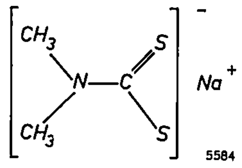
[9] 1-Cl-3,5-diethyl triazine . . . . .



[10] 2,6-Dichlorobenzonitrile . . . .



[11] Sodium dimethyl dithiocarbaminate (NaDDC) . . .



BIBLIOGRAPHY

[1] C. R. Darwin, On the origin of species by means of natural selection, 1859.

[2] J. F. Martinet, Katechismus der Natuur, published by J. Allart, Amsterdam 1778.

[3] J. Bonner, The role of toxic substances in the interactions of higher plants, Bot. Rev. **16**, 51-65, 1950.

[4] R. van der Veen, Wortelconcurrentie in de koffie- en rubbertuinen, Archief koffiecult. **9**, 65-102, 1935.

[5] J. H. Uhlenbroek and J. D. Bijloo, Proc. 4th int. Congr. crop protection, Hamburg 1957, Part I, 579-581, 1959.

[6] F. W. Went, The dependence of certain annual plants on shrubs in Southern California deserts, Bull. Torrey Bot. Club **69**, 100-114, 1942.

[7] See e.g. R. van der Veen, De weg van de energie bij levende organismen, Vakbl. Biol. **39**, 173-181, 1959.

[8] R. van der Veen and J. S. C. Wessels, Biochim. biophys. Acta **19**, 548, 1956.

[9] A. Kaars Sijpesteyn, M. J. Janssen and G. J. M. van der Kerk, Biochim. biophys. Acta **23**, 550, 1957.

For information on the subjects discussed, reference may also be made to the following articles that have appeared in this journal:

R. van der Veen, Influence of light upon plants, Philips tech. Rev. **11**, 43-49, 1949/50.
R. van der Veen, Photosynthesis, ibid. **14**, 298-303, 1952/53.
R. van der Veen, "Boekesteyn", the agrobiological laboratory of N.V. Philips-Roxane, ibid. **16**, 353-359, 1954/55.
J. Meltzer, Research on the control of animal pests, ibid. **17**, 146-152, 1955/56.
M. J. Koopmans, Fungicide research, ibid., pp. 222-229.
R. van der Veen, Growth substances in plants, ibid., pp. 294-298.
W. Duyfjes, The formulation of pesticides, ibid. **19**, 165-176, 1957/58.

Summary. Principal contents of the inaugural lecture delivered by the author as professor extraordinary of botany at the University of Utrecht. In e.g. cultivating the land, man participates in the constant evolution of the "balance of nature". The chemical agents used to this end should not be seen merely as a means of pest control; they can also be of value to the fundamental study of living matter. Their action relies on the disturbance of essential processes in the living organism, and because of that fact they may well be the key to understanding those processes. The author advocates close cooperation between biologists, biochemists and agriculturists with a view to more intensive research into this "negative biology", and mentions various chemical agents that might be studied along these lines.

THE APPLICATION OF CONTROL THEORY TO LINEAR CONTROL SYSTEMS

by M. van TOL *).

621-53.001

Up to the beginning of the 'forties, control engineering was approached largely on an empirical basis; measures taken to improve a control system tended to rely more on practical experience and intuition than on theoretical insight. Improvement came when it was recognized that the theory of control systems and servomechanisms could be derived directly from the theory of negative-feedback amplifiers.

The article below discusses the present-day treatment of control problems with the aid of complex variables, a method commonly used in alternating-current theory. As one example, an analysis is given of the frequency response characteristics of the temperature-control system used in the subcritical suspension reactor at Arnhem (Netherlands), previously described in this journal.

Introduction

Equipping a plant with a system for automatically controlling a variable such as temperature, liquid level or gas flow amounts essentially to giving it negative feedback and amplification. The output signal Θ_0 of the instrument that measures the variable is compared with a reference (input) signal Θ_i, and their difference ε is amplified and used to control the regulating device (*fig. 1*). It is therefore not surprising that control theory runs parallel with the theory of negative-feedback amplifiers and network analysis, developed in the 'twenties for the purposes of electronic engineering [1]).

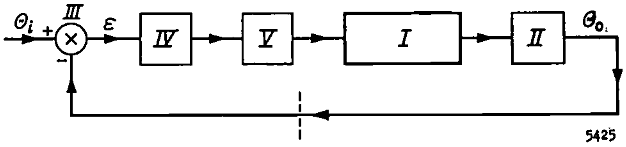


Fig. 1. Block diagram of an automatic control system. *I* plant. *II* instrument which measures the quantity to be controlled. The output signal Θ_0 from *II* is compared in *III* with the reference signal Θ_i. The difference ε between Θ_0 and Θ_i is the input signal of the controlling unit *IV*, the output signal of which controls the regulating unit *V*. The latter in turn acts directly on the quantity to be kept constant.

An analysis of the characteristics of an automatic control system is concerned with the closed loop (cf. fig. 1) formed by the plant in which the process to be controlled takes place, the measuring unit, the controlling unit, and the regulating unit. If all the elements composing the control loop are *linear*, the problem can be treated exactly with the aid of *complex variables*. Other methods will not be considered here.

When applying this method to the analysis of a negative-feedback amplifier, the starting point is to take the characteristics of the amplifier *without* feedback. The same procedure is applied to a control system. The loop is regarded as opened at some point — e.g. at the dotted line in fig. 1 — and an *open-loop transfer function* is assigned to it. This function, as in AC theory, is the ratio of the complex quantities which can be formed from a sinusoidal input and output signal — both having an angular frequency ω — by replacing the trigonometric function by an exponential function with an imaginary exponent. The symbols Θ_i and Θ_0 used above for the input and output signals in fact denote these complex quantities.

The transfer function is normally split into a constant real factor K and a complex variable $G(j\omega)$. The latter describes the variation with ω both of the gain (amplitude response) and of the phase shift φ between input and output signals. Since all elements of the loop are assumed to be linear, KG is equal to the product of the transfer functions K_1G_1, K_2G_2, etc. of the individual elements. It is therefore immaterial where the loop is imagined to be opened.

In this article we shall consider three examples of the way in which a transfer function is found in practice from an analysis of the physical data of a given process. The first example relates to controlling the level in a tank, a problem frequently encountered in the chemical industry; the second concerns the control of temperature in a glass-furnace feeder which was recently described in this journal [2]), and the third the system of controlling the suspension temperature in the subcritical nuclear reactor at the

*) Research Laboratories, Eindhoven.

[1]) This theory will be found in e.g. H. W. Bode, Network analysis and feedback amplifier design, Van Nostrand, New York 1945.

[2]) See P. M. Cupido, Philips tech. Rev. **22**, 311-319, 1960/61 (No. 9/10).

N.V. KEMA laboratories in Arnhem, also recently described in this journal [3]). It will be seen that the physical analysis referred to must often rely on approximations.

For the convenience of readers unfamiliar with the subject, we shall first examine the *KG* function at somewhat greater length in order to show how fundamental it is to control theory.

The transfer function

The importance of the *KG* function in control theory becomes immediately evident when we try to answer the following two questions: how far does the control system reduce a disturbance Θ_s, and is the closed loop sufficiently stable? We will first consider the case where the disturbance can only occur behind the last block in fig. 1. The control loop can then be reduced to the diagram in *fig. 2*. In that case

$$\Theta_\mathrm{o} = KG\varepsilon + \Theta_\mathrm{s}, \quad . \ . \ . \ . \ . \quad (1)$$

or, since $\varepsilon = \Theta_\mathrm{i} - \Theta_\mathrm{o}$,

$$(1 + KG)\Theta_\mathrm{o} = KG\Theta_\mathrm{i} + \Theta_\mathrm{s}\,,$$

hence

$$\Theta_\mathrm{o} = \frac{KG}{1+KG}\,\Theta_\mathrm{i} + \frac{1}{1+KG}\,\Theta_\mathrm{s}. \quad . \quad (2)$$

If we only want to know the extent to which a sinusoidal disturbance is suppressed, we may put $\Theta_\mathrm{i} = 0$, giving

$$\frac{\Theta_\mathrm{o}}{\Theta_\mathrm{s}} = \frac{1}{1+KG}. \quad . \ . \ . \ . \ . \ . \quad (3)$$

The absolute value of this ratio is called the *deviation ratio* [4]).

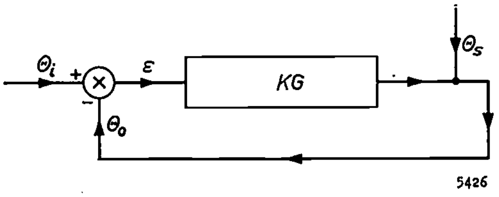


Fig. 2. Block diagram of control loop in which the disturbance Θ_s can only occur behind the last unit (cf. fig. 1).

It will thus be clear that the effect of the disturbance is not eliminated, but reduced by the factor given in eq. (3). The remaining effect is called the *deviation*. (In theory a zero deviation can be achieved by inserting an integrator in the loop, for which element *G* approaches infinity as ω approaches zero.)

As regards the *stability* of the system the magnitude of Θ_s is unimportant, and (2) reduces to

$$\frac{\Theta_\mathrm{o}}{\Theta_\mathrm{i}} = \frac{KG}{1+KG}. \quad . \ . \ . \ . \ . \ . \quad (4)$$

The significance of this equation will be made clear presently.

We now turn to the case where the disturbance occurs not at the end of the loop but at a point somewhere in the middle (*fig. 3*). We represent the transfer function of the blocks in front of that point by K_1G_1, that of the others by K_2G_2. Writing $K_1G_1K_2G_2$ as *KG*, the output signal in this case is

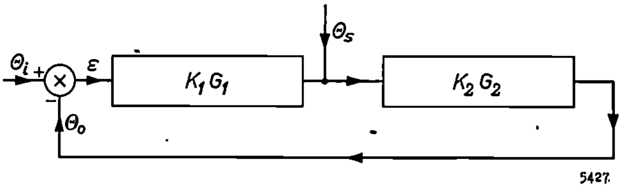


Fig. 3. Control loop where the disturbance signal Θ_s occurs somewhere in the middle. The transfer function of the units in front of this point is K_1G_1, that of the others K_2G_2.

$$\Theta_\mathrm{o} = KG\varepsilon + K_2G_2\,\Theta_\mathrm{s}\,, \quad . \ . \ . \quad (5)$$

which, in the same way as above, leads to

$$\Theta_\mathrm{o} = \frac{KG}{1+KG}\,\Theta_\mathrm{i} + \frac{K_2G_2}{1+KG}\,\Theta_\mathrm{s}. \quad . \quad (6)$$

From this relationship we again first derive the deviation ratio, i.e. the quotient of the values of the output signal with and without control action for the case where $\Theta_\mathrm{i} = 0$. We find

$$\frac{\Theta_\mathrm{o}}{\Theta_\mathrm{s}K_2G_2} = \frac{1}{1+KG}. \quad . \ . \ . \ . \quad (7)$$

For the stability of the system we again assume $\Theta_\mathrm{s} = 0$ in equation (6), which once more yields

$$\frac{\Theta_\mathrm{o}}{\Theta_\mathrm{i}} = \frac{KG}{1+KG}. \quad . \ . \ . \ . \ . \ . \quad (8)$$

Apart from the fact, which is trivial, that equations (4) and (8) are identical, we see that the right-hand side of eq. (7) is identical with that of (3). *The general rule is that to analyse both the stability of the system and the factor by which disturbances are reduced it is necessary to have the transfer function of the whole assemblage of units in series, i.e. the KG function of the open loop.*

The behaviour of an open control loop can also be represented graphically, either by a polar plot of the *KG* function in the complex plane (Nyquist diagram) or by first deriving from *KG* the separate

[3]) See B. L. A. van der Schee and M. van Tol, Philips tech. Rev. **21**, 121-133, 1959/60.

[4]) J. M. L. Janssen, Trans. Amer. Soc. Mech. Engrs. (ASME) 76, 1303, 1954.

amplitude and phase characteristics and plotting these in some convenient way. The Nyquist diagram is useful for finding quickly the absolute value P of the function $1 + KG$ at a particular frequency. This is done by joining the relevant point on the curve to the point (—1,0) on the axis, as illustrated in *fig. 4*.

The frequency response curves are often plotted on a logarithmic scale: the amplitude characteristic in logarithmic coordinates and the phase characteristic in semilogarithmic coordinates (Bode diagram). This method offers advantages when the maximum permissible gain is to be determined.

As far as *stability* is concerned it will be sufficient here to note that the frequency response curves should be such that the gain $|KG(j\omega)|$ is less than unity at the frequency ω_c where φ is —180°, i.e. where the negative feedback in the loop has changed to positive feedback (regeneration). In the Nyquist diagram this means that the curve must not enclose the point (—1,0) [5]).

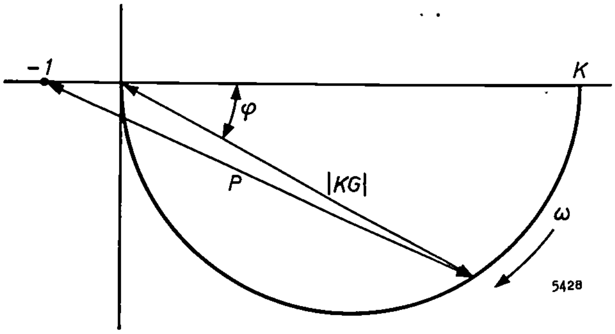


Fig. 4. Nyquist diagram of the function $KG = K(1 + j\omega\tau)^{-1}$. The distance from a point on the curve to the origin gives the absolute value of KG at the relevant frequency, i.e. the gain. Disturbances are reduced by the factor P, found from the distance of a point on the curve to the point (—1,0). The phase shift φ between output and input signal runs here from 0 to —90°.

In addition to Bode and Nyquist diagrams the graphical methods shown in *fig. 5* are sometimes used. The plot in fig. 5*a*, called a Nichols chart, is in effect a Nyquist diagram drawn in semilogarithmic rectangular coordinates. In fig. 5*b*, called the inverse Nyquist diagram, the reciprocal of KG is drawn in the complex plane. This method is useful for quickly determining the response of the closed loop, since it follows from (4) that $\Theta_i/\Theta_o = (1 + KG)/KG = KG^{-1} + 1$.

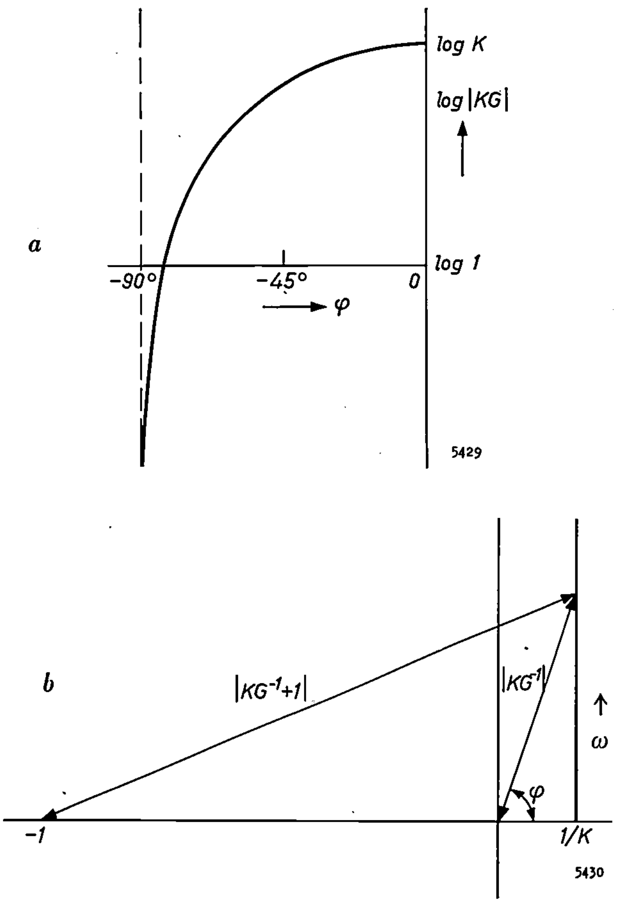


Fig. 5. *a*) Relation between log $|KG|$ and φ for the transfer function $KG = K(1 + j\omega\tau)^{-1}$ (cf. fig. 4) plotted in rectangular coordinates (Nichols chart).
b) Graph of the function $1/KG$ in the complex plane, likewise for the case $KG = K(1 + j\omega\tau)^{-1}$.

Example of the determination of the transfer function of a single element

As an example of the method used to determine the transfer function of an element we take an open tank into which water is run from a tap (rate of flow I_1) and at the same time run off through an outlet (rate of flow I_2). The tank is assumed to be of uniform width, area of base F, so that the volume of water V in the tank is proportional to the height H to which the tank is filled (*fig. 6*). It is further assumed that I_2 is proportional to H, which we may write $I_2 = c \times H$. If I_1 is constant (value I_0),

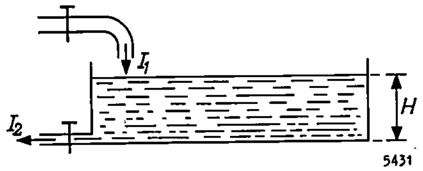


Fig. 6. For calculating the transfer function of a tank filled with water from a tap (flow rate I_1, independent variable) and emptied through an outlet where the flow rate I_2 is proportional to the height H.

[5]) For complicated control systems the stability criterion as formulated here is oversimplified, but it is adequate for the relatively simple systems discussed in this article. The stability problem is dealt with at some length e.g. by G. S. Brown and D. P. Campbell, Principles of servomechanisms, Wiley, New York 1948, and by G. J. Thaler and R. G. Brown, Servomechanism analysis, McGraw-Hill, New York 1953.
The relation between stability and gain will be dealt with in the next number of this journal.

the value H_0 which H has in the steady state is such that I_2 is equal to I_0, and therefore $H_0 = I_0/c$. We now want to ascertain how H will fluctuate around the value H_0 when I_1 is not constant but shows a sinusoidal ripple of amplitude $\hat{i}_1$ and angular frequency ω:

$$I_1 = I_0 + \hat{i}_1 \sin \omega t.$$

The problem is defined by the differential equation

$$I_1 - I_2 = \frac{\mathrm{d}V}{\mathrm{d}t} = F\frac{\mathrm{d}H}{\mathrm{d}t},$$

or, substituting $H_0 + h$ for H and given $I_0 = cH_0$,

$$F\frac{\mathrm{d}h}{\mathrm{d}t} + c\,h = \hat{i}_1 \sin \omega t. \quad . \quad . \quad (9)$$

We can readily solve an equation of this kind by finding a solution of the form $\bar{h} = \hat{h}\,\mathrm{e}^{\mathrm{j}\omega t}$ for the equation

$$F\frac{\mathrm{d}\bar{h}}{\mathrm{d}t} + c\,\bar{h} = \bar{i}_1, \quad . \quad . \quad . \quad . \quad (10)$$

where $\bar{i}_1 = \hat{i}_1\,\mathrm{e}^{\mathrm{j}\omega t}$. Substituting the expression for $\bar{h}$ in eq. (10), we obtain

$$F\mathrm{j}\omega\bar{h} + c\,\bar{h} = \bar{i}_1,$$

or

$$\bar{h} = \frac{\bar{i}_1}{F\mathrm{j}\omega + c} = \frac{\bar{i}_1/c}{1 + \mathrm{j}\omega\tau}, \quad . \quad . \quad (11)$$

where τ represents the quotient F/c, which is easily seen to have the dimension of time. Since finding the KG function consists in finding the ratio of the complex quantities corresponding to the input and output signals, it follows from equation (11) that

$$KG = \frac{1/c}{1 + \mathrm{j}\omega\tau},$$

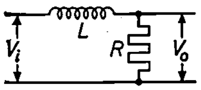
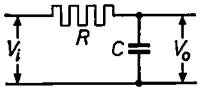
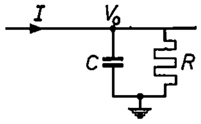
whence

$$K = 1/c \quad \text{and} \quad G = \frac{1}{1 + \mathrm{j}\omega\tau}. \quad . \quad (12)$$

(The factor $1/c$ in the numerator of (11) is not a gain factor but a dimensional factor, due to our choice of h as the starting quantity instead of the flow I_2. If we had started with I_2, K would have been unity; there is no amplification of the input signal.)

The form of the G function found here is frequently encountered in practice. The operation of most circuit elements used in control systems can be defined by an arrangement of blocks in series — as many blocks as there are time constants — having a transfer function of the form $K(1 + \mathrm{j}\omega\tau)^{-1}$.

Table I. Electrical circuits whose transfer function can be expressed in the form $K(1 + \mathrm{j}\omega\tau)^{-1}$. In the third example the input and output quantities have different dimensions, resulting in a dimensional factor R. The gain factor is in each case equal to unity.

Circuit	Input and output quantities	Transfer function	τ
	V_i, V_o	$\frac{R}{R + \mathrm{j}\omega L} = \frac{1}{1 + \mathrm{j}\omega L/R}$	L/R
	V_i, V_o	$\frac{1/\mathrm{j}\omega C}{1/\mathrm{j}\omega C + R} = \frac{1}{1 + \mathrm{j}\omega RC}$	RC
	I, V_o	$\frac{1}{1/R + \mathrm{j}\omega C} = R\frac{1}{1 + \mathrm{j}\omega RC}$	RC

Some familiar electrical circuits that have this transfer function are shown in *Table I*. The last one may be regarded as the equivalent electrical circuit of the tank in our example.

Control of the glass temperature in a tank-furnace feeder

As our second example we shall consider a system for controlling the temperature of molten glass in the feeder from a tank furnace to a glass-working machine. In the article cited [2]) it was explained that it is important to ensure that the conditions under which glass is produced are kept as uniform as possible. The first prerequisite is to keep the temperature of the glass in the melting tank carefully constant. Since this cannot be done directly, the various quantities which affect the glass temperature are stabilized separately. These quantities are the temperature of the fuel oil for the burners, the pressure of the atomizing air, the flow rates of oil and air to the burners (or possibly the flow rate of the air and the ratio of oil flow to air flow), the pressure of the combustion gases above the glass, and the level in the tank. From the melting tank the glass flows through the "working end" of the furnace into one or more feeders for the glass-working machines. Since the flow of glass to the machines must be constant, the temperature in the feeder is separately controlled; this control must be very rigorous, because the viscosity of the glass, which has a great effect on its rate of flow, is highly temperature-dependent. We shall now examine the control system used for this purpose.

The temperature of the glass in the feeder is controlled with the aid of an electrical thermometer, which is immersed in the molten glass and whose output signal is used to regulate the burners — gas burners in this case (*fig.* 7); we shall not discuss the practical details of this control here. A block diagram of the control system is shown in *fig. 8*. Block *I* is the feeder with molten glass, which receives a specific flow of heat and acquires the temperature Θ_t. Block *II* is the thermometer, whose output signal Θ_0 is compared with the desired value Θ_i. The difference ε is amplified in block *III* — we assume provisionally that the amplification A is linear and independent of frequency. The output from *III* controls the valve *IV* which regulates the gas supply and thus indirectly the flow of heat to the gas. Let us now try to find the transfer function of this system.

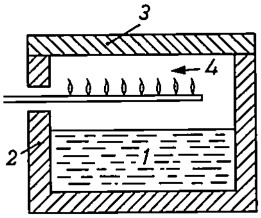


Fig. 7. Heating of molten glass in a glass-furnace feeder. *1* glass, *2* wall of feeder. *3* roof. *4* gas burner.

We shall argue presently that the feeder with molten glass, like the water tank discussed, behaves in principle like the electrical circuit of a resistance and capacitance in parallel (see example 3 in Table I). The current I represents the flow of heat from the burners to the glass, the capacitance C the heat capacity of the glass, R the thermal resistance over which the heat is dissipated, and V_0 the temperature to be kept constant. Disturbances that arise may be due to fluctuations in I, C or R. Fluctuations in C may be due to variations in the level of glass, fluctuations in R to variations in the chimney draught. We shall confine ourselves here to variations in I. Putting $RC = \tau_I$, the frequency-dependent part of the transfer function for block I is given by

$$G_I = \frac{1}{1 + j\omega\tau_I} . \quad . \quad . \quad . \quad . \quad (13)$$

Because of the very high values assumed by the heat capacity of the glass and the thermal resistance, τ_I is of the order of magnitude of half an hour.

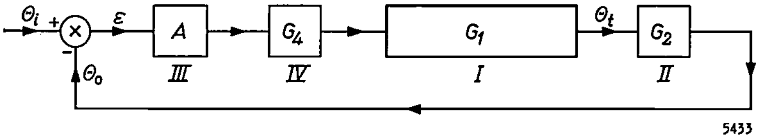


Fig. 8. Block diagram (simplified) of system for controlling the temperature of the molten glass flowing through a feeder from the glass furnace to the glass-working machine. *I* feeder. *II* thermometer. *III* linear amplifier. *IV* regulating valve.

In order to justify the assumption that this case may in fact be identified, as a first approximation, with the third example in Table I, we shall look at what happens in the feeder (see fig. 7). The heat delivered by the burners is partly transferred to the glass and partly to the walls and roof of the feeder. The ratio between the two heat flows will be roughly independent of the magnitude of the total heat flow. The way in which the heat is transported from the flame to the glass need not be considered here. All we are interested in is the fact that there is a flow of heat to the glass that depends on the flow of gas to the burners, and which responds almost without any lag to a change in the gas flow.

It remains to be shown that the process of heat removal may be regarded as the flow of current through a resistance as a result of a potential difference V_0 which represents the temperature difference ΔT between the glass and the surroundings of the furnace. The heat is transported partly by radiation and partly by conduction and convection: the glass radiates heat to the roof and walls, heat is conducted through the glass and walls to the atmosphere, and finally heat is lost from the outside of the feeder by radiation, conduction and convection. The dependence of the heat flow on the temperature difference ΔT can, of course, be expressed in a power series. This series, owing in particular to the presence of a radiation component, will also contain higher powers of ΔT. It follows, then, that the thermal resistance, i.e. the derivative of the heat flow with respect to ΔT, is not independent of ΔT and in an equivalent electrical circuit may only be represented by a pure resistance provided ΔT is practically constant. In our case this condition is satisfied, the temperature variations of the molten glass in the feeder being of the order of only 1 °C.

Not so much need be said about the other units of the loop. The thermometer, block *II*, can be treated as a capacitor (heat capacity) c which is charged by a current supplied by a voltage source via a resistance r. Here the input quantity is the voltage which represents the temperature of the glass tank (Table I, example 2). The transfer function for G_{II}, like that for G_{I}, therefore again has the form $(1 + \mathrm{j}\omega\tau_{\mathrm{II}})^{-1}$, but the value of the time constant τ_{II} ($= rc$) is very much smaller (e.g. half a minute).

The transfer function of block *IV*, the regulating valve, must define the relationship between the output signal from *III* and the flow of heat to the tank. There are two steps involved. The signal from *III* controls the valve in the gas pipe. The relationship between this signal and the corresponding change in the flow of gas will not, in general, be exactly linear. Secondly, the flow of gas to the burner determines the flow of heat to the glass; the relationship between these two quantities, which is partly governed by the size of the luminous cone in the flame [6]), will not always be linear either. For small variations, however, the behaviour of a non-linear element like block *IV* may permissibly be regarded as linear. Since a valve usually shows a certain lag, which may be characterized by one time constant, we have $G_{\mathrm{IV}} = (1 + \mathrm{j}\omega\tau_{\mathrm{IV}})^{-1}$. Here again, τ_{IV} is small compared with τ_{I}.

Let us now briefly review the simplifications we have adopted. Firstly, of course, there were the approximations introduced to simplify the formulae for G_{I} and G_{IV}. Further, the block diagram as such (fig. 8) contains various approximations. In the first place we have disregarded the fact that, owing to the heat capacity of the burner, the heat flow to the glass does not respond instantaneously to a change in the gas supply. To take this into account we must add the factor $(1 + \mathrm{j}\omega\tau_{\mathrm{V}})^{-1}$ to the transfer function of the whole loop. It is also a simplification to represent the glass by a single capacitance and to neglect the heat capacity of the wall. Finally, the temperature distribution of the glass is certainly not entirely uniform. There is no space to allow for all these factors here.

As a first approximation, however, our simplified treatment of the system's behaviour is justified. We may even go a step further and, since τ_{II} and $\tau_{\mathrm{IV}} \ll \tau_{\mathrm{I}}$, equate G_{II} and G_{IV} to unity. The block diagram then reduces to that of *fig. 9*, and the transfer function of the open loop is

$$KG = \frac{A}{1 + \mathrm{j}\omega\tau_{\mathrm{I}}}, \quad . \quad . \quad . \quad . \quad . \quad (14)$$

where $K = A$, and $G = (1 + \mathrm{j}\omega\tau_{\mathrm{I}})^{-1}$.

As far as the *stability* of the system is concerned, however, the presence of the small time constants can *not* be neglected, for the variations of the controlled quantity are required to be very small, and therefore K must be given a very high value (see eq. (3)). The extent to which that is possible without endangering the stability of the system is in fact governed by the value of these small time constants [7]).

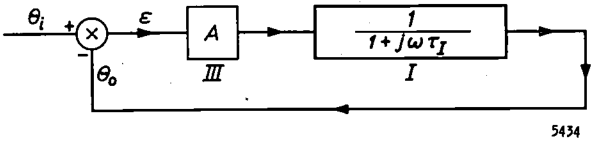


Fig. 9. Simplified form of block diagram of fig. 8, obtained by equating with unity the G function of all blocks for which $\tau \ll \tau_{\mathrm{I}}$. This approximation is permissible only if the gain factor A of *III* is not too high.

Control of the suspension temperature in the KEMA subcritical suspension reactor at Arnhem

Before finding the open-loop transfer function for the subcritical suspension reactor at Arnhem, we shall briefly recapitulate the operation of the control system, which has already been described in this journal [3]). A suspension of uranium oxide is pumped at a constant rate around the reactor circuit (see *fig. 10*). The pump P supplies 3-7 kW of energy to the fluid, and therefore some cooling is necessary even at the highest operating temperature. Stabilization of the temperature against variations in the power supplied by the pump and withdrawn by the cooler is effected by applying somewhat excessive cooling and by compensating the excess with the aid of a variable heater element V, which

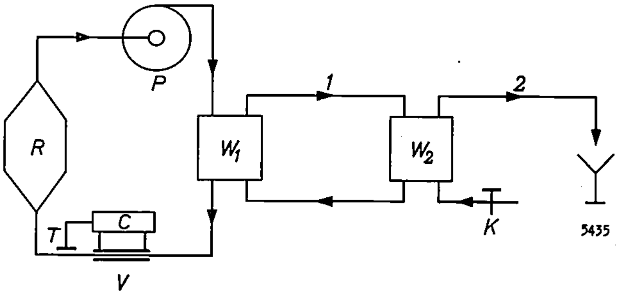


Fig. 10. Schematic representation of reactor circuit (left) and cooling circuits of the KEMA subcritical suspension reactor at Arnhem. R reactor vessel. P pump which circulates the suspension of uranium oxide around the reactor circuit in the direction of the arrows. T resistance thermometer. C control unit. V heater (regulating circuit). W_1 first heat exchanger. *1* primary cooling circuit. W_2 second heat exchanger. *2* secondary cooling circuit. K cooling-water valve.

[6]) See page 313 of reference [2]).

[7]) To be discussed in the forthcoming article mentioned in footnote [5]).

surrenders its heat directly to the suspension. The heater element is controlled by a thermometer T, which measures the temperature of the suspension. At equilibrium, the heater gives up about half its maximum power (5 kW), making it possible to offset variations of more than 2 kW in the power to be dissipated. The heater is simply a length of metal piping through which an electric current is passed. The temperature is measured with a Wheatstone-bridge resistance thermometer which delivers a voltage proportional to the difference between the measured and the desired temperature.

A block diagram of the control system is shown in *fig. 11.* The first block (A) again represents a linear amplifier. The transfer functions of blocks I and II together define the behaviour of the heater (i.e. the way in which the temperature Θ_t of the effluent suspension reacts to a variation in the power supplied to the heater), that of the last block (III) the behaviour of the thermometer itself. The disturbance Θ_s is a change in the temperature of the suspension, which is not detected until liquid of the wrong temperature reaches the thermometer. A fluctuation in the mains voltage must also be introduced at this point.

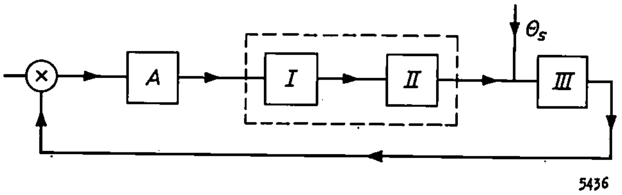


Fig. 11. Block diagram of temperature-control system in the Arnhem subcritical suspension reactor. Blocks I and II together represent the heater, block III the thermometer.

It may be asked whether fig. 11 does in fact represent the control loop, for there is another closed loop involved, viz. that of the circulating suspension. In order to show that the latter loop may be disregarded for our purposes, we recall that the control system has been designed to offset variations in the power delivered to the suspension by the pump and removed by the cooler [3]). The control equipment need only ensure that the fluid leaves the heater at the correct temperature. It is immaterial whether or not the fluid itself flows in an open or closed circuit.

Nor are our considerations affected by disregarding the influence of heat losses in other parts of the reactor circuit. Any such variation can be treated as a disturbance signal entering the loop at the same point as Θ_s, and can thus be regarded as a component of the latter.

To determine the transfer function of the heater it is convenient to divide the heating process into two stages, represented in fig. 11 by blocks I and II. The transfer function K_IG_I defines the response of the temperature of the wall of the heater to a variation in the supplied power; the function $K_{II}G_{II}$ defines the response of the temperature of the effluent to a change in the temperature of the wall of the heater. In this connection we make two closely related simplifying assumptions. Firstly, we assume that the temperature of the heater pipe is uniform at all points, so that we may in fact refer to *the* temperature of the pipe. Secondly we assume that variations in the temperature of the effluent are negligible compared with the temperature variations shown by the pipe wall; in other words, we regard the heat capacity of the fluid as extremely high. In fact the suspension circulates fast enough for these assumptions to be correct. Otherwise, our division of the heater into two blocks would not be permissible, unless feedback were applied from block II to block I.

The transfer function K_IG_I can most quickly be found from the equivalent electrical circuit shown in *fig. 12a.* The current I represents the power supplied, C_1 the heat capacity of the pipe, R_2 the thermal resistance at the boundary between pipe and suspension, C_2 the heat capacity of the suspension, V_1 the temperature of the pipe and V_2 the temperature of the fluid. Resistances R_1 and R_3 represent heat losses due to conduction and convection in the pipe and fluid mass in the rest of the reactor circuit. Measurements show that $C_1 \approx 500$ J/°C, $C_2 \approx 400\ C_1$, $R_1 \approx 25\times10^{-3}$ °C/watt and $R_2 \approx 2.2\times10^{-3}$ °C/watt. We see that C_2 is in fact much larger than C_1. We can thus regard C_2 as short-circuiting R_3, in which case the diagram in

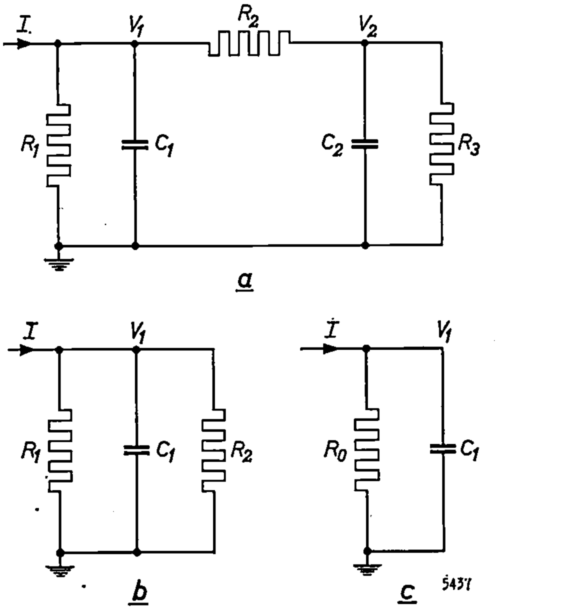


Fig. 12. *a*) Equivalent electrical circuit of heater operation. As $C_2 \gg C_1$, this can be simplified to *b*) or, by substituting R_0 for R_1 and R_2 in parallel, to *c*).

fig. 12*a* reduces to that in fig. 12*b*, and the latter may in turn be simplified to fig. 12*c* by replacing R_1 and R_2 by a single resistance R_0.

The transfer function of the first block is therefore equal to (see Table I, example 3):

$$\frac{V_1}{I} = R_0 \frac{1}{1 + j\omega\tau_I}, \quad . \; . \; . \; . \quad (15)$$

i.e. $K_I = R_0 = 2$ °C/kW and $G_I(j\omega) = (1 + j\omega\tau_I)^{-1}$. The time constant τ_I is equal to R_0C_1 (one second).

The transfer function $K_{II}G_{II}$ (the change in the temperature of the fluid due to the change in the heater temperature) is found by determining the way in which the temperature of the effluent suspension fluctuates when the temperature of the pipe varies sinusoidally around the equilibrium value T_0 with an amplitude T_1 which is small compared to the average temperature difference between the pipe wall and the suspension. (Since the flow is highly turbulent, the temperature of the fluid is virtually identical in all points of a cross-section.)

If we consider a "disc" of suspension of thickness dx, moving at a constant velocity v through the pipe, wc see that the disc must constantly encounter a different wall temperature in its passage through the pipe. The diffcrence betwcen this temperature and the equilibrium temperature of the wall is $T_1 \sin \omega(t_0 + x/v)$. Here t_0 is the time at which the disc entered the pipe, and x the distance which the disc has travelled in the pipe at the given moment. We shall now consider how much extra heat is supplied to the disc in its passage through the pipe as a result of this temperature difference. In the time dx/v during which the disc remains between x and x + dx it takes up an additional quantity of heat which gives rise to a temperature change proportional to

$$T_1 \sin \omega\left(t_0 + \frac{x}{v}\right) \times \frac{dx}{v}. \quad . \; . \; . \quad (16)$$

The integral of this expression over the whole length l of the heater is proportional to the difference between the actual temperature which the disc has acquired by the time it reaches the end of the pipe and the temperature it would have acquired if the wall temperature had permanently been equal to T_0. This difference is thus proportional to

$$\int_0^l T_1 \sin \omega \left(t_0 + \frac{x}{v}\right) \frac{dx}{v}, \quad . \; . \; . \quad (17)$$

which reduces to:

$$2\tau_{II}\, T_1 \frac{\sin\omega\,\tau_{II}}{\omega\,\tau_{II}} \sin \omega\,(t_0 + \tau_{II}), \quad (18)$$

where $\tau_{II} = l/2v$ (about 0.15 second). If we now consider the entire flow, i.e. a continuous series of discs for which the value of t_0 is successively larger by an amount dx/v, we may conclude that the variation of the temperature difference with time is given, apart from a constant factor, by expression (18) if t_0 is replaced by $(t - l/v)$, which is equal to $(t - 2\tau_{II})$. Expression (18) then becomes:

$$2\tau_{II}\, T_1 \frac{\sin\omega\,\tau_{II}}{\omega\tau_{II}} \sin \omega(t - \tau_{II})\,. \quad . \; . \quad (19)$$

The corrcsponding complex quantity (we shall henceforth disregard the constant factor $2\tau_{II}$) is:

$$T_1 \frac{\sin\omega\,\tau_{II}}{\omega\tau_{II}}\, e^{j\omega(t-\tau_{II})}. \quad . \; . \; . \; . \quad (20)$$

To find $G_{II}(j\omega)$ we divide (20) by the complex quantity corresponding to the input signal, $T_1e^{j\omega t}$, giving:

$$G_{II}(j\omega) = \frac{\sin\omega\,\tau_{II}}{\omega\tau_{II}}\, e^{-j\omega\,\tau_{II}}. \quad . \; . \; . \quad (21)$$

It can be seen from eq. (21) that the time τ_{II} (which is known as the distance velocity lag) is of the nature of a transit time. This transit time causes a phase shift $\varphi = -\omega\tau_{II}$ proportional to ω (*fig. 13*). Further, there is a frequency-dependent amplitude factor $(\sin\,\omega\tau_{II})/\omega\tau_{II}$ which is equal to unity at very low frequencics and is zero when $\omega\tau_{II} = \pi$. In the latter case each disc traverses the heater in exactly onc period, and thus takes up just as much energy in the one half as it gives up in the other half.

Block *III* in the diagram defines the behaviour of the resistance wire of the thermometer. This behaviour can be characterized to a good approximation by a single time constant τ_{III} (about 0.10 sec), so that

$$G_{III}(j\omega) = \frac{1}{1 + j\omega\tau_{III}}. \quad . \; . \quad (22)$$

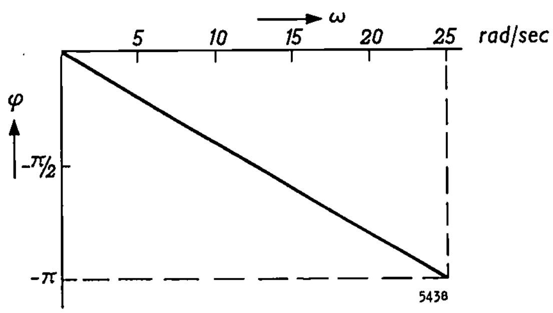


Fig. 13. Phase characteristic of the dissipation of heat by the heater in the circulating suspension (cf. eq. (21)). Because of the distance velocity lag τ_{II}, there is a phase shift proportional to ω. At the frequency $\omega = \pi/\tau_{II}$ ($\approx$ 25 rad/sec) the phase shift φ is $-\pi$.

The transfer function of the open control loop as a whole, which is equal to the product $AK_{\rm I}G_{\rm I}({\rm j}\omega) \times K_{\rm II}G_{\rm II}({\rm j}\omega)K_{\rm III}G_{\rm III}({\rm j}\omega)$, is thus:

$$\frac{\Theta_0}{\varepsilon} = AK^{\rm I}\, K_{\rm II}K_{\rm III}\, \frac{1}{1+{\rm j}\omega\tau_{\rm I}} \times \frac{\sin\, \omega\tau_{\rm II}}{\omega\tau_{\rm II}}\, {\rm e}^{-{\rm j}\omega\tau_{\rm II}} \times \frac{1}{1+{\rm j}\omega\tau_{\rm III}}. \quad . \ (23)$$

(Since the thermometer is not exactly at the end of the heater but somewhat further up in the pipe line, we should introduce into eq. (23) a second distance velocity lag, equal to the time $\tau_{\rm IV}$ which the fluid needs to cover the relevant distance. However, since $\tau_{\rm IV} \ll \tau_{\rm II}$, this factor may be neglected.)

The frequency response curves derived from this transfer function are shown in *fig. 14*, and the Nyquist diagram in *fig. 15*. In *fig. 16* the factor *P*, the reciprocal of the deviation ratio, is plotted as a function of ω.

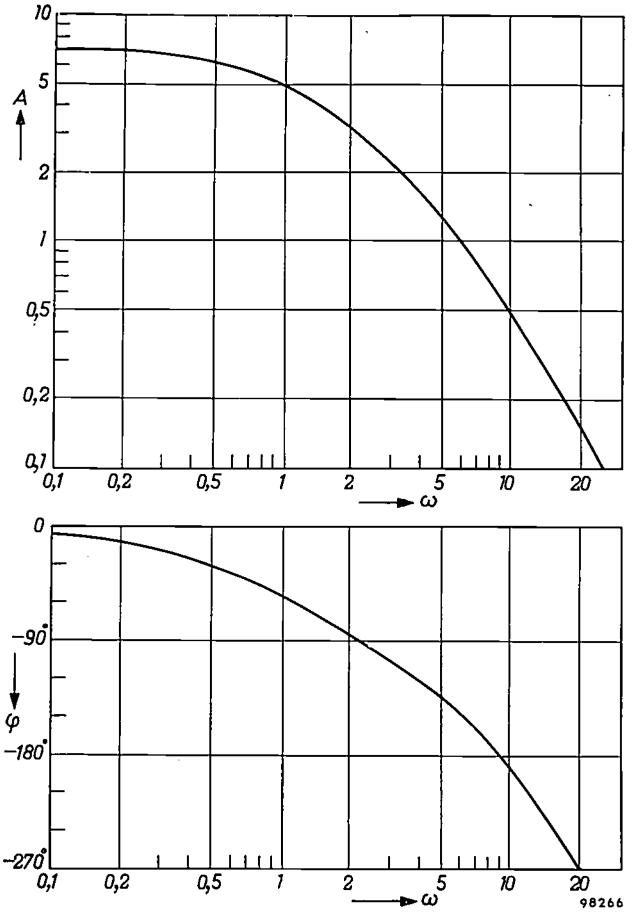


Fig. 14. Frequency response curves of the open control loop. *Top:* amplitude characteristic. *Bottom:* phase characteristic. At very low frequencies the gain is roughly 7. At the value of ω where the phase shift φ is 180°, the gain is less than unity.

A word should be added about the fact that the system is unable to correct high-frequency disturbances (see fig. 16). A sudden change in the valve controlling the cooling water in the secondary cooling circuit will not result in an immediate change in the temperature of the liquid flowing over the thermometer, nor will a sudden change in the mains voltage, i.e. in the power fed back to the heater (cf. fig. 12). Rapid variations of this kind thus give rise to relatively slow disturbances in the temperature of the suspension — which will, of course, be corrected — and are therefore unable to do any harm.

In conclusion, some general remarks. The above examples have shown that it is very often not possible to calculate a loop transfer function exactly, and that approximations have to be made. In any case, exactly linear systems are never in fact encountered, so that our assumption of linearity is in itself an approximation. In cases where it is not possible to calculate the *KG* function at all, the frequency response curves can in principle be found from measurements. A sinusoidal disturbance is deliberately introduced, and the resultant variation of the output signal (Θ_0 in fig. 1) is then recorded. This method cannot of course be used where the

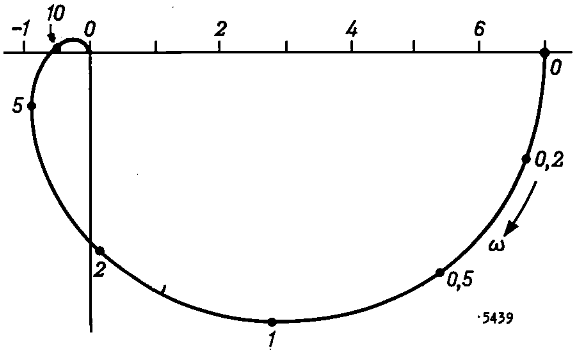


Fig. 15. Nyquist diagram pertaining to fig. 14.

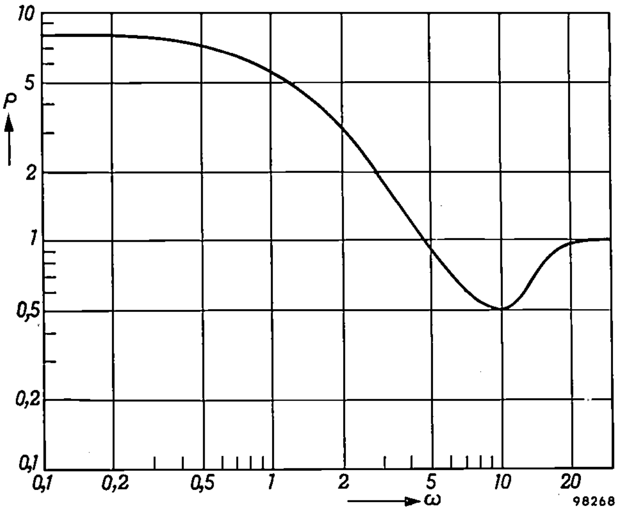


Fig. 16. The factor *P* (cf. fig. 4) as a function of ω, derived from fig. 15.

system is non-linear, or where the control loop cannot be opened. Considerable practical difficulties arise where long time constants are involved, because one then has to wait a very long time in each measurement before the switching transient has ended. In the latter case the object can sometimes be achieved by introducing a step-function disturbance and seeing how the system then behaves. From the step-function response thus obtained one can derive the frequency response curves, and from these find the values of the various time constants.

Summary. The theory of an automatic control system is formally similar to that of a negative-feedback amplifier. The behaviour of a feedback system, provided all its elements are linear, can be derived from the open-loop transfer function, $KG(\mathrm{j}\omega)$. This function is the quotient of the complex quantities corresponding to a sinusoidal input signal (angular frequency ω) and the resultant sinusoidal output signal. K is a constant factor. Disturbances are reduced by the factor $|1 + KG|$. The system is stable when the curve representing KG in the complex plane does not enclose the point (—1,0). The KG functions are calculated for 1) a tank with water flowing in and out, 2) the temperature control system for a glass-furnace feeder, and 3) the temperature control system for the subcritical suspension reactor at Arnhem. The elements of a control loop frequently have a transfer function of the form $KG(\mathrm{j}\omega) = K(1 + \mathrm{j}\omega\tau)^{-1}$.

A SIMPLE ANALOGUE COMPUTER FOR DETERMINING THE COLOUR POINT OF A LIGHT SOURCE

by B. van der WAAL *).

535.651.1

Determining the "colour point" of a light source calls for a series of simple but laborious computations. The computer described below performs this work rapidly and automatically; the CIE-distribution coefficients, which define the spectral sensitivity of the eye of a "standard observer", are built into the machine.

One of the principal characteristics of a light source is the colour of the light it emits. The colour can be objectively specified[1]) with the aid of three quantities X, Y and Z, defined as

$$\left.\begin{aligned} X &= \int S(\lambda)\,\bar{x}(\lambda)\,\mathrm{d}\lambda,\\ Y &= \int S(\lambda)\,\bar{y}(\lambda)\,\mathrm{d}\lambda,\\ Z &= \int S(\lambda)\,\bar{z}(\lambda)\,\mathrm{d}\lambda. \end{aligned}\right\} \quad . \; . \; . \; . \quad (1)$$

In these equations $S(\lambda)$ is the relative spectral energy distribution of the radiated light; $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are the CIE-distribution coefficients, which define the spectral sensitivity of the eye as laid down in 1931 for the "colorimetric standard observer" by the International Commission on Illumination (CIE). The coefficients are so chosen that the values of $\bar{y}(\lambda)$ are identical with those of the relative luminous efficiency $V(\lambda)$. The maximum values of these coefficients occur at wavelengths in the red, green and blue parts of the spectrum, respectively (see *fig. 1*).

It is however usually more convenient to specify the colour of a light source by the ratios of X, Y and Z to their sum:

$$\left.\begin{aligned} x &= \frac{X}{X+Y+Z},\\ y &= \frac{Y}{X+Y+Z},\\ z &= \frac{Z}{X+Y+Z}. \end{aligned}\right\} \quad . \; . \; . \; . \; . \quad (2)$$

As the sum $x + y + z$ is 1, two of these values are of course sufficient to specify a colour. Any colour can thus be represented by a point in an x-y diagram, as in the CIE *chromaticity diagram* or colour triangle (*fig. 2*). The colour points of monochromatic light lie on a horseshoe-shaped curve called the *spectrum locus*. The straight line joining the ends of the spectrum locus, the "purple line", represents the pure (saturated) purples, mixtures of red and violet radiations. In the centre lies the "white point" E, where x, y and z are equal. The other points may be regarded as colour points of a mixture of white light with a spectral colour or with a colour on the purple line.

*) Lighting Division, Eindhoven.

[1]) See W. de Groot and A. A. Kruithof, The colour triangle, Philips tech. Rev. 12, 137-144, 1950/51; F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. 19, 86-97, 1957/58; P. J. Bouma, Physical aspects of colour, Philips Technical Library, 1947.

The notation used in these articles differs somewhat from current usage, which conforms more or less with the official recommendations of the International Commission on Illumination; see "International Lighting Vocabulary", publication CIE 1-1-1957.

system is non-linear, or where the control loop cannot be opened. Considerable practical difficulties arise where long time constants are involved, because one then has to wait a very long time in each measurement before the switching transient has ended. In the latter case the object can sometimes be achieved by introducing a step-function disturbance and seeing how the system then behaves. From the step-function response thus obtained one can derive the frequency response curves, and from these find the values of the various time constants.

Summary. The theory of an automatic control system is formally similar to that of a negative-feedback amplifier. The behaviour of a feedback system, provided all its elements are linear, can be derived from the open-loop transfer function, $KG(\mathrm{j}\omega)$. This function is the quotient of the complex quantities corresponding to a sinusoidal input signal (angular frequency ω) and the resultant sinusoidal output signal. K is a constant factor. Disturbances are reduced by the factor $|1 + KG|$. The system is stable when the curve representing KG in the complex plane does not enclose the point $(-1,0)$. The KG functions are calculated for 1) a tank with water flowing in and out, 2) the temperature control system for a glass-furnace feeder, and 3) the temperature control system for the subcritical suspension reactor at Arnhem. The elements of a control loop frequently have a transfer function of the form $KG(\mathrm{j}\omega) = K(1 + \mathrm{j}\omega\tau)^{-1}$.

A SIMPLE ANALOGUE COMPUTER FOR DETERMINING THE COLOUR POINT OF A LIGHT SOURCE

by B. van der WAAL *).

535.651.1

Determining the "colour point" of a light source calls for a series of simple but laborious computations. The computer described below performs this work rapidly and automatically; the CIE-distribution coefficients, which define the spectral sensitivity of the eye of a "standard observer", are built into the machine.

One of the principal characteristics of a light source is the colour of the light it emits. The colour can be objectively specified[1]) with the aid of three quantities X, Y and Z, defined as

$$\left.\begin{aligned} X &= \int S(\lambda)\,\bar{x}(\lambda)\,\mathrm{d}\lambda,\\ Y &= \int S(\lambda)\,\bar{y}(\lambda)\,\mathrm{d}\lambda,\\ Z &= \int S(\lambda)\,\bar{z}(\lambda)\,\mathrm{d}\lambda. \end{aligned}\right\} \quad \ldots\ldots \quad (1)$$

In these equations $S(\lambda)$ is the relative spectral energy distribution of the radiated light; $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ are the CIE-distribution coefficients, which define the spectral sensitivity of the eye as laid down in 1931 for the "colorimetric standard observer" by the International Commission on Illumination (CIE). The coefficients are so chosen that the values of $\bar{y}(\lambda)$ are identical with those of the relative luminous efficiency $V(\lambda)$. The maximum values of these coefficients occur at wavelengths in the red, green and blue parts of the spectrum, respectively (see *fig. 1*).

It is however usually more convenient to specify the colour of a light source by the ratios of X, Y and Z to their sum:

$$\left.\begin{aligned} x &= \frac{X}{X+Y+Z},\\ y &= \frac{Y}{X+Y+Z},\\ z &= \frac{Z}{X+Y+Z}. \end{aligned}\right\} \quad \ldots\ldots \quad (2)$$

As the sum $x + y + z$ is 1, two of these values are of course sufficient to specify a colour. Any colour can thus be represented by a point in an x-y diagram, as in the CIE *chromaticity diagram* or colour triangle (*fig. 2*). The colour points of monochromatic light lie on a horseshoe-shaped curve called the *spectrum locus*. The straight line joining the ends of the spectrum locus, the "purple line", represents the pure (saturated) purples, mixtures of red and violet radiations. In the centre lies the "white point" E, where x, y and z are equal. The other points may be regarded as colour points of a mixture of white light with a spectral colour or with a colour on the purple line.

*) Lighting Division, Eindhoven.

[1]) See W. de Groot and A. A. Kruithof, The colour triangle, Philips tech. Rev. 12, 137-144, 1950/51; F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. 19, 86-97, 1957/58; P. J. Bouma, Physical aspects of colour, Philips Technical Library, 1947.
The notation used in these articles differs somewhat from current usage, which conforms more or less with the official recommendations of the International Commission on Illumination; see "International Lighting Vocabulary", publication CIE 1-1-1957.

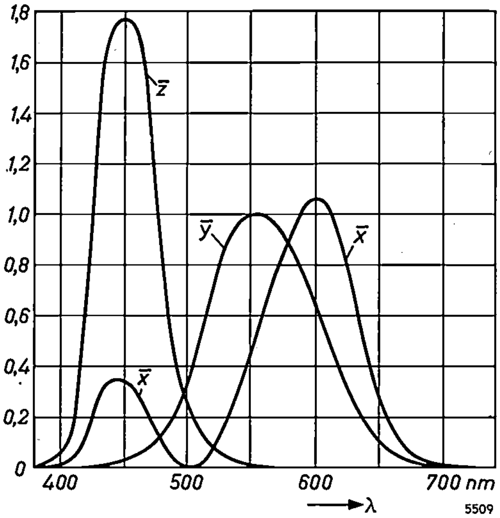


Fig. 1. The CIE-distribution coefficients $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$, giving the equal-energy distribution curves in the CIE system.

(A chromaticity chart showing the actual colours corresponding to various points can be found in the first article referred to in note [1]).

The determination of the colour point of a light source

The method of finding the colour point of a given light source follows directly from equations (1) and (2). The value of $S(\lambda)$ for the light is measured with a spectrophotometer at a series of wavelengths rising in steps of 10 nm from 385 to 735 nm. (1 nm = 10^{-9} m.) To find X the values thus measured are multiplied by the values of $\bar{x}(\lambda)$ corresponding to the same wavelengths and the products are then added. (One should really multiply by the width of the wavelength interval used in the measurement, but as we are only concerned with the ratios of X, Y and Z, we have omitted to do so here.) The values of Y and Z are calculated in the same way, after which x and y are determined by two division operations from equation (2). The computation of a single colour point thus involves about 100 multiplications, followed by three additions and two divisions.

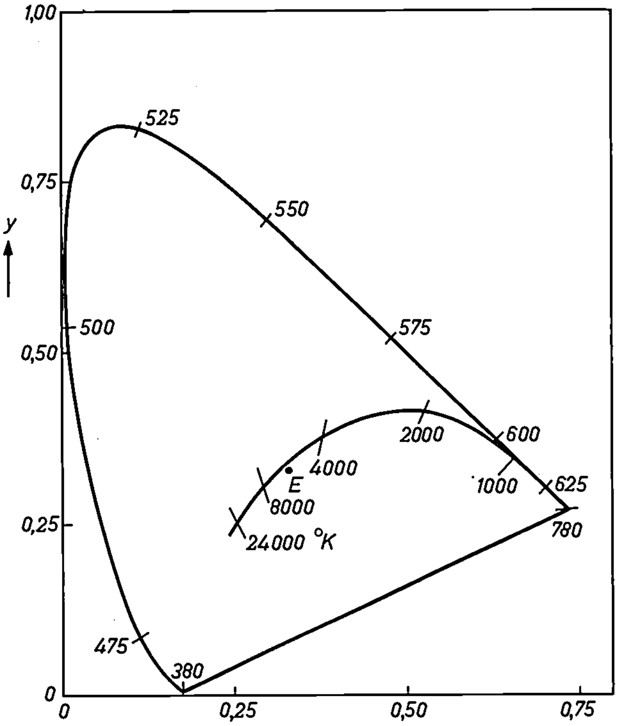


Fig. 2. The CIE chromaticity diagram in x-y coordinates. On the spectrum locus (curve of the spectral colours) the corresponding wavelengths are given in nm. (1 nm = 10^{-9} m.) E is the "white point". The curve inside the spectrum locus joins the colour points of a black-body radiator at different temperatures.

A simple analogue computer has been designed for performing these operations automatically. The principle is illustrated in *fig. 3*. For every wavelength λ_n at which $S(\lambda)$ is measured there is a circuit consisting of three resistors R_{xn}, R_{yn} and R_{zn}, whose respective values are inversely proportional to $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ at that wavelength. A voltage derived from an accumulator B_1 via a potentiometer R_{pn} is applied to each resistor, the value of the voltage being proportional to the measured value of $S(\lambda)$ (for this purpose the meter M, used as a voltmeter, can be connected to each potentiometer in turn). The currents through the resistors are then proportional to $S(\lambda_n)\bar{x}(\lambda_n)$, $S(\lambda_n)\bar{y}(\lambda_n)$ and $S(\lambda_n)\bar{z}(\lambda_n)$, respectively. The sum of the currents through all resistors R_{xn} is

$$\Sigma S(\lambda_n)^{-}(\lambda_n)$$

and is therefore proportional to X. The sums of the currents through R_{yn} and R_{zn} are similarly proportional to Y and Z respectively, the constant of proportionality being the same in all cases.

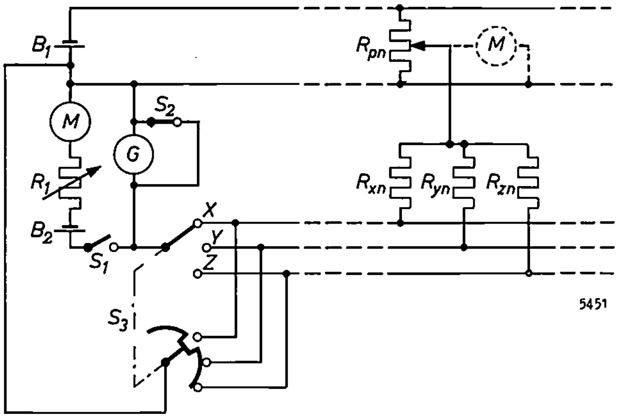


Fig. 3. Principle of the computer circuit for determining the colour point of a light source.

In measurements on tubular fluorescent lamps the energy at the wavelengths of the five mercury lines is determined in addition to $S(\lambda)$ at the wavelengths mentioned. The computer contains similar circuits for these wavelengths too, giving a total of 41.

When all 41 potentiometers have been set to the appropriate values, the three currents are determined by a null method, using a circuit consisting of a galvanometer G, a battery B_2, the meter M (now used as an ammeter) and a variable resistance R_1. The switch S_1 is closed, S_2 opened, and S_3 turned first to position X. The current through G is now made zero by adjusting R_1; X can then be read off from meter M. The values of Y and Z are found in the same way, with S_3 in positions Y and Z respectively.

Switch S_3 was given the form shown in fig. 3 to ensure that all resistors pass current both during the adjustment of the voltages and the measurement of the currents. If current were to flow through only one of the three sets of resistors (in which case only the top half of S_3 would be needed) the currents through the potentiometers would change when S_3 was switched over, and the voltages would have to be reset. With this arrangement, however, one setting serves for all three measurements. A photograph of the computer can be seen in *fig. 4*.

Other uses for the computer

The computer can be used for many other calculations that involve multiplication operations with the CIE-distribution coefficients or simply with the relative luminous efficiencies $V(\lambda)$. For example, to find the distribution of luminous flux over various wavelength regions, the measured values of the relative spectral energy distribution $S(\lambda)$ are multiplied by the values of $V(\lambda)$ for the wavelengths concerned, and the products for the various wavelengths are added.

The computer is also useful for "normalizing" the relative spectral energy distribution $S(\lambda)$. This is done when it is necessary for irradiation purposes (e.g. the irradiation of plants) to know the *absolute* value of the radiant power delivered to a surface by a light source in a particular range of wavelengths. A spectral energy distribution is then needed which gives the absolute instead of the relative values for the surface in question.

These values might be determined by the following procedure. The total radiant power P incident on the surface is measured and divided by $\int S(\lambda)\mathrm{d}\lambda$, an integral proportional to the total radiant power received by the photometer in the measurement of $S(\lambda)$. Multiplying the values of $S(\lambda)$ by this factor $P/\int S(\lambda)\mathrm{d}\lambda$ gives the absolute distribution required:

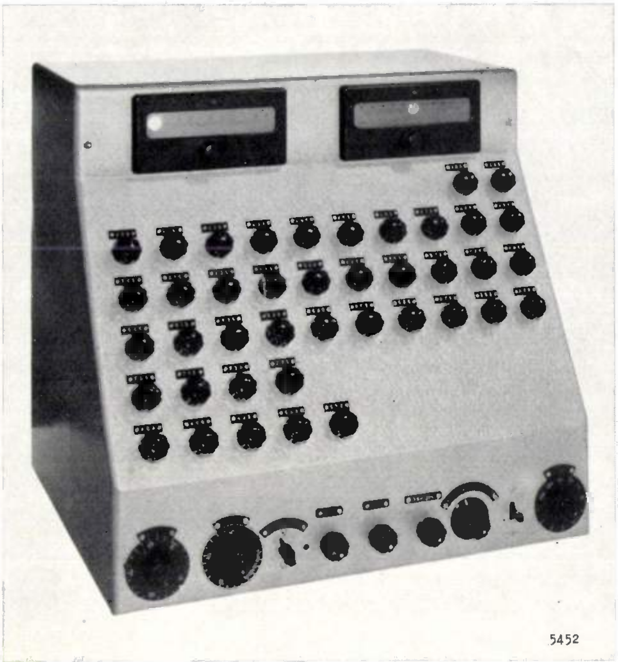


Fig. 4. Photograph of the colour-point computer. The sloping panel contains the potentiometer controls for presetting the measured energy values. Above left, the meter M shown in fig. 3, and right, the galvanometer G. The bottom row contains, among other things, the controls of the variable resistors used in the determination of X, Y and Z.

$$\frac{S(\lambda)}{\int S(\lambda)\mathrm{d}\lambda}\,P. \qquad (3)$$

$S(\lambda)/\int S(\lambda)\mathrm{d}\lambda$ is called the *normalized distribution*. To find from this the absolute distribution on an arbitrary surface the values are multiplied by the measured total radiant power through that surface.

The actual procedure is somewhat different. For practical reasons it is easier to measure the *luminous flux* on a surface than the radiant power. The normalized energy distribution is then found by dividing $S(\lambda)$ by

$$K\int S(\lambda)V(\lambda)\mathrm{d}\lambda. \qquad (4)$$

(This integral is proportional to the total luminous flux received by the photometer in measuring $S(\lambda)$; K is the luminous efficiency of radiation, which is about 680 lumen/W.) The normalized energy distribution found in this way is expressed in μW/lumen per 10 nm. The absolute distribution is found from it by multiplying by the measured luminous flux on the surface.

Since $V(\lambda)$ and $\bar{y}(\lambda)$ are identical, the integral in eq. (4) is equal to the integral Y in eq. (1), and can thus be determined with the computer in the same manner as the colour point. Now, however, the absolute value of Y must be known, and therefore the computer is calibrated by determining Y with it for an energy distribution in which this quantity is known. For this purpose we used the equi-energy spectrum, for which Y can be easily computed.

Summary. After a brief explanation of the CIE chromaticity chart, a simple analogue computer is described with which the colour point of a light source can be determined from the relative spectral energy distribution $S(\lambda)$, measured at a series of wavelengths rising in steps of 10 nm. These values are multiplied by the corresponding CIE-distribution coefficients $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$. The products $S(\lambda)\bar{x}(\lambda)$, $S(\lambda)\bar{y}(\lambda)$ and $S(\lambda)\bar{z}(\lambda)$ are all added to find the values of X, Y and Z which define the chromaticity of the light. For every wavelength at which $S(\lambda)$ is measured, the computer contains three resistors whose values are proportional to $1/\bar{x}(\lambda)$, $1/\bar{y}(\lambda)$ and $1/\bar{z}(\lambda)$. A voltage proportional to the measured value of $S(\lambda)$ is applied to these resistors. The currents then flowing through them are proportional to the products mentioned, and are added in the computer. Other operations which can be carried out with the aid of the computer include the normalizing of a relative spectral energy distribution.

ANALYSIS OF RESIDUAL GASES IN TELEVISION PICTURE TUBES WITH THE AID OF THE OMEGATRON

by J. van der WAAL *) and J. C. FRANCKEN *).

621.385.832.032.94:621.384.83

The omegatron, in a new design recently described in this journal, is a particularly useful mass spectrometer for the quantitative analysis of residual gases. Analyses of this kind have provided valuable information on processes inside a television picture tube during manufacture and in operation.

In a properly functioning television picture tube the total residual gas pressure should be less than 10^{-5} torr (1 torr = 1 mm Hg). The composition of the residual gas is also an important consideration, for one kind of gas can be more harmful than another.

The life of the cathode in a picture tube is particularly dependent on the types of gas present. The composition of the residual gas in an evacuated and sealed-off tube changes continuously, owing to the desorption and adsorption of gases by the various components. The composition of this residual gas can be determined with a mass spectrometer.

It is important that the spectrometer used should have a small volume and that it should release little gas compared with the picture tube. A mass spectrometer found especially suitable for this purpose is the omegatron.

The omegatron had already been used in 1953 for *qualitative* analysis of residual gases in sealed-off cathode-ray tubes [1]). As recently mentioned in this journal, an omegatron has now been designed which is also suitable for quantitative analyses [2]). Its use for analysing the residual gases in television picture tubes will be described in this article. The operation of the omegatron was discussed at length in the above-mentioned article. For convenience, we shall briefly recapitulate this discussion here.

Principle of the omegatron

A perspective sketch of the omegatron in its simplest form is shown in *fig. 1*. The narrow beam of electrons from the cathode K is parallel to a uniform magnetic field. The collision of electrons with gas molecules gives rise to ions. If the latter have a velocity component perpendicular to the direction of the magnetic field, they describe circular paths perpendicular to the static magnetic field. The radii of these paths are given by the equation:

$$r = \frac{m}{e}\,\frac{v_0}{B}, \qquad \ldots\ldots (1)$$

where m is the mass of the ion, e its charge, v_0 the component of the velocity perpendicular to the direction of the magnetic field, and B the magnetic induction. The angular frequency ω_c of the revolution of the ion is

$$\omega_c = \frac{e}{m} B, \qquad \ldots\ldots (2)$$

and the period of revolution is therefore independent of the velocity v_0. An RF field $\hat{E}_{hf} \sin \omega_0 t$ is applied at right angles to the magnetic field. If the angular frequency ω_0 of that field is made equal to the angular frequency ω_c of a given kind of ion, the result

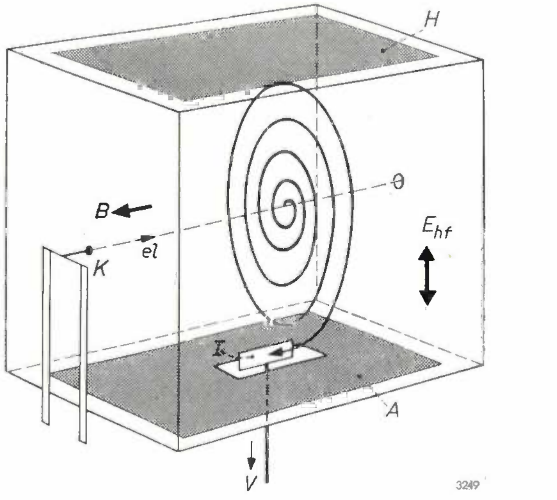


Fig. 1. Perspective sketch of a simplified omegatron. B indicates the direction of the magnetic induction, E_{hf} the direction of the RF field. The figure illustrates the spiral path described by an ion. K cathode, which emits the ionizing beam of electrons *el*. A and H are the electrodes to which the RF voltage is applied. I ion collector. V connection to amplifier.

*) Electron Tubes Division, Eindhoven.

[1]) J. Peper, Philips tech. Rev. **19**, 218-220, 1957/58.

[2]) A. Klopfer and W. Schmidt, Philips tech. Rev. **22**, 195-206, 1960/61 (No. 6).

is a condition of resonance. The resonating ions then describe spiral paths of uniformly increasing radius. They are caught on a suitably placed collector electrode, and the ion current, i.e. the number of ions collected per unit time, is detected with the aid of a sensitive amplifier.

Ions having a different mass but the same charge as the resonating ions, i.e. ions with a different value of e/m, are periodically accelerated and decelerated by the RF field. The radius of the orbits described by these ions alternately increases and decreases, but the maximum value generally remains smaller than the distance between the point of origin of the ions and the collector. Such ions do not therefore contribute to the measured current. To discriminate in this way between two kinds of ions, their mass difference Δm must exceed a specific minimum value. The ratio $m/\Delta m$ is called the resolution of the omegatron.

The pressure of each component of the gas mixture can be found by varying the frequency ω_0 of the RF field and measuring the ion current for the various mass numbers. If the relationship between ion current and pressure is reproducible, the partial gas pressure for any mass number can be determined from the ion current. This relationship will be reproducible when all resonating ions reach the collector electrode. That is only the case when the total gas pressure in the omegatron is less than 10^{-5} torr; otherwise ions are lost by collision with gas molecules. Further, ions in resonance should not strike other electrodes. This can be prevented by fitting side plates and appropriately choosing the potential distribution inside the omegatron, as described in the article quoted [2]). A photograph of the omegatron used for this investigation is shown in *fig. 2*.

Experimental set-up

In all the experiments described here, use was made of a permanent magnet having a magnetic induction of 0.44 Wb/m². The RF voltage was 1 volt r.m.s. This voltage was taken from a Philips signal generator, type GM 2653. The electron current of the omegatron was stabilized. After amplification, the ion current was measured with a highly stable DC amplifier with negative feedback [3]). The smallest measurable current was 10^{-14} A.

The resolution $m/\Delta m$ of the omegatron depends markedly on the magnitude of the magnetic induction B, as appears from the equation:

$$\frac{m}{\Delta m} = \frac{e\,B^2 S_0}{2\,\hat{E}_{\mathrm{hf}}\,m}, \qquad (3)$$

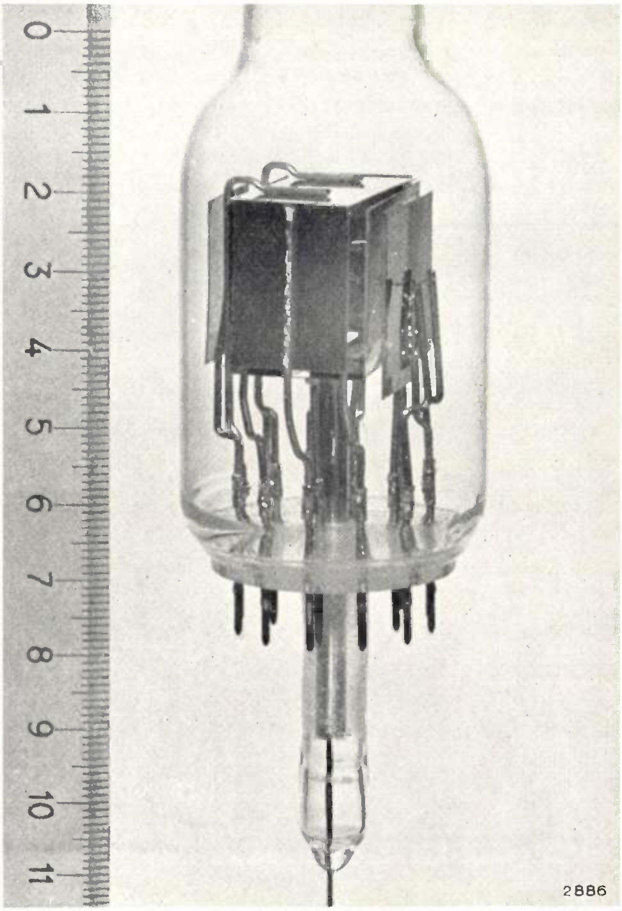


Fig. 2. The omegatron with side plates in a glass envelope. The cathode is on the right, and on the extreme left is the collector for the electrons which have traversed the omegatron. The ions formed in the space inside the omegatron move to the ion collector. This is connected to the central metal pin fused into the glass tube and projecting below it.

where S_0 is the distance from the point where the ion originated to the ion collector and $\hat{E}_{\mathrm{hf}}$ is the amplitude of the RF field.

The resolution of the omegatron can easily be derived from the equation given in the literature for the radius r of the ion orbits [4]):

$$r = \frac{\hat{E}_{\mathrm{hf}}}{B\varepsilon} \sin \frac{\varepsilon}{2} t, \qquad (4)$$

where ε is the difference between the resonance frequency ω_c of an ion having mass m and charge e and the frequency ω_0 of the applied RF field: $\varepsilon = |\omega_0 - \omega_c|$. It is assumed that $\varepsilon \ll \omega_c$. The ions cannot pass beyond a circle of radius

$$r_{\mathrm{max}} = \hat{E}_{\mathrm{hf}}/B\varepsilon\,. \qquad (5)$$

If $\varepsilon \leqslant \hat{E}_{\mathrm{hf}}/BS_0$, then $r_{\mathrm{max}} \geqslant S_0$ and ions of resonance frequency ω_0 reach the ion collector even though the frequency ω_0 of the RF field differs from ω_c. If $\varepsilon > \hat{E}_{\mathrm{hf}}/BS_0$, $r_{\mathrm{max}} < S_0$ and these ions do not arrive on the collector. It follows, then, that ions having a resonance frequency ω_c will continue to fall on

[3]) J. Pelchowitch, Philips Res. Repts **9**, 1, 1954.

[4]) H. Sommer, H. A. Thomas and J. A. Hipple, Phys. Rev. **82**, 697, 1951.

the collector while the frequency of the RF field is varied between $\omega_0 + \varepsilon$ and $\omega_0 - \varepsilon$. It will then only just be possible to discriminate completely between a mass m and a mass $m + \Delta m$, provided the difference $\Delta\omega_c$ between the resonance frequencies of the two masses is equal to 2ε. We can thus write the resolution as

$$\frac{m}{\Delta m} = \frac{\omega_c}{\Delta\omega_c} = \frac{\omega_c}{2\varepsilon} = \frac{eB^2S_0}{2\,\hat{E}_{\mathrm{hf}}\,m} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

In the experimental arrangement used the resolution is theoretically (i.e. according to equation (6)) such that two kinds of ions whose mass difference is $\Delta m = 1$ can be completely separated if the mass number m is not greater than 34. In actual fact this was found to be the case up to $m = 23$. In order to improve the resolution, we have recently obtained a permanent magnet whose magnetic induction B is 0.92 Wb/m^2. *Fig. 3* shows a photograph of this magnet together with the other equipment used. The picture tube under investigation is held in an adjustable yoke. The distance between the poles of the magnet is 5.0 cm, and the diameter of the pole pieces is 13.0 cm. At a distance of 2.0 cm from the centre the variation in the magnetic induction is only 0.1%. Theoretically the resolution is now such that two kinds of ions with $\Delta m = 1$ can be completely separated if m is not greater than 73. Experimentally, effective separation is found up to $m = 49$, which means that for example propane gas (C_3H_8, mass number 44) can be fully distinguished from its "neighbours" using the large

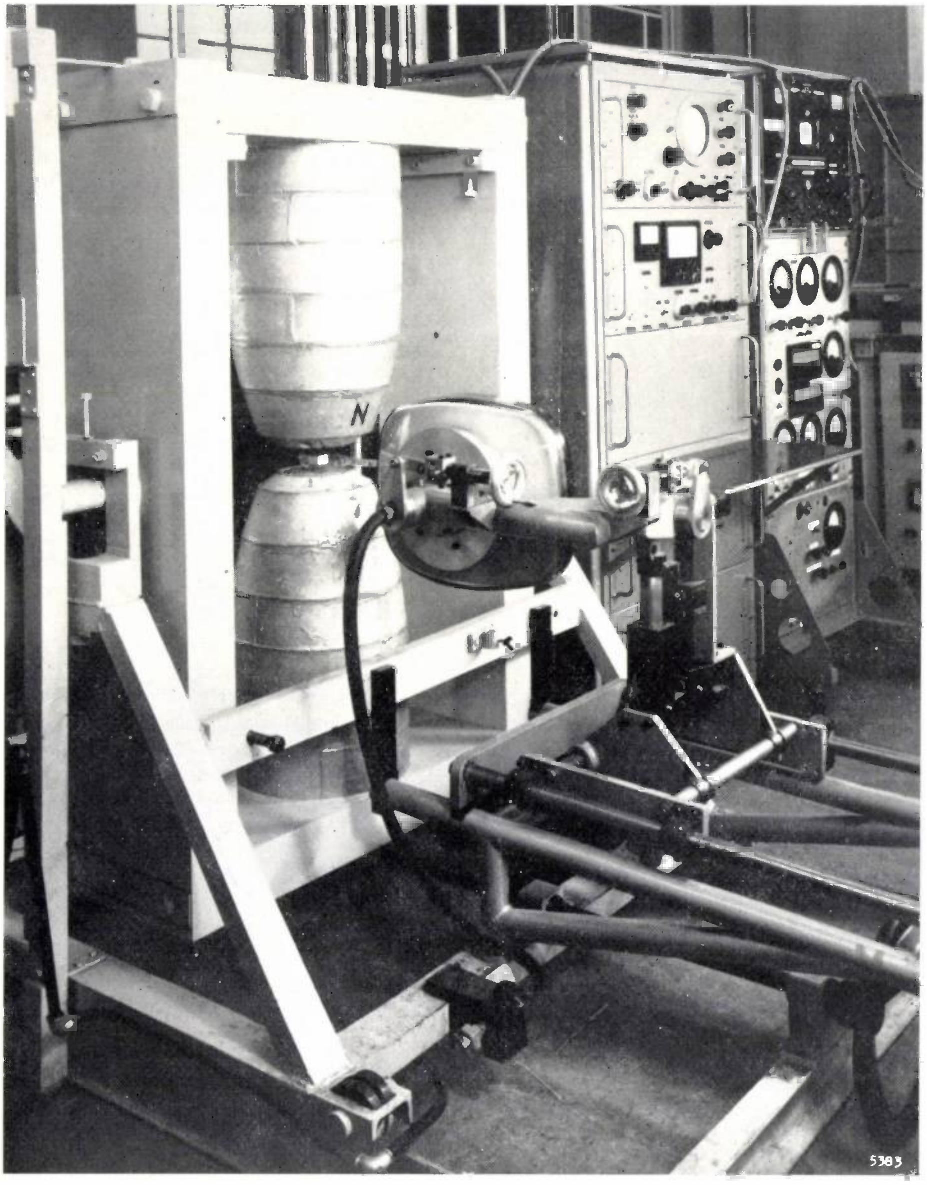


Fig. 3. Experimental set-up, consisting of a permanent magnet (magnetic induction B = 0.92 Wb/m^2), a DC amplifier for the ion current, a Philips signal generator, type GM 2653, which supplies the RF voltage, and ancillary equipment for the other voltages needed for the omegatron. The television picture tube under investigation is connected to an omegatron located between the pole pieces of the magnet.

magnet; but not when using the small magnet. *Fig. 4* shows the spectrum of a mixture of propane and a trace of methane. This spectrum was recorded using the magnet with induction $B = 0.44$ Wb/m².

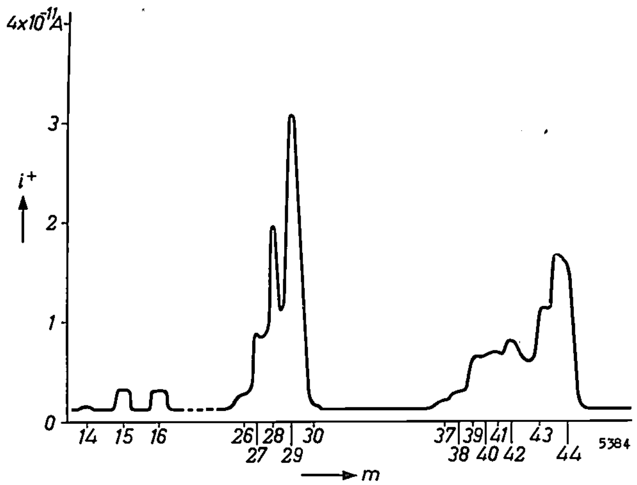


Fig. 4. Mass spectrogram of residual gases in a television picture tube, showing ion current i^+ of omegatron as a function of the preset mass number m. The tube contains propane C_3H_8 and a small quantity of methane CH_4. The RF voltage was 1 V. The magnet used for this recording has a magnetic induction $B = 0.44$ Wb/m².

Fig. 5 shows the same spectrum, but now with an induction $B = 0.92$ Wb/m². The improved resolution is particularly noticeable for the mass numbers 43 and 44.

Gas analyses on picture tubes in various stages of manufacture

In the manufacture of television picture tubes it is important to know not only the composition of the residual gases in the finished tube, but also which parts of the tube contribute most to this residue, and what gases they contribute. This information is obtained by studying unfinished tubes, i.e. tubes which have not yet been fitted with a fluorescent screen and/or other components. Analyses are also carried out after the various operations following seal-off.

In all cases an omegatron is fixed to the side of the tube immediately below the screen. First of all we investigate the glass bulb of a picture tube which contains only an unevaporated barium getter, but otherwise none of the "normal" components, such as fluorescent screen and electron gun. In addition to the omegatron a Philips ionization gauge, type 53 EM [5]), is sealed to the tube. The only operation undergone by the tube is evacuation for two hours on an oil-diffusion pump. During this process the tube is heated for 20 minutes at 400 °C and allowed to cool to about 100 °C, after which it is sealed off. It is assumed that a more or less stationary state sets in 20 hours after seal-off. The gas composition is then measured with the omegatron. The gases found are hydrogen, methane, water vapour, carbon monoxide and carbon dioxide. Next, the getter is evaporated and a further measurement is made. The chemically active gases have now been absorbed by the getter, and in their place the tube contains a great deal of hydrocarbons, probably due to reactions between water vapour and carbon in the getter. The hydrocarbons are not taken up by the getter. A certain amount of argon, absorbed by the barium getter during its manufacture, is also found. The ionization gauge is now switched on for 15 minutes. The pumping action of this gauge depends on two effects. In the first place, hydrocarbons are decomposed at the hot cathode, which is at a temperature between 1500° and 2000°C. The fragments react with the getter, thereby reducing the pressure. Secondly, the electron current of 2.5 mA produces ions in the gas, which are captured by the collector and thus removed from the gas. This too results in a pressure drop. At the end of the 15 minutes the ion gauge is switched off and the tube is left for 14 days without anything being done to it. At the end of this period only a very slight increase in pressure is found.

In *fig. 6* the results of gas analyses done at four different times are compared: *a*) before and *b*) after evaporation of the getter, *c*) after 15 minutes of drawing current in the ionization gauge, and *d*) after 14 days. In the right-hand column the getter had been preheated (subjected to previous degassing) in the left-hand column it had not. The initial pres-

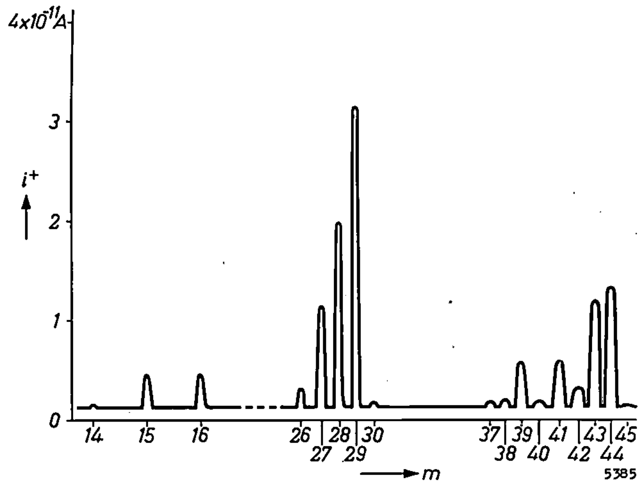


Fig. 5. The same mass spectrum as in fig. 4, but now using a magnet with $B = 0.92$ Wb/m². The voltages applied to the omegatron were the same as in the previous analysis. The resolution is substantially improved.

[5]) E. Bouwmeester and N. Warmoltz, Philips tech. Rev. 17, 121-125, 1955/56.

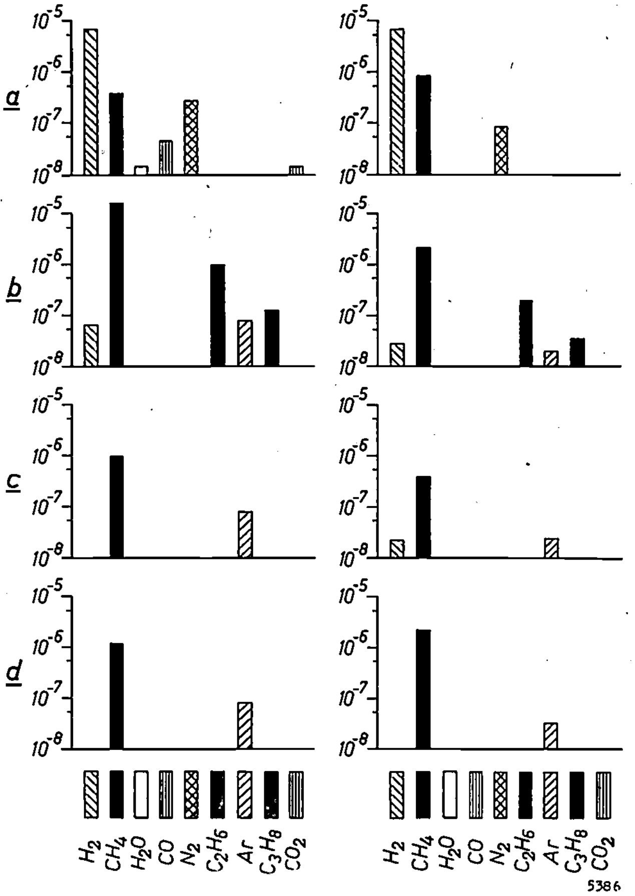


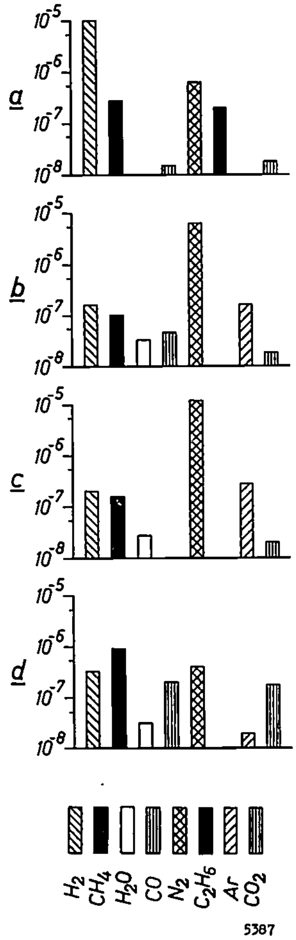
Fig. 6. Composition of the residual gases in unworked bulbs of television tubes. *a*) Before evaporation of barium getter; *b*) after evaporation of getter; *c*) after Philips 53 EM ionization gauge had drawn a current of 2.5 mA for 15 minutes; *d*) 14 days later. In the left column the getter had not been preheated; in the right column it had. Scale for partial pressures in torrs (mm Hg), likewise in following figures.

sure is lower when the getter has been preheated, but after 14 days no difference is to be found.

Fig. 7 shows a comparison between four different bulbs analysed at four different phases in the production process, without the getter having been evaporated. The first bulb (*a*), as in the previous case, contains no screen or components. The second (*b*) has been provided with a fluorescent screen, and in the third (*c*) the screen has been given the conductive coating of aluminium. The last bulb (*d*) also contains the electron gun. The cathode in the gun was degassed and activated when the bulb was still on the oil-diffusion pump. All four bulbs were subjected to the pumping process described above. Twenty hours after seal-off the gas composition was determined with the omegatron. It is noticeable that the hydrogen pressure is lower and the nitrogen and argon pressures higher in the second bulb. The hydrogen has probably been taken up by the screen, and the higher nitrogen and argon pressure is presumably due to the uptake of these gases from the atmosphere during the production of the screen. Bulb *c* differs hardly at all from bulb *b*, but the finished tube with the electron gun has a lower nitrogen pressure and a higher carbon monoxide pressure. The carbon dioxide pressure has also gone up, probably as a result of the gas given off by the oxide cathode.

After evaporation of the getter in the finished tube the gas composition changes radically, and in most tubes the residual gases consist solely of methane and argon. In some tubes, however, other hydrocarbons are found, but with a much lower partial pressure than methane.

Fig. 8 shows how the composition of the residual gases in a finished picture tube changes during operation. Measurements were started twenty hours



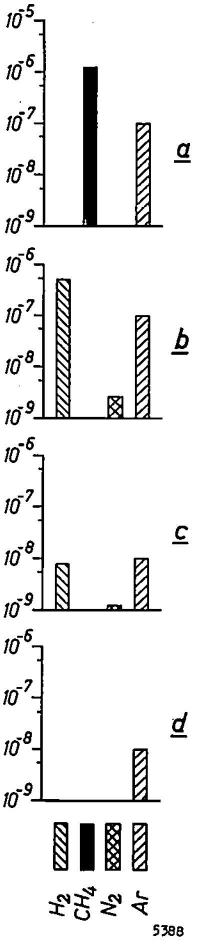


Fig. 7

Fig. 8

Fig. 7. Composition of residual gases in picture tubes in various stages of production. *a*) Unworked glass bulb; *b*) bulb with fluorescent screen; *c*) bulb with fluorescent screen and conductive aluminium coating; *d*) tube complete with electron gun. The getter was not evaporated in any of the tubes.

Fig. 8. Composition of residual gases in a finished picture tube at four different times. *a*) 20 hours after seal-off, with the cathode cold; very little change occurs when the cathode is raised to working temperature. *b*) during the first hour of screen bombardment; *c*) after 250 hours of screen bombardment, measured during operation; *d*) immediately after switching the tube off.

after seal-off (*a*). While the cathode is heating up, very little change is found, but as soon as the screen is scanned by an electron beam, the nitrogen and hydrogen pressures increase and the hydrocarbons vanish (*b*). After scanning of the screen for 250 hours with an electron current of 275 μA — the electron energy being 16 keV — the hydrogen and nitrogen pressures are found to have decreased (*c*). After scanning is stopped, the only residual gas which can be detected is argon (*d*). The pumping speed of the getter film is apparently not high enough to deal initially with the large production of hydrogen and nitrogen. We shall return later to the very sharp drop in the quantity of hydrocarbons at the beginning of the scanning of the screen.

Breaking open the picture tubes under vacuum

We have seen how the composition of the residual gases in picture tubes can be determined by connecting an omegatron to the tube at the outset. It may happen, however, that a tube from normal production is not working properly because the gas pressure is too high. In such a case we want to be able to attach the omegatron to the rejected tube to ascertain the gas responsible for the fault.

This is done in the following way. First of all an area 5 mm in diameter in the neck of the bulb near the electron gun is ground down with an ultrasonic drill to a thickness of only 0.5 mm [6]). A thin glass tube is cemented round the weakened area with epoxy resin. The glass tube is connected to a pump system incorporating an omegatron, and a metal weight is introduced into the tube. The resin is hardened by heating at about 100 °C for half an hour, resulting in a strong vacuum-tight joint. At a suitable moment the weight, which has a sharp steel point and can be moved with a magnet, is made to break the thin glass wall, thus establishing communication between picture tube and omegatron. *Fig. 9* shows a diagram of the whole arrangement. The part shown enclosed in the dotted lines is degassed for 15 hours at 400 °C before being connected to the rest of the system. At the same time the glass connecting tube is heated by hot wires to 200 °C. During this process the epoxy resin does not get hotter than 100 °C, and the picture tube as a whole remains at room temperature. After 15 hours on the diffusion pump and outgassing, the pressure is reduced to a value between 1×10^{-8} and 5×10^{-8} torr. The residual gases are then hydrogen, nitrogen, carbon dioxide and sometimes a small amount of water vapour.

After the system has been cooled to room temperature, the high-vacuum valve K_1 is closed. This is a glass valve containing no grease and is closed by means of a magnet. Once the glass membrane between the system and the picture tube has been broken, the gas content of the tube can be analysed with the omegatron.

For this method to have any value the gas desorption of the epoxy resin must obviously be low. In the system without epoxy resin obtained by sealing off the glass tube above the connection to the picture tube, a pressure of 10^{-9} torr can

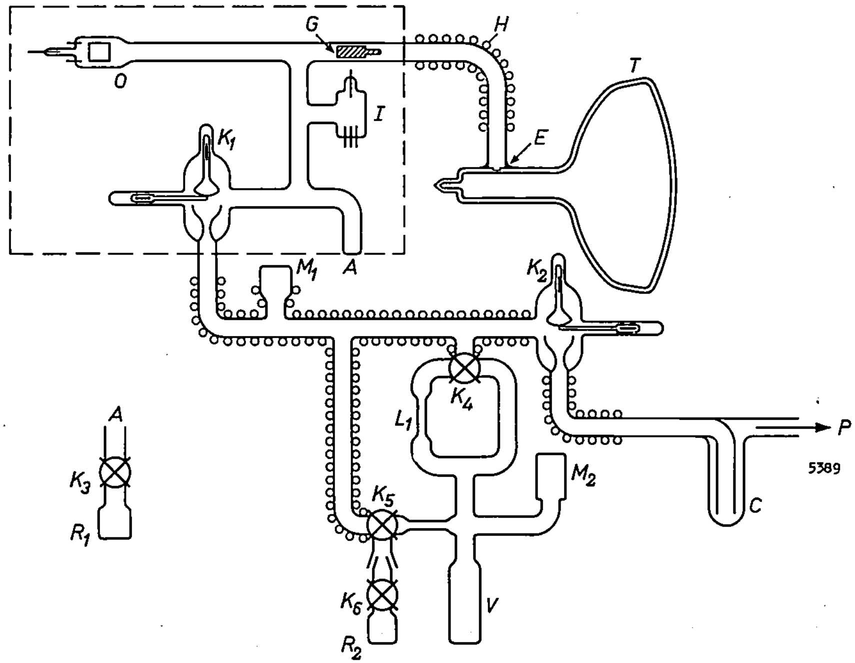


Fig. 9. System for breaking open a television tube and analysing the residual gases. A glass tube is cemented with epoxy resin *E* round a spot in the neck of a picture tube *T* where the wall has been ground very thin. The connecting tube is degassed by a heater coil *H*. The part of the system inside the broken line can be outgassed in an oven at 400 °C. This part comprises an omegatron *O*, an ionization gauge *I*, a weight *G* and a glass high-vacuum valve K_1. M_1 Penning gauge. K_2 second high-vacuum tap. K_4-K_6 grease cocks. M_2 Pirani (hot-wire) gauge. R_2 reservoir containing carbon monoxide. *V* expansion volume of 5.2 l. *C* liquid-nitrogen trap. L_1 capillary. The mercury diffusion pump is connected to *P*. The control valve K_3 and methane-filled bulb R_1 can be connected to point *A*.

[6]) R. H. Collins and J. C. Turnbull, Vacuum **10**, 27, 1960 (No. 1/2).

easily be achieved. The residual gas is then primarily nitrogen. The lower pressure must therefore be attributed entirely to the absence of gas given off by the epoxy resin. Provided the pressure in the picture tube is much higher than 5×10^{-8} torr, which is always the case in practice, this method is therefore permissible.

Reasonable agreement is found between the composition of the residual gases in the picture tube before it is broken open (measured with a second omegatron fixed directly to the bulb) and the composition found after the tube is broken open. The only difference found after breaking open the tube is that the pressures of the water vapour and carbon dioxide are lower, owing to absorption on the well-degassed components of the pump system.

Fig. 10 shows the composition of the residual gases in two picture tubes that were rejected on the grounds of excessive gas pressure and were broken open in the manner described. These tubes were found to contain hydrocarbons, water vapour and argon. The hydrocarbons were probably due to reactions between water vapour and carbon compounds in the getter film [7]. In other picture tubes the excessive pressure was sometimes found to be attributable to argon. The latter tubes showed no leakage, as was proved by storing them for a considerable period in plastic envelopes filled with helium. At the end of this time the omegatron registered no higher helium pressure. It is probable that air entered these tubes *during the pumping process.* The active gases would then be taken up by the getter film, but not the inert gas argon.

The remainder of the system represented in fig. 9 serves for investigating the gettering capacity of the barium getter in the picture tube [8]. For this measurement K_2 is closed and K_1 opened. The carbon monoxide is admitted from a bulb having a known volume V of 5.2 l. The pressure p_2 in V is measured with a Pirani (hot-wire) gauge, which is specially calibrated for carbon monoxide. In this way a known quantity of gas can be admitted to the getter film. The pumping speed S of the getter can be found from the equation

$$S = L \frac{p_2 - p_1}{p_1}, \quad \ldots\ldots \quad (7)$$

where p_1 is the carbon-monoxide pressure in the picture tube and L the conductance between the bulb and the picture tube. The conductance is almost entirely determined by the capillary L_1. The pressure p_1 in the picture tube can be measured by using the electron gun of the tube itself as an ionization gauge [9]. The procedure is illustrated in *fig. 11.*

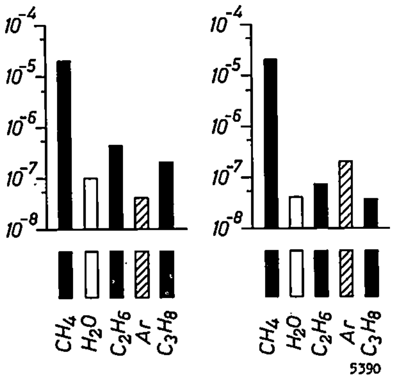


Fig. 10. Composition of residual gases in two picture tubes from normal production, rejected because of their excessive gas pressure.

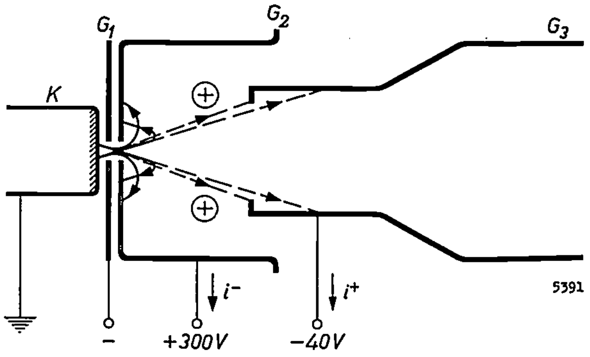


Fig. 11. Pressure measurement in a picture tube using the electron gun itself as an ionization gauge. K cathode at earth potential. G_1 grid negative to earth. G_2 positive grid which collects the electrons. The ion current, which is a measure of the pressure, is measured with G_3.

The cathode is kept at earth potential. With the aid of the first negative grid the electron beam is directed on to the second grid, which has a potential of 300 V positive to the cathode. All other electrodes are given a potential of 40 V negative to the cathode. These electrodes are used to determine the ion current, which in this case is a measure of the pressure.

The total amount of carbon monoxide that a getter film in a picture tube can take up depends on the evaporated quantity of getter material. A representative value is 1 torr-litre when 45 mg of barium is evaporated in a 21″ picture tube.

Other kinds of electron tubes that can be broken open and analysed in this way are oscilloscope tubes, transmitting tubes and television pick-up tubes, such as vidicons and image orthicons. Generally

[7] F. H. R. Almer, H. J. R. Perdijk and P. G. van Zanten, Advances in vacuum science and technology, Proc. 1st int. congress on vacuum techniques, Namur, 10th-13th June 1958, Part 2, p. 676, Pergamon Press, Oxford 1960.

[8] J. J. B. Fransen and H. J. R. Perdijk, Philips tech. Rev. **19**, 290, 1957/58.

[9] J. de Gier, Ingenieur **65**, O 36, 1953.

speaking it can be said that any electron tube can be investigated in this way, provided the quantity of gas in the tube is greater than 10^{-7} torr l.

An example of the analyses of an image orthicon can be seen in *fig. 12.* In this case the residual gases were analysed four months after manufacttre, and were found to consist of methane, ethane and argon from the barium getter. Also found was a partial pressure of 2×10^{-6} torr for helium, which had diffused through the glass wall from the outside atmosphere. The quantity Q of helium that diffuses per unit time through a surface of area A and thickness d, at a pressure difference across the wall Δp, is given by:

$$Q = \frac{KA\Delta p}{d} \quad . \; . \; . \; . \; . \; . \; . \quad (8)$$

Here K is the permeability of helium of the borosilicate glass used, given by [10]):

$$K = 1.5\times10^{-13} \frac{\text{torr-litre}\times\text{mm (thickness glass wall)}}{\text{cm}^2\text{ (surface)}\times\text{sec}\times\text{torr (He pressure diff.)}}.$$

In the case of the image orthicon, A was 600 cm^2, Δp was 3×10^{-3} torr, and the average wall thickness d 2.2 mm. The volume of the orthicon was 740 cm^3, giving a calculated pressure rise after four months of 1.9×10^{-6} torr. This is in excellent agreement with the measured helium pressure of 2×10^{-6} torr.

Decomposition of hydrocarbons in a picture tube during operation

When the electron gun of a picture tube is used as an ionization gauge in the manner described above, the partial pressures of the hydrocarbons measured with the omegatron are found to decrease very rapidly. This is due to the disintegration of hydrocarbons on the hot cathode, which is at a temperature of 800 °C, and to the subsequent uptake of the fragments in the getter film. Furthermore, the collision of electrons with gas molecules gives rise to ions which are captured by the negative electrodes. Both mechanisms contribute to the reduced partial pressures of the hydrocarbons. A plot of the decrease measured in a picture tube is shown in *fig. 13.* The heavier hydrocarbon molecules decompose more rapidly than the light ones, so that the partial pressures of the heavier molecules decrease faster than that of methane molecules, for example, which are light. The pumping speed for the latter can be found from the formula:

$$V\frac{\mathrm{d}p}{\mathrm{d}t} = -Sp\,, \quad . \; . \; . \; . \; . \; . \quad (9)$$

where V is the volume, p the pressure, $\mathrm{d}p/\mathrm{d}t$ the pressure change per unit time and S the pumping speed. After integration we find:

$$\ln\frac{p}{p_0} = -\frac{S}{V}t, \quad . \; . \; . \; . \; . \; . \quad (10)$$

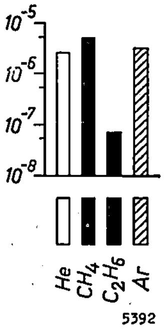


Fig. 12. Composition of the residual gases in an image orthicon four months after manufacture. Note the high helium pressure.

where p_0 is the pressure at $t = 0$. The pumping speed can thus be found from the slope of the curve obtained by plotting the logarithm of the pressure against time. The pumping speed found in this way for methane is roughly 0.004 l/sec.

The pumping speed for hydrocarbons is much higher during normal operation of the tube, when a screen is bombarded by electrons of energy 16 keV. The exceptionally high pumping speed cannot be explained by the longer path of the electrons through the tube and the consequent larger number of ions formed. The production of ions does not even increase linearly with the length of the electron

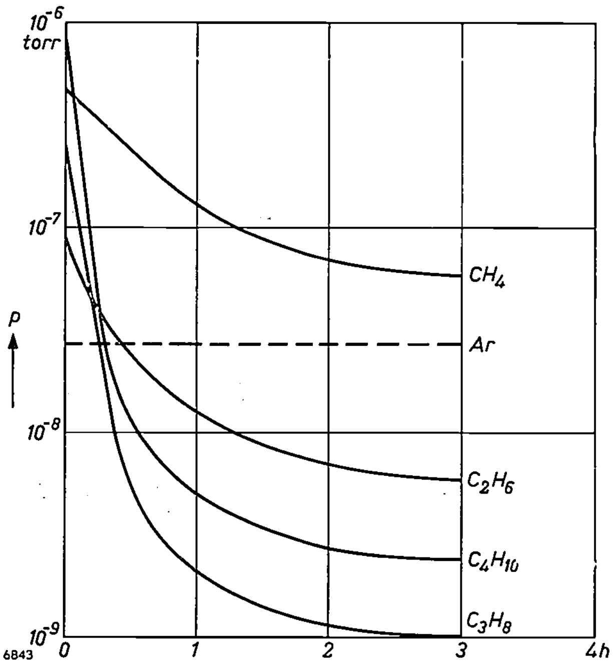


Fig. 13. Decrease in the partial pressure of hydrocarbons in a picture tube. The logarithm of the pressure is plotted versus time in hours. Electron current 400 μA. The electron gun of the tube was used as an ionization gauge.

[10]) A. Klopfer and W. Ermrich, Vacuum 10, 128, 1960 (No. 1/2).

trajectories, since the electrons now have a higher energy and therefore their ionization probability is less than when the electron gun is used as an ionization gauge. The conclusion, then, is that the high pumping speed for hydrocarbons is attributable to the *fluorescent screen.* This is clearly demonstrated in *fig. 14*, where the logarithm of the methane pressure is plotted as a function of time from the beginning of the screen bombardment. Curve 1 relates to a finished picture tube, and curve 2 to a tube without a fluorescent screen but containing the conductive aluminium coating. In the latter tube the pumping speed for methane is again 0.004 l/sec, compared with 2 l/sec in the finished tube. A similar comparison of two tubes, one of which had a silicate binder added to the fluorescent screen and the other not, revealed that the silicate binder is responsible for the high pumping speed for hydrocarbons. The form in which the silicate binder is present after preparation of the screen shows much resemblance to silica gel, a substance possessing a large internal surface and capable of adsorbing large amounts of gas.

The fact that the fluorescent screen is responsible for the pumping effect is again clearly apparent from the effect of the accelerating voltage. The effect is not perceptible until 1.5 kV, i.e. the same potential at which the electrons are able to penetrate through the aluminium layer to the phosphor coating. Up to 4 kV the pumping speed continues to rise, and above this value it is independent of the voltage.

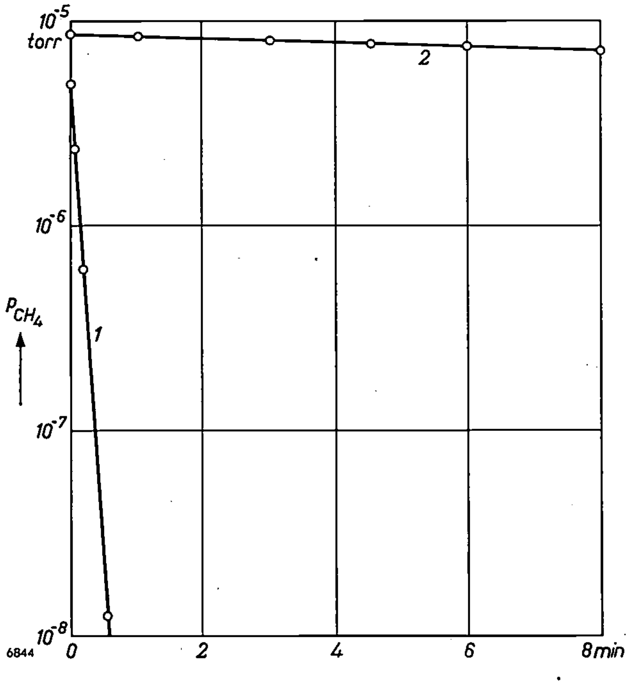


Fig. 14. Adsorption of methane during bombardment of the screen in a finished picture tube (*1*) and in a tube without fluorescent screen (*2*). Constant pumping speeds are inferred from the slopes of the two curves.

Various experiments were done to investigate the mechanism of this pumping effect. A finished picture tube was connected by a glass tube to the pumping system shown in fig. 9. The connecting tube was temporarily closed off by a thin-walled glass bulb. The bulb R_1, filled with methane, was connected to the pump system at A via a control valve K_3. The part of the system shown within dotted lines in the figure was degassed for 15 hours at 400 °C. The picture tube was kept at room temperature. At the end of this period the total pressure measured with the omegatron was roughly 10^{-8} torr. The residual gases consisted almost entirely of hydrogen, nitrogen and carbon dioxide. Communication with the picture tube was then established by breaking the thin-walled bulb with a steel ball. The high-vacuum valve K_1 was closed, after which the picture tube was given a methane pressure of about 5×10^{-6} torr by adjustment of the control valve K_3. The omegatron was adjusted so as to bring ions having a mass 16 into resonance, making it possible to determine the methane pressure continuously throughout the experiment. During this time the electron beam scanned a variable area on the fluorescent screen. *Fig. 15* shows the logarithm of the methane pressure plotted versus time from the moment of switching on the tube, for five different scanning areas. The current density of the electron beam at the screen was at all times 1 μA/cm^2, the accelerating voltage 16 kV. Whether the barium getter was evaporated or not had no effect on the results.

Similar experimental runs were carried out for different current densities at the screen. They all gave results resembling those in fig. 15, except that the pressure decreased more rapidly the higher the current density.

The characteristic shape of the curves in fig. 15 may be explained qualitatively as follows. We suppose that the electrons, in bombarding the screen, give rise to active centres in the silicate binder of the fluorescent coating, as a result of which the incident methane molecules are trapped and at the same time hydrogen is desorbed. During the initial bombardment of the screen the number of centres increases and so does the pumping speed. Assuming that the number of centres reaches a maximum after a certain time, it follows that the pumping speed must then become constant. This is the case in the linear portion of the curve in fig. 15. After some considerable time a state of equilibrium is

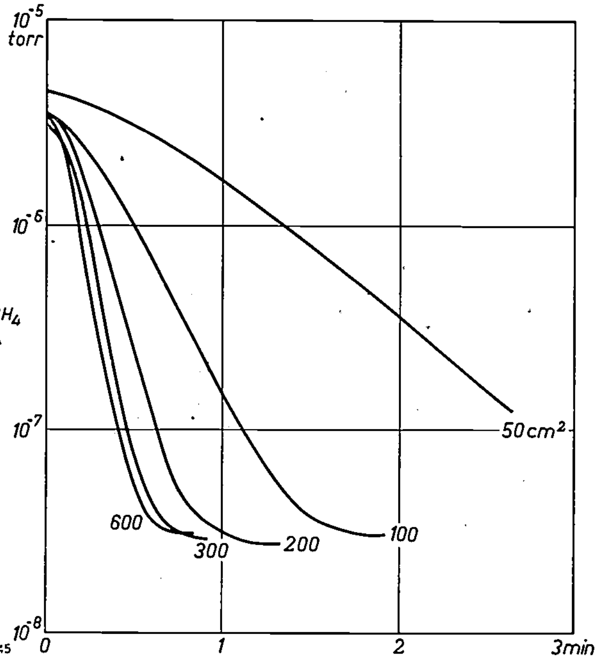


Fig. 15. Variation of methane pressure in a picture tube. The logarithm of the pressure is plotted as a function of time in minutes for five different values of the scanned area. The current density of the electron beam was kept constant ($j = 1\ \mu$A/cm² on the screen). $V_a = 16$ kV.

reached in which the pressure remains constant. In this state the adsorption of methane is equal to its desorption.

The mechanism may be described in somewhat greater detail as follows. Let n_0 be the maximum number of centres formed per unit area and N the number of centres already produced on the scanned area A. The number of centres that can still be produced is then $An_0 - N$. Experiment shows that the rate of change of $An_0 - N$ is proportional to the current density j at the screen. It is also found that the active centres have a certain lifetime and that the number vanishing per unit time is proportional to the number present. The number of centres that disappear as a result of methane adsorption is negligibly small. The number of centres appearing or vanishing per unit time is therefore given by:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = k_1 j\,(An_0 - N) - k_2 N. \quad . \quad . \quad . \quad (11)$$

where k_1 and k_2 are constants. At equilibrium we have:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = 0\,, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

and therefore, putting $k_1/k_2 = k$, we find:

$$N = \frac{k_1 j A n_0}{k_1 j + k_2} = \frac{kjAn_0}{kj+1}\,. \quad . \quad . \quad . \quad . \quad . \quad (13)$$

The number of methane molecules adsorbed by the screen per unit time, i.e. $V\,\mathrm{d}p/\mathrm{d}t$, is proportional to the number of active centres and to the pressure:

$$V\frac{\mathrm{d}p}{\mathrm{d}t} = -aNp\,, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

where a is a proportionality factor. Combining this expression with equation (9) it follows that the pumping speed S is equal to $+aN$, or, denoting the maximum pumping speed per unit area, an_0, by s_0:

$$S = \frac{kjAs_0}{kj+1}\,. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

In the linear portion of the "characteristic" in fig. 15 the pumping speed is constant and the number of centres formed under the given conditions is maximum. The maximum pumping speed per cm² is $s_0 = 0.01$ l/sec, and the constant $k = 2.0$. The formula derived shows good agreement with the experimental values of the pumping speed found for different values of scanned area A and current density j at the screen. The formula is valid up to a scanned area of 300 cm², above which the increase in pumping speed is less than follows from the formula.

The high pumping speed for methane in a picture tube during operation considerably reduces the partial pressure of the methane. A finished picture tube, when not in operation, often contains more methane than any other gas. This is due to the fact that methane may be produced at room temperature by reactions between water vapour and carbon in the getter film. There is no uptake of methane in the barium getter itself. If the methane were not to disappear during operation, the ionic bombardment of the cathode would be very much stronger and the life of the tube would therefore be shortened. This illustrates clearly how important it is to know the composition of the residual gases and to investigate the various processes that may take place inside a television picture tube.

Summary. In the manufacture of television picture tubes it is important to know not only the total pressure in the tubes but also the composition of the residual gases. This information can readily be obtained with an omegatron. After briefly describing the principle of the omegatron for quantitative analyses and the experimental system used, the authors discuss the desorption of gas by various parts of the picture tube. A method of analysing the composition of the gas in sealed-off tubes is described and some results are given. Finally, experiments are discussed which demonstrate the high uptake of methane by the fluorescent screen in a picture tube during electron bombardment. This removal of desorbed methane takes place continuously during the normal operation of a picture tube. It is attributable to the silicate binder used in the screen, and it guards the cathode against damage by ion bombardment.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2865: F. N. Hooge: Relation between electronegativity and energy bandgap (Z. phys. Chemie Neue Folge (Frankfurt a.M.) **24**, 275-282, 1960, No. 3/4).

It is shown that there exists a relation between the energy gap of crystals with formula AB and the electronegativities of the atoms A and B. The values of the electronegativities are the only quantities needed for the estimation of the width of the energy gap.

2866: G. Thirup: The application of phase-locking techniques to the design of apparatus for measuring complex transfer functions (J. Brit. Instn. Radio Engrs. **20**, 387-396, 1960, No. 5).

The theory of phase-lock synchronization is given briefly. Two devices for measuring complex voltage ratios are described. In the first, which has a frequency range 1-110 Mc/s, the voltage ratio to be measured is converted to a constant intermediate frequency of 450 kc/s. The frequency difference between the signal generator and the local oscillator is kept constant by coupling the tuning shafts mechanically and by a phase-lock synchronizing device; special means for extending the pull-in range, making use of a coarse and fine electronic adjustment, are described. The second instrument has a frequency range of 30-700 Mc/s, and here the local and signal generators are free-running and are separately tuned manually. A double frequency conversion system is used. The frequency of the second local oscillator, by means of a phase-lock device, is kept 450 kc/s below the frequency difference of the signal and first local oscillator. A description is also given of how the complex measuring results can be displayed on a cathode-ray tube.

2867: M. Avinor: Gold-activated (Zn,Cd)S phosphors (J. Electrochem. Soc. **107**, 608-611, 1960, No. 7).

Gold is shown to produce three emission bands in CdS, at 640, 800 and 1150 mμ. The long-wave band appears only when a coactivator is used, while the short-wave bands are observed with activation by gold alone. The 1150 mμ band in CdS is shown to correspond to the 530 mμ gold band in ZnS.

2868: H. Bremmer: The propagation over an inhomogeneous earth considered as a two-dimensional scattering problem (Electromagnetic wave propagation, int. Conf. sponsored by the Postal and Telecomm. Group of the Brussels Universal Exhibition, edited by M. Désirant and J. L. Michiels, pp. 253-260, Academic Press, London 1960).

This article deals with a two-dimensional integral equation for the distribution of field strength across a flat inhomogeneous earth. The equation in question is based on an approximative boundary condition, which accounts for the local distribution of the electrical constants. An infinite series representing the solution is interpreted in terms of scattering effects. The result applying to a special situation (two homogeneous regions separated by a straight boundary) is compared with that derived from the conventional one-dimensional equation which constitutes a saddlepoint approximation of the above two-dimensional equation.

2869: C. Z. van Doorn and Y. Haven: Anisotrope kleurcentra in KCl (Ned. T. Natuurk. **26**, 216-220, 1960, No. 7). (Anisotropic colour centres in KCl; in Dutch.)

When a KCl crystal is coloured by heating in K vapour (F absorption band at 5340 Å) and then irradiated at room temperature with light absorbed in the F band, a new (M) band appears at 8080 Å. After irradiation at 77 °K with polarized light absorbed in the F band, both the F and M bands are found to be dichroic. The authors interpret this effect by assuming the M band to be caused by two neighbouring F centres. The merits of this model are compared with those of similar models proposed by Seitz and by Knox.

VOLUME 23, 1961/62, No. 5 pp. 133-160 Published 7th February 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

AN EXPERIMENTAL FLUORESCENT SCREEN IN DIRECT-VIEWING TUBES FOR COLOUR TELEVISION

by R. R. BATHELT *) and G. A. W. VERMEULEN *).

535.371.07:621.397.62:621.397.132

The screen of dir ect-viewing tubes for colour television consists of three phosphors, giving red, green and blue fluorescence. The phosphors hitherto used have been, respectively, a phosphate, a silicate and a sulphide. A serious drawback of the phosphate and silicate is their relatively long afterglow. Red and green fluorescent phosphors are known that have a much shorter afterglow, but their use has been barred by the difficulty of applying them to the face plate. Methods of overcoming this difficulty have been studied at Philips since 1957. The technique described in this article, although still in the experimental stage, offers good prospects of direct-viewing screens composed of three sulphide phosphors.

The principles of colour television have already been dealt with in this journal [1]). Briefly, the camera produces three images of the scene in the primary colours red, green and blue, and the three video signals thus obtained (R, G and B respectively), after being coded, are transmitted modulated on a carrier; after decoding in the receiver, three corresponding video signals R, G and B are available, and it is the task of the display device to modulate a red, a green and a blue light source in accordance with these signals.

A system developed for this purpose is the *colour-television projector* earlier described [2]). This uses three projection tubes having phosphors which fluoresce red, green and blue respectively. The tubes are separately driven by the signals R, G and B, producing a red, a green and a blue image. The three images are magnified by an optical system and projected on to a screen in such a way that they accurately coincide.

Although this system can also be adopted for smaller pictures, *direct-viewing colour picture tubes* are preferable for use in the home.

Direct-viewing colour picture tubes

Several types of these tubes exist. They are all designed with the common object of producing a complete colour picture on the screen by means of the signals R, G and B. They might be said to be a combination of three ordinary picture tubes — but each having a screen which fluoresces in a different primary colour — in a single envelope.

As in all cathode-ray tubes, the fluorescent screen is a coating on the inside surface of the tube face. It consists of a regular array of large numbers of colour cells (several hundred thousand in the shadow-mask tube to be mentioned later). Each colour cell contains a specific geometrical arrangement of three primary phosphors, which fluoresce red, green and blue, respectively, upon bombardment by electrons. The arrangement of these phosphors may differ widely: they may be applied as closely packed parallel lines in the sequence red-green-blue-red-green-blue-etc., but usually they are dots with the red, green and blue of each triad contiguous to one another. The shortest distance between two dots of the same colour must be small enough for the structure to be indistinguishable by the eye at the normal viewing distance, where it is thus seen only as a mixture of the light emitted by the individual phosphors (cf. colour printing).

*) Electron Tubes Division, Eindhoven.

[1]) F. W. de Vrijer, Fundamentals of colour television, Philips tech. Rev. **19**, 86-97, 1957/58.

[2]) T. Poorter and F. W. de Vrijer, The projection of colour-television pictures, Philips tech. Rev. **19**, 338-355, 1957/58.

The phosphors in each cell must be bombarded by electrons in accordance with the instantaneous values of the three primary signals. This can be done in two ways:

1) *simultaneously*, by means of three electron guns (one for each primary colour) which are mounted side by side in the neck of the tube, or
2) *sequentially*, by means of one electron gun, the beam striking in turn a red, a green and a blue phosphor. In this method electronic means are needed to ensure that the gun is driven at every instant by the appropriate signal.

Further, the tube must also contain a colour-selecting device, which directs the electrons on to the appropriate primary phosphors. It is particularly in this respect that the various types of colour tubes differ one from the other. We shall not go into these differences here; later on in this article, onc particular type — the shadow-mask tube — will be dealt with at greater length.

Conventional method of applying the phosphors

For applying the three phosphors in a specific pattern to the inside face of thc picture tube, a kind of photographic process is nowadays employed. Use is made of a lacquer which is polymerized and hardened under ultraviolet irradiation. The lacquer is usually a solution of polyvinyl alcohol in water, sensitized with a dichromate. The maximum sensitivity lies at a wavelength of 365 nm, i.e. in the long-wave ultraviolet; the boundary wavelength is in the blue, in the region of 470 nm (1 nm = 10^{-9} m) (*fig. 1*).

The first phosphor to be applied to the face plate is suspended in the lacquer and the resultant slurry is coated uniformly over the tube face. After drying, the coating is irradiated only at those places where the phosphor is required. Depending on the type of tube, this is done by producing an optical image of a particular negative on the face plate by means of ultraviolet radiation, or by the shadow effect obtained by irradiating a negative with a point source of ultraviolet radiation; these negatives contain the appropriate pattern of lines or dots. The lacquer is hardened only at the places irradiated, so that a latent image is produced which is developed in the next operation. This operation is a water treatment, the unwanted phosphor being removed by causing the lacquer at the non-irradiated areas to swell up and dissolve. At the irradiated places the lacquer is insoluble and remains adhering to the glass together with the phosphor. The result, after drying, is the first phosphor pattern, still mixed with the polyvinyl alcohol.

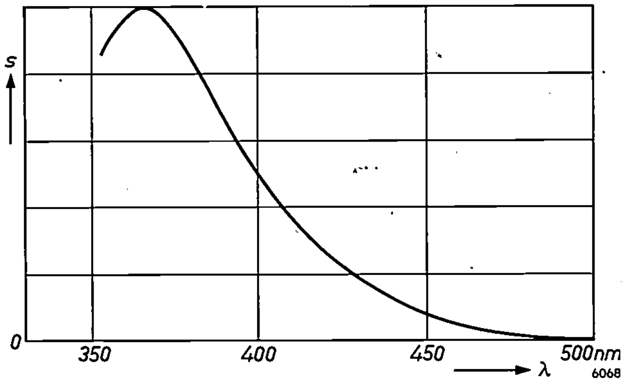


Fig. 1. Spectral sensitivity characteristic of the hardening of a polyvinyl-alcohol solution sensitized with dichromate. The maximum lies in the long-wave ultraviolet at 365 nm. (1 nm = 10^{-9} m.)

The process is now repeated with the second phosphor, and then again with the third phosphor. The irradiation must be done with other negativcs or from other points, since of course the three patterns must be displaced with respect to one another.

Once the fluorescent screen is completed, it is aluminized in the same way as in black-and-white tubes, being coated with some organic material (generally a methacrylate) on which a film of aluminium is then vacuum-deposited. The aluminium backing serves, among other things, as a reflector. Finally the entire screen is baked out to remove the polyvinyl alcohol and the methacrylate.

Choice of phosphors

The choice of the primary colours was dealt with at some length in the article quoted [3]). Apart from the colour of the emitted light and the luminous efficiency, important considerations in the choice of phosphors are the afterglow (persistence) and the technological properties of the phosphors in regard to their application in colour tubes. Afterglow in moving television pictures must not be too long, for if the luminance of one field (i.e. of one of the two equal parts into which the picture is divided in the interlaced scanning method normally used in television) is too high after 20 milliseconds, when the next field is being traced, the definition will be reduced, and moreover "trailing" will be noticeable at marked transitions of brightness.

The following are the main phosphors hitherto used in direct-viewing colour picture tubes:

for red, a zinc phosphate activated with manganese;
for green, willemite (a zinc silicate) activated with manganese;
for blue, a zinc sulphide activated with silver.

The colours in which these phosphors fluoresce are the primary colours of the colour television system adopted in the United States — the N.T.S.C. system

[3]) See ref. [1]), p. 90, fig. 8.

(named after the National Television System Committee)[4].

Table I gives the major properties of the above-mentioned phosphors, and for comparison the data of the normal phosphors used in black-and-white tubes (a mixture of yellow and blue). As appears from the table, the red and green phosphors have a much higher persistence than the blue phosphor, too high in fact. This causes the above-mentioned loss of definition in moving images and also, depending on the colours in the scene, gives rise to "trailing" — streaks of red, orange or green hue behind sharp transitions.

Red and green fluorescent phosphors are known, however, that have a lower persistence. As a green phosphor, for example, willemite can be used with a greater activator content. The persistence is then much lower, although still not as low as might be wished; moreover the luminous efficiency is somewhat lower.

A better solution is offered by the zinc-cadmium sulphides having a suitably chosen cadmium-sulphide content. With increasing cadmium-sulphide content the colour of the light emitted by this phosphor changes from blue through green, yellow and orange to red. The dot-dash line in the chromaticity diagram shown in *fig. 2* gives the colour points for the fluorescence of zinc-cadmium sulphides, the cadmium-sulphide content being indicated at some points along the curve. A yellow-fluorescent (ZnCd)S, as mentioned in table I, is used as the yellow component in black-and-white tubes. All (ZnCd)S phosphors have a low persistence.

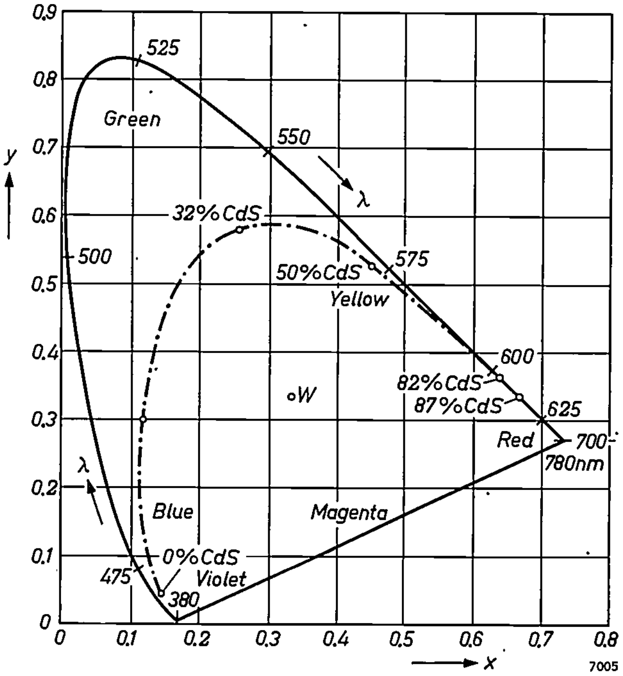


Fig. 2. Chromaticity diagram. The dot-dash line gives the colour points of the fluorescent light from (ZnCd)S phosphors with varying cadmium-sulphide content. *W* is the colour point of white.

In addition to a low persistence the zinc-cadmium sulphides have the advantage of a high efficiency. *Table II* gives the data for the strongly activated willemite just mentioned and for the zinc-cadmium sulphides suitable for green and red. With an appropriately chosen cadmium-sulphide content it is possible to give red-fluorescent (ZnCd)S the same colour point as the $Zn_3(PO_4)_2$-Mn used hitherto. The green of (ZnCd)S is less saturated than that of willemite; we shall return to this point at the end of the article.

In view of the high efficiency and low persistence of the green-fluorescent and red-fluorescent sulphide, it might be asked why these long familiar phosphors

[4] On direct-viewing tubes the resultant colour of the three primary rasters will be somewhat less saturated than the colours of the individual phosphors, owing to the fact that some of the fast secondary electrons liberated upon the bombardment of one of the three phosphors strike the two other phosphors.

Table I. Data of conventional phosphors in colour picture tubes and black-and-white tubes.

Type of tube	Colour	Phosphor	Colour co-ordinates		Efficiency *) cd/W	Persistence **) %
			x	y		
colour tube	red green blue	$Zn_3(PO_4)_2$-Mn Zn_2SiO_4-Mn ZnS-Ag	0.670 0.210 0.140	0.330 0.710 0.080	1.5 5.5 1.4-2.3	29 18 <1
black-and-white tube	yellow blue	(ZnCd)S-Ag ZnS-Ag	0.450 0.145	0.530 0.100	6.5	<1 <1

*) Luminous intensity in a direction perpendicular to an aluminized fluorescent screen radiating in accordance with Lambert's law, per watt of electric power supplied. The actual efficiencies of the phosphors are in fact higher: the figures given here allow for the 72% transmission of the tube face.

**) Luminance 20 milliseconds after electron-beam cut-off, as a percentage of the initial value.

Table II. Data of strongly activated willemite and of green- and red-fluorescent zinc-cadmium sulphide.

Colour	Phosphor	Colour co-ordinates		Efficiency *) cd/W	Persistence **) %
		x	y		
green	Zn_2SiO_4-Mn, strongly activated	0.250	0.700	4.5	4.5
green	(ZnCd)S-Ag 32% CdS	0.255	0.575	8.4	1-1.5
red	(ZnCd)S-Ag { 87% CdS 82% CdS	0.670 0.640	0.330 0.360	1.8 2.3	<1 <1

*) and **): see Table I.

were not used right from the beginning in colour picture tubes. The reason is that the red-fluorescent (ZnCd)S cannot be applied to the face plate by the method described above, whereby the lacquer containing the phosphor is irradiated with ultraviolet from the side where the electron gun will later be fitted. The zinc-cadmium sulphides get a deeper yellow colour as the cadmium-sulphide content increases, and strongly absorb ultraviolet, violet and blue. Upon irradiation a layer of the necessary thickness would therefore not be thoroughly hardened right up to the glass, and would flake off to a greater or lesser extent during the development process. The green-fluorescent sulphide is so weakly yellow that it can still just be applied in the normal manner, but with increasing cadmium-sulphide content the hardening radiation is so strongly absorbed that the method can no longer be used for the red-fluorescent sulphide.

Efforts have been made to find other methods of applying the sulphide phosphors. In principle, ultraviolet-absorbent phosphors can be applied photographically by the "sticky-dots" method [5]), in which the phosphor is made to stick to a previously applied lacquer pattern. The necessary stickiness is obtained by irradiating the lacquer for so short a time that it does not harden completely.

In this way, however, only an extremely thin layer of phosphor can be obtained. It appears to be difficult to prevent holes forming in the layer, as a result of which the luminous intensity is both inadequate and non-uniform over the screen. The whole procedure, starting from the application of the lacquer coating and ending with the development and drying of the phosphor layer, would therefore have to be repeated. In our experience, however, the result is even then unsatisfactory, and moreover it is extremely difficult to meet the requirements in regard to the average size of the grains and the grain-size distribution of the phosphors.

A new method of applying sulphide phosphors

In a new method developed here, use is made of the circumstance that it is not necessary to fix all three phosphors with the aid of a negative, but only two of them; the spaces left open by these two can be filled with the third phosphor, which is fixed by irradiation through the glass (without a negative), i.e. from outside the tube. Since the ultraviolet now has its maximum intensity at the interface between glass and lacquer, powerful adhesion to the glass is obtained.

The last phosphor to be applied (the red-fluorescent one) will now however be not only between but also (seen from outside) behind the green and the blue phosphors, and as a result the green and in particular the blue fluorescent light will be contaminated by red. Steps are therefore needed to prevent the red-fluorescent phosphor from being deposited behind the other phosphors. For this purpose, before the lacquer suspension containing red-fluorescent phosphor is applied, the green-fluorescent and blue-fluorescent phosphors are temporarily coated with an ultraviolet-absorbent dye on the gun side. At these positions, then, the lacquer with the red-fluorescent phosphor does not harden upon irradiation through the glass and is therefore removed in the development process.

Fig. 3 illustrates schematically in column *a* the method of fixing a red-fluorescent phosphate, which absorbs virtually no ultraviolet, and in columns *b* and *c* the fixing of a red-fluorescent sulphide that does absorb ultraviolet, respectively without and with a dye on the other phosphors. In case *b* the red-fluorescent phosphor is also deposited behind the other phosphors; in case *c* this is prevented by the protective dye.

After the last phosphor has been applied, the dye coating is dissolved.

Application of the dye

The dye can be applied in very much the same way as the phosphor. The ultraviolet-absorbent dye is

[5]) M. Sadowski and P. D. Payne Jr., Photodeposition of luminescent screens, J. Electrochem. Soc. **105**, 105-107, 1958.

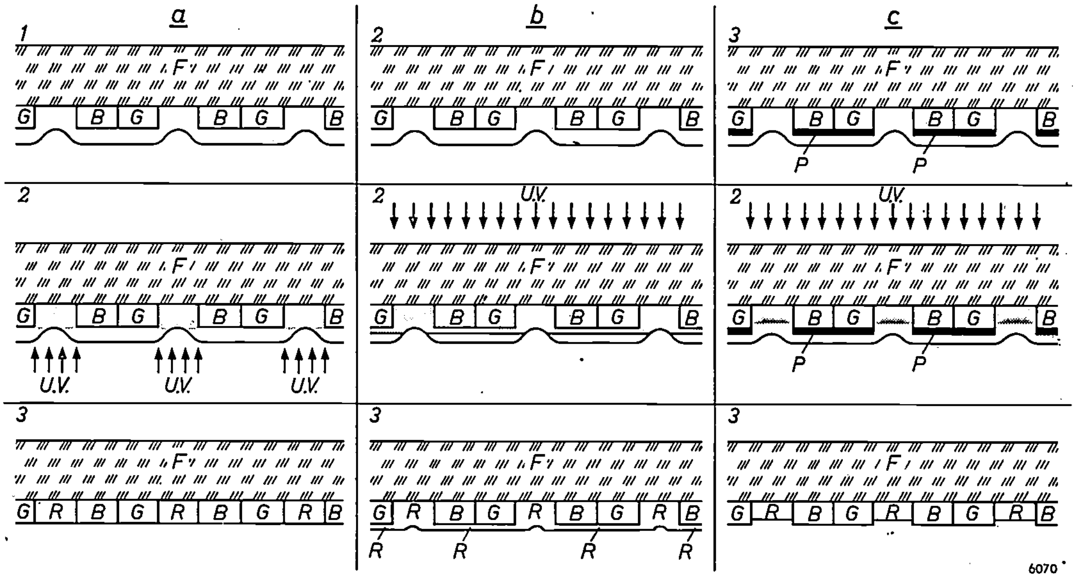


Fig. 3. Sketch representing schematically the application of the red-fluorescent phosphor (R) between the green-fluorescent and blue-fluorescent phosphor patterns (G and B).

Column a: The conventional method (red-fluorescent phosphor absorbing little ultraviolet).
1) Tube face F with G and B phosphor patterns, the latter being covered with a lacquer containing the R phosphor in suspension.
2) The places where the R phosphor dots are to come are irradiated from the gun side with ultraviolet. The shading indicates where the lacquer hardens.
3) Upon development the unhardened lacquer is removed, leaving behind the R phosphor dots.

Column b: The R phosphor is now a sulphide and absorbs ultraviolet. No dye is yet applied to the other phosphors.
1) As a, 1) above.
2) Irradiation with ultraviolet through the tube face. The shading indicates where the lacquer hardens.
3) After development the R phosphor remains, not only between but also (seen from outside) *behind* the G and B phosphors. As a result, the green and blue emissions are contaminated with red.

Column c: The new method.
1) The G and B phosphors are coated with a dye P which absorbs ultraviolet.
2) Hardening by irradiation through the glass.
3) After development the R phosphor remains only *between* the G and B phosphors.

suspended in a solution of polyvinyl alcohol sensitized with dichromate, and a coating of this suspension is applied to the screen, to which the green-fluorescent and blue-fluorescent patterns have already been fixed; this coating is then irradiated in such a way that both patterns are entirely covered with the dye after development.

For the purpose of analysing the principle of this method we shall assume first of all that the dye and afterwards the red-fluorescent phosphor are fixed with radiation of the same spectral composition (effective range from about 450 to 350 nm) and that the dye does *not* absorb selectively in this range.

The curve in *fig. 4* illustrates qualitatively the decrease in the intensity I of the radiation within the layer to be hardened, as a function of the depth of penetration d. At $d = d_\mathrm{p}$ the intensity I has dropped to the minimum value I_min which, at a given irradiation time T_p, is needed to produce hardening. What remains after development is thus a layer of dye of thickness d_p. The transmittance of this layer

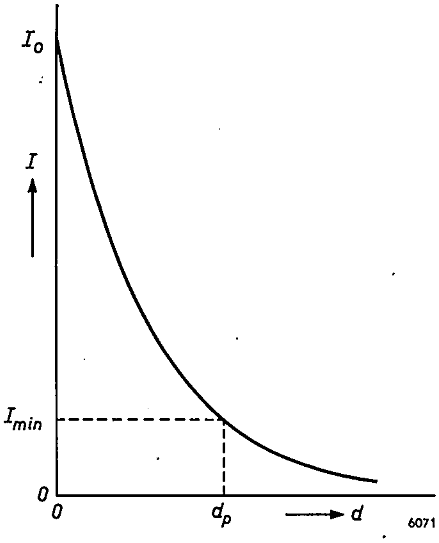


Fig. 4. Approximate variation of the intensity I of ultraviolet as a function of the depth of penetration d in a lacquer coating containing a dye in suspension. At $d = d_\mathrm{p}$ the intensity has dropped to the minimum value I_min which, at a given exposure time T_p, is necessary to harden the lacquer. The result after development is a layer of dye of thickness d_p and transmittance I_min/I_0.

is not zero but equal to $I_{\min}/I_0$, where I_0 is the intensity at $d = 0$. When the suspension with the red-fluorescent phosphor has been applied and is irradiated through the glass, this radiation reaches the layer where it joins the glass with the intensity $a_g I_0$, and where the layer is separated from the glass by the green-fluorescent or blue-fluorescent phosphor, plus the dye, with the lower intensity $a_g a_f a_p I_0$. (The coefficients a allow for the loss by reflexion at an interface and by absorption in a medium; the suffixes g, f and p relate to glass, phosphor and dye respectively.) In order to fix the red-fluorescent phosphor solely at the places first mentioned, the irradiation time T_r must be so chosen that the exposure dose $a_g I_0 T_r$ causes hardening, but the smaller dose $a_g a_f a_p I_0 T_r$ does not. The result is therefore markedly dependent on T_r and on the factor a_p. Now a_p depends on the dye thickness d_p, and hence on the irradiation time T_p needed to fix the dye. It therefore depends critically on the ratio between the irradiation times T_p and T_r.

The value of this ratio becomes less critical if we choose a dye that absorbs selectively, such that a_p is small at the wavelength 365 nm to which the lacquer is most sensitive (see fig. 1), and much larger at longer wavelengths. For fixing the dye this makes it possible to use radiation mainly consisting of longer waves (from a different source) than the radiation used for fixing the red-fluorescent phosphor.

It should be noted that the dye is indispensable on the blue-fluorescent phosphor dots, but if necessary may be dispensed with on the green-fluorescent dots, provided the latter are composed of a somewhat yellowish (ZnCd)S; this phosphor itself absorbs the ultraviolet to a sufficient extent.

The new method applied to the shadow-mask tube

The new method is suitable for any type of direct-viewing colour picture tube. We shall presently discuss its application in the shadow-mask tube [6]), but first it will be useful to give a brief description of this tube.

The shadow-mask tube contains three electron guns and a fluorescent screen with the primary phosphor applied as dots arranged in a hexagonal pattern (*fig. 5*). Each colour cell consists of one red, one green and one blue phosphor dot. The centres of these dots lie on the corners of equilateral triangles. In the optimum arrangement the dots touch one another without overlapping. The open spaces between the dots are later covered by the aluminium backing.

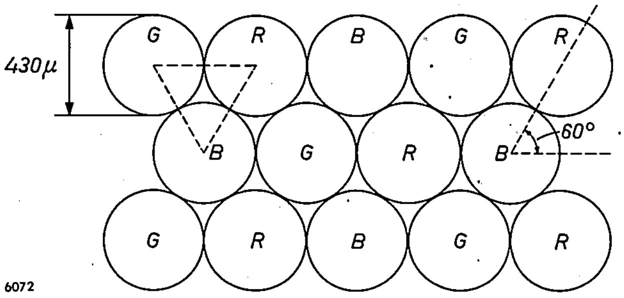


Fig. 5. Hexagonal configuration of the phosphor dots in the fluorescent screen of a shadow-mask tube. Each triad of phosphor dots fluorescing red, green and blue (*R*, *G*, *B*) forms a colour cell.

The three guns are mounted side by side in the neck of the tube. Colour selection is effected with the *shadow mask*, which is a perforated metal plate fitted in the tube about 13 mm behind the screen (*fig. 6*). Each colour cell has a corresponding hole in the shadow mask, which means that there are three times as many phosphor dots on the screen as there are holes in the mask. The guns are tilted slightly, so that their axes intersect on the shadow mask. The phosphor dots are so disposed on the tube face that the beam from each gun, after being deflected in the deflection coils, can pass through the hole in the mask only on to the correct phosphor. Electrons from the "red" gun, for example, strike only red-fluorescent dots, electrons from the "blue" gun strike only blue-fluorescent dots, and so on. Colour selection is therefore effected solely by masking.

For applying the phosphor dots by the photographic method no negative is needed during exposure in this case, since the shadow mask itself can serve as a "jig". Beyond the bend which they undergo in the deflection coils the electron trajectories from one

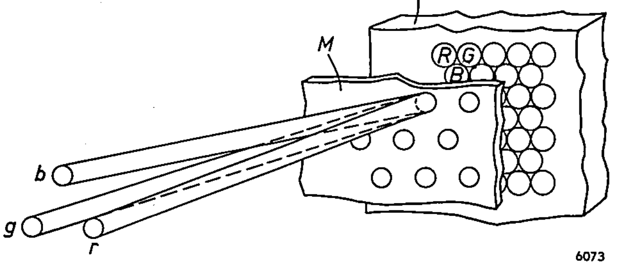


Fig. 6. Illustrating the principle of the shadow-mask tube. *F* tube face with hexagonal phosphor-dot pattern *R-G-B* as in fig. 5. *M* shadow mask. *r*, *g*, *b* electron beams from the "red", "green" and "blue" guns respectively. The configuration is such that each of the three beams strikes only phosphor dots of the corresponding colour.

[6]) H. B. Law, A three-gun shadow-mask color kinescope, Proc. Inst. Radio Engrs. **39**, 1186-1194, 1951.
N. F. Fyler, W. E. Rowe and C. W. Cain, The CBS-Colortron: a color-picture tube of advanced design, Proc. Inst. Radio Engrs. 42, 326-334, 1954.

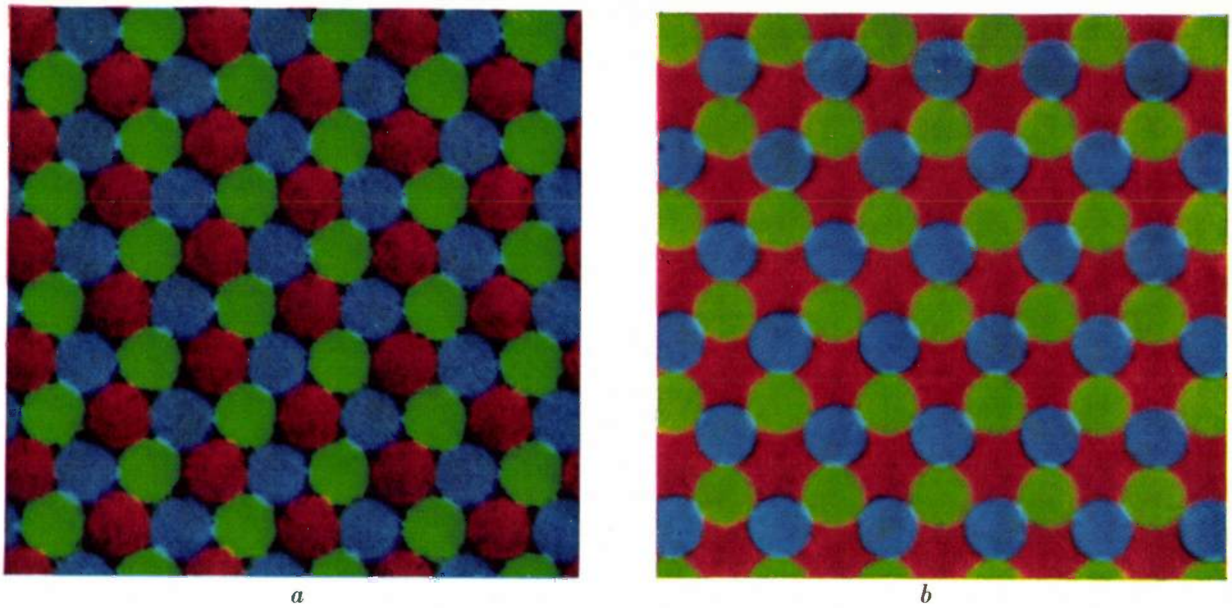


Fig. 7. Colour photographs of the screen of shadow-mask tubes. *a*) Screen with normal phosphors. *b*) Screen with three sulphide phosphors applied by the new method.

To make the structure of the screen visible, the guns were strongly defocused. In correct focus the electrons strike only parts of the phosphor dots that have a diameter of about 0.8 times the dot diameter.

gun in the tube are straight lines which, continued backwards, intersect at one point, the deflection point. The rays from an ultraviolet source situated at the deflection point thus pass through the shadow mask and strike those places on the screen where the phosphor dots of the relevant colour have to come, and only there do they harden the lacquer. In the successive irradiations the source must of course be situated at the deflection point for the primary colour in question.

In its present version [7]) the shadow-mask tube is a round glass tube having a deflection angle of 70° and a convex screen of 53 cm diameter. The screen has roughly 350 000 colour cells. The holes in the shadow mask are made as large as is consistent with the requirement that, under normal conditions, the electrons are only just prevented from striking phosphor dots of the wrong colour. Nevertheless, a large proportion of the electron current from each gun is lost in the mask itself, where it generates heat; the average transmittance of the mask is only 15%. *Fig. 7a* reproduces a 20-times enlarged colour photograph of the colour screen. To make the entire screen structure visible, the guns were strongly defocused (normally the electrons strike only parts of the phosphor dots that have a diameter of about 0.8 times the total diameter of the dot).

In order to provide such a tube with three sulphide phosphors by the new method, the first procedure is to apply the green-fluorescent and blue-fluorescent phosphor dots in the normal way. Next, the dye suspension is similarly applied, after which it is dried and irradiated through the shadow mask in the same manner as for the blue-fluorescent phosphor. A yellow layer of dye forms behind each blue-fluorescent dot upon development. The screen is now coated with lacquer containing the red-fluorescent phosphor in suspension, and the lacquer is irradiated through the tube face. During this exposure, in which the shadow mask is not used, the lacquer hardens everywhere except behind the blue-fluorescent dots (owing to absorption in the dye) and behind the green-fluorescent dots (which are sufficiently absorbent themselves). The result after development is that all open spaces between the green-fluorescent and the blue-fluorescent pattern are covered with red-fluorescent phosphor. This can be seen from the photograph in fig. 7*b*, taken under similar conditions as for fig. 7*a* [8]).

An important advantage of the new method in tubes with phosphor dots (as opposed to those with phosphor lines) is the following. The phosphor-dot

[7]) C. P. Smith, A. M. Morrell and R. C. Demmy, Design and development of the 21CYP22 21-inch glass color picture tube, RCA Rev. **19**, 334-348, 1958.

[8]) During the preparation of this article it came to the authors' attention that the Radio Corporation of America had brought out a shadow-mask tube using three sulphide phosphors. The fluorescent pattern is a three-dot configuration, which means that the red dots are not applied in the manner described here.

pattern of a normal colour picture tube (fig. 5) shows open spaces between the dots. Through these openings light from lamps in the room may be reflected by the aluminium backing. In the new method the openings between the green and blue phosphors are completely filled by the red phosphor, and the aluminium backing is thus unable to cause any troublesome reflections.

Tubes with two and three sulphides

Tubes using a zinc-cadmium sulphide for red, zinc sulphide for blue, and strongly activated willemite — with its relatively low persistence — for green, have not proved to be entirely satisfactory. Although the persistence is appreciably lower than in a tube using normal phosphors, so that moving images are sharper, the dark green fringes that appear during movement (owing to the fact that willemite still has a longer persistence than the sulphides, see Table II) have proved to be disadvantageous. Another drawback is that willemite has a lower efficiency than phosphors of the sulphide type.

Using three sulphides, i.e. a sulphide also for the green, gives the advantages of excellent definition for moving images and a higher luminance. Compared with the normal phosphors the gain in luminance is 20 to about 53% in the red (depending on the colour co-ordinates), 49% in the green and about 43% in the white ($x = 0.287, y = 0.316$).

Set against these substantial advantages is the drawback that green-fluorescent (ZnCd)S gives a less saturated colour than willemite. Generally speaking, where one of the primary colours (in this case green) has a lower saturation the result is a less faithful reproduction in two respects:

a) It narrows the range of colours reproducible in the picture.

b) It gives rise to colour errors, owing to the fact that the N.T.S.C. system hitherto used for colour television takes the fluorescence of willemite as the basis for green.

To conclude, we shall examine these points in more detail.

The dependence of the reproducible range of colours on the colour point of the green

The chromaticity diagram shown in *fig. 8* indicates the colour points R, G and B of the three normal phosphors. As will be known, all colours that can be produced by additive mixing of these three primaries lie within and on the periphery *1* of the triangle RGB.

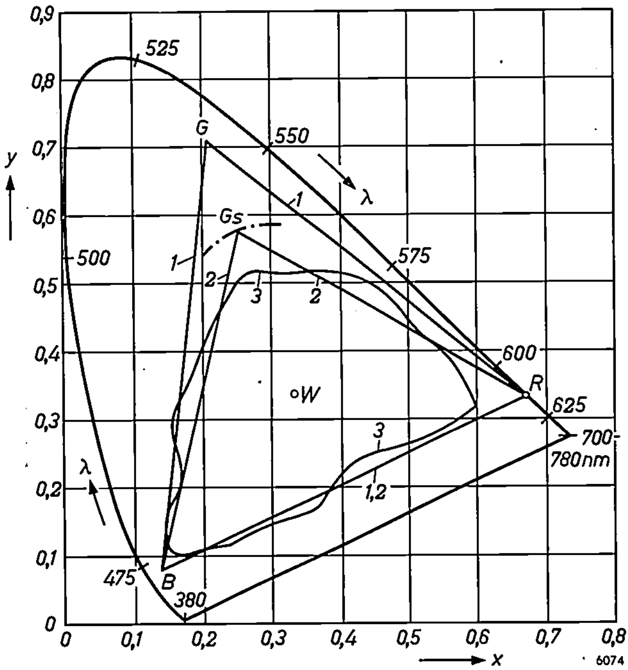


Fig. 8. Chromaticity diagram with the colour points R, G and B of the normal phosphors of direct-viewing colour picture tubes. All colours reproducible with these phosphors lie within or on the periphery *1* of triangle RGB. A sulphide phosphor with 87% CdS has its colour point also at R, a sulphide phosphor with 32% CdS at G_s. A screen containing both these sulphides and having the normal sulphide for blue can produce all colours lying within or on the periphery *2* of triangle RG_sB. Contour *3* marks the boundary of all reflection colours of the natural and artificial dyes. W is the white point.

With increasing cadmium-sulphide content in the green-fluorescent sulphide, G_s moves to the right along the dot-dash line.

A tube using sulphide phosphors has the same primary blue. Both for red and green there is a choice from a series of phosphors of widely differing colours. In fig. 8 a red-fluorescent sulphide phosphor has been chosen (with 87% CdS) whose colour point coincides with R. For green we have taken a (ZnCd)S with 32% CdS, which has roughly the same dominant wavelength as willemite. The colour point of this sulphide phosphor is denoted by G_S. Contour *2* thus bounds all colours that can be reproduced with these three sulphide phosphors. The contour *3* bounds all reflection colours of natural and artificial dyes and printing inks [9]).

It can be seen from fig. 8 that nearly as many colours can be reproduced with the three sulphide phosphors as with the conventional phosphors. The difference, which is due to the lower saturation of the green, is in reality even smaller than the figure might lead one to suppose, for as far as the green hues are concerned the change-over from G to G_S implies no loss in colour reproduction, since G_S is still outside the contour *3* of the colours which actually occur. (Even in the greenest parts of the picture, therefore, purely green fluorescence will never be present, but these parts

[9]) W. T. Wintringham, Color television and colorimetry, Proc. Inst. Radio Engrs. **39**, 1135-1172, 1951.

will also contain red and blue.) There is, however, a slight loss in the reproduction of orange, yellow and greenish bluc. The loss in yellow is smaller than might be inferred from fig. 8, because of the fact that the difference between colours whose chromaticity points are the same distances apart are much less noticeable in yellow than, for example, in blue [10]); see *fig. 9.*

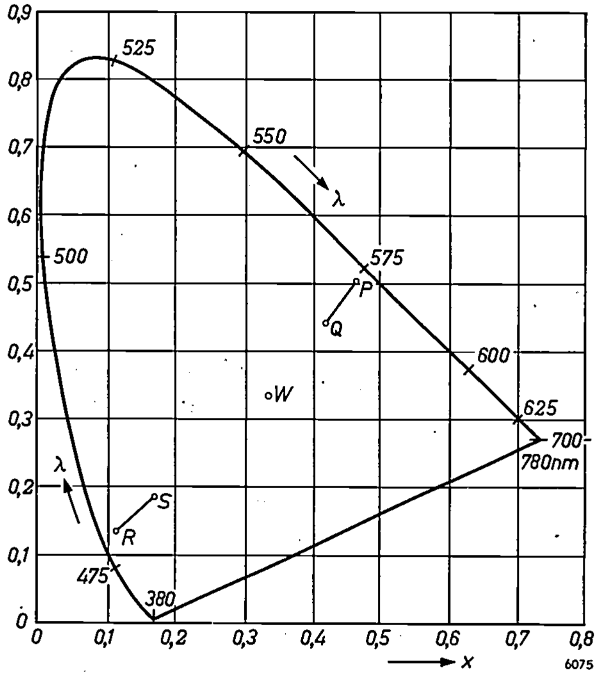


Fig. 9. Chromaticity diagram with two points in the yellow, P and Q, which are just as far apart as the points R and S in the blue. Between the yellow colours corresponding to P and Q the eye sees less difference than between the blue colours corresponding to R and S [10]).

Since yellow colours occur much more frequently than greenish blue, some improvement can be obtained by giving the green-fluorescent sulphide a somewhat higher cadmium-sulphide content, which shifts its colour point slightly towards yellow (in fig. 8 from G_S along the dot-dash line to the right).

Colour errors on a three-sulphide screen with an N.T.S.C. signal

A colour television system based, as far as green is concerned, on the colour point of willemite gives rise to colour errors on a screen that contains a green-fluorescent sulphide. These errors have been extensively studied, both in theory and practice, by Jackson [11]). His conclusion is that the errors can be almost completely compensated by means of a linear matrix network: all that is necessary is to subtract a specific percentage of the "green" signal from the "red" and "blue" signals. Especially in the absence of such compensation, it is advisable to shift the colour point of the red slightly towards orange. This has the additional advantage of improving the efficiency.

A favourable choice appears to be a combination of sulphides having the following cadmium-sulphide contents:

Colour	Phosphor	Colour co-ordinates	
		x	y
red	(ZnCd)S with 84% CdS	0.655	0.345
green	(ZnCd)S with 32% CdS	0.255	0.575

Blue remains unchanged.

The colour errors appearing on a three-sulphide tube under N.T.S.C. operation should not in any case be exaggerated. On repeated occasions we have compared such tubes with others containing the conventional phosphors. The same signal was fed to all monitors, and no correction was applied for differing phosphors. Although it could clearly be seen during the display of the primary colours that the green was less saturated on the three-sulphide tubes and that the red was more orange than on the normal tubes, hardly any difference was noticeable in a colour picture.

To produce white on the new tubes, almost the same current ratios $R : G : B$ are needed as in existing tubes.

Summary. In direct-viewing tubes for colour television, use has hitherto been made of a phosphate for the red-fluorescent phosphor and of a silicate for the green-fluorescent phosphor. The sulphide (ZnCd)S-Ag, which fluoresces red or green depending on the CdS content, has substantially better properties (lower persistence, higher efficiency). Applied by the conventional method, however, the red-fluorescent sulphide does not adhere well to the tube face.

A new method giving good adhesion is described, in which a suspension containing the red-fluorescent sulphide is hardened by ultraviolet irradiation through the glass instead of from the gun side. The other phosphor patterns are coated with a dye to prevent the red phosphor from sticking to them. The result is a screen on which moving images are much sharper and which has 40 to 50% higher luminance.

[10]) D. L. MacAdam, Proc. Inst. Radio Engrs. **39**, 479 (fig. 12), 1951.

[11]) Unpublished investigation by R. N. Jackson of Mullard Research Laboratory, Salfords (England).

CIRCUITS FOR DIFFERENCE AMPLIFIERS, I.

by G. KLEIN *) and J. J. ZAALBERG van ZELST *).

621.375:621.317.725.083.6

Pursuant to an article recently published in this journal, which dealt with difference amplifiers in general, part I of the present article gives details of circuits designed to achieve the very high rejection factors that are frequently required. Part II, to be published in the next number, will offer hints on the efficient use of difference amplifiers. Some cases will also be described where difference amplifiers can be used with advantage even where the object is not to amplify a potential difference between two points.

Introduction

The behaviour of a difference amplifier is largely governed, as discussed in an earlier article [1]), by the characteristics of the first stage. In dealing now with various circuits used for difference amplifiers, we shall therefore be primarily concerned with single-stage amplifiers. The problems that arise when stages are added, some of which will also be touched on here, are as a rule not difficult to solve.

It was shown in the previous article that difference amplifiers are mainly used for amplifying low-frequency or DC signals. Our considerations will therefore be confined to such amplifiers, and we shall disregard the problems encountered at higher frequencies, connected for example with the capacitances of valves and other components.

The nature of the applications of difference amplifiers makes it necessary to stipulate for the most important characteristics — the rejection factor and the discrimination factor — a lower limit which differs from case to case. *The circuits that we shall deal with have been designed so as to be able to guarantee these minimum values without readjustment of the amplifier, even when the parameters of the valves and other components, which always show some mutual disparity, have the maximum deviation from the normal value and these deviations are all operative in the same (adverse) direction.*

As explained at some length in the above-mentioned article [1]), the problem of the difference amplifier consists in amplifying the voltage between two points which may both have a much higher potential with respect to earth. An obvious method of amplifying the potential difference between the two points would be to connect them with the input terminal and the "earth terminal" of a normal single-ended amplifier. The latter terminal is then not connected to earth; it can be said that electrically the whole amplifier "floats". Without going into details, it may be noted that amplifiers made electrically floating in this way are usually complicated and unwieldy in construction, and moreover usually have to be screened against interfering induction voltages. The principle described in this article, of *a balanced amplifier with a very high common cathode resistance*, offers in almost every case a much simpler solution.

Circuit with common cathode resistance

It was shown in the article quoted [1]) that a balanced amplifier, consisting of two independently operating sections, cannot be used as a difference amplifier because its discrimination factor is equal to unity, whereas for most purposes a value of at least 100 is required. In principle, a circuit can be used where the two valves of a balanced amplifier are given a common cathode resistance without decoupling (*fig. 1*). As we shall see, this resistance, R_k, must be very much higher than that needed for giving the grid the normal negative bias. For

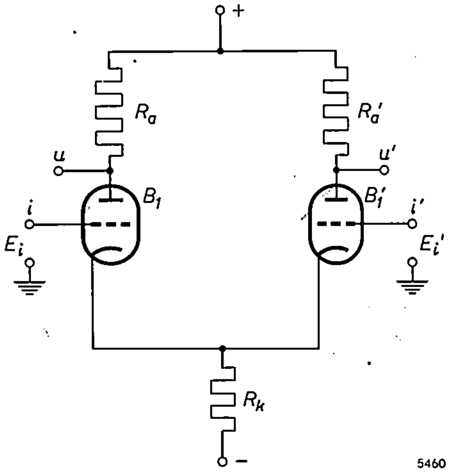


Fig. 1. Difference amplifier with a resistance R_k incorporated in the cathode lead common to both valves. Input terminals i and i', output terminals u and u'.

*) Philips Research Laboratories, Eindhoven.

[1]) G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. 22, 345-351, 1960/61 (No. 11).

this reason, R_k is not usually connected to earth but to a voltage source which supplies a negative voltage with respect to earth. In this way the grids of valves biased to their normal operating point can be more or less at earth potential, which simplifies connection to the points whose voltages are to be measured, especially in the case of DC amplifiers.

If both valves are perfectly identical, and identically biased, there will be no change in the current through R_k when two equal but opposite small changes [2]) are made to the two grid voltages E_i and E_i'. The resistance R_k then has no effect on the amplification of the two valves. If, however, the changes in E_i and E_i' have the same sign, the two valves function as if they were connected in parallel. The presence of R_k now gives rise to negative feedback. As a result the amplification of in-phase signals (see [1])) is smaller than that of anti-phase signals, and the discrimination factor F is therefore greater than unity. Assuming perfect symmetry, the value of F is easy to calculate. From the familiar expression for the gain of a triode the anti-phase gain is found to be

$$A = \mu \frac{R_a}{R_i + R_a}, \quad \ldots \quad (1)$$

where μ is the amplification factor and R_i the internal resistance of the valves, and R_a is the anode resistance. The in-phase gain C, where negative feedback due to R_k occurs, is given by

$$C = \mu \frac{R_a}{R_i + R_a + 2(1+\mu)R_k}. \quad \ldots \quad (2)$$

The discrimination factor is thus

$$F = \frac{A}{C} = 1 + 2\frac{(1+\mu)R_k}{R_i + R_a}. \quad \ldots \quad (3)$$

In the circuits we shall be dealing with, the value of R_i is always large compared with R_a, and $2\mu R_k$ large compared with R_i. Since moreover the amplification factor μ is large compared with unity for normal valves, we can write F with negligible error as

$$F = 2SR_k, \quad \ldots \ldots \quad (4)$$

where S is the transconductance of both valves. Where the transconductance of the valves is not exactly the same, the discrepancy can be taken into account by inserting in eq. (4) the average value of S for both valves. Small discrepancies in the values of R_i and R_a are found to have an entirely negligible influence on F.

To compute the rejection factor H we must start from differences between the parameters, which cause a degree of asymmetry in the amplifier. If the amplifier were perfectly symmetrical, the value of H would of course be infinite. To find the rejection factor of a circuit as in fig. 1, we should therefore assume that the two halves of the amplifier differ in the transconductance and amplification factor of the valves and also in the resistance in the anode leads. The result of this calculation [3]), provided the differences are not excessive, can be presented to a good approximation by the formula:

$$H = \frac{4}{\left(\frac{\Delta S}{S} + \frac{\Delta R_a}{R_a}\right)\frac{1}{SR_k} + \frac{1}{\mu}\left(\frac{\Delta\mu}{\mu}\right)\left(2 + \frac{R_a}{R_k}\right)}, \quad \ldots \quad (5)$$

where ΔS, $\Delta\mu$ and ΔR_a are the differences in the transconductance, amplification factor and anode resistance of the valves respectively. Distinguishing the relevant values for the one valve from those for the other by a prime, we may write $\Delta S = S' - S$, $\Delta\mu = \mu' - \mu$ and $\Delta R_a = R_a' - R_a$.

From equation (5) we see that the smallest value of H occurs when $\Delta S/S$, $\Delta\mu/\mu$ and $\Delta R_a/R_a$ have the same sign (i.e. when S, μ and R_a of one valve are greater than those of the other valve), and moreover have the maximum values that can be expected from the normal tolerances of valves and resistors. The equation is greatly simplified if we assume that the maximum relative difference of the above three quantities have the same value, denoted by δ. (In practice, δ may for example be 0.1.) Inserting this in eq. (5) we find the minimum rejection factor that can occur for a given magnitude of δ:

$$H_{\min} = \frac{4}{\delta\left\{\frac{2}{SR_k} + \frac{1}{\mu}\left(2 + \frac{R_a}{R_k}\right)\right\}}. \quad \ldots \quad (6)$$

From this expression we see that $H_{\min}$ is greater the larger the value of R_k in the common cathode lead. If the value of R_k is so high that R_a/R_k is negligible compared to 2, we can simplify eq. (6) to:

$$H_{\min} = \frac{2}{\delta\left(\frac{1}{SR_k} + \frac{1}{\mu}\right)}. \quad \ldots \quad (7)$$

Equations (6) and (7) show that to obtain a high value of $H_{\min}$ it is necessary, though not sufficient,

[2]) The changes must be small enough to prevent the curvature of the characteristics from playing any part.

[3]) See G. Klein, Rejection factor of difference amplifiers, Philips Res. Repts **10**, 241-259, 1955.

to have a large R_k. If SR_k is of the same order of magnitude as μ, then R_k must be very considerably increased to achieve a relatively slight increase in $H_{\min}$. For a particular value of $H_{\min}$, both SR_k and μ must have a specific minimum value which is larger the larger is $H_{\min}$. If SR_k is small compared to μ, the minimum rejection factor is given by:

$$H_{\min} = 2SR_k/\delta\,, \quad \ldots\ldots \quad (8)$$

or, putting $\delta = 0.1$:

$$H_{\min} = 20\,SR_k\,. \quad \ldots\ldots \quad (9)$$

From the expressions (4) and (9) we see that in this case F is roughly equal to 0.1 $H_{\min}$.

It may also happen that SR_k is large compared to μ, in which case $H_{\min}$ is approximately given by:

$$H_{\min} = 2\mu/\delta\,, \quad \ldots\ldots \quad (10)$$

or, with $\delta = 0.1$:

$$H_{\min} = 20\,\mu\,. \quad \ldots\ldots \quad (11)$$

In order to guarantee a rejection factor of, for example, 10 000 one must — assuming that R_k is sufficiently large — use valves having an amplification factor of at least 500. Where SR_k is not large compared to μ, the latter value must be even higher. For instance, $H_{\min}$ can also be made equal to 10 000 when both SR_k and μ are equal to 1000. Given a transconductance of 1 mA/V [4]) for both valves, R_k must then be equal to 1 MΩ.

In certain cases where an even higher rejection factor is required, a much higher value of R_k is needed, e.g. 10 MΩ. The use of such a resistor of normal construction in the common cathode lead is invariably inadvisable, in view of the abnormally high negative biasing voltage then called for. It is a fortunate circumstance, however, that the only requirement imposed on R_k is that the quotient of the voltage change and the resultant current change should be high; in other words, a high *differential* resistance is wanted. The DC resistance (voltage divided by current) may permissibly be very much lower and indeed should be so, having regard to our remarks on the supply voltage needed.

In one of the following sections we shall consider circuits which in fact combine a very high differential resistance with a low DC resistance. First, however, we shall examine in more detail various problems arising from the necessity of using valves having a high amplification factor.

Difference amplifiers using pentodes

Most pentodes have a much higher amplification factor than conventional triodes. Where valves with a high amplification factor have to be used in a difference amplifier, our thoughts first turn therefore to the use of pentodes. The circuit diagram of a difference amplifier with pentodes is shown in *fig. 2*. (For the present it is sufficient to show a normal resistance R_k in the common cathode lead.)

The very high amplification factor of a pentode is only used to full advantage when, given a variable control-grid voltage, the potential difference between screen grid and cathode is kept constant and when changes in screen-grid current flow only partly or not at all through R_k. If the difference amplifier is intended solely for alternating voltages, this can be achieved by connecting a capacitor C between the screen grids and the cathodes in the usual way (see fig. 2). Where the amplifier is also required to deal with DC signals, a voltage-stabilizing valve can be used instead of a capacitor (see *fig. 3*). Since the operating voltage of such a valve is very little dependent on the current, a practically constant potential is maintained in this way between the screen grids and the cathode, and changes in screen-grid current do not flow through R_k to any significant extent.

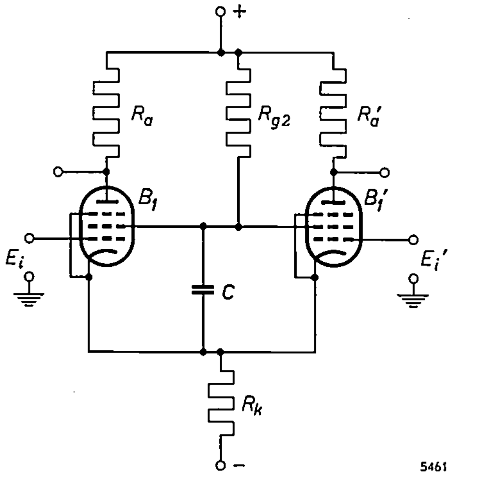


Fig. 2. Difference amplifier for AC signals, with two pentodes.

The above two conditions are never entirely fulfilled, for which reason $H_{\min}$ is smaller than follows from equation (7). The denominator of the complete equation in fact contains two extra terms, the first of which is due to the fact that the control-grid voltage and the screen-grid voltage do not necessarily influence the anode current in the same ratio in both valves. The second term is due to the possible spread in the distribution of cathode-current changes over anode and screen grid.

[4]) The valves are usually biased to obtain a low anode current, and hence a low mutual conductance, so that at the required value of R_k a lower negative supply voltage can be used. When the anode current is raised (and R_k correspondingly reduced) the mutual conductance increases relatively less, so that in spite of the higher S the product SR_k is no greater than when S is smaller.

A serious difficulty entailed by the use of pentodes arises from the flow of direct current to the screen grid. To obtain the required screen-grid potential the resistance R_{g2} must not exceed a specific value. (In fig. 3 the current of the voltage-stabilizing valve B_2 also flows through R_{g2}. In this case, therefore, the resistance in question must be even smaller than when a capacitor is used.) Since a constant potential difference is maintained between the screen grids and the cathodes, and also between the points marked + and —, the resistor R_{g2} may be regarded as in parallel with R_k as far as voltage and current changes are concerned. The differential resistance between cathode and earth is consequently reduced, and this causes, as we have shown above, a smaller value for the guaranteed rejection factor and discrimination factor. Methods that largely overcome this drawback will be discussed in the following sections.

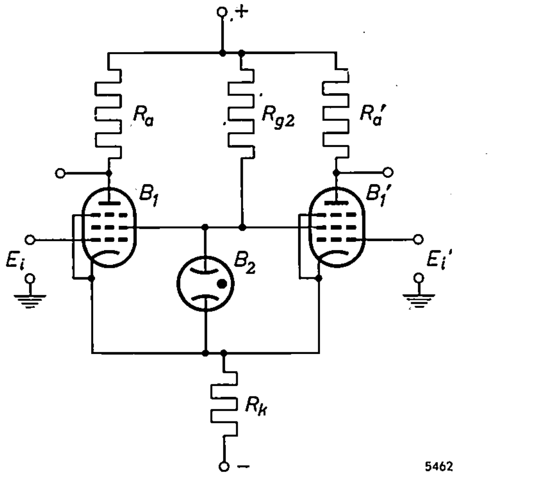


Fig. 3. Difference amplifier for DC signals, with two pentodes. A voltage-stabilizing valve B_2 is inserted between the screen grids and the cathodes.

The use of cascodes

An amplifier stage giving an amplification factor much higher than that of a triode can also be obtained by connecting two triodes in such a way as to produce a "cascode" circuit. The principle of such an arrangement is shown in *fig. 4*. The cathode of triode B_2 is connected to the anode of B_1. Provided the biasing voltages are so chosen that the valves operate in the normal part of their characteristics, a cascode exhibits properties closely resembling those of a pentode. The grid of triode B_2 functions in the circuit very much like the screen grid of a pentode. An important difference, however, is that screen-grid current always flows in a pentode, whereas in the "upper" valve in a cascode circuit no more than the usual, very low, grid current flows. This accounts for a marked advantage gained by using cascodes in difference amplifiers, to which we shall presently return.

A simple calculation shows that the transconductance of a cascode is practically identical with that of the "lower" valve:

$$S_{\text{casc}} = S_1 . \quad \ldots\ldots \quad (12)$$

The amplification factor μ_{casc} of the whole circuit is given by

$$\mu_{\text{casc}} = \mu_1(\mu_2 + 1) , \quad \ldots\ldots \quad (13)$$

where μ_1 and μ_2 are the amplification factors of the "lower" and "upper" valves, respectively. Since the amplification factors of normal valves are always very much larger than unity, we can write (13) to a very good approximation as

$$\mu_{\text{casc}} = \mu_1\mu_2 . \quad \ldots\ldots \quad (14)$$

The amplification factor of a cascode is thus high compared with that of each of the two valves of which it is composed.

A higher amplification factor, if required, can be obtained by using a cascode with more than two triodes. *Fig. 5* shows an example using three triodes. The gain of the cascode in this case is

$$\mu_{\text{casc}} = \mu_1(\mu_2 + 1)(\mu_3 + 1),$$

or, to a close approximation,

$$\mu_{\text{casc}} = \mu_1\mu_2\mu_3 .$$

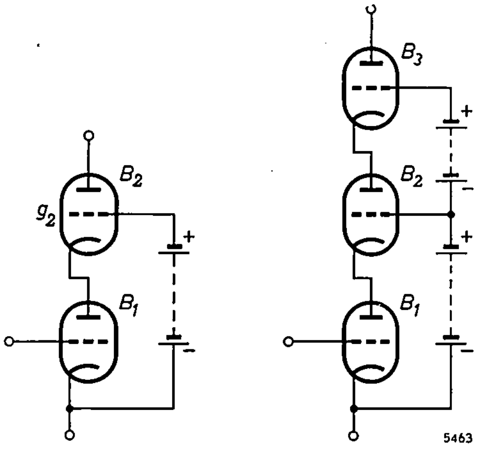


Fig. 4 Fig. 5

Fig. 4. Cascode consisting of two triodes.

Fig. 5. Cascode consisting of three triodes.

To obtain an amplification factor as in (13) with a circuit as shown in fig. 4, the grid g_2 of B_2 must have a constant potential with respect to the cathode of B_1. This is represented in the figure, for simplicity, as being produced by a battery. As a rule, of course, the bias for g_2 will be derived from the anode-voltage source by means of a voltage divider. *Fig. 6* shows a voltage divider for this purpose in a difference amplifier consisting of two cascodes. Because of the extremely low grid current flowing in the valves, the resistance R_1 can be given a very high value, e.g. a few megohms. To maintain the required constant potential between the grids of B_2 and B_2' and the cathodes of B_1 and B_1' a capacitor C is introduced, the effect of which is to transfer practically all sufficiently rapid voltage variations from the lower cathodes to the upper grids. For alternating voltages, then, resistor R_1 can be regarded as being in parallel with R_k, so that the differential resistance in the common cathode lead is again smaller than R_k. In view of the fact, however, that R_1 can be given a very high value, this drawback is less serious here than with the screen grids of pentodes.

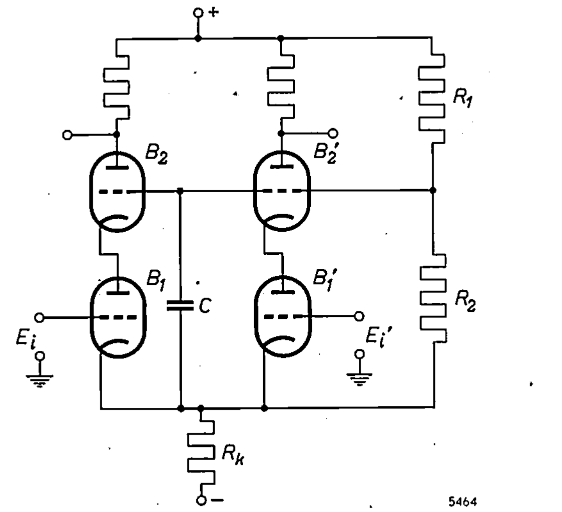


Fig. 6. Difference amplifier for AC signals, with two cascodes.

Cascodes are not so superior to pentodes when the difference amplifier is to be used for amplifying DC potential differences in the same way as the pentode circuit in fig. 3. The capacitor C will then be replaced by a voltage-stabilizing valve, and because of the direct current flowing in this valve the value of R_1 has to be made very much smaller. As in pentode circuits, arrangements can again be made that largely overcome this drawback. We shall touch on this under another heading.

It should be noted that the gain in H_min obtained by using cascodes instead of pentodes is less than would follow from equations (7) and (14). This is because the maximum spread in the amplification factor of a cascode is greater than that shown by the amplification factors for each valve individually. If μ_1 and μ_2 may each have a relative deviation δ from the nominal value, then the maximum relative deviation of μ_casc is equal to 2δ. When cascodes are used, therefore, the minimum guaranteed value of H is given not by equation (7) but by

$$H_\mathrm{min} = \frac{2}{\delta\left(\dfrac{1}{S_1R_\mathrm{k}} + \dfrac{2}{\mu_\mathrm{casc}}\right)}. \quad . \quad . \quad (15)$$

The term accounting for the influence of the amplification factor on H_min is thus only reduced by half the ratio between the amplification factors when we change over from triodes to cascodes.

It should be remarked that equation (15), like equation (7) in the case of pentodes, is only valid on the assumption that there are no voltage variations between the grids of the upper valves and the cathodes of the lower ones. In order to allow for the fact that this is never the case, and assuming that the voltage variations on the grids are k times those on the cathodes ($k < 1$), the value for μ_casc to be inserted in (15) should not be calculated from (13) or (14) but from:

$$\mu_\mathrm{casc} = \frac{\mu_1(\mu_2+1)}{(1-k)\mu_2+1}.$$

Since hardly any grid current flows in a cascode, H_min is not reduced by any spread in the distribution of the cathode current, as it is in the case of pentodes.

It may sometimes be regarded as a disadvantage of cascodes that the power-supply circuit has to deliver a higher voltage than for pentodes. This is a particular objection in multi-stage DC amplifiers with direct coupling between the stages (see part II of this article). It is even more of a drawback, of course, when cascodes consisting of more than two valves are used.

Circuits for obtaining a high differential resistance in the cathode lead

We have seen from a numerical example that the use of a normal resistance in the common cathode lead of the valves in a difference amplifier seldom deserves consideration, but that a component or circuit is required for this purpose whose DC resistance is low and whose differential resistance is very high. This requirement can largely be met by incorporating a pentode in the cathode lead (*fig.* 7). If the control

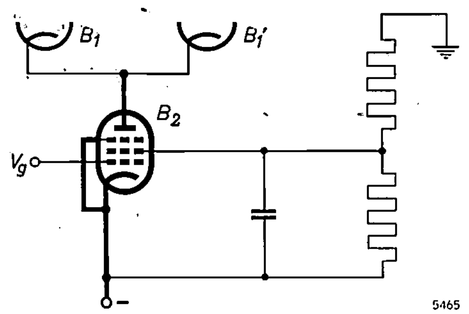


Fig. 7. Use of a pentode B_2 for producing a high differential resistance in the common cathode lead of the tubes B_1 and B_1'.

grid and screen grid are kept at the stipulated constant potentials with respect to the cathode, the differential resistance of a pentode is equal to its internal resistance, which may be more than 1 MΩ.

Another method of producing a high differential resistance in the common cathode lead is illustrated in *fig. 8*. The cathode lead of triode B_2 contains a resistance R_{k2}. If the grid of B_2 has a constant potential, the differential resistance R_d of the branch represented by thick lines in fig. 8 is given by

$$R_d = R_{i2} + (1 + \mu_2)R_{k2}, \quad . \quad . \quad . \quad (16)$$

where R_{i2} is the internal resistance and μ_2 the amplification factor of triode B_2. From eq. (16) we see that R_d is greater than $\mu_2 R_{k2}$. If, for example, the triode used has an amplification factor of 50 and R_{k2} is equal to 0.1 MΩ, then R_d will be greater than 5 MΩ.

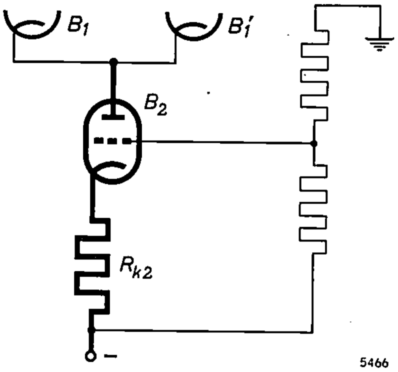


Fig. 8. Use of a triode B_2 with cathode resistance R_{k2} for producing a high differential resistance in the common cathode lead of the tubes B_1 and B_1'.

A pentode has a much higher amplification factor than a triode, and therefore R_d can be given a higher value if the triode B_2 in fig. 8 is changed for a pentode. We should bear in mind, however, that equation (16) is only valid for a pentode provided the changes in the anode current are equal to the changes in the current through the resistance in the cathode lead. Since there is always some screen-grid current flowing in a pentode, the only way to fulfil this condition is to insert between the screen grid and cathode either a capacitor of sufficiently high capacitance or a gas-discharge tube. In many cases it is then simpler to use a cascode arrangement, where there is no grid current and therefore no need for the above measure.

Fig. 9 shows a cascode, formed from two triodes B_2 and B_3, incorporated in the common cathode lead of valves B_1 and B_1'. The cathode lead of B_2 contains the resistor R_{k2}, and the grid voltages are kept constant by a voltage divider R_1-R_2-R_3. We have seen that the amplification factor of a cascode is greater than the product of the amplification factors of the two valves. For this reason the differential resistance of the branch represented by thick lines in fig. 9 is greater than $\mu_2\mu_3 R_{k2}$ (μ_2 and μ_3 being the amplification factors of triodes B_2 and B_3). Here too it is possible to use cascodes built up from more than two triodes, so that there is practically no limit to the value of the differential resistance which can be obtained in this way.

As explained above, the differential resistance in the common cathode lead of valves B_1 and B_1' does not consist solely of the thickly drawn branch in figures 7, 8 and 9, but also of a resistance in parallel with this branch and represented by R_{g2} in figures 2 and 3 and by R_1 in fig. 6. Since R_{g2} is much smaller than the required differential resistance (in DC amplifiers this also applies to R_1, as we have seen above), this is a severe obstacle to a high rejection factor. Two circuits that largely overcome this obstacle are represented in *figs 10* and *11*.

In fig. 10 the required constant potential difference between the cathodes and screen grids of B_1 and B_1' is not obtained by interposing a capacitor or voltage-stabilizing valve, but by means of a cathode follower B_3 and a coupling capacitor C. (Where the difference amplifier is also to amplify DC potential differences, a voltage-stabilizing valve should be substituted for the capacitor C.) If B_3 were an ideal cathode follower and the reactance of the capacitor C were negligible compared with the resistance R_{g2}, the voltage on the screen grids would completely "follow" the voltage variations on the cathodes. In reality, of course, this is never so. For that to be possible the differential resistance between the

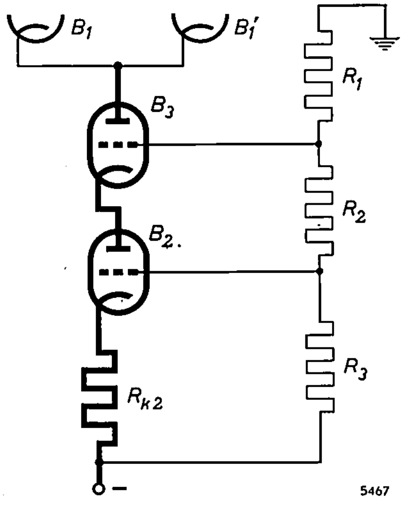


Fig. 9. Cascode with cathode resistance R_{k2} used for producing a high differential resistance in the common cathode lead of the tubes B_1 and B_1'.

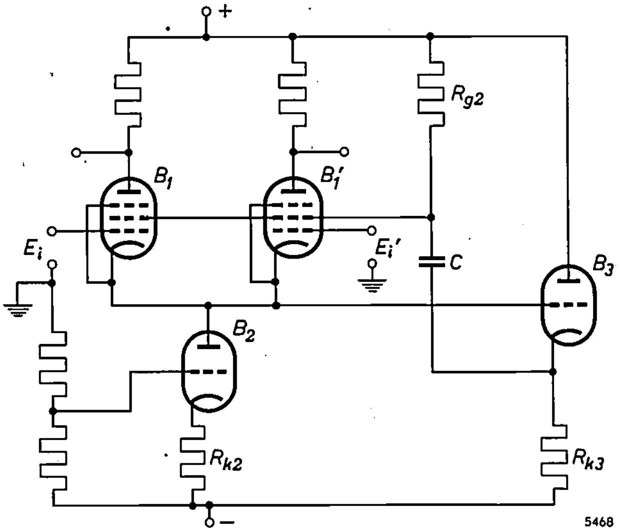


Fig. 10. Difference amplifier with pentodes B_1 and B_1'. The common cathode lead is given a high differential resistance by means of the triode B_2 and the resistor R_{k2}. The cathode follower B_3 enables the screen grids to follow the voltage fluctuations on the cathodes of B_1 and B_1'.

cathode of B_3 and earth would have to be infinitely high, and B_3 itself would have to have an infinitely large amplification factor. There will still, therefore, be slight voltage variations between the screen grids and the cathodes of B_1 and B_1', but owing to the fact that R_{g2} in fig. 10 has no influence on the common cathode resistance of these valves the cathode follower B_3 does in fact considerably improve the rejection factor and the discrimination factor.

The situation is even more improved if the cathode follower B_3 in fig. 10 is replaced by an amplifier whose gain approaches still closer to unity. In this way the influence of R_{g2} on the rejection and discrimination factors can be almost entirely eliminated.

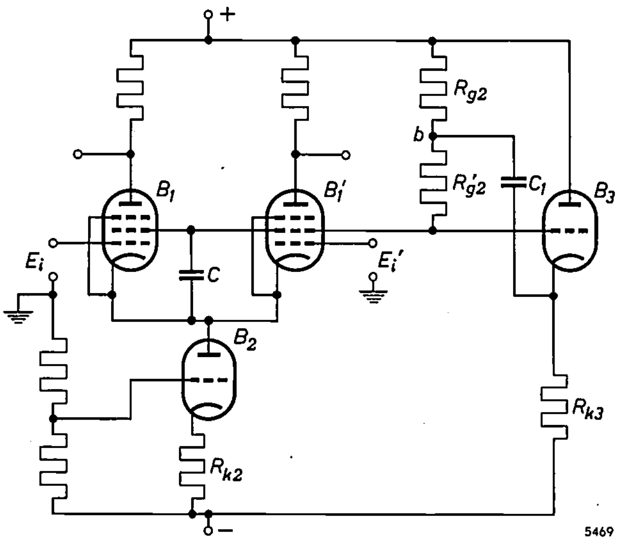


Fig. 11. Difference amplifier with pentodes B_1 and B_1'. The common cathode lead is given a high differential resistance by means of the triode B_2 and the resistor R_{k2}. The cathode follower B_3 raises the differential resistance between the screen grids and the anode-voltage source to a very high value.

In fig. 11 a capacitor C is again shown incorporated between the cathodes and screen grids of pentodes B_1 and B_1'. Given a high enough value of C, the voltage on the screen grids will almost completely follow the voltage variations on the cathodes. In that respect, then, this circuit has an advantage over that in fig. 10. The disadvantage, that the DC resistance of the screen grids reduces the differential resistance in the common cathode lead of B_1 and B_1', is largely overcome by the fact that the screen grids are here connected to the anode-voltage source by a circuit which has a very high differential resistance but a much smaller DC resistance. This circuit consists of the two resistors R_{g2} and R_{g2}' and a cathode follower B_3, which, via capacitor C_1, transmits the voltage variations on the screen grids almost entirely to point b. As a result the differential resistance between the screen grids and the anode-voltage source is much greater than $R_{g2} + R_{g2}'$, which can mean a considerable increase in the rejection and discrimination factors.

In this circuit too (fig. 11) the cathode follower B_3 can be changed for an amplifier whose gain is closer to unity. Use can also be made of circuits combining the principles of figs 10 and 11. Details of these circuits, however, are beyond the scope of this article.

The methods illustrated in figs 10 and 11 for increasing the differential resistance in the common cathode lead can also be adopted, of course, when cascodes are used in the differential amplifier instead of pentodes. As mentioned above, when cascodes are used these methods will generally be needed only when the differential amplifier is required to amplify DC potentials, in which case voltage-stabilizing valves must be used instead of capacitors, and the grids of the "upper" valves (see fig. 6) are therefore connected via a relatively small resistance to the anode-voltage source.

Some results of measurements

It will be clear from the foregoing that we can choose from a wide variety of circuit arrangements in order to design a difference amplifier whose rejection and discrimination factors can be guaranteed to have very high values. From the numerous possibilities, we have chosen three examples with a view to comparing the measured rejection factors with the minimum values calculated for these circuits.

Fig. 12 shows the circuit of a difference amplifier equipped with two E 80 F pentodes. The high differential resistance required in the common cathode lead is obtained with the triode B_2 and the resistor R_{k2}. By means of the cathode follower B_3 the voltage variations on the cathodes of B_1 and B_1' are transmitted to point b. The influence of the screen-grid resistance on the rejection factor is thus reduced here by a circuit which combines the principles illustrated in the figures 10 and 11. The two halves of a double triode ECC 81 are used for B_2 and B_3.

The minimum value of the rejection factor, assuming $\delta = 0.1$, is calculated to be 20 000 for this circuit. In *fig. 13* the calculated minimum is represented on a logarithmic scale together with the rejection factors measured on 25 arbitrary combinations of valves. The measurements were done at a frequency of 1 kc/s. The lowest value measured was 24 000.

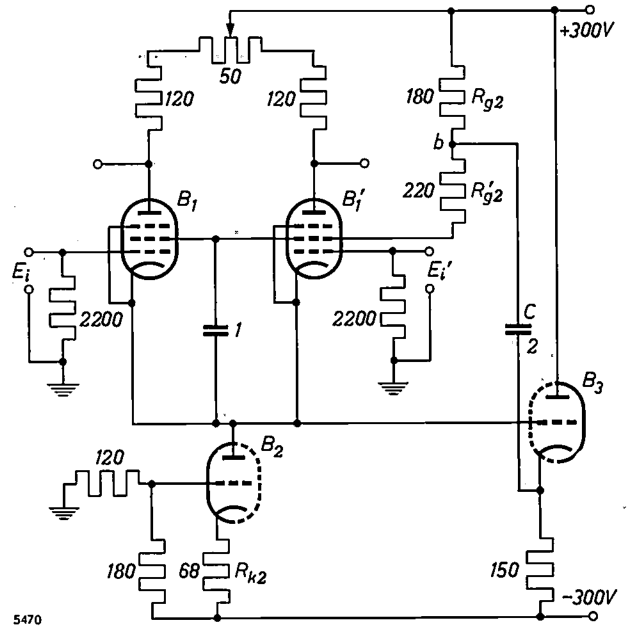


Fig. 12. Difference amplifier with two pentodes, type E 80 F. Triodes B_2 and B_3 are formed from the two halves of a double triode ECC 81. Resistances are given in kΩ, capacitances in μF. The 50-kΩ potentiometer is adjusted in such a way that in the steady state the anode voltages of the pentodes do not differ too much, giving both tubes roughly the desired operating point.

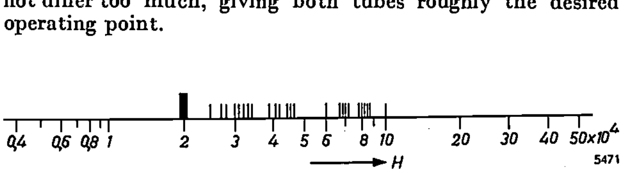


Fig. 13. Measured values of rejection factor H (thin marks) and calculated minimum value $H_{\min}$ (thick mark) for the circuit of fig. 12.

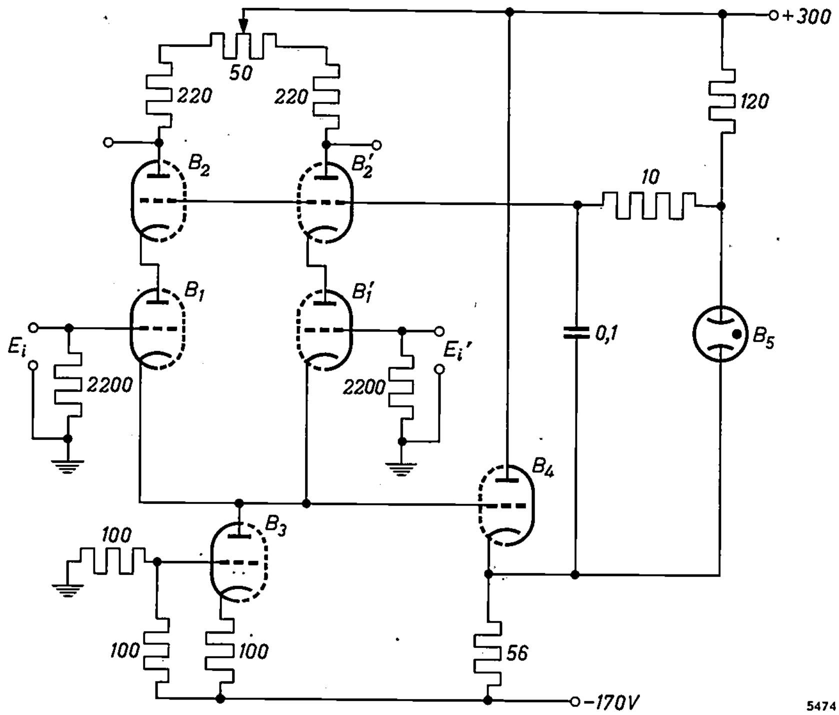


Fig. 14. Difference amplifier for DC signals, with two cascodes composed of two double triodes type E 80 CC. Triodes B_3 and B_4 are formed from the two halves of a double triode type ECC 81. B_5 is a voltage-stabilizing valve type 85 A 2. Resistances in kΩ, capacitances in μF.

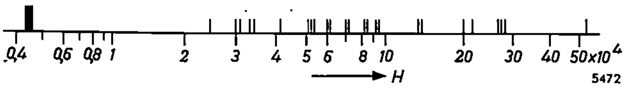


Fig. 15. Measured values of rejection factor H (thin marks) and calculated minimum value $H_{\min}$ (thick mark) for the circuit of fig. 14.

Fig. 14 gives the circuit diagram of a difference amplifier with two cascodes consisting of two double triodes, type E 80 CC. One half of a double triode ECC 81 is used for B_3 and the other half as a cathode follower, B_4, for transmitting the voltage variations on the cathodes of B_1 and B_1' to the grids of B_2 and B_2'. The transmission is effected by a voltage-stabilizing valve B_5 of type 85 A 2. Since the differential resistance of such a valve increases with increasing frequency, a capacitor of 0.1 μF is connected in parallel with B_5. The resistance of 10 kΩ between this capacitor and B_5 serves to correct instability effects likely to occur in a circuit of this kind.

Assuming that all quantities involved may show mutual deviations of 10%, we calculate a minimum value of 4500 for the rejection factor of this circuit. The values measured on 25 arbitrary combinations of valves are again shown on a logarithmic scale in *fig. 15*. The lowest value measured, 24 000, is much higher than the calculated minimum value.

The reason is that in fact the mutual disparity in the transconductance and amplification factor of a type E 80 CC double triode very seldom approaches 10%. Consequently there is a good chance that higher values of H will be found with circuits using this valve rather than other types whose S and μ are higher but show a greater mutual disparity.

The measurements whose results are represented in fig. 15 were carried out with direct voltage. If the difference amplifier is intended to handle alternating voltages only, the circuit can be considerably simplified. As already shown in fig. 6, the voltage for the grids of B_2 and B_2' can then be obtained from a voltage divider between the anode-voltage source and the cathodes of B_1 and B_1'. The voltage variations on the cathodes of the latter valves can be transmitted to the grids of B_2 and B_2' by means of a capacitor. *Fig. 16* represents the diagram of a circuit on which, with various combinations of valves, measurements were carried out at a frequency of 100 c/s. Two E 80 CC double triodes were used for B_1, B_1', B_2 and B_2', and half of a double triode ECC 81 was used for B_3. Again assuming 10% deviations in the values of all quantities involved, the minimum value calculated for the rejection factor was 6500. In *fig. 17* the measured values are again set out together with the calculated minimum value of H. The lowest value measured on 25 arbitrary combinations of valves was 24 000. Here, too, owing to the constancy of the E 80 CC double triode, this measured minimum is appreciably higher than the calculated value.

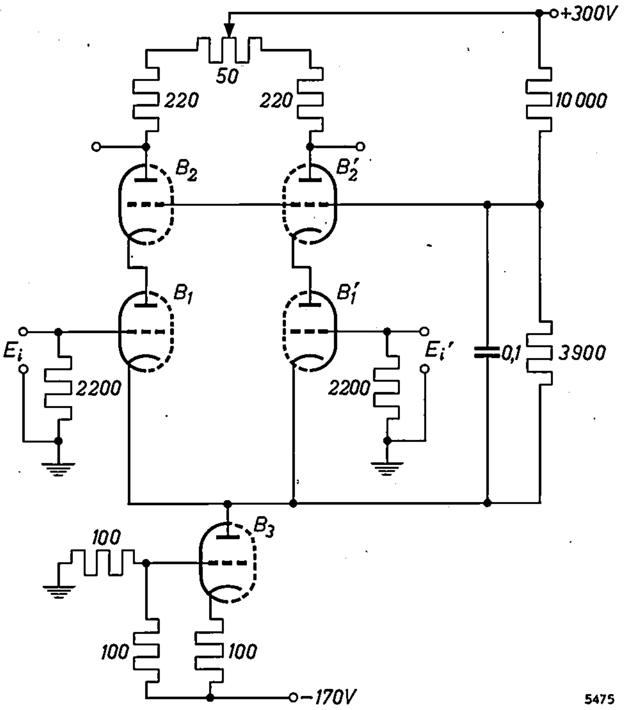


Fig. 16. Difference amplifier for AC signals, with two cascodes composed of two E 80 CC double triodes. B_3 is one half of an ECC 81 double triode. Resistances in kΩ, capacitances in μF.

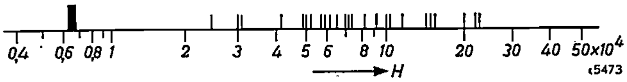


Fig. 17. Measured values of rejection factor H (thin marks) and calculated minimum value $H_{\min}$ (thick mark) for the circuit of fig. 16.

Part II of this article will deal with various problems that arise in the design of multi-stage difference amplifiers. Indications will also be given of the proper method of connecting the amplifier to points between which the potential difference is to be measured. Finally, some cases will be mentioned where difference amplifiers can also be used with advantage where the object is not to amplify the potential between two points.

Summary. An efficient single-stage difference amplifier can be produced by incorporating in the common cathode lead of the two valves of a balanced amplifier an element which has a very high differential resistance. The latter is necessary to obtain a high discrimination factor. In order to be able to guarantee a high rejection factor, even though the parameters of valves and other components show the maximum (adverse) deviation from their rated value, not only must the differential resistance in the cathode lead be high but the valves must also have a high amplification factor. The valves used may be either pentodes, or triodes in a cascode arrangement. The voltage on the screen grids of the pentodes or on the grids of the "upper" triodes in the cascodes must follow the voltage variations on the cathodes closely. Some circuits which achieve this result are described. The high differential resistance in the common cathode lead can be obtained with circuits using pentodes, triodes or cascodes. Measurements on three circuits are discussed, and the calculated minimum rejection factor is compared with the values measured on 25 arbitrary combinations of valves.

LOOP GAIN AND STABILITY OF SIMPLE CONTROL SYSTEMS

by M. van TOL *).

621-53.001

In the development of control theory during the last twenty years, two main lines may be distinguished. On the one hand refined mathematical methods have been introduced, such as the use of Laplace transforms, and on the other hand approximations have led to simple rules of thumb, which are of considerable practical value.

In the following article such a rule of thumb is given for finding the relationship between the time constants of a control system and the maximum loop gain which the stability of the system permits. This rule, which has not as far as we know been explicitly formulated elsewhere, makes it possible to show the relationship between the various methods employed to improve the performance of a control system.

The factor P by which a disturbance is reduced in an automatic control system depends on the transfer function $KG(j\omega)$ of the open control loop. In a previous article [1]) it was shown that P is equal to the absolute value of the complex quantity $1 + KG(j\omega)$ and therefore approaches $1 + K$ for very low frequencies. (We assume provisionally that $G(0)$ is equal to unity.) A step-function disturbance is therefore ultimately corrected by the system except for a portion $1/(1+K)$. Since this portion is smaller the larger the valuc of K, the aim is to make K as large as possible.

There is a limit, however, to the possible increase of K. Since the loop gain at any given frequency is proportional to K, a higher K results in a rise in gain at the frequency ω_c at which the phase shift φ is —180°. When this gain exceeds unity the system is no longer stable. In designing a control system — and the same applies to negative-feedback amplifiers and servo-mechanisms — it is therefore necessary to make $KG(0)$ high enough to reduce disturbances effectively, at the same time ensuring that $|KG(j\omega_c)| < 1$. The problem thus amounts to finding a suitable form for the function $G(j\omega)$. The measures taken to that end must also ensure that the response of the system is not impermissibly slow.

Control engineers have long used certain established methods of improving the performance of control systems. In this article we shall show that all these methods, between which there has hitherto seemed to be little or no connection, may be regarded as applications of one and the same principle. That principle is directly derived from a very simple formula which defines the relationship between the maximum permissible value of K and the two longest characteristic times τ that define the behaviour of the elements of the control loop [2]). Although this formula is not exact unless the control loop meets certain requirements, it can be shown to be of practical value in numerous cases where these requirements are not completely satisfied.

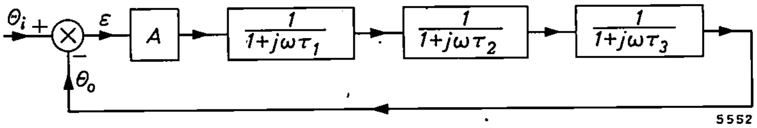

Θ_i
+
ε
−
A
$\frac{1}{1+j\omega\tau_1}$
$\frac{1}{1+j\omega\tau_2}$
$\frac{1}{1+j\omega\tau_3}$
Θ_o
5552

Fig. 1. Block diagram of a control system consisting of a linear amplifier, gain factor A, and elements having the transfer function $(1 + j\omega\tau_n)^{-1}$. In the usual terminology Θ_i is the desired value, Θ_o the instantaneous value of the controlled condition (output quantity) which is compared with Θ_i, and ε, the deviation, is their difference.

The basic formula

Let us consider a control loop that can be represented by the block diagram in *fig. 1*. Apart from the

*) Research Laboratories, Eindhoven.

[1]) M. van Tol, Application of control theory to linear control systems, Philips tech. Rev. **23**, 109-118, 1961/62 (No. 4).

[2]) See also: M. van Tol, Stability and optimal loopgain of simple control circuits, Transactions of the 5th International Instruments and Measurements Conference, Stockholm 12-16 Sept. 1960, or M. van Tol, Control engineering **8**, 1961, in the press.

controlling unit, a linear amplifier with gain factor A, we may write for all blocks:

$$K_n = 1, \quad \ldots \ldots \ldots \quad (1)$$

$$G_n = \frac{1}{1 + j\omega\tau_n}. \quad \ldots \ldots \quad (2)$$

The amplitude-frequency response characteristics of each of these blocks (which we shall here refer to as amplitude characteristics for short) when plotted logarithmically in a Bode diagram, approximate to two straight lines intersecting on the vertical $\omega = 1/\tau_n$. The first, which approximates to the low-frequency characteristic, is horizontal; the other has a slope of —1. The phase shift φ between output and input signal varies with increasing ω from 0 to —90°, and at $\omega = 1/\tau_n$ is exactly —45° (*fig. 2*).

The slope of these two lines can be found by writing the differential quotient d log $|G|$/d log ω as a function of ω, and then ascertaining how this function behaves where $\omega \ll 1/\tau_n$ and where $\omega \gg 1/\tau_n$. Using the formula

$$|G| = 1/\sqrt{1 + \omega^2\tau_n^2},$$

we find

$$\frac{\mathrm{d} \log |G|}{\mathrm{d} \log \omega} = \frac{-\omega^2\tau_n^2}{1 + \omega^2\tau_n^2}.$$

Where ω is small the differential quotient is thus ~ 0; where ω is large it approaches —1. The line $|G| = 1/\omega\tau_n$ intersects the other at the frequency for which the value of the ordinate is unity. As can be seen, that frequency is $\omega = 1/\tau_n$.

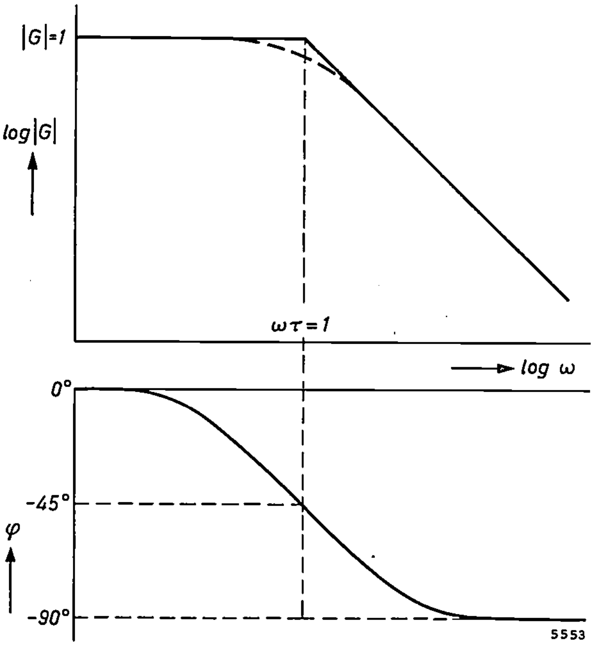


Fig. 2. Bode diagram of an element having the transfer function $G = (1 + j\omega\tau)^{-1}$. The amplitude-frequency response characteristic is approximated by two straight lines intersecting at the vertical $\omega = 1/\tau$. The phase shift φ is exactly —45° at the frequency $1/\tau$ and has a maximum value of —90°.

The amplitude characteristic in logarithmic coordinates of the complete open loop can very simply be derived from that of the individual blocks, every multiplication being an addition on logarithmic paper. This characteristic, too, can thus be approximated by straight lines.

If the various time constants τ_n are found to differ considerably — e.g. by a factor of 5 or more — the phase-frequency response characteristic (or phase characteristic for short) can also be found to a good approximation by adding the phase characteristics of the individual blocks. The phase shift is then roughly —45° for $\omega = 1/\tau_1$, about —135° for $\omega = 1/\tau_2$, about —225° for $\omega = 1/\tau_3$ and so on (*fig. 3*).

As mentioned above, for a stable system $|KG|$ should be less than unity at the frequency ω_c where φ is —180°. In practice, this limit should not be too closely approached; the closer the phase shift at the frequency for which $|KG| = 1$ approaches —180°, the poorer is the damping of the system and therefore the longer the oscillations last. For practical purposes a phase "margin" of 45° is found to be sufficient to guarantee stability. In other words, the loop gain drops below unity at the frequency where the phase shift is —135°. In the event of a step-function transient the overshoot Θ_s' is then no more than about 15% (see *fig. 4*). As can be seen from fig. 3, a phase shift of —135° occurs at exactly $\omega = 1/\tau_2$. To meet the stability requirement, then,

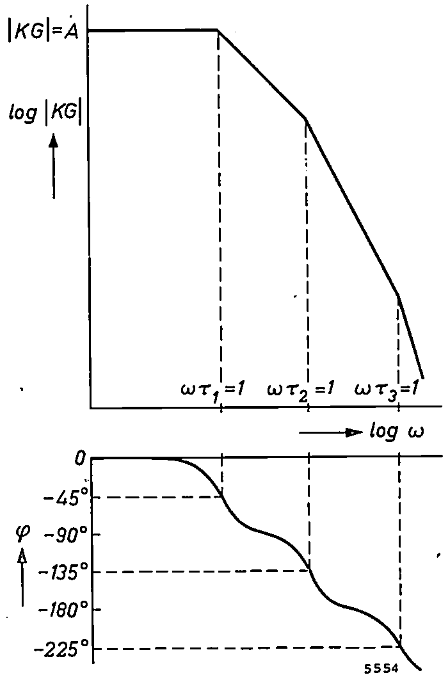


Fig. 3. Bode diagram (approximation) of an open control loop consisting of a linear amplifier (gain A) and various elements having the transfer function $(1 + j\omega\tau_n)^{-1}$, where $\tau_1 \gg \tau_2 \gg \tau_3$ etc.

A must not be greater than the value $A_{\max}$ which causes the response to reach unity at the second break in the amplitude characteristic (*fig.* 5).

As we have seen, the schematic amplitude characteristic of a single block has a slope of —1 for frequencies greater than $1/\tau_1$. The same therefore holds for the section PQ (fig. 5) of the amplitude characteristic of the entire (open) control loop. The triangle PQR is thus isosceles: $PR = QR$. Hence:

$$\log A_{\max} - \log 1 = \log 1/\tau_2 - \log 1/\tau_1, \qquad (3)$$

or
$$A_{\max} = \frac{\tau_1}{\tau_2}. \qquad (4)$$

Put into words: *to ensure stability, the open loop gain must not be greater than the quotient of the longest two time constants.*

The principle underlying nearly all methods that can be used to improve the performance of an automatic control system follows directly from equation (4). If the system is not sufficiently stable, then A is too large, or in other words the quotient τ_1/τ_2 is too small. The latter must then be increased until it is at least equal to A. If disturbances are not adequately suppressed, then A must be increased, but at the same time, to maintain stability, τ_1/τ_2 must also be raised so that it remains at least equal to A. *In both cases, then, the rule is to try to increase the quotient* τ_1/τ_2.

Although, as appears from the derivation of eq. (4), this rule applies only to control loops where all blocks have the transfer function $K_n/(1 + j\omega\tau_n)$, and where $\tau_1 \gg \tau_2 \gg \tau_3$ etc., it is nevertheless a useful starting point for many practical cases.

Applications

The methods commonly used to improve an unsatisfactorily operating control system are: to change one of the time constants (if applicable, an exceptionally simple method); to introduce an element giving a negative phase shift or one giving a positive phase shift; to introduce an integrating element or a differentiating element, or a combination of both.

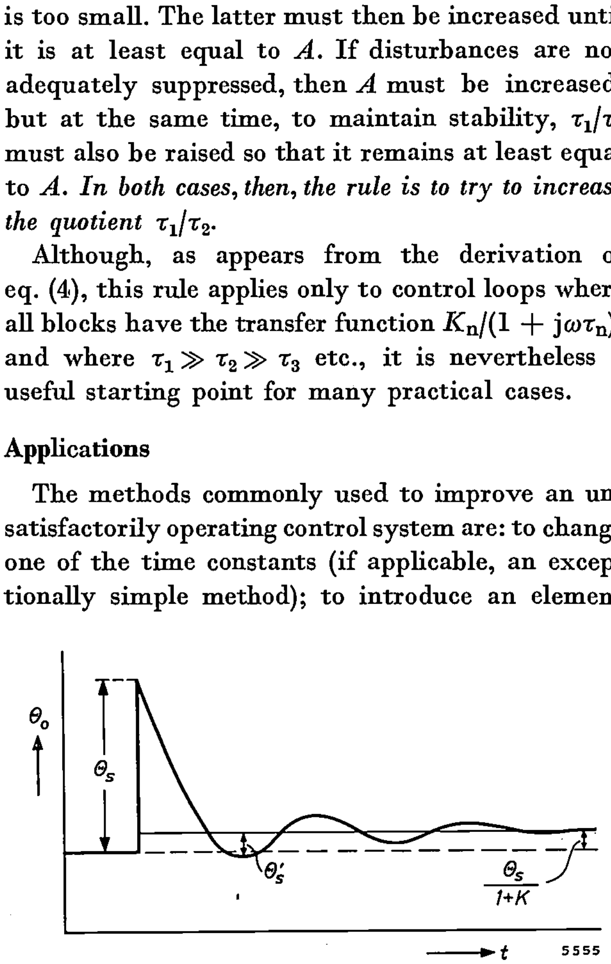


Fig. 4. With a "phase margin" of 45°, a step-function disturbance of magnitude Θ_S gives rise to an overshoot Θ_S' no greater than about 0.15 Θ_S.

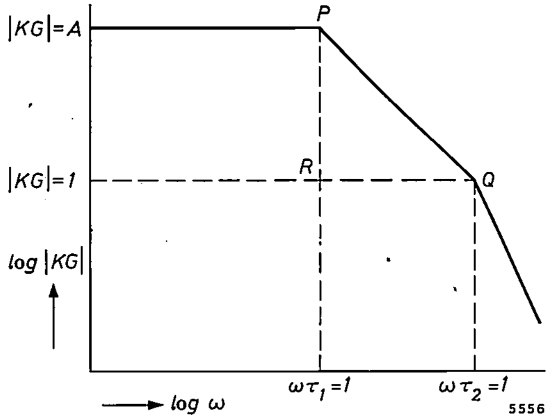


Fig. 5. With a phase margin of 45° the loop gain $|KG|$ is exactly equal to unity when $\omega = 1/\tau_2$, i.e. at the second break Q in the logarithmic amplitude-frequency response characteristic approximated by straight lines. The slope of the part between the first break (P) and the second is —1.

We shall now show that all these measures amount to increasing the quotient τ_1/τ_2. As far as the first method is concerned — changing one of the time constants — this is immediately evident: τ_1 is increased or τ_2 decreased (the latter is only possible, of course, provided τ_3 is sufficiently small.) The other methods require somewhat more detailed analysis.

Introduction of a phase-shifting element

The electrical or pneumatic circuits used for producing a negative phase shift for the above purpose — the electrical ones are known as phase-lag networks [3]) — have the transfer function

$$G' = \frac{1 + j\omega\tau_b}{1 + j\omega b\tau_b}, \qquad (5)$$

where $b > 1$. If we make τ_b equal to τ_1, the longest time constant of the control loop, the transfer function of the combination of the block with time constant τ_1 and the phase-lag network is:

$$G_1' = G' \times G_1 = \frac{1 + j\omega\tau_1}{1 + j\omega b\tau_1} \times \frac{1}{1 + j\omega\tau_1} = \frac{1}{1 + j\omega b\tau_1}. \qquad (6)$$

The unit with time constant τ_1 has thus as it were

[3]) A description of these networks will be found in books dealing with the theory of electrical circuits. See also e.g. G. S. Brown and D. P. Campbell, Principles of servomechanisms, Wiley, New York 1948, chapter 7; or R. A. Bruns and R. M. Saunders, Analysis of feedback control systems, McGraw-Hill, New York 1955, chapters 6 and 13.

been replaced by a similar block having the time constant $b\tau_1$. After introducing such a network, we can therefore increase the gain by a factor b without reducing stability. If τ_1 is already very great it may not be possible to use this method, since it may not be possible to make a phase-shifting element with the even greater time constant $b\tau_1$.

Using a "phase-lead network" — also called a damping network because it generally increases the damping in the closed loop — the value of τ_2 is reduced in a corresponding manner. The transfer function of such an arrangement [3]) is:

$$K''G'' = \frac{1}{c}\,\frac{1+\mathrm{j}\omega\tau_c}{1+\mathrm{j}\omega\tau_c/c}, \quad . \; . \; . \; . \quad (7)$$

where $c > 1$. Choosing $\tau_c = \tau_2$, the effect is as if the time constant τ_2 in the open loop had been changed to τ_2/c. The loop gain may therefore be increased by the factor c.

It is a condition here, as in the method where τ_2 itself is reduced, that τ_3 should not approach too closely to τ_2.

Since the introduction of the network in itself decreases the gain by a factor c, when this method is used the total extra gain must be c^2. The fact that the phase-lead network attenuates the signal may sometimes be a drawback; in some cases the attenuated signal may no longer be strong enough in relation to the input noise of the amplifier.

In practice, b and c in these two methods are given values of the order of 10.

The introduction of an integrating or differentiating element

The conventional method of introducing an integrating or differentiating element in a control loop is illustrated in *fig. 6*. The signal ε_o applied to block IV is not simply the A-times amplified deviation $\Theta_i - \Theta_o$, which we shall call ε_i, but contains a component originating from the integrating or differentiating element (block *III*). If block *III* is an integrating element, we speak of P.I. control (P = proportional, I = integral); if *III* is a differentiating element, we speak of P.D. control (D = derivative).

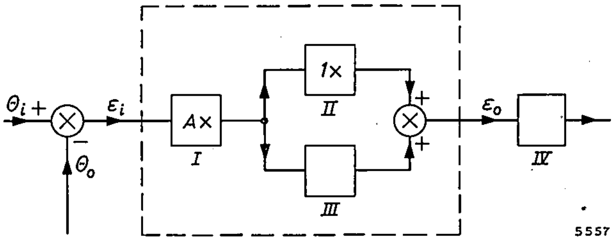


Fig. 6. Method of introducing an integrating or differentiating element (block *III*) in a control loop (P.I. control and P.D. control, respectively). *I* is a linear amplifier, gain factor *A*. *II* ditto, with gain factor equal to unity.

In the case of P.I. control we have:

$$\varepsilon_o = A\left\{\varepsilon_i + \frac{1}{\tau_i}\int_0^t \varepsilon_i \; dt\right\}. \quad . \; . \quad (8)$$

For sinusoidal signals the transfer function of the whole assembly of units inside the dotted square is given by:

$$KG_{\text{P.I.}} \equiv \frac{\varepsilon_o}{\varepsilon_i} = A\left(1+\frac{1}{\mathrm{j}\omega\tau_i}\right) = A\,\frac{1+\mathrm{j}\omega\tau_i}{\mathrm{j}\omega\tau_i}. \quad (9)$$

Now the integrating action of practical networks is never perfect. Given a constant input signal, the output signal does not go on increasing indefinitely but finally reaches a constant value. In many cases the gain is not infinite at zero frequency, but has a finite value which we shall call B. The transfer function is then not $1/\mathrm{j}\omega\tau_i$ but

$$\frac{1}{(1/B)+\mathrm{j}\omega\tau_i}.$$

The transfer function $KG_{\text{P.I.}}$ is therefore not given in such cases by eq. (9) but by

$$KG_{\text{P.I.}} = A\left\{1+\frac{1}{(1/B)+\mathrm{j}\omega\tau_i}\right\} =$$

$$= AB^*\,\frac{1+\mathrm{j}\omega\tau_i^*}{1+\mathrm{j}\omega B^*\tau_i^*}, \quad (10)$$

where $B^* = B + 1$ and $\tau_i^* = B\tau_i/(B+1)$.

It will be seen that this transfer function has the same form as that of a phase-lag network. If we choose τ_i such that $\tau_i^* = \tau_1$, then τ_1 in the loop is again apparently increased to $B^*\tau_i^*$, i.e. to $B\tau_i$. A difference is that the factor B can be much larger than the maximum feasible value of b.

The non-ideal behaviour of practical integrators can be briefly explained with the aid of two examples. The first is a capacitor charged by a current. The voltage cannot go on rising indefinitely for two reasons: *a*) the sheathing materials of the capacitor are not perfectly insulated from one another, and at very high tensions the leakage current is no longer negligible; *b*) no current source can continue to supply current under unlimitedly high inverse voltage. The same applies, *mutatis mutandis*, to pneumatic integrators. It makes no formal difference which of the two causes dominates in a given case. The second example is an electric motor that cannot go on turning indefinitely but must stop as soon as e.g. the sliding contact of the potentiometer which it operates reaches the end of the resistance. The cause of the non-ideal behaviour is of a different nature in both examples. In the first case we have an unwanted extra time constant, but the element remains linear; in the second example non-linearity occurs.

In both cases ideal operation is approached only when the output signal is not unduly large. In arriving at equation (10) we envisaged an integrator of the type of the first example.

In the case of P.D. control, the transfer function $KG_{\text{P.D.}}$ is found from:

$$\varepsilon_o = A\left(\varepsilon_i + \tau_d \frac{d\varepsilon_i}{dt}\right). \quad . \quad . \quad . \quad (11)$$

For sinusoidal signals this gives:

$$KG_{\text{P.D.}} \equiv \frac{\varepsilon_o}{\varepsilon_i} = A(1 + j\omega\tau_d). \quad . \quad . \quad (12)$$

Here again, the actual situation differs somewhat from the theoretical, for the differentiating action ceases when the frequency exceeds a particular value. The transfer function in practice is therefore given by:

$$KG_{\text{P.D.}} = A\,\frac{1 + j\omega\tau_d}{1 + j\omega\tau_d/C}, \quad . \quad . \quad . \quad (13)$$

where $C \gg 1$.

Just as with the phase-lead network, the apparent effect is a reduction of τ_2. Again the difference is that C can be very much greater than the factor c. Both B and C are usually of the order of 100.

As mentioned, a combination of two or more of the methods discussed above can be used. A familiar example is a P.I.D. controller, a device which has both an integrating and a differentiating element in parallel with the linearly amplifying element. In this case we should choose $\tau_i = \tau_1$ and $\tau_d = \tau_2$. This too constitutes a convenient rule of thumb for adjusting the controller in many of the situations encountered in practice.

A simple rule can also be given for the effect which the measures discussed have on the speed with which the control system reaches the new steady state after a step-function disturbance. The rule is: *if the gain A is kept equal to τ_1/τ_2, an increase in τ_1 at constant τ_2 does not change the speed of response, whereas a decrease in τ_2 at constant τ_1 makes the response faster.* The truth of this can quickly be verified. In broad terms, the response time of low-pass filters — to which category controlled systems belong — is equal to the reciprocal of the highest frequency f_0 passed through[4]). Now if A is τ_1/τ_2 the highest transmitted angular frequency ω_0 is given by $\omega_0\tau_2 = 1$. This applies not only to the open loop (cf. fig. 5) but equally to the closed loop, whose amplitude characteristic is given by the formula:

$$\frac{\Theta_o}{\Theta_i} = \frac{KG}{1 + KG}.$$

The response time is therefore proportional to τ_2 and is not affected by variations in τ_1 — at any rate as long as τ_1 is several times larger than τ_2.

Summarizing, it may be said that the basic formula $A_{\max} = \tau_1/\tau_2$ gives a quick indication of what can be done in a particular case to improve the system, and in addition gives a good general insight into the merit of currently used corrective methods and the relationship between them.

[4]) See p. 352 of the book by Brown and Campbell, ref. [3]), or W. C. Elmore and M. Sands, Electronics, McGraw-Hill, New York 1949, p. 139.

Summary. For a control system to be stable the open loop gain $|KG(j\omega)|$ must be less than unity at the frequency where the phase shift φ is $-180°$. In practice the control engineer adopts a safety margin of 45° and attempts to make $|KG| < 1$ when $\varphi = -135°$. In control loops where all the elements, apart from a linear amplifier, have a transfer function $(1 + j\omega\tau_n)^{-1}$, and the time constants $\tau_1 \gg \tau_2 \gg \tau_3$ etc., this requirement is found to be fulfilled if $|KG(0)| \leqq \tau_1/\tau_2$. The formula is also applicable in cases where the above conditions are not entirely met. All known methods of improving a control system, viz: changing one of the time constants, using a phase-lag network or a phase-lead network, integrating element (with $\tau_i = \tau_1$) or differentiating element (with $\tau_d = \tau_2$), are shown to boil down to increasing τ_1/τ_2. If $|KG(0)|$ is made equal to τ_1/τ_2, the response time of the controller is roughly equal to τ_2 and independent of τ_1.

AN ELECTRICALLY SCREENED ROOM FOR MICROWAVE EXPERIMENTS

621.317.2:621.319.74

Experimental work in radio engineering is often hampered by interference caused by radiation from installations in other parts of the laboratory. This can be avoided by doing the experiments in a space completely enclosed by a conducting screen (metal sheet or gauze), which strongly attenuates extraneous electromagnetic radiation, partly by reflection and partly by absorption (joule heating due to alternating currents induced in the screen).

The first constructions of this kind made in our laboratories consisted simply of double wire-netting fixed around a wooden frame, which enclosed a space of a few cubic metres [1]). For certain purposes, however, the screening provided by these "cages" proved to be inadequate, particularly in the decimetre and

[1]) Part of a screened room of this kind can be seen in Philips tech. Rev. 14, 121, 1952/53 (fig. 4).

In both cases ideal operation is approached only when the output signal is not unduly large. In arriving at equation (10) we envisaged an integrator of the type of the first example.

In the case of P.D. control, the transfer function $KG_{\mathrm{P.D.}}$ is found from:

$$\varepsilon_\mathrm{o} = A\left(\varepsilon_\mathrm{i} + \tau_\mathrm{d}\frac{\mathrm{d}\varepsilon_\mathrm{i}}{\mathrm{d}t}\right). \quad . \; . \; . \quad (11)$$

For sinusoidal signals this gives:

$$KG_{\mathrm{P.D.}} \equiv \frac{\varepsilon_\mathrm{o}}{\varepsilon_\mathrm{i}} = A(1 + \mathrm{j}\omega\tau_\mathrm{d}). \quad . \; . \quad (12)$$

Here again, the actual situation differs somewhat from the theoretical, for the differentiating action ceases when the frequency exceeds a particular value. The transfer function in practice is therefore given by:

$$KG_{\mathrm{P.D.}} = A\,\frac{1 + \mathrm{j}\omega\tau_\mathrm{d}}{1 + \mathrm{j}\omega\tau_\mathrm{d}/C}, \quad . \; . \; . \quad (13)$$

where $C \gg 1$.

Just as with the phase-lead network, the apparent effect is a reduction of τ_2. Again the difference is that C can be very much greater than the factor c. Both B and C are usually of the order of 100.

As mentioned, a combination of two or more of the methods discussed above can be used. A familiar example is a P.I.D. controller, a device which has both an integrating and a differentiating element in parallel with the linearly amplifying element. In this case we should choose $\tau_\mathrm{i} = \tau_1$ and $\tau_\mathrm{d} = \tau_2$. This too constitutes a convenient rule of thumb for adjusting the controller in many of the situations encountered in practice.

A simple rule can also be given for the effect which the measures discussed have on the speed with which the control system reaches the new steady state after a step-function disturbance. The rule is: *if the gain A is kept equal to τ_1/τ_2, an increase in τ_1 at constant τ_2 does not change the speed of response, whereas a decrease in τ_2 at constant τ_1 makes the response faster.* The truth of this can quickly be verified. In broad terms, the response time of low-pass filters — to which category controlled systems belong — is equal to the reciprocal of the highest frequency f_0 passed through[4]. Now if A is τ_1/τ_2 the highest transmitted angular frequency ω_0 is given by $\omega_0\tau_2 = 1$. This applies not only to the open loop (cf. fig. 5) but equally to the closed loop, whose amplitude characteristic is given by the formula:

$$\frac{\Theta_\mathrm{o}}{\Theta_\mathrm{i}} = \frac{KG}{1 + KG}.$$

The response time is therefore proportional to τ_2 and is not affected by variations in τ_1 — at any rate as long as τ_1 is several times larger than τ_2.

Summarizing, it may be said that the basic formula $A_{\max} = \tau_1/\tau_2$ gives a quick indication of what can be done in a particular case to improve the system, and in addition gives a good general insight into the merit of currently used corrective methods and the relationship between them.

[4] See p. 352 of the book by Brown and Campbell, ref. [3]), or W. C. Elmore and M. Sands, Electronics, McGraw-Hill, New York 1949, p. 139.

Summary. For a control system to be stable the open loop gain $|KG(\mathrm{j}\omega)|$ must be less than unity at the frequency where the phase shift φ is $-180°$. In practice the control engineer adopts a safety margin of 45° and attempts to make $|KG| < 1$ when $\varphi = -135°$. In control loops where all the elements, apart from a linear amplifier, have a transfer function $(1 + \mathrm{j}\omega\tau_\mathrm{n})^{-1}$, and the time constants $\tau_1 \gg \tau_2 \gg \tau_3$ etc., this requirement is found to be fulfilled if $|KG(0)| \leqq \tau_1/\tau_2$. The formula is also applicable in cases where the above conditions are not entirely met. All known methods of improving a control system, viz: changing one of the time constants, using a phase-lag network or a phase-lead network, integrating element (with $\tau_\mathrm{i} = \tau_1$) or differentiating element (with $\tau_\mathrm{d} = \tau_2$), are shown to boil down to increasing τ_1/τ_2. If $|KG(0)|$ is made equal to τ_1/τ_2, the response time of the controller is roughly equal to τ_2 and independent of τ_1.

AN ELECTRICALLY SCREENED ROOM FOR MICROWAVE EXPERIMENTS

621.317.2:621.319.74

Experimental work in radio engineering is often hampered by interference caused by radiation from installations in other parts of the laboratory. This can be avoided by doing the experiments in a space completely enclosed by a conducting screen (metal sheet or gauze), which strongly attenuates extraneous electromagnetic radiation, partly by reflection and partly by absorption (joule heating due to alternating currents induced in the screen).

The first constructions of this kind made in our laboratories consisted simply of double wire-netting fixed around a wooden frame, which enclosed a space of a few cubic metres[1]. For certain purposes, however, the screening provided by these "cages" proved to be inadequate, particularly in the decimetre and

[1] Part of a screened room of this kind can be seen in Philips tech. Rev. 14, 121, 1952/53 (fig. 4).

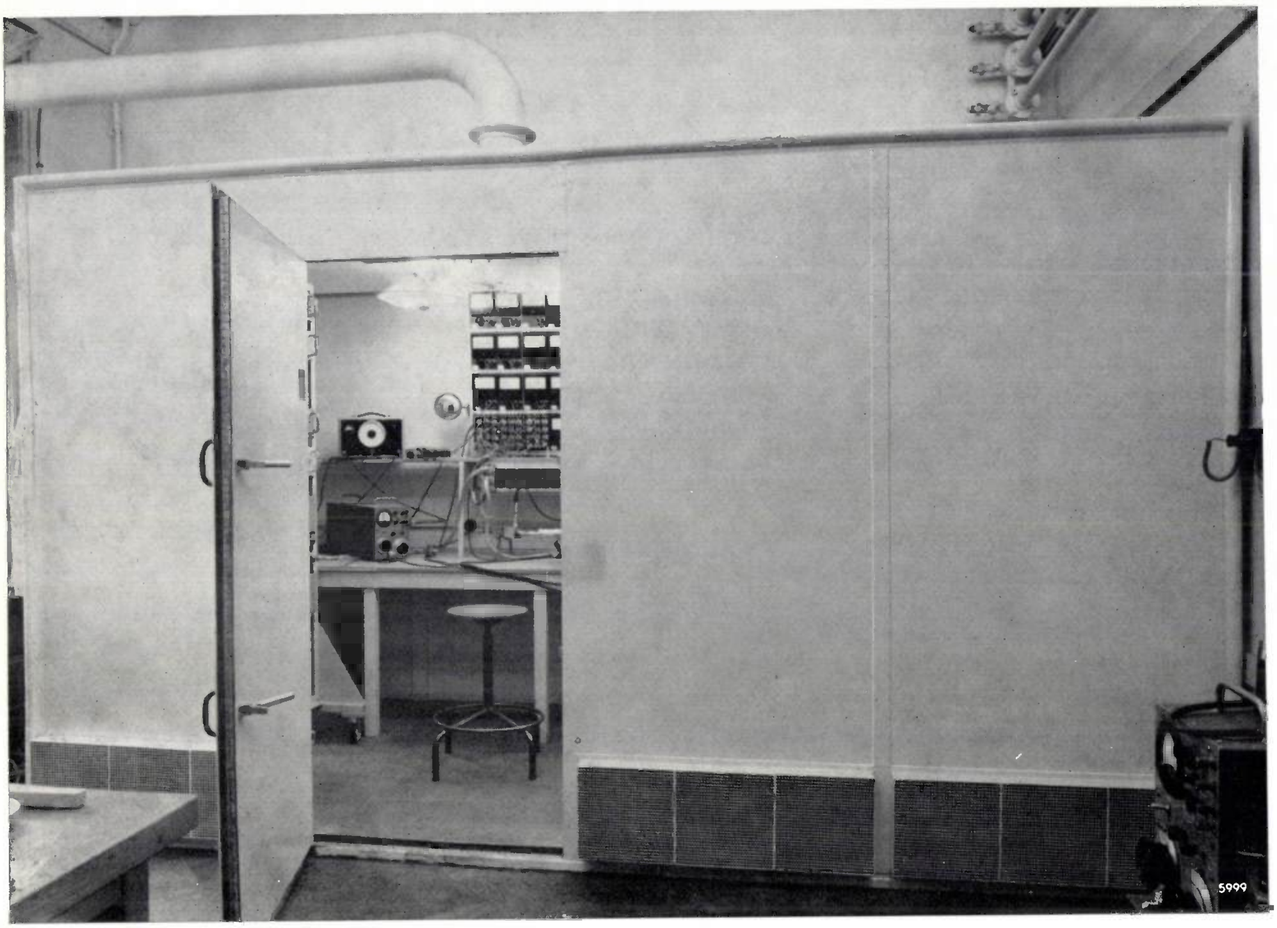


Fig. 1. General view of the electrically screened room. Floor, walls and ceiling are built up from sections each consisting of a welded angle-steel frame to which 1-mm copper sheet is screwed. The frames too are screwed together. To ensure good screening, the joints between the sections and the gaps between the door and the door-frame had to be sealed. To this end lengths of brass strip (2 mm thick, 60 mm wide) were soldered to the joints between the sections, first being screwed on in order to avoid strains during soldering. The gaps round the door were dealt with as follows. The copper sheet that forms the door was fixed to a welded angle-brass frame, around the whole periphery of which were soldered sprung strips of silver-plated phosphor-bronze (see photograph). When the door is closed, these strips press against a silver-plated copper cornice around the door frame. Good electrical connection between the cornice and the copper wall of the room is also ensured by soldering.

centimetre wavebands. A room has been in use for some time now that affords effective electrical screening up to wavelengths of about 3 cm, and is moreover designed to resemble a normal test room as much as possible. The room in question (see *fig. 1*) is 4 m long, 2 m broad and 2.30 m high; it provides ample facilities for experiments and space for three persons. The double gauze has been replaced by a single, completely closed screen of copper sheets [2]).

In such an enclosure ventilation is essential, both for the persons inside it and for cooling the equipment. The problem presented by the ventilation system, which must not impair the electrical screening, has been solved by using ventilation grids consisting basically of a number of narrow, parallel waveguides disposed at right angles to the walls. The waveguides attenuate all radiation at frequencies below a specific limit, in this case 15 000 Mc/s. The actual form of the grids is shown in *fig. 2*. They

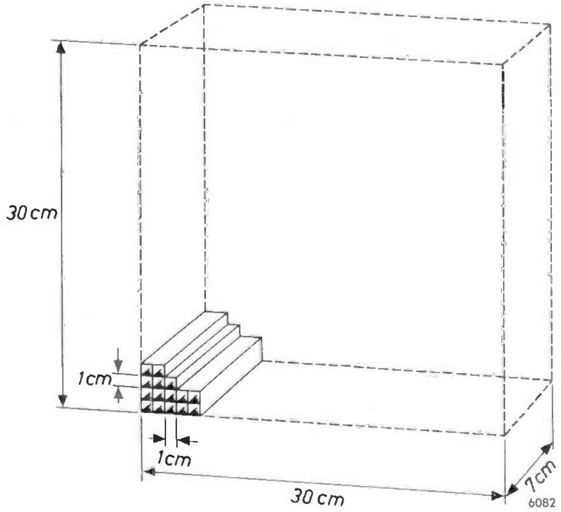


Fig. 2. Construction of a ventilation grid.

[2]) Double *perforated* sheets proved to be an equally effective electrical screen, but they are expensive and entail a heavy, unwieldy door of intricate structure. Moreover the walls present a disturbing background to the eye during observations (moiré patterns).

are situated near the bottom of one long and one short side wall and consist of perpendicularly intersecting copper strips 1 mm thick and 7 cm wide. The cross-section of the air channels, i.e. of the waveguides, is 1 cm^2. Air is extracted through a similar grid in the ceiling and an exhaust pipe. Space is left between the walls and the work benches to help distribute the upward air stream; in these gaps the supply cables are laid. As a further measure to ensure effective air distribution and to reduce air noise, a horizontal sheet of sound-absorbent material is suspended just below the ceiling.

The volume of the enclosure is 18 m^3, the density of air 1.3 kg/m^3. If we assume the specific heat to be 1000 J/kg °C then the mean temperature rise per watt of power dissipated in the *air* is $1/(18\times 1.3\times 1000)$ °C/sec $\approx 4\times 10^{-5}$ °C/sec. If the power consumed in the room, say 2.5 kW, had to be entirely dissipated in the air — which is certainly not the case — it would be necessary, if we take a mean temperature rise of 6 °C as permissible, to refresh the air completely (i.e. all 18 m^3) in a time of $6/(2500\times 4\times 10^{-5})$ sec = 1 min. The exhaust capacity of the ventilator employed is 20.5 m^3/min, which is thus more than adequate.

To avoid draughts, the flow rate of the indrawn air has been put at 0.25 m/sec. The total cross-section of the inlet grids must therefore be roughly $20.5/(60\times 0.25) = 1.37$ m^2. This is achieved with 15 grids measuring 30×30 cm^2, nine of which can be seen in fig. 1.

To comply with safety regulations, a person working in the screened room must be observable from outside. For this purpose the door has an observation window, of the same construction as the grids and measuring $10\times 10\times 7$ cm. A convex mirror is fixed to the wall directly opposite the window, enabling anyone looking in to see practically the whole of the interior (*fig. 3*).

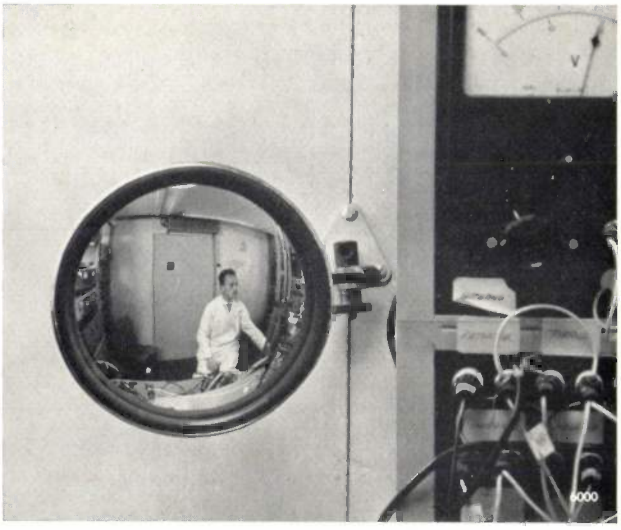


Fig. 3. Photograph taken through one of the openings in the observation window in the door. A convex mirror (see fig. 1) gives an overall picture of the interior. The window is seen in the mirror as a small dark block in the door.

In order to prevent interfering energy entering the room through the supply cables, the latter are provided with a low-pass filter consisting of a network filter combined with two coaxial filters of very simple design (see caption to *fig. 4*).

An idea of the effectiveness of the screening was obtained by setting up a transmitter outside the room, opposite those places thought likely to allow most energy to enter, i.e. the door (through gaps round the edge and through the window), the ventilation grids and the filter. For each position the

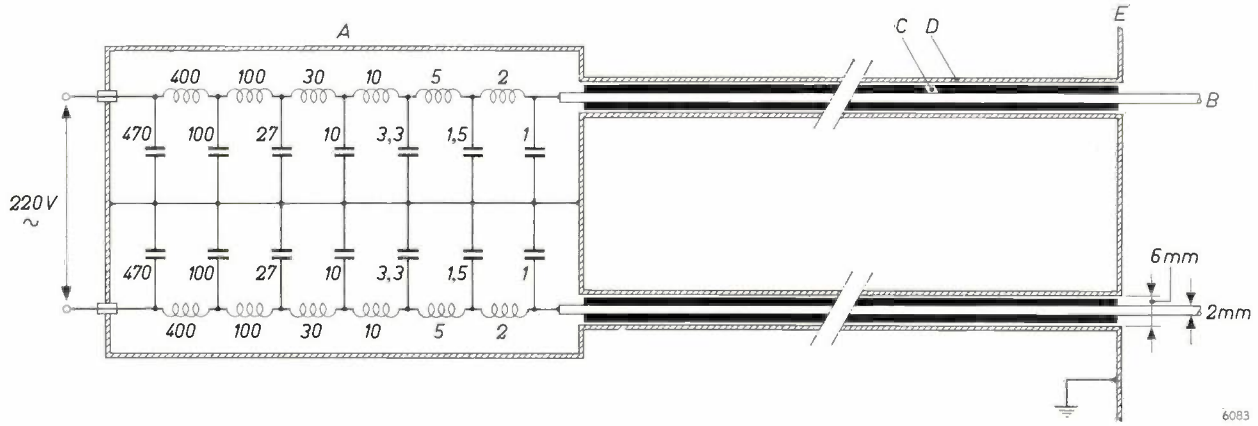


Fig. 4. Principle of the low-pass filter unit for the supply cables. At the input (left) the mains voltage is applied; *E* is the wall of the room (earthed). The brass box *A* contains a network filter for 6 A; inductances are given in μH, capacitances in nF. In the frequency range from 0.6 to 100 Mc/s the network filter gives more than 120 dB attenuation. Frequencies above 100 Mc/s, where the network filter is less effective, are attenuated by means of a coaxial filter, some decimetres long, incorporated in each of the two leads. *B* inner conductor: enamelled copper wire. *C* absorbent sheath: ferroxcube 4B. *D* outer conductor: brass pipe. A 10-cm-long test specimen of this coaxial filter gave about 20 dB attenuation at 100 Mc/s, and about 50 dB at 4000 Mc/s. The total voltage drop across the whole filter is 5 V at a load of 220 V, 6 A.

Table I. The degree of screening obtained.

Wavelength	Attenuation in dB		
	door	vent. grid	filter unit
300 m	> 150	> 150	> 150
3 m	> 150		
7.5 cm	45		
3 cm	25		

output power of a sensitive receiver inside the room measured with the door open was compared with the value found with the door shut. The results at four different wavelengths are given in the table alongside.

A. J. F. de BEER *).

*) Research Laboratories, Eindhoven.

Summary. Description of an electrically screened room (4×2×2.30 m) suitable for microwave experiments. Floor, walls and ceiling are constructed from 1-mm copper sheets. Ventilation is provided by clusters of air channels, 7 cm long and 1 cm^2 in cross-section, which act as waveguides with a limiting frequency of 15 000 Mc/s and thus do not impair the screening. The interior of the room can be observed from outside through a similarly designed window with the aid of a convex mirror. The penetration of interference through the supply cables is prevented by means of a network filter combined with two simple coaxial filters. Radiation from a transmitter outside the closed door is attenuated by at least 150 dB at wavelengths of 300 m and 3 m, by 45 dB at 7.5 cm, and by 25 dB at 3 cm. With the transmitter opposite a ventilation grid or the filter unit the attenuation is > 150 dB at all the wavelengths mentioned.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, The Netherlands, where a limited number of reprints are available for distribution.

R 404: J. B. Davies: Theoretical study of non-reciprocal resonant isolators (Philips Res. Repts **15**, 401-432, 1960, No. 5).

Modern microwave techniques make frequent use of the directional isolator, an element which attenuates the microwaves very little in one direction, but very strongly in the opposite direction. The resonance isolator consists in principle of a thin strip of ferrite placed in the longitudinal direction in the waveguide with a magnetic field across it adjusted to give resonance. Although various usable models of this isolator have already been developed, the mathematical treatment of its behaviour is still incomplete. In this article, the attenuation of the isolator is calculated for various shapes of the ferrite, using a perturbation theory which is only valid for thin strips of ferrite. It is found that the optimum shape of the ferrite is not, as was previously thought, a vanishingly thin strip, but a rod of elliptical cross-section. The results of these calculations emphasize the fact that a knowledge of the dielectric losses in the ferrite is of great importance for the determination of its optimum shape. Good results have also been obtained in practice with a thin strip of ferrite placed next to a strip of a material with a high dielectric constant. The behaviour of this combination is also calculated to a first approximation.

R 405: S. van Houten: Thermal conductivity in $Li_xNi_{(1-x)}O$ (Philips Res. Repts **15**, 433-436, 1960, No. 5).

Measurements of the thermal conductivity of semiconductors of the formula $Li_xNi_{(1-x)}O$ as a function of the lithium content. The results can be explained on the basis of the impurity scattering of phonons. The mean free path of the phonons is found to be a linear function of the mean distance between neighbouring Li atoms. This is in agreement with Kittel's theory of glasses, but not with Klemens' equation for the scattering at point defects. It is also found that this material is not so suitable for thermo-electric generators as other authors had thought.

R 406: M. L. Verheijke: The carbon content of low carbon martensite (Philips Res. Repts **15**, 437-444, 1960, No. 5).

The carbon content of low-carbon martensite can be determined dilatometrically or by means of resistivity measurements. Both methods are based on the assumption that the decreases in length and resistivity due to the first stage of tempering are linear functions of the carbon content. Experiments show that this is an acceptable assumption.

R 407: H. J. Heijn: Representations of switching functions and their application to computers (Philips Res. Repts **15,** 448-491, 1960, No. 5).

Continued from **R 399.**

R 408: D. J. Kroon: Line shape of proton magnetic resonance in paramagnetic solids (Philips Res. Repts **15,** 501-583, 1960, No. 6).

The method of nuclear magnetic resonance is being increasingly used in solid-state research of recent years. This technique has proved particularly useful in the determination of the positions of light nuclei in a crystal lattice. The information about the structure of a polycrystalline substance is mainly derived from the second moment of the resonance line.

In this thesis (Amsterdam, June 1960) an investigation of the effect of paramagnetic ions in a crystal on the second moment of the resonance line is described. The shape of the resonance line is also studied.

On the assumption that the paramagnetic ions may be regarded as point dipoles, broadening of the resonance lines may be analysed into three components:

a) composition broadening, due to the non-identical surroundings of nuclei in different unit cells of the specimen,

b) broadening due to the non-cubic surroundings of the nuclei,

c) broadening due to the fact that the crystallites constituting the porous specimen are not spherical.

The contributions to the second moment from these three sources are calculated. In order to test the theory, experiments were carried out on samples of the composition $NH_4Fe_xAl_{1-x}(SO_4)_2$. The proton magnetic resonance lines of these samples were determined at 77 °K. The variation of the second moment as a function of the iron content is in good agreement with the theory developed in this thesis. An extra broadening found especially at low iron concentrations is ascribed to a short spin-lattice relaxation time. It is found possible to indicate the positions of the protons in the crystal. It is found that even at low temperatures (20 °K) the protons are in rapid motion. The contribution of the anisotropy of the crystal to the second moment is calculated on the basis of this fact. The calculated value is of the same order of magnitude as the measured value, and the difference is probably due to the finite extent of the magnetic moment of the paramagnetic ions. The broadening of the resonance line due to the form of the crystallites was experimentally determined by some measurements on diluted samples. Finally, the shape of the resonance line was calculated and compared with the measured shape. The good agreement between the observed and calculated shaped supports the assumption made concerning the positions of the dipoles.

In the experimental part of this thesis, a simple spectrometer for nuclear magnetic resonance measurements on solids is described. In this connection, a rapid and accurate method for the design of permanent magnets is mentioned in some detail. The distortion of the resonance lines due to the method of measurement is also calculated, and means of correcting the measurements for this distortion are discussed.

R 409: J. J. Scheer: Some preliminary experiments concerning the influence of band bending on photo-electric emission (Philips Res. Repts **15,** 584-586, 1960, No. 6).

The theory that *p*-type semiconductors should exhibit a higher photo-electric quantum efficiency than *n*-type semiconductors is verified by measurements of the photo-emission from cesium-covered silicon surfaces.

R 410: G. E. G. Hardeman: Nuclear dynamic polarization in irradiated polytetrafluorethylene (Philips Res. Repts **15,** 587-597, 1960, No. 6).

When a magnetic interaction exists between the electrons and the nuclei in a substance placed in a constant magnetic field, the magnetic polarization of the nuclei can be increased by means of radiation of a suitable frequency. The extent of this "dynamic" polarization of the nuclear moments can be determined from the intensity of the nuclear magnetic resonance produced by other radiation, whose frequency is equal to the resonance frequency of the nuclei. This paper describes an investigation of the dynamic polarization of ^{19}F nuclei in samples of polytetrafluoroethylene which has been previously irradiated with fast electrons in order to form paramagnetic centres. The dynamic polarization is found to be a function of the concentration of paramagnetic centres, the intensity of the radiation, and the line shape of the paramagnetic resonance. It spreads outwards from the paramagnetic centres, which are present in much smaller numbers than the nuclei, throughout the whole spin system.

A 26: A. Klopfer, S. Garbe and W. Schmidt: Residual gases in vacuum systems (6th nat. Symp. Vacuum Technol. Trans., Philadelphia, October 7-9, 1959, Ed. C. R. Meissner, pp. 27-33, Pergamon, Oxford 1960).

Investigations of the gas composition in sealed-off high and ultra-high vacuum systems with the aid of the omegatron have shown that the gas evolution of the materials and interactions between the gases and the surfaces determine the kind of gases which are present. Such interactions are: absorption, chemical and exchange reactions. The results obtained on radio tubes, television picture tubes and systems consisting of glass and metal only are given in this paper. It appears possible to classify the composition of residual gases on the basis of the conditions prevailing in the systems in question.

A 27: K. J. Planker and E. Kauer: Bestimmung der effektiven Masse freier Ladungsträger in Halbleitern aus der Ultrarotabsorption (Z. angew. Phys. **12**, 425-432, 1960, No. 9). (Determination of the effective mass of free charge carriers in semiconductors from the infrared absorption; in German.)

The possibility of determining the effective mass of free charge carriers in semiconductors by means of infrared measurements is discussed. The main points of the existing theories of absorption are discussed, and their results are expressed in a common mathematical form. Comparison with experiment shows that the slope of all experimentally determined absorption curves is correctly predicted by Schmidt's theory of thermal scattering and Meyer's theory of scattering by defects; the absolute value of the absorption coefficients cannot however be calculated very exactly.

Absorption measurements on *n*-CdTe give an effective mass of about 0.24 of the free-electron mass.

A 28: H. G. Reik, H. Risken and G. Finger: Theory of hot-electron effects in many-valley semiconductors in the region of high electric field (Phys. Rev. Letters **5**, 423-425, 1960, No. 9).

The quantitative description of hot-electron effects in many-valley semiconductors is obtained by means of the solution of a properly formulated Boltzmann equation for high fields. In this formulation the actual band structure and the various scattering mechanisms are taken into account. The theory can qualitatively account for the experimentally observed anisotropy of hot electrons in a Sasaki-type experiment. For quantitative agreement intervalley scattering has to be taken into account.

A 29: S. Garbe, A. Klopfer and W. Schmidt: Some reactions of water in electron tubes (Vacuum **10**, 81-85, 1960, No. 1/2).

Traces of water which are not removed during the degassing of vacuum tubes will later be evolved, especially from the glass walls. It has been shown that this does not in fact lead to a great increase in the partial pressure of water vapour during the operation of the tube: most of the water reacts with the hot cathode or the getter mirror, forming principally hydrogen, methane and carbon dioxide.

A 30: A. Klopfer and W. Ermrich: Properties of a small titanium-ion pump (Vacuum **10**, 128-132, 1960, No. 1/2).

A small getter-ion pump is described which has a capacity of several torr-litres for chemically active gases. The maximum pumping speed for CO is about 50 l/sec. The lowest pressure obtained with this pump is less than 10^{-10} torr. The composition of the residual gases during the process of evacuation has been measured. The factors determining the lowest pressure that can be obtained are discussed. (See also Philips tech. Rev. **22**, 260, 1960/61, No. 8.)

A 31: R. Groth: Über die Temperaturabhängigkeit der kurzwelligen Ausläuferabsorption von MgO im Ultraroten (Ann. Physik **6**, 328-344, 1960, No. 5/6). (The infra-red absorption of MgO on the short-wave tail of the infra-red band as a function of temperature; in German.)

The absorption of MgO single crystals was measured for wavelengths between 5 and 10 μ (i.e. on the short-wave tail of the infra-red absorption band) at temperatures between 20 and 1500 °C. Two heaters were designed to cover the entire temperature range. The absorption coefficient at a given wavelength increases linearly with the temperature at high temperatures, as predicted by Born and Huang for absorption by second-order electrical moments. The behaviour at lower temperatures suggests that both possible excitation processes contribute to the absorption.

VOLUME 23, 1961/62, No. 6 pp. 161-188 Published 13th March 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

ZONE MELTING OF OXIDES IN A CARBON-ARC IMAGE FURNACE

by C. KOOY *) and H. J. M. COUWENBERG *).

66.049.4:548-31:621.365.24

In solid-state research the method of heating in a radiation furnace, using a carbon arc as the primary source of heat, yields advantages similar to those of the conventional method of high-frequency heating. Unlike the latter, however, radiation heating can also be applied to materials of very high resistivity, which include many oxides. By means of special devices it is consequently possible to subject these materials to the process of zone melting, familiar from semiconductor technology (where it is applied in the special forms of zone refining and zone levelling). In this way large single crystals have been prepared from a series of oxides. The composition of the crystals proved to be reasonably close to the stoichiometric composition.

Technologists and experimenters early hit upon the idea of producing very high temperatures by means of a *solar furnace*, in which the sun's rays are focused into an intense image by a large parabolic mirror. An object placed in that image can be heated to extremely high temperatures, theoretically to the temperature of the sun itself. Towards the end of the 17th century Von Tschirnhaus in Dresden had several large lens systems built for this purpose [1]), with which he is believed to have achieved temperatures up to 3000 °C.

Solar furnaces are still used today [2]), partly because they are one of the means of making the sun's energy directly of service to man, partly because of the high temperatures which they bring within easy reach. Some of their additional advantages compared with conventional furnaces are: the atmosphere in which the charge is heated can be independent of the source of heat; the irradiation is confined to a very small area where melting can occur; as a consequence, the non-molten material can itself act as a "crucible" and there is no risk of the melt being contaminated by foreign matter from this source; finally, the temperature can be raised in an extremely short time and the charge kept under observation during the heating process.

Radiation heating shares these advantages with *RF induction heating*, which is therefore widely used in solid-state research, particularly for the zone melting of silicon and germanium [3]). RF heating falls down, however, where the material to be melted has been developed for the very object of causing minimum energy dissipation in high-frequency fields, as is the case with ferrites. More generally the same applies to materials that have a very high electrical resistivity, such as the oxides NiO, TiO_2 and others.

For the zone melting of such materials we have used the method of radiation heating with good results — although not with a solar furnace but with a *carbon-arc image furnace*. In our changeable climes the solar furnace scarcely commends itself for practical use, and the mechanism needed for the lens or mirror system to follow the course of the sun makes it relatively expensive as a laboratory instrument. The use of an intense terrestrial source

*) Research Laboratories, Eindhoven.

[1]) See page 182 of this number.

[2]) One of the world's largest installations, using a parabolic mirror of 90 m^2 surface area, has been erected in the Pyrenees. See: F. Trombe, Le laboratoire de l'énergie solaire de Mont-Louis, Bull. Soc. Chim. France 1953, pp. 353-368. See also: Applications thermiques de l'énergie solaire dans le domaine de la recherche et de l'industrie, Colloque international à Mont-Louis, June 1958, Centre National de la Recherche Scientifique, Paris (about 50 articles).

[3]) See J. Goorissen, Segregation and distribution of impurities in the preparation of germanium and silicon, Philips tech. Rev. 21, 185-195, 1959/60.

of radiation, e.g. the carbon arc, instead of the sun's rays makes a much simpler construction possible, which still offers most of the advantages of the solar furnace. A description will be given here of an arc image furnace which has been in use for the zone melting of oxides in this laboratory for about two years. The working procedure will be discussed and some results mentioned [4]).

Description of the apparatus

The radiation from the carbon arc can be concentrated upon an object very effectively with an arrangement of two elliptical mirrors (see *fig. 1c*). This arrangement is better than one using a single elliptical mirror (fig. 1*a*), since the symmetrical path of the rays results in an image having less optical aberration and therefore a greater radiant flux in the focus used for heating. For practical purposes, two elliptical mirrors are preferable to two parabolic mirrors (fig. 1*b*) in view of the additional possibilities offered by the small cross-section of the beam at the central "focus": diaphragms can easily be introduced near this point, or a plane mirror to alter the light path. The latter may be useful to avoid interrupting long heating periods when new carbons have to be fitted: in that case *two* carbon arcs are used and alternately ignited and focused on to the object by reversing the plane mirror [5]).

With an arrangement using two elliptical mirrors, Null and Lozier [6]) have shown that the radiant flux at the focus used for heating can reach 10 to 14 watt/mm^2 (depending on the type of carbons). This value is only slightly lower than the maximum value of about 15 W/mm^2 hitherto achieved in solar furnaces. With their apparatus Null and Lozier were able to melt small quantities of such materials as zirconium oxide (melting point 2970 °K), titanium carbide (3400 °K), tungsten (3680 °K) and even zirconium carbide (3800 °K).

The radiation source in our apparatus is a cinema-projection arc lamp (Philips type EL 4455). This lamp is already fitted with an elliptical mirror, which serves as one of the two mirrors for the arrangement shown in fig. 1*c*. The second mirror used is identical, see *fig. 2*. The diameter of the mirrors is about 36 cm, and the focal lengths are 14 and 68 cm. The carbons in the lamp are rated for an arc current of 80 A; the carbon feed is continuous and automatic.

The radiant flux at the focus used for heating in this furnace has not been measured but is estimated to be roughly 5 W/mm^2. The temperatures obtainable depend to a marked extent, of course, on the absorption and reflection by the irradiated substance. If the material is transparent or highly reflecting, the actual increase in temperature will be slight. Glass, for example, hardly gets hot, and the fairly transparent Al_2O_3 (melting point 2320 °K) cannot be melted by the radiant flux mentioned — unless a little chromium is added to increase absorption. On the other hand NiO (melting point 2360 °K), which is a good absorber, is easily melted: a 5 mm thick sintered NiO rod melts within a few seconds of being placed in the focus. This incidentally is a

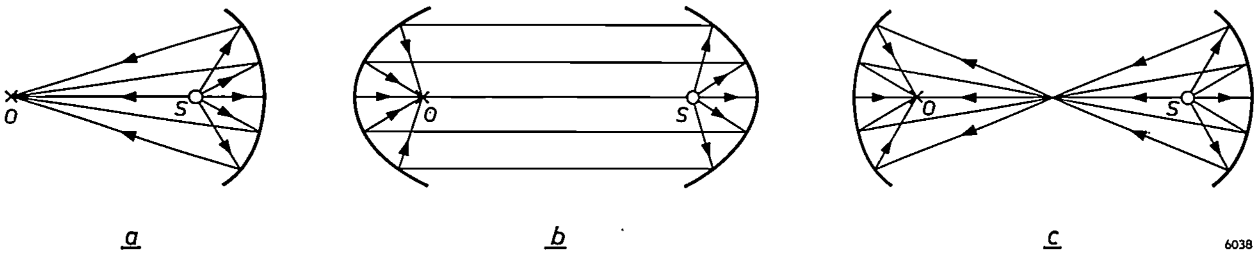


Fig. 1. Three possible mirror arrangements for heating an object *O* with the aid of a radiation source *S*.
a) Single elliptical mirror, with *S* and *O* in the two foci of the ellipse.
b) Two parabolic mirrors, with *S* and *O* each in the focus of one of the parabolas.
c) Two elliptical mirrors, so arranged that they have one focus in common and with *S* and *O* in the two other foci.

[4]) Similar equipment on the market (cf. [5])) has been used for a number of other applications. As far as we know, however, zone melting using an image furnace has hitherto only been applied to small quantities of organic compounds, i.e. with relatively low temperatures up to say 200 or 300 °C. The arrangement consisted of a single elliptical mirror; see E. F. G. Herington, Zone refining, Endeavour **19**, 191-196, 1960.

[5]) See R. E. De la Rue and F. A. Halden, Arc-image furnace for growth of single crystals, Rev. sci. Instr. 31, 35-38, 1960. Technical particulars of the carbon-arc image furnace will be found in: P. E. Glaser, Imaging-furnace developments for high-temperature research, J. Electrochem. Soc. **107**, 226-231, 1960.

[6]) M. R. Null and W. W. Lozier, Carbon arc image furnaces, Rev. sci. Instr. **29**, 163-170, 1958.

good illustration of the speed at which the radiation furnace heats up its charge. Other oxidic materials, such as TiO_2 and $MnFe_2O_4$ (a spinel) are also readily melted.

The mechanism used for introducing the charge into the focus of our carbon-arc image furnace is specially designed for the purpose of zone melting and more specifically for zone melting by the *floating-zone technique.* Two rotatable holders, which grip two rods of the oxide to be melted, are mounted in a frame in such a way that the axes of the rods are vertically in line. The upper holder is capable of slight vertical displacement, and the whole frame can be shifted over a fair distance in the same direction. By moving the frame the tip of the lower oxide rod is introduced first into the focus and melted. Next, the upper rod is dipped into the molten tip, and the surface tension of the melt causes the formation of a molten zone between the two rods. The zone can now be made to move up and down the rods by lowering or raising the frame.

The material is heated only at the side of the nearer mirror. To prevent this unilateral heating from causing irregular solidification and melting of the moving zone, the holders of both rods are made to *rotate.* A rotational speed of 60 to 120 r.p.m. is found in general sufficient to produce a solid-liquid interface of good rotational symmetry. If moreover the two holders are rotated in opposite directions (as in the present case), the opposing motions of rotation at either side of the zone give rise in the melt to a stirring action, which promotes a uniform temperature distribution and a homogeneous composition. The surface tension of the melt is high enough in the case of the oxides mentioned to keep a zone between 5 and 10 mm long intact with rods about 5 mm thick. The zone length can be regulated by varying the arc current or by slightly defocusing the second elliptical mirror.

Since the oxidic materials in question are heated so quickly to their melting point, relatively high zone speeds are possible. For example, the molten zone in the rod of $MnFe_2O_4$ remained intact in our apparatus at a speed of 20 cm per hour. Normally we work at a zone speed of a few centimetres per hour.

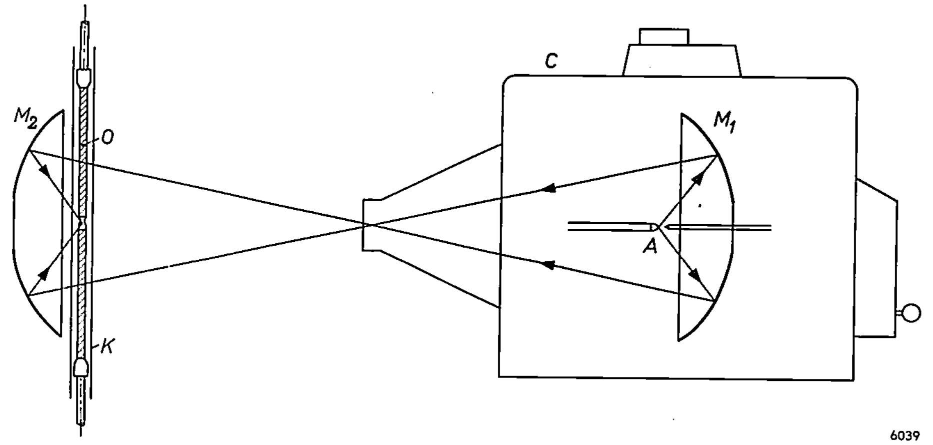


Fig. 2. Sketch of the carbon-arc image furnace. C cinema arc lamp, with carbon arc A and elliptical mirror M_1. The oxide rod O to be heated is situated in the proximate focal point of the second elliptical mirror M_2, and may be surrounded by a tube K of glass or quartz.

The rods can be surrounded by a tube of glass or quartz without interfering with the heating. In this way the entire process can be carried out in any desired gas atmosphere.

A photograph of the apparatus is shown in *fig.* 3, in which some of the details described can be seen.

Preparation of single crystals

The floating-zone technique offers an elegant possibility of producing single crystals. This method of producing crystals is widely used for metals, and has been developed into a highly refined technique for germanium and in particular for silicon, using RF heating. Hitherto, however, it has found scarcely any application for oxides of high melting point.

We have investigated the possibility of producing single crystals of oxides with a high melting point by zone melting in our carbon-arc image furnace. The results have proved to be very satisfactory. We have for example made single crystals of NiO and TiO_2, and the following series of mixed crystals of $MnFe_2O_4$ with other compounds having the spinel structure:

$Mn_{1+x}Fe_{2-x}O_4$, with $x = -0.1$ or 0 or 0.1 or 0.2;

$MnTi_xFe^{II}_xFe^{III}_{2-2x}O_4$, with $x = 0.15$ or 0.30 or 0.45;

$MnTi_{0 \cdot 15}Co_{0 \cdot 15}Fe_{2 \cdot 7}O_4$;

$Mg_{0 \cdot 45}Mn^{II}_{0 \cdot 55}Mn^{III}_{0 \cdot 23}Fe_{1 \cdot 77}O_4$.

The starting material used was obtained by prefiring powdered metal oxides, mixed in the right proportions, and by pressing these prefired powders into the shape of rods and sintering them. The sintered rods are still porous but sufficiently easy to handle, and have enough resistance to thermal shock to withstand zone melting without disintegrating.

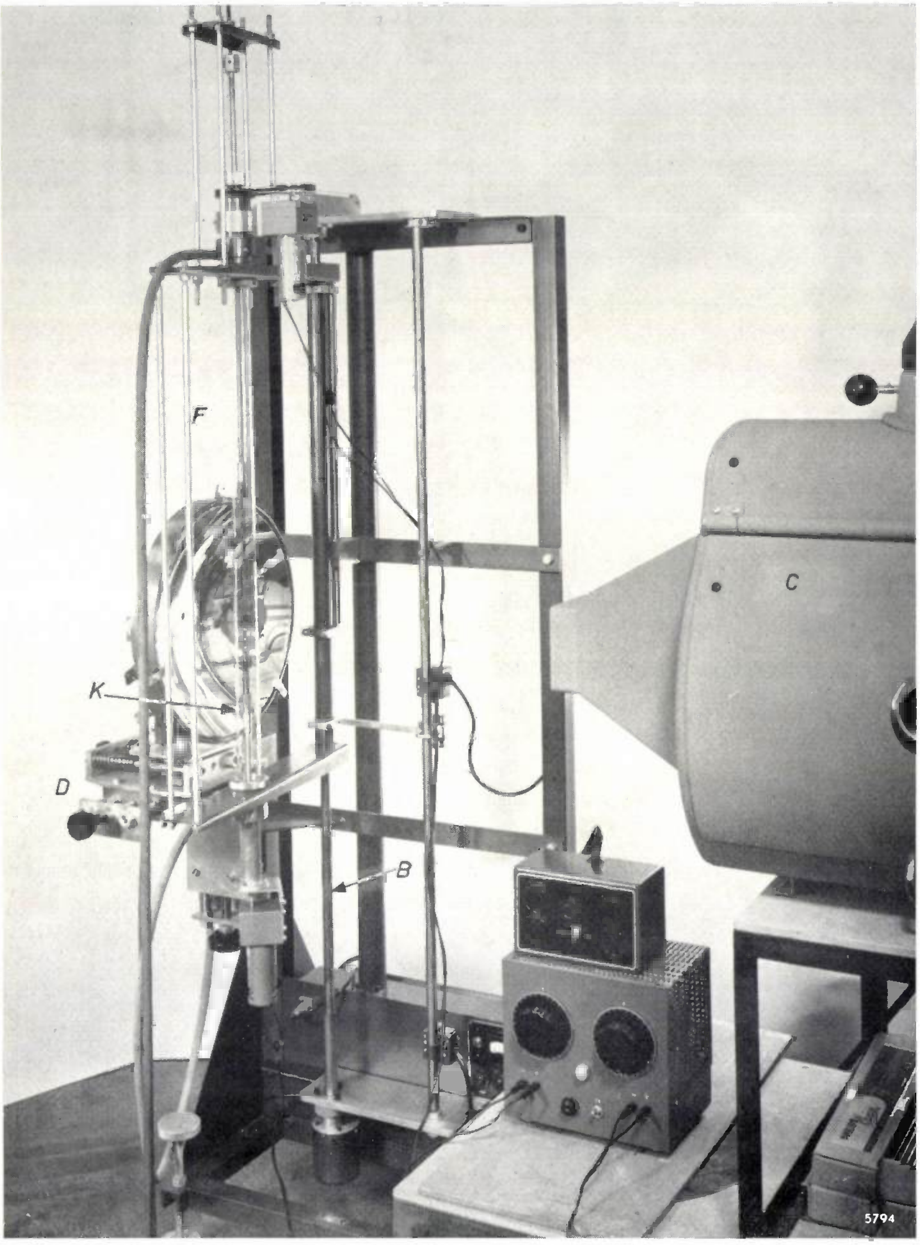


Fig. 3. Carbon-arc image furnace used in Philips laboratories at Eindhoven for the zone melting of oxides. Partly visible on the right is the cinema arc lamp *C* (type EL 4455), which has a built-in elliptical mirror as condenser. On the left can be seen the second elliptical mirror, mounted on a carriage *D*, adjustable in three directions. In the upper and lower parts of the frame *F*, which can be shifted vertically at variable speed along the guide rod *B*, are mounted on their bearings the rotary holders in which the oxide rods are clamped. The rotational speed of each holder is independently regulated by the lower control box, the movement of the frame by the upper one. Surrounding the oxide rods is a quartz tube *K*, to both ends of which rubber tubing is connected through which the gas in which the zone melting process is carried out is admitted to the tube.

Our crystal-growing techniques are similar to the standard ones as applied, for instance, to silicon. The growing crystal is initially narrowed to a very small diameter by increasing the distance between the two rods. Pieces of the single crystal thus produced may later be used as seeds for growing further single crystals. The total length of the single crystals obtained is limited by the burning life of the carbons (we have not yet used interchangeable carbon arcs, see above). At currents from 50 to 70 A this is between two hours and half an hour which, at a zone speed of a few centimetres per hour, yields crystals several centimetres long — sufficient for most purposes in solid-state research. Some of these single crystals can be seen in *fig. 4*. *Fig. 5* shows the molten zone during the growth of a single crystal of $MnFe_2O_4$.

The processes of zone melting and crystal-growing are much more complex for oxidic compounds than for elements such as Ge and Si, owing to the fact that the oxides may decompose and the equilibrium pressure of the oxygen in an oxide is critically dependent on temperature. The partial pressure

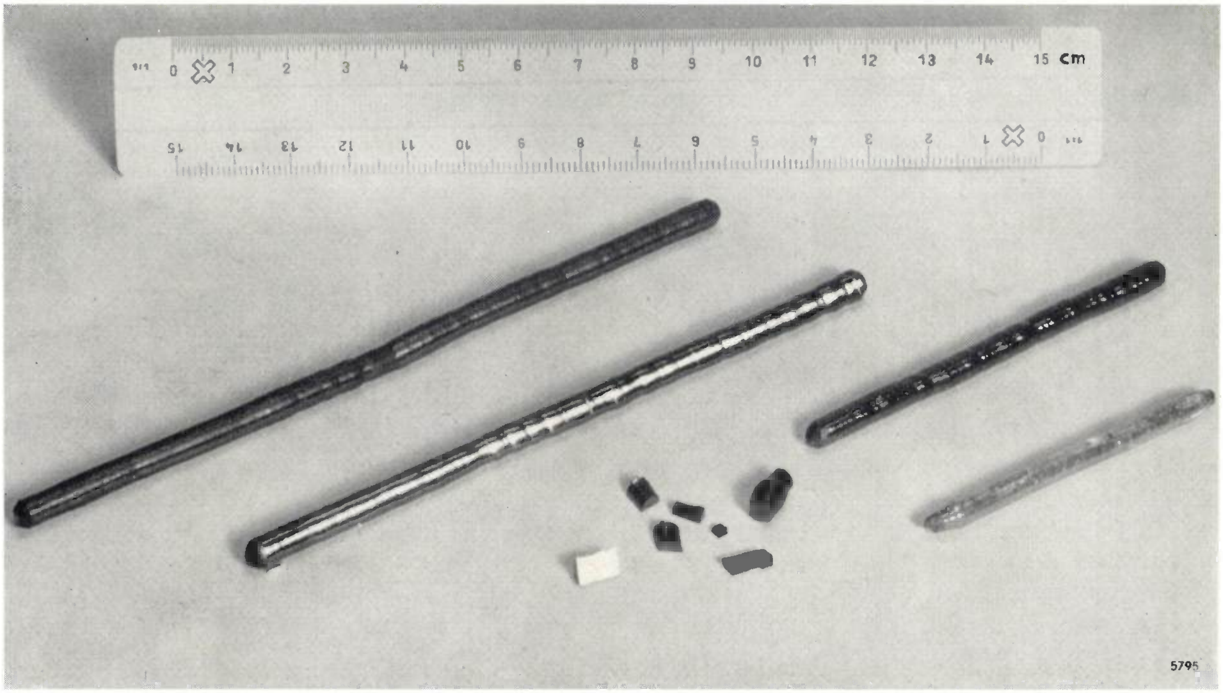


Fig. 4. Some monocrystalline rods of oxides produced in different atmospheres with the apparatus described. The zone speed in all cases was 6 cm per hour. From left to right: $MnFe_2O_4$ in air; $Mn_{1.1}Fe_{1.9}O_4$ in nitrogen with 0.2% oxygen; $Mn_{1.2}Fe_{1.8}O_4$ in nitrogen with 0.2% oxygen; TiO_2 in air. In the foreground are some fragments of an NiO rod, showing good cleavage planes; this rod was grown in pure oxygen.

chosen for the oxygen in the gas atmosphere will prevent both oxidation and decomposition at one temperature only. Within the molten zone the temperature is fairly constant because of the stirring effect caused by the rotation, but in the solidified part the temperature decreases very steeply in the axial direction. In traversing the zone, therefore, the material is bound to be successively reduced and oxidized.

When zone-melting NiO in air, for example, a considerable reduction occurs in the melt: upon abrupt cooling the melt can be shown to contain metallic nickel. Nevertheless, the material that crystallizes from the melt is almost exactly stoichiometric NiO, and the cooled rods are even slightly oxidized. When TiO_2 is zone-melted in air, the material solidifying from the melt is the reduced black modification which, during cooling, is completely re-oxidized into the light-yellow form of virtually stoichiometric TiO_2. As our last example we mention $MnFe_2O_4$ and the substituted manganese ferrites referred to above: when these are zone-melted in air the outer layer oxidizes during cooling. After removal of this skin, however, the core is again found to have a composition which is in fairly good agreement with the proportions by weight of the

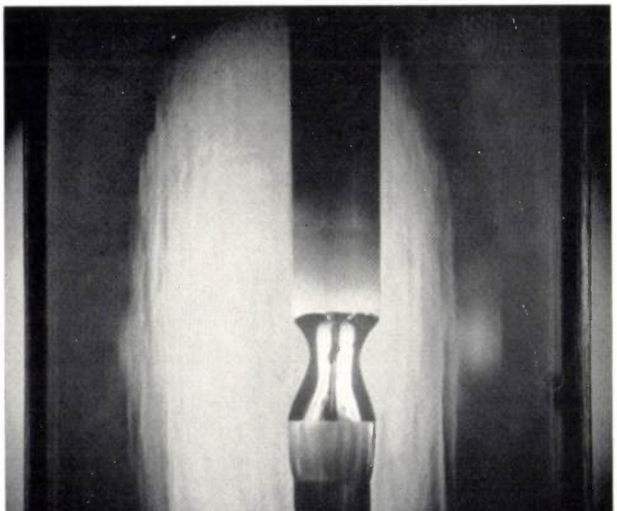


Fig. 5. Molten zone, photographed through the 3.5 cm wide quartz tube, during the growth of a single crystal of $MnFe_2O_4$. The frame with both rods is moving slowly downwards, which in effect means that the zone in the material moves upwards. The part above the zone is the sintered polycrystalline rod, 5 mm thick, and the part below is the solidified single crystal.

oxides in the sintered rods. When these manganese ferrites are zone-melted in nitrogen plus 0.2% oxygen, there is scarcely any oxidation of the outer layer (cf. fig. 4).

Thus, in spite of the complicating factor of decomposition, it is possible by zone-melting these oxides to produce single crystals that show no marked deviation from the stoichiometric composition. This applies to the crystal rod as a whole, since the constant supply of fresh material to the melt soon gives rise to a stationary state during the actual process of zone melting. This contrasts with two other methods of producing oxide crystals, namely the Bridgeman-Stockbarger method and the flux method [7]), where a gradual change in the chemical composition is unavoidable during crystal growth. Yet another method, based on the old Verneuil process [7]), has in common with zone melting the advantage of reaching a steady state, after which the composition of the solidifying material remains in principle constant. This process can also be carried out with radiation heating [5]), making a free choice of atmosphere possible. In the Verneuil process, however, the preparation of the starting material involves many more difficulties than in the method of zone melting described here.

In conclusion, mention may be made of a problem that arises in our process, and which is connected with the above-mentioned high temperature gradient in the material immediately after solidification. This is an inherent feature of the highly localized heating in the radiation furnace, the gradient being increased by the low thermal conductivity of many oxides. True, the local temperatures in the material cannot be accurately measured: the results found with the optical pyrometer (the only feasible method of measurement in this case) are difficult to interpret owing to the reflection of the intense light from the carbon arc. It may be assumed, however, that axial temperature gradients occur of the order of 1000 °C per cm. At a zone speed of 5 cm per hour this means that the single crystal actually cools down at a rate of roughly 100 °C per minute. Owing to the relatively low plastic deformability of the oxides, this rapid cooling often gives rise to *cracks* in the single crystals. If one wishes to avoid this, without unduly sacrificing the rate of growth, one must e.g. use additional heat sources to keep the solidified material longer up to temperature.

Summary. For the purpose of zone-melting oxides possessing a high electrical resistivity the use of RF induction heating — as applied to silicon and germanium — is not suitable. Such oxides can be effectively zone-melted, however, with the aid of radiation heating. An apparatus employed for this purpose is described, in which two elliptical mirrors produce an image of an intense carbon arc, and an oxide rod to be melted is placed in this image. The floating-zone technique is applied, so that the equipment is eminently suited for producing single crystals of oxides having a high melting point. The authors describe the preparation of rod-shaped single crystals of NiO, TiO_2, $MnFe_2O_4$ and of various substituted manganese ferrites. The composition is constant throughout the single-crystal rod and is in good agreement with the stoichiometric composition.

[7]) See e.g. F. W. Harrison, The growth of oxide single crystals containing transition metal ions, Research 12, 395-403, 1959.

INFLUENCE OF THE NON-LINEAR BEHAVIOUR OF A RECORDING INSTRUMENT ON THE PROPERTIES OF A CONTROL SYSTEM

by C. H. LOOS *).

621-53.001:621.317.7.087.6

Pursuant to the articles on the application of control theory to linear systems published in the two preceding numbers of this journal, the article below deals with an element which behaves non-linearly when the input signal undergoes very rapid variations. In qualitatively analysing the stability characteristics of a control loop containing this element, use is made of a rule of thumb formulated in the second of the articles mentioned. The surprising conclusion is that the stability of the control loop in question is not a monotonic function of the speed of variation, but shows a minimum.

Introduction

For measuring and recording important variables in industrial plants, increasing use is being made of recording instruments (recording millivoltmeters) whose operation is based on automatic compensation of the measured quantity (voltage) by means of a servomechanism. Fundamentally, the circuit of these instruments stems from Poggendorf's well-known compensation method (*fig. 1a*), except that the null instrument here is replaced by an amplifier which drives the motor that moves the sliding contact of the potentiometer (see fig. 1*b*). Attached to the sliding contact is a stylus or pen. Non-electric quantities to be measured are first converted into an electrical signal.

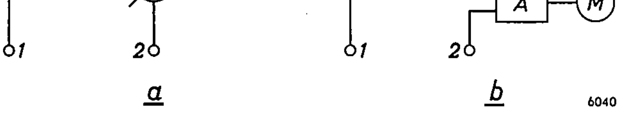


Fig. 1. *a*) Poggendorf compensation method of measuring an e.m.f. The voltage is applied to the terminals *1* and *2*. The contact *3* is then shifted until the highly sensitive meter *G* no longer shows a deflection. The potential difference between points *3* and *4* is then equal to the e.m.f. to be measured and can be calculated from the current flowing through *R* and the resistance between *3* and *4*.
b) In a recording instrument, *R* is a potentiometer whose sliding contact is moved by a motor *M*. The latter is fed by the potential difference between *2* and *3*, highly amplified by *A*. As soon as the potential difference is zero, the sliding contact remains stationary.

Frequently the recorded quantities also have to be automatically controlled, in which case the recording instrument itself can sometimes be used as part of the controller, i.e. as an amplifier. For this purpose a second potentiometer is employed, which is fed with a constant voltage much higher than the voltage across the measuring potentiometer, and whose sliding contact moves synchronously with that of the other. In this way a gain of e.g. 5000× can readily be achieved, offering a particularly simple method of effecting the control action.

A recording instrument thus modified behaves as a linear element only when the changes in the input signal are slow enough for the sliding contact (the pen) to follow. In this article we shall examine what happens when this condition is not fulfilled. It will be shown that instability effects may arise in a control loop which contains, in addition to the recording instrument, two elements having the transfer function $(1 + j\omega\tau)^{-1}$. If the recorder were an ideal amplifier, a control system of this kind ought to be stable for every value of the loop gain [1]). Remarkably enough, the extent to which the stability is endangered — we shall express this presently in a more rigorous form — does not increase monotonically with the discrepancy between the desired speed of the pen and the maximum possible speed. In fact, where this discrepancy is very large, the danger of instability decreases!

The sliding contact will no longer follow a varying signal when the amplifier *A* (fig. 1*b*) is overdriven. Irrespective of the magnitude of the potential between *2* and *3*, the amplifier

*) Research Laboratories, Eindhoven.

[1]) Examples of control loops with linear elements will be found in the article by M. van Tol, Philips tech. Rev. **23**, 109, 1961/62 (No. 4), where the transfer function is also discussed. The relation between loop gain and stability is dealt with by M. van Tol in Philips tech. Rev. **23**, 151, 1961/62 (No. 5).

then delivers its maximum output signal and the pen consequently moves at a constant speed. Even when the amplifier is not overdriven, the pen does not of course follow a varying signal exactly: the motor can only turn when the potential between *2* and *3* differs from zero. The deviation is smaller the higher the gain factor of A; theoretically it approaches zero for an infinitely high gain factor. If a *constant* signal is applied between *1* and *2*, and the motor behaves like an ideal integrator, the deviation will of course be zero in the long run.

With the aid of two figures we shall now try to show qualitatively the way in which the output voltage of the instrument varies when the variations in the input voltage are too fast. We at once introduce the approximations necessary to simplify the theoretical treatment of the instrument's behaviour. The most general case is represented in *fig. 2*. The broken curve represents the input signal and the solid line the output signal. For convenience the gain is assumed to be unity. When the input-signal variations are too fast for the pen to follow, the pen moves at a constant speed (portion AB). This continues until the speed of variation has dropped sufficiently to allow the pen to catch up again (point B). The pen now follows the variation of the input voltage until (at point C) the signal again changes too rapidly. Thereupon the output signal again varies linearly with time (portion CD) and so on.

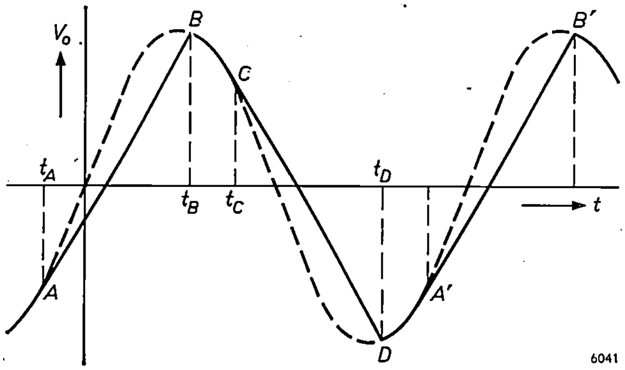


Fig. 2. Output voltage of a recording instrument adapted as a controller, when the maximum speed at which the input signal varies is too fast for the pen to follow. The pen follows the input signal only between B and C (and D and A', etc.). Outside these regions the pen moves uniformly at its maximum speed.

If we now increase the frequency or the amplitude of the output signal, points B and C etc. come closer together, until finally the signal acquires a triangular waveform (*fig. 3*).

Summarizing, then, we note that with rising frequency and/or amplitude the gain is initially linear. When a certain limit is exceeded, we obtain the case represented in fig. 2, and finally, after passing a second limit, the case in fig. 3. We shall now put this into mathematical form, after which we shall examine the behaviour of the instrument as an element in a control loop, with the aid of *describing functions*. In this method the problem is treated as if the element were a linear one, the output signal following from a sinusoidal input signal being approximated by its fundamental Fourier component. For these sinusoidal signals we can then establish a transfer function — the describing function — in the same way as for linear elements. Unlike the transfer function of a loop consisting solely of linear elements, however, this describing function may contain the amplitude as well as the frequency of the input signal.

The describing function is especially useful for analysing the stability of a control loop containing a non-linear element. Since the higher Fourier components of an output signal whose frequency is near the cut-off frequency are usually strongly attenuated in the other elements of the loop, the signal when it appears again at the input of the non-linear element, having passed once around the loop, has in fact become virtually sinusoidal, and may to a very good approximation be regarded as solely due to the fundamental Fourier component.

Calculation of the transfer function

To calculate the transfer function of the recorder we start from a sinusoidal input signal of amplitude U. Disregarding the limited speed of the recorder, we consider the instrument to be an ideal amplifier having a gain factor A. Since the magnitude of A has no effect on the behaviour of the instrument — although of course it does affect the control loop to which it belongs — we shall henceforth assume A to be equal to unity.

The rate of change of the input signal $U \sin \omega t$ is $\omega U \cos \omega t$. As long as ωU is smaller than the

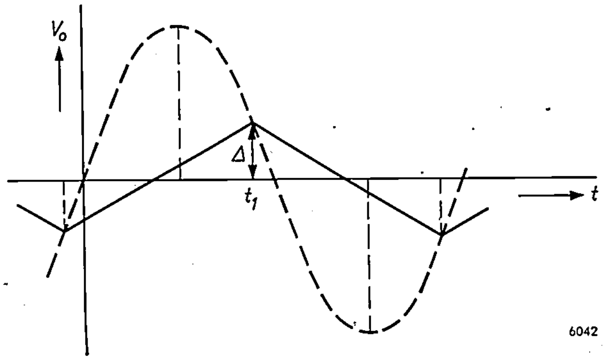


Fig. 3. When the maximum speed at which the input signal changes substantially exceeds the maximum writing speed, the output signal has a triangular waveform.

value corresponding to the maximum speed B of the pen, the behaviour of the instrument is linear.

We shall first consider the case represented in fig. 3, where the pen is no longer able to follow the input signal at all, and the output signal has a triangular waveform. The slope of the straight lines is $+B$ and $-B$ respectively, and the amplitude $\varDelta$ of the signal (half the peak-to-peak value) is $\pi B/2\omega$. The first Fourier component (first harmonic) of a triangular signal of amplitude $\varDelta$ (see appendix) is:

$$\frac{8\varDelta}{\pi^2}\sin(\omega t+\varphi)\,. \quad \ldots\ldots \quad (1)$$

The amplitude ratio of the first harmonics of output and input signals is thus:

$$\frac{4}{\pi}\times\frac{B}{\omega U}\,. \quad \ldots\ldots\ldots \quad (2)$$

The phase angle φ is quickly found when it is remembered that the peak value of the output signal, and hence of its first harmonic, occurs at the moment at which the input signal (in the second quadrant) also has the value $\pi B/2\omega$, and that the peak value of the input signal occurs when $\omega t = \pi/2$. It follows from this that:

$$\varphi=\frac{\pi}{2}-\left(\pi-\arcsin\frac{\pi B}{2\omega U}\right)=-\frac{\pi}{2}+\arcsin\frac{\pi B}{2\omega U}\,. \quad \ldots\ldots \quad (3)$$

Before considering the case of fig. 2, we shall consider what are the "limits" mentioned above. We have seen that the behaviour of the instrument is linear if $\omega U < B$, i.e. if $B/\omega U > 1$.

The case of fig. 3 — triangular output voltage — occurs where $-\omega U\cos\omega t_1 > B$, that is where $\omega U\cos\arcsin\dfrac{\pi B}{2\omega U} > B$, i.e. where

$$\sqrt{1-\left(\frac{\pi B}{2\omega U}\right)^2}>\frac{B}{\omega U}\,.$$

This is the case when

$$\frac{B}{\omega U}<\frac{1}{\sqrt{1+\pi^2/4}}\,,$$

i.e. when

$$\frac{B}{\omega U}<0.538.$$

The case of fig. 2 thus occurs in the region

$$0.538<\frac{B}{\omega U}<1. \quad \ldots\ldots \quad (4)$$

If we now calculate the first Fourier component of the output signal when the latter is partly sinusoidal and partly linear with respect to time (see appendix), we find for the amplitude of the first sine term:

$$b_1=\frac{U}{\pi}\{\pi-\arccos p-k(p)-\sin k(p)\cos k(p)+ + p\sqrt{1-p^2}+2p\sin k(p)\}\,, \quad \ldots \quad (5)$$

and for that of the first cosine term:

$$a_1=\frac{U}{\pi}\{\cos k(p)-p\}^2, \quad \ldots \quad (6)$$

where $p = B/\omega U$ and $k = \omega t_{\rm B}$ (cf. fig. 2). From this we can directly derive the amplitude ratio $(\sqrt{a_1^2+b_1^2}/U)$ and the phase shift $(\arctan a_1/b_1)$ for the relevant range of p values. Combining the result with those for the regions $p > 1$ and $p < 0.538$, and plotting a Bode diagram — with the quantity $\omega U/B$ or $1/p$ as the abscissa — we arrive at *fig. 4*. As can be seen, the characteristics closely resemble those of a linear element having one time constant $\tau_{\rm R}$ of magnitude $\pi U/4B$. This time constant is thus proportional to the amplitude of the input signal and inversely proportional to the maximum writing speed.

A characteristic difference is that for p values greater than unity the gain is exactly constant and the phase shift exactly zero. In this region the

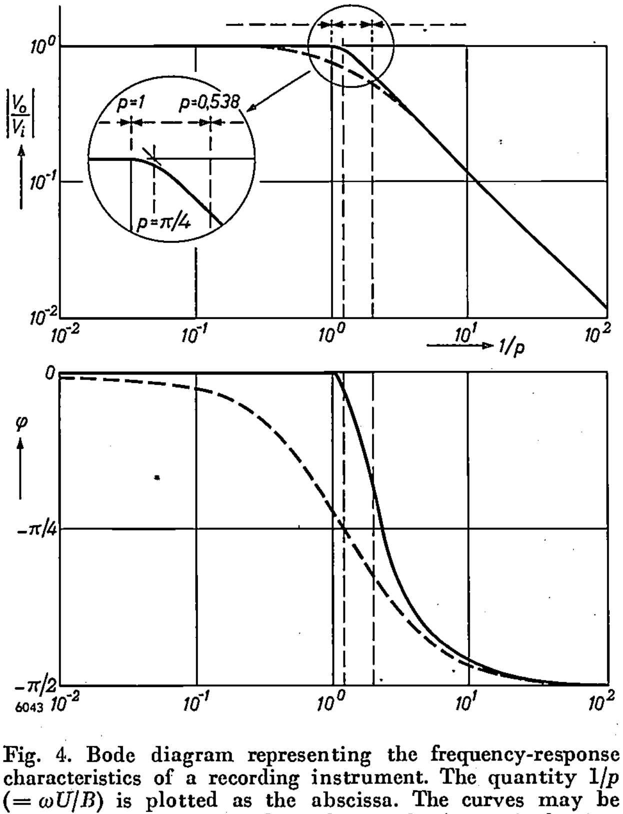


Fig. 4. Bode diagram representing the frequency-response characteristics of a recording instrument. The quantity $1/p$ $(=\omega U/B)$ is plotted as the abscissa. The curves may be approximated by those of an element having a single time constant $\tau_{\rm R}$ of the value $\pi U/4B$ (broken lines).

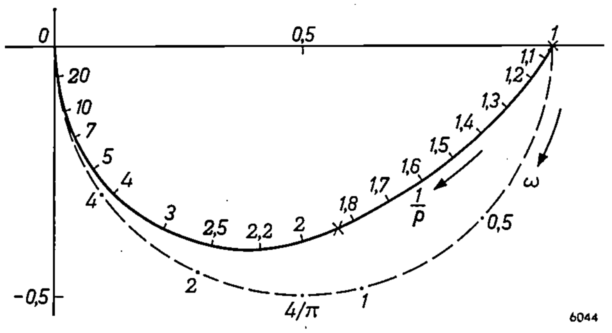


Fig. 5. Nyquist diagram appertaining to fig. 4. The quantity $1/p$ is again chosen as the variable in order to draw one curve instead of a set of curves. The crosses mark the points $p = 1$ and $p = 0.538$. As in fig. 4, the broken curve relates to an element having the transfer function $(1 + \mathrm{j}\omega\pi U/4B)^{-1}$. The figures given beside this curve relate, of course, to ω.

instrument does not behave like an element with a single time constant but in fact like an ideal amplifier. The relevant Nyquist diagram is shown in *fig. 5.* This too is represented in such a way that a point on the curve is not applicable to a certain value of ω but to $1/p$. It should be noted that all points for which $0 < 1/p < 1$ coincide on the real axis.

A recording instrument in a control loop with two time constants

We shall now examine the characteristics of a control loop as shown in *fig. 6.* Here R and A together constitute the recorder; R represents the behaviour of the instrument as such and A the gain independent of R, which is solely determined by the voltage applied to the second potentiometer.

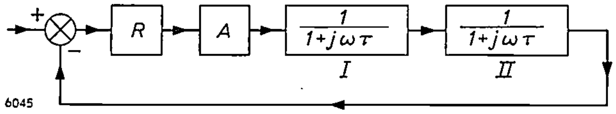


Fig. 6. Block diagram of a control loop consisting of a recorder R, an ideal amplifier with gain factor A, and two elements whose transfer function is $G(\mathrm{j}\omega) = (1 + \mathrm{j}\omega\tau)^{-1}$.

Blocks I and II both have a transfer function of the form $(1 + \mathrm{j}\omega\tau)^{-1}$, but markedly different values of τ. Where $p > 1$, the recorder behaves like an ideal amplifier and the control loop is stable. Where $p < 1$, however, R also contributes to the phase shift, which may therefore in principle be greater than 180°, so that instability may occur. We shall examine this point presently.

First, however, we shall emphasize that, for the purposes of stability considerations, the situation in a loop containing a nonlinear element differs somewhat from that of a loop containing nothing but linear elements. For in this case the Nyquist diagram does not contain simply one curve, which may or may not enclose the point (—1,0), but a set of curves whose parameter is an amplitude — in our case the amplitude of the signal (which may consist only of noise) appearing at the input of the recorder. To be sure of stability, the loop gain should be chosen such that that curve in the complete set of curves which cuts off the largest section of the negative real axis does not enclose the point (—1,0). If that section has no finite value, then stability is out of the question.

A good qualitative insight into the behaviour of the control loop in fig. 6 can be obtained by using the rule of thumb arrived at in the above-mentioned article [2]). This stated that in a control loop which, besides an ideal amplifier, contained solely elements having the transfer function $G(\mathrm{j}\omega) = (1 + \mathrm{j}\omega\tau)^{-1}$, the maximum permissible loop gain $A_{\max}$ is equal to τ_1/τ_2. Here τ_1 is the longest and τ_2 the next longest time constant. If $A = \tau_1/\tau_2$, then the gain drops to unity at the second break in the double-logarithmic amplitude characteristic, i.e. at a phase shift of 135° (90° due to the block with τ_1 and 45° due to that with τ_2). Although the break in the case of the recorder corresponds to a phase shift somewhat smaller than 45°, that does not affect the validity of the argument.

If τ_{R} is initially the longest of the time constants ($\tau_1 = \tau_{\mathrm{R}}$), then τ_{R} determines the position of the *first* break (*fig. 7a*) and an increase in U — we call the initial value U_1 — leads to greater stability, and a decrease to reduced stability. If τ_{R} is the next largest time constant (fig. 7*b*), then τ_{R} determines the position of the *second* break in the curve, and the stability reacts in precisely the opposite way to variations in U.

If we now let the amplitude U pass through a range of values such that τ_{R} begins with the next largest time constant and ends with the largest, and if we start from a stable state ($A = \tau_1/\tau_{\mathrm{R}}$), we then see that as U increases the stability decreases — and may finally result in instability — but that the stability of the system increases again as soon as τ_{R} has become the largest time constant. The stability is therefore not a monotonic function of U, but shows a minimum when τ_{R} is roughly equal to the longer of the two fixed time constants.

An important consequence of this effect is that when a control loop of the type in fig. 6 becomes unstable it does not start to oscillate with ever-increasing amplitude, but enters into a *stationary* state (see below). By measuring the amplitude and frequency occurring in this state for various values of the two fixed time constants we have been able to verify experimentally the theory described above.

We shall work this out quantitatively for the case where the two fixed time constants are identical. Let the transfer function of the recording instrument be approximately $(1 + j\omega\tau_R)^{-1}$ where $\tau_R = \pi U/4B$ (see above), then the transfer function KG of the entire (open) control loop is given by:

$$KG = \frac{A}{(1 + j\omega\tau)^2\,(1 + j\omega\tau_R)} =$$

$$= \frac{A}{(1-\omega^2\tau^2-2\omega^2\tau\tau_R) + j(2\omega\tau + \omega\tau_R-\omega^3\tau^2\tau_R)}\,.$$

The Nyquist diagram for this function, for the case where $A = 10$ and both fixed time constants are equal to one second, is shown in *fig. 8*. Here again, there is not just one curve but a set of curves with τ_R as parameter. The three curves shown relate to cases where τ_R is equal to 0.1, 1.0 and 10 seconds, respectively. As can be seen, the closed loop *is* stable in the two extreme cases, but not when τ_R is 1 second.

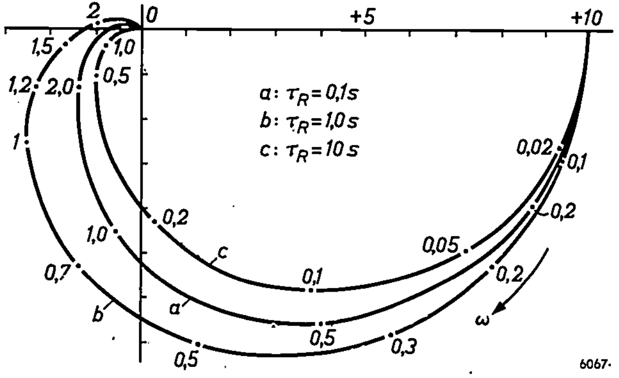


Fig. 8. Nyquist diagram of the control loop in fig. 6, for the case where the two fixed time constants are both 1 second and A is 10. A set of curves is found whose parameter is the time constant τ_R of the recorder. For $\tau_R = 0.1$ sec and 10 sec the closed loop is stable; for $\tau_R = 1.0$ sec it is unstable. The figures beside the curves again give the relevant values of ω.

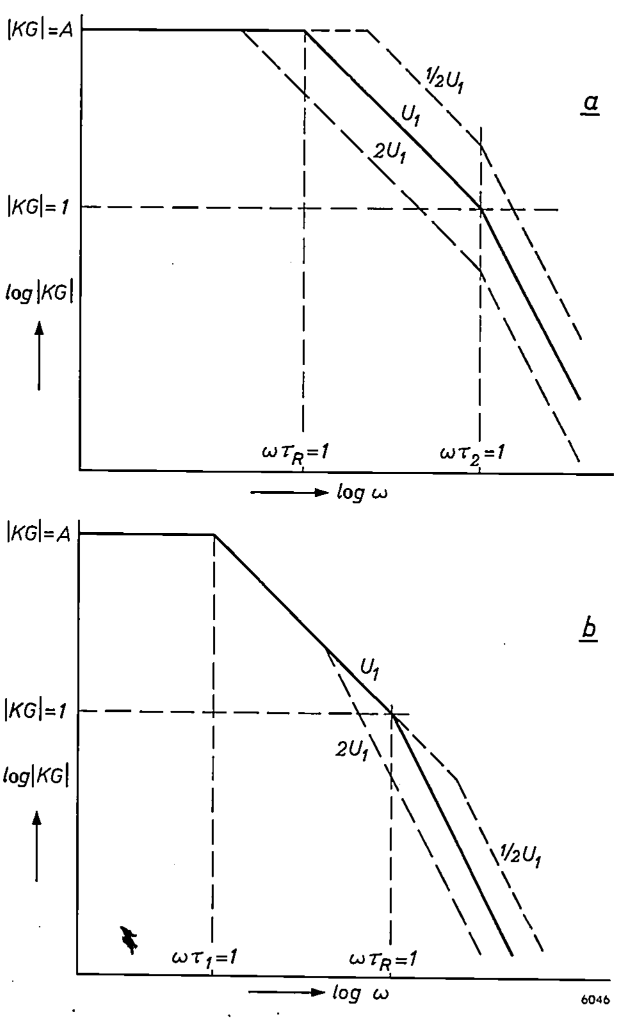


Fig. 7. Bode diagram of the control loop in fig. 6, approximated by straight lines. *a*) The time constant τ_R of R is the longest of the three time constants. When the amplitude U increases (the other parameters remaining constant) the first break shifts to the left and the gain at $\omega = 1/\tau_2$ decreases, as a result of which the stability increases. *b*) τ_R is the next largest time constant and determines the position of the second break in the curve. With increasing amplitude the stability decreases.

The behaviour of the stability as a function of τ_R — i.e. as a function of U/B — can be derived from the displacement of the point where the KG curve intersects the negative real axis. For this purpose we equate the imaginary part with zero:

$$\omega(2\tau + \tau_R - \omega^2\tau^2\tau_R) = 0\,. \quad . \quad . \quad (7)$$

The curve therefore intersects the negative real axis ($\omega \neq 0$) when:

$$\omega^2 = \frac{2\tau + \tau_R}{\tau^2\tau_R}. \quad . \quad . \quad . \quad . \quad . \quad (8)$$

The coordinate of the point of intersection is:

$$\frac{A}{1-\{\tau^2 + 2\,\tau\tau_R\}\left\{\dfrac{2\tau + \tau_R}{\tau^2\tau_R}\right\}} = \frac{A\tau\tau_R}{-2(\tau+\tau_R)^2}. \qquad (9)$$

It follows directly from eq. (9) that the point of intersection tends to the origin when τ_R is very small or very large. The absolute value of the real coordinate is maximum when $\tau_R = \tau$. Substituting this in eq. (9) we find that this maximum value is equal to $A/8$. Where $A > 8$, as in the example of fig. 8, there is therefore a region of τ_R values at which the system is unstable. The limits τ_R' and τ_R'' of that region can be calculated with the aid of eq. (9). We find

$$\tau_R' = \tfrac{1}{4}\tau\,\{(A-4) - \sqrt{A(A-8)}\}, \qquad (10a)$$

and

$$\tau_R'' = \tfrac{1}{4}\tau\,\{(A-4) + \sqrt{A(A-8)}\}. \qquad (10b)$$

If we let τ_R — or the amplitude U, where B is fixed — increase from a low value, the system begins to oscillate as soon as τ_R exceeds the value $\tau_R{}'$. As a result, the amplitude goes on increasing of its own accord to a value corresponding to $\tau_R{}''$. The system then remains oscillating at this amplitude with a frequency given by:

$$\omega^2 = \frac{2\tau + \tau_R{}''}{\tau^2\,\tau_R{}''}. \quad . \; . \; . \; . \; (11)$$

Experiments have shown the measured oscillation frequencies and amplitudes to be in good agreement with the relations derived theoretically.

Appendix: Calculation of the first Fourier components of the output signal

The amplitudes b_n and a_n of the nth sine and cosine terms in the Fourier expansion of the periodic function f(x) are given by the equations:

$$b_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} \mathrm{f}(\xi)\sin n\xi\,\mathrm{d}\xi, \quad . \; . \; . \; . \; . \; (12a)$$

and

$$a_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} \mathrm{f}(\xi)\cos n\xi\,\mathrm{d}\xi. \quad . \; . \; . \; . \; . \; (12b)$$

Writing the nth harmonic in the form

$$\sqrt{a_n{}^2 + b_n{}^2}\,\sin(\omega t + \varphi_n),$$

we find that the phase angle φ_n is equal to arc tan a_n/b_n. If f(x) cannot be described by one analytical function in the whole region from $-\pi$ to $+\pi$, the integrals in (12) must be split into separate integrals whose limits are those within which the relevant expression for f(x) is applicable. In calculating the first Fourier component of the triangular output voltage (fig. 3) we shall disregard the phase — which has already been found by other means — and choose the zero point of the time axis so as to enable us to use a sine series. We then find

$$b_1 = \frac{2}{\pi}\int_0^{\pi/2}\frac{2\Delta}{\pi}x\sin x\,\mathrm{d}x + \frac{2}{\pi}\int_{\pi/2}^{\pi}\frac{2\Delta}{\pi}(\pi - x)\sin x\,\mathrm{d}x = 8\Delta/\pi^2. \quad (13)$$

In order to calculate the first Fourier component of the partly sinusoidal and partly linear output signal (the case of fig. 2) we have to ascertain the moments at which a particular part changes to the next. We shall first consider the transition from a sinusoidal to a linear part (fig. 2, point A). For the relevant moment of time t_A we can write:

$$\omega U \cos \omega t_A = B\,,$$

or

$$\omega t_A = -\,\mathrm{arc}\cos B/\omega U. \quad . \; . \; . \; . \; . \; . \; (14)$$

Putting $B/\omega U = p$, the value V_A of the output voltage V_0 that occurs at $t = t_A$, and which is equal to U sin ωt_A, can be reduced using eq. (14) to:

$$V_A = -U\sqrt{1-p^2}. \quad . \; . \; . \; . \; . \; . \; (15)$$

The equation of the line section AB is then:

$$V_0 = -U\sqrt{1-p^2} + B(t + \frac{1}{\omega}\,\mathrm{arc}\cos p). \quad . \; . \; . \; . \; (16)$$

The point B where the output voltage again becomes sinusoidal is found by ascertaining the value of t at which the line defined by (16) intersects the sinusoidal line:

$$U \sin \omega t_B = -U\sqrt{1-p^2} + B(t_B + \frac{1}{\omega}\,\mathrm{arc}\cos p). \; (17)$$

Putting $\omega t_B = k$, equation (17) transposes to:

$$\sin k + \sqrt{1-p^2} = +kp + p \;\mathrm{arc}\cos p. \; . \; . \; (18)$$

This equation cannot be solved analytically, and therefore no formula can be derived from it for t_B. We have therefore adopted a graphic solution. In *fig. 9* can be seen how k ($=\omega t_B$) varies with p ($=B/\omega U$) in the region $0.538 < p < 1$. The four expressions for f(ξ) to be used here, and the limits between which these expressions are valid, are given in the following table.

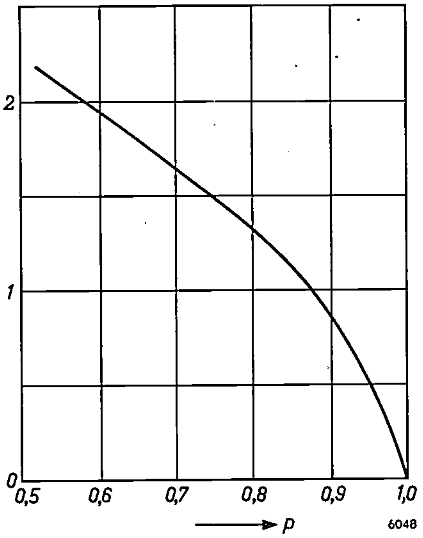


Fig. 9. Variation of the quantity k ($=\omega t_B$) as a function of p ($=B/\omega U$).

function	lower limit $\omega t =$	upper limit $\omega t =$
$Bt + Up\,\mathrm{arc}\cos p - U\sqrt{1-p^2}$	$-\,\mathrm{arc}\cos p$	k
$U \sin \omega t$	k	$\pi - \mathrm{arc}\cos p$
$-Bt + Up\pi - Up\,\mathrm{arc}\cos p + U\sqrt{1-p^2}$	$\pi - \mathrm{arc}\cos p$	$\pi + k$
$U \sin \omega t$	$\pi + k$	$2\pi - \mathrm{arc}\cos p$

The fact that k can only be calculated numerically does not make it impossible to carry out the integrations analytically (see (5) and (6)). Numerical calculation is required only when it is necessary to determine the variation of the coefficients a_1 and b_1 with p.

Summary. When the speed at which the input signal varies exceeds a certain value, the pen of a recorder is no longer able to follow the signal, but moves uniformly at its maximum speed B. In such a case the recording instrument may no longer be regarded as a linear and lag-free element. The frequency-response characteristics found when the output signal is approximated by its first Fourier component are found to resemble closely those of an element having a single time constant. The value of this time constant, however, is here proportional to the amplitude U of the input signal and inversely proportional to B. When this element is included in a control loop having a further two time constants, the stability of the loop is a function of U. The maximum stability is found at very small *and* at very large values of U.

CIRCUITS FOR DIFFERENCE AMPLIFIERS, II

by G. KLEIN *) and J. J. ZAALBERG van ZELST *). 621.375:621.317.725.083.6

*Part II of this article deals with some of the problems that arise in the application of the circuits discussed in part I**). Some other uses for difference amplifiers are described, in particular as an "electronic voltage microscope" and as a logarithmic voltmeter.*

Effect of the input network on the rejection factor

Where a potential difference between two points is to be measured, and a difference amplifier having a high rejection factor is used for this purpose because both points have a high voltage with respect to earth, careful attention should also be paid to the network by which the amplifier is coupled to the points in question. Any asymmetry in that network can ruin the results obtained with a good difference amplifier.

If the measurement is concerned solely with alternating voltages, the amplifier is usually coupled to the points by two capacitors, C and C' (*fig. 18*). Voltage division then occurs across these capacitors and across the input resistances R_i and R_i' of the amplifier. If the products R_iC and $R_i'C'$ are not equal, an in-phase component of E_i and E_i' gives rise to an anti-phase component in the voltage on the input terminals. This coupling network may then be said to have its own finite rejection factor. Provided the discrepancy is not too great, this factor is given by:

$$H = \frac{4\pi f R_i C}{\delta}, \quad \ldots\ldots \quad (17)$$

where f is the signal frequency and δ the relative difference between the products R_iC and $R_i'C'$. Using components for R_i, R_i', C and C' that may show a maximum deviation of 5% from the nominal value, the products R_iC and $R_i'C'$ may show a maximum discrepancy of 20% ($\delta_{\max} = 0.2$). In this case the minimum value of the rejection factor is:

$$H_{\min} = 20\pi f R_i C. \quad \ldots\ldots \quad (18)$$

Given the requirement $H_{\min} = 50\,000$ at $f = 50$ c/s, for example, then the product R_iC must be at least equal to 16 seconds. If R_i and R_i' are rated at 1 MΩ, the capacitors used for C and C' must therefore have a rating of at least 16 μF.

This is a much higher value than would be needed for simply keeping the voltage drop in such a network reasonably low. Simple calculation shows that, for the input signals of the amplifier to differ by no more than 1% from E_i and E_i', the capacitance of C and C' under the conditions mentioned need be only 0.022 μF.

Another important quantity to be considered when using a difference amplifier is the internal resistance of the voltage sources that supply E_i and E_i'. Here too even a slight difference may reduce the rejection factor considerably. In *fig. 19* the internal resistances of the signal sources are denoted by R_b and R_b'. The network sketched can again be said to have a rejection factor, which, if R_i and R_i' are identical and R_b and $R_b' \ll R_i$, is given by:

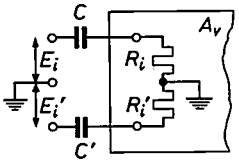


Fig. 18.

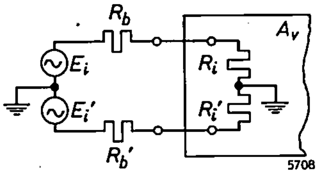


Fig. 19.

Fig. 18. Network for coupling a difference amplifier A_v to the points between which the potential difference is to be measured. Mutual disparities between C and C' and between R_i and R_i' can have a marked effect on the rejection factor.

Fig. 19. Connection of a difference amplifier A_v to two points regarded electrically as voltage sources having internal resistance R_b and R_b'. Any difference in these resistances may considerably reduce the rejection factor.

$$H = 2\,\frac{R_i}{\Delta R_b}. \quad \ldots\ldots\ldots \quad (19)$$

Here ΔR_b is the absolute value of the difference between R_b and R_b'. Where the amplifier is to be used for a variety of purposes, it must be taken into account that in some cases the internal resistance of one of the voltage sources, e.g. R_b', is zero. In that case ΔR_b is equal to the internal resistance of the other voltage source, and therefore:

$$H = \frac{2R_i}{R_b}. \quad \ldots\ldots\ldots \quad (20)$$

*) Research Laboratories, Eindhoven.
) Philips tech. Rev. **23, 142, 1961/62 (No. 5).

When R_b has a specified value, then in order to allow for this unfavourable situation the input resistances of the difference amplifier must be equal, according to (20), to at least:

$$R_i = \tfrac{1}{2} H R_b \, . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

If R_b is 1 kΩ and the minimum acceptable rejection factor of the input circuit is 50 000, then according to (21) the input resistances R_i and R_i' must be at least 25 MΩ. A value of this order is nearly always to be found in DC amplifiers, where the grids of the first valves are directly coupled to the input terminals and where no grid leaks are necessary for these valves. In AC amplifiers, where capacitors are connected between the grids and the input terminals and therefore grid leaks must be used, special measures are sometimes needed in order to obtain the high input impedance required.

In practice, cases are frequently encountered where the rejection factor of the input network is governed both by coupling capacitors and by the internal resistance of the voltage sources. Here again, it is a fairly simple matter to calculate the values which the various resistances and capacitances must have in order to be able to guarantee a specific minimum rejection factor.

Multi-stage difference amplifiers

Hitherto we have been concerned solely with single-stage difference amplifiers. We shall now briefly consider various problems that arise in the design of multi-stage amplifiers. In a previous article [5]) it was shown that as a rule the rejection factor of a difference amplifier is primarily governed by that of the first stage. It should be noted that the rejection factor of a multi-stage amplifier can also be influenced by the coupling elements between the stages: asymmetry in these elements may reduce the rejection factor that can be guaranteed for a given circuit. The considerations applicable to the coupling elements between the stages are similar to those mentioned in regard to the circuit elements used for connecting the difference amplifier to the measuring points. Since lower demands are made on the part of the circuit following the first stage, however, the requirements are not so rigorous.

In AC amplifiers the stages are nearly always coupled in the conventional way by means of capacitors and resistors. In this case, then, the above remarks also apply to these circuit elements. In DC amplifiers, where coupling capacitors obviously cannot be used, the grids of the valves in the second stage can be directly connected to the anodes of the valves of the first stage (see *fig. 20*). In order for the second-stage valves to be biased to their normal operating point, their anodes and cathodes must have a higher potential than the corresponding electrodes in the previous stage. The higher supply voltages then needed may be felt as a drawback. To get around this difficulty, voltage dividers can be used for the coupling between the various stages (*fig. 21*), thereby lowering the "voltage level" of the second and successive stages. Of course, this has the effect of reducing the sensitivity of the amplifier. An even greater objection to the use of voltage dividers is that they increase the output resistances of the first stage and lower the input

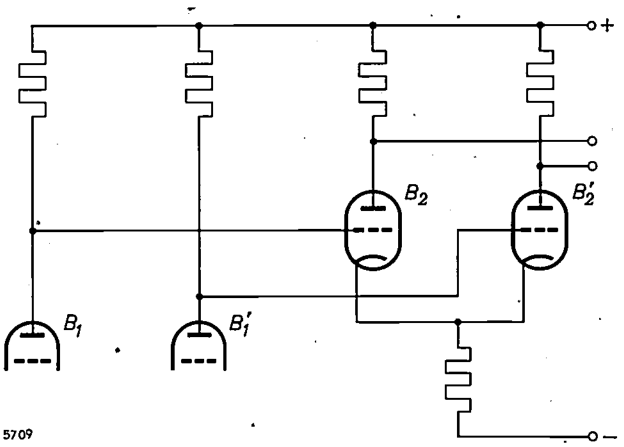


Fig. 20. Difference amplifier for DC voltages with direct interstage coupling.

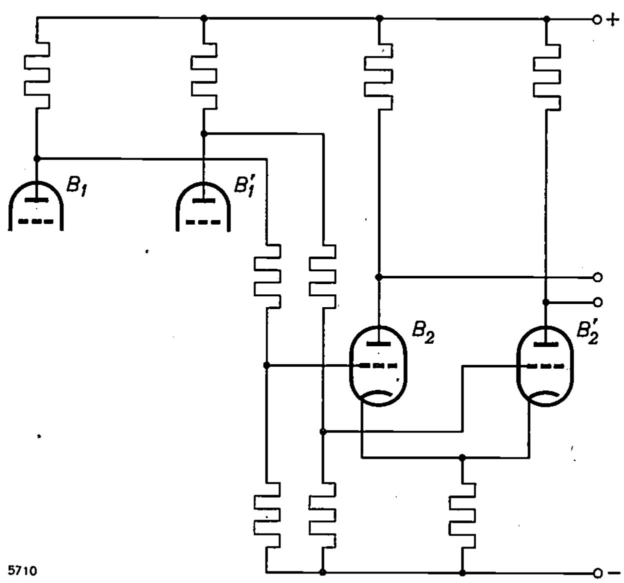


Fig. 21. Difference amplifier for DC voltages, with the two stages coupled via voltage dividers.

[5]) G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. 22, 345-351, 1960/61 (No. 11).

resistances of the second stage; in connection with the mutual disparity between these resistances, the result is that the guaranteed rejection factor for the coupling network is lower (see eq. (20)). For this reason, in DC difference amplifiers where very high demands are made on $H_{\min}$ the grids of the valves in the second stage are frequently connected directly to the anodes of the valves in the first stage. In the further stages the coupling can be as shown in fig. 21 (see also fig. 26).

Reducing the DC voltage level without any appreciable loss in sensitivity can be achieved with a circuit using elements whose differential resistance is much higher than their DC resistance. A circuit of this type is shown in *fig. 22*. The elements having a very high differential resistance are formed by triodes B_2 and B_2' with resistances R_{k2} and R_{k2}' incorporated in the cathode leads. If the resistances R_2 and R_2' are small compared with these differential resistances, the voltage level of points *c* and *d* can be much lower than that of points *a* and *b*, although the signal voltages are passed with virtually no attenuation. An arrangement as sketched in fig. 22 can be used with particular advantage where the difference amplifier is required to deliver strong output signals, e.g. for deflecting the beam in a cathode ray tube. In this way it is possible to avoid the difficulties that may arise from the use of a voltage divider built up from normal resistors, owing to the fact that the last stage then has to supply a signal voltage several times higher than the voltage taken from the divider.

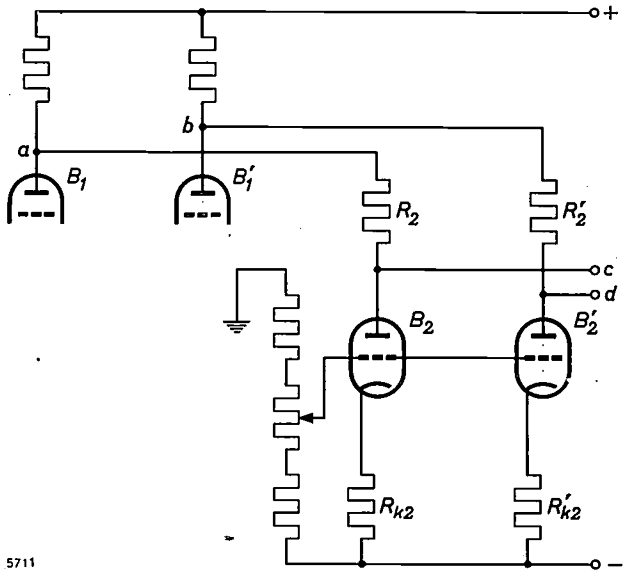


Fig. 22. Circuit in which the DC voltage level of points *c* and *d* is lower than that of points *a* and *b*, although there is scarcely any attenuation of the signal voltage.

Gain control

When a difference amplifier is built up from several stages, gain control will generally be wanted. When choosing the appropriate circuit it should be borne in mind that the gain control too may reduce the rejection factor. For this reason it is usually inadvisable to apply the gain control to the first stage, the rejection factor of which has to meet the highest demands.

A widely used method of gain control — varying the transconductance of the valves by changing the negative grid bias — is not effective here in view of the high differential resistances in the cathode leads.

A severe drawback also attaches to the method represented in *fig. 23*. If it is used in a DC amplifier, its effect is also to alter the operating point of the valves in the next stage. Here too, the guaranteed rejection factor is lowered, owing to the mutual disparity in the voltage-division ratios of the potentiometers.

Fig. 24 indicates an arrangement with which the gain of the difference amplifier can be varied without altering the operating points of the valves. The gain for anti-phase signals is controlled by the variable resistance between the two anodes. This resistance does not, however, affect the gain for in-phase signals, and therefore the discrimination factor *F* varies in the stage whose gain is controlled. Consequently the rejection factor *H* of the next stage has to meet higher demands.

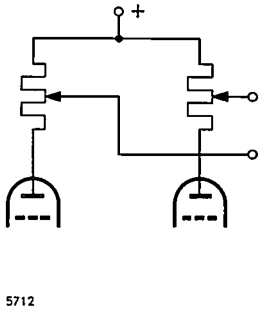


Fig. 23.

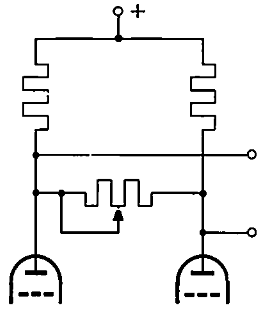


Fig. 24.

Fig. 23. Method of gain control. A discrepancy in the voltage-division ratios of the two potentiometers may result in a lower rejection factor. If this circuit is used in a DC amplifier, the setting of the gain control affects the operating point of the valves in the following stage.

Fig. 24. Method of gain control where the discrimination factor is dependent on the value of the variable resistance.

The same can be said of the circuit sketched in *fig. 25a*, where a variable resistance is inserted between the two cathodes. Here again, the magnitude of this resistance determines the gain for anti-phase signals, but has no influence on the gain for in-phase signals. Increasing the gain therefore again

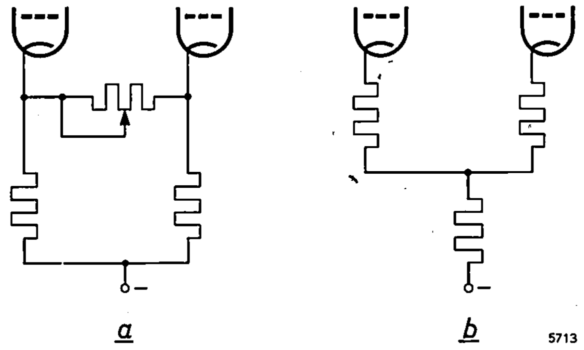


Fig. 25. *a*) Method of gain control, where both the discrimination factor and the rejection factor are dependent on the value of the variable resistance.
b) Equivalent circuit, with the delta resistance network replaced by a star network.

reduces the discrimination factor. Moreover, the variable resistance in this circuit can also influence the rejection factor. This can be seen most readily if we replace the delta network of resistances by an equivalent star network, as in fig. 25*b*. In this arrangement there is negative feedback as a result of the resistances in the cathode leads of the valves. These resistances are roughly equal to half the resistance between the cathodes in fig. 25*a*. The result of this negative feedback is to reduce the effective transconductance of the valves, which again reduces the guaranteed rejection factor. This network too should therefore preferably be applied to one of the last stages of a difference amplifier, where the rejection factor is not so critical.

As a further illustration of the methods of controlling the gain, *fig. 26* shows the circuit diagram of a 3-stage difference amplifier. The second stage contains a 3-step volume control as shown in fig. 24, and the gain of the third stage is controlled on the principle represented in fig. 25. Further particulars of this circuit will be found in the caption to the figure.

Influence of supply voltages; stability

In a sensitive, unbalanced amplifier designed for signals of very low frequency the constancy of the supply voltages, particularly for the first stage, is always an important consideration. A fluctuation in the anode supply voltage, for example, can produce a change in the output voltage from the valves in the first stage which is amplified by the following stages and thus occurs as an interference component in the amplified signal. It is important to note that, for a given sensitivity, the demands made on the constancy of the supply voltages for a difference amplifier need not be as high as in the case of a normal amplifier. This can be understood by considering a difference amplifier in its simplest form, with triodes whose control grids in the quiescent state are at earth potential (see fig. 1). A change in the positive and negative supply voltages by the same amount in the same direction, corresponds to an in-phase signal at the input terminals. This in-phase signal appears at the output terminals attenuated by the rejection factor with respect to the anti-phase signal to be amplified.

If only one of the two supply voltages changes, the effect on the output signal is not so simple to analyse. It can be shown that a change in the *positive* supply voltage of the first stage appears in the output signal as an anti-phase signal which is attenuated with respect to the input signal by a factor

$$\frac{2\mu^2}{\Delta\mu},$$

and that the corresponding factor for a change in the *negative* supply voltage is:

$$\frac{4SR_k}{\frac{\Delta S}{S}+\frac{\Delta R_a}{R_a}+\frac{1}{4}\frac{R_a}{R_k}\frac{\Delta\mu}{\mu^2}}.$$

From equation (5) we see that the sum of the reciprocals of these two factors is equal to $1/H$, which confirms the effect deduced above of a simultaneous change of both supply voltages in the same direction. Since the rejection factor is at least equal to $H_{\min}$, both the above attenuation factors are always greater than the minimum value of the rejection factor.

More complicated circuits also involve auxiliary voltages, which are often derived for simplicity from the positive and negative supply voltages and are therefore affected by changes in the latter (see e.g. figs 10 and 11). It can be shown that the disturbances thus introduced are always a few orders of magnitude smaller than those occurring in an unbalanced amplifier.

As the thermionic emission of a valve depends on the *heater voltage*, and this dependency differs from one valve to another, changes in heater voltage will also appear at the output terminals as an anti-phase signal, obviously of very low frequency. The magnitude of this anti-phase signal depends on the construction of the valves, and its maximum value is therefore dependent on the type of valve used. Experiments with numerous valves have shown that in this respect the type E 80 CC triode gives the best results. It was found that in a difference amplifier fitted with this type of valve a change of

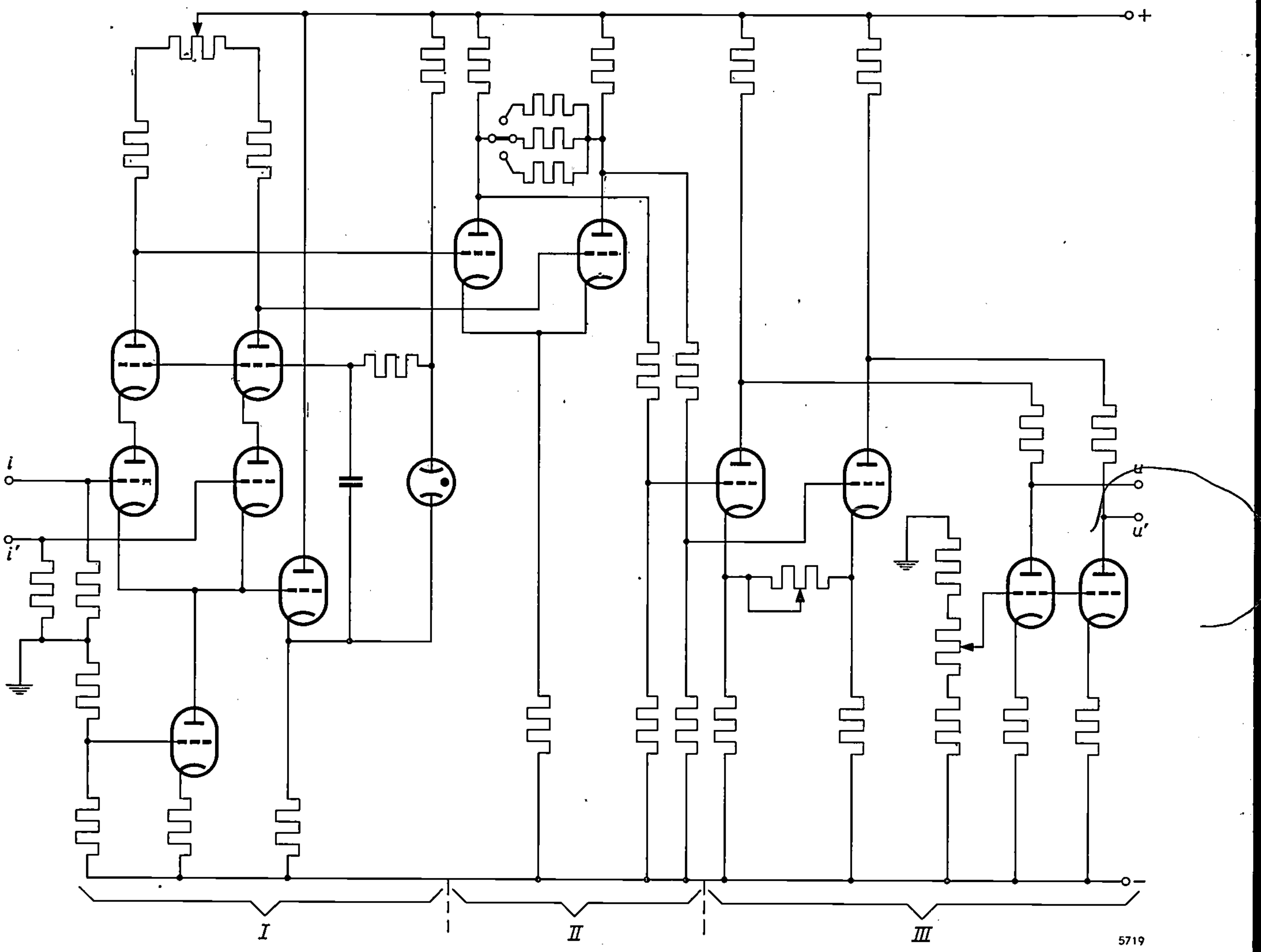


Fig. 26. Circuit diagram of a three-stage difference amplifier. The stages are denoted by *I*, *II* and *III*. Stage *I* is designed as in fig. 14; stages *II* and *III* are simple circuits, each with two amplifying triodes. Stages *I* and *II* are directly coupled (cf. fig. 20); stages *II* and *III* are coupled via voltage dividers to reduce the DC voltage level (fig. 21). Stage *II* uses three-step gain control as in fig. 24; the gain control in *III* is as in fig. 25. The output terminals are given earth potential in the quiescent state by the circuit shown in fig. 22. The different heights at which the valves are drawn correspond to the differences in their DC voltage levels.

10% in the heater voltage caused an interfering anti-phase signal the maximum value of which corresponded to an anti-phase signal of 10 mV at the input. Although this disturbance is small compared with the corresponding disturbance in an unbalanced amplifier (100 to 200 mV), it is still excessive in many cases. For difference amplifiers too, therefore, it may be necessary to ensure that the heater voltage remains reasonably constant, with variations considerably less than 10%.

Another complication frequently encountered with sensitive amplifiers is the occurrence of feedback via the supply circuit, which may even give rise to oscillation. In a *balanced amplifier* there is generally much less feedback of this kind than in an unbalanced amplifier, owing to the fact that the current variations in the output valves are in anti-phase and the supply circuit need therefore deliver hardly any current varying with the signals. A *difference amplifier* is even more favourable in this respect, because, as shown above, a signal voltage returned from the output via the supply circuit to one of the previous stages undergoes very little amplification.

To a considerable extent the advantages mentioned can often be obtained by designing only the *first stage* as a difference amplifier and the remainder as normal amplifying stages. The effect of supply-voltage fluctuations and any tendency to instability can frequently be substantially suppressed in this way. Usually, however, the full advantages of a difference amplifier are only obtained by designing the amplifier with *all* its stages as difference amplifiers. Since the stages following the first stage can usually be fairly simple in circuitry, an entirely balanced amplifier may even in fact be simpler than an amplifier which is partly unbalanced.

The above-mentioned advantages of a difference amplifier enable such an amplifier to be used in cases where it is not a question of amplifying the voltage difference between two arbitrary points, but the potential of one point to earth. One of the two input terminals is then earthed (*fig. 27*) and the difference amplifier is used as a "normal" amplifier. The output voltage may be taken either in push-pull from the output terminals or between one of these terminals and earth, as desired.

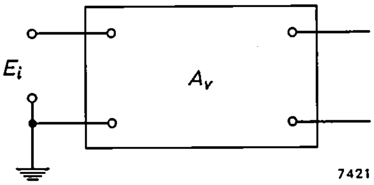


Fig. 27. Difference amplifier used as an unbalanced amplifier.

Negative feedback

It is easily seen that the discrimination factor of a difference amplifier is lowered if a simple form of negative feedback is introduced, the in-phase and anti-phase signals being returned in the same ratio from the output terminals to the input. Since the in-phase signals undergo much less amplification, the feedback also reduces the gain for these signals much less than for the anti-phase signals. In a difference-amplifier with negative feedback, then, the feedback ratio for in-phase signals should be much larger than for antiphase signals. Because of the interaction of in-phase and anti-phase signals, it is not so easy to see what effect the feedback has on the rejection factor. For this reason we shall not be concerned in this article with the problems arising from the use of feedback in a difference amplifier.

Quite another matter is the fact that a difference amplifier can be used as a means of producing highly effective negative feedback in an unbalanced amplifier. In this procedure a signal voltage derived from the output is returned in the usual way to the input stage of the amplifier, the aim being to amplify the difference between the input signal and the feedback signal. In a commonly used circuit the latter signal is applied to the cathode of the first valve, which then in fact functions as a difference amplifier. After what has been said it will be clear that the rejection factor of such a "difference amplifier" will as a rule be very small. Although the object of the feedback, i.e. to reduce distortion and minimize the extent to which the parameters of valves and other components influence the gain, may be satisfactorily achieved, better results are possible if a good difference amplifier is used as the first stage (see *fig. 28*). The feedback is then more effective, because the input signal and the feedback signal contribute almost equally to the output signal. This also offers advantages in circuitry, since there is hardly any load on the feedback network and because the feedback signal can be applied to the first stage at earth potential.

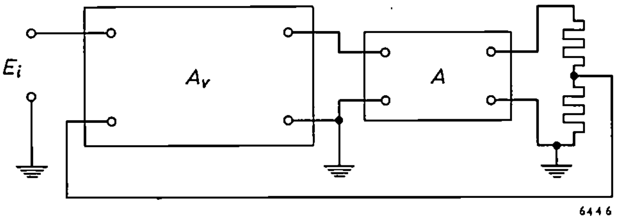


Fig. 28. Difference amplifier used as the first stage in a feedback amplifier.

The use of difference amplifiers with large anti-phase signals

Where high voltages are to be measured or otherwise investigated, the use of a high-gain amplifier is seldom considered. Nevertheless a sensitive difference amplifier offers advantages here that are not so easily obtained by other means. To make this clear, we should first of all point out that in the above theory on the operation of a difference amplifier we assumed that the anti-phase signal on the grids is small enough to allow a reasonable current to flow in both valves, this signal then being amplified as in a conventional balanced amplifier. There is no amplification, however, if the anti-phase signal exceeds the above-mentioned limit, for in that case the anode current in one of the valves is cut off, with the result that the other valve functions as an unbalanced amplifier with a very high cathode impedance. The gain of this valve is then extremely low. It may further be said that a difference amplifier gives amplification only when the difference in potential between the two input terminals does not exceed a specific value. As soon as this potential difference exceeds that value the difference amplifier is "overdriven". This does not mean, however, that the valves are overloaded. If the input signal of a multi-stage amplifier is progressively increased, it will generally be the valves in the last stage that are "overdriven" first, since it is here that the signals are strongest. If the gain for anti-phase signals is high, only a small potential difference between the input terminals is sufficient to overdrive the last stage. The amplifier may be so designed, for example, that amplification occurs only if the potential difference between the input terminals amounts to no more than a few millivolts, and if necessary even less.

This property of a difference amplifier can occasionally be turned to good use, particularly for the purpose of very accurately comparing two differently time-dependent signals at the moments when they are almost identical. Suppose, for example, that the one input voltage, E_i, is a DC signal and the other, E_i', a large AC signal (see *fig. 29*), then the latter signal will only be amplified at the moments at which its instantaneous value differs only very slightly from the magnitude of the DC signal. By connecting the output of the difference amplifier to an oscilloscope, very small portions of the waveform of the alternating voltage can then be displayed distinct from the remainder of the wave form. The portions concerned are drawn thick in fig. 29. The combination of a difference amplifier and oscilloscope in this way may be described as an "electronic voltage microscope". Using a high-gain difference amplifier, it is thus possible to display a detail of a few millivolts of a waveform whose amplitude is ten or more volts. If E_i is made roughly equal to the peak value of E_i', the "voltage microscope" displays only the peaks of the AC signal (*fig. 30*). This constitutes a highly accurate method of checking the constancy of the signal amplitude. *Fig. 31* shows an example of such an oscillogram, obtained by applying to one input terminal of the difference amplifier an alternating voltage of 10 V amplitude and 80 c/s frequency, and to the other a DC voltage of 10 V. The height of a square in the figure corresponds to 2 mV. It can be seen that amplitude variations of roughly 4 mV occur, i.e. 0.04%. *Fig. 32* shows the top portion of a 10 V square-wave voltage. We see here that the tops are not perfectly flat but show variations in amplitude of about 2 mV, i.e. 0.02%.

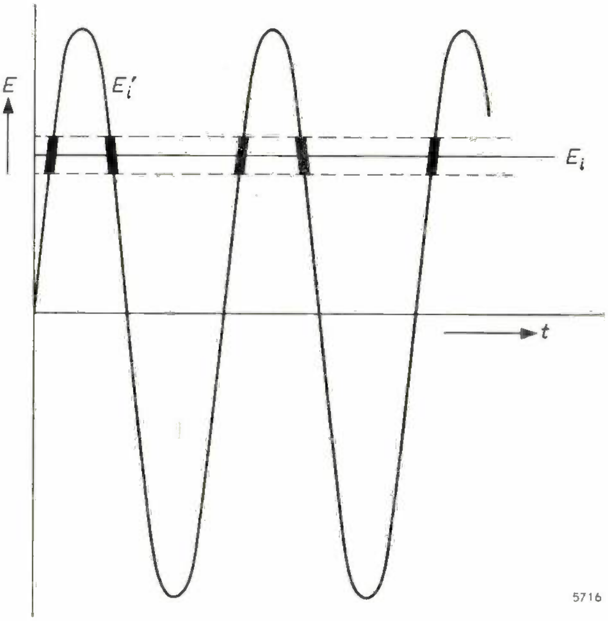


Fig. 29. Amplitude-versus-time plot of the voltages E_i and E_i' on the input terminals of a difference amplifier when the latter is used in combination with an oscilloscope as an "electronic voltage microscope". Only the thickly outlined portions of E_i' are displayed on the oscilloscope.

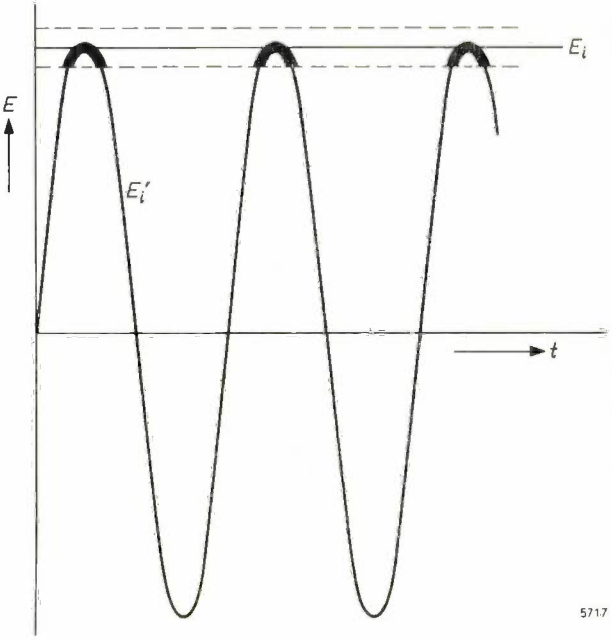


Fig. 30. When the DC voltage E_i is made roughly equal to the amplitude of E_i', only the peaks of E_i' appear on the oscilloscope.

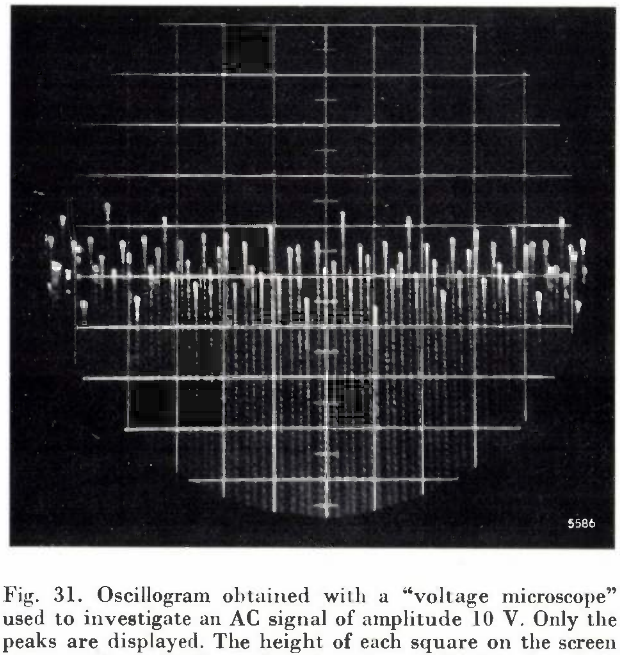


Fig. 31. Oscillogram obtained with a "voltage microscope" used to investigate an AC signal of amplitude 10 V. Only the peaks are displayed. The height of each square on the screen corresponds to 2 mV, so that in order to display the whole waveform the paper would have to be 35 m in height! Very small variations in amplitude can be demonstrated in this manner.

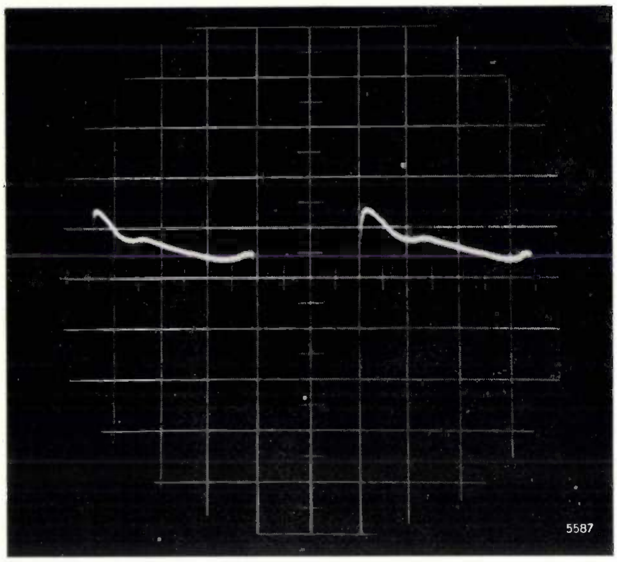


Fig. 32. Top of a 10-V square-wave voltage. One square corresponds in height to 2 mV. The tops are not flat but show variations in amplitude of roughly 2 mV.

Since a DC voltage can as a rule be measured directly with greater precision than an AC voltage (e.g. using a compensator), a difference amplifier also makes it possible to determine the amplitude of an AC signal in a very accurate but simple manner. The procedure is simply to make E_i equal to the amplitude of the AC signal to be measured, E_i', as illustrated in figs 30 and 31, and to measure E_i.

To conclude, we shall describe another application of a difference amplifier where a high DC voltage E_i is applied to one input terminal and a periodically varying voltage E_i' is applied to the other. In fig. 29 the waveform of E_i' is sinusoidal, but we shall now assume that E_i' is a different periodic function of time. As an example, it is assumed in *fig. 33* that E_i' decays exponentially in each cycle:

$$E_\mathrm{i}' = E_\mathrm{i0}'\,\mathrm{e}^{-\frac{t}{T_0}}, \qquad (22)$$

where T_0 is a constant.

Anode current now flows in both valves of the last stage only as long as the signal voltage on the grid of the relevant valve is greater than that on the grid of the other valve. During these times the anode currents are practically constant; the output signal E_o of the difference amplifier thus has a square-wave form, as shown in fig. 33 below. A simple calculation shows that the mean value E_om of E_o is a logarithmic function of E_i:

$$E_\mathrm{om} = \frac{E_1}{T}(2T_0 \ln E_\mathrm{i0}' - 2T_0 \ln E_\mathrm{i} - T) \qquad (23)$$

(E_1 and T are explained in fig. 33). The value E_om, measured with an integrating circuit, is a measure of E_i on a logarithmic scale, and the whole arrangement thus constitutes a *logarithmic voltmeter*.

If E_i' is a periodic exponentional function of time, we thus obtain an output signal whose mean value is the *inverse* (logarithmic) function of E_i. It is not difficult to see that, even if E_i' is some other periodic function of time, this method still produces an output signal which is the inverse function of E_i. This can be turned to use in various ways [6]).

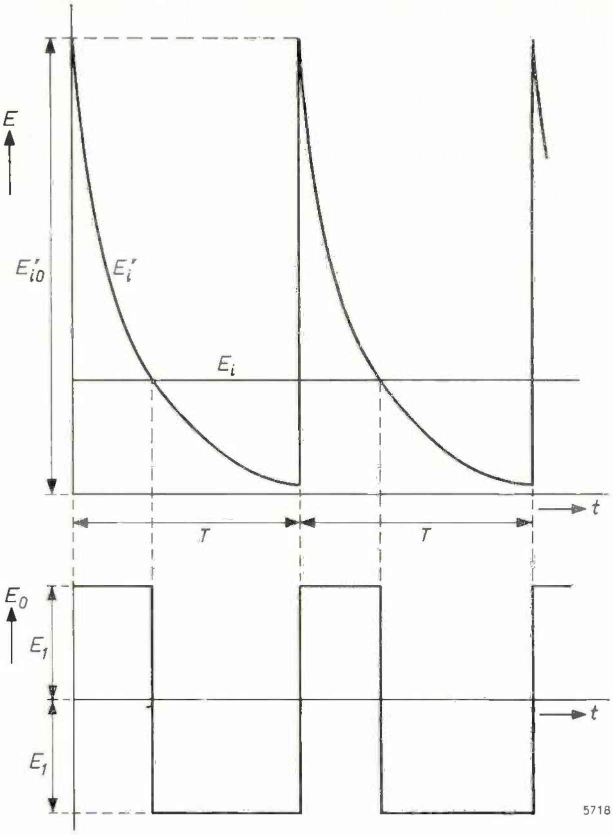


Fig. 33. Use of a difference amplifier as a logarithmic voltmeter. *Above:* The two input signals, E_i and E_i', as functions of time. *Below:* Output signal as a function of time.

Without going deeper into these and other possible applications of difference amplifiers, it is hoped that the above examples have shown that there are many more uses for these amplifiers than simply the amplification of small potential differences.

[6]) See G. Klein and J. M. den Hertog, A sine-wave generator with periods of hours, Electronic Engng. 31, 320-325, 1959.

Summary. The minimum value of the rejection factor of a difference amplifier is affected by the elements coupling the amplifier to the points between which the potential is to be measured. In a multi-stage amplifier, the coupling elements between the stages also have an important influence. An incidental advantage of a difference amplifier compared with normal types is that less rigorous demands are made on the constancy of the supply voltage; a difference amplifier also shows much less tendency to oscillate. Further applications for difference amplifiers are discussed, in particular as a means of introducing highly effective feedback in an unbalanced amplifier, as an "electronic voltage microscope" and as a logarithmic voltmeter.

INSPECTION OF NEGATIVES FOR PRINTED CIRCUITS

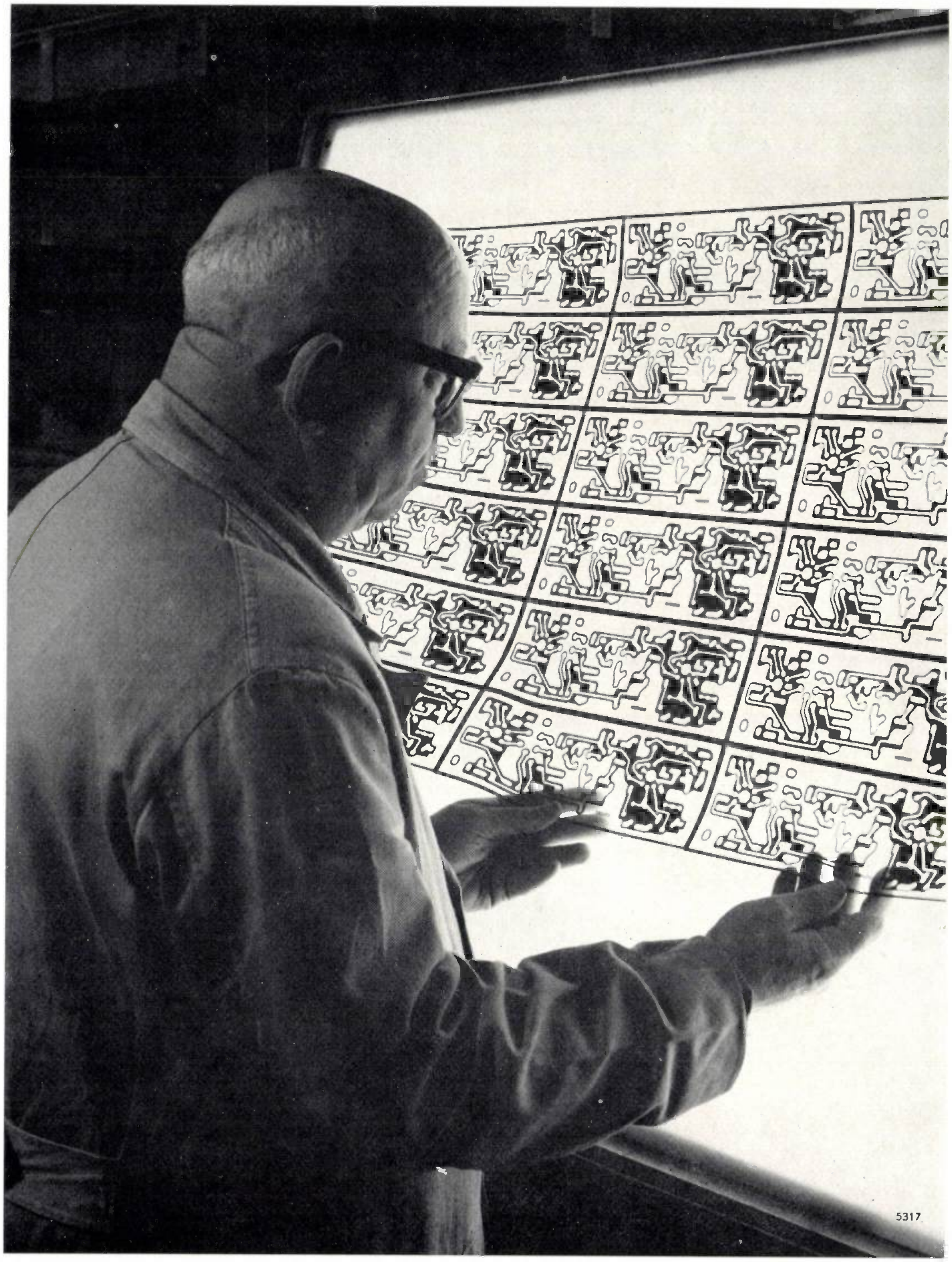


Photo Maurice Broomfield

Inspection of a master negative used in the manufacture of printed circuits by the photographic etched-foil process. Prints of this negative are made on a copper-plated panel of laminated board coated with photographic emulsion. The unexposed parts of the copper foil are removed by etching. This method lends itself particularly well to the reproduction of fine detail. Since the slightest fault is reproduced in the finished product, the negatives are subjected to careful scrutiny all the time they are in use.

RADIATION FURNACES OF PAST CENTURIES

662.997:621.472

The carbon-arc image furnace, whose use as a laboratory instrument is described in this issue (page 161), has had numerous predecessors. The use of lenses or concave mirrors to focus the sun's rays for lighting kitchen fires or altar flames was known to antiquity [1]). In the 16th and 17th centuries radiation furnaces had found a place in the equipment of metallurgists and alchemists. In 1540 Biringuccio [2]) mentioned a mirror roughly a foot in diameter with which a gold ducat could be melted. Glauber [3]) in c. 1650 indicated the size of mirror required for specific purposes: 1 span (= 9 inches) for igniting wood, 2 spans for melting tin, lead and bismuth, 4 or 5 spans for melting gold and silver and also for forging iron. Robert Hooke even planned to base a kind of temperature scale on this scheme.

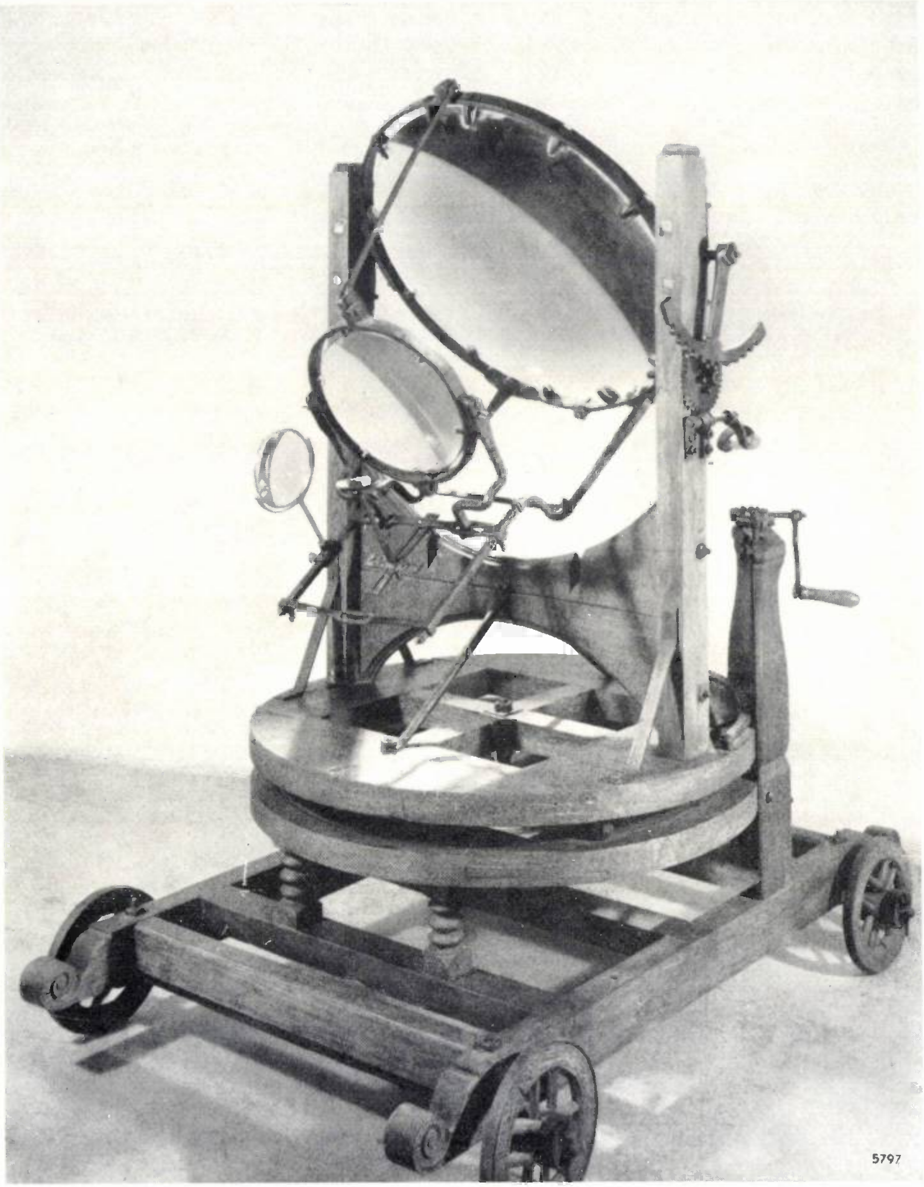


Fig. 1. Burning-glass instrument built by E. W. von Tschirnhaus towards the end of the 17th century. (Reproduced by courtesy of the Deutsches Museum, Munich.)

[1]) Various sources are mentioned by R. J. Forbes, Studies in ancient technology, VI, Brill, Leyden 1958.
For some of the particulars mentioned here we are indebted to Professor Forbes personally.

[2]) V. Biringuccio, De la Pirotechnia libri X, Roffinello, Venice 1540.

[3]) J. R. Glauber, Opera mineralis, Amsterdam 1651.

At the end of the 17th century, the century in which the first lens systems for telescopes had been built and considerable progress made in the grinding of lenses, mirrors were for a time outstripped by lenses. This was largely due to the work of the German mathematician and physicist E. W. von Tschirnhaus (1651-1708)[4]. He had occupied himself with the improvement of "burning mirrors" since 1679, and in 1687 he is reported to have made a gigantic mirror 130 cm in diameter, beaten from sheet copper. He then apparently realized that he could do better with lenses than with the imperfectly focussing mirrors (the art of grinding parabolic mirrors had not yet been perfected), which were moreover difficult to handle. He set up a glass works, where larger pieces of glass were cast than any one had managed to do before, he developed special grinding methods, and he hit on the idea of combining a large lens with a smaller one (the "collective"), which improved focussing. In February 1694 he stated in a letter to Leibniz that he had succeeded in melting with his lenses several substances which were previously considered to be unmeltable. Among these substances was clay; this discovery enabled him to make porcelain, which had been imported from China for centuries but which no one in Europe had previously known how to imitate[5].

Tschirnhaus' great lens systems were greeted with much enthusiasm, and were sold to physical societies in various European countries, where they were used for scientific or pseudo-scientific ends. One went to Holland, apparently (Leibniz wrote about it to Huygens in 1694); Tschirnhaus sent two to Paris — one of them with an enormous lens 94 cm in diameter and weighing 74 kg; other lenses went to London, St. Petersburg etc. Several of these lens systems are preserved. The Lomonosov Museum in Leningrad still houses the lens of the equipment Tschirnhaus had sent there, with a diameter of 57.5 cm and still in its original wooden housing. The Mathematisch-Physikalische Salon in Dresden also has several pieces of Tschirnhaus equipment[6]. But particularly well preserved is the complete Tschirnhaus "burning glass" in the Deutsches Museum, Munich, which is shown in *fig. 1*. This has an objective lens 75 cm in diameter.

A somewhat smaller instrument (objective lens about 37 cm diameter) is to be found in the Museo di Storia della Scienza in Florence[7]. Benedetto

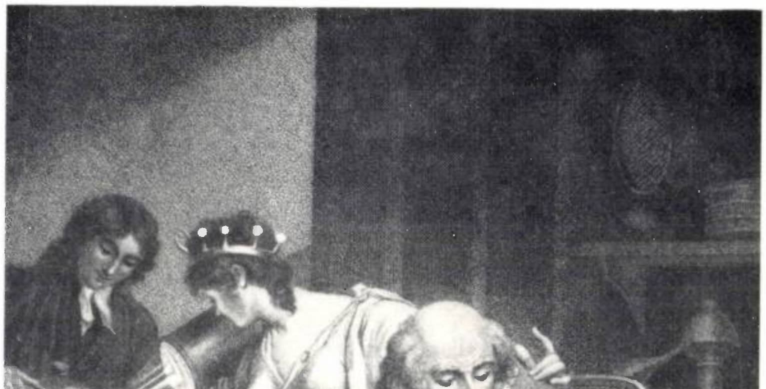


Fig. 2. Mezzotint engraving by John Chapman after a painting by Richard Corbould from 1805, symbolizing the Science of Chemistry at the crossroads. The chemist in the background is demonstrating the preparation of oxygen by decomposition of an oxide in a solar furnace. (Reproduced by courtesy of Prof. John Read, University of St. Andrews, Scotland. See: J. Read, The alchemist in life, literature and art, Nelson, London 1947.)

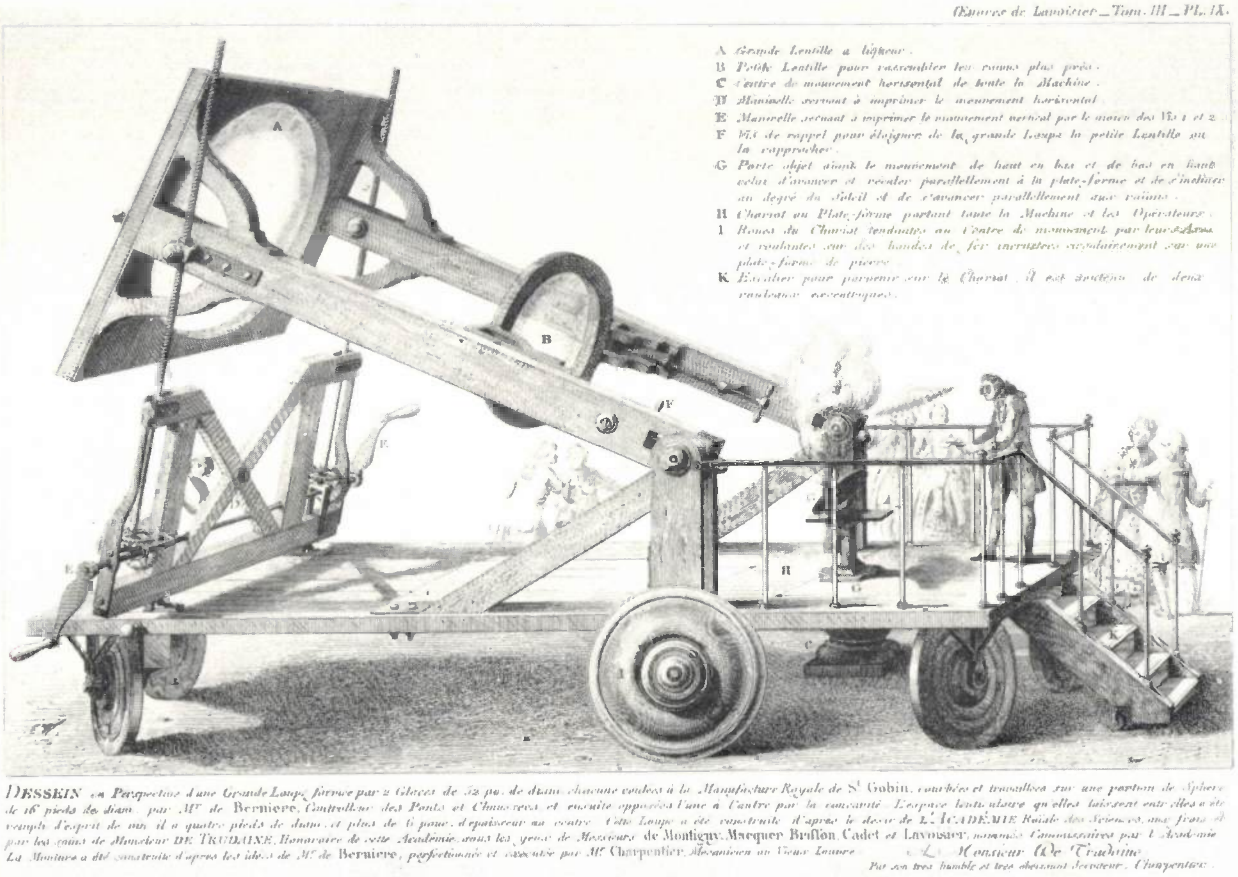
[4] For the details given below see especially: J. S. T. Gehler, Physikalisches Wörterbuch, Leipzig 1787, and the more recent publication: R. Wunderlich, Brenngläser als Hilfsmittel chemischen Forschens, Chymia 2, 37-43, 1949.

[5] E. W. von Tschirnhaus, De magnis lentibus seu vitris causticis eorumque usu et effectu, Acta eruditorum (Leipzig) 1697, pp. 414 et seq.

[6] For details see: E. W. von Tschirnhaus und die Frühaufklärung in Mittel- und Osteuropa, Ed. E. Winter, Akademieverlag Berlin 1960, especially the articles by O. Volk (pp. 247-265) and V. L. Cenakal (pp. 285-307).

[7] Maria L. Bonelli, The burning glass of Benedetto Bregans of Dresden, "Florence" 11, No. 2, p. 22, 1960.

Bregans brought this from Dresden in 1690; it thus probably came from Tschirnhaus' workshop, too. In 1710 Bregans presented his burning glass as a gift to Grand Duke Cosimo III of Tuscany. It is interesting to note that the Florence instrument was put to use a century later for a scientific investigation undertaken by Humphrey Davy and his assistant at that time, Michael Faraday. Davy and Faraday visited Florence in 1814, and Davy profited from the opportunity to use the burning glass for studying the combustion of diamond — a subject which had already occupied Cosimo III, and which Lavoisier *et al.* had thoroughly investigated starting in 1772, partly with the aid of the Tschirnhaus burning glass in Paris [8]). Faraday gave a detailed description of the experiment in his diary. His report begins: "Today we made the grand experiment of burning the diamond, and certainly the phenomena presented were extremely beautiful and interesting". The diamond in the experiment was placed in a perforated platinum crucible in the middle of a glass sphere filled with pure oxygen (volume 22 cubic inches). Under the heat of the solar image produced by the "Duke's burning glass" for three quarters of an hour — with interruptions to cool the glass sphere — the diamond seemed gradually to shrink and become opaque, and finally caught fire, whereupon it burnt away in a few minutes without further heating.


Œuvres de Lavoisier _ Tom. III _ Pl. IX.
A Grande Lentille à liqueur.
B Petite Lentille pour rassembler les rayons plus près.
C Centre de mouvement horizontal de toute la Machine.
D Manivelle servant à imprimer le mouvement horizontal.
E Manivelle servant à imprimer le mouvement vertical par le moyen des Vis 1 et 2.
F Vis de rappel pour éloigner de la grande Loupe la petite Lentille ou la rapprocher.
G Porte objet ayant le mouvement de haut en bas et de bas en haut celui d'avancer et reculer parallèlement à la plate-forme et de s'incliner au degré du Soleil et de s'avancer parallèlement aux rayons.
H Chariot ou Plate-forme portant toute la Machine et les Opérateurs.
I Roues du Chariot tendantes au Centre de mouvement par leurs axes et roulantes sur des bandes de fer incrustées circulairement sur une plate-forme de pierre.
K Escalier pour parvenir sur le Chariot. Il est soutenu de deux rouleaux excentriques.
A
B
F
E
H
DESSEIN en Perspective d'une Grande Loupe formée par 2 Glaces de 52 po. de diam. chacune coulées à la Manufacture Royale de St Gobin, courbées et travaillées sur une portion de Sphère de 16 pieds de diam. par Mr de Bernière, Contrôleur des Ponts et Chaussées et ensuite opposées l'une à l'autre par la concavité. L'espace lenticulaire qu'elles laissent entre elles a été rempli d'esprit de vin; il a quatre pieds de diam. et plus de 6 pouc. d'épaisseur au centre. Cette Loupe a été construite d'après le desir de L'ACADÉMIE Royale des Sciences, aux frais et par les soins de Monsieur DE TRUDAINE, Honoraire de cette Académie, sous les yeux de Messieurs de Montigny, Macquer Brisson, Cadet et Lavoisier, nommés Commissaires par l'Académie. La Monture a été construite d'après les idées de Mr de Bernière, perfectionnée et exécutée par Mr Charpentier, Mécanicien au Vieux Louvre.
A Monsieur De Trudaine
Par son très humble et très obéissant serviteur, Charpentier.

6706

Photo Deutsches Museum, Munich

Fig. 3. Etching by Charpentier from 1775, showing the burning device with a great liquid lens built in 1774 to order of the Académie Royale des Sciences, Paris, for the investigations of Lavoisier *et al.* The lens consisted of two spherically curved cast-glass plates which fitted closely together, the space between them (4 feet in diameter, more than 6 inches thick in the middle) being filled with alcohol, and later with turpentine. (From: Oeuvres de Lavoisier, Vol. III, Paris 1865.)

The etching reproduced in *fig. 2*, after a painting by Corbould from 1805, gives a nice idea of what the experiment described may have looked like. At that time the science of chemistry was on the threshold of a new era, the phlogiston theory of combustion

[8]) See Abbé Rozier, Observations sur la Physique, sur l'Histoire Naturelle et sur les Arts, Vol. 2 (1772), pp. 108-111 ("Résultat de quelques expériences faites sur le diamant par MM. Macquer Cadet et Lavoisier de l'Académie Royale des Sciences").

having just been discredited. The picture shows Chemistry, allegorically represented as a beautiful maiden wearing a coronet, hesitating between the old and the new, symbolized as age and youth. The young chemist is demonstrating the preparation of oxygen in a solar furnace using a lens system. This experiment was done in a very similar way by Priestley [9]) — himself still an obstinate supporter of the phlogiston theory — in 1775.

The burning glass had beaten the concave mirrors for a while, because the mirrors were even more difficult to make, and more difficult to handle: it was found to be virtually impossible to keep the mirrors continually focussed on the object to be heated in long experiments. It was however realized that the performance of the burning glasses was limited by their chromatic and other aberrations and by the optical flaws in the great masses of glass. In order to eliminate these flaws in the glass, the Académie Royale des Sciences commissioned a big *liquid lens* (*fig. 3*) for Lavoisier *et al.* as late as 1774. But by that time the technique of grinding *parabolic* mirrors was beginning to be mastered (J. Short 1710-1768, W. F. Herschel 1738-1822 [10])), and although, as we have seen above, use was still made of the existing solar furnaces using burning glasses for a long time, the days of the picturesque big lens systems were drawing to a close.

S. GRADSTEIN *).

[9]) J. Priestley, Philosophical Transactions Vol. **65**, 1775, letters of 15th March, 1st April and 29th May. The burning glass used by Priestley is still preserved at Dickinson College, Carlisle, Pa. (U.S.A.).

[10]) An interesting alternative was developed by Buffon, who showed in 1747 that a large number of suitably arranged *flat* mirrors could be used instead of a concave mirror (Kircher had put this idea forward in 1646), and that the same experiments could be done with these as with the great burning glasses. He managed to melt a silver object at 20 feet with an arrangement of 117 mirrors. An arrangement of this type used by Buffon is preserved in the Conservatoire des Arts et Métiers, Paris. This method is again being used today, notably for a gigantic Russian installation for making use of solar energy.

*) Research Laboratories, Eindhoven.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of those papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, The Netherlands, where a limited number of reprints are available for distribution.

2870: K. Reinsma: The inherent filtration of X-ray tubes (Radiology **74**, 971-972, 1960, No. 6).

If the inherent filtration of an X-ray tube is known, it is possible to determine the extra filtration required to reach the prescribed total. For a given voltage the radiation quality of an X-ray beam, expressed in mm Al half-value layer, depends markedly on the waveform of the voltage (constant potential or full-wave rectified). Further, there is a marked difference in the equivalent inherent filtration, depending on whether it is determined from dose measurements or from the radiation quality. For the latter case, a set of curves is given showing radiation quality as a function of equivalent inherent filtration (mm Al) for various voltages from 50 to 150 kV constant potential, and 50-100 kV full-wave rectified.

2871: H. F. L. Schöler, E. H. Reerink and P. Westerhof: The progestational effect of a new series of steroids (Acta physiol. pharmacol. neerl. **9**, 134-136, 1960, No. 1).

In order to study the influence of stereochemical changes upon the pharmacological properties of steroids, a number of steroid hormone analogues have been prepared which have the same configuration of the C-9 hydrogen atom and the C-10 methyl group as present in lumisterol$_2$ (see No. **2856**). The activities of the compounds administered subcutaneously and orally at three dosage levels were compared with that of progesterone administered subcutaneously, also at three dosage levels. Photographs taken of the uterine sections were ranked according to the degree of change in the shape of the mucous-membrane epithelium.

2872: F. C. de Ronde, H. J. G. Meyer and O. W. Memelink: The *P-I-N* modulator, an electrically controlled attenuator for mm and sub-mm waves (IRE Trans. on microwave theory and techniques **MTT-8**, 325-327, 1960, No. 3).

The construction and performance of a millimetre-wave modulator are described. The main part of the modulator consists of a *P-I-N* germanium structure inserted into a rectangular waveguide. A modulation depth of 11 dB could be obtained at frequencies up to 5 kc/s, this modulation being caused for the greatest part by attenuation.

2873: H. C. Hamaker: Attribute sampling in operation (Bull. Inst. Int. Statistique **37**, 265-281, 1960, No. 2).

The features of practical interest in attribute sampling procedures are discussed and on the basis of the arguments brought forward some modifications in existing sampling tables are proposed which, it is believed, would render these tables of still greater practical value. Some of the topics considered are sample size efficiency, the AQL (acceptable quality level) concept, the relation between lot size and sample size, advantages of a constant sample size, and tightened and reduced inspection. At the end the main conclusions are summarized and a modified sampling standard is proposed as a basis for discussion.

2874: B. Okkerse: De bereiding van dislocatievrije germaniumkristallen (Ingenieur **72**, O 21-O 26, 1960, No. 29). (The preparation of dislocation-free germanium crystals; in Dutch.)

The paper gives a method for the preparation of dislocation-free germanium crystals. Dislocations in germanium crystals are generated by sources which are activated by the thermal stresses during the growth of the crystal. By decreasing the diameter of the seed crystal to about 1 mm these stresses can be reduced. Consequently the seed crystal can be made dislocation-free, and then a dislocation-free crystal may grow on this seed crystal. The relevant properties of dislocations and various techniques for detecting dislocations are reviewed. Some properties of dislocation-free germanium crystals are discussed. See also Philips tech. Rev. **21**, 340-345, 1959/60.

R 411: J. H. N. van Vucht: Ternary system Th-Ce-Al (Philips Res. Repts **16**, 1-40, 1961, No. 1).

A report of an investigation of the ternary system Th-Ce-Al, including a review of data on the binary systems Th-Al, Ce-Al and Th-Ce and on the element cerium. No ternary compounds were found in the system Th-Ce-Al. This investigation was primarily undertaken to determine the structure of "Ceto", a non-evaporating getter; this structure is described.

R 412: P. C. Newman, J. C. Brice and H. C. Wright: The phase diagram of the gallium-tellurium system (Philips Res. Repts **16**, 41-50, 1961, No. 1).

The phase diagram of the gallium-tellurium system has been investigated by differential thermal analysis and direct observation of melting points under controlled tellurium pressures. The results of these investigations, which are confirmed by X-ray analysis, show that besides the two compounds already known (GaTe and Ga_2Te_3), there exist two other compounds Ga_3Te_2 and $GaTe_3$. These two compounds, however, are not stable at room temperature. A hexagonal unit cell for $GaTe_3$ with $a = 6.43$ Å and $c = 14.20$ Å is reported. The melting points of GaTe and Ga_2Te_3 are 835 ± 2 °C at 6×10^{-2} torr Te-pressure and 792 ± 2 °C at 2 torr, respectively. Upper decomposition limits for Ga_3Te_2 and $GaTe_3$ are 753 ± 2 °C and 429 ± 2 °C.

R 413: J. D. Fast and M. B. Verrijp: Internal friction in lightly deformed pure iron wires (Philips Res. Repts **16**, 51-65, 1961, No. 1).

The internal friction (damping of free torsional vibrations) of pure (99.99%) iron wires is measured before and after they have been subjected to a very small plastic deformation. The damping after deformation is found to be strongly dependent on the amplitude of the deformation, the temperature at which it was carried out, and the temperature of measurement. Wires which are deformed at temperatures below —30 °C show a spontaneous increase of damping with time, while those deformed at higher temperatures show a spontaneous decrease. In the latter case the logarithm of the damping is found to be a linear function of t^p, where t is the time and p a constant between 0.2 and 1.0 which varies from experiment to experiment. Further experiments carried out with increased concentrations of vacancies and carbon atoms have shown that the spontaneous decrease of the damping is due to the diffusion of point defects towards dislocations, and the anchoring of the latter by the former. The spontaneous increase of the damping is probably due to the dispersion of local concentrations of dislocations.

R 414: L. Schmieder: The behaviour of the mercury high-pressure arc under mechanical vibrations (Philips Res. Repts **16**, 66-84, 1961, No. 1).

If a high-pressure mercury-vapour lamp is allowed to vibrate sinusoidally in a direction perpendicular to the axis of the discharge, the arc voltage rises. If both the frequency and the velocity amplitude of the vibration exceed certain values, the arc will be quenched. This can be explained by assuming that the mechanical vibration gives rise to forced gas flow, which results in a ccrtain loss of energy from the arc by convection. If the lamp is suddenly given a (constant) acceleration, it takes some time before the resulting Poiseuille flow pattern becomes constant. This transient time explains the existence of a critical frequency for the quenching of the discharge. Above this frequency, the gas behaves like a frictionless fluid because of its inertia. The convection losses are then proportional to the velocity amplitude of the mechanical vibration. The "mechanical stability" of the arc is defined as the velocity amplitude needed to quench the arc at frequencies above the critical frequency. Calculations show that this quantity is proportional to the diameter of the tube and inversely proportional to the mass of gas per unit length of tube. The experimental results agree quite well with these calculations.

R 415: J. A. W. van der Does de Bye: Signal-to-noise ratio of a *p-n*-junction radiation counter (Philips Res. Repts **16,** 85-95, 1961, No. 1).

X-ray quanta can be counted with the aid of a *P-N* diode of low capacitance, biased in the reverse direction. The quanta give rise to pairs of holes and electrons in the space-charge layer between the *P* and *N* regions. The charges on these particles can be completely collected and measured, thus making X-ray spectroscopy possible. This paper deals chiefly with the influence of the noise on the detection of the signal produced with the aid of a thermionic amplifier with *RC* pulse shaping. For a given choice of the circuit involved, the noise is mainly determined by the diode current I_g and the shot noise of the first amplifier tube. The quantity $C^2R_{eq}I_g$, where C is the total input capacitance and R_{eq} the equivalent noise resistance of the first amplifier tube, may be used as a quality factor for the *P-N* counter. This quality factor can be used to calculate the minimum quantum energy at which detection of X-rays is possible (6 keV) and the width of the spectral lines (3 keV), i.e. the minimum energy difference between two quanta which can still be distinguished. The experimentally determined values were 9 and 4½ keV respectively.

R 416: M. Koedam: Cathode sputtering by rare-gas ions of low energy (Philips Res. Repts **16,** 101-144, 1961, No. 2).

It is well known that a metal emits atoms when it is bombarded with gas ions; this phenomenon is known as cathode sputtering. It was first observed, more than a hundred years ago, in a gas discharge between two (cold) electrodes. This thesis (Utrecht, March 1961) describes measurements carried out with mono-energetic gas ions striking the metal surface at right angles. The energy of the ions varied between 40 and 1500 eV. Chapter I contains a summary of the most important literature. Chapter II describes the apparatus, and some measurements carried out to determine the energy distribution of the gas ions. Chapter III is devoted to the determination of the number of atoms released by the cathode sputtering. After a summary of previous methods follows a detailcd description of the method used in the present investigation: the sputtered metal is collected on a glass plate, and the amount determined from the optical transmission of glass plates plus metal layer. This chapter also contains a comparison of the structure of layers of silver (and copper) formed by sputtering and by evaporation. The experimental results are given in chapter IV. The sputtering yield (atoms/ion) was determined as a function of the energy of the gas ions (which varied between 40 and 250 eV) for polycrystalline silver. The angular distribution of the sputtered atoms was determined for monocrystalline copper bombarded with gas ions of energies up to 1500 eV. It was found that the ⟨110⟩ and ⟨100⟩ directions are preferred directions, while the angular distribution also depends on the nature and energy of the bombarding ions. The experimental results are discussed and compared with those of other authors in chapter V. The existence of preferential directions for the sputtering and the variation of the angular distribution with the ion energy are explained.

R 417: A. J. W. Duijvestijn and A. J. Dekkers: Chebyshev approximations of some transcendental functions for use in digital computing (Philips Res. Repts **16,** 145-174, 1961, No. 2).

Description of iterative methods for finding the best approximation to continuous functions in a given interval by means of a truncated polynomial or a truncated continued fraction. The article also describes direct methods for obtaining approximations to such best approximations.

R 418: B. H. Schultz: On the study of volume recombination of excess charge carriers in semiconductors with the aid of photoconductance (Philips Res. Repts **16,** 175-181, 1961, No. 2).

A method is described for the elimination of surface effects in determinations of the recombination time of holes and electrons in semiconductors by means of photoconductance measurements. Moreover, the measurements indicate directly whether the value found for the recombination time is reliable or not.

R 419: B. H. Schultz: Recombination at copper and at nickel centres in *p*-type germanium (Philips Res. Repts **16**, 182-186, 1961, No. 2).

The rate of recombination of excess holes and electrons at copper centres in germanium depends on the temperature. If the germanium also contains some antimony, which partially compensates the acceptor action of the copper, a different dependence on the temperature is found. All the experimental results can be explained by assuming that recombination occurs not only at copper ions, but also at neutral copper atoms. The contradiction between the results obtained by previous workers is hereby resolved. A similar contradiction which is also found with nickel could not be resolved. The data obtained in the present investigation agree with those of Wertheim, but not with those of Kalashnikov and Tissen.

R 420: L. J. van der Pauw: Determination of resistivity tensor and Hall tensor of anisotropic conductors (Philips Res. Repts **16**, 187-195, 1961, No. 2).

The resistivity tensor of an anisotropic conductor with respect to an arbitrarily chosen rectangular coordinate system can be described by six constants. It is shown that these six constants are related to the "sheet resistivities" of six plane-parallel samples by six linear equations. The plane-parallel samples may be of arbitrary shape and cut in arbitrary but known directions. The Hall effect can most generally be described by nine constants. For the determination of these nine constants only three such samples are required, combined however with three different orientations of the magnetic induction.

H 9: K. Böke: Kapazitätsmessungen an der Grenzfläche Silicium-Elektrolyt (Z. Naturf. **15a**, 550-551, 1960, No. 5/6). (Capacitance measurements on silicon-electrolyte interfaces; in German.)

A short report of capacitance measurements on the interface between silicon and an electrolyte. These experiments form part of a fundamental investigation of the phenomena occurring at the surface of semiconductors. See also **H 10.**

H 10: H. U. Harten: The surface recombination on silicon contacting an electrolyte (Phys. Chem. Solids **14**, 220-225, 1960).

A silicon disc of thickness equal to the recombination length for diffusion is immersed in an electrolyte. A voltage applied between the electrolyte and the sample is used to vary the charge density on the surface of the silicon. The rate of surface recombination and the surface photoelectric effect are measured as functions of the applied voltage. The results are in qualitative agreement with the theory, and are dependent on the oxidation state of the silicon.

H 11: G. Schulten and H. Severin: Dämpfungsarme Leitungen für Millimeterwellen (Nachr. techn. Fachber. **23**, 20-23, 1961). (Low-damping transmission lines for millimetre waves; in German.)

A brief survey of surface-wave transmission lines, which can have considerably lower attenuation for millimetre waves than the more common type of waveguide in which the wave is propagated in the interior of a hollow conductor.

H 12: H. Severin: Neuere Mikrowellenferrite und ihre Anwendungen (Nachr.techn. Fachber. **23**, 24-27, 1961). (New microwave ferrites and their uses; in German.)

A brief survey of ferrites which can be used at higher and lower frequencies in the microwave region than has been hitherto possible. A number of factors determining the upper and lower frequency limits for such applications are discussed.

VOLUME 23, 1961/62, No. 7 pp. 189-220 Published 19th April 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

OPTO-ELECTRONICS

by G. DIEMER *).

535:539.124

*On 6th October 1961 Dr. G. Diemer delivered an address under the above title upon his inauguration as professor extraordinary at the Technische Hogeschool, Eindhoven. The text is reproduced below more or less verbatim. In accordance with custom, the speaker made no use of such visual aids as drawings or formulae written on the blackboard or of slides. Professor Diemer's deliberate and extensive use of verbal imagery, however, more than made up for their lack. For this reason the editors refrained from supplementing the text with figures. An interlude, in which an opto-electronic experiment was demonstrated with the aid of a burning cigar, has been omitted, but we felt we should not withhold from our readers another interlude in which the speaker expatiates on the function of language and imagery. For the rest, the style of this speech will doubtless reveal the affinity of spirit between the speaker and other authors such as G. Remedi **) who have previously graced the columns of our journal.*

My aim in the following considerations is to tell you something about my particular pursuit, opto-electronics — a small domain in the territory of the technical sciences. I shall tour the ground with you and try to describe the state of its cultivation. Now and then we shall take a look at the larger surrounding fields of physics and technology so that we can see our own land in its right proportions.

The nature of our work resembles that of gardeners and nurserymen; certain shrubs are planted and tended with care, others are pruned back or even eradicated. Eagerly and patiently the cultivator seeks for new varieties, and the skilled worker feels content when the fruits of his labour are appreciated by others. Let me first try to make you acquainted with some of the more proficient cultivators that we shall meet on our tour and with a few of the more outstanding plants that embellish our garden.

That delicate plant, which has spread into the farthest corners of our estate, is called Planck's constant. It bears the name of the famous grower who, with his deep ploughing, has permanently changed the structure of our land. This species of moss, whose spores are also to be found nearly everywhere in the "hortus physicus", is known briefly as "electron". Personally I would rather it had been named after "old J.J.", who was the first to grow this species in a pure culture. Over there, in that ancient corner, you can see a couple of stout Newton oaks — still massive, although Albert with his youthful zest lopped them a bit about half a century ago when he put up that trim hedge of relativity privet. Albert left untouched, however, those tall and graceful Maxwell beeches which, majestically mature, still tower above nearly everything grown here. That row of waving Schrödinger poplars was planted only a few decades ago. They shot up quickly, and now their roots are threatening the base of Albert's hedge. Some of our soil experts are rather worried about this, but otherwise we gardeners like the shade from all these trees and we feel safe on this protected estate. As long as we are kept usefully and agreeably busy in our own garden, tending, planting and cultivating, I think we may rightly regard ours as a dignified and worthwhile occupation. Unfortunately there are some who, in their enthusiasm for the job, cannot resist the temptation to spread their plants by single-crop cultivation all over the earth's surface, where, outside their

*) Philips Research Laboratories, Eindhoven.
**) G. Remedi, Science goes to the fair, Philips tech. Rev. 17, 362-368, 1955/56.

own domain, they tend — much to the alarm of their specialized growers — to degenerate into rank weeds. Unthinkingly they often stamp with heavy boots over the carefully laid flowerbeds of others, and even I have to tread cautiously here lest I step over the bounds of my own ground.

I admit that I have sustained my metaphor longer than is customary in discourses of this kind, and I can imagine that some of you will be wondering uneasily whether you have strayed into a school of horticulture instead of an institute of technology. But what is metaphor? All language is metaphor, even that of science and technology. There is only a difference in the degree of lameness, and to escape the accusation that my metaphor no longer has any leg to stand on, I promise that from now on I shall deal in more concrete terms with the phenomena of light and colour observable in my field of study.

Opto-electronics has become established in the specialist literature of recent years as the term denoting that part of the physics and technology of the solid state which is concerned with the interaction of light and electrons in matter. More particularly we can distinguish here two groups of phenomena which are in a certain sense complementary: on the one hand there is the change that takes place in the electrical properties of some semiconductors and insulators when they are irradiated by light, and on the other there is the change in the optical properties of these substances when they are subjected to the action of electric or magnetic energy. If we disregard for a moment the third form of energy — thermal energy or heat — which plays an accompanying or fundamental part in nearly all solid-state phenomena, these two complementary groups, as defined, are distinct from two other closely related branches. These are: solid-state *electronics*, where both the energy supplied and the effects thereby produced are electric or magnetic in nature, and solid-state *spectroscopy*, where the energy supplied and the effects are both of an optical nature.

The branches mentioned are in practice — and fortunately — not rigorously distinct from one another, and it is precisely the give and take between them that often leads to the discovery of interesting new effects or to their practical application. In the case of fluorescence, for example — a purely optical effect where light phenomena are produced by optical energy — the interaction of light and electrons in the solid is an essential intermediate stage in the process. The knowledge gained of these electronic processes has quite recently made it possible to generate, in what is called a "laser", a form of fluorescence where, unlike all other forms of light production hitherto known, the light waves are coherent, that is they are emitted in a regular rhythm. The underlying principle of this device is "stimulated emission", an effect that was known more than fifty years ago from the theoretical analysis of the laws of radiation and which has found practical application during the last ten years in the "maser" for generating and ampliying microwaves.

I shall confine myself here, however, to the two complementary fields of opto-electronics which I have mentioned. Where optical radiant energy results in a variation of electrical properties I shall speak, for the sake of brevity, of an r.e. effect (*r* for radiation, *e* for electricity), and, in the alternative case, of an e.r. effect. Here too, the phenomena involved are so multifarious that it will be wise to select from each field one representative effect for further discussion.

Let me first draw your attention to the r.e. effect called photoconduction. Certain solids, for example a cadmium sulphide crystal, which by nature contain very few freely mobile electrons, exhibit when illuminated a marked increase in their electric conductivity, proportional to the number of free charge carriers produced by the irradiation. The energy that enables the electrons to break out of their bound state is supplied by the light quanta or photons incident on the crystal — hence the term photoconduction. A photoconductor, then, may be regarded as an electric-current switch having a finite internal resistance, the mechanism being operated by the incident light signal. The efficiency of the operation can be expressed in terms of power gain, i.e. the ratio of the electric power that can be switched on and off by the crystal to the radiant power that must fall on the crystal for that purpose. The power gain of cadmium sulphide may havê the surprisingly high value of many millions. Opto-electronic "straight" amplifiers can be built by combining such a sensitive r.e. element with an e.r. element. There are two ways in which this can be done.

In the (e.r. - r.e.) combination, where the coupling between the two elements is effected by radiation, the driving signal and the delivered power are both electrical, so that the whole can be regarded as an electrical amplifier. In the (r.e. - e.r.) combination the coupling is electrical and both the driving and delivered power are radiation signals. This makes it possible, therefore, to amplify radiation, and if many such combinations are stacked in the form of layers, image intensification can be achieved. However, as we shall see presently, the e.r. elements hitherto available convert electric energy into light with such poor efficiency that as far as this element alone is concerned we have to speak of attenuation instead of amplification. It is for that very reason that a high power gain in the photoconductor is of such exceptional importance for practical applications.

The requirements for achieving this high gain are: in the first place it must be possible to expose the substance to a strong electric field; next, the liberated charge carriers must have a high mobility and a long life in their free state, that is there must be a long interval of time between the moment of their liberation and their return to the bound state. As a result the switching on and off of the photoconductor is accompanied by a time lag that can never be shorter than the interval mentioned. Plainly, then, photoconductors that owe their high sensitivity to the longevity of the liberated electrons must possess an inherent response lag. In the case of cadmium sulphide the lag is of the order of 0.1 to 0.01 second; with the related cadmium selenide, switching times of about 1 millisecond are possible, but the gain here is correspondingly lower. As I have said, the poor efficiency of the e.r. elements at present available makes it necessary for the photoconductor to have a very high gain factor if the combination is to be effective. The requirements mentioned indicate the direction in which we should search for materials and preparation conditions that will supply us with sensitive and fast photoconductors. The search is important because if the switching speed can be increased the potential applications will be considerably widened.

In spite of their moderate switching speed these devices already have so many attractive aspects that it will certainly not be long before they are put to certain practical uses. One aspect is that switching by means of light signals offers opportunities of coupling and decoupling signals in ways that would be difficult to achieve in an entirely electrical circuit. This can be simply illustrated by pointing to the fact that, unlike electrical conductors, two light rays cause no short-circuit if they intersect. Another aspect is covered by the American coinage "molecular electronics": the opto-electronic circuits discussed here function by virtue of the properties of the substance itself which, to handle the signals as required, only needs to be supplied with electrodes to which a voltage is applied. Additional lumped coupling elements, such as capacitors, resistors and inductors, are not necessary, and this makes it quite possible to build a complete circuit in a very compact form by using simple layer structures on which more or less complicated electrode systems are impressed (e.g. by a printing technique). But further improvements are also to be expected in the field of e.r. elements, and I should now like to discuss their most important representative for practical purposes at the present time — the electroluminescent panel.

In a review article *) I once described the mechanism of the electroluminescent panel as follows: "The functioning of a luminous panel differs from other familiar kinds of luminescence, such as cathodoluminescence (as in television picture tubes) and photoluminescence (as in fluorescent lamps) only in the way in which the excitation of the solid substance takes place. In cathodoluminescence, excitation is brought about by bombardment with a beam of fast electrons which, in penetrating the substance, raise certain bound electrons to a higher energy state; the bound electrons are thereby rendered capable of emitting light quanta $h\nu$ (h = Planck's constant = 6.62×10^{-34} joule sec; ν = frequency of light = number of vibrations per second) on reverting to the ground state. In photoluminescence the same kind of excitation is caused to take place in the substance by its absorption of short-wave radiation, such as ultra-violet quanta, falling upon it. It seems that electroluminescence, as in a luminous panel, can best be understood by supposing that mobile electrons are present in the electroluminescent substance, that these electrons are able to acquire kinetic energy from the electric field and, when they have sufficient energy, are able to bring bound electrons in the crystal into an excited state by colliding with them (impact excitation)."

So much for the quotation. May I now assume that you know what we mean by electroluminescence? I can hardly expect it of the uninitiated among you, even though in writing this description an effort was made to use no more jargon than was necessary, and moreover the article went through the hands of a professional editor, experienced in the simple and clear exposition of scientific and technical information. I therefore propose to analyse this piece of prose and to comment on the statements it contains; in doing so I shall have the opportunity to say more about the background of our research and the methods employed.

By way of interlude I should like at this point to present some general observations on the language we use and on matters connected with it. This is a subject about which I feel strongly, and I do not want to let this chance go by without saying a few words about it in my own circle. The physicist busy with his daily work, with his thoughts, formulating ideas

*) Philips tech. Rev. 19, 1, 1957/58.

and talking to colleagues and associates, uses a language familiar to him, usually his mother tongue. In the style of the extract I have just cited, however, something emerges of the typical manner in which we tend to use language in our publications. Unlike our practice in daily discussions and contrary to what has really happened in the laboratory, our propensity in writing with our lavish use of the passive form and the "pluralis physicorum", is towards impersonal narration which is meant to suggest objectiveness. We thus humbly sacrifice our personality on the altar of "repeatability". But is language indeed merely a code system for communicating information?

It has often been said that nothing is really repeatable. Indeed, the adoption of new ideas and concepts, originated by our great authoritative researchers, is essentially *un*repeatable. They influence the way of thinking of their contemporaries and later investigators to such an extent that if only for that reason science and technology pass through a non-repetitive process of historical evolution. Newton's concept of mass, Planck's constant, Bohr's model of the atom, Einstein's theory of relativity, they will always leave their traces in our soil, however much we go on ploughing and working the same ground. Even the language of science, the vehicle, stimulant and reflection of our scientific thinking, undergoes a directed non-repetitive process of growth. And if only because the postulation of new concepts, the formation of fresh imagery, is a creative process, just as all creative work springs more from imagination, intuition and courage than from reasoning, the body of the so-called exact sciences will always contain essentially non-rational elements. The genius of the great men of science in fact lies in a combination of the qualities I have mentioned; they are not afraid, in the face of the apparently irrefutable logic of the existing, rationally ordered data of experience, to say how they see and think about things. It was not logical of Planck to refrain from letting the energy of his "packets" of radiation go to zero when he was trying to explain the experimentally determined laws of radiation; it was illogical of Bohr to postulate that an electron circling in its orbit does not radiate. In technology we find clear recognition of the value of such non-rational factors in the practice of patent grants, where a true invention is expected to be based rather on an unexpected, surprising idea than on a conclusion arrived at by logical reasoning from existing data known from experience.

But there is more to it than that. For our research, our growing up into physicists and our work as such, take place less than ever before in ivory towers — in spite of our enforced seclusion in establishments from which the outside world is rather sternly excluded. They are bound by many threads of social and other human activities to life as a whole. In learning and making use of what we have learned, in determining the direction of our research and in attempting, as part of a group, to go farther in that direction — in all these stages language is more than merely a code system for communicating information.

In the beginning was the Word. Goethe's Faust, for me in many respects *the* representative of scientific investigators, hesitates at this text and, after reflection, writes down instead, one after the other: the meaning, the power, the deed! Going back over the train of our thoughts, however, I arrive at another beginning: the *image*, the remarkable capacity of the human mind to order the chaos of its impressions and perceptions into images and concepts. On this genesis of all thought, more conjurative than rational, depends the conveyance of language from mother to child and our ability to communicate our experiences and ideas and to receive those of others. I have found that learning to recognize and understand the non-rational elements in the physical sciences and technology has been a great help in advancing my development as a scientific investigator.

I should like to begin my discussion of the extract quoted on electroluminescence by commenting on the concept of luminescence.

Luminescence is a particular form of light emission for which the necessary energy is *not* supplied by heating. Fire, the candle flame, the incandescent gas mantle, the electric filament, and also the sun and the stars are all familiar forms of light sources whose ability to emit light derives from heating. The lightning flash, the glow from marine forms of life and from fireflies, neon signs, yellow sodium lighting from street lanterns, the light from fluorescent lamps and from television screens, Čerenkov radiation and also the light from the electroluminescent panel, on the contrary, are not due to heating and are therefore examples of luminescence.

Why do we place such emphasis in this definition on whether or not the generation of the light is of thermal origin? To me the most important reason is the wish to dissociate luminescence from the coarse mode of excitation that heating essentially is. As I remarked, we have known since Planck that light is emitted in discrete energy quanta of the magnitude $h\nu$, corresponding in the range of visible radiation to electron transitions or "jumps" of 2 or 3 volts. If we heat a solid we increase the thermal agitation energy of the ensemble of its atomic nuclei and electrons. The movements of the atomic nuclei and electrons are entirely random, however, and the energy that we can supply to the electrons in this way fluctuates around an average value which increases in direct proportion to the temperature. At the temperatures which available solids can withstand without all too quickly evaporating, this average energy is a lot lower than is needed for electron transitions of 2 or 3 volts. Only a few energetic outliers of the chaotic movements are able to produce an electron transition sufficient for a visible quantum to be emitted. Most of the transitions are too small and result in infra-red or heat radiation. The direct conversion of thermal energy into light is therefore just about as efficient as if we were to hurl a piano down the stairs in order to get a certain note out of it. With luminescence our aim is to learn the art of striking the right note.

To the physicist, however, luminescence is not merely the emission of visible radiation. *Electro*-luminescence, for example, encompasses radiation in the ultra-violet and infra-red as well as in the visible spectrum. Physically these radiations are closely related: for us they differ only in frequency, which is highest in the ultra-violet and lowest in the infra-red. The ease with which we physicists generalize concepts derived from the experience of the senses lends our research a flexibility that may sometimes sadly let down those who are eager to turn our inventions to practical use. The electroluminescent panel — in its simplest form a thin layer of zinc-sulphide paint coated on a base and provided with electrode layers to which the mains voltage is applied — was originally hailed as a potentially interesting flat light-source of extremely simple construction. When we found, however, that such panels were not so very suitable for the purposes of lighting — for rather fundamental reasons which I shall presently deal with — all we had to do was to shift the wavelength range up a bit in order to find an entirely different use for them, that is in the opto-electronic circuits already mentioned. For switching electric signals with the aid of radiation quanta it is evidently of no immediate consequence whether the radiation is visible or not, and it now turns out that the near infra-red offers the best compromise in regard to the speed and sensitivity of the circuit. To the physicist it is a small matter to vary the wavelength by hardly a factor of two, but the lighting engineer may feel that he is again being sold a pup.

Here in Eindhoven, under the smoke of the lamp factories, I can hardly avoid telling you something more about the reasons why we expect this "light-source of the future" to keep that name for the time being in a strictly literal sense. It is true that the luminous panel conforms to the definition given, in that the light emission is not due to heating, but that does not alter the fact that about 99% of the electrical energy supplied is converted into heat and only 1% into light. If we wanted to light a whole room with these panels, for instance by covering the walls and ceiling with them, we should discover that our "luminous tiles" had turned our room into something like a tiled stove.

The fact that low efficiency is a typical feature of an electroluminescent panel is bound up with the mechanism of excitation, mentioned in the last sentence of the extract quoted. In introducing my subject I borrowed imagery from the horticultural sector, but that is now so far behind us that in order to explain to you the mechanism of electroluminescence I shall transplant the action of this play of electrons to a sports field. The emission of light takes place during a jumping contest between electrons. Successful jumps are celebrated by the ignition of light signals, and these nimble quanta are presented with colours denoting the height cleared. The highest jumps are rewarded with violet or blue, the lower ones successively with green, yellow and red, while awkward little hops close to the ground are dismissed with infra-red. To enjoy visible light of a certain colour, then, we have to compel the electrons to make correspondingly high jumps. The stick used for measuring the height cleared by the electrons is graduated in volts, and we recall that to produce visible light we must make the hurdles between two and three volts high. For our chemists this is not a particularly difficult chore: all they have to do is to dissolve a spot of manganese in the zinc sulphide, which has the effect of producing hurdles of the right height all over the zinc-sulphide field. It is now up to the physicist to persuade the electrons to jump well. This he does, following the Frenchman Destriau, by applying an electric potential, thus setting the whole sports field on a tilt. Now the excitement really starts. The electrons break into a trot, and a few of them, those that don't stumble too much over the bumps in the ground, get up such a speed that they are able to clear the jump and we are rejoiced with a useful light quantum. Most of the little scamps, however, were born lazy and are content to jog-trot calmly down the slope, dodging craftily between the hurdles and hopping over mole hills, supplying us only with very long-wave infra-red quanta (i.e. heat radiation). The experimenter, determined to teach the recalcitrants manners, now makes the slope steeper and steeper, but this makes them so wild that very soon they completely destroy the sports field. "The panel has broken down again," sighs the physicist. As a trainer of jumpers, then, the physicist is ultimately not a great success, and he tells the lighting engineer that, as far as luminous panels based on the Destriau effect are concerned, he will have to be content with a low efficiency.

For a more effective electron trainer we again call in the help of the chemist, hoping that one of the numerous varieties in his bulky book of recipes will provide us with a better sports field. Recollecting a luminous effect discovered by the Russian Lossev whilst playing with an old-fashioned crystal detector in the days of Rapallo, we agree to have another try, this time with a proper semiconductor instead of the "quarter-conductor" zinc sulphide. Since the semiconducting carborundum of the crystal detector is not so very easy for our chemists to handle, we take the rather less wayward gallium phosphide, whose electro-optical properties resemble in many respects those of carborundum. In a pure state gallium phosphide forms a sports field the entire surface of which is completely occupied by electrons; above their heads there extends a system of scaffolding at just so many volts above the ground that electrons jumping down are able to produce light quanta of the required colour. In pure gallium phosphide, however, this superstructure is empty. The chemist therefore sets off to build on one half of the field, let us say the east half, a large number of donor pillars. Seated on these pillars are electrons that can easily step over to the scaffolding where, egged on by the notorious agitator Brown, they can take a walk around. For the present they dare not jump down to the fully occupied ground bencath them, for the great trainer Pauli has flatly forbidden them to perch on the heads of their own kind. Our chemist now prepares an opportunity for them to jump by removing individuals from the west half of the field (to be exact he lets them take a seat on fairly low acceptor bar-stools whose legs take up hardly any room on the grass). Vacant spaces are thus created where the jumpers can land without defying the dreaded Pauli. At this point the front ranks of the scaffold-walkers on the east side show an inclination to take the plunge towards the inviting vacant spaces in the west half. Simultaneously the eastern ground electrons start to make for them too, leaving similar gaps in their own ranks near the half-way line. (Although orthodox spectators might find all this difficult to follow, we broad-minded physicists regard it as perfectly normal and understandable that the eastern linesmen should consider these gaps in their own ranks as individuals, as aliens with subversive tendencies.) Those leaving their own territory, however, are punished by Coulomb with such a nostalgia that the crossing incidents very soon come to an end: Coulomb lets the aliens on both sides feel his powers by raising the whole of the west half to a level at which the occupiers of the bar stools are at the same height as the squatters on the pillars. In this deadlock the chemist sees no glimmer of light, and the physicist has to assume the initiative. Taking a simple battery of a few volts, he connects the positive pole to the west half and the negative pole to the east half, thereby crushing Coulomb's opposition and restoring the original difference in level between the eastern scaffold-walkers and the ground in the west half. Electrons now come forward in a steady stream to risk the jump, and every jump (in this ideal case) is successful. The electrical connections between the negative and positive poles of the battery are responsible for the supply and removal of the jumpers.

The ideal, which is to convert electrical energy into light with a high efficiency, has not yet in fact been realized. Working in close cooperation, physicists and chemists have been able in recent years to raise the efficiency of the Lossev effect to roughly the same level as that of the Destriau effect. Unlike the situation with the zinc-sulphide panel, however, the excitation mechanism in the case of the gallium-phosphide diode sets no fundamental limit to the efficiency of the conversion. The main difficulty in improving the efficiency is now to improve our control of the material. The gallium-phosphide crystals hitherto available, for example, are still very inferior in their physical and chemical purity to the transistor materials, germanium and silicon. There are many arguments that might be put forward to excuse this state of affairs; I shall mention here only two. A crystal consisting of a compound of two elements like gallium and phosphorus has more degrees of freedom by which it can escape our control than a single-element transistor crystal. Further, the amount of labour devoted to the perfecting of transistor materials may certainly be estimated as at least a hundred times greater than that devoted to gallium phosphide. The fact that the Lossev effect opens up interesting perspectives in many fields of application is now increasingly recognized. The moment it is possible to raise by a factor of ten the efficiency with which visible rays can be generated with this diode "lamp", it can become the "light-source of today". It might then in principle also be possible to replace the voluminous picture tubes in television sets by a "picture on the wall".

To make predictions is difficult, even for a physicist, whose profession prides itself on being able to describe with exactitude the future course of many phenomena from a given initial situation. Whether we shall ever in fact obtain from the Lossev diode an efficiency of 100 per cent or even higher — which, incidentally, would not be in conflict with physical laws, since a diode of this kind, biased in the forward direction, is able through the Peltier effect to draw heat from its surroundings and convert it into other forms of energy — whether we shall ever achieve that is a question you cannot expect me to answer today.

What is more to the point is that a Lossev source with an efficiency of only one per cent already makes very useful opto-electronic circuits possible. The very fact that the radiation is emitted less in the visible range than in the near infra-red is, as we have seen, an advantage in this connection. Moreover, the use of a DC supply of low voltage allows much better matching to the photoconductive part of the opto-electronic circuit and to other solid-state devices, such as transistors, diodes and magnetic cores, than is possible with the Destriau panel.

We trust that our chemists will press on bravely with their efforts to perfect their equipment for preparing pure gallium phosphide, equipment which, because of its combination of high pressure and high temperature, resembles in many respects a phosphor bomb. We physicists will meanwhile try to track down the evil-doers, the killers, who still stifle 99 per cent of the potential quanta at birth.

This brings me to the end of the survey I wanted to give you of my own working territory. The poet Hölderlin expressed in the following lines, freely translated, just how difficult it is for the uninitiated to grasp the essence of a subject unfamiliar to him: "He who merely smells my plant knows it not — nor yet he who plucks it merely to learn of it". To what extent have I succeeded as a guide on this tour in leading the uninitiated among you really *into* my garden, as I intended, and not "up the garden path"? I gladly leave the answer to this question to my fellow members of the guild, who may check the adequacy of the imagery which I have used against the imagery of our professional symbols and formulae.

Summary. Principal contents of the address delivered by the author upon his inauguration as professor extraordinary at the Technische Hogeschool, Eindhoven. Opto-electronics is concerned with the interaction of light and electrons in solids, and in particular with "r.e. effects", where optical radiant energy causes a change in electrical properties, and with the alternative "e.r. effects". After discussing photoconductivity, an r.e. effect, the author describes the potential applications and properties — such as power gain and response lag — of (e.r.-r.e.) combinations, and finally deals at greater length with electroluminescence, the most important practical representative of the e.r. effects. The mechanisms of the two known forms, i.e. the Destriau effect and the Lossev effect, are explained with a somewhat parodied version of the band scheme, and it is made clear why a high luminous efficiency is possible in principle with the Lossev effect but not with the Destriau effect.

PREPARATION OF GRANULAR PLASTIC FOR GRAMOPHONE RECORDS

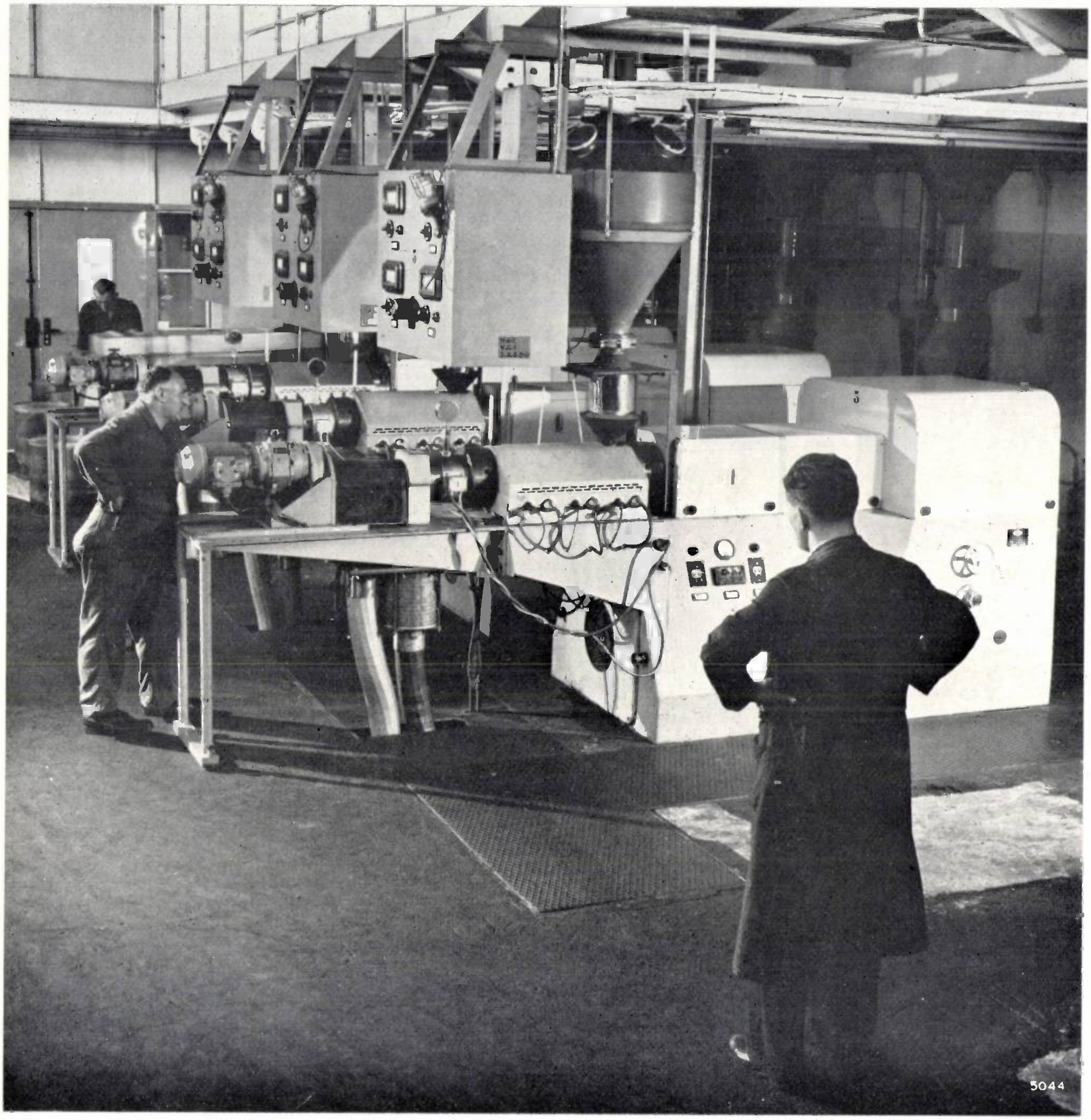


Photo Maurice Broomfield

The photograph shows part of the plant at Philips' Phonographische Industrie, Baarn, Netherlands, that converts a mixture of plastics in powder form to the granular material from which microgroove gramophone records are made. A powder mixture, prepared on the platform just visible at the top of the picture, descends through a hopper past a permanent magnet, which removes any iron particles. The mixture then enters an electric heating chamber. This is divided into three zones of different temperatures, which are kept constant by automatic controllers (in the console near the hopper). Two conveyor screws conduct the powder through the three zones. Thereafter it is compressed and compounded, i.e. kneaded into a homogeneous mass. The mass then passes to an extruding head and is extruded through a perforated plate in the form of strands about 5 mm in diameter.

At the left-hand end of the machine is the granulator. Here the still warm strands are reduced to grains by rotating cutters. The grains drop into a fast, filtered air stream, which cools them and conveys them to collecting bins. From this granular material the microgroove records are pressed [1]).

[1]) Philips tech. Rev. 17, 108, 1955/56.

A 4 MeV INDUSTRIAL RADIOGRAPHY INSTALLATION

by T. R. CHIPPENDALE *).

621.386.8:621.384.62:620.179.152

For radiography of very great thicknesses of steel a harder radiation is required than can be delivered by conventional X-ray tubes. For generating a radiation of the required hardness, the electrons bombarding the target must have an energy of several MeV. An attractive method of producing electrons of such energy is by means of a linear accelerator. This makes for a manageable installation and an intense X-ray output which permits of rather short exposure times· With the installation here described, the exposure time for the radiography of a 12" (30 cm) steel wall with 1 m focus-film distance, and using a slow, sensitive film, is less than half an hour; flaws about 0·75 mm across in the wall can then be detected.

Industrial radiography calls for a wide range of equipment capable of detecting flaws or irregularities in a considerable range of materials and specimen thicknesses. Conventional X-ray equipment operating at up to 300 or 400 kV has been commonly available for some time and has been found satisfactory in many engineering applications. However, the radiations produced at such voltages are too strongly absorbed for a useful penetration of heavier sections of materials, e.g. 7 cm or more of steel. For such a purpose, radiation of a much higher quantum energy than can be obtained with electrons accelerated to 400 keV is needed. Equipment has been designed successfully in which resonant transformers (up to 1 and even 2 MV) and Van de Graaff generators supply the voltage. More recently, radioactive isotopes have been used as radiation sources. The quantum energy of their γ radiation generally lies somewhere between 1 and 2 MeV, but with radioactive sources there is the disadvantage that the ratio between the size of the source and its intensity is unfavourable. To provide optimum radiographic definition an approximation to a point source is required; this cannot be achieved with a radioactive source of the required intensity.

For the radiography of steel above 7 cm in thickness X-rays generated by means of electrons of about 4 MeV are particularly suitable [1]). Great difficulties would be encountered in an attempt to obtain this energy of 4 MeV by making the electrons traverse a potential difference of 4 MV between cathode and target. There are, however, many kinds of particle accelerators, such as cyclotrons, betatrons, linear accelerators, etc., in which it is possible, without the application of very high voltages, to accelerate the particles to high energies. X-ray installations with betatrons supplying electrons with energies of 15 and 31 MeV have indeed been used [2]). Linear accelerators, however, also have very attractive properties for this purpose [1]). With a linear accelerator an intensely radiating X-ray focus of small dimensions can be obtained, which makes for short exposure times and high-definition radiographs.

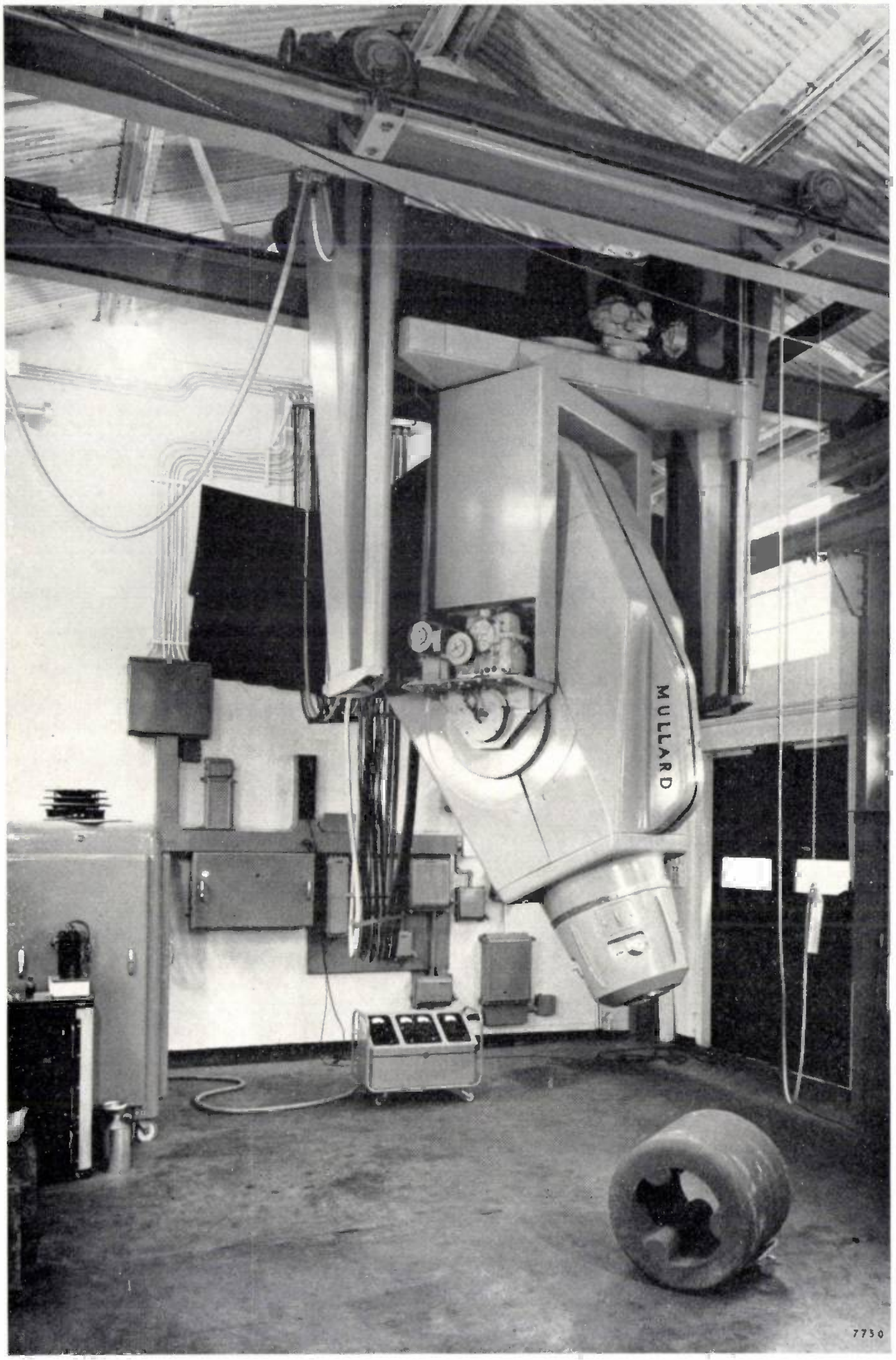
In a linear accelerator, the acceleration results from the electrons "surf-riding" an electromagnetic wave travelling in a straight accelerator tube of special construction, known as a corrugated guide. The wave is generated by a magnetron that is coupled to the corrugated guide by a system of waveguides. To reach an energy of 4 MeV a corrugated guide with a length of 1 m is sufficient. As a linear accelerator is fairly light (magnets with heavy iron cores are not required) it can be used in the construction of an easily manoeuvrable installation in which the X-ray beam can be brought quickly into the desired position with respect to the specimen. In view of the fact that the specimens are generally heavy and difficult to handle, this is very important, particularly when it is a question of routine inspections. A rapid setting up makes it possible, furthermore, to take full advantage of the short exposure times required, especially for materials which are not too thick.

An X-ray installation for the inspection of heavy metal sections has been designed and constructed at Mullard Research Laboratories on behalf of Mullard Equipment Ltd, and delivered to the British Armament Research and Develop-

*) Mullard Research Laboratories, Salfords, England.

[1]) C. W. Miller, Industrial radiography and the linear accelerator, J. Brit. Instn. Radio Engrs. **14**, 361-375, 1954.

[2]) Bau und Anwendungsmöglichkeiten von Betatron-Geräten, Stahl und Eisen **73**, 705-721, 1953; W. Lückerath, K. Fink and R. Flossmann, Durchstrahlen von heissen Vorblöcken aus Stahl mit einem Betatron und Sichtbarmachen des Durchstrahlungsbildes mit einem Röntgenbildverstärker und einer Fernseheinrichtung, Stahl und Eisen **79**, 1637-1646, 1959; W. J. Oosterkamp, J. Proper and M. C. Teves, X-ray inspection of hot steel billets during rolling, Philips tech. Rev. **21**, 281-285, 1959/60.



British crown copyright reserved

Fig. 1. General view of the 4 MeV X-ray installation constructed by Mullard Research Laboratories for the British Armament Research and Development Establishment. Apart from the accelerator and its suspension one can also see the modulator (in the cabinet with castors against the rear wall of the room) and the trolley with meters and controls of the three vacuum gauges, viz. Pirani, ionisation and Penning manometers. These components are discussed in the text,

ment Establishment. This installation incorporates a 4 MeV linear accelerator. The experience obtained previously by the firm's laboratory in the manufacture of linear accelerators [3]), in particular of various clinical X-ray installations that were also equipped with 4 MeV linear accelerators, proved of great value [4]). This equipment which, as far as the author knows, was the first to be designed specifically for industrial radiography, was installed in June 1956 (*fig. 1*). Since then it has been used regularly and has proved its value. *Fig. 2* shows the exposure times required with this installation for various thicknesses of steel.

The corrugated guide with its magnetron and associated waveguide system and an adjustable X-ray beam collimator (X-ray head) are incorporated in a single unit — the accelerator assembly. This

[3]) C. F. Bareford and M. G. Kelliher, The 15 million electron-volt linear electron accelerator for Harwell, Philips tech. Rev. **15**, 1-26, 1953/54.

[4]) An illustration with a short description of a clinical X-ray installation equipped with a 4 MeV linear accelerator was published in Philips tech. Rev. **17**, 31, 1955/56; a more complete description will be found in T. R. Chippendale and M. G. Kelliher, A linear accelerator for X-ray therapy, Discovery **15**, 397-404, 1954.

accelerator assembly is mounted in a suspension system that makes possible the requisite rapid and accurate setting up with respect to the specimen. Other main components are the modulator that supplies the accelerator assembly with 50 kV pulses, a trolley containing the meters required for the vacuum system (all shown in fig. 1), and the control desk and the power supply units (*fig. 3*) which are placed in a separate room protected against radiation. In *fig. 4* the relationship between the main components of the equipment is shown schematically.

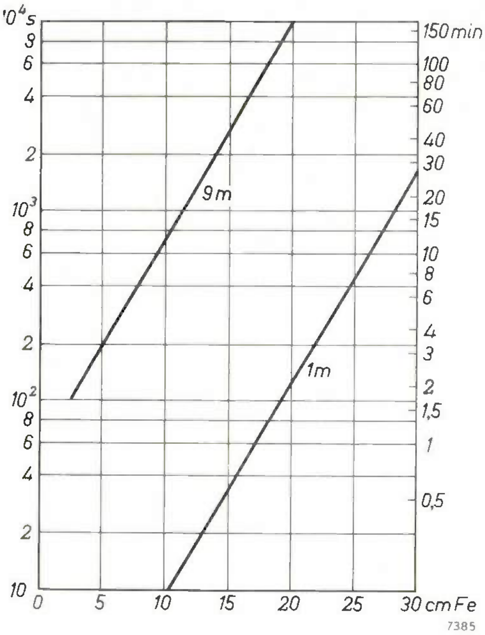


Fig. 2. Exposure time (in seconds and minutes) as a function of the thickness (in cm) of steel for the installation depicted in fig. 1 at target-film distances of 1 m and 9 m. Film used: "Ilford" Industrial F. Density: 2.5. Development 8 min at 20 °C in PQX 1. Lead screens: 0.1 mm front and 0.25 mm back. "Ilford" Industrial F is a slow, high-resolution film; for the faster "Ilford" Industrial C film the exposures should be multiplied by 0.7.

The present paper deals with the suspension of the accelerator assembly, the adjustable X-ray head, the vacuum system, the focussing of the electron beam on the target and the X-ray output, and it presents some information on attainable results. Next some features of the electron accelerator equipment are discussed, viz. particulars about the modulator, the electron gun, the rectangular waveguide system, and the corrugated guide. At thc end of the article there is brief mention of a more recent installation, equipped with a similar accelerator, which is fully transportable.

Suspension

The suspension (*fig. 5*) of the accelerator assembly is reminiscent of a travelling crane. On its front the accelerator housing carries the adjustable X-ray head, from which the X-ray beam emerges. The accelerator has five independent electrically motorised movements.

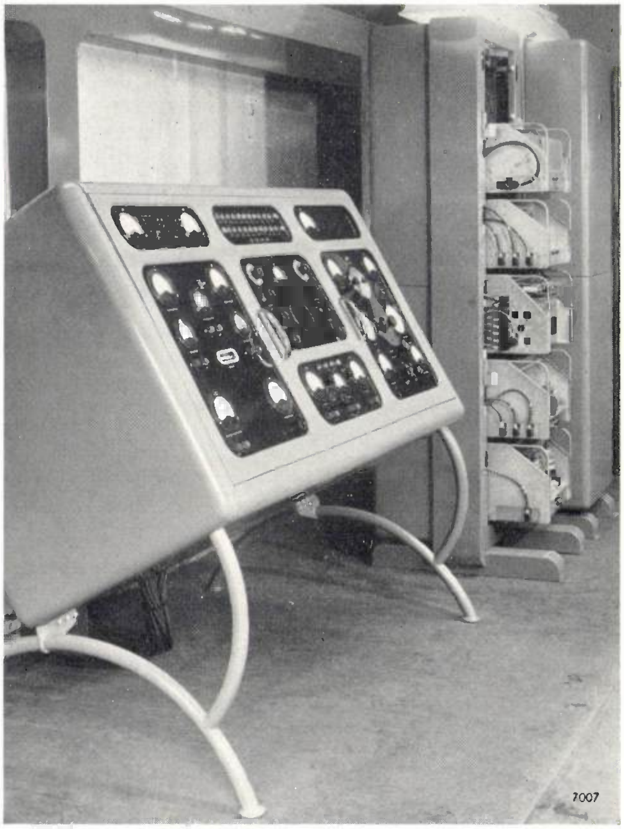


Fig. 3. Control desk. In background, two cabinets housing the power supplies, one opened. The room housing these units of the X-ray installation is shielded against radiation.

First of all the accelerator assembly is slung from two horizontal stub axles in a U-frame *3* and can thus tilt from 5° above the horizontal to 5° beyond the downward vertical. To minimise the mechanical requirements of the motor and transmission, accommodated in the U-frame (see fig. 1), the accelerator assembly with its X-ray head and all connecting cables and hoses are balanced about the stub axles. The stub axle *1* visible in fig. 5 contains the vacuum line that connects the corrugated guide to the oil-diffusion pump. A special rotatable vacuum-tight coupling in

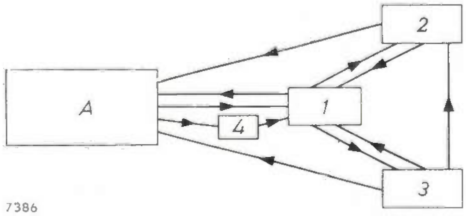


Fig. 4. Schematic diagram of the installation. *A* accelerator. *1* control desk. *2* modulator. *3* power supplies. *4* vacuum gauge trolley.

this line (to be discussed presently) permits the pump to remain vertical when the accelerator assembly is tilted.

Secondly the U-frame is suspended from the lift tray *7* by a short hollow shaft 30 cm in diameter. Thus the accelerator assembly can rotatc through 45° to either side of the central position.

Thirdly the lift tray can be moved up and down over a distance of 2.5 m by means of fixed nuts, on two diagonally opposed corners, which pass over vertical lead-screws *2*. These $1\frac{1}{2}''$ lead-screws are mounted on the drop members *4* of the suspension frame forming the traversing trolley. The whole is constructed of 6 mm steel sheet in internally webbed box formation. The lead-screws are driven via a worm gear by a motor mounted on the traversing trolley. Two hollow 10 cm guide shafts *5*, and 45 cm long bushes *6*, on the other diagonal of the traversing trolley ensure smooth movement.

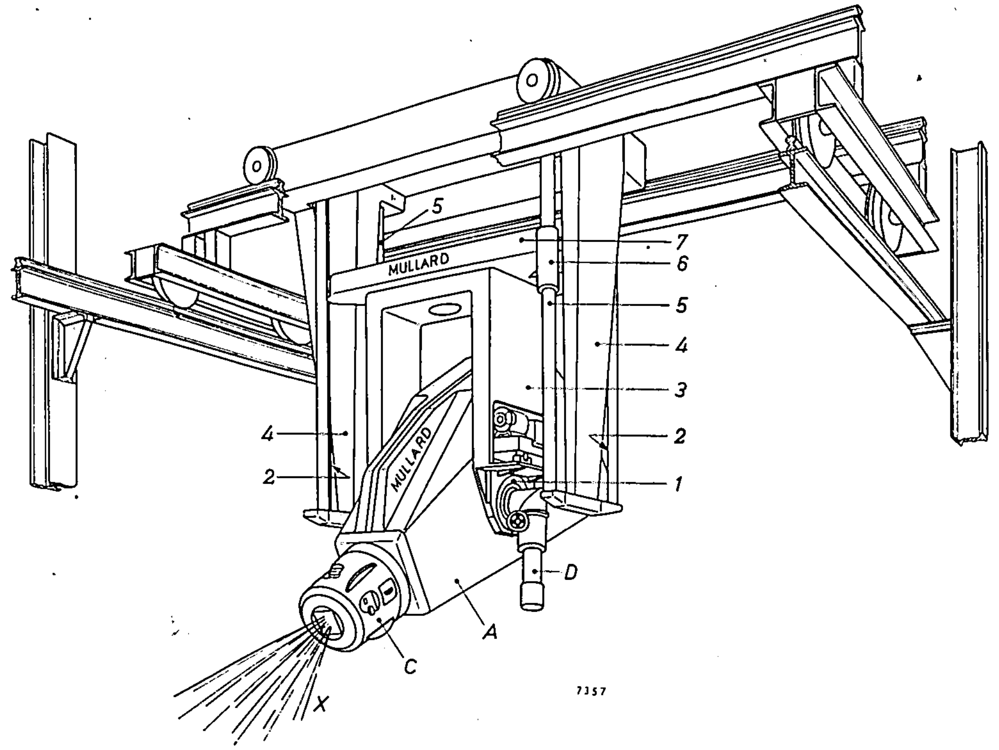


Fig. 5. Diagram of the suspension of the accelerator assembly. *X* X-ray beam. *C* X-ray head providing an adjustable rectangular radiation field. *A* accelerator. *D* oil-diffusion pump. *1* one of the stub axles about which the accelerator assembly can be tilted. *2* vertical lead-screws. *3* U-frame. *4* vertical drop members. *5* guide shafts. *6* guide bush. *7* lift tray. Speeds of the five motorised movements: tilt and swivel, both 2°/sec; vertical displacement, about 1.2 m/min; both horizontal displacements, about 1.8 m/min. The X-ray head can be rotated manually about the beam axis through 360° (see fig. 7).

Fourthly the traversing trolley runs on rails that are mounted on I-section girders; these connect the gantry trolleys.

Fifthly the gantry travels the length of the room on rails mounted on stanchions. In the installation here discussed the distances over which gantry and traversing trolley can ride are limited by the construction of the building to about 2.5 m. All movements are provided with limit switches and mechanical safety stops.

The contactors for the motors operate for safety at 24 volts. The gantry control switches are carried in a small unit suspended from the traversing trolley. To bring the accelerator assembly into the desired position, only one hand is required. The weight of the gantry and all the components it supports is about $3\frac{1}{2}$ tons (3400 kg). Of this weight the X-ray head accounts for about 360 kg, while the accelerator housing and its contents contribute about 720 kg.

X-ray head

When fast electrons are intercepted by a target, X-rays are produced which tend increasingly to concentrate in the forward direction the higher the energy of the electrons. For this reason X-ray installations for very hard radiation have a transmission target (*fig. 6*) instead of a target such as is used in conventional X-ray tubes. The transmission target must be of such a thickness that the X-ray output is at its maximum, i.e. it must be thick enough to intercept all electrons but not so thick that the output is reduced unduly as a result of absorption of the X-rays. There remains, however, considerable radiation both laterally and towards the rear. An X-ray head has therefore been developed that ensures that the radiation intensity in all directions but that of the useful beam is attenuated to an arbitrary level of less than 0.1% of that along the axis of the beam.

Fig. 7 shows the construction of the X-ray head, which is basically similar to that used in the above-mentioned clinical version from which it was developed. The transmission target is fixed at the end of the target snout, an extension of the vacuum envelope of the corrugated guide that projects some way into a lead block *4*, fixed rigidly to the accelerator. Close to the target there is an additional cylinder *3* of "heavy alloy", a strongly absorbing tungsten-copper alloy (density = 17 g/cm^3). The blocks *3* and *4* contain a conical hollow for the X-ray beam. The cone defined by the hollow of the blocks has, at a distance of 1 m from the target, a diameter of 26 cm, and halfway through the target, which has a thickness of 3 mm, a diameter of 2 mm. (The focus of the electron beam *on* the target has a diameter of less than 2 mm.) There are, however, still a fair number of stray electrons in the corrugated guide that impinge on the target outside the X-ray focus. It is in order to absorb the X-ray radiation produced by these stray electrons, and thus to maintain the quality of the focus, that the tungsten-copper block *3* is used.

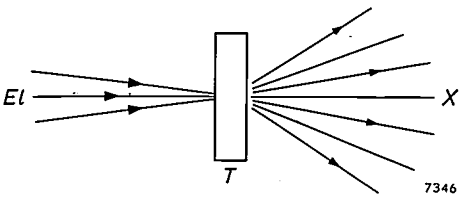


Fig. 6. Transmission target. If use is made of the X-ray radiation *X* emitted by a target *T* in the direction of the incident electron beam *El*, it is customary to speak of a transmission target. If the energy of the electrons is very high, e.g. 4 MeV, most of the X-rays are emitted in this direction.

At the lower end of the fixed collimator are mounted three ionisation chambers, one after the other. One of these provides a measure of the dose

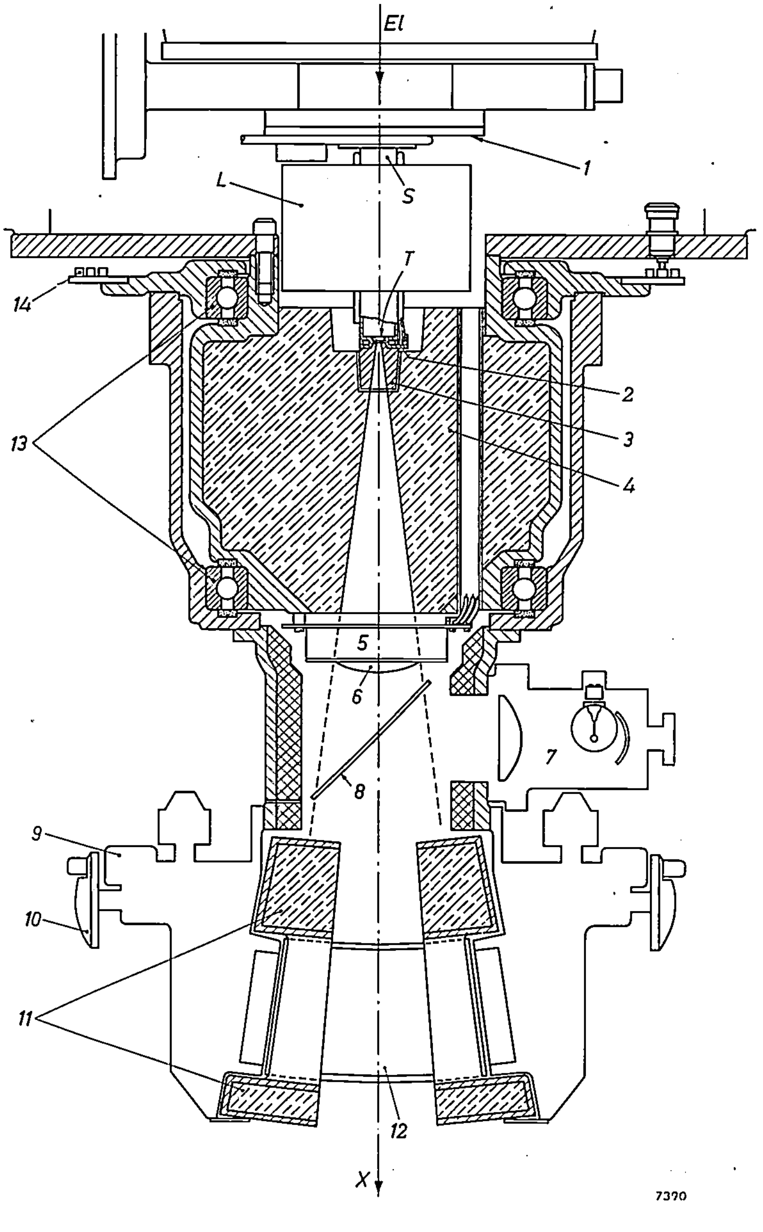


Fig. 7. Diagram of the X-ray head. *El* electrons from the accelerator. *1* end face of vacuum envelope. *S* target snout carrying at its end a 3 mm tungsten target *T*, set in the water-cooled copper block *2*. *L* magnetic quadrupole lenses focussing the electrons upon the target (cf. fig. 12). *3* block of heavy tungsten-copper alloy. *4* lead collimating block. *5* three ionization chambers for the measurement and the control of the radiation. *6* flattening filter (rarely used). *7* optical system used with mirror *8* to provide a parallel light beam for the positioning of the accelerator with respect to the object. *9* digital counters indicating the dimensions of the irradiated field at 1 m from the target *T*. *10* handwheels for the adjustment of the lead screens (jaws) *11* and *12*, that provide an adjustable rectangular diaphragm. *X* X-ray beam. *13* ball races enabling the rectangular diaphragm to be rotated. *14* slip-rings for electrical connections.

rate, a second gives the integral dose, while the third is part of a device to be described later for the automatic centring of the electron beam. Just below the ionisation chambers there is space for a correction filter for flattening the X-ray field distribution. It was found, however, that such a beam-flattening filter (frequently used in medical applications) is hardly ever necessary in industrial radiography. The reason will be discussed presently.

The X-ray cone defined by blocks *3* and *4* (fig. 7) is further limited to a beam of rectangular section by two pairs of lead screens. Each of these screens is 10 cm thick in all; in one of the pairs this total is achieved in two stages for constructional reasons. By means of handwheels each pair may be opened or closed like the jaws of a pair of pliers about an axis through the focus. In this way the inner face of the diaphragm always coincides with the outer edge of the required X-ray beam, thus minimising scatter. By the side of each handwheel there is a digital indication of the dimensions (in mm) of the associated field at a distance of 1 m from the focus (cf. fig. 8). This field is adjustable between the limits of 3×3 cm and 15×25 cm. Because of experience since obtained, the maximum field in the new version of the installation has been increased to 25×30 cm.

The lower part of the X-ray head, and thus the rectangular beam collimators, can be rotated about the beam axis.

To permit of ready positioning of the accelerator a source of light is incorporated which, via a lens and mirror system, projects a parallel beam of light along the axis of the X-ray beam. Alternatively the system could be so arranged that the light source appears to be situated at the focus. The light beam would then illuminate exactly the field covered by the X-ray beam. At normal levels of room illumination this system only works well at small focus-specimen distances unless use can be made of reflective devices (e.g. adhesive tape with lenticular reflecting surfaces) that can be affixed to the specimen.

Vacuum system

The vacuum in the corrugated guide (pressure about 3×10^{-6} mm Hg) is maintained by continuous pumping with a 10 cm oil-diffusion pump (*fig. 8*), which is combined with a baffle valve. The pumping speed above the baffle valve is 150 l/sec. A rotary backing pump is mounted on the U-frame of the suspension. A fork (*10* in *fig. 9*) fixed to the U-frame holds a pin projecting from the baffle valve and so keeps the diffusion pump vertical as the accelerator assembly is tilted. To make tilting possible, the

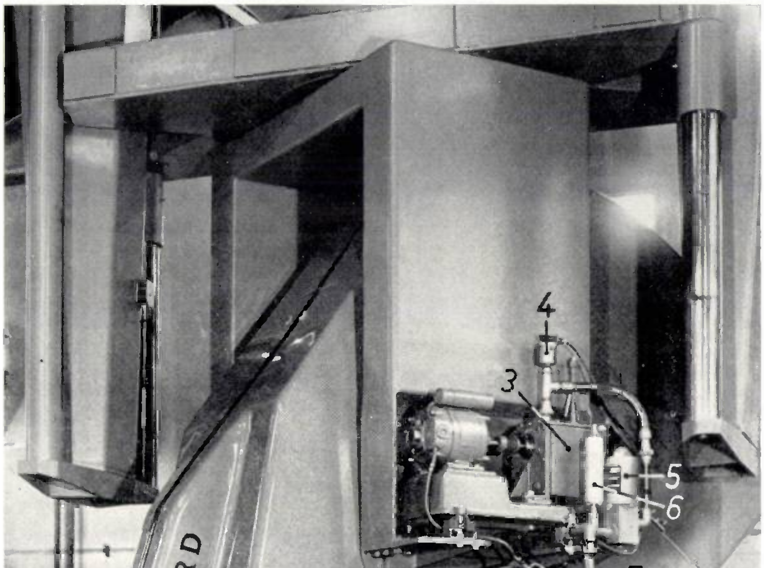


Fig. 8. Arrangement of the vacuum pumps. *D* oil-diffusion pump. *1* baffle valve. *2* ionisation gauge for the high vacuum. *VJ* rotatable vacuum-tight joint (cf. fig. 10) enabling the pump *D* to remain vertical when the accelerator is tilted. *3* backing pump. *4* and *5* magnetic safety valves. *6* Pirani manometer for the backing vacuum. *7* rigid connection between *D* and *3*. *B* backing line to electron gun (discussed on p. 203). *PJ* rotatable joint in *B* (cf. fig. 10). *8* Penning manometer operating a vacuum safety interlock device. *9* control panel.

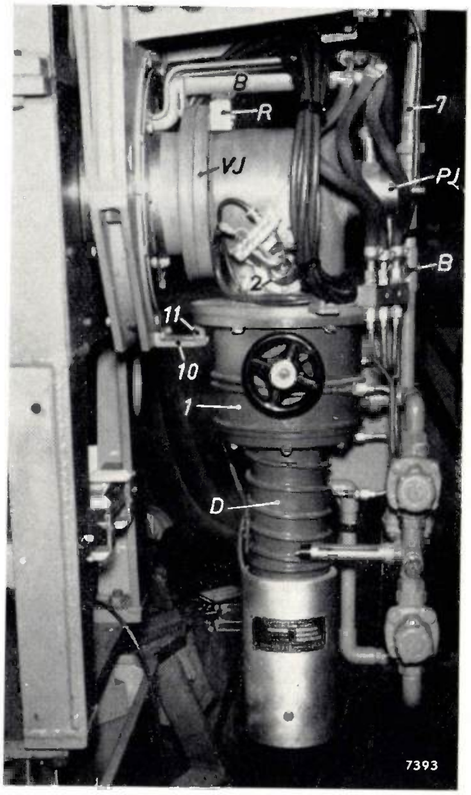


Fig. 9. Detail of the vacuum pump installation showing the fork *10* and the pin *11* that hold the pump *D* in a vertical position when the accelerator assembly is tilted. *R* is a reservoir with vacuum oil for the rotatable joint *VJ* (cf. fig. 10). The other characters have the same significance as in fig. 8.

15 cm vacuum line incorporates a rotating joint. Its construction is explained in *fig. 10*. Joints of this type have been operated successfully in lines of up to 20 cm diameter for clinical accelerators. As there is no relative movement between backing pump and diffusion pump, the line connecting them is rigid. A small rotating joint is however mounted in a branch *B* of the backing line leading to the electron gun of the accelerator. This branch is used if the cathode of the gun has to be replaced, as will be discussed presently. The construction of this small joint is also shown in fig. 10. The control panel for the vacuum pumps is mounted alongside the diffusion pump. Magnetic valves protect the diffusion pump against mains failure. The backing pressure is measured by means of a Pirani gauge, while the high vacuum is measured by means of a triode ionisation gauge. A Penning gauge is incorporated for the operation of a relay circuit that actuates a safety system if the pressure becomes too high. The controls and meters of these three vacuum gauges are accommodated on a trolley (fig. 1) that is connected to the gauges by means of a long flexible cable. The whole arrangement is rather similar to that of the 15 MeV linear accelerator described previously in this journal [3]).

Beam focussing and target

Fig. 11 shows the electron distribution in the practically parallel beam supplied by the accelerator. It will be seen that over 90% of the electrons fall on the target within a circle of 1 cm diameter. For many applications of the accelerator such a relatively large beam size is not a disadvantage, for example if the accelerator is part of a neutron generator, or if the

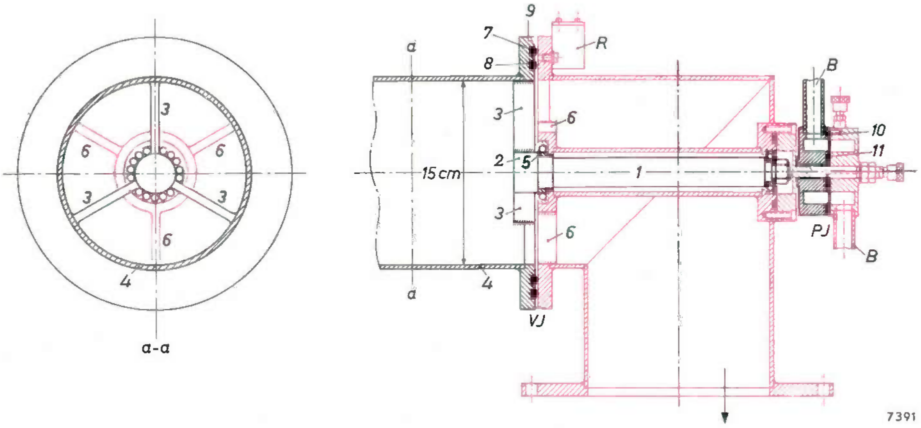


Fig. 10. Construction of the rotatable joints *VJ* and *PJ* (figs 8 and 9) in the 15-cm high-vacuum line and in the backing line *B* to the electron gun. The parts printed in black turn with the accelerator when the latter is tilted. The elbow-piece, printed in red, under which the oil-diffusion pump is fixed, remains stationary. The red part on the right is fixed to the elbow-piece. The thicker part *2* of the stub axle *1* is held, by means of spokes *3*, in the high-vacuum line *4* to the accelerator. The ball race *5* fits in a mount held in the elbow piece by means of spokes *6*. The two concentric O-rings *7* and *8*, with vacuum-oil lubricant in between, form a vacuum-tight seal. The rings bear on the polished chromium-plated flange *9* of the elbow-piece. The force with which the rings are compressed is adjusted by means of distance pieces at the ball races. *R* is a reservoir containing vacuum oil.

In the case of the coupling *PJ* in the backing line *B*, sealing is performed by means of the O-rings *10* and *11*, smeared with vacuum grease.

electron beam itself is to be used for irradiation. A strong concentration of the beam is indeed unfavourable in the latter case, because this would only lead to a high local load on the output window and would make its construction unnecessarily difficult. For radiographic purposes, however, a smaller focal spot is required. In clinical linear accelerators a focal spot size of 5 mm is considered small enough for limiting the penumbra, i.e. to produce a sharp enough cut-off at the edge of the treatment area.

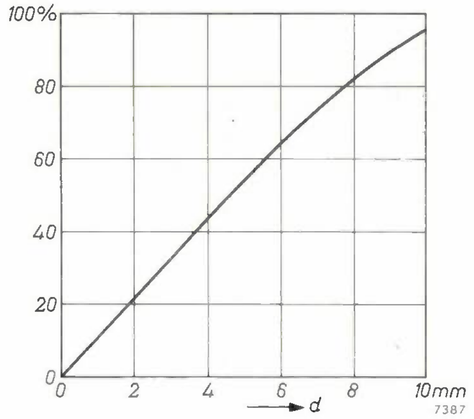


Fig. 11. Percentage of the electrons in the beam that fall on the target within a circle of diameter *d*.

For industrial radiography a focal spot not larger than 2 mm is desirable, viz. sufficiently small to reduce geometrical unsharpness in the radiographic image to the same order as that produced by other sources of unsharpness. A focus of 5 mm can be obtained by means of a system of stops in either the electron beam or in the X-ray beam. The argument has been that for clinical work the loss of about 40% of the output (see fig. 11) is of no consequence, as the remainder is so large that exposure times are still very short for the usual therapeutic doses. If, however, this method of collimation were to be extended to a focal spot of 2 mm diameter, fig. 11 shows that 80% of the output would immediately be wasted. For industrial radiography, where one is always fighting long exposure times, this is not acceptable.

Magnetic lenses were therefore fitted on the target snout. As only little space was available, we have made use of a system of two magnetic quadrupole lenses (*fig. 12*). Such a system, first described in 1952 by Courant *et al.* [5]), requires less space and power than a conventional axially symmetrical magnetic lens.

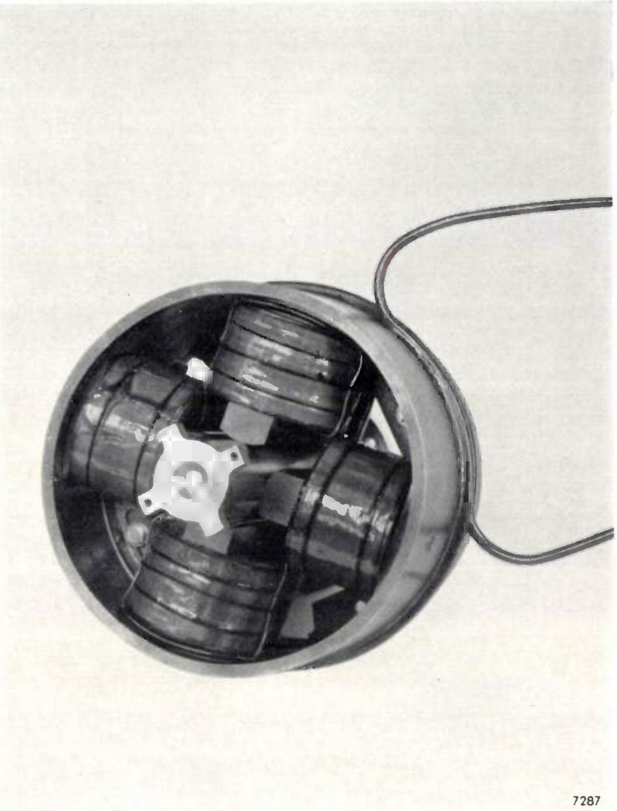


Fig. 12. Magnetic quadrupole lenses on the target snout of the accelerator. The illustration shows the four poles of the front lens. The second lens is immediately behind.

The performance of the lenses was investigated in a number of ways. An accurate measurement of the high-power high-energy electron distribution proved by no means easy. In the first place an extension of the method described by Bareford and Kelliher [3]) was attempted. The measuring device shown in *fig. 13* was used. The measured 90% transmission

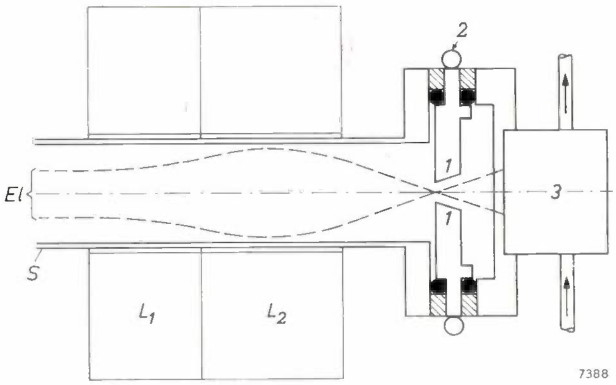


Fig. 13. Set-up for the calorimetric determination of the percentage of the electron beam *El* concentrated by the magnetic quadrupole lenses L_1 and L_2 inside a 2 mm diameter circle. The copper disc *1*, thermally insulated and mounted at the position of the target, is provided with a 2 mm hole. From the temperature rise of both the water in the cooling circuit *2* of disc *1* and that in the calorimeter *3*, the required percentage can be deduced. O-rings provide vacuum sealing.

[5]) E. D. Courant, M. S. Livingston and H. S. Snyder, The strong focusing synchrotron. A new high energy accelerator, Phys. Rev. **88**, 1190-1196, 1952; see also: R. M. Sternheimer, Double focusing of charged particles by a system of two magnets with non-uniform fields, Rev. sci. Instr. **24**, 573-585, 1953.

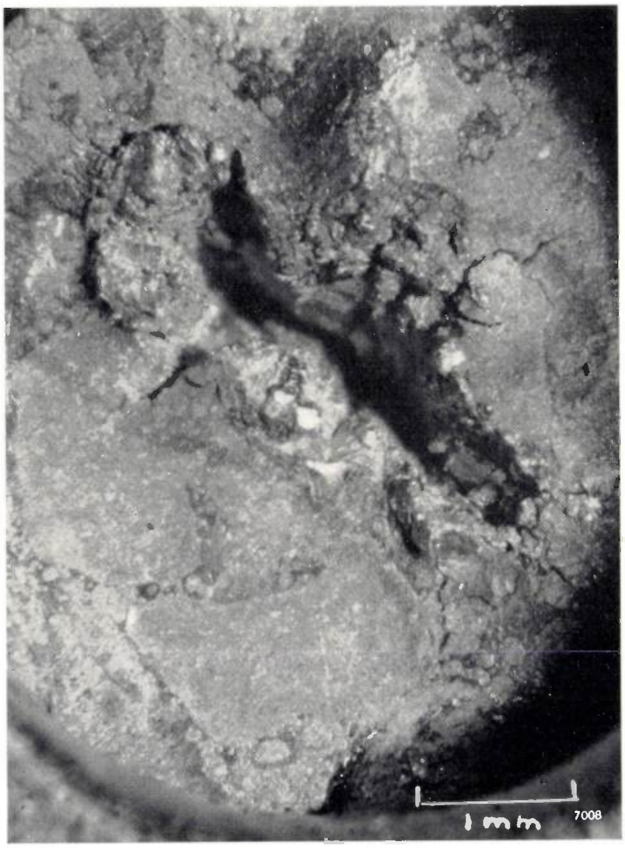


Fig. 14. Effect of the focussed electron beam on a gold target after exposure for about 1.5 hours.

through the 2 mm hole in the copper plate replacing the target is, however, probably erroneous owing to an unknown amount of reflected scatter from the innermost sections of the calorimeter assembly. We estimate that in fact about 95% came through.

In a second method a piece of photographic paper was placed at the focus. From this it was found that the focus was elliptical and certainly not greater than 2×1 mm. The most accurate estimate was, however, obtained from an examination of a gold target exposed for about 1·5 hours to the electron beam (*fig. 14*). From the shape of the hole at a depth of 2 mm, convincing proof was obtained that the dimensions of the focus were 1.5×0.5 mm. The larger dimensions on the surface are the result of an initially incorrect adjustment of the lenses.

In a quadrupole lens the north poles and south poles alternate, so that similar poles face each other diametrically. The optical equivalent of a quadrupole lens is a lens pair consisting of two mutually crossed cylindrical lenses of equal focal length, one positive and the other negative. The system consists of two such lens pairs L_1 and L_2, with focal lengths $\pm f_1$ and $\pm f_2$. L_1 and L_2 are mutually rotated by 90°, so that *fig. 15* represents the optical equivalent of the complete system. Rays of light such as l', parallel to the axis and in the vertical plane of symmetry, are therefore subjected first to a converging and then to a diverging effect. In the case of the rays like l'' in the plane of horizontal symmetry it is just the other way round. Rays l' and l'' at the same time represent the projections of the ray l on the planes of symmetry. Thus the path of an arbitrary ray parallel to the axis is also readily conceived. By means of the elementary lens formula, applied for both planes of symmetry, it is a simple matter to deduce that the rays l' and l'' intersect the axis in the same point F, i.e. that the foci in the two planes of symmetry coincide, provided that

$$f_1^2 - f_1 f_2 = d^2,$$

where d is the distance between L_1 and L_2. Any arbitrary ray parallel to the axis will now also pass through F, so that the beam is focussed in this point.

The condition of coincident foci does not yet completely determine the magnitudes of f_1, f_2 and d. There is still a certain degree of freedom in the design, which in this case has been used for accommodating the lens system in the limited available space (in the X-ray head a cylindrical space of 15 cm diameter and 15 cm length was available).

To make the focus F lie exactly in the target, it is necessary to make the power of the lenses adjustable in such a way that the condition of coincident foci is maintained. This is achieved by fine trimming of the current in the two lenses by means of mechanically coupled variable resistors.

Even if the two foci coincide, the system of fig. 15 still produces distorted images because of the fact that the magnification in the vertical plane differs from that in the horizontal plane. If the incident beam is exactly parallel, this is not noticeable, but in our case the beam contains rays making angles of up to about 0.005 radians with the axis. As a result the focal spot takes on an elliptical form. A spread in the electron energy of ± 0.25 MeV ("chromatic aberration") also contributes to this effect. From a theoretical estimate of the effect of the various factors on the sharpness of the focus it follows that the two above-mentioned factors are the most important, and that the focus can be expected to be not larger than 2×1 mm. As stated above, it has proved to be 1.5×0.5 mm from the experiment with the gold target.

It will be clear from fig. 14 that gold, preferred as a target material for ease of soldering and its high thermal conductivity, cannot withstand the heavy specific loading of over 1000 W/mm² (850 W on the

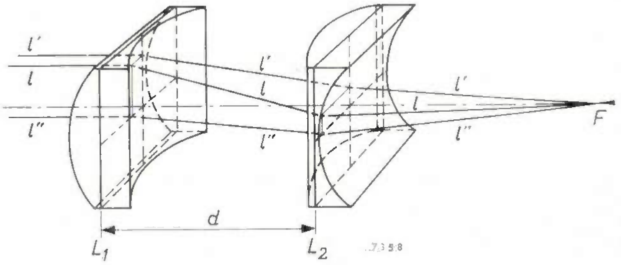


Fig. 15. Optical equivalent of the two magnetic quadrupole lenses. *Each* lens corresponds to a positive and a negative cylindrical lens (of identical focal lengths) in a mutually crossed configuration; the two lenses are also mutually crossed. A beam of light incident in a direction parallel to the axis is concentrated in a point F, provided that a certain relationship (see text) is maintained between the distance d and the focal lengths associated with L_1 and L_2.

1.5 × 0.5 mm focus). The target actually used therefore consists of a tungsten disc of 5 mm diameter and 3 mm thickness. This disc is mounted in a copper block in which an annular canal for cooling water is formed as close to the tungsten as is permitted by mechanical reliability. A 0.1 mm layer of the copper block is retained on the atmospheric side of the target to form a continuous vacuum enclosure, as large-diameter tungsten rod (from which the target is cut) can exhibit porosity along the length of the rod.

An average loading of 1000 W/mm^2 of the focus is several times the usual loading in conventional X-ray tubes. That the tungsten target can nevertheless withstand this load is due to the higher electron energy, giving a *volume* loading rather than a purely *surface* loading.

It is found that the position of the focus on the target depends to some extent on circumstances, e.g. the beam current and the high-frequency power. As it is necessary for the focus to be exactly behind the opening in the collimator block *3* (fig. 7), one of the three ionisation chambers *5* incorporated in the X-ray head is divided into four quadrants. Each quadrant supplies a separate current, and these currents (via amplifiers) control the currents energizing a pair of mutually crossed deflection coils that keep the electron-beam focus exactly aligned with the collimator aperture.

X-ray output

In any given linear accelerator the maximum X-ray output is governed by the maximum permissible duty cycle of the radio-frequency source. With the VX 4061 magnetron (2999.5 Mc/s) used, this maximum is fixed at a value of 0.001, corresponding to a pulse length of 2 μsec at a repetition frequency of 500 pulses per second. In practice, X-rays are only emitted for a part (1.8 μsec) of each RF pulse because of the filling time of the system. The repetition frequency can be selected at 100, 200, 300, 400 or 500 per second, which provides a directly proportional step-wise control of the X-ray output. Fine control, normally only used for tuning to maximum output at the selected repetition frequency, may be achieved by adjusting the electron-gun filament current.

It is common practice to characterise an X-ray installation by giving the *dose rate* in r/min (röntgens per minute) on the axis of the X-ray beam at a distance of 1 m from the target. This quantity is more readily measurable than the *intensity* (watt/cm^2). With only the heavy alloy block *3* (fig. 7) in position, and without the magnetic lens, the above-mentioned dose rate amounts to 750 r/min for the installation here described. With the lens operating, this figure is reduced to 600 r/min. This does not imply the loss of useful X-ray photons. When the lens is used, the electron beam is made to converge on the target, which results in a more divergent X-ray beam. This is an advantage as the available radiation is now distributed more uniformly over the irradiated field (*fig. 16*). It will in fact hardly ever be necessary to insert a special flattening filter (*6* in fig. 7) into the X-ray beam: it is found that with the cone angle chosen, the radial change in attenuation due to both the change in effective thickness traversed by the X-rays and the non-flat polar diagram still gives a tolerable range of film densities across the full field.

In fig. 15 it is seen that the rays show a different convergence towards the focus in the two mutually perpendicular planes of symmetry. Fig. 16 shows how this difference also manifests itself in the distribution of the dose rate over the irradiated field.

Performance

Fig. 17 gives an idea of the size of inclusions that can be detected in steel with the installation described in this article. These results are obtained using "Ilford" Industrial C film. On average they are very similar to the results obtained by Möller and Weeber using betatrons [6]. However, the X-ray output of the linear accelerator is so large that it is possible to take advantage of the high resolution of slow, ultra-fine grain films (e.g. "Ilford" Industrial F) without the necessity for uneconomically long exposure times. The resolution is thereby increased [7]. For example, the size of inclusions which can be detected in a 30-cm steel wall is about 0·9 mm with Industrial C and about 0·75 mm with Industrial F. Data on the exposure times for "Ilford" Industrial F have already been given in fig. 2. We see from this figure that e.g. the exposure time for 30 cm of steel with a focus-film distance of 1 m is less than half an hour. The times given are those when lead intensifying screens are used, 0.25 mm thick behind the radiographic film and

[6] H. Möller, W. Grimm and H. Weeber, Arch. Eisenhüttenw. **25**, 279-291, 1954 and **26**, 603-609, 1955. Möller and Weeber (Arch. Eisenhüttenw. **32**, 107-112, 1961, No. 2) also investigated a Mullard 4.3 MeV linear accelerator of the type described here (with an elliptical focus of 1.5 × 0.5 mm, not a circular one 5 mm in diameter as erroneously assumed by them). Their results were, however, based on radiographs which were not really representative of the best performance of this equipment. It is not surprising, therefore, that they found a poorer wire perceptibility than we did.

[7] The improvement in resolution is about two wires and one wire respectively, in the case of the standardised penetrameter tests DIN Fe 2 and DIN Fe 3.

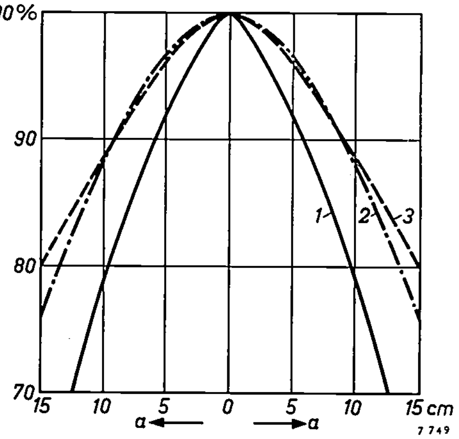


Fig. 16. Relative intensity of the X-radiation in a plane normal to the beam axis at 1 m from the target, as a function of off-axis distance *a*. These measurements were made with only the heavy alloy slug (*3* in fig. 7) in position (lead block absent). Curve *1*: without focussing by the magnetic quadrupole lenses. Curves *2* and *3*: with focussing, and measured in two directions at right angles in the two planes of symmetry of the quadrupole lenses. The difference between curves *2* and *3* reflects the unequal convergence of the electrons in the two planes of symmetry (cf. fig. 15).

0.1 mm thick in front of it. This combination gives a 2- or 3-fold intensification due to photo-electrons liberated by the X-rays in the lead foils. Intensifying screens consisting of fluorescing salts are not suitable for the very hard radiation delivered by the accelerator.

"Ilford" Industrial F film is also advantageous because no complications arise with regard to development. For some types of film, developing conditions may require adjustment in order to minimise the blotchy development characteristic of high-energy radiation. Moreover, development conditions vary the relative speeds of films. It was found, for instance, that exposure doses for any given density were in the ratio of 1.0 for "Ilford" Industrial C film, 1.75 for "Ilford" Industrial F and 6.0 for "Kodak" M when developed in ID 19 for five minutes at 20 °C, but when developed in PQX 1 for eight minutes at 20 °C the ratio was 1.0:1.5:3.9. If it is also borne in mind that a high-contrast developer such as PQX 1 tends to give better penetrameter sensitivity results, one can appreciate the dangers and difficulties of comparing performance measurements made by different workers.

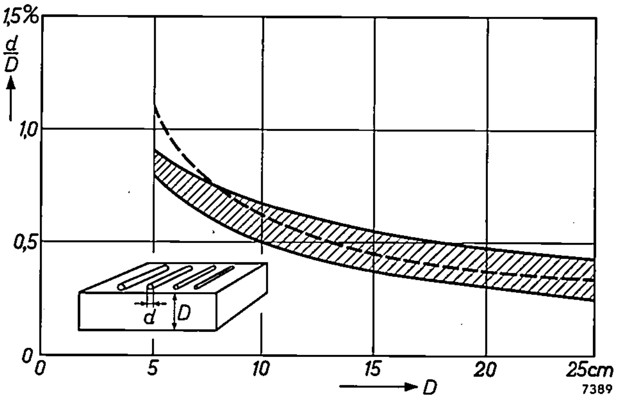


Fig. 17. Wire penetrameter sensitivity in steel for the 4.3 MeV linear accelerator, for "Ilford" Industrial C film (shaded area). This sensitivity, expressed as the percentage thickness (*d*/*D*) of the thinnest wire still just perceptible, is plotted as a function of the thickness *D* of the steel. The broken line refers to results obtained with a 31 MeV and a 15 MeV betatron installation by Möller and Weeber [6]).

As the linear accelerator works with very short pulses (1.8 μsec), it is ideally suited to stroboscopic radiography (for example, of an internal-combustion engine [8])). Facilities are provided to allow the modulator to be triggered from an external source. One may, therefore, synchronise the pulse repetition frequency of the accelerator to some rotating or reciprocating device. A relative velocity between subject and film of 100 m/sec causes a blurring of approximately 0.1 mm on the film.

There follow now a few particulars concerning the linear accelerator proper.

The modulator

The modulator (see fig. 1) supplies 50-kV 2-μsec pulses to the magnetron and simultaneously to the electron gun of the accelerator. The modulator is basically similar to those used for magnetrons in radar transmitters [9]). As stated, the repetition frequency is stepwise adjustable from 100 to 500 per second. Because of the mobile suspension of the accelerator, the modulator is connected to it by a flexible coaxial cable of 2.5 cm diameter, insulated with polythene. In comparison with the modulator reported in 1953 in this Review [3]), only two important modifications have been incorporated. Firstly, to permit of movement of the accelerator assembly, the 5 : 1 ratio output pulse transformer has been removed from the modulator cabinet and installed alongside the magnetron in the accelerator unit. The advantage of this is that the cable between the modulator and the accelerator requires to be insulated for a maximum voltage of only 10 kV instead of 50 kV. Furthermore it is better that the extra capacitance due to the long length of pulse cable is inserted on the input side of the pulse transformer rather than on the output side. The second modification has been to take advantage of the development of suitable high-power hydrogen thyratrons to replace the ignitron originally used as a switching valve. As already stated in the above-

[8]) Stroboscopic radiographs of a rotating internal-combustion engine, made with the installation here reported, are discussed by B. J. Vincent, Brit. J. appl. Phys. **11**, 132-135, 1960.

[9]) Cf., for example, Philips tech. Rev. **19**, 28, 1957/58.

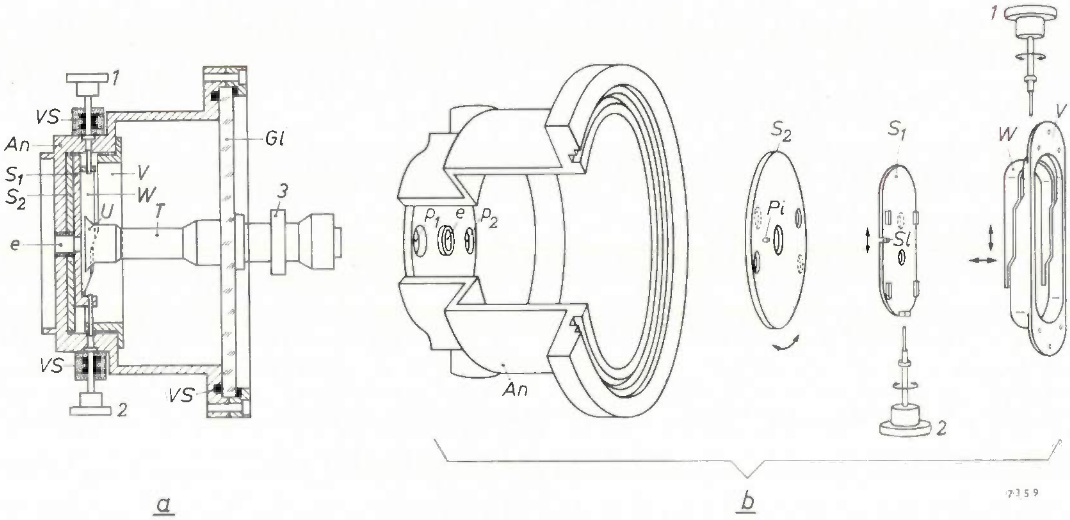


Fig. 18. *a*) Diagram and *b*) exploded view of the vacuum lock between electron gun and the main vacuum system. *U* focussing electrode, guiding the electrons issuing from the cathode (mounted within the support tube *T*) through the hole *e* in the anode *An*. p_1 and p_2 are extra pumping holes. To close *e*, p_1 and p_2, wedge *W* is unclamped by means of knob *1*. By turning knob *2*, plate S_1 is shifted, which causes plate S_2 to rotate by means of pin *Pi* and slot *Sl*. Next S_1 and S_2 are reclamped with *1*. Vacuum sealing is ensured by O-rings (not drawn). *V* clamping bridge piece. *Gl* glass plate carrying the cathode and insulating this from the (earthed) anode. Vacuum seals *VS* are provided for *1*, *2* and *Gl*. Knob *3* is used when replacing the cathode.

mentioned article [3]), normal ignitrons are not really suitable for repetition frequencies of 500 per second. The deionisation time is too long, resulting in ionic bombardment that reduces the life to a fraction of the normal. Sometimes the life amounted to less than 100 hours. The CX 1119 hydrogen thyratron now used has a life that is normal for a valve (i.e. of over 1000 hours). Another advantage is that the thyratron has a lower jitter characteristic than the ignitron. This is of importance if the installation is to be used for stroboscopic radiography.

The electron gun

The electrons to be accelerated are supplied by an electron gun, of which *fig. 18* is a diagram and *fig. 19* a photo. The anode is earthed, as are the corrugated guide and the associated waveguide system. The directly-heated cathode, consisting of a flat spiral of 0.3 mm tungsten wire (*fig. 20*), receives from the modulator the same 50 kV negative voltage pulses as the magnetron cathode. The electrons then attain a velocity of about 0.4 *c* (*c* = velocity of light). A focussing electrode directs the electron beam through a hole in the anode into the corrugated guide.

The exact moment of a cathode failure cannot be predicted and it may therefore occur during the course of making a radiograph. To avoid the necessity of admitting air into the whole evacuated system when replacing the cathode — after which it would take about an hour to re-establish the vacuum — a vacuum lock is incorporated between the electron gun and the main vacuum system. The construction of this lock (fig. 18) is simplified as compared with the previous design [3]). If replacement of the cathode

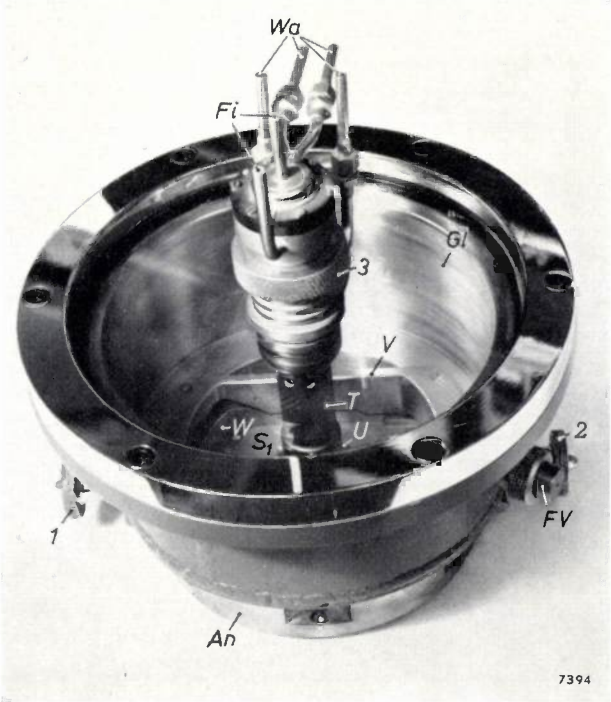


Fig. 19. The electron gun. *Fi* cathode connections. *Wa* connections for water cooling. *FV* connection for backing line. The remaining characters have the same significance as in fig.18.

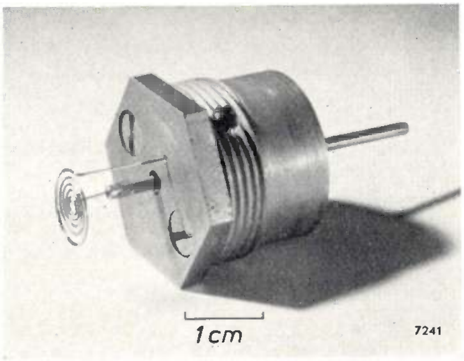


Fig. 20. Cathode holder with the directly-heated cathode of the electron gun. If the cathode has to be replaced, the complete component illustrated here is exchanged. After replacement the new cathode will then automatically be mounted in the correct position.

is necessary, the vacuum lock is closed, and air admitted to the electron gun. By unscrewing a knurled knob *3* the cathode assembly may then be withdrawn. A new pre-focussed cathode unit (fig. 20) having been screwed into position, the electron gun lock is first evacuated to the pressure of the backing pump by means of the special branch of the backing line mentioned on page 203. Connection is then re-established with the high vacuum; owing to the negligible volume of the lock relative to the volume of the accelerator proper, the high vacuum is virtually unaffected. Replacement of the cathode need only interrupt the working of the accelerator for five minutes or less.

The filament current for the cathode is supplied by a 10V-2A mains transformer (visible in fig. 22), with a secondary whose insulation can withstand 50 kV with respect to earth, and whose capacitance to earth is very low (15 pF). This capacitance must be small to limit the rise time of the pulses.

The cathode temperature, and thus the beam current, is controlled by means of a variable autotransformer in the primary of the filament transformer. This autotransformer is accommodated on the control desk.

Rectangular waveguide system

Fig. 21 is a schematic diagram of the accelerator [10]). Of the RF power entering the corrugated guide, part is used for accelerating the electrons, another part is necessary for covering losses, and the remainder is fed back to the entrance of the corrugated guide by way of a feedback bridge. In the present case this remainder amounts to 50%. In the feedback bridge, power from the magnetron makes the total up to 100%. Thus the feedback ratio of the bridge is unity. *Fig. 22* shows the system of rectangular waveguides used for the conduction of the RF power.

The feedback bridge and the waveguides connecting it with the magnetron and with the corrugated guide are filled with air at 2 atm gauge. At the RF powers under consideration, electrical discharges would occur in the guides at atmospheric pressure: either the guides must be evacuated (thus eliminating the need for windows between the waveguide system and the corrugated guide), or they must be pressurised. Pressurising has the advantage that

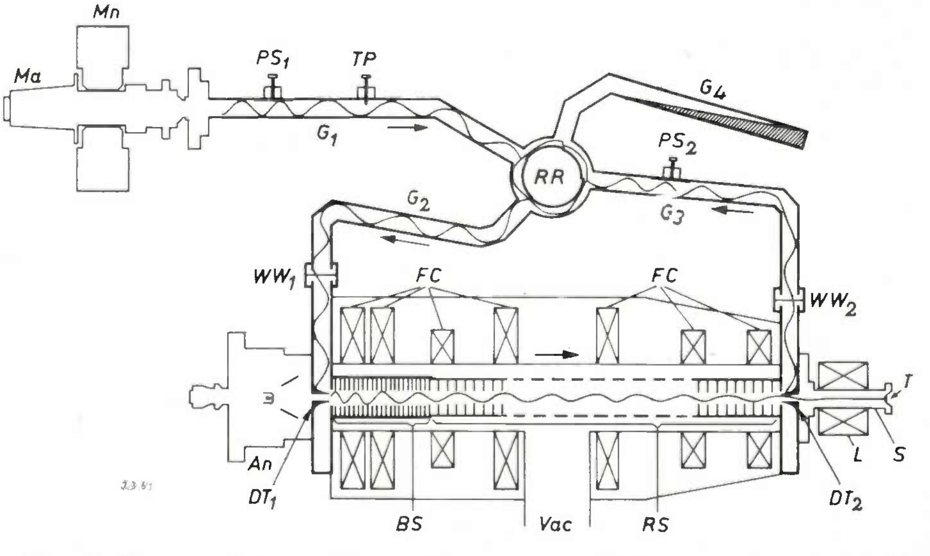


Fig. 21. Diagram of the accelerator. *Ma* magnetron with its permanent magnet *Mn*. PS_2 phase shifter in the rectangular waveguide G_3, serving to ensure correct phasing of the wave arriving at the rat race *RR*, which acts as a feedback bridge. PS_1 tuning phase shifter and *TP* tuning probe for trimming the magnetron frequency. G_4 fourth arm of the rat race, fitted with a matched water load. *BS* and *RS* are respectively the bunching section and the relativistic section of the corrugated guide. Waveguide windows WW_1 and WW_2 separate the vacuum in the corrugated guide from the 2 atm gauge pressure in the remaining rectangular guide system. *An* anode. *FC* coils supplying an axial magnetic field that precludes divergence of the electron beam. DT_1 and DT_2 are respectively the in- and output doorknob transformers [10]). *Vac* connection to the vacuum pump. *S* target snout carrying the system *L* of two magnetic quadrupole lenses that focus the electron beam on the target *T*.

[10]) The principle of the linear accelerator is discussed by D. W. Fry, The linear electron accelerator, Philips tech. Rev. **14**, 1-12, 1952/53. The transformers DT_1 and DT_2 are discussed on p. 8 of this article. For a recent survey, reference is made to: L. Smith, Linear accelerators, Encyclopedia of Physics, Vol. **44**^{1} (Nuclear Instrumentation), 341-389, Springer, Berlin 1959.

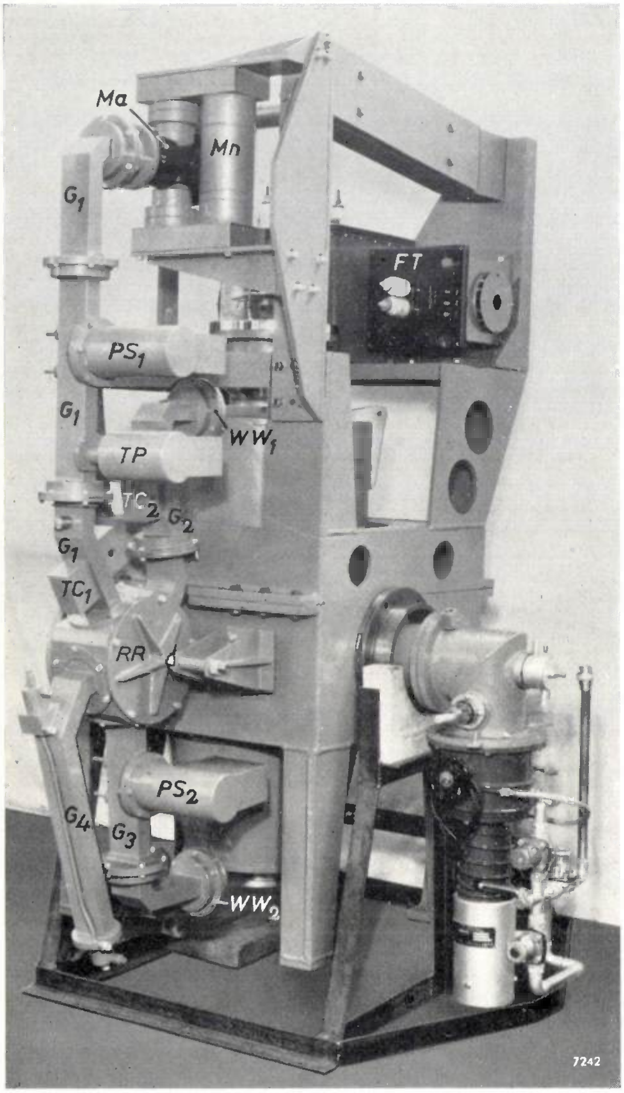


Fig. 22. Photograph of a partly-assembled accelerator, giving a view of the system of rectangular waveguides. TC_1 and TC_2 are thermocouples that measure the RF power [11]). *FT* is the mains transformer for the cathode filament. The remaining letters have the same significance as in fig. 21.

sealing problems are less troublesome and that removal of the magnetron does not affect the vacuum in the corrugated guide. Furthermore pressurising is more reliable: it has happened in waveguide systems working under vacuum conditions that the magnetron window has been melted by a glow discharge at this part of the waveguide, remote from the pump, where the vacuum is poorest.

That in the 15 MeV accelerator described earlier [3]) evacuation was chosen was due to the fact that at that time no suitable waveguide windows were available for separating the pressurised part from the vacuum in the corrugated guide. Such a window has to withstand a pressure of three atmospheres while transmitting high-power microwave signals (for the window in the line leading to the beginning of the waveguide this power may amount to an average of 4 kW, viz. 4 MW peak power with a duty cycle of 10^{-3}). In the meanwhile, however, a simple demountable window has been developed for this purpose (*fig. 23*). It consists of a polished quartz disc 5.5 cm in diameter, mounted in a brass plate 6 mm thick. An O-ring ensures a good seal. To protect this ring from possible radio-frequency heating effects, the periphery of the disc is metal-coated. Matching of the window is achieved by adjusting the internal diameter of the brass plate to suit a given quartz disc. In this way, close tolerance working of the quartz thickness is not required, nor is the diameter of the disc critical, as the O-ring satisfactorily accommodates reasonable variations in this diameter.

The application of a pressure of 2 atm gauge has called for a new design of the phase shifters (fig. 21). In the 15 MeV accelerator the phase shifters each consisted of a ceramic wedge on an adjustable mount on steel pins in the evacuated waveguide. At a pressure of two atmospheres gauge such a wedge was found to give rise to electric discharges. This has been overcome by using a lozenge-shaped body of polystyrene that can be moved on stainless-steel supports. Just as previously, the phase shifters are each provided with a small electric motor to permit remote adjustment. The protective covers over the driving mechanism are visible in fig. 22.

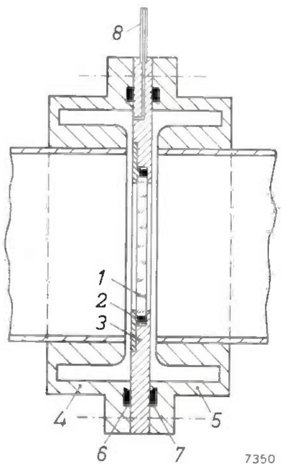


Fig. 23. Cross-section of the waveguide windows (WW_1 and WW_2 in figs 21 and 22). *1* quartz-glass disc mounted in brass flange *3* vacuum-sealed by means of O-ring *2*. *4* and *5* are choke flanges as commonly used in waveguide connections. O-rings *6* and *7* provide vacuum sealing. To the left of the quartz disc the pressure is 2 atm gauge, maintained via the line *8*; to the right is the vacuum of the corrugated guide.

[11]) Cf. the article of note [3]), page 19.

In the 15 MeV accelerator a "circular magic tee" was used as the feedback bridge; it has been discussed in detail previously [3]). It works successfully but, because the waveguides connected to it do not all lie in the same plane, the construction demands a great deal of space. With a view to manoeuvrability, it was desirable to limit the size of the accelerator as much as possible. A feedback bridge was therefore developed in which the waveguides all lie in one plane. This device is known as a "rat race". This rat race (*fig. 24*) is so constructed that the radio-frequency couplings are separate from the gasketted joints for sealing the device. This construction can also be used in an evacuated system.

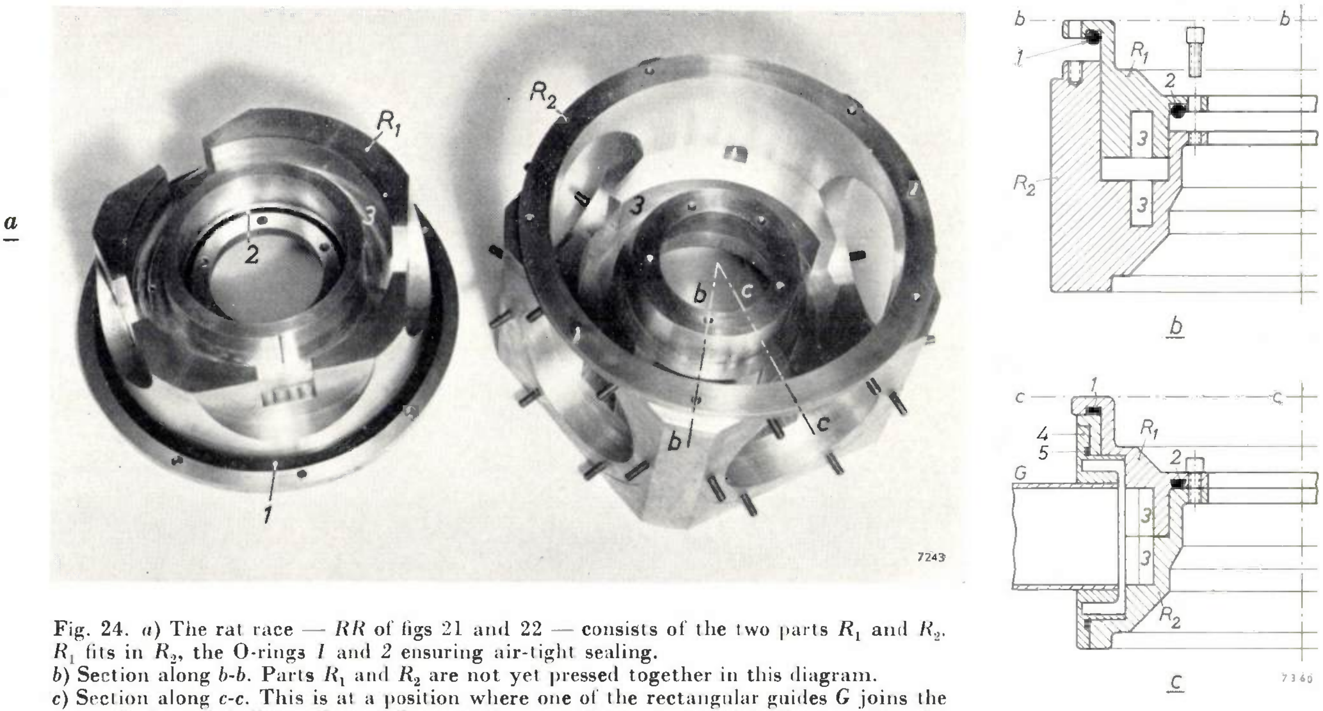


Fig. 24. *a*) The rat race — *RR* of figs 21 and 22 — consists of the two parts R_1 and R_2. R_1 fits in R_2, the O-rings *1* and *2* ensuring air-tight sealing.
b) Section along *b-b*. Parts R_1 and R_2 are not yet pressed together in this diagram.
c) Section along *c-c*. This is at a position where one of the rectangular guides *G* joins the rat-race channel *3*. Parts R_1 and R_2 are fully pressed together in this diagram. For the high-frequency coupling of guide *G* and the rat race a choke flange *4* is used, the O-ring *5* providing the seal.

Corrugated guide

A travelling electromagnetic wave that is to be used for the acceleration of electrons must in the first place have an electrical component in the direction of propagation. In the second place the velocity of the wave in the guide (v_p) must increase in such a way along the guide that the wave and the electrons carried along by it always remain in phase. The electrons enter the corrugated guide with a velocity of about 0.4 *c*. As their energy increases, their velocity increases and approaches *c* asymptotically; v_p must therefore also increase in the same way. An electromagnetic wave with such properties can be obtained in a round copper tube divided by copper diaphragms into separate cells. Such an accelerator tube, commonly called a corrugated guide, is generally used in linear electron accelerators. In the diagram of our accelerator (fig. 21), which is of conventional construction, one can see that the corrugated guide (1 m length, 9 cm external diameter) is surrounded by an envelope 12.5 cm in diameter, to which the vacuum pump is connected and with which all the cells (cavities) are connected by holes. The cavities are thus pumped radially to obtain a good vacuum everywhere. Around the vacuum envelope are the coils supplying the axial magnetic field that precludes divergence of the electron beam. To give mechanical robustness, the whole assembly is encased in a steel drum which also acts as a return path for the magnetic flux.

The way in which v_p increases along the corrugated guide can be controlled by suitably varying the radius *b* of successive cavities (*fig. 25*). The smaller the diaphragm pitch *s*, the more accurately can v_p be made to vary as required. For this reason *s* is made smaller at the beginning of the guide (where v_p must increase quickly) than further along, where the velocity of light has almost been reached. The diaphragms all contribute to the ordinary ohmic losses, so that the number of diaphragms must not be made unnecessarily great.

In the first part of the corrugated guide the electrons that are carried along by the wave are bunched into groups (one bunch per wavelength). The first part is therefore called the "bunching section". The second part is known as the "relativistic section", because the energy here taken up by the electrons mainly results in a relativistic increase of their mass. In the present equipment the corrugated guide consists of a bunching section of 22 cm, with $s = 1$ cm, and a relativistic section — in which the electrons take up by far the greater part of their final energy — of 78 cm with $s = 2$ cm.

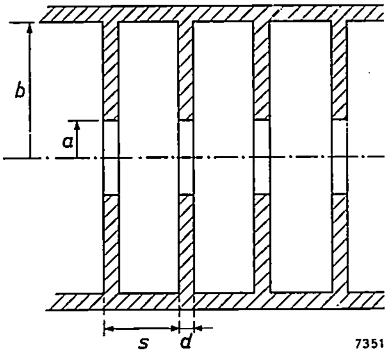


Fig. 25. Dimensions of the corrugated guide sections. *a* aperture radius. *b* radius of the resonant cavities. *s* pitch of the diaphragms. *d* diaphragm wall thickness.

The parameter a/λ_0

Of great importance is the choice of the radius *a* of the diaphragm apertures (fig. 25), or, more precisely, of the ratio a/λ_0, where λ_0 is the wavelength in free space corresponding to the frequency f (i.e. $\lambda_0 f = c$). This ratio is the main parameter determining the energy finally taken up by the electrons with a given magnetron power. Once *a* has been chosen, *b* is fixed by the specified value of v_p. One would like to make *a* small, because this results in a high electron energy and hence also in a large X-ray output, because the X-ray output per mA beam current increases sharply with this energy (this can be seen from fig. 26*b*). However, the smaller *a*, the more sensitive v_p becomes to variations in the magnetron frequency f and to tolerances in the dimensions of the corrugated guide. As the wave and the electrons must be kept very accurately in phase in order to obtain effective operation of the corrugated guide, both dimensional tolerances and constancy of frequency become more and more critical as *a* is made smaller. The ratio a/λ_0 thus determines the accuracy to be demanded of the frequency and the dimensions. For the corrugated guide of the present equipment, the bunching section has the ratio $a/\lambda_0 = 0.168$, while the relativistic section has the value $a/\lambda_0 = 0.13$. To compensate for the effect on v_p of the discontinuities in *a* and *s* at the junction of the two sections, *b* at this place also shows a discontinuity (fig. 21).

As the electrons gather by far the greater part of their energy in the relativistic section, it is of special importance to have a small value of a/λ_0 in this section.

The selected values of a/λ_0 originate in calculations based on a tolerance of 1 in 10^4 for the frequency and for certain dimensions. Experience since obtained has shown, however, that with these accuracies the value of a/λ_0 for the relativistic section might well have been reduced to 0.1. It would therefore be possible to obtain from a corrugated guide of one metre length a still greater X-ray output than from the present installation. It can be read from fig. 26*c* that about a 30 percent increase of the pulse dose rate can be expected.

Build-up phenomena

In the design of an accelerator, account should be taken of the warming up of the magnetron (effect on frequency) and of the corrugated guide (effect on its dimensions) when the installation is switched on. It is moreover desirable that the frequency is appropriate to the dimensions of the guide not only under steady-state conditions, but also in the warming-up period that precedes it. If this is not so it will be some time after switching on before the full X-ray output is produced. This is known as the build-up effect. Thanks to careful design of the water cooling circuit for the corrugated guide, the full X-ray output is obtained in 2 or 3 seconds.

Design for maximum X-ray output

Once the lengths of the bunching and relativistic sections and their values of a/λ_0 have been chosen, and when the magnetron power is given, there remains to decide how v_p, and hence the radius *b* of successive cavities, must vary along the guide. This variation should of course satisfy the condition that wave and electrons are in phase all along the guide. A suitable variation, however, satisfies this condition only for one value of the beam current. The design of the guide thus depends on the beam current selected. As will be shown, it is possible to estimate theoretically the X-ray output as a function of beam current. A curve is then obtained which shows a maximum and the corrugated guide should obviously be designed for the beam current that corresponds to this maximum.

How the above-mentioned curve is obtained, is shown in *fig. 26*. In fig. 26*a* the energy *E* of the electrons is plotted as a function of the beam current I, for the case in which the variation in b matches this current [12]). This has been done for two values of a/λ_0 in the relativistic section, viz. $a/\lambda_0 = 0.13$ and $a/\lambda_0 = 0.1$. Every point on such a curve (not measured, but calculated) thus corresponds to a different variation of b, i.e. to a different corrugated guide. In fig. 26*b* the X-ray dose rate per mA of beam current is plotted as a function of the electron energy for a tungsten target. This curve has been obtained from measurements of various workers, including our own measurements. From fig. 26*a* and *b*, fig. 26*c* has been derived, whence it is found that a maximum X-ray output may be expected if a variation of b is chosen that matches the corrugated guide at a beam current of about 0.2 A. It follows from fig. 26*a* that with this beam current an electron energy of about 4.3 MeV may be expected when a relativistic section with $a/\lambda_0 = 0.13$ is used. To find the X-ray dose rate in the beam axis at a distance of 1 m from the target, with no focussing of the electron beam, the value of 8.9×10^5 r/min found from fig. 26*c* still requires to be multiplied by the duty cycle. With 500 pulses of 1.8 μsec per second (see page 206) the duty cycle amounts to 9×10^{-4}, so that 800 r/min may theoretically be expected. The actual accelerator comes very close to these expectations.

As a further illustration of the robustness and versatility of this type of equipment it can be mentioned that a modified version of the equipment described in this article has been supplied by Mullard to the United Kingdom Atomic Energy Authority for use in the radiography on site of nuclear-power-station pressure vessels (*fig. 27*).

The main modification was the elimination of all external water cooling circuits. A further difference was the accommodation of all the power supplies, modulator and switchgear in a weatherproof control cabin (fig. 27*b*) linked to the accelerator by

[12]) The method of calculating these curves is discussed in the article quoted in note [3]), pages 3-4.

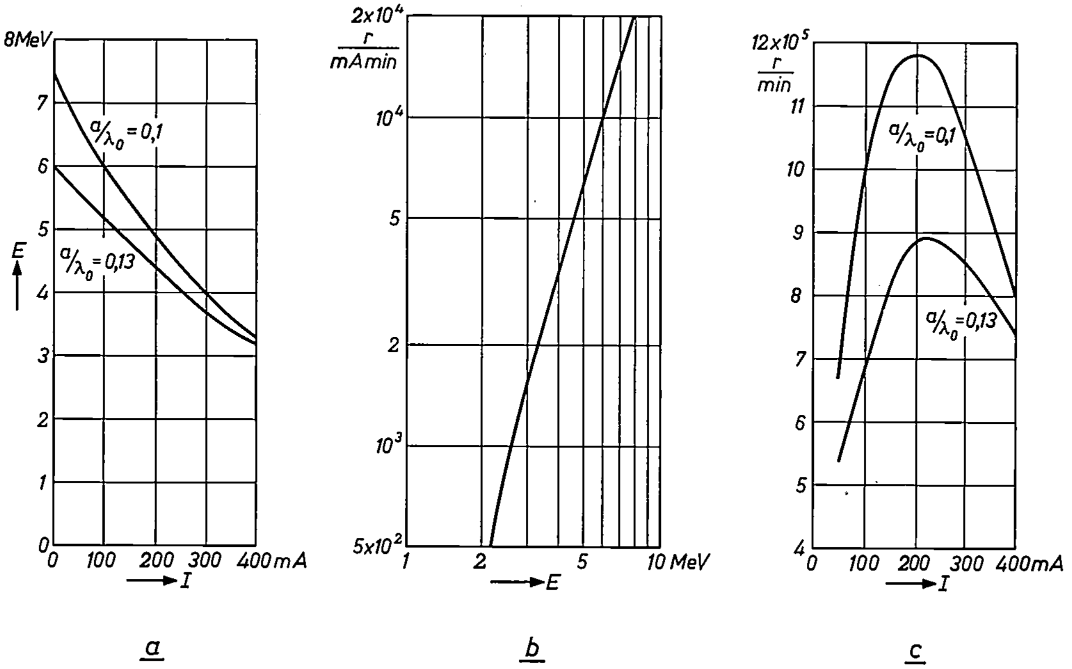


Fig. 26. *a*) Calculated energy E attainable by the electrons in a corrugated guide with feedback, if the variation of the radius b (cf. fig. 25) along the guide exactly matches the electron-beam current I. Each point of the curves therefore refers to a different guide. The curves are calculated for guides with a 22 cm long bunching section of 1 cm pitch and $a/\lambda_0 = 0.168$, followed by a 78 cm relativistic section of 2 cm pitch with both $a/\lambda_0 = 0.13$ and $a/\lambda_0 = 0.1$. The electrons enter the guide with a velocity of 0.4 *c*, the magnetron power is 2 MW (in the pulse), and the feedback ratio is unity.
b) Measured X-ray output as a function of the electron energy in MeV of a parallel electron beam falling perpendicularly on a tungsten target. The curve represents the dose rate on the axis of the generated X-ray beam at 1 m from the target, for a mean electron beam current of 1 mA.
c) Curves obtained by combining the data of *a* and *b*, representing the predicted X-ray intensity from a linear accelerator during a pulse, on the beam axis at 1 m from the target, as a function of the beam current I for which the accelerator is designed. To obtain maximum X-ray output, the accelerator should obviously be designed for $I \approx 200$ mA.

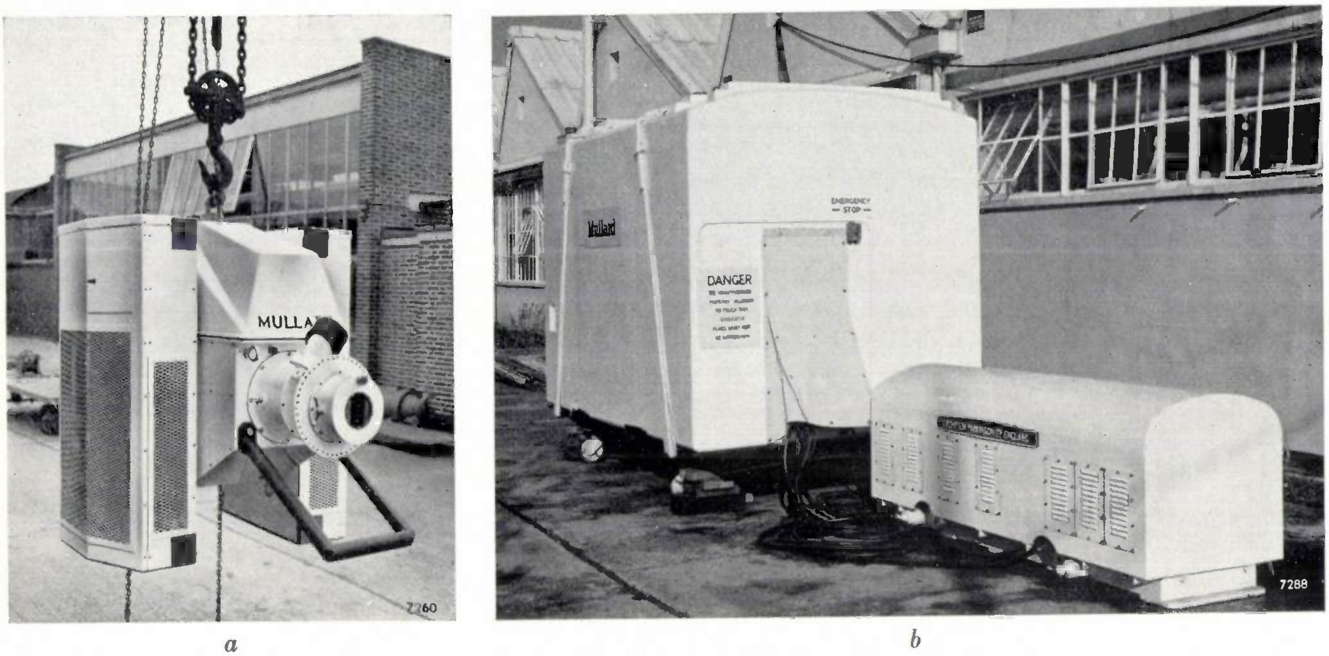


Fig. 27. A fully transportable version of the linear-accelerator radiographic unit. The installation is wholly independent of external electricity and water supplies.
a) The accelerator X-ray generator proper. Suspended from a crane it can be manoeuvred into any desired location.
b) Cabin containing power supplies, control equipment and modulator. Foreground: rotary alternator (for mains stabilisation on building sites where large fluctuations are encountered).

cables 250 feet long. The suspension system described in this article is not needed in the present case: the accelerator can easily be handled and manoeuvred by site cranes in prefabrication shops and behind biological shields. The high output of the accelerator also permits it to be mounted on a turntable at the centre of curvature of the pressure vessel, so minimising setting-up time and allowing considerable lengths of weld to be radiographed at one exposure (*fig. 28*).

Depending upon the type of film used, it has been reported that up to 10 feet of pressure-vessel weld

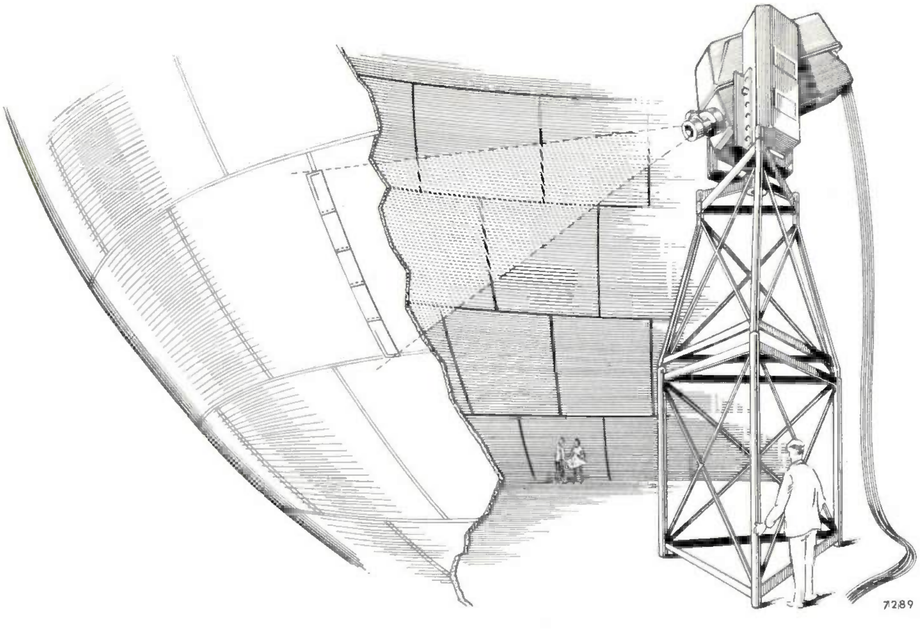


Fig. 28. Sketch of the transportable unit set up at the centre of a spherical steel reactor pressure vessel (diameter 20 yards, wall thickness 3½″, or 9 cm). Such pressure vessels, used in nuclear-reactor power stations, are constructed on site and radiographed during construction.

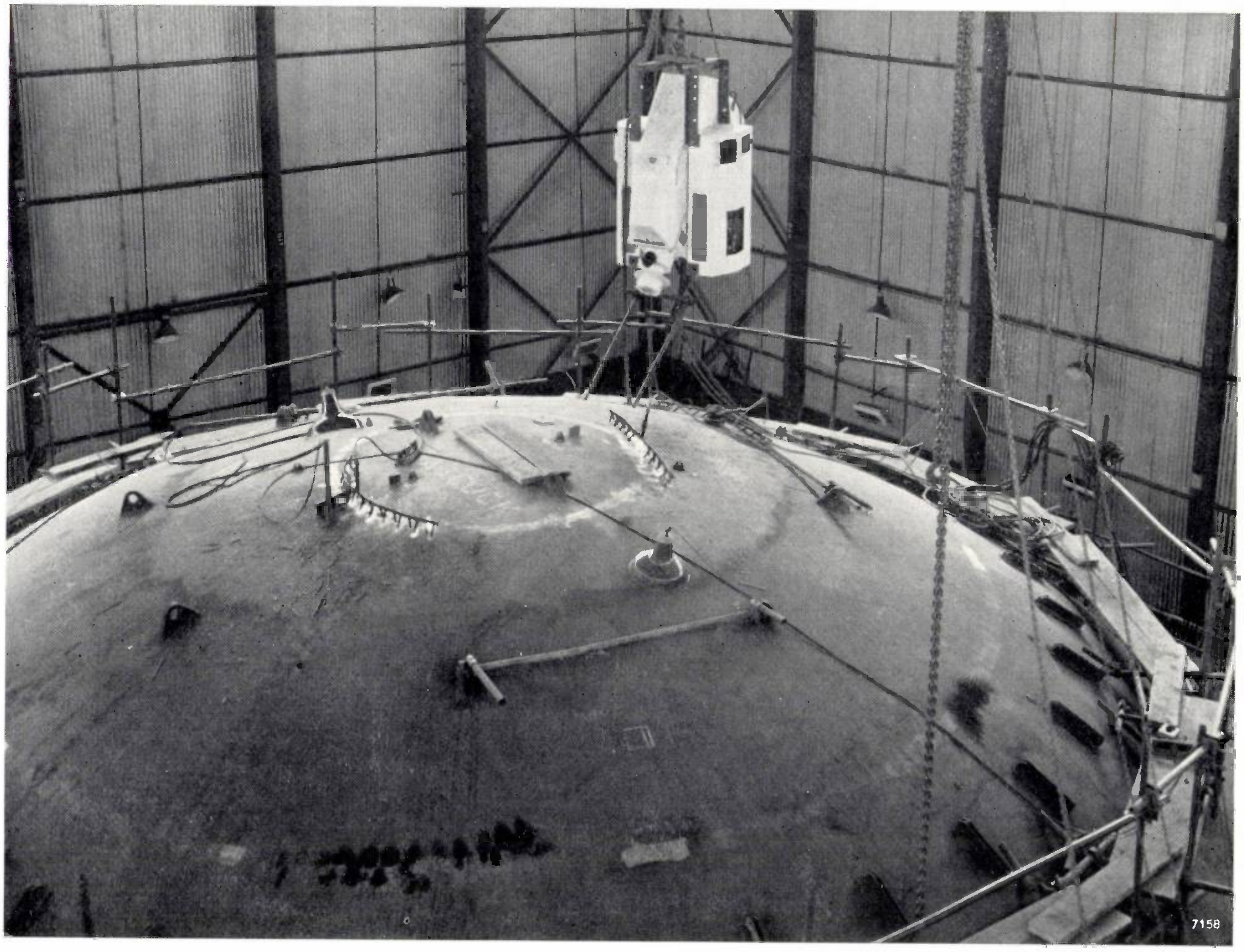


Fig. 29. The transportable accelerator in operation. It is shown here above one of the pressure vessels of the nuclear power station at Trawsfynydd (Wales). Photograph by courtesy and with permission of Atomic Power Constructions Ltd.

can be radiographed in 15 minutes. *Fig. 29* shows the accelerator in operation, suspended above a reactor pressure vessel. The reduced radiographic schedule on projects of this type represents a very considerable economic advantage over all other known radiographic equipments suited to applications of this character.

The author wishes to thank the Director of Mullard Research Laboratories for permission to publish this article. He also wishes to thank the staff of the Armaments Research and Development Establishment and his colleagues at Mullard Research Laboratories for their invaluable cooperation and assistance.

Summary. With an electron linear accelerator, based on a corrugated waveguide of 1 m length, it is possible to construct a very intensive source of hard X-rays that is particularly suitable for the radiography of heavy steel sections (up to 16″). An X-ray installation for such purposes, in regular use since 1956, has been designed and constructed by Mullard Research Laboratories in England. Easy manoeuvrability cf the X-ray beam is a special requirement of such industrial radiographic equipment in view of the unwieldiness and weight of many specimens. This has been achieved by means of a suspension which permits of swiveling and tilting and of varying the height of the accelerator. The accelerator is mounted under a trolley riding on a gantry, rather like a gantry crane. The electrons are accelerated to 4.3 MeV and focussed to a spot of 1.5×0.5 mm on a tungsten target by means of a system of magnetic quadrupole lenses. The X-ray output is variable in steps; the maximum dose rate at 1 m from the target amounts to 600 r/min in the centre of the X-ray beam. A specially developed X-ray head limits the beam to a rectangular field whose dimensions at 1 m from the tungsten target are adjustable to between 3×3 cm and 15×25 cm. Results depend on radiographic technique; on the average one may expect to detect flaws of about 0·75 mm in a 30 cm (1 foot) wall, when using a slow, sensitive film.

Various details of the design are described: the rotatable joints in the high-vacuum and the backing lines, enabling the oil-diffusion pump to remain vertical when the accelerator is tilted; the electron gun with its simplified vacuum lock, between it and the high vacuum of the corrugated guide, which permits cathode replacement in less than five minutes; the air-tight waveguide windows, enabling the system of rectangular waveguides — that couples the magnetron (the high-frequency energy source) to the corrugated guide — to sustain a pressure of 2 atm gauge; the "rat race" serving as a feedback bridge and providing a compact construction. The design of the corrugated guide for maximum X-ray output is explained.

At the end of the article brief mention is made of a more recent X-ray unit of this type, also with a 4.3 MeV linear accelerator, that is fully transportable. This version has proved of great value in the construction of pressure vessels for nuclear reactors.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of those papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2875: Th. P. J. Botden: The silica gel leak detector (Advances in vacuum science and technology, Proc. 1st int. congress on vacuum techniques, Namur, Belgium, June 1958, editor E. Thomas, Vol. 1, pp. 241-244, Pergamon, Oxford 1960).

The silica-gel leak detector consists of a sensitive Penning ionization gauge connected to a tube filled with grains of silica gel, the latter at liquid-air temperature. For a flow of H_2, He and Ne through the silica gel, the sorption is less than 5%, while for air only the fraction 3×10^{-4} can pass the silica gel. Because of this selective sorption and the use of the Penning gauge, leaks of 10^{-9} Torr l/s can be readily detected. With some precautions it is possible to detect leaks of 10^{-11} to 10^{-12} Torr l/s. The sensitivity is equal for H_2, He and Ne when these gases are used as detection gas. An indication of a leak is, in general, obtained within 10 s. From the plot of gauge current versus rate of leak and the sensitivity curve of the gauge, the approximate values of pumping speed of the gauge for H_2, He and Ne have been calculated.

2876: N. Warmoltz: Spectromètre de masse pour la détection des fuites, fonctionnant avec un mélange à faible teneur en hélium (as **2875**, Vol. 1, pp. 257-259). (Mass spectrograph for leak detection, using a mixture of helium and air; in French.)

Description of a leak detector using helium as detector gas. The presence of He is demonstrated with a simple mass spectrograph. The ions, drawn from a hot-cathode source, are initially accelerated by an electrostatic lens and focussed on the first dynode of a photomultiplier tube, which gives considerable amplification of the ion current to be measured. If the air pressure in the apparatus is fairly high, the peaks at 6 (C^{++}), 7 (N^{++}) and 8 (O^{++}) may overlap the peak at 4 (He^{+}). This is avoided by using a double spectrograph (analogous with a double monochromator). The multiplier tube increases the sensitivity by a factor of 500. This makes it possible to use a mixture of helium and air with only 1-5% helium, which is much more economical.

2877: J. J. Opstelten and N. Warmoltz: Un manomètre à membrane à échelle linéaire pour des pressions entre 10^{-5} et 10 mm Hg (as **2875**, Vol. 1, pp. 295-298). (A linear-scale diaphragm manometer for pressures between 10^{-5} and 10 mm Hg; in French.)

The essential part of the manometer described is a stainless-steel diaphragm 20 μ thick. The bending of the diaphragm is a function of the pressure drop across it, and is measured electrically. For this purpose the diaphragm is sandwiched between two metal plates mounted at a distance of 25 μ on either side, with ceramic spacers and glass lead-in insulators. The two capacitors thus produced are incorporated in a Wheatstone bridge with AC supply (frequency 0.5 Mc/s). A null method of measurement is used: two voltages, $V + v$ and $V - v$ respectively, are applied to the two capacitors and v is varied so as to bring the diaphragm into the balanced position. The pressure on one side is kept constant at about 10^{-7} mm Hg with a Penning gauge. A linear relation then exists between v and the pressure on the other side. The choice of V depends on the pressure to be measured in a range from 10^{-5} mm Hg to 1 mm Hg, beyond which a troublesome gas discharge occurs. Higher pressures can be measured by omitting V and determining the capacitance of one of the capacitors in a bridge. This extends the range up to 10 to 20 mm Hg. The manometer can if necessary be heated up to 150 °C. A special circuit is needed for gases that form absorbed layers on the capacitor plates. In the case of highly reactive gases (e.g. UF_6) the chamber to which the gas is admitted is not provided with an electrode, so that the ceramic and glass parts (which are the most liable to attack) can also be omitted.

2878: S. Woldring: The mechanics of breathing — general principles and technique of measurement (Proc. Tuberculosis Res. Council (Roy. Netherl. Tuberculosis Ass.) No. 46, 1959, pp. 5-27, published 1960).

A survey is given of the basic principles of the mechanics of lung ventilation. Lung volume, rate of flow and intrathoracic pressure are shown to be interrelated and to depend on the elastic properties of the lungs and on the resistance to flow of the airways. The factors which determine the behaviour of these quantities are discussed.

An instrumental set-up is described for the simultaneous recording of volume, pressure and flow. It also enables the investigator to plot these variables against each other or against time. The characteristic properties of these instruments are discussed.

2879: H. C. Burger: The significance of the flow/volume diagram in the study of the mechanics of breathing (as **2878**, pp. 28-47).

Using examples of normal and abnormal cases, the author demonstrates the diagnostic value of flow-volume diagrams, as recorded by the equipment described by Woldring in No. **2878** above. Some of the phenomena observed are attributed to the fact that the respiratory passages during rapid exhalation may become constricted or almost closed by a check-valve mechanism. An aerodynamic explanation is given to supplement the mechanical theory.

2880: G. D. Rieck and H. A. C. M. Bruning: Subcrystals in large vapour-grown crystals of tungsten (Acta metallurgica **8**, 97-104, 1960, No. 2).

Single crystals of tungsten, grown by decomposition of the chloride in the vapour phase, were investigated. A substructure has been found both with X-ray and microscopic techniques. The disorientation between the subcrystals is random and is therefore different from that found in single crystals in recrystallized doped tungsten wires. The occurrence of a substructure or even dendritic branches depends upon the circumstances during the growth. The rows of etch pits on the photomicrographs of etched surfaces are of the same nature as those found by other authors on tungsten prepared in a different manner. Electron-microscopc pictures of etched surfaces sometimes reveal pyramid-shaped etch hills and whisker-like needles, which are supposed to be subcrystals grown with the highest perfection, during temporarily favourable conditions.

2881: P. Westerhof and E. H. Reerink: Investigations on sterols, XV. The syntheses and properties of $9\beta,10\alpha$-progesterone and 6-dehydro-$9\beta,10\alpha$-progesterone (Rec. Trav. chim. Pays-Bas **79**, 771-783, 1960, No. 7).

The preparation of steroid hormone analogues with abnormal configuration has been intensively investigated. This paper describes the synthesis of $9\beta,10\alpha$-progesterone and 6-dehydro-$9\beta,10\alpha$-progesterone from lumisterol$_2$. Both compounds show a pronounced progestational activity, even when administered orally. Numerous intermediates are described too. The structures of the compounds prepared were confirmed by ultraviolet and infrared absorption data. The biochemical activities are recorded briefly, and will be reported fully elsewhere.

2882: P. Westerhof and E. H. Reerink: Investigations on sterols, XVI. The syntheses and properties of $9\beta,10\alpha$-androstanes (Rec. Trav. chim. Pays-Bas **79**, 794-806, 1960, No. 7).

Pursuing the investigations described in No. **2881**, the authors describe the syntheses of a number of $9\beta,10\alpha$-androstanes, starting from 22-(1′-piperidyl)-$9\beta,10\alpha$-bisnorchola-4,20(22)-dien-3-one or from $9\beta,10\alpha$-bisnorchol-4-en-3-on-22-al. The physiological properties of the $9\beta,10\alpha$-androstanes prepared will be reported elsewhere. The paper closes with a detailed description of the methods of preparation. The chemical structures assumed for the compounds are confirmed by absorption and optical-rotation data.

2883: B. G. van den Bos, M. J. Koopmans and H. O. Huisman: Investigations on pesticidal phosphorus compounds, I. Fungicides, insecticides and acaricides derived from 3-amino-1,2,4-triazole (Rec. Trav. chim. Pays-Bas **79**, 807-822, 1960, No. 7).

The preparation and the biological properties of a number of compounds with fungicidal, insecticidal and acaricidal activities are described. The compounds were prepared by substitution of a phosphorus-containing group for a hydrogen atom of the 1,2,4-triazole nucleus in 3-amino-1,2,4-triazole or a 5-substituted derivative thereof. Biological activity was highest when the phosphorus-containing group was a bis(dimethylamido)phosphoryl group.

2884: A. H. Boerdijk: Diagrams representing states of operation of a general thermocouple (J. appl. Phys. **31**, 1141-1144, 1960, No. 7).

The state of operation of a thermocouple of which (*a*) the bars have an arbitrary shape, (*b*) the properties of the materials are arbitrary functions of temperature, and (*c*) the composition is, under certain restrictions, inhomogeneous and anisotropic, depends on three independent parameters: the current I and the temperatures T_1, T_2 of the junctions. If T_2 is kept constant, operating characteristics, such as curves of constant output power or efficiency, can be plotted in an $I,(T_2{-}T_1)$ diagram. The existence of regions of generation of electricity and of cooling is proved. These regions are investigated. Possible generalization and reduction of the diagram are discussed. As an illustrative example, the cooling region of a general couple with temperature-independent properties is dealt with.

2885: J. W. L. Köhler: The gas refrigerating machine and its position in cryogenic technique (Progress in Cryogenics **2**, 41-67, 1960).

After a brief review of the various cryogenic techniques, the theory of the Stirling process is described. The ideal cycle (without losses) and the departures from the ideal in practical cycles are discussed, and an example is given. Applications mentioned are the liquefaction of air, gas separation, the cold box, etc. The position of the gas refrigerating machine is considered in relation to existing refrigerating methods, in which connection the subject of efficiency is discussed. It is shown that the Stirling process has opened up many new possibilities in the cryogenic field, especially in the range between —80 °C and —180 °C. This applies in particular to small gas refrigerating machines, although the refrigerating capacity varies in a wide range and at present an upper limit is not in sight. See also Philips tech. Rev. **16**, 69 and 105, 1954/55, and **20**, 177, 1958/59.

2886: G. J. M. Ahsmann and Z. van Gelder: La chute cathodique normale pour des cathodes monocristallines (Le Vide **15**, 226-233, 1960, No. 87). (The normal cathode drop in the case of monocrystalline cathodes; in French and in German.)

The authors describe reproducible measurements of the normal cathode drop in the case of monocrystalline platelets of germanium, silicon and copper in neon at a pressure of 40 mm Hg, using the sputtering method. Changes in the surface sometimes give rise to changes in the cathode drop. Differences between crystal faces are attributed to differences in work function φ. The formula $V_n = C\varphi$ is discussed. Also dealt with are temperature effects due to (*a*) changes in density distribution, and (*b*) surface changes brought about by the penetration of gas ions. Monocrystalline materials, because of their stable and reproducible cathode drop, are in principle well-suited for the production of glow-discharge stabilizers.

2887: A. Kats and Y. Haven: Infra-red absorption bands in α-quartz in the 3 μ region (Phys. Chem. Glasses **1**, 99-102, 1960, No. 3).

In quartz crystals, H^+ ions occur as impurities. The most common impurity is Al^{3+}, which replaces Si^{4+}; the resultant charge deficiency is compensated by Li^+, Na^+ or H^+. The presence of the H^+ ions gives rise to absorption bands at 3311, 3371 and 3435 cm^{-1}. It is shown from deuterium exchange experiments that some bands in this part of the absorption spectrum are due to O-H vibrations. The exchange is effected by heating quartz crystals at 1000 °C in D_2O vapour. The absorption referred to then disappears and new O-D bands appear in the region of 2500 cm^{-1}.

2888: L. Heijne: Photoconductive properties of lead-oxide layers (thesis Amsterdam, June 1960).

In this thesis a study is made of the photoelectrical properties of vapour-deposited polycrystalline layers of very pure PbO in the yellow, orthorhombic modification. These layers can be heated to about 350 °C without phase change, unlike single crystals, which have a tendency to change to the red, tetragonal modification, which is stable below 489 °C. Special attention is paid to the properties of semiconductor contacts, in connection with the space-charge layers associated with them. The mathematical analysis of the flow of charge carriers and the kinetics of their generation and recombination is presented. The methods of preparation and measurement are described and experimental results are given. A separate study is devoted to trapping levels and the determination of their properties using the electrical glow-curve method. The last chapters deal with the effect of impurities on the photoelectrical characteristics, and discuss the use of photoconductive layers in television camera tubes.

2889: N. V. Franssen: Some considerations on the mechanism of directional hearing (thesis Delft, July 1960).

This thesis deals with problems of directional hearing in connection with stereophonic sound reproduction. Like recent investigations by Cherry, Licklider and David, it is concerned with the mechanism of directional hearing as well as with observable phenomena. An attempt is made to design and build an electronic model of the hearing mechanism capable of explaining the experiments, whose purpose, among other things, is to give a clearer insight into the concepts of intensity and time differences. It is shown that the apparent direction of the sound is mainly determined at the start. The behaviour of the model is in reasonably good agreement with measurements made by K. de Boer. The experiments indicate that front-back discrimination is mainly achieved by movements of the head as far as low tones are concerned, whereas the influence of the external ears predominates for high tones. The relation between head movements and the projection of the perceived signals in space is discussed. Experiments relating to the discrimination of binaurally presented frequency differences are described. The results are explained in terms of a second electronic model. The combination of this model with the earlier one makes it possible to give a plausible explanation of the phenomena produced when two loudspeakers emit speech or music with a relatively large time difference. In the last chapter it is shown that the model can also cast some light on hearing phenomena not directly related to direction. This is discussed with respect to the reception of pitch and timbre. A possible mechanism for the analysis of sound by the hearing is considered, and the physical quantities related to the subjective evaluation of a given sound are suggested.

2890: W. Verweij: Probe measurements and determination of electron mobility in thc positive column of low-pressure mercury-argon discharges (thesis Utrecht, September 1960).

This thesis describes measurements and calculations relating to the positive column of electrical discharges in mixtures of argon and mercury vapour. The investigation was particularly concerned with the electrical properties of the plasma, which are mainly determined by the mobility of the electrons. Using electric probes, the electron concentration and its radial distribution, the electron temperature and the field strength (gradient) along the axis were measured for numerous current values and partial pressures of argon and mercury vapour. From the probe characteristics, obtained as described by Langmuir, it was possible to determine the concentration and the velocity distribution of the electrons in the plasma. Measurements with radially displaceable probes provided the spacial distribution of these data. The field strength was determined with two probes located at a known distance from each other on the tube axis. Reliable and reproducible results were ensured by taking steps to avoid oscillations at the electrodes, by using very thin and short probes (20 μ thick, and a few mm long), and by heating the probes before each measurement. In each series of measurements one variable was changed and the others held constant. The argon pressure was varied from 0 to 20 mm Hg, the mercury-vapour pressure from about 0.5×10^{-3} to 90×10^{-3} mm Hg. The current varied from 0 to 800 mA. The velocity distribution of the electrons could be characterized by an electron temperature. Two methods of determining electron concentration gave values in good agreement with one another for argon pressures below about 9 mm Hg. The experimental temperature values of electron mobility were reduced to the case of a homogeneously distributed gas filling. It is shown that the electron mobility in the gas mixture can be satisfactorily calculated from the collision cross-sections of the gas atoms according to the Lorentz electron theory, but not from the mobility values in the separate components.

2891: E. F. de Haan: Signal-to-noise ratio of image devices (Adv. in Electronics and Electron Physics **12**, 291-306, 1960).

After giving the general formulae for noise in induced currents and noise sources in television, the author analyses the noise in various image devices, i.e. in image intensifiers (for X-rays and light) and in multiplier tubes. He then discusses methods by which the signal-to-noise ratio S/R can be improved. Finally, it is shown how these methods can be applied to pick-up equipment using vidicons and image orthicons. A distinction is made between noise where S/R is proportional to the square root of the radiation intensity, and where S/R increases linearly with the radiation intensity.

2892: W. Kwestroo and A. Roos: Compounds in the system TiO_2-Cr_2O_3-Fe_2O_3 (J. inorg. nucl. Chem. **13**, 325-326, 1960, No. 3/4).

The three-phase system TiO_2-Cr_2O_3-Fe_2O_3 was investigated by firing various mixtures of these oxides at 1300 °C and quenching them to room temperature. Some of the compounds, which Andersson found in the two-phase system TiO_2-Cr_2O_3, were examined as regards their stability in the presence of Fe_2O_3. It is shown that the compound $TiCr_2O_5$, which was not found by Andersson, can be made by replacing at least a sixth part of the Cr^{3+} ions by Fe^{3+} ions. The structure of this compound is discussed.

2893: R. J. Meijer: The Philips Stirling thermal engine — analysis of the rhombic drive mechanism and efficiency measurements (thesis Delft, November 1960).

This thesis is devoted to the Philips hot-gas engine with rhombic drive mechanism. Chapter I contains a short history of the hot-gas engine, and discusses the Stirling process and a new drive mechanism for the displacer-piston engine which offers great advantages for balancing the engine and reducing the gas forces. Chapter II contains the analysis of this "rhombic drive", viz. the derivation of the conditions for balancing; the derivation of various quantities needed for further calculations from the dimensions of the drive mechanism; the calculation of the pressure variations in the common buffer space of a multi-cylinder engine with arbitrarily chosen crank angles; the calculation of the torque due to gas forces and inertia forces; the calculation of the forces in the drive mechanism; an estimation of the friction energy and the torque due to friction. Some applications of the equations involving the torque are also given. This chapter ends with a sample calculation for a single-cylinder hot-gas engine which has been built and tested in the Research Laboratories, Eindhoven. Chapter III describes efficiency measurements made on the test engine mentioned in chapter II. After a short description of the construction of the engine the measuring equipment is described; the measurements are summarized in tables and graphs. Finally, some properties of the hot-gas engine are compared with those of the internal-combustion piston engine. The equations derived in this thesis are collected in an appendix; three further appendices give the coefficients of the series expansions used in the calculations of chapter II. See also Philips tech. Rev. **20**, 245, 1958/59.

2894: C. A. de Bock and A. M. Worst-Van Dam: A method for the measurement of antibodies against *Hemophilus pertussis* in sera (Antonie van Leeuwenhoek **26**, 73-76, 1960, No. 1).

The antibody content of *Hemophilus pertussis* sera is usually determined with the tube-agglutination technique, wherein the agglutinogen consists of a fresh *H. pertussis* phase I bacteria suspension, prepared from a culture on the well-known Bordet-Gengou agar. The authors frequently needed such a fresh suspension in order to determine agglutination titers from sera. As the preparation of such fresh suspensions takes up much time, they decided to search for a stable *H. pertussis* antigen, with the same properties as a fresh bacterial suspension. To obtain a further saving of time, a modification of the usual technique of tube-agglutination was desirable. For the purpose the method with porcelain tiles known from the field of virology was adapted.

In the present paper the preparation of coloured *H. pertussis* antigen and a new agglutination technique are described and their usefulness is shown.

2895: C. A. de Bock and A. M. Worst-Van Dam: Fixation of the agglutinogen from *Hemophilus pertussis* on the bacterial surface (Antonie van Leeuwenhoek **26**, 126-128, 1960, No. 1).

In the preparation of an antigen from *Hemophilus pertussis* (see No. **2894**) it was found that with some methods the agglutinogen is almost lost from the bacterial surface. Further investigation showed that the agglutinogen is rather soluble in some suspending media but can be fixed on the bacterial surface with formaldehyde in saline.

Now available

T. J. Kroes: Tube and semiconductor selection guide 1960/61, Philips Technical Library, 157 pp.

This book, now available in the third, revised edition, is compiled with the aim of helping the user to find, among the many thousands of types of tubes and semiconductors, the type that is best suited for the purpose he has in mind. It will also be useful to the dealer who has to plan or extend his stocks, and who wants to limit the number of different types as much as possible. The book contains an extensive "interchangeability and replacement list" (50 pp.) i.e. a list of tubes and semiconductors together with the equivalent or nearly equivalent Philips types. Further details are tabulated in seven sections: Tubes for receivers and amplifiers; Cathode-ray tubes; Transmitting tubes; Tubes for microwave equipment; Industrial tubes; Miscellaneous; Semiconductors.

The text accompanying the tables is in English. The book begins with an alphabetical list of the terms used, with their translations in French, German and Spanish, and explanatory introductions in many languages.

K. Hinkel: Magnetrons, Philips Technical Library, 1961, 93 pp., 55 figures.

Experience has shown that many people who work with magnetrons and microwaves have only a vague idea of how these tubes work, and all the uses to which they can be put. This book has been written to explain the operation of these tubes to such people, and to set out the physical principles on which the operation is based. The book is also suitable for students. The six chapters are entitled: Introduction; The electrical mechanism; The circuit; Conditions for oscillation; Examples of practical delay lines and cathodes; The characteristics of magnetrons.

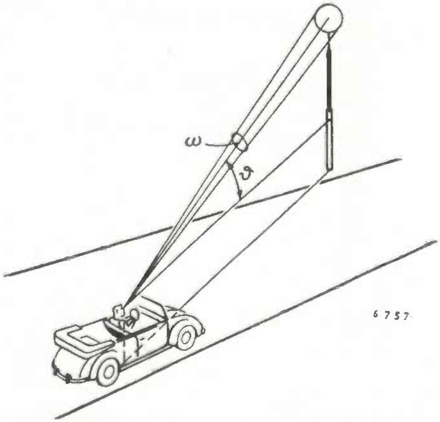
Spanish, German and French editions of this book will also appear.

ERRATA

1) In the formula at the bottom of the left-hand column of page 238, the numerator and the denominator have been exchanged. The correct formula is thus:

$$\frac{P_{\mathrm{WX}_n}}{P_{\mathrm{W}}(P_{\mathrm{X}})^n} = K\,,$$

2) In the right-hand column of page 261, the sketch belonging to fig. 5 has been omitted. It is printed below and the reader is requested to stick it in its proper place.



VOLUME 23, 1961/62, No. 8/9 pp. 221-292 Published 27th June 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

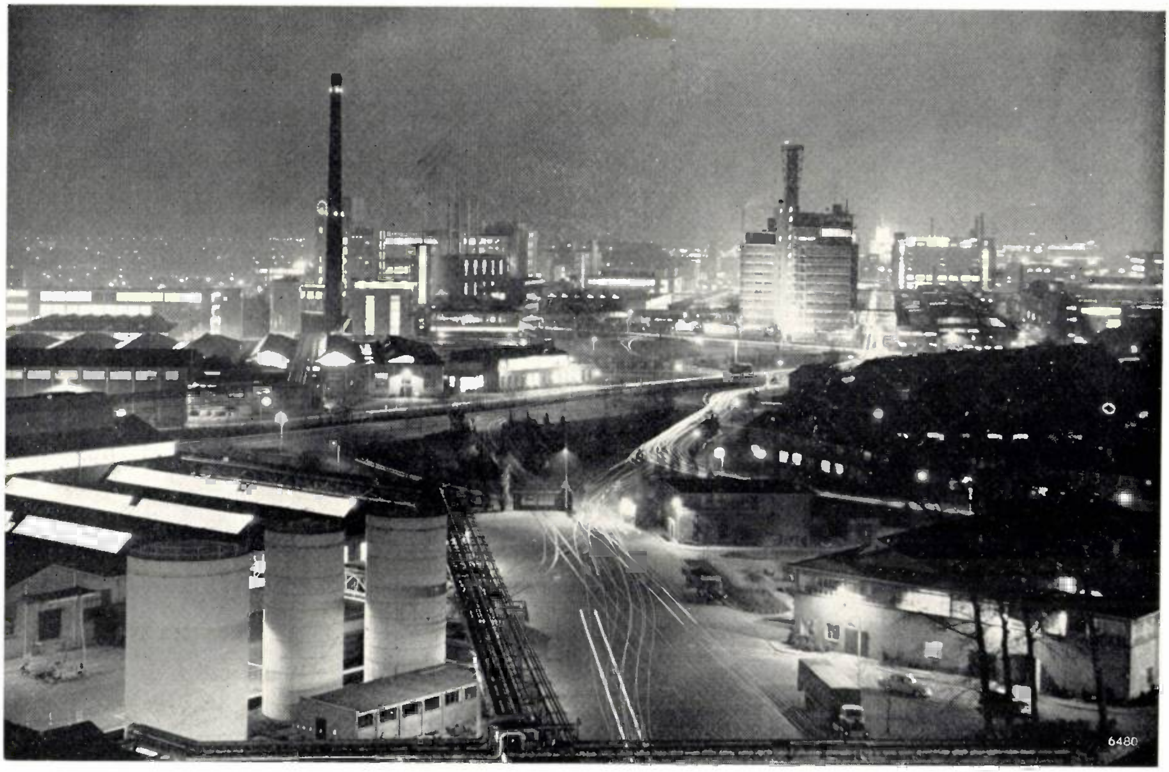


Photo M. Broomfield

Electric lamps were the first product of the Philips factories, and for nearly thirty years their only one. Although since then the research and production programme of the enterprise has expanded into numerous other fields, lamps and lighting have remained a corner-stone of Philips' activity. Some aspects of that activity are discussed in this number of our journal. The incandescent lamp, a product whose potentialities one would have thought had been exhaustively studied, is now seen to be open to further development, radically affecting the whole structure of the lamp. The sodium lamp too, whose construction had settled into a standard mould for many years, is again making headway as a result of fresh technical development; luminous efficiencies of 130 lumens per watt (disregarding ballast losses) or more can now be achieved with these lamps. Various modern light sources are reviewed in an article dealing with the present situation of road lighting. In the development of control gear (ballast) — a side-line of lamp manufacture — use is being made of "solid-state thyratrons", thus widening the useful scope of tubular fluorescent lamps. The concluding article deals with an example of the auxiliary equipment employed in the development of new types of lamp: a device is described which automatically plots the isocandela diagrams of beamed light sources.

This series of articles on episodes in the present work of Philips in the field of lighting is prefaced by an account of episodes from the period when it all started: a short history of the foundation of Philips' factories in 1891, documented by a selection of letters written by the founder, Gerard Philips.

THE BIRTH OF A LAMP FACTORY IN 1891

by N. A. HALBERTSMA *).

621.326.6(091)

The writing of history depends on a certain degree of good fortune. Anyone taking a momentous initiative, a man obsessed with an idea and struggling to give it form and substance, does not generally feel the urge to set down his motives on paper or to describe the conditions in which he embarks on his work, revealing the opportunities, desiderata and limitations of his day and age. Consequently the historian, who obviously cannot be content with extracting bare facts and relevant data from official documents, but who seeks to reconstruct motives and circumstances, usually has to rely on implicit information from many and various sources, in particular from contemporary records that chance to have been preserved.

A rich source of such information, bearing on the foundation of the Philips factories, was found in a letter-book containing copies of correspondence addressed to numerous people by the founder, Gerard Philips, in the period from April 1889 to April 1892. The book itself was destroyed in the bombing of the Eindhoven factories in 1942. Shortly before, however, typed copies had been made of all the letters in it that were still at all legible, the enterprise having by then already attained an age and size at which it could suitably meditate on its beginnings.

An attempt is made here, drawing on those early letters, and also of course on the fund of general data available on those years, obtained in part from books (see the references given at the end of the article), to present a broad picture of the circumstances in which a lamp works, the Philips factory, came into being in 1891. Some of the letters, which are interesting as historical documents, are reproduced *in extenso*.

The dawn of electrical engineering

The evolution of engineering since the 18th century has been based to a considerable extent on the exploitation of natural sources of energy (steam power, water power, gas, oil, etc.) and on the use of such energy sources which was made possible by the introduction of electricity. Steam and water power set the wheels of industry turning. Oil, gas and electricity, on the other hand, were used in the first place for *lighting*.

Oil lamps were of course known in antiquity. At the end of the 18th and the beginning of the 19th century a variety of refined forms came on the scene. In 1813 gas light made its appearance in the streets of London and Paris, and Berlin followed in 1826. Gas lighting in streets and houses was made possible by the *distribution* of the gas from the gasworks through a network of pipes all over the town. In London some 25 miles of gas mains had been laid by the end of 1815.

The fundamental discoveries relating to the generation of electricity and the production of electric light had already been made by this time. In about 1800 Davy had observed the intense light of the electric arc between two carbon electrodes, and in the first quarter of the 19th century there were repeated demonstrations of the light produced when a platinum wire is heated by an electric current. But as long as the current had to be supplied by a few hundred or thousand copper-zinc cells there could be no question of any practical application. Not until the second half of the 19th century, after the invention and perfection of the "dynamo-electric machine" (by Nollet, Holmes, Gramme, Siemens, etc.) was it possible to consider making use of the wide possibilities of the electric arc, with its great brightness, as a practical light source. Arc lamps were successfully introduced in lighthouses (for the first time in 1862 in the South Foreland Lighthouse), in military searchlights and for stage lighting.

In all these cases an electrical generator was used in conjunction with one single light source. The idea of *distributing* electricity in the same way as gas — i.e. to feed a number of dispersed public or private light sources from one generator — lay ready to hand, but was not so easy to realize as might now be thought. Jablochkoff took the first step in this direction in Paris, where he lit the Grands Magasins du Louvre in 1877 and the Avenue de l'Opéra and other streets in 1878 with large numbers of series-wired arc lamps of the kind he himself invented (the "Jablochkoff candle"). But his system proved to be far from ideal; the light fluctuated too much, and the use of arc lamps in series was certainly not a feasible method of domestic lighting.

*) Previously with N.V. Philips, Eindhoven, and emeritus professor of illuminating engineering at Utrecht.

In 1879 at Munich a new journal appeared under the name "Zeitschrift für angewandte Elektrizitätslehre", i.e. Journal of Applied Electrical Theory (or, as we would say nowadays, Journal of Electrical Engineering). This term was first used at that time by Werner Siemens in his proposal to found an association of workers in this field. He expected that such an association would help a large and solid structure to grow up on the foundations of the "applied electrical theory" already present. The association was indeed founded (the Elektrotechnische Verein Berlin, which merged with the Verband Deutscher Elektrotechniker in 1893), but before it could make any further progress in the application of electrical theory, one man — Thomas Alva Edison — startled the world towards the end of 1879 with a complete solution to the problem of "the distribution of electric light".

Edison's system of electric lighting

Edison had taken a comprehensive view of the problem and tackled all its aspects at the same time. Understandably he made no effort to improve the arc lamp, the smallest unit of which was still too powerful for domestic lighting and whose use involved so many complications. Instead he concentrated his attention on the *incandescent lamp*. He was not deterred by prevailing doubts, as e.g. expressed in the pronouncement of the British physicist, Sylvanus Thompson, in 1878 that "any system depending on incandescence will fail". After years of experiments, initially with filaments of platinum and later with carbonized bamboo fibres, with which Göbel had already experimented in 1846, he finally succeeded in making lamps possessing reproducible properties: a luminous efficiency of 0.2 candles per watt, a life [1]) of 200-300 hours, and a luminous intensity of either 10 or 16 candles. But this was only one side of Edison's work. The other was the detailed elaboration of the system of *parallel wiring* (in which direction, incidentally, Brush had already taken a tentative step at the end of 1878 in the arc-lighting of Wanamaker's Store at Philadelphia). Edison had realized that parallel wiring, whcre the current and not the voltage is sub-divided, was the only practical system in which lamps or other current-consuming devices could be separately switched on and off and which could thus lead to successful distribution. Edison's lamps were adapted to this system, having long thin filament wires to give them a *high* electrical resistance. This essential idea distinguished his lamps from those of Swan and others, who were still working on the idea of series wiring, and was set down by Edison in his most important patent, lodged on 10th Nov. 1879 ("... electric lamps giving light by incandescence, which lamps shall have a high resistance so as to allow of the practical sub-division of the electric light ..."). Under Edison's scheme the current from an electricity plant, where several dynamos, according to requirements, could be connected in parallel, would be carried by insulated copper rods, tubes or cables laid in the streets, and thence conducted by thinner branch lines into factories, shops and houses, just as in the case of the gas mains from the gasworks. Edison had to design every component of the installations required for this scheme. The lamp cap (or base) by which millions of electric lamps are still today fixed in their holders is the Edison screw cap. He invented the fuse, which automatically breaks the circuit in any branch of the network where serious overloads or short-circuiting occur. He improved the magnetic circuit of dynamos to such an extent that their efficiency rose from about 50% to 90% and more. There was also an electricity meter, which measured consumption by weighing electrolytic copper deposits, among the series of inventions with the aid of which Edison at the end of 1879 was able to present the world with a serviceable system of electric lighting.

While Edison was pushing ahead in the United States with the manufacture of all these components for his installations, he attracted attention to his system in Europe by a spectacular exhibit at the first International Electrical Exhibition at Paris in 1881 (where incidentally several other electric lamp manufacturers including Swan, Lane-Fox, Maxim and Siemens — see *fig. 1* — had their products on display). Edison demonstrated there a lighting system comprising 1000 lamps of 16 candle-power each, fed by his first large dynamo which, with its associated steam engine, weighed 25 tons and delivered 70 kW. His lamps at that time were already in mass production — in the first 15 months he had sold 80 000 at an initial price of about 3 dollars each. On 4th Sept. 1882 Edison started in New York to supply electricity on a large scale through the Pearl Street power station, first with one and later with six of the large aggregates mentioned. In 1890 this station, which had become world-famous, was destroyed by fire as a result of a short-circuit in

[1]) The prccise meaning of these values is doubtful, since the concept "life" had not then been properly defined. In later years the "useful life" of carbon-filament lamps was specified as the number of hours the lamp burned until its light output had dropped to 80% of the initial value (owing to blackening of the bulb). The lamp would often burn for very much longer before its filament gave up.

the cables. Meanwhile, however, Edison Electric Illumination Companies had begun to supply electricity in many other large towns in the United States. In the years between, electric lamps had dropped in price to between 15 and 25 cents, their life had increased to 450 hours, and their luminous efficiency to 0.3 candles per watt. These were still the standard properties of carbon filament lamps in 1913 [2]).

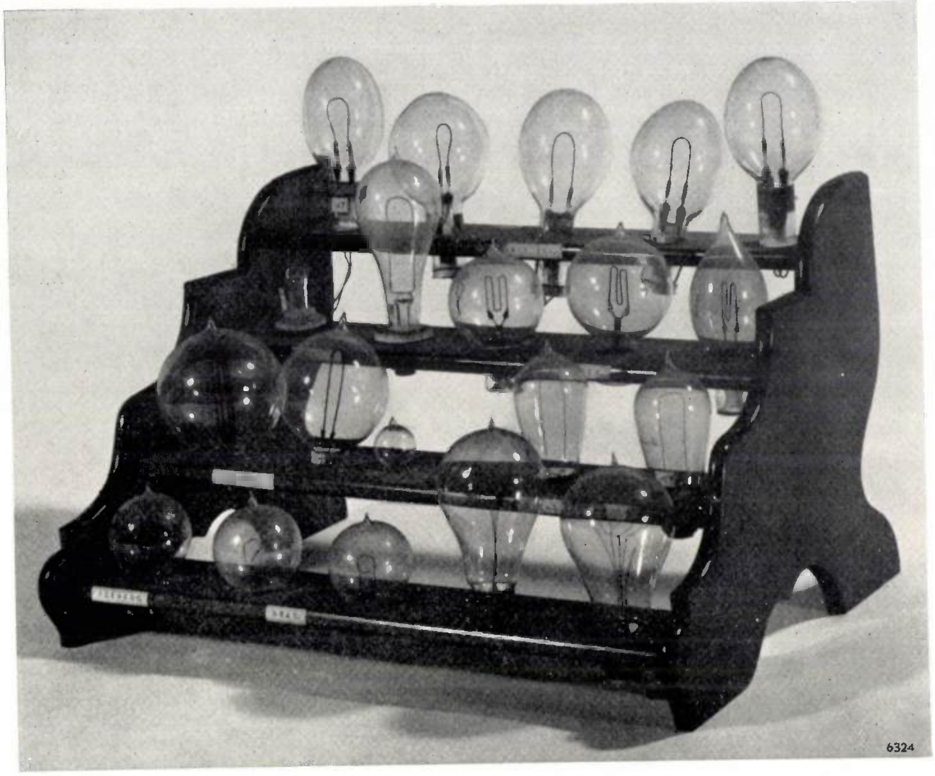


Fig. 1. Collection of 20 incandescent electric lamps, bought by the Teyler Foundation in 1881 at the Paris Exhibition, and including lamps made by Edison, Swan, Lane-Fox, Maxim and Siemens. (Reproduced by courtesy of Teyler's Museum, Haarlem.)

Examples of various types of incandescent electric lamps produced by different manufacturers in the first ten years are shown in *fig. 2*.

The development of electric lighting in Europe, more particularly in the Netherlands

After the Paris Exhibition of 1881 electrical engineering made vigorous headway in Europe too. Edison designed the first electric power station for London (Holborn) in 1882. Emil Rathenau, the German industrialist, acquired Edison's patents and in 1884 founded the "Deutsche Edison Gesellschaft für Angewandte Elektrizität", which later became the "Allgemeine Elektrizitätsgesellschaft" (A.E.G.). The existing firm of Siemens & Halske, specialists in telegraphic equipment (which they had started producing in 1847) embarked on the manufacture of dynamos and electric motors. Schuckert, Bergmann and Kolben, who had helped Edison in the development of installation equipment, went into production in Europe. Wherever electric mains had been laid, arc lamps too stood a better chance where larger light sources than the 16-candle-power incandescent lamps were needed. In Germany factories thus sprang up for making arc lamps and carbons (Schuckert), electricity meters (Aron), switches and lamp holders (Staudt and Voigt), and so on.

Elsewhere in Europe, notably in Great Britain, France, Switzerland and Italy, central power stations built on the Edison system paved the way to the distribution of electricity in the larger towns.

[2]) See A. Wilke, Die Elektrizität, 6th Edn. (edited by W. Hechler), Spamer, Leipzig 1914. — Lamps containing a "metallized" filament delivered 0.4 candles per watt.

A rough idea of the situation of electric lighting in 1891 is given by the following table of the numbers of lamps then in use:

	Incandescent lamps	Arc lamps
United States	2 800 000	23 500
London	600 000 *)	?
Paris	118 000	6 800
Berlin	70 000	3 000

*) A central power station was under construction for this number of lamps.

Although this brief review must naturally remain pitifully incomplete, let us also take stock very briefly of the development of electrical engineering in other fields than lighting. The large-scale use of electricity for chemical processes had long been under consideration. In 1789 Deiman and Paets van Troostwijk had discovered the electrolysis of water [3]), in 1807 Davy had prepared sodium and potassium by an electrolytic method, and in 1854 Bunsen had made aluminium in this way. By about 1840 the technique of electroplating, developed by several researchers, was in use in industry. The application of electricity for transmitting and distributing mechanical energy also began to come to the fore in the 'eighties. Here, however, the situation was still uncertain. Attempts were also being made to meet local requirements of mechanical energy by means of small prime movers, e.g. small hot-air engines. And even as regards the possibility of *centralized* power supply, electricity had its competitors: in Paris at that time, Popp built an installation for supplying compressed air to users all over the city through a network of pipelines.

A milestone was reached in the transmission of mechanical energy by electricity when A.E.G. and the Oerlikon Engineering Works demonstrated the transmission of 100 HP by means of three-phase alternating current at the 1891 International Electrical Exhibition at Frankfurt-on-Main. The generator was 110 miles away at Lauffen on the Neckar. This ushered in the era of alternating current, and the D.C. system adopted by Edison was gradually eclipsed.

To conclude this introduction let us take a look at events in the Netherlands.

In 1878 Willem Smit, whose name is still connected with three major electrical firms in the Netherlands, built his first dynamo at the age of eighteen, after having seen in the Leygraaf Hotel at Rotterdam an arc lamp run from a dynamo built by Gramme. With his own dynamo, and a Hefner-Alteneck arc lamp, Smit was able to light his father's rivet factory at Slikkerveer. This aroused such interest among neighbouring factory-owners that he received orders for similar installations.

In 1882 this pioneer of electrical engineering in the Netherlands, together with his brother-in-law Adriaan Pot, founded a company for producing electric lighting equipment and trading in electrical appliances and related articles. The new enterprise received orders for lighting ships with parallel-wired incandescent electric lamps, and it hired out installations for public festivities, exhibitions, etc.

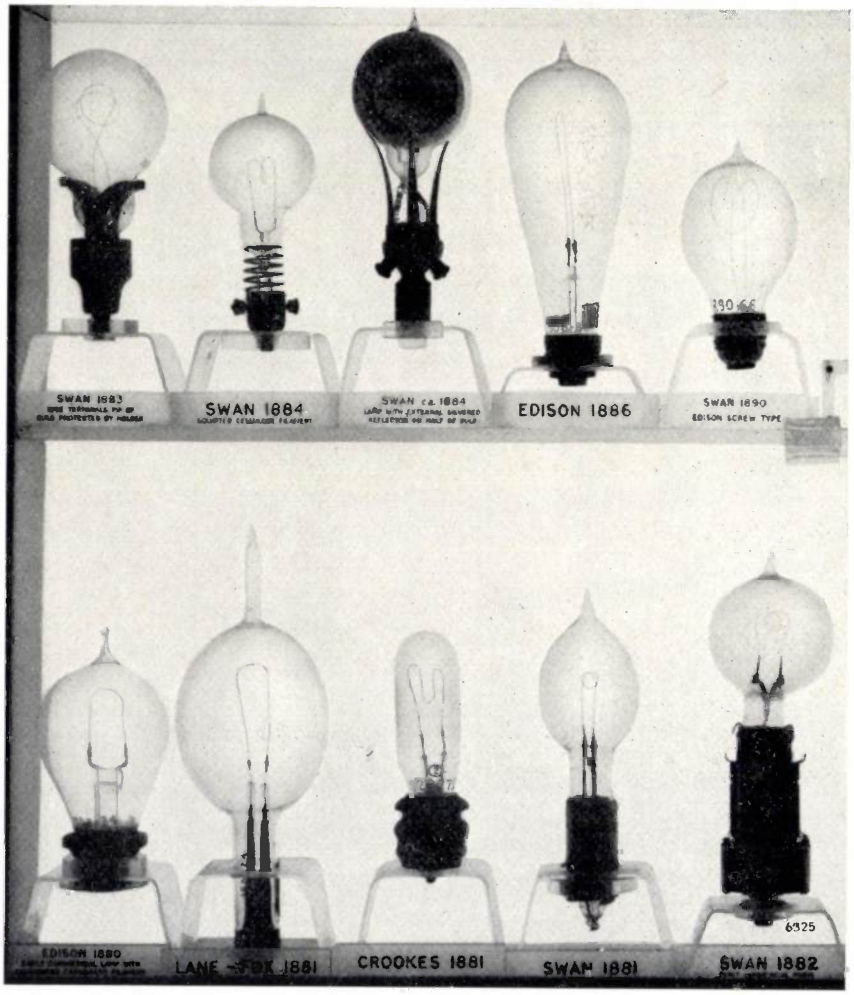


Fig. 2. Electric lamps of types made between 1880 and 1890, in chronological order. (By courtesy of the Science Museum, London.)

[3]) J. R. Deiman and A. Paets van Troostwijk, Beschrijving van eene Elektrizeer-Machine en van Proefnemingen met dezelve in het werk gesteld (Description of an Electricity Machine and of experiments performed with the same), Amsterdam 1789.

Even the lighting system installed in the Coomans Hotel at Rotterdam (1884), comprising one arc lamp of 60 A and 120 Swan incandescent lamps of 35 candle power, was on hire. The hotel paid a monthly rent corresponding to the amount it would otherwise have spent on gas lighting.

In 1886, the power station of Willem Smit & Co. — the first in the Netherlands — was put into operation at Kinderdijk. The station started with two DC dynamos of 110 V, 200 A, and later an alternating-current generator was added for lighting the neighbouring village of Krimpen via a cable under the river Lek. The station operated on very simple lines. At midnight the lights went out, but if there was a celebration somewhere a few hours' extension could be obtained "if requested in good time". The tariff charged was based on the number of lamps installed, and amounted to 2.50 guilders monthly for every lamp of 35 candle-power.

Other details that throw light on the situation in the Netherlands and Europe in the years that Gerard Philips' plans matured will emerge from the following account of his life.

Gerard Philips' first steps in electrical engineering

Gerard L. F. Philips, born in 1858, was the elder son of B. F. D. Philips, a manufacturer, banker and owner of the gas-works at Zaltbommel. From 1876 he studied first civil engineering and later mechanical engineering at what was then called the "Polytechnische School" at Delft. He showed keen interest in the developments of applied electricity, both in the Netherlands and abroad. While he was still at secondary school in Arnhem he had attended a course of lectures and demonstrations on this subject at the "Wessel Knoops" Physical Society. The lecturer, who held him enthralled, was a young teacher of the name of Dr. H. A. Lorentz — the later renowned physicist and Nobel Prize winner.

At Delft the new subject was not yet on the curriculum [4]), and when Gerard, after graduating in mechanical engineering, decided to apply himself to electrical engineering, his obvious course was to try to acquire more knowledge through practical experience. In 1883 he therefore set out as a young engineer for Glasgow, where he supervised the installation of an electric lighting system on the S. S. Prins Willem van Oranje for the Zeeland Shipping Company. After that he worked for nearly a year in the laboratory directed by the distinguished physicist William Thomson, later Lord Kelvin, whom he assisted in the development of electrical measuring instruments.

He was then invited by the managing director of the Brush Electrical Company, manufacturers of dynamos and arc lamps, to go to Berlin and prospect the market there for the Brush products. It is not surprising, in view of the strong position which the German firms A.E.G. and Siemens occupied on their home market, that this assignment yielded few positive results.

Back in England, Gerard sat in 1887 the examination of the City and Guilds of London Institute in "Electric lighting and transmission of power and telegraphy", and was awarded first prize, a silver medal (*fig. 3*). He then settled in London for a time as the representative of several German electrical firms.

The next episode in the life of Gerard Philips, and the experience it gave him, undoubtedly did much to form in his mind the plan of making electric lamps in the Netherlands. Emil Rathenau, the managing director of A.E.G., appointed him the agent of that firm in Amsterdam. His assignment went further than simply procuring orders for electrical machines and installations. The municipal corporation of Amsterdam intended to grant a concession for the building and operation of a power station, and Rathenau was particularly keen to obtain that concession. Gerard had to win over both the corporation and leading figures in financial circles for Rathenau's tender. It was no easy task, for on the one hand there were differences of opinion

Letter I (letter-book p. 146).
(Translated from the original German.)

Hotel Mille Colonnes, Amsterdam. 24.9.1890

Herrn Direktor E. Rathenau,
A. E. G.,
Berlin-N.

The proposal of the Municipal Executive to grant a concession to "Electra" was approved at today's meeting of the City Council.

This therefore concludes my efforts to procure this concession, and not in the way I had hoped.

Several councillors, members of the Public Works Committee, informed me confidentially that they really regretted having to grant a concession to "Electra", and that if your price had been not more than a few cents higher than that quoted by "Electra" they would certainly have given you preference. The price difference now, however, was too (considerable *)).

Yours faithfully,

G. L. F. Philips

*) Not clearly legible.

[4]) The first course of lectures on electrical engineering was given at the Technische Hochschule at Darmstadt in the winter term of 1882/1883, by Professor Dr. E. Kittler.

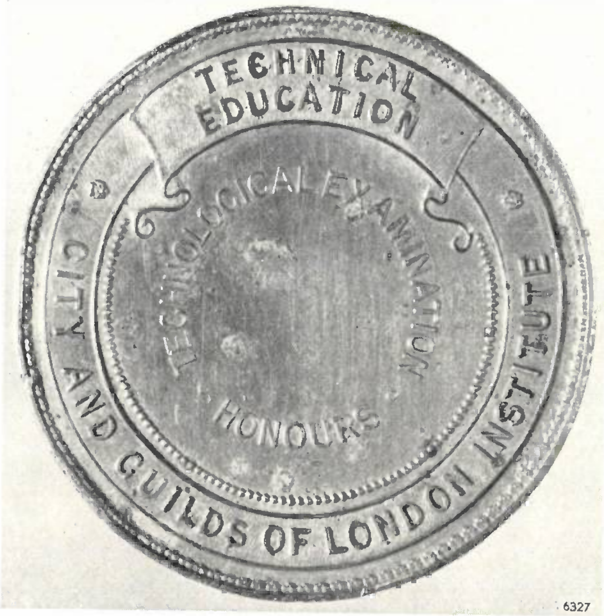

TECHNICAL EDUCATION
TECHNOLOGICAL EXAMINATION
HONOURS
CITY AND GUILDS OF LONDON INSTITUTE
6327

Fig. 3. Silver medal awarded to Gerard Philips in 1887 for passing an examination of the City and Guilds of London Institute with honours.

in the corporation about the form and terms of the concession and about the relative merits of municipal control and private enterprise, and on the other hand Rathenau put his demands very high as regards the electricity rate. He asked for 60 cents per kWh, and in letters to Berlin — which were preserved in the letter-book referred to — Gerard fought stubbornly to persuade Rathenau that he ought to reduce this price. Rathenau, however, refused to yield. When the concession was dealt with by the City Council, they decided to grant it to a competitor who was prepared to supply the electricity at 46 cents per kWh (see *letter I*, of 24th September 1890).

Soon afterwards Gerard's work for the A.E.G. came to an end.

Prelude to the foundation of the factory

Gerard Philips was now 32 years of age, and it was time to decide how and where he was to earn his living by electrical engineering in a way that would give him personal satisfaction. His difficult assignment under A. E. G. must have strengthened his self-confidence and also have aroused the desire to be independent and to carry the full responsibility for an enterprise himself. Would not the Netherlands, in view of the rapid advancement of electrical engineering, soon need an electrical industry of its own? And would there not in the first place be a demand for the production of electric lamps, which have a limited life and therefore called for constant replacements?

Gerard Philips was not the only one to whom this idea occurred; the optimistic expectations for electric light were too widespread for that. In 1890 four electric lamp factories were already established in the Netherlands, but the correspondence shows that Gerard regarded only one of them as a serious competitor, namely Goossens & Pope, a factory founded in March 1889 at Venlo by the Englishman Pope. The other three: Boudewijnse at Middelburg, De Khotinsky at Rotterdam and Roothaan & Alewijnse of Nijmegen, soon succumbed to the same difficulties that Gerard Philips was later to come up against — and to overcome.

At the Brush Electrical Company, Gerard had gained some experience of the manufacture of incandescent lamps, and he knew that the most important factor was the uniformity of the cellulose carbon filament. His first work towards the realization of his aim (which he started before he left the employment of A.E.G.) was therefore in the chemical field. He equipped a wash-house in his parents' garden at Zaltbommel as a "laboratory" (*fig. 4*) and as soon as he thought he had succeeded in producing a carbon filament of good quality he started advertising in the "Elektrotechnische Zeitschrift" and the "Electrician" for a works manager.

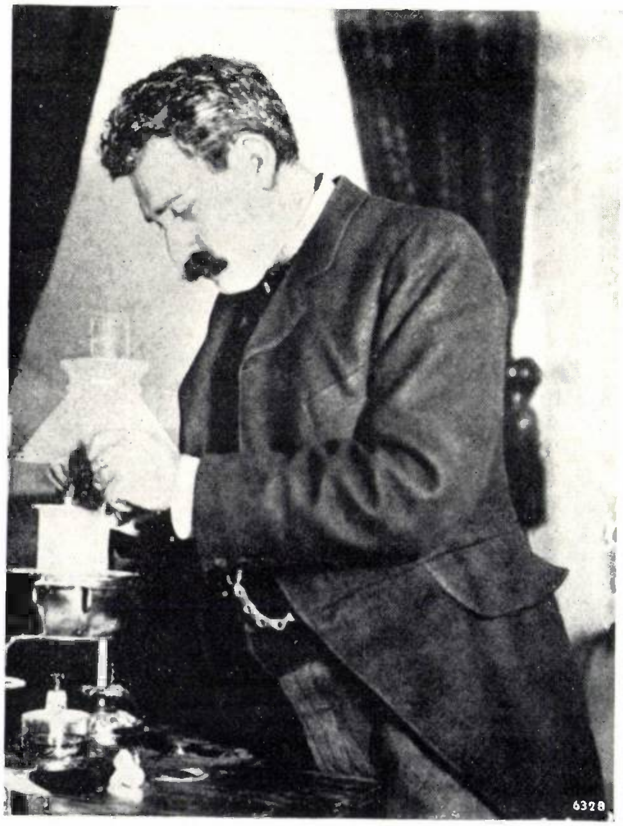


Fig. 4. Gerard Philips experimenting in his home-made laboratory at Zaltbommel in 1890.

Letter II (letter-book pp. 149/150).

Hotel Mille Colonnes, Amsterdam. . . *) November 1890

C. J. Robertson Esq.,
Middelburg.

Sir,
I duly received your favour of the 6th inst. I still fail to see any decrepancy between my advertisement and conversation, in fact, I am inclined to think, that, if any, the want of agreement is on your side.

It is perfectly true that our plans are not yet settled but we thought that your knowledge of Dutch lamp making works and prices would assist us herein, and that at the same time, we would be able to enter into an agreement with you. We turned to you because you have a knowledge of the country and the language, and for no other reason, for, as I frankly told you, Mr. Boudewijnse's lamps hardly enjoy as good a reputation overhere as those of several other Dutch and German makes.

We have received quite a large number of applications where amongst from gentlemen who are very well qualified.

It is our intention to start a factory of 1000 lamps daily output, but we would at first make only 500 lamps a day, making the building and engine large enough for 1000 lamps. For a factory of this size we can supply the required capital ourselves, and we estimate this, when the output of 1000 lamps daily has been reached, at F. 75.000.— viz.:

total cost of building and plant	F. 37.000.—
total working capital	F. 38.000.—.

An increase of the working capital, above this cipher, would not . . . *)

As advertised, we want to contract with an electrician (or as I would now put it, after our recent conversation, an electrical engineer), able to plan and fit up the factory and to manage the works. We are prepared to pay him, in addition to a fixed salary, a certain part of the net profits, but we are quite willing to consider any other basis of agreement.

I am, however, aware that it offers important advantages to start at once on a large scale, say with a daily output of 2500 lamps. Since we have very good connections, it is quite possible that we may be able to form a syndicate for this purpose. In such case it would be advisible to allot the expert a certain number of shares in the Company to . . . *) for his goodwill and his patents or what is equivalent to this. This, however, must be done by special separate contract since the Dutch laws don't allow such seemingly gratuitous allotments to be incorporated with the articles of association.

But in order to be able to form a defined opinion of the relative advantages of the manufacture of lamps on different scales, it would interest us to have your calculation of the rentability of the lamp making business for an output of:

1) 1000 lamps daily (at first only 500 daily but building an engine large enough for 1000),
2) 2500 lamps daily.

We assume that lamps of all current C.P. (up to 100 C.P.) and E.M.F. have to be made, and we premise certain unit prices of materials, coals and labour. In the first case I would leave out any remuneration of the partners, accept any salary to be claimed by you in accordance with . . . *) proposals. In the second case a remuneration of the management in accordance to your own views or proposals should be drawn into account.

If we start a lamp factory of 500-1000 lamps daily on our own account, we will probably chose Breda, where we can buy a plot of land for F. 1000.—. The wages there are about as in Middelburg.

I shall be pleased to know your terms and I hope there is no objection on your part to supply an estimate on the lines above mentioned. In such case I could hardly ask you to do so at your earliest convenience.

Yours truly,
G. L. F. Philips

Letter III (letter-book p. 154).

Hotel Mille Colonnes, Amsterdam. 14.11.1890

M. E. Bailey Esq.,
London, E. C.

Dear Sir,
I duly received your application of Oct. 20, in answer to my advertisement in the Electrician of October 17, address 943 Electrician's Office.

You will understand that I have received a large number of applications both from England and the Continent, and I am now in correspondence with several lampmaking electrical engineers on this matter.

I intend to start an incand. lampfactory in Holland, with a daily output of 1000 lamps, starting however with less, but working up to this number.

I am able to produce extremely homogeneous and equal cellulose filaments on a business scale, and I am not so much in want of special details of manufacture, patented or not, of which, of course, any man, engaged in lamp-making, has quite a number to himself; what we want is a man quite competent to carry out and work a modern lamp factory on thorough business principles, in other words to make first-class lamps on a sound business scale in as cheap and efficient a way as possible.

In your letter you state that you were able to make lamps for four pence three farthings each. How do you understand this price? Are you prepared to say — and eventually to guarantee — that this is the net price inclusive of (London) labour, salaries, rent, amortisation, etc.? Today, of course, allowance has been made for the higher prices of platinum. You will excuse my feeding some doubt, and more so since you work (?) with platinum and mercury glass airpumps, two things, which in the States they are trying to do away with.

I shall however be glad to receive some information on these points and to know whereon your data are based. And also what the cost of establishment of a factory for 1000 lamps daily would be on your plans. I shall also be glad to know your terms for an engagement.

I can give you many references with regard to myself.

I shall have to run over to London before long and would seize the opportunity to call upon you, if you are able to give me reliable data and facts concerning your ways of lampmaking.

I hope to receive your reply and proposals at an early date.

Yours truly,
G. L. F. Philips

This appears from several of his letters addressed to applicants for the post, and his letters also show that he was contemplating a factory with a final output of 1000 lamps daily. After abortive negotiations with an English candidate, who demanded co-partnership in the firm (*letter II* of about 9th November 1890), and with yet another Englishman (*letter III* dated 14th November 1890), he entered into more decisive arrangements with the "production chief" or foreman of a small lamp factory at Brussels, which had had to close down. In a letter dated 30th December 1890 Gerard invited this candidate, E. Woschke, to come to Zaltbommel for discussions (*letter IV*). It was evidently his intention — as in previous contacts — to test his ideas on production against those of people with practical experience The letter in question also reveals that J. J. Rees

*) Illegible.

Letter IV (letter-book p. 163).
(Translated from the original German.)

Hotel Mille Colonnes, Amsterdam. 30.12.90

Herrn E. Woschke,
Brüssel.

I duly received your favour of the 10th inst. but was unable to let you know anything definite earlier. I am at present very busy working out various details of lamp making. First I am now making a clean and chemically pure filament. Then, as regards the fabrication of . . . . *) there are several details which do not appeal to me at all. Like most lamp makers I want to make the connection between the carbon and platinum wires electrically; this makes much faster evacuation necessary. The Seel pumps take an abnormally long time; I had the opportunity recently to inspect and make sketches of the pumps in the Roothaan and Alewijnse factory. These take less than half an hour . . . . *) in Venlo evacuation takes only half an hour. This means that considerably less machine power is needed than in the Seel factory. On this part I have exact specifications for the Nijmegen and Venlo factories. I also intend to use accumulators, etc.

I shall not run the enterprise with Mr. Reese, but with my father's assistance. Of course, quite differently from the Seel people, as economically as possible, with no nonsense, and with the least possible loss at the start.

I should now like to have a decisive talk with you, in my father's presence. You may be sure that you are dealing with decent people. My father would like you to come to his house at Zaltbommel. Any day suits us, and of course we shall pay your expenses.

If you take the 6.27 a.m. train from Brussels, you travel via Antwerp and arrive at 8.27 in Roosendaal, at the Dutch customs. You change there and take the 8.50 to 's-Hertogenbosch, arriving at 10.02. You wait there for the 10.33 train to Zaltbommel or Bommel, arriving at 10.51 in the morning. Altogether, then, only 4½ hours. You will find an omnibus at the station.

I hope to hear from you soon what day you will be coming. You can leave again in the evening at 6.37 and will be back in Brussels at 11.14.

Yours truly,

G. L. F. Philips

*) Illegible.

of Amsterdam, an acquaintance of Gerard's who was to be a partner in the projected enterprise, had dropped out, but that Gerard could now count on financial support from his father.

The discussion at Zaltbommel that took place shortly afterwards led to agreement. With effect from 1st January 1891, Woschke was engaged at a monthly salary of 120 guilders. Pending a decision on the place of establishment and the purchase either of a plot of land or, even better, of existing premises, Woschke remained in Brussels where he received a sort of retaining wage.

One of the places Gerard had in mind as a likely place of establishment was Breda (see letter II), where a plot of land was available for 1000 guilders and where wages were apparently not very high. In February 1891, however, he came upon the

Letter V (letter-book pp. 168/169).
(Translated from the original German.)

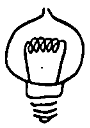
Hotel Mille Colonnes, Amsterdam. 28.2.1891

Herrn E. Woschke,
Brüssel.

I duly received your letters of the 10th and 26th inst. and thank you for your answers to my various questions. I still do not need you at present. I am not nearly as far as you seem to think, and you will certainly help in setting up the factory, when I can make good use of your services. For various reasons matters have been postponed, and I am still negotiating the purchase of existing factory premises. I have the chance of buying a very nice factory, only a few years old, a single-story, rectangular building, equipped with steam engine and boiler; but the building is entirely of brick and iron, and I am now afraid that the iron may influence the instruments; I should like to have your opinion on that point. The factory measures 18 × 20 M. I am off to Zalt Bommel today and shall write to you from there.

I am sorry that I cannot ask you, as I had definitely hoped, to come to Holland yet; the situation where you are is certainly not pleasant, but I could not let you come to Holland before I have rented or purchased a proper building somewhere. That should soon be possible; I hope to be able by the middle of next week to invite you to come here to give your opinion on the building. Apart from the premises mentioned above, I am negotiating for two other buildings in another town, but I would like to rent or buy steam power at the same time.

I now have much further information regarding the Venlo factory. One man there evacuates 500 to 600 lamps daily, i.e. he turns the valves, for there is no question of lifting weights. He receives 9 . . . . *) per 100 lamps. I shall try to become acquainted with these pumps, for you appreciate that this is very different from Seel. The man concerned has one bench of 36 lamps under him; three small or two large lamps on each pump. They are Sprengel pumps. The Venlo lamps are well known, and are much used here; the factory was doubled in size last summer; I know the lamps, and they have spiral carbon filaments that look like this:



That has the effect that the filament is not seen so sharply, the light appears to be more concentrated, more like a gas flame. I can have the man who made these carbons for 12 guilders a week; he is skilled carpenter and a clever fellow.

You'll probably still have some of that uncarbonized wire from Venlo. I should like to have a small piece to compare with my wire. In Venlo the carbon filaments are left all night over white hot . . . . *). They use blast-furnace coke.

As I said, I hope I shall need you next week for deciding on the factory premises. Very soon after that you will be able to move here, but at present, as you will understand, that is not yet possible.

Yours truly,

G. L. F. Philips

P.S.
In Venlo they had two German glass-blowers at the beginning, but not now. Everything today is done by girls. That makes a considerable difference. We shall have to try that ourselves later in order to remain thoroughly competitive. The Venlo factory is a sound business, quite different from Seel. They do good business and we must try to follow their example.

*) Illegible.

vacant buckskin factory of Schroeder & Weijers in Eindhoven (*fig.* 5), which was equipped with a boiler and a 40-HP steam engine. On 28th February 1891 he mentioned this opportunity to Woschke (*letter V*) and on 16th May 1891 he bought the factory for 12 150 guilders.

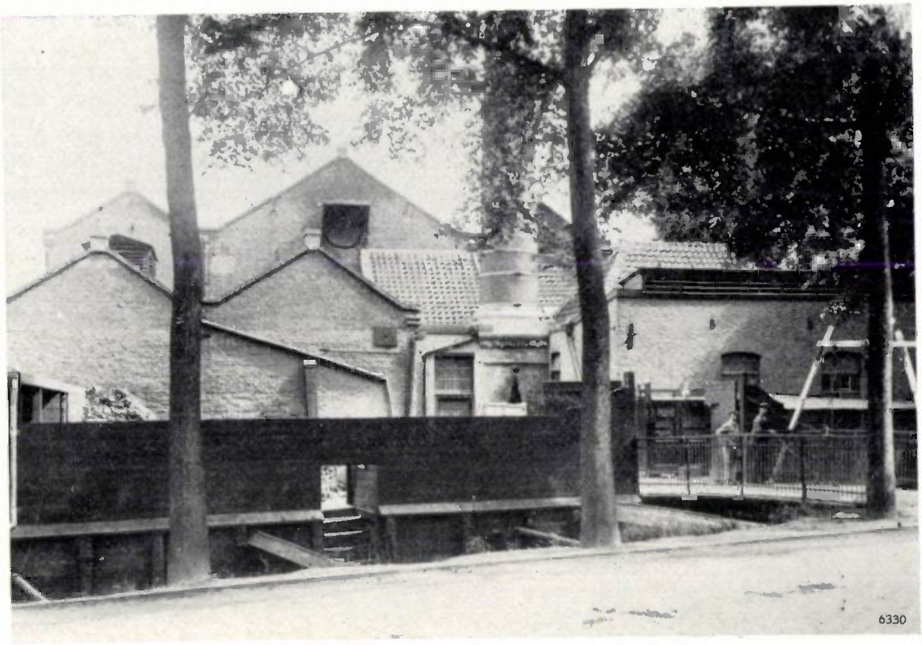


Fig. 5. The factory of Philips & Co. in 1891.

That letter reveals how very busy Gerard had been meanwhile with the technical problems, whose solution he regarded as essential to the success of his enterprise. Apart from fabricating a homogeneous carbon filament, he was particularly concerned with the evacuation of the bulbs and with glass-blowing. At that time lamp manufacture was an industry in which wages represented a relatively high proportion of the costs of production. It was therefore imperative to shorten the production time per lamp and to simplify the operations, so that as much as possible of the work could be done by girls as cheap labour.

The group photograph of the entire personnel taken in the first year of business (see *fig.* 6), shows that Gerard lost no time in putting the latter principle into effect.

The beginnings of Philips & Co.

On 15th May 1891 the firm of Philips & Co. was established at Eindhoven (*fig.* 7) with G. L. F. Philips as the working partner and his father, B. F. D. Philips, as sleeping partner. The capital provided by the latter was 75 000 guilders. Woschke was at last able to move to Eindhoven and Gerard at once set about overhauling the steam installation, which had been idle for more than a year, and to place the necessary orders, first of all for electrical equipment.

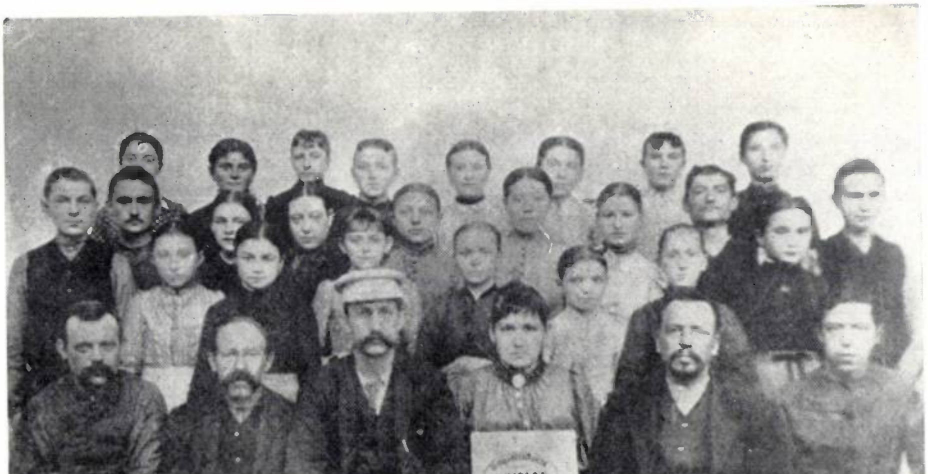
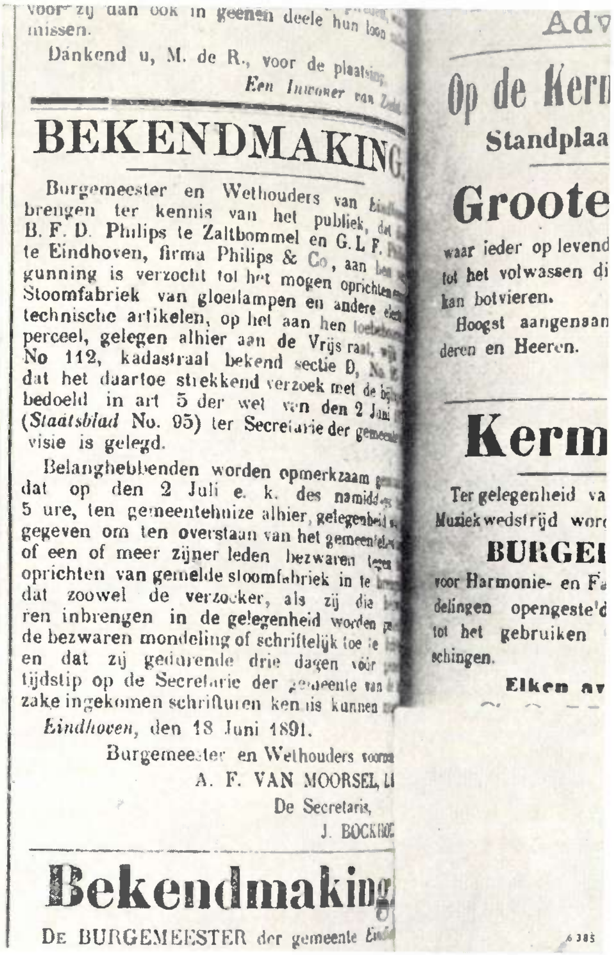


Fig. 6. Group photograph of the entire personnel of Philips & Co. in 1891. Woschke — who was dismissed at the end of 1892 — is the man wearing a cap.

He asked for quotations from several firms, including Mijnssen & Co. of Amsterdam, the Dutch agents of A.E.G. Their quotation arrived on 22nd May, and on 23rd May he sent off an order for three dynamos, one large and two smaller ones (see *letter VI*). He asked for shipment within fourteen days. On 25th May he ordered from Hartmann & Braun of Frankfurt-on-Main a whole series of measuring instruments, including an aperiodic precision voltmeter, four ordinary voltmeters, a portable test voltmeter with two scales, a wide-range ammeter (200 A) and four narrow-range ones (5 A), a "Wasservoltameter", a direct-reading tension galvanometer and a measuring bridge. Gerard gave a reference and held out the prospect of further regular orders, but demanded shipment of the instruments within three weeks. Everything points to the go-ahead lines on which Gerard was determined to build up his enterprise. When it appeared that the A. E. G. needed a longer delivery time for the smaller dynamos, Gerard cancelled that part of his order and promptly turned to other suppliers who, as he wrote, "took more trouble with the smaller machines".



voor zij dan ook in geenen deele hun loon
missen.

Dankend u, M. de R., voor de plaatsing
Een Inwoner van Zuid

BEKENDMAKING.

Burgemeester en Wethouders van Eindh
brengen ter kennis van het publiek, dat
B. F. D. Philips te Zaltbommel en G. L. F. Ph
te Eindhoven, firma Philips & Co, aan hen
gunning is verzocht tot het mogen oprichten
Stoomfabriek van gloeilampen en andere elek
technische artikelen, op het aan hen toebehoo
perceel, gelegen alhier aan de Vrijs raat, wijk
No 112, kadastraal bekend sectie D, No
dat het daartoe strekkend verzoek met de bij
bedoeld in art 5 der wet van den 2 Juni
(*Staatsblad* No. 95) ter Secretarie der gemeente
visie is gelegd.

Belanghebbenden worden opmerkzaam ge
dat op den 2 Juli e. k. des namiddags
5 ure, ten gemeentehuize alhier, gelegenheid
gegeven om ten overstaan van het gemeentebe
of een of meer zijner leden bezwaren tegen
oprichten van gemelde stoomfabriek in te
dat zoowel de verzoeker, als zij die be
ren inbrengen in de gelegenheid worden ge
de bezwaren mondeling of schriftelijk toe te
en dat zij gedurende drie dagen vóór
tijdstip op de Secretarie der gemeente van
zake ingekomen schrifturen kennis kunnen

Eindhoven, den 18 Juni 1891.

Burgemeester en Wethouders voorn
A. F. VAN MOORSEL, L
De Secretaris,
J. BOCKHO

Bekendmaking.

De BURGEMEESTER der gemeente Eind

Adv

Op de Kern
Standplaa

Groote
waar ieder op levend
tot het volwassen di
kan botvieren.
Hoogst aangenaan
deren en Heeren.

Kerm
Ter gelegenheid va
Muziekwedstrijd word
BURGEI
voor Harmonie- en F
delingen opengeste'd
tot het gebruiken
schingen.
Elken av

6385

Fig. 7. Announcement of the application by the firm of Philips & Co. for a licence to "set up a steam factory for incandescent lamps and other electrotechnical articles", dated 18th June 1891 and published in the "Peel- en Kempenbode".

Nowadays a man starting up in business would buy a car. The automobile was still right at the beginning

Letter VI (letter-book p. 170).
(Translated from the original Dutch.)

Philips & Co. 23.5.1891

Den Heeren Mijnssen & Co.,
Amsterdam.

Dear Sirs,

We are in receipt of your favour of 22nd inst., and request you to supply us with the dynamos mentioned below, on condition however that we receive within eight days the necessary drawings with dimensions for laying the foundation and arranging transmission, and further that the dynamos hereby ordered are dispatched to us within fourteen days.

We order:

1 G 200, shunt, for 150 Volts and 170 Amperes, complete with regulator.

1 NG 25, compound, for 300 Volts and 10 Amperes, complete with regulator.

1 S 15, for 200 Volts and 7 Amperes, complete, but without regulator.

All at the terms of payment mentioned in your letter.

The speed of the above machines should not exceed the figures mentioned in the price list; in this connection we request you to send us the correct data together with the drawings referred to.

The compounding of the machine must be very good, as good as can be expected from a first-class compound dynamo.

We trust that you will execute this order to our complete satisfaction.

Yours faithfully,

Philips & Co.

Letter VII (letter-book p. 208).
(Translated from the original Dutch.)

13.7.1891

Den Heer Deumer Cramer,
Utrecht.

Dear Sir,

I have your bicycle back in my possession, but the bell is missing. I am still not satisfied with the repair, however, and I am therefore returning the machine once again. I am not prepared to accept a machine which is defective in one of its main parts. I am paying good money and expect a good machine for it, without the slightest defect. That seems to me to be no more than reasonable. On the whole I cannot say that your machine behaves as a first-class machine should. I know several people here who have not yet had any trouble after long use, and your machine already shows defects after a short trial. The saddle bar has a tendency to twist, and pulls the connecting rod between the bars backwards, so that the frame bars under the saddle bend over backwards; it seems to me that this is due to a weak (or faulty) construction. At the top of the front fork a seam is already noticeable where the paint . . . . *) obviously caused by play in those parts.

I request you therefore to supply me as soon as possible with a well-built machine, without a single defect, as otherwise I shall have to turn elsewhere. I have been asked by various people whether I can recommend the machine, but what am I to answer under these circumstances?

Yours etc.

G. L. F. Philips

*) Illegible.

Letter VIII (letter-book p. 248).
(Translated from the original German.)

30.7.1891

Herrn W. Müller,
Brüssel, N.

We duly received your letter of 27th inst. We are prepared to pay you a wage of Frs. 45 per week for making the new pumps. Incidentally, you may depend on it that you are dealing with decent people. If we are satisfied with you, as we expect to be after what we have heard from Mr. Woschke, your position here will certainly be a very assured one, and we shall pay you a very fair wage. You are the first here, and that has its advantages! In Holland we do not like constant changes of personnel; when we have good people we keep them. Binding arrangements are not entered into here for the good reason that it is unpleasant to have to go on working together when one party is not pleased with the other. We believe, however, going by what Mr. Woschke says, that this will not be the case with us.

Looking forward to your arrival, we remain,

Yours truly,

Philips & Co.

Letter IX (letter-book p. 356).
(Translated from the original German.)

29th October 1891

Herrn H. Jahrke,
Frauenwald.

We have an immediate vacancy for an efficient and steady glass-blower for making and repairing our mercury pumps. We should fix his salary at 20 guilders a week and would pay his travelling expenses. It is essential that he shonld be able to start immediately. We should be very obliged to you if you could help us find the man we want. He must, of course, be of sober and decent habits. If he is satisfactory, his position with us would certainly be a very assured one.

Thanking you in advance for a reply, we remain,

Yours faithfully,

Philips & Co.

Letter X (letter-book p. 315).
(Translated from the original German.)

1st October 1891

Herrn Albert Voss,
Ellrich a/Harz.

For cementing-in our incandescent lamps we require a special plaster of Paris which must be very hard, but should not take too long to harden properly. Nor should the plaster expand too much.

Furthermore we shall soon need a new floor for our mercury-pump department, and may perhaps decide to give your floor plaster a trial.

We therefore request you to supply us with information, prices and samples for both purposes.

How long can the lamp plaster be stored? Please let us have your keenest prices.

Yours faithfully,

Philips & Co.

Reference banker J. H. Stein, Cologne.

of its development in those days (if we disregard the cumbersome steam carriage), but Gerard did order a modern means of transport — a bicycle. The result was not to Gerard's liking, and we can see from his letter to the Utrecht importer who delivered the machine just how angry he could be over slovenliness and unreliability (*letter VII*).

After the electrical equipment he turned his attention to the pumps for evacuating the lamp bulbs. The equipment then available in this line was intended more for physical experiments on a laboratory scale, and was not suitable for evacuating large numbers of lamps simultaneously and quickly. Gerard therefore tried to engage someone with experience in this field, and on 30th July 1891 he wrote to a certain W. Müller in Brussels, who had been recommended by Woschke (*letter VIII*). The letter gives the impression that Müller was very keen on obtaining a permanent post, but Gerard considered that a satisfactory employer-employee relationship was more important than a binding contract. Apparently he was right, for three months later, on 29th October 1891, Gerard was again urgently seeking a man for the same post, this time through a business relation in Germany (*letter IX*). In the letter in question Gerard insists rather significantly that the man should be "steady" and "sober" (presumably in the meaning of not drinking to excess) and we can only guess that this was perhaps the reason for the previous disappointment. As appears from the letters which Gerard wrote on 16th and 21st November to two applicants for the post, this was evidently a stumbling block in Gerard's work. On 29th February 1892 he was again writing to one of these applicants, because the post — meanwhile occupied by someone else — had again fallen vacant.

Fig. 8 gives an idea of what the evacuating system for carbon-filament lamps must have looked like in those days.

The lamp bases — originally made of plaster of Paris, cast in zinc moulds in which the brass screw-thread mantle and the central contact had been placed beforehand — were not made by Gerard himself. The Boudewijnse lamp factory at Middelburg, which had failed as such, had turned to the specialized production of lamp caps, one improvement introduced being the sealing together of the shell and cap contact by means of "Vitrite", an insulating glass compound. On 1st October 1890 Gerard asked for a quotation from "The Vitrite Works" — apparently with good results, for the caps of nearly all Philips incandescent lamps are still made today by the Middelburg factory.

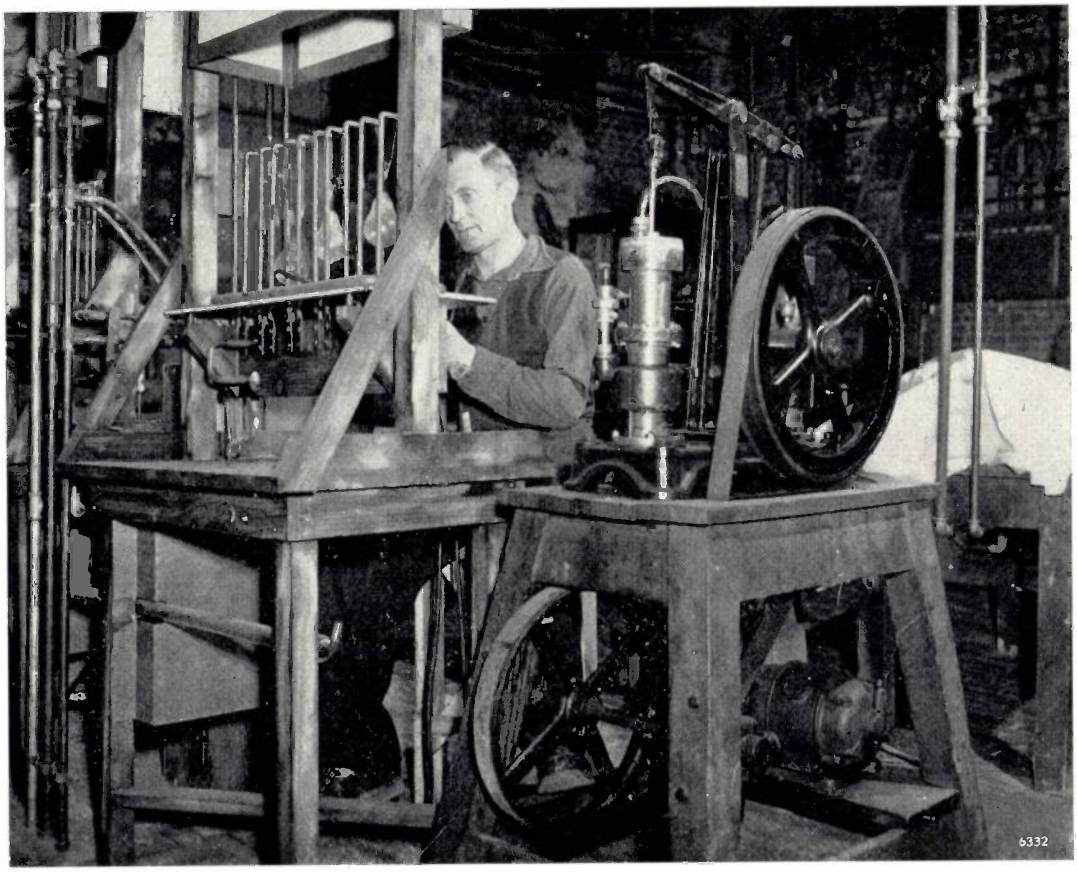


Fig. 8. Reconstruction of a pumping system for carbon-filament lamps, as must have been used by Gerard Philips when he first started up the factory. This reconstruction was made in 1951 on the occasion of the 60th anniversary of the Company, and the system was then in use for several weeks.

Plaster of Paris remained necessary for fixing the vitrite lamp caps to the bulb, and Gerard made enquiries amongst suppliers for special types of plaster, which had to be very hard and not to cxpand too much, otherwise the neck of the bulbs might be damaged (*letter X*).

An order placed on 3rd October 1891 with a printer at Krefeld for about 150 000 gummed and perforated labels of 15 × 10 mm, which were to be stuck on the lamps, was based on an estimate of the kinds of lamps and voltages for which there might be a demand. The order comprised:

17 500 labels for lamps of 8 candle-power
32 000 labels for lamps of 10 candle-power
70 000 labels for lamps of 16 candle-power
17 500 labels for lamps of 32 candle-power
17 500 labels for lamps of 50 candle-power.

The distribution according to voltage was as follows:

22 500 labels for 100 V
22 500 labels for 105 V
19 500 labels for 110 V
9000 labels for each of the voltages 65-98-102-108-112-115-120-125 V
4500 labels for each of the voltages 50-60-70-75 V.

The work of equipping the factory, making the pumps and various rather primitive aids to production, took almost a year. A fire insurance specification dated 5th February 1892 gives a clear picture of the inventory; see *letter XI*. The value of the buildings was assessed at 16 000 guilders (without foundations and "steam chimney" which were apparently not covered by insurance), and the machinery and equipment at about 25 000 guilders.

The early years of production

The first consignment of lamps (one hundred) was dispatched in May 1892. Remarkably enough, these lamps went to the stearin-candle factory at Gouda, a competitor in the lighting business which, notwithstanding all the improvements since made to electric lamps, has by no means had to quit the field. The fire hazard in the candle factory was the reason for the early change-over from gas lighting with open flames to safe lighting with electric lamps.

Other customers at that time were mainly steamship companies, theatres, restaurants and hotels, and shops selling luxury goods.

Letter XI (letter-book p. 401).
(Translated from the original Dutch.)

5th February 1892

Insurance specification for the lamp factory of Messrs. Philips & Co., Eindhoven.

Factory, office and adjoining buildings, excluding foundations and steam chimney	F. 16.000.00
Boiler, steam engine and steam pipes	F. 6.000.00
Machinery comprising: 3 dynamos with regulator, 1 bellows, 1 mechanical iron lathe with accessories, 1 automatic air-pump, 1 vacuum tank	F. 4.000.00
Mechanical instruments, apparatus, glass-blowing, forging and other tools, work benches, glass-blowers benches, woodwork, cupboards and factory equipment	F. 10.000.00
Office equipment including: one Chadwood safe, and office requisites and stationery . .	F. 600.00
Stocks including: platinum, mercury, glassware, brass mantles, chemicals required for lamp production	F. 5.000.00
	F. 41.600.00

The sales in 1892 were no more than 11 000 lamps. That was far too low. The plans for the factory were based on the estimate that an annual production of 150 000 lamps (500 daily) would be just sufficient to defray costs, and that a target of 1000 lamps daily should be aimed at to run the factory at a profit.

The Dutch market was too small, and competition on the international market was fierce (the A.E.G. in 1891 was already producing more than 3000 lamps a day, the Edison General Electric Company about 8000 lamps a day) and prices were dropping. Hopes of introducing the electric lamp for domestic lighting — which was to be the foundation of mass production — seemed to be dashed by the spectacular development of gas lighting. In 1885 Auer von Welsbach had patented his incandescent gas mantle, and its subsequent improvement raised the luminous intensity of gas burners from 16 candle-power to 70 or 80 candle-power, resulting in an increase in "luminous efficiency" from 0.09 to 0.65 candle-hours per litre of gas. Wherever there was a gas mains the number of consumers in the early 'nineties went up in leaps and bounds, and the reddish-yellow light from thc carbon filament lamp contrasted miserably with the brilliant white, slightly greenish, gas light.

Added to all this were the early difficulties that Gerard experienced in running his factory. He had to be buyer and salesman as well as head of production, and after dismissing some of his employees who had proved unequal to their task he was obliged to shoulder an increasing variety of responsibilities himself. There was indeed a wide gulf between the work of an engineer, as he had imagined it, and that of a manufacturer. His knowledge of electrical engineering could only be applied sporadically. The purity of the raw materials and the proper operation of the Sprengel vacuum pumps (mercury-drop types) caused him the most trouble. All-round technician that Gerard was, he grappled competently with the numerous problems of production and was constantly introducing improvements: he replaced zinc-chloride cellulose as the conventional basic material for the carbon filaments by collodion acetate, which could be better controlled, and the mercury pump by the faster and more reliable oil-diffusion pump. But Gerard had his hands full with such technical problems, and could not pay sufficient attention to the selling side of the business. Sales remained far too low.

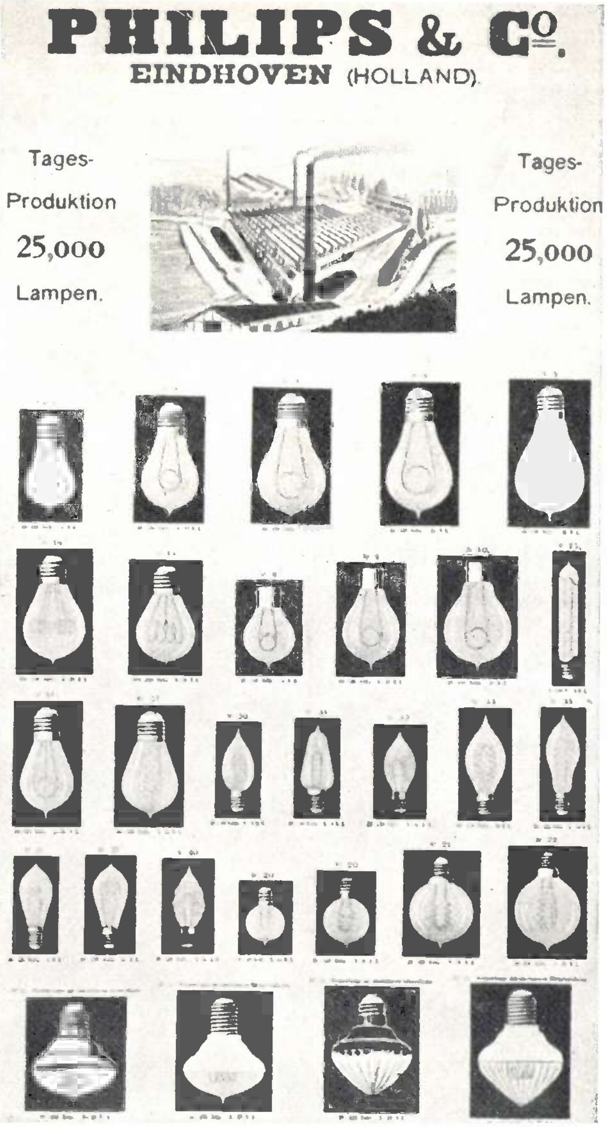

PHILIPS & Co.
EINDHOVEN (HOLLAND).
Tages-Produktion 25,000 Lampen.
Tages-Produktion 25,000 Lampen.

Fig. 9. Advertising poster of Philips & Co. in about 1903.

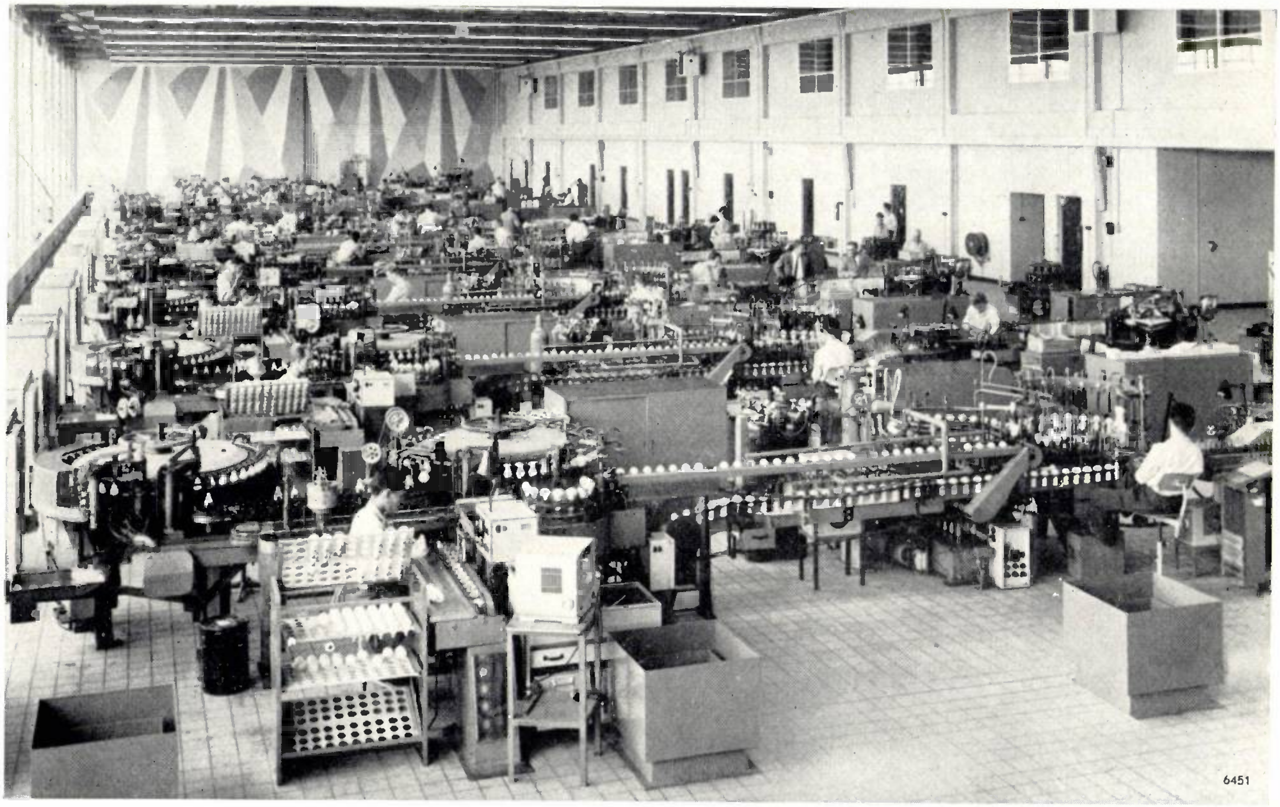


Fig. 10. A modern lamp factory: hall with a number of production groups parallel to each other, in the Philips lamp works at Weert (Netherlands), opened in 1961. Each production group covers the whole width of the hall and constitutes an independent unit. Each group, which is operated by three or four persons, is supplied with tungsten filaments, support rods, bulbs, lamp caps etc., and delivers a continuous stream of finished incandescent lamps (more than 2000 per hour). Here too the lamps of each production unit are subjected to the necessary inspection. The stream of lamps coming from each unit is carried on a conveyor belt through the floor to the packing machines under the hall.

Consequently, in 1894 father and son reached the painful decision to sell the factory and wind up the business. Once again a project to which Gerard had devoted all his energy was to come to nought. Would he now, at the age of 36, have to start afresh elsewhere? What had become of the independence he had set his heart on?

Actually it was a trifle that turned the tide at the critical moment. Someone who wanted to buy the factory tried to beat down the moderate asking price of 25 000 guilders by a further 1000 guilders, and that so irritated Gerard's father that he broke off negotiations and decided to risk another attempt at putting the enterprise on its feet. His confidence in his son's technical capabilities was unshaken, but he saw clearly where the shoe pinched: Gerard had to be relieved of the commercial work. When it proved difficult to find a suitable man for the job, it occurred to Gerard's father to offer it to his younger son, Anton Frederik.

Young Anton, till then engaged in a London banking house, entered his elder brother's small business in the beginning of 1895, at the age of 21. So quickly did he boost sales, largely by travelling to countries abroad — including Russia, where there were as yet few gas-works — that at the end of the same year Philips & Co. were no longer operating at a loss [5]).

In the following years production went from strength to strength:

1895:	100 000	lamps
1896:	280 000	,,
1897:	630 000	,,
1898:	1 200 000	,,
1900:	2 700 000	,,
1902:	3 600 000	,,

The firm hold which Philips & Co. had managed by 1902 to acquire even in Germany, in spite of powerful German competition, is illustrated by the fact that the Düsseldorf Gewerbeausstellung (Industrial Exhibition) of that year was lit entirely by Philips lamps.

[5] It is perhaps interesting to mention that the first electric lamp factory to go into regular production — originally 1000 lamps daily — founded by Edison in 1881, was also run at a loss for the first three years.

Even before the advent of the metal filament (the tungsten lamp entered the field in 1906) the factory had to be considerably expanded, a variety of special lamp types were developed, and production rose to 25 000 and more lamps daily (*fig. 9*). Meanwhile the production process was being increasingly mechanized and made more efficient. From the moment that Anton Philips took over the commercial side, Gerard Philips was able to give full expression to his technical talents as manufacturer and designer — he was after all a trained mechanical engineer — and mechanization claimed his undivided attention. When Gerard retired from the management in 1922 the production process was so highly mechanized that the output per man-hour was fifty times greater than in 1892.

It is outside the scope of this article to describe this and the subsequent progress up to the present day. Let it suffice to indicate the point it has now reached by an illustration (*fig. 10*): a recently opened factory turning out more than 2000 lamps an hour per production unit.

Our last figure (*fig. 11*) brings us back to the beginning of our story: it is a bas-relief showing the founder Gerard Philips, which has been built into the foot of the chimney of the (still existent) little factory of 1891.

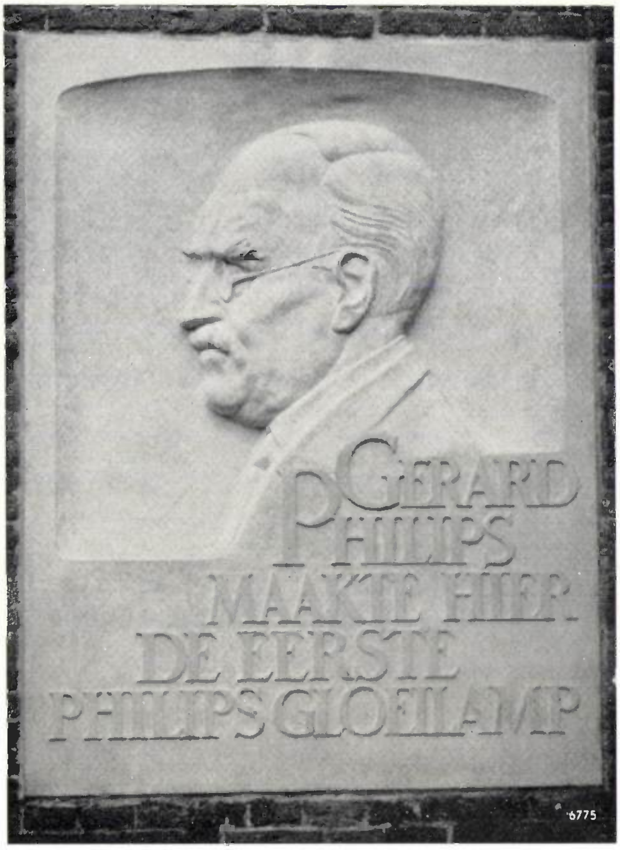


Fig. 11. Commemorative plaque affixed to the factory of 1891: "Gerard Philips made the first Philips incandescent lamp here".

Short bibliography

Anon. (under the pseudonym "Electron"), De Bougie Jablochkoff, Electrische Verlichting, published by Van Es, Amsterdam 1878 (in Dutch).

This booklet acclaims the demonstration at Paris of "a new form of lighting which has entered into its infancy", and aptly illustrates the topicality which the "distribution of the electric light" then possessed.

J. Dredge (Ed.), Electric Illumination, 1882 (Part I) and 1885 (Part II).

Mainly contributions compiled from "Engineering" with extracts from patents granted in America and England.

A. von Urbanitzky, Electricity in the service of man (adapted from the German by R. Wormell), Woodall, London 1886.

Idem, Die elektrische Beleuchtung und ihre Anwendung in der Praxis, 2nd Edn., Hartleben, Vienna 1890.

G. S. Ram, The incandescent lamp and its manufacture, The Electrician Publ. Comp., London 1893.

The author describes his experience of many years in the making of lamps, presenting his knowledge "with as little mathematical embellishment as, under the circumstances, is possible".

H. Weber, Die elektrischen Kohlenglühfadenlampen, ihre Herstellung und Prüfung, Jänecke, Hannover 1908.

This book contains among other things a useful historical review. Also of interest is a chapter on the repair of electric lamps, i.e. the removal of carbon deposits or the replacement of a burnt-out filament. The author is opposed to this practice and remarks: "The user will, if the lamp comes up to reasonable expectations, gladly buy a new lamp at an appropriate price after the old one has become defective, and will not want to have cheaper repaired lamps".

A chronological history of electrical development, Nat. Electr. Manuf. Assoc., New York 1946.

F. A. Lewis, The incandescent light; a review of its invention and application, The Thomas Alva Edison Foundation, West Orange (N.J.) 1949.

G. F. Westcott, Mechanical and electrical engineering (Classified lists of historical events, The Science Museum), H. M. Stat. Off., London 1955.

P. J. Bouman, Anton Philips of Eindhoven, Weidenfeld and Nicolson, London 1958.

Summary. A brief account is given of the early period of electrical engineering, and a picture is presented of the situation that existed in the field of lighting in Europe, and particularly in the Netherlands, when Gerard Philips started manufacturing lamps at Eindhoven. A valuable source of historical data concerning the creation of the factory was found in a letter-book containing copies of the correspondence addressed by Gerard Philips to numerous persons between April 1889 and April 1892. Eleven of these letters, which contain interesting details of that period, are printed here *in extenso*.

IODINE INCANDESCENT LAMPS

I. PRINCIPLE

by J. W. van TIJEN *).

621.326.79

A short time ago a new type of incandescent lamp possessing very attractive properties was introduced. The lamp owes these properties to the presence of a small quantity of iodine inside the bulb, hence the name "iodine lamp" by which it is generally known.

The value of this invention can best be made clear against the historical background of the incandescent lamp, with special references to the gradual improvement of luminous efficiencies. In an incandescent lamp electrical energy is converted into light by heating a filament to incandescence by the passage of an electric current. The higher the temperature of the filament the more efficient is the energy conversion (greater luminous efficiency), but the faster too, unfortunately, is the rate at which the filament material evaporates. Only if the rate of evaporation can be reduced is it possible to raise the temperature without at the same time shortening the life of the filament.

The improvement in the luminous efficiency of the incandescent lamp that has been achieved in the eight decades of its existence may therefore be described as the result of constant efforts to overcome filament evaporation. The iodine filling is the latest tool for this purpose.

Short history of the incandescent lamp [1])

The incandescent lamp made by Edison in 1879 consisted of a carbon filament mounted in an evacuated glass bulb. This type of lamp (rated for about 50 W) had a luminous efficiency in the region of 3 lumens per watt [2]).

Subsequent developments were dominated at first by the search for filament materials less subject to evaporation than carbon. Following Nernst's ceramic filament and the use of osmium and tantalum, the drawn tungsten filament was introduced (by Coolidge) in 1910. Using this filament, also mounted in an evacuated bulb, luminous efficiencies of about 9 lm/W were achieved with a consumption of 40 W. This improvement by a factor of 3 compared with a carbon-filament lamp was chiefly the result of being able to raise the filament temperature from about 2100 °K to 2400 °K. (Another reason for the improvement was that the spectral energy distribution of the radiation emitted by tungsten is more favourable than from carbon at the same temperature, in that the visible radiation from tungsten constitutes a larger percentage of the total power dissipated by radiation.)

Since then no better material than tungsten has been found. All subsequent developments have therefore been based on the tungsten filament, efforts being concentrated on minimizing its rate of evaporation. Langmuir took this development a good step forward by filling the bulb with inert gases, such as argon and nitrogen (1913). This had the effect of returning to the filament a certain percentage of the evaporated tungsten by the collision of tungsten atoms with the gas molecules, enabling the temperature of the filament to be raised to about 2800 °K while maintaining the same useful life. The gas filling has the drawback, however, that energy is lost by heat transfer to the gas. Langmuir reduced these "gas losses" by coiling the filament, thereby achieving a luminous efficiency of 11 lm/W from a 60 W lamp with a gas filling of about one atmosphere. In the 'thirties this was raised to 12 lm/W by introducing a double coil (the coiled-coil filament), which reduced the gas losses still further.

Efficiency was again improved by using instead of the conventional argon an inert gas of greater molecular weight, such as krypton or xenon. The greater molecular weight has the effect of reducing the tungsten evaporation and also the heat transfer to the gas. Krypton and xenon are relatively scarce and therefore more expensive, for which reason they are used only in special lamps where they are economically justified.

For a long time it looked as if the evolution of the incandescent lamp's efficiency had come to a standstill, until in 1959 the iodine filling was introduced. It was found possible by means of a chemical process involving iodine to return to the filament that part of the evaporated tungsten which is not returned by

*) Philips Lighting Division, Eindhoven.

[1]) For a comprehensive history of the incandescent electric lamp see: A. A. Bright Jr., The electric-lamp industry, MacMillan, New York 1949. See also the article by N. A. Halbertsma in this number (p. 222) and further W. Geiss, Philips tech. Rev. 1, 97, 1936 and 6, 334, 1941.

[2]) The theoretical maximum for white light (equi-energy spectrum) is approximately 220 lumens per watt.

the filling gas and which settles on the bulb wall.

In the following an attempt will be made to give some idea of the chemical and physical effects that occur in an iodine lamp.

The regenerative iodine cycle

The principle underlying the iodine lamp is a regenerative cycle whereby the tungsten deposited on the bulb wall is converted at the temperature of the wall into a volatile compound, which then decomposes on or in the neighbourhood of the filament. This kills two birds with one stone:

a) it prevents blackening of the bulb wall;

b) the tungsten filament suffers no weight loss.

Various tungsten compounds enter into consideration for such a cycle. The halides and the carbonyl compound $W(CO)_6$ have the necessary volatility. As regards the other properties required, the halides are more suitable than the carbonyl compound, since the decomposition temperature of the latter is rather low. Moreover its use involves the risk of tungsten-carbide formation, which might cause brittle spots to appear in the filament.

The fundamental idea of a regenerative halogen cycle is fairly old. As long ago as 1916, Hamburger published particulars of an imperfect cycle produced with the aid of chlorine, on the basis of experiments done in our research laboratories [3]). It was not until 1959, however, that Zubler and Mosby in the General Electric laboratory at Cleveland succeeded in making technically useful lamps on this principle, using a regenerative iodine cycle [4]). One of the difficulties they had to overcome was to produce bulbs capable of withstanding higher temperatures than those found in normal incandescent lamps (see also p. 240).

We shall now consider the various tungsten halides in more detail. Their formation and decomposition takes place according to the following type of reaction:

$$W + nX \rightleftarrows WX_n ,$$

where X represents a halogen atom and n is a small integer. The situation of the equilibrium depends strongly on temperature, as we shall show with the aid of the graph in *fig. 1*. The position of the equilibrium is given by the equation

$$\frac{P_W\ (P_X)^n}{P_{WX_n}} = K,$$

where P_W, P_X and P_{WX_n} are the partial pressures of the components in equilibrium, and K is the equilibrium constant. In fig. 1 the logarithm of the calculated equilibrium constant K is plotted as a function of temperature. At values of $K \ll 1$ the partial pressure of the atoms W and X is much higher than that of the molecules WX_n; at values of $K \gg 1$ it is much lower. At about the temperature where the K line passes the value 1 (or $\log K = 0$) the equilibrium changes over from one side to the other.

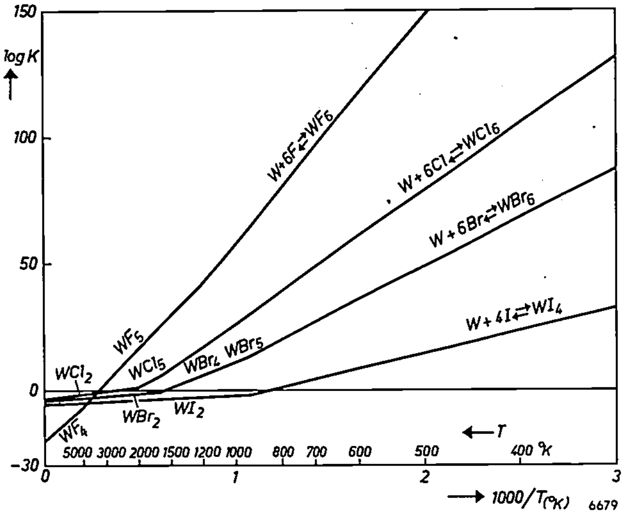


Fig. 1. The thermodynamically calculated equilibrium constant K for the formation of all tungsten halides, as a function of temperature T. The graph shows that, at the usual bulb-wall temperatures (200-600 °C) the equilibrium lies on the side of the compound, and at the usual filament temperatures (3000 °K) on the side of the constituent elements. These are the necessary conditions for a regenerative tungsten cycle.

Fig. 1 shows that at the high temperature at which tungsten filaments are incandescent (approx. 3000 °K) most tungsten halides will be almost completely dissociated. A possible exception is the fluoride, which will only be partly dissociated at the temperature in question. In the case of the chloride the temperature of the equilibrium cross-over lies at about 2000 °K, for the bromide at about 1500 °K and for the iodide at about 1000 °K.

From the above data we may conclude that:

1) at conventional filament temperatures all tungsten halides are dissociated to the extent required for the regenerative cycle;

2) at all normal bulb-wall temperatures (200-600 °C) the equilibrium lies at the tungsten-halide side, which is also a necessary condition for the regenerative cycle.

[3]) L. Hamburger, Chem. Weekbl. 13, 535, 1916.

[4]) E. G. Zubler and F. A. Mosby, Illum. Engng 54, 734, 1959.

We have been considering only the reaction with halogen *atoms* and not that with molecules. The reason is that, at the high temperature of the filament, all halogen molecules dissociate completely into atoms. Although these atoms diffuse to regions of lower temperature, they have little chance to associate there owing to the low probability of collision at the gas pressures chosen in practice. Moreover, the reactivity of the halogen atoms is many times greater than that of the halogen molecules. The formation of tungsten halides will therefore be primarily due to reaction with atoms.

It appears from the foregoing that, in principle, all tungsten halides are suitable for use in incandescent lamps. Nevertheless, it is not by chance that the iodine cycle has had the most success. One of the main reasons is the fact that in a regenerative cycle the aim must be to attack the tungsten deposits on the relatively cold bulb wall while leaving the relatively cold ends of the tungsten filament intact. It can be seen from the graph that the filament ends are least subject to attack when iodine is used. In that case the temperature of the filament ends need be no higher than about 1500 °K to remove the danger of chemical attack. It is not difficult to find practical constructions that fulfil this requirement.

Whilst this may seem to open the way to the making of lamps of higher than normal efficiency, or possibly of longer life, further consideration shows that there are other conditions that must also be satisfied.

Analysis of the reasons for the limited life of normal incandescent lamps and iodine lamps

For a proper understanding of the phenomena in an iodine lamp, we shall take a closer look at the process by which the filament deteriorates in operation and finally burns out. Let us first consider a vacuum lamp. If the filament in such a lamp had a perfectly constant diameter and underwent completely uniform evaporation, the filament, growing steadily thinner, would pass less and less current and become steadily colder. The result would be an infinitely long filament life and a steadily declining light output. The actual finite life of the filament is due to the occurrence of thin or weak spots which are loaded with the full current. At these spots the tungsten wire gets very hot and finally breaks.

This is a random process but there is nevertheless a fixed and entirely reproducible relation between the rate of evaporation and the life of a filament. To illustrate this, *Table I* presents data [5]) relating to uncoiled tungsten wire in vacuo, together with the appertaining luminous efficiencies.

Table I. The life of an uncoiled tungsten filament in vacuo.

Temperature °K	Luminous efficiency lm/W	Rate of evaporation v g/cm² s	Life H of wire of 0.01 mm diameter hours	$v \times H$
2000	2.93	15.5×10^{-15}	1.04×10^{7}	16.1×10^{-8}
2200	5.71	22.4×10^{-13}	7.20×10^{4}	16.1×10^{-8}
2400	9.77	13.8×10^{-11}	1.17×10^{3}	16.1×10^{-8}
2600	14.8	41.7×10^{-10}	3.86×10^{1}	16.1×10^{-8}
2800	20.9	83.3×10^{-9}	1.9	15.8×10^{-8}
3000	27.8	10.5×10^{-7}	0.15	15.7×10^{-8}

Apart from the very rapid decline in life with increasing temperature, we notice especially that the life is almost exactly inversely proportional to the rate of evaporation. This means that the filament will nearly always burn out as soon as a particular constant percentage of tungsten has evaporated. This percentage is known as the end-of-life weight loss, and for vacuum lamps it is 10 to 15% of the weight of the wire.

In a gas-filled lamp the situation is very much the same. The only difference is that, owing to the tungsten atoms constantly evaporating and returning to the filament, making the wire more and more irregular, the end-of-life weight loss is smaller by a factor of 2 or 3 than in vacuum lamps. This effect, which in itself would impair the life of the filament, is more than compensated by the retarding effect of the gas filling which, at a pressure of about 1 atm, may easily slow down the rate of evaporation by a factor of 50.

What now is the situation in an iodine lamp? When the lamp is burning the tungsten evaporates in the normal way, part of the evaporated tungsten being reflected back to the filament by the gas molecules. The part not so reflected reaches the bulb wall, where it reacts with the iodine vapour and is converted into volatile tungsten iodide; this dissociates near the filament, as a result of the high temperatures prevailing there, and thus reinforces the local concentration of tungsten vapour. This in turn increases the deposition of tungsten on the filament, until just as much tungsten is deposited on the filament as the latter loses by evaporation.

In the ideal case of uniform tungsten evaporation and deposition, this would mean that the iodine lamp would have an infinitely long life and a constant light output (unlike the "ideal" vacuum lamp, whose light output would gradually decline). This ideal case is never found, for there will always be temperature differences in the incandescent filament. We shall see below what consequences this has.

[5]) C. Zwikker, Physica 5, 252, 1925.

Owing to diffusion, the tungsten vapour around the filament shows a homogeneous distribution. The concentration of the tungsten vapour is adjusted, as it were, to the *average* temperature of the filament. On the relatively hot spots of the tungsten surface more tungsten will evaporate than will be deposited; conversely, on the relatively cold spots more will be deposited than evaporated.

Thus, although the filament, seen as a whole, suffers no weight loss in an iodine lamp, tungsten can certainly "migrate" from relatively hot parts of the filament to relatively cold parts. This process can in fact be observed, and its effect is that the relatively hot spots become gradually thinner and hotter until finally the filament burns out.

If, by analogy with the behaviour of normal incandescent lamps, we may reckon with a more or less constant "end-of-life migration", we can show that the following formula [6]) holds for the life H of an iodine lamp:

$$H \propto \frac{T_{\mathrm{h}}^{-32}}{\Delta T},$$

where ΔT is the temperature difference between relatively hot and cold parts of the filament and T_{h} is the temperature of the hottest part.

For $\Delta T = 0$ the life of the filament would be infinitely long. This ideal, as we have said, cannot be achieved because small irregularities are always present or develop in a tungsten filament. Again, for a given temperature difference the life decreases very rapidly with increasing T_{h}, because the tungsten migration process is greatly accelerated at higher temperatures.

Compared with the tungsten evaporation in normal incandescent lamps, however, tungsten migration is a slow process provided that special precautions are taken in the construction and filling of the iodine lamp (see points (5) and (6) in the following section). In fact, therefore, the iodine lamp offers a longer life and/or a higher efficiency than a normal incandescent lamp. For the same life some types of lamp, filled at a pressure of about 1 atm, show an improvement in efficiency of about 25%. Another advantage of the iodine lamp is that its light output remains constant throughout life (there being no blackening of the wall and no weight loss in the filament).

Some technical details of the iodine lamp

To sustain an iodine cycle in an incandescent lamp, the lamp must meet certain requirements which cause it to differ in construction from normal incandescent lamps.

1) Size and temperature of the bulb.

The bulb of a normal incandescent lamp should preferably not be too small, as otherwise the blackening that occurs will absorb too much light, and, more especially, the bulb will get too hot.

In the case of the iodine lamp the situation is different; for one thing there is no blackening, and for another the bulb temperature *must* be higher than normal. At too low temperatures the rate of formation of tungsten iodide would be too low, and moreover the compound would condense. The dimensions of the bulb are often chosen so as to obtain a bulb temperature of about 600 °C. This means that the bulb must be made of quartz glass or some other type of glass having a high softening point.

2) Lamp shape.

Iodine lamps are made in a variety of shapes. In the case of high-wattage iodine lamps (from 500 W) the most usual shape so far is a cylindrical bulb with a coiled filament along the axis (see *fig. 2*).

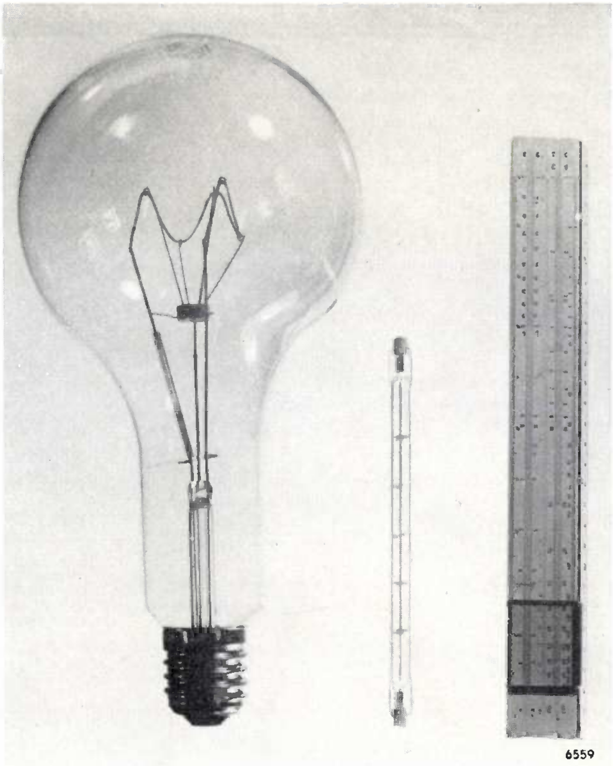


Fig. 2. A 1000 W iodine lamp, with a 1000 W incandescent lamp of normal construction for comparison.

[6]) This formula, which we derived with the aid of the kinetic gas theory, applies to the temperature range from about 2700 to about 2900 °K. Outside this range the formula holds with an exponent of T_{h} differing somewhat from that used here. The reliability of the formula has been confirmed by numerous experiments.

A tubular lamp of this kind, whose length is large compared to its diameter, must be operated horizontally, otherwise the iodine will accumulate by thermal diffusion at the bottom of the bulb and the cycle will no longer be maintained.

3) Iodine content.

The lamps are filled with an inert gas to which about 0.25×10^{-6} moles/cc of iodine is added. Too much iodine causes loss of light as a result of absorption by the iodine vapour; too little iodine makes the iodine cycle too slow.

4) Mounting and support of the filament.

The usual materials, such as nickel, molybdenum, tantalum and iron, are not suitable for the filament supports, unless measures are taken to prevent them from reacting with the iodine. The metals must therefore either be given a protective coating, e.g. glazed, or noble metals like platinum must be used, or special alloys. Tungsten can also be employed, but the components must then be designed so as to ensure that they reach a sufficiently high temperature (see also p. 239).

5) Purity of the gas filling.

It is a known fact that traces of water vapour in an incandescent lamp rapidly transport tungsten by a cyclic process from the filament to the bulb wall. Although in an iodine lamp the tungsten deposited on the bulb wall is returned to the filament by the iodine cycle, this additional transport of tungsten leads to very irregular deposition on the filament (see *fig. 3*), resulting in a relatively short life. Other impurities can also have an adverse effect on the life of the lamp. In the manufacture of iodine lamps, therefore, measures are necessary to ensure that the gas filling has a high degree of purity.

6) Reduction of temperature gradients along the filament.

Since tungsten has the tendency to migrate from relatively hot to relatively cold spots, steps must be taken to reduce the temperature gradients along the filament. For example, the components used for mounting and supporting the filament should be as light as possible in construction, so as to minimize the loss of heat through them.

Further improvement of efficiency by increased filling pressure

Since iodine lamps, as we have seen, are relatively small, they can readily be filled with gases at pressures higher than one atmosphere. The small bulbs are relatively strong and easily capable of withstanding a higher pressure. Under favourable conditions, filling pressures up to about 10 atm are feasible in very small bulbs.

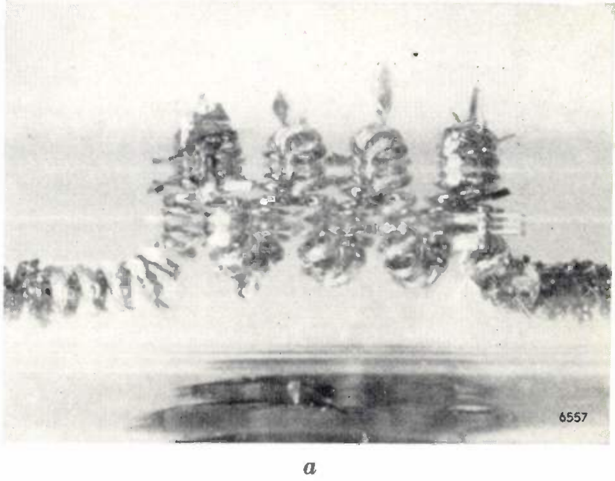


Fig. 3. *a*) Tungsten filament of an iodine lamp in which traces of water vapour were present, *b*) tungsten filament of an iodine lamp containing no water vapour. Both lamps had burnt for the same number of hours.

A high-pressure filling in iodine lamps is particularly advantageous in view of the virtual absence of convection currents in the small bulbs. In standard gas-filled incandescent lamps, where these currents do occur, a higher pressure is a double-edged weapon: on the one hand it slows down the evaporation of tungsten, on the other it increases the heat loss via the gas. The latter, however, is entirely due to convection. In a "stationary" gas the thermal conduction is independent of pressure. All we are left with in small iodine lamps, therefore, is the retarding effect on filament evaporation. According to a formula calculated by us and confirmed by experiment, the luminous efficiency in that case is proportional to the filling pressure to the power 0.12. With higher pressures up to about 10 atm, which can be used under appropriate conditions, luminous efficiencies can thus be achieved which are about 30% higher than in corresponding iodine lamps filled at a pressure of 1 atm [7]).

[7]) About half the efficiency improvement of 25% at a filling pressure of 1 atm, mentioned on p. 240, is estimated to be due to the increase of pressure that takes place when the lamp is burning.

To sum up, it can be said that the iodine lamp compares favourably with the normal incandescent lamp by maintaining full lumen output throughout life. A few types of iodine lamp, having a "normal" life, offer efficiency gains up to 25%. This applies to lamps filled at a pressure of 1 atm. In those types where very high pressures are technically feasible, a further improvement of up to 30% can be achieved, which represents a gain in efficiency of 60% over normal incandescent lamps.

IODINE INCANDESCENT LAMPS

II. POSSIBLE APPLICATIONS

by J. J. BALDER *).

621.326.79

The iodine lamp is usually cylindrical in shape, with a diameter of about 10 mm, and has a coiled filament mounted along the axis of the bulb. We shall discuss first some possible applications of iodine lamps for wattages higher than 500 W, which are long in proportion to their diameter, and then those of relatively short iodine lamps of lower wattage [1]).

The first category of lamps is particularly suited for use in conjunction with cylindrical reflectors. The beam of light from this combination of lamp and reflector can fairly easily be given the desired form in the plane perpendicular to the long axis, the light distribution in this plane being controlled by the shape of the reflector cross-section. If necessary, a beam-spreading lens may be added in front of the reflector.

Effective control of the light in planes through the axis is more difficult. If nothing is done to prevent it, the beam in such axial planes fans out over the whole 180°. The width of the beam in those planes can only be limited to some extent by using specially shaped end mirrors or other more complicated devices.

The obvious applications are therefore to be found where a beam is wanted which is broad in one direction and narrow in the direction perpendicular thereto. Beams of this kind are required for flood-lighting, the lighting of sports fields and similar flat areas, poster lighting etc. Other possible applications are the lighting of factories, churches, shopping arcades, warehouses, theatres, studios and so on. To function properly the lamps must be operated horizontally (see Part I of this article). In nearly all the cases mentioned, the horizontal position of the lamp is not in conflict with the use for which the lighting is intended.

The small transverse dimensions of the lamp make it possible to obtain the required beams with good optical efficiency using fairly small reflectors, provided the materials and the design of the fittings are well chosen with a view to the possibility of high temperatures. *Fig. 1* shows a fitting for a 1000 W iodine lamp, suitable for sports-field lighting, for example, or for flood-lighting.

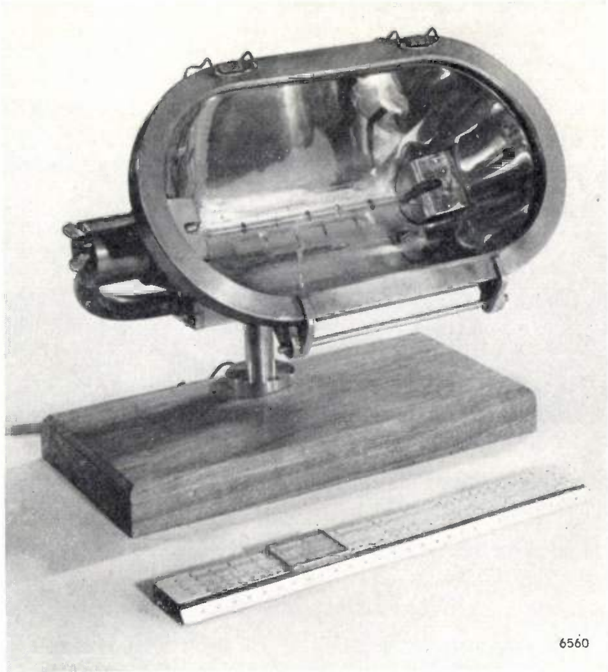


Fig. 1. A 1000 W iodine-lamp fitting, suitable for sports-field lighting for or flood-lighting.

*) Philips Lighting Division, Eindhoven.

[1]) For potential applications of the iodine lamp in general, see also C. J. Allen and R. L. Paugh, Illum. Engng **54**, 741, 1959.

that case is proportional to the filling pressure to the power 0.12. With higher pressures up to about 10 atm, which can be used under appropriate conditions, luminous efficiencies can thus be achieved which are about 30% higher than in corresponding iodine lamps filled at a pressure of 1 atm [7]).

To sum up, it can be said that the iodine lamp compares favourably with the normal incandescent lamp by maintaining full lumen output throughout life. A few types of iodine lamp, having a "normal" life, offer efficiency gains up to 25%. This applies to lamps filled at a pressure of 1 atm. In those types where very high pressures are technically feasible, a further improvement of up to 30% can be achieved, which represents a gain in efficiency of 60% over normal incandescent lamps.

[7]) About half the efficiency improvement of 25% at a filling pressure of 1 atm, mentioned on p. 240, is estimated to be due to the increase of pressure that takes place when the lamp is burning.

IODINE INCANDESCENT LAMPS

II. POSSIBLE APPLICATIONS

by J. J. BALDER *).

621.326.79

The iodine lamp is usually cylindrical in shape, with a diameter of about 10 mm, and has a coiled filament mounted along the axis of the bulb. We shall discuss first some possible applications of iodine lamps for wattages higher than 500 W, which are long in proportion to their diameter, and then those of relatively short iodine lamps of lower wattage [1]).

The first category of lamps is particularly suited for use in conjunction with cylindrical reflectors. The beam of light from this combination of lamp and reflector can fairly easily be given the desired form in the plane perpendicular to the long axis, the light distribution in this plane being controlled by the shape of the reflector cross-section. If necessary, a beam-spreading lens may be added in front of the reflector.

Effective control of the light in planes through the axis is more difficult. If nothing is done to prevent it, the beam in such axial planes fans out over the whole 180°. The width of the beam in those planes can only be limited to some extent by using specially shaped end mirrors or other more complicated devices.

The obvious applications are therefore to be found where a beam is wanted which is broad in one direction and narrow in the direction perpendicular thereto. Beams of this kind are required for flood-lighting, the lighting of sports fields and similar flat areas, poster lighting etc. Other possible applications are the lighting of factories, churches, shopping arcades, warehouses, theatres, studios and so on. To function properly the lamps must be operated horizontally (see Part I of this article). In nearly all the cases mentioned, the horizontal position of the lamp is not in conflict with the use for which the lighting is intended.

The small transverse dimensions of the lamp make it possible to obtain the required beams with good optical efficiency using fairly small reflectors, provided the materials and the design of the fittings are well chosen with a view to the possibility of high temperatures. *Fig. 1* shows a fitting for a 1000 W iodine lamp, suitable for sports-field lighting, for example, or for flood-lighting.

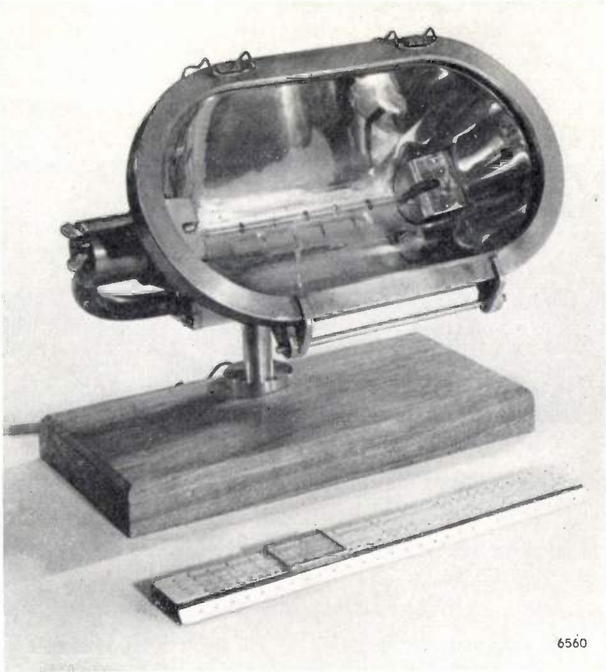


Fig. 1. A 1000 W iodine-lamp fitting, suitable for sports-field lighting for or flood-lighting.

*) Philips Lighting Division, Eindhoven.

[1]) For potential applications of the iodine lamp in general, see also C. J. Allen and R. L. Paugh, Illum. Engng **54**, 741, 1959.

In regard to the applications of the low-wattage category of lamps (up to a few hundred watt), the filament of which may be from a few millimetres to a few centimetres in length, long cylindrical reflectors need no longer be considered. More or less bowl-shaped, possibly faceted, reflectors are suitable for some purposes; where the filaments are only a few millimetres in length use may even be made of reflectors having rotational symmetry, with a parabolic or other, more complicated form of cross-section. Here, too, the smallness of the light source makes it possible to use small reflectors.

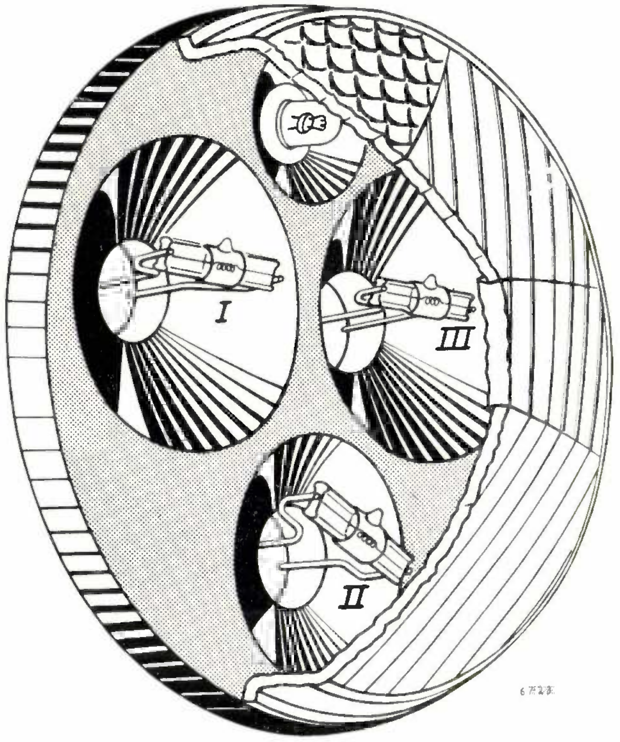


Fig. 2. Example of a car headlamp fitted with three 35 W iodine lamps, *I*, *II* and *III*. The individual reflectors are paraboloids of revolution, with diameters of 75, 60 and 75 mm respectively. In this design the three reflectors are contained in a space equal in diameter to that of many normal headlamps (170 mm). The front glass consists of segments which each give a different beam spread. The dipped beam (passing beam) is obtained from lamps *I* and *II*, the main beam (driving beam) from lamps *I* and *III*. The small normal incandescent lamp fitted in a fourth, small reflector at the top functions together with the appertaining segment of the front glass as a side light.

As an example of the application of low-wattage iodine lamps, we shall consider their use in car headlamps.

Car headlamps must be able to illuminate the road in two different ways: by a main beam (driving beam) and, in the presence of approaching traffic, by a dipped beam (passing beam). The passing beam can with advantage be made asymmetrical. Both forms of lighting have to comply with statutory requirements [2]). We shall now discuss one of the many possible designs of a car headlamp using iodine lamps. In this case three lamps of 35 W each are used; see *fig. 2*.

The three iodine lamps are contained in a space having the dimensions of a normal car headlamp measuring 170 mm across. Their common front glass consists of four segments, each producing a different beam spread. The first lamp (*I*) is mounted in a reflector 75 mm in diameter which, like the others, is a paraboloid of revolution. The light which it reflects is given a considerable spread by the segment of the front glass through which it passes. In this way a broad beam is obtained as shown in diagram *I* in *fig. 3*. The second iodine lamp (*II*) gives a concentrated beam as represented in diagram *II* in fig. 3, with the aid of a reflector only 60 mm in diameter and a segment of the front glass giving little spread. The third iodine lamp (*III*) is mounted like the first in a reflector 75 mm in diameter, but in this case the front glass gives a moderate spread. The resultant distribution can be seen in diagram *III*.

Beams *I* and *II* combined produce a good asymmetrical passing beam (see diagram *A*), and *I* and *III* together give a good main beam (see diagram *B*). When switching from passing to main beam and vice versa, lamp *I* continues to burn, thus ensuring a certain continuity in the lighting of the road.

We have not considered in this example such details as the positioning of the filament, the screening in directions where oncoming traffic might be troubled by glare etc. In these respects there is a choice from many possible designs. *Fig. 4* shows as an example a lamp intended for axial mounting of the filament in the reflector. There is also a much

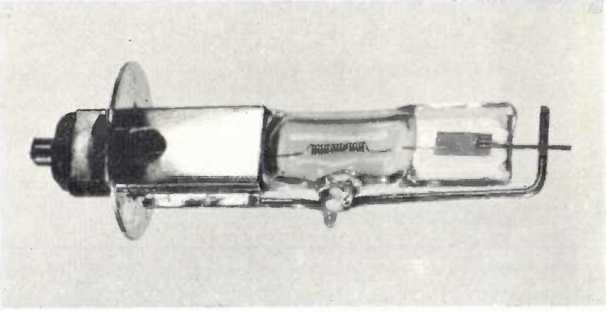


Fig. 4. Example of an iodine car lamp (14 V, 70 W) designed for axial positioning of the filament in the reflector. For comparison, see the headlamp in fig. 2, where the filaments are perpendicular to the reflector axis.

[2]) See the introduction to the article by W. Bähler in this number (p. 278), and J. B. de Boer and D. Vermeulen, Philips tech. Rev. 12, 305-317, 1950/51.

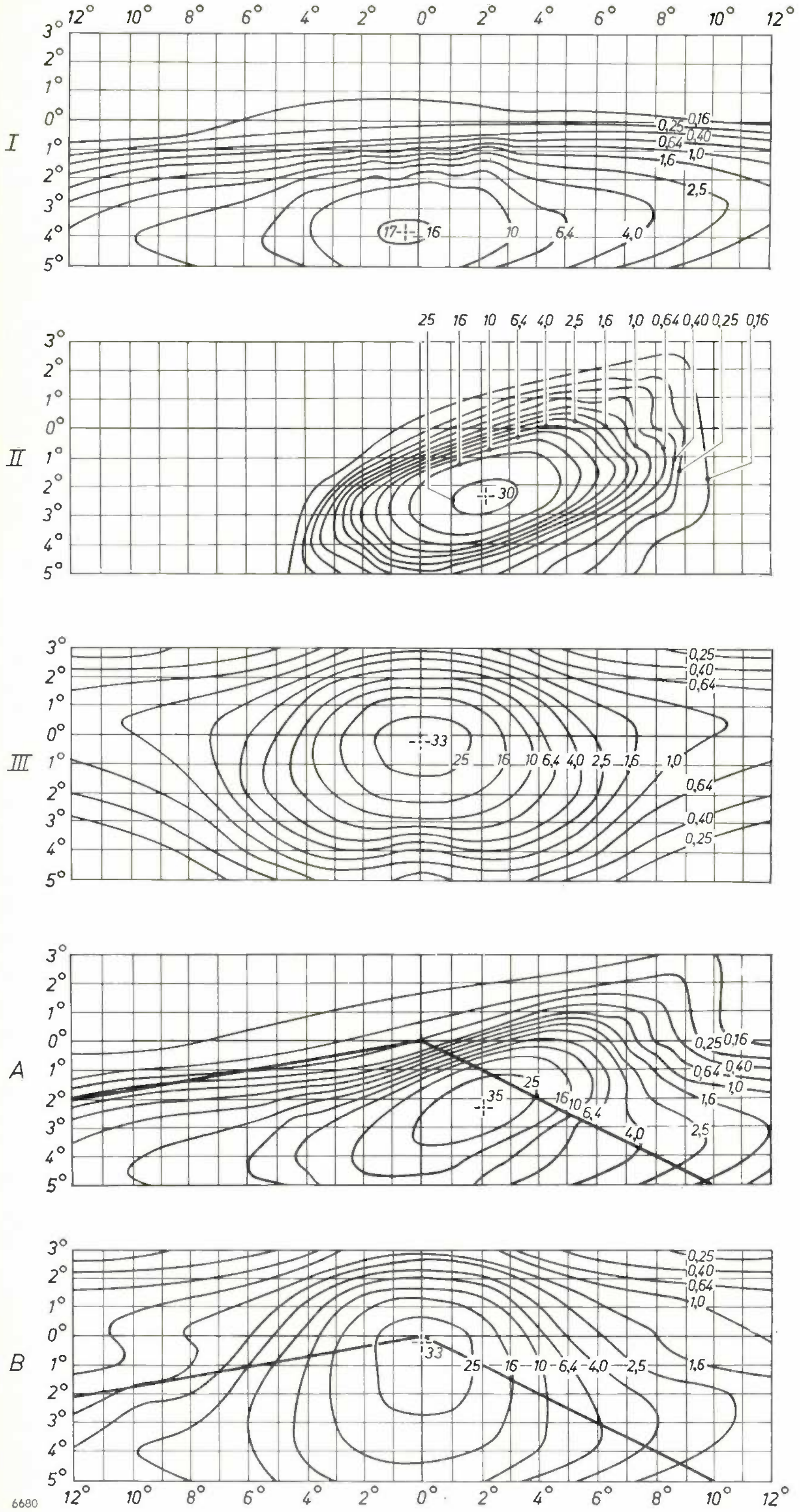


Fig. 3. Isocandela diagrams of the car headlight of fig. 2, measured at a distance of 25 metres. The figures on the contours give luminous intensities in kilocandela. Diagrams *I*, *II* and *III* relate to the iodine lamps *I*, *II* and *III* respectively. Diagram *A* represents the sum of the beams in diagrams *I* and *II*, which is a European asymmetrical passing beam. The diagram also shows the left-hand and right-hand kerbs of a road 6 metres wide as seen from a car headlamp situated at a height of 0.75 metres above the middle of the right-hand half of the road [2]). Diagram *B* gives the sum of the beams from diagrams *I* and *III*: this is the driving beam. Here too the sides of the road are indicated as in diagram *A*.

wider choice than has been mentioned here in regard to wattage, the size of reflectors, the configuration and combination of reflectors, and hence of the composition, character and intensities of the beams. All car headlamps designed for operation with iodine lamps are superior to standard headlamps by reason of either a higher optical efficiency for the same size of headlamp, or a smaller size for the same optical efficiency, and also by reason of the better luminous efficiency of the iodine lamp itself.

Summary of I and II. The history of the incandescent lamp has been marked by continuous efforts to overcome filament evaporation, with the object of raising the luminous efficiency and/or lengthening the life of the lamp. A new tool for that purpose is the regenerative iodine cycle, achieved by adding a small quantity of iodine to the filling gas. The evaporated tungsten is deposited on the bulb wall and converted into a volatile iodide, which compound is again decomposed in the high-temperature region near the filament. In this way all the evaporated tungsten can be returned to the filament. The article describes the conditions for maintaining such a cycle, which in principle is also possible with other halides, and explains why the best results have been obtained with iodine. Iodine lamps are relatively small and have a higher bulb-wall temperature than normal incandescent lamps (about 600 °C). For this reason they are made of quartz glass or of other types of glass having a high softening point. Iodine lamps for high power ratings (500 W and more) are generally long and cylindrical in shape and are operated horizontally. Filled with inert gas at a higher pressure than the one atmosphere usual in normal incandescent lamps, they give an even higher luminous efficiency, owing to the absence of convection currents in the small bulbs; as a result there is scarcely any increase in the heat losses of the gas with increasing pressure. Under favourable conditions the resultant gain in efficiency compared with normal incandescent lamps is about 60%.

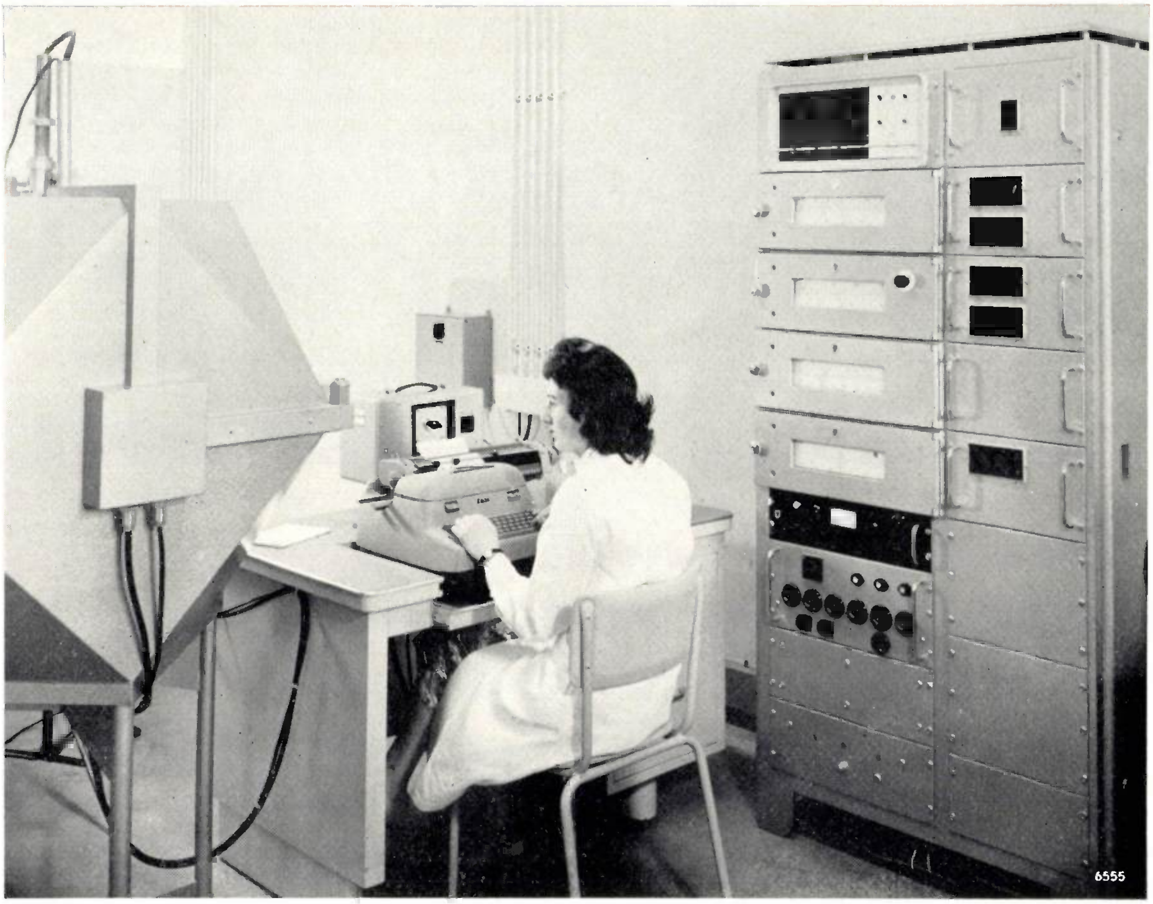
Article II discusses some possible applications of iodine lamps, in particular their use in car headlamps. An example is described in which the headlamp is fitted with three 35 W iodine lamps, each in its own small reflector.

AUTOMATIC DIGITAL PHOTOMETER FOR TESTING INCANDESCENT LAMPS

In an incandescent-lamp factory regular photometric measurements have to be made on large numbers of lamps by way of sample inspection. Done by conventional methods these measurements call for constant attention of a high order from the operator. After placing the lamp in the photometer she must accurately adjust the voltage on the lamp to the required value by means of a meter reading (and keep it at that value). She must then carefully read the luminous flux and the power consumption from two meters, enter the readings for each lamp without error in the inspection sheet, divide one value by the other to find the luminous efficiency (lumens per watt), divide this in turn into the luminous-efficiency figure specified for the particular type of lamp, and again record both results.

The equipment illustrated here, which is in use at Philips' new incandescent lamp factory at Weert (Netherlands), performs practically all these operations automatically. The operator has no meters to read, no calculations to perform and very little to record. She selects the voltage on a keyboard at her right on the table (not visible in the photo): there is a choice from 80 values between 90 and 287.5 V. An automatic control system keeps the lamp voltage at the selected value to within 0.1%. On the same keyboard she selects the luminous efficiency specified for the type of lamp under test and the appropriate (coupled) ranges on the watt and lumen meters (choice of six ranges). The selected values of voltage and specified luminous efficiency are automatically printed on the inspection sheet which the operator has inserted in the electric typewriter and on which, where necessary, she has typed a few identifying particulars of the type of lamp tested. After placing each lamp in the photometer (on the extreme left in the photograph) the operator presses a button. The system automatically allows the lamp to burn for a preset time, e.g. 10 seconds, sufficient for it to reach a steady state, after which the system automatically carries out the measurements, performs the calculations mentioned and prints out the results on the inspection sheet. The progress of these automatic operations is signalled by pilot lights on the keyboard.

The installation is composed of series-produced industrial equipment. The cabinet seen on the right contains in the left-hand part two automatic potentiometers for measuring the lumens and watts, one automatic potentiometer which acts as an analogue computer for the division operations, and — above — an analogue-digital converter, which



translates the continuously variable results of the measurements and calculations into the discontinuously variable signals required for driving the electric typewriter. A fourth automatic potentiometer works as a servo-system for the final adjustment of the lamp voltage (which thus takes only a few seconds). The standard of accuracy required makes it necessary to operate the lamps from a DC supply. The DC generator used (not shown in the photograph) is fed by a stabilized power pack, bottom left in the cabinet.

The right-hand part of the cabinet contains the programme switch and relays for controlling the automatic measuring and computing cycle, and all switches for adjusting the voltage, meter ranges and so on. The switches are operated by remote control from the above-mentioned push-buttons on the keyboard. Grouping them in one cabinet together with the appertaining meters made it possible to simplify the wiring. The equipment as a whole is checked twice a day by measurements on standard lamps.

RECENT IMPROVEMENTS IN SODIUM LAMPS

by M. H. A. van de WEIJER *).

621.327.532

Introduction

The development of sodium lamps in recent years has led to marked improvements in their luminous efficiency and to a much smaller decline in luminous flux during the life of the lamp. Better thermal insulation largely accounts for the higher efficiency; the more constant light output has been achieved by making the discharge tube from a new kind of glass, which is less susceptible to attack by sodium, and by improving its design.

Until recently the only type of sodium lamp in common use had a detachable vacuum jacket (Dewar flask) for thermal insulation (see *fig. 1a* and *b*). An advantage of this construction is that the old vacuum jacket can be used again when it becomes necessary to replace the discharge tube.

Although this type proved satisfactory for many years, and is still widely used, it has certain practical disadvantages. The first is the increasing absorption of light by the dust and dirt accumulated on the vacuum jacket, not merely on the outside but on the inside too. The latter is due to the far from hermetic seal between lamp and vacuum jacket. Every time the lamp heats up and cools down again, air is expelled and then again drawn into the jacket, bringing dirt with it.

A second drawback is the gradual decline in the thermal insulation provided by the vacuum jacket, owing to deterioration of the vacuum. Unlike the accumulation of dirt, this trouble is not immediately perceptible, but it can seriously impair the efficiency of the lamp, whether old or new.

To avoid these difficulties, designs were introduced some years ago in which the discharge tube was hermetically sealed within the insulating

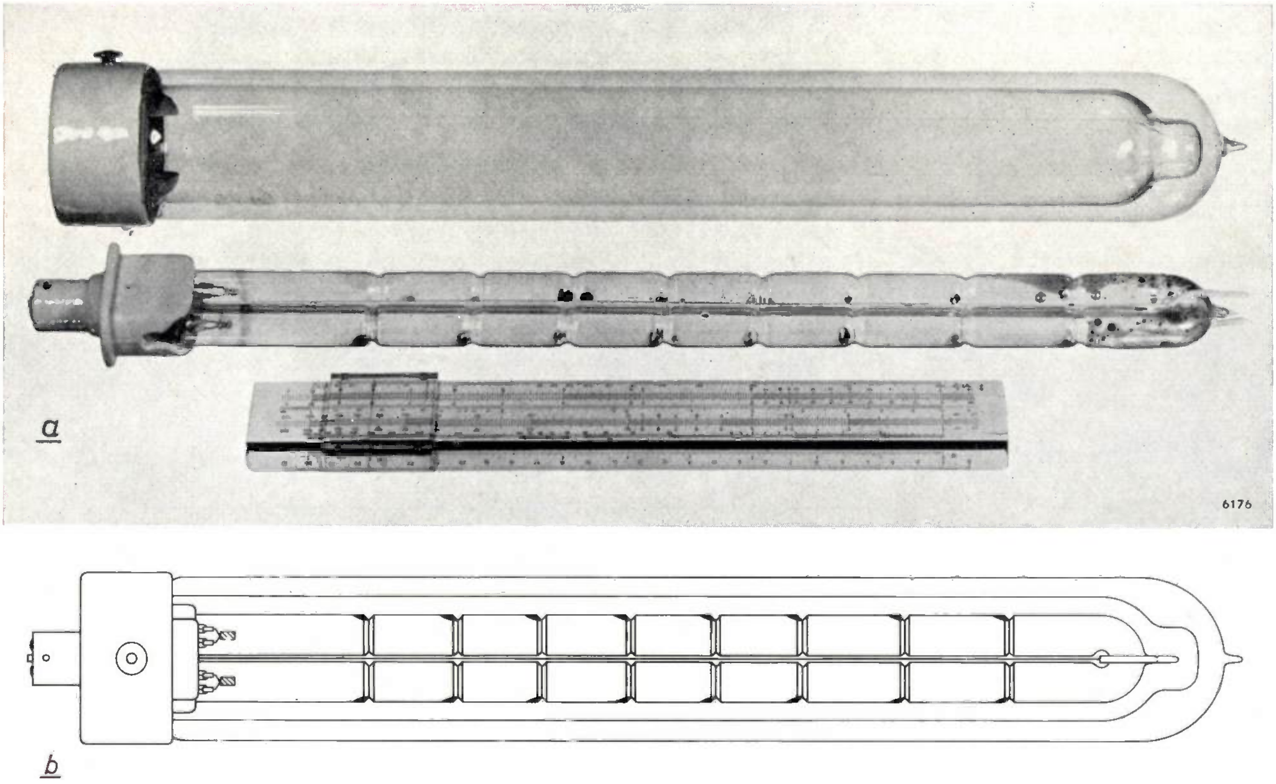


Fig. 1. *a*) Photograph and *b*) cross-section of a type SO 140 W sodium lamp. The U-shaped discharge tube is enclosed in a detachable vacuum jacket.

*) Philips Lighting Division, Development Laboratory, Turnhout (Belgium).

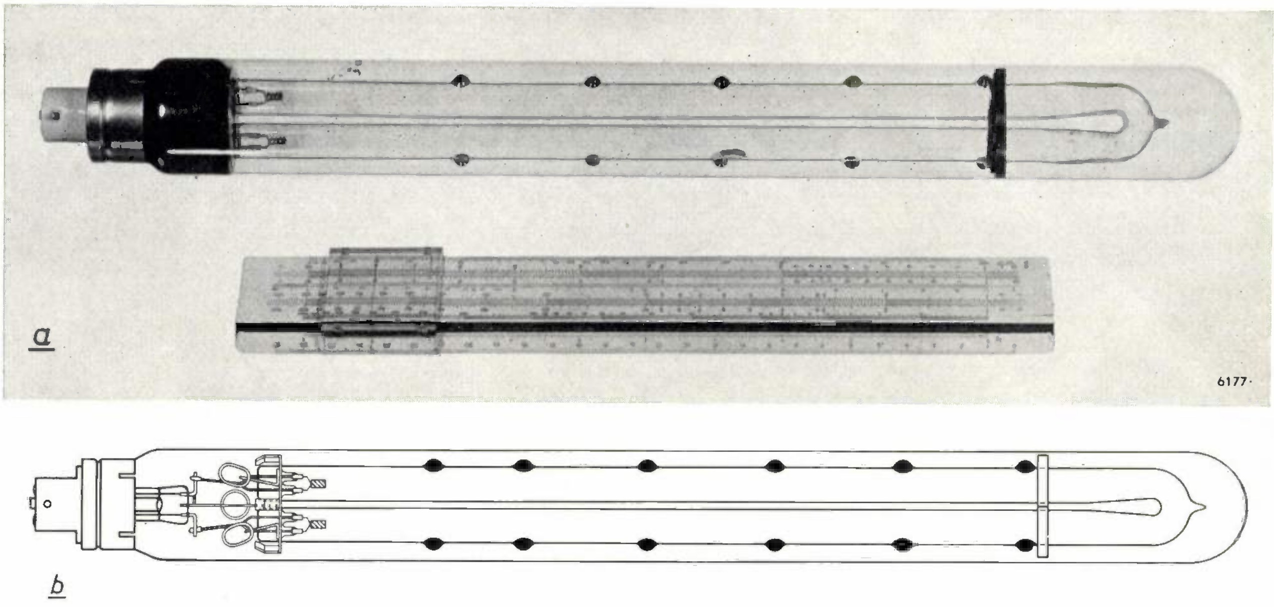


Fig. 2. *a*) Photograph and *b*) cross-section of an integral type SO 140 W sodium lamp, with single-walled non-detachable vacuum jacket. The discharge tube contains protuberances for holding the sodium in place.

jacket. This idea was by no means new, "integral" lamps of this kind having been marketed as long ago as the 'thirties [1]). It is only in recent years, however, that full benefit has been derived from the advantages of this more expensive construction, new techniques having made it possible to achieve higher efficiencies and a more constant luminous flux.

In this "integral" type of lamp, marketed by Philips in a more modern form in 1958, the separate double-walled vacuum jacket was replaced by a single-walled tubular envelope in which the discharge tube was permanently mounted. The space between discharge tube and envelope was evacuated. A modern gettering method made it possible to maintain the high vacuum throughout the life of the lamp. The discharge tube was provided with small protuberances for containing the sodium (*fig. 2a* and *b*) [2]). A range of these lamps was developed for ratings of 45, 60, 85 and 140 W. Lamps having a detachable vacuum jacket were immediately replaceable by lamps of the new type, giving a much more constant luminous flux.

In the further development of the integral sodium lamp it was found that a somewhat more complicated design offered even greater gains, both as regards efficiency and constancy of luminous flux. As a result of better thermal insulation, for example, it proved possible to achieve, and indeed exceed, the unprecedented efficiency of 100 lm/W in a sodium lamp suitable for practical use.

The improved thermal insulation was obtained by means of a separate glass "sleeve" fitted around the U-shaped discharge tube inside the evacuated bulb (*fig. 3a* and *b*). This sleeve radiates part of the heat emanating from the discharge tube back into the interior [3]).

In the lamps illustrated in fig. 2 the outer bulb also radiated energy back to the discharge tube. With the double-walled form shown in fig. 3, however, the effect is considerably greater, owing to the fact that the reflecting glass wall (of the sleeve) can get hotter than in the single-walled construction. Consequently the heat loss is smaller and the efficiency higher. This improved design is used in lamps rated for 45, 60, 85, 140 and 200 W, which were first marketed in 1960 under the type designation SOI.

Table I surveys the luminous efficiencies obtained with sodium lamps in the forms introduced in 1956, 1958 and 1960. The figures relate to 140 W lamps.

[1]) W. Uyterhoeven, Elektrische Gasentladungslampen, Springer, Berlin 1938, p. 216.

[2]) For further particulars of this type of lamp, see W. Verwey and M. H. A. van de Weijer, New sodium lamps, Communic. P-59.22 of the 14th Session of the International Commission on Illumination, Brussels 1959.

[3]) In 1955 the (British) General Electric Company brought out a sodium lamp with a narrow sleeve fitted around each limb of the U-shaped discharge tube.

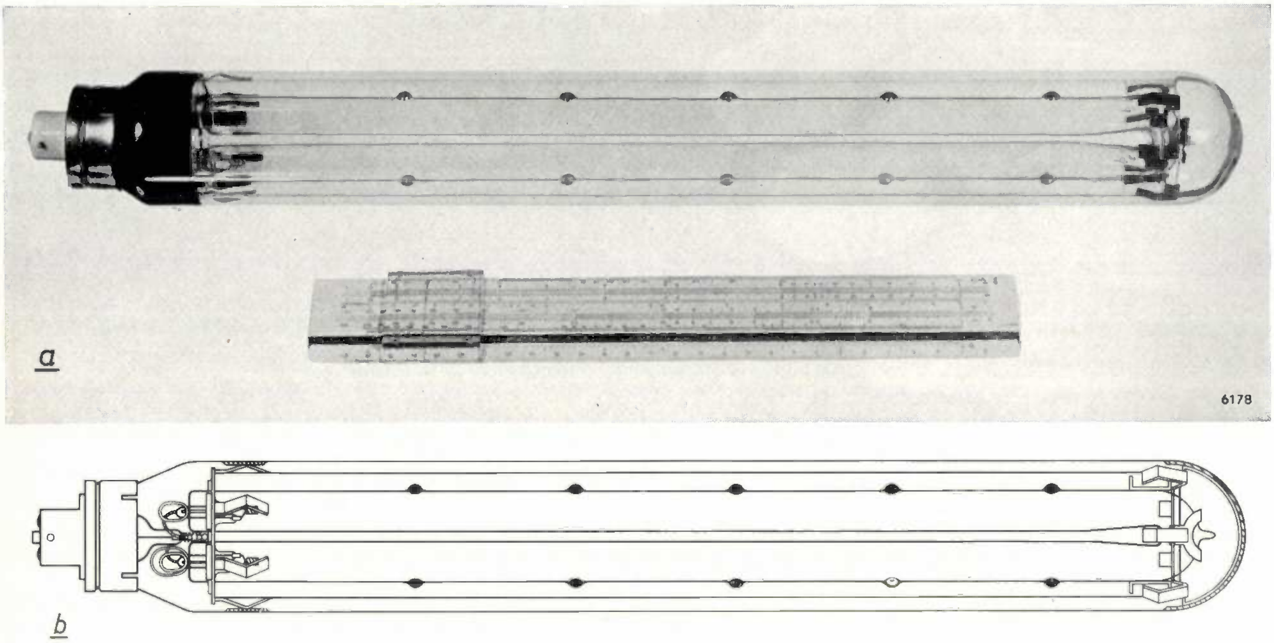


Fig. 3. *a*) Photograph and *b*) cross-section of a type SOI 140 W sodium lamp, with vacuum jacket and protuberances as in fig. 2. A sleeve for improving the thermal insulation surrounds the discharge tube.

Table I. Luminous efficiency of 140 W sodium lamps in designs of recent years.

Year	1956	1958	1960
Design	discharge tube in detachable vacuum jacket	discharge tube in single-walled vacuum jacket	discharge tube with sleeve in vacuum jacket
Luminous efficiency (lm/W)			
after 100 h	79	82	100
after 4000 h	52 *)	68	88

*) In a clean vacuum jacket.

To explain this development and give some idea of the further prospects of the sodium lamp, we must go somewhat deeper into the various factors governing the luminous flux, its decline during operation, and the life of the lamp. Principal among these factors are:

1) the thermal insulation around the discharge tube,
2) the rare gas added (for initiating and maintaining the discharge),
3) the shape and size of the discharge tube,
4) the composition of the glass of which the discharge tube is made.

Effect of thermal insulation on luminous efficiency and luminous flux

The operating principle of the sodium lamp is to excite as efficiently as possible the resonance radiation of sodium (wavelengths 589.0 and 589.6 nanometres). (1 nanometre (nm) = 10^{-9} m.)

The luminous efficiency is closely dependent on the vapour pressure of the sodium. Since all sodium lamps operate with saturated sodium vapour, the vapour pressure is determined by the temperature of the discharge tube. If the vapour pressure is too low (temperature too low), the number of sodium atoms capable of being excited is too small. If the vapour pressure is too high, self-absorption predominates, i.e. the sodium atoms absorb too much of the resonance radiation themselves. There is consequently one optimum vapour pressure, and that amounts to roughly 4×10^{-3} torr (1 torr = 1 mm Hg), corresponding to a tube temperature of about 270 °C [4]).

The temperature of the discharge tube is governed on the one hand by the power consumed, and on the other by the thermal insulation. Assuming that the same tube temperature of, say, 270 °C is desirable in all forms of sodium lamp, this means that if the thermal insulation is changed the power consumption must also be changed in order to maintain the optimum temperature: improved insulation calls for less power, and vice versa. In comparing the properties of particular lamp constructions, we take each lamp at its op-

[4]) Uyterhoeven, loc. cit. p. 205. Other lamp parameters affect this optimum temperature, e.g. the pressure of the added rare gas (see p. 252 of this article). This explains why other optimum values are sometimes mentioned in the literature.

timum working point. The construction with the best thermal insulation then has the highest luminous efficiency. To obtain good thermal insulation it is necessary to take measures to counteract losses due to convection and conduction as well as the total loss due to radiation.

Convection and conduction losses depend primarily on the vacuum around the discharge tube. In *fig. 4a* and *b* it can be seen how the efficiency η and the optimum power P depend on the residual pressure p_0 in the vacuum jacket when the lamps are not exposed to wind and when they are. The graph shows clearly that the vacuum inside the jacket has to meet high requirements. It is maintained by a barium film deposited on the inside of the outer bulb. Practically all the gases gradually released during operation are removed by chemical combination with the barium. The film is of course deposited at a position where it does not obstruct the emission of light.

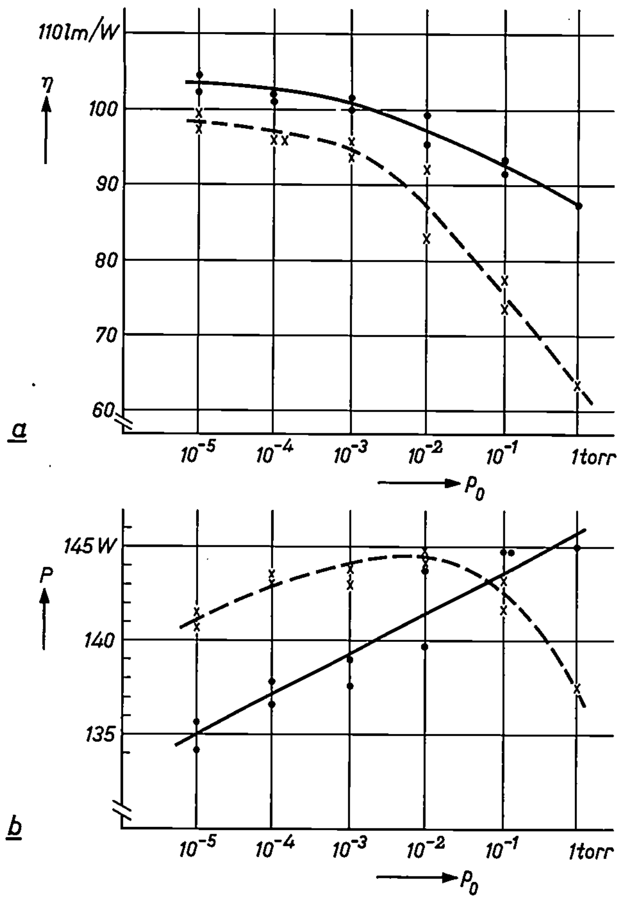


Fig. 4. *a*) Luminous efficiency η, and *b*) optimum power consumption P, of a sodium lamp (type SOI 140 W), as a function of the pressure p_0 of the gas (argon) in the outer bulb. The solid curves relate to operation out of the wind, the broken curves to exposure to wind.

The radiation losses can be limited in two ways:

a) Thermal energy radiated by the discharge tube is absorbed by a surrounding jacket. This consequently gets hot and radiates heat back to the discharge tube. The sleeves mentioned work largely along the same lines.

b) Thermal energy radiated by the discharge tube is directly reflected by the surrounding jacket. As we shall see, the discharge tube radiates mainly in the infrared, with a peak at about 5 μ. Since glass reflects only about 14% of radiation at these wavelengths, it is necessary in this case to coat the glass with some substance which is an effective reflector of infrared rays.

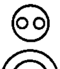
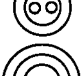
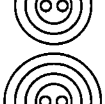
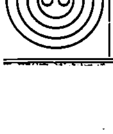
Both methods are discussed below.

Limitation of radiation losses by means of sleeves

The function of the sleeves, then, depends mainly on absorption and to a slight extent on the reflection of infrared radiation. The hot sleeve radiates outwards as well as inwards. The outward radiation can in turn be intercepted by a second sleeve around the first, and so on. Radiation losses can thus be progressively reduced by increasing the number of sleeves. In each case, however, the reduction becomes successively smaller, whereas the absorption of light (2.5% to 3% per sleeve) becomes more and more of a nuisance. These effects are illustrated in *Table II*.

The table clearly demonstrates that, as mentioned above, an improvement in insulation is accompanied by a smaller optimum power loading. As a result of this effect the luminous flux decreases more steeply than it would be increased by the greater efficiency if the power were constant. The

Table II. Influence of the number of sleeves, under optimum load (P_{opt}), on the luminous flux Φ_{la} and the luminous efficiency η_{la} of the lamp (including sleeves) and on the luminous flux Φ_{t} and efficiency η_{t} of the discharge tube. The figures relate to a U-shaped discharge tube having an inside diameter of 12 mm and an arc length of 666 mm.

Cross-section	number of sleeves	P_{opt}	lamp		disch. tube alone	
		W	Φ_{la} lm	η_{la} lm/W	Φ_{t} lm	η_{t} lm/W
	0	127	11300	89	11600	91.5
	1	96	9400	98	9900	103
	2	80	8150	102	8800	110
	3	70	7200	103	8000	114

sleeves thus improve the efficiency but reduce the luminous flux, and at the same time they make the lamp larger and more fragile. For these reasons no more than one sleeve is used in present-day sodium lamps.

Limitation of radiation losses by means of an infrared-reflective coating

For further improvement of the thermal insulation better results can be expected from the second method mentioned, i.e. the application of a coating around the discharge tube for effectively reflecting the infrared rays. A coating of this kind, applied to the sleeve or to the inside of the outer envelope, must of course readily transmit the sodium light.

The radiant power emitted in the infrared by the discharge tube is roughly equivalent to that from a black body having a temperature of about 545 °K (270 °C). At this temperature the spectral distribution of the radiant power from a black body is as shown in *fig. 5*. The coating to be applied must therefore be an especially good reflector at wavelengths of about 5 μ.

In about 1930 investigations were made at Philips into the usefulness of metallic layers as infrared reflectors for sodium lamps [5]). This work has been continued by Kauer of Philips Zentrallaboratorium GmbH, Aachen, and by Van Alphen of Philips Research Laboratories, Eindhoven. Without going into the theory, we shall mention here some of the results of the study made of metallic and metal-oxide layers.

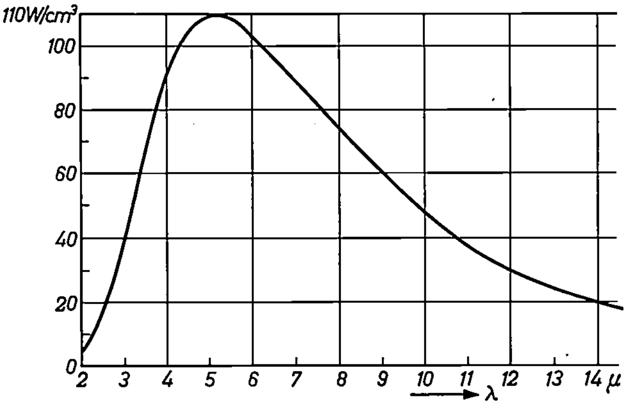


Fig. 5. Spectral distribution of the radiant power from a black body at 545 °K. The power in watts per square centimetre of radiating surface and per centimetre wavelength interval is plotted against the wavelength.

1) *Metallic layers. Fig. 6* shows the reflection coefficient of a layer of gold at a wavelength of 5.29 μ and the transmission coefficient at a wavelength of 0.59 μ (sodium light), as functions of the thickness of the layer.

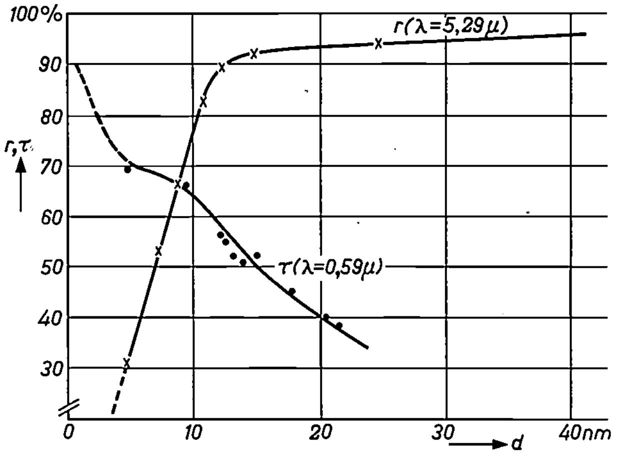


Fig. 6. The reflection coefficient r of plane gold layers at a wavelength $\lambda = 5.29$ μ, and the transmission coefficient τ at $\lambda = 0.59$ μ (sodium light), both as a function of the thickness d of the layers.

To decide on the thickness of the layer it is necessary to strike the most favourable compromise between the infrared reflection and the transmission of light. For this purpose a series of sodium lamps were made, differing only in the thickness of the layer of gold on the sleeve in each lamp. The lamps thus differed in their thermal insulation and hence in their optimum power loading. *Fig. 7* shows how the optimum power, the maximum luminous efficiency and the total luminous flux vary as functions of the thickness of the gold layer.

The variation of the luminous efficiency can be explained broadly as follows. At thicknesses less than 4 nm the reflection in the infrared is practically zero, although the layer already absorbs some light. Even layers as thin as this, therefore, reduce the efficiency. Between 5 and 15 nm there is a marked increase in infrared reflection. In spite of the accompanying increase in the absorption of light, the efficiency nevertheless rises with increasing thickness up to a maximum which, in this case, had the high value of 125 lm/W at 15 nm. In layers thicker than 15 nm the reflection coefficient shows no further rise of any significance and the increasing absorption of light predominates, as a result of which the efficiency declines.

Thus, although the application of a gold layer of the proper thickness can raise the luminous efficiency from about 100 to 125 lumens per watt, the gain is accompanied by a severe decline in light output, namely from about 14 000 to some 4000 lumens. That explains why lamps of this kind have not been put on the market. Efforts are still being

[5]) Austrian patent number 134 018, granted in 1933 in the name of W. de Groot.

made, however, to improve the transmission of light by coating the gold layer with another substance to reduce its reflection of light.

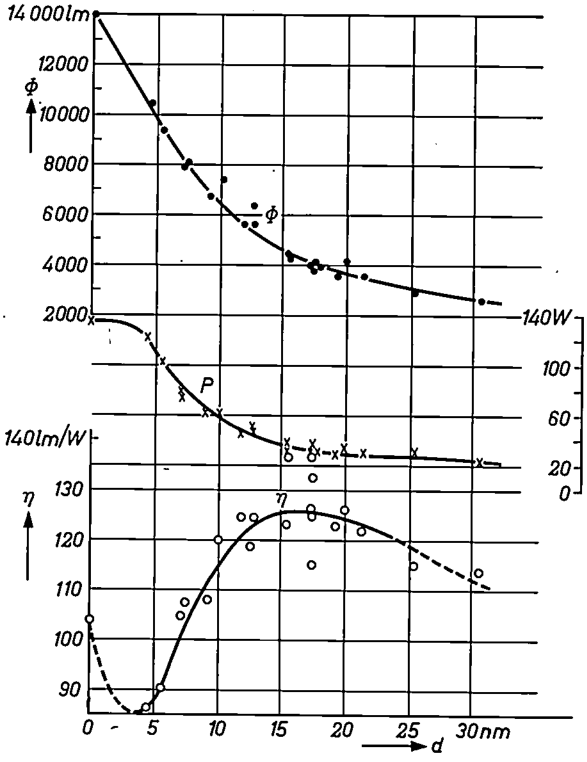


Fig. 7. Luminous flux Φ, lamp power P and efficiency η under optimum loading, as functions of the thickness d of a gold layer coated on the sleeve of a series of otherwise normal SOI 140 W sodium lamps (diameter of discharge tube 15 mm, arc length 820 mm, sleeve diameter 50 mm, outer bulb diameter 60 mm).

2) *Metal-oxide layers.* From the classical electron theory of metals it was already known that substances containing a fairly high concentration of free charge carriers would be good infrared reflectors. Substances exhibiting the required property were therefore to be expected amongst the semiconductors, provided that the concentration of free charge carriers could be made sufficiently large.

Investigations have been carried out on thin layers of metal oxides (on glass) which were known to combine transparency with a fairly high electrical conductivity [6]. The reflection coefficient is then dependent on the electrical conductivity as well as on the wavelength of the radiation. *Fig. 8* shows the reflection coefficient r of a layer of stannic oxide (SnO_2) at a wavelength of 5 μ, as a function of conductivity, together with the absorption coefficient a of the layer for sodium light.

A comparison of these curves for stannic-oxide layers with those for gold layers (fig. 6) shows that the light transmission of the stannic-oxide layers is markedly superior. It has thus proved possible with lamps of the type just described, but in which the sleeve is coated with stannic oxide instead of gold, to achieve the equally high efficiency of 125 lm/W [7]) with the considerably higher light output of about 8800 lumens (power consumption 70 W). A lamp of the same size, containing a gold layer, gives an equal efficiency with a light output of no more than 4400 lm (power consumption 35 W). It therefore looks as if the transparent semiconductor coatings offer better practical prospects than the metal layers, which are more reflective but absorb more light.

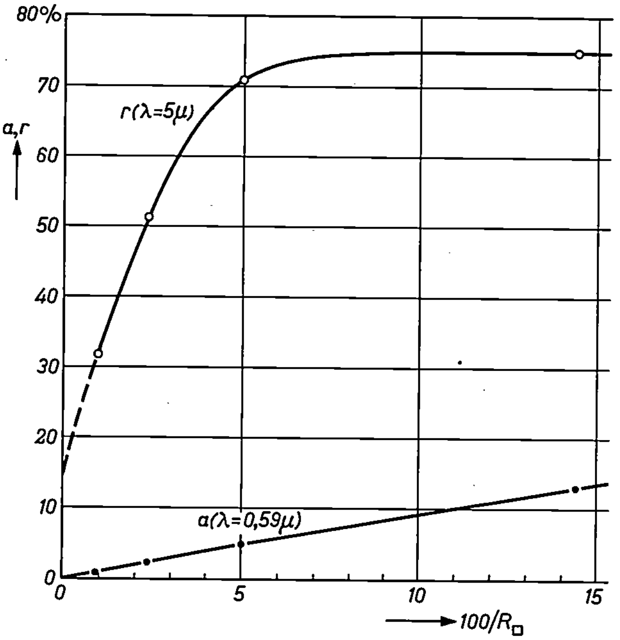


Fig. 8. Reflection coefficient r at $\lambda = 5\,\mu$ and absorption coefficient a for sodium light, as functions of the conductivity of stannic-oxide layers. The abscissa is $100/R_\square$, where $R_\square$ is the "resistance in ohms per square".

Finally it should be noted that the gain in efficiency obtained by using an infrared reflector always involves a lower lumen output, even if the infrared reflector were to transmit light for 100 %. This appears from *fig. 9a,* in which the measured luminous flux is corrected for the light absorption in the reflector (in this case a layer of gold). For this purpose separate measurements had previously been made of the light transmission of sleeves coated with gold layers of different thicknesses (fig. 9*b*).

[6]) This property makes such layers useful for other purposes, e.g. in electroluminescent panels; see Philips tech. Rev. **19**, 1-11, 1957/58.

[7]) In larger lamps, rated for 200 W, efficiencies as high as 140 lm/W have been reached.

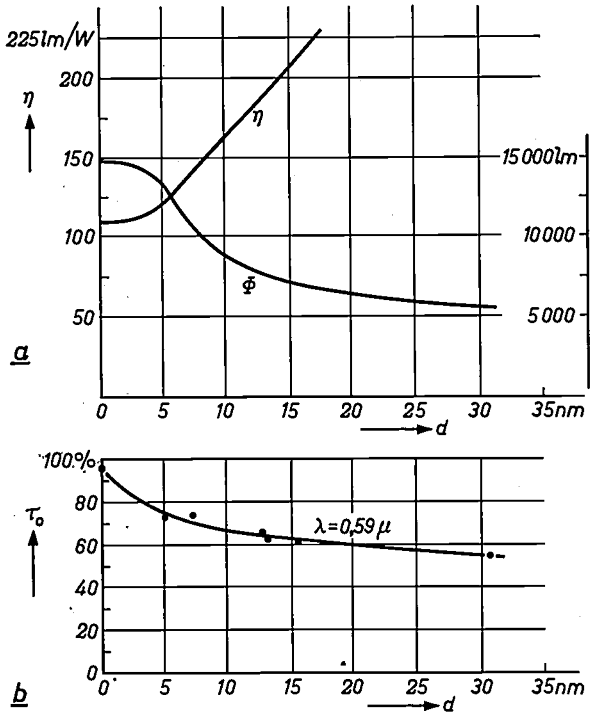


Fig. 9. *a*) Luminous flux Φ and efficiency η of a sodium lamp with a layer of gold coated on the sleeve, after correction for the absorption of light in the gold layer. The curves are derived from fig. 7 and fig. 9*b*.
b) Sodium-light transmission coefficient τ_0 of sleeve coated with gold layer, as a function of layer thickness *d*. The curve differs from that in fig. 6, which relates to a plane layer (single reflection), whereas the above curve relates to a sleeve, in which the light undergoes multiple reflections.

Influence of rare-gas pressure on luminous efficiency

The rare gas, which serves to initiate and maintain the discharge in sodium lamps, is given a pressure of about 10 torr. Under optimum power loading (temperature about 270 °C) the vapour pressure of the sodium is of the order of only 10^{-3} torr. In that case, then, there are roughly 10^4 times as many rare-gas atoms as sodium atoms present in the gas mixture. It is therefore evident that the kind of rare gas used and its pressure will have a considerable influence on the properties of the lamp. We shall confine ourselves here to mentioning some established relations between rare-gas pressure and certain properties of the lamp, such as luminous efficiency and power consumption.

The rare gas nowadays used in sodium lamps is neon, with a small admixture of argon and/or xenon to lower the ignition voltage. The effect of the pressure of a mixture of neon and argon on luminous efficiency has been investigated quantitatively by measurements on a type SOI 140 W lamp in the pressure range from 1 to 15 torr. The measurements showed that as the rare-gas pressure is decreased the maximum efficiency of the lamp increases fairly steeply: lamps with 15 torr reached 94 lm/W, lamps with 1 torr 114 lm/W. A striking circumstance is that the lamp, when operated at maximum efficiency, consumes less power at low rare-gas pressures than at high. This means that the optimum temperature of the discharge tube must also differ in these two cases, being lower at low powers than at high [8]).

We see, then, that if the rare-gas pressure is reduced, the sodium-vapour pressure — which is of course governed by the temperature of the discharge tube — must also be reduced in order to obtain maximum efficiency.

Fig. 10 illustrates how the luminous efficiency, the power consumed, the arc voltage and the lamp current vary with the rare-gas pressure under optimum loading conditions. The luminous flux Φ is also represented, and it can be seen that Φ in this case rises with increasing luminous efficiency, in

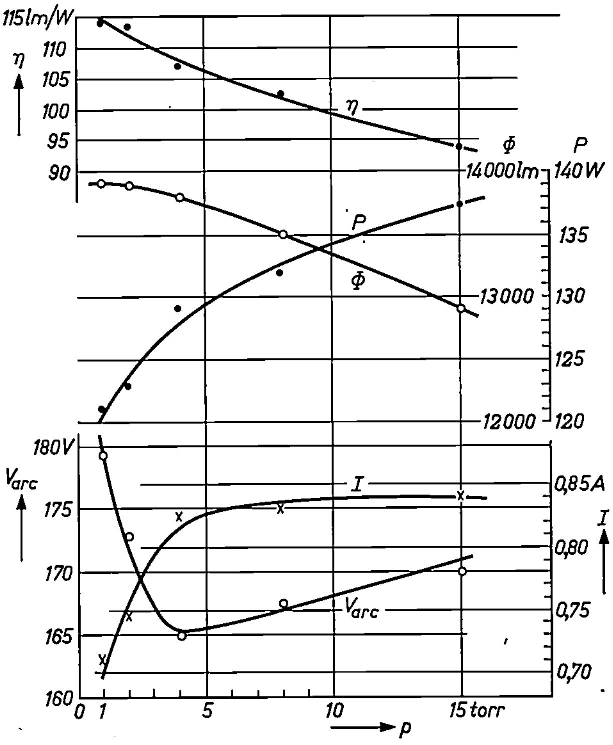


Fig. 10. Luminous flux Φ, lamp power P, efficiency η, arc voltage V_{arc} and lamp current I of an optimally loaded sodium lamp, type SOI 140 W, in dependence on the rare-gas pressure p (neon plus 1% argon).

[8]) We have not yet confirmed this by direct measurements, but the conclusion is reasonable, especially considering that the lowest power loading gives the highest efficiency, so that a smaller fraction of this low power is converted into heat.

spite of the lower power needed for optimum loading [9]).

The improvement in luminous efficiency which, as shown by these experiments, can be achieved by lowering the rare-gas pressure is unfortunately possible only within strict limits. It will presently be made clear why the pressure of the filling cannot be arbitrarily low.

Effect of arc length on luminous efficiency

If only the length of a sodium discharge is varied and the current is kept constant, the efficiency increases as the arc is made longer. This familiar effect is due to the fact that the electrode losses are relatively less significant in a long arc than in a short one. *Table III* shows how various properties of the lamp vary as functions of arc length. As we shall presently see, it is particularly important to take this effect into account when there is any reason to make the lamp shorter.

Table III. Arc voltage V_{arc}, power consumption P and luminous efficiency η of a sodium discharge tube as functions of arc length. Diameter of the tube 12 mm, current 0.6 A.

Arc length mm	V_{arc} V	P W	η lm/W
280	82	45.5	80.7
405	109	61.2	89.7
615	156	85.6	98.0

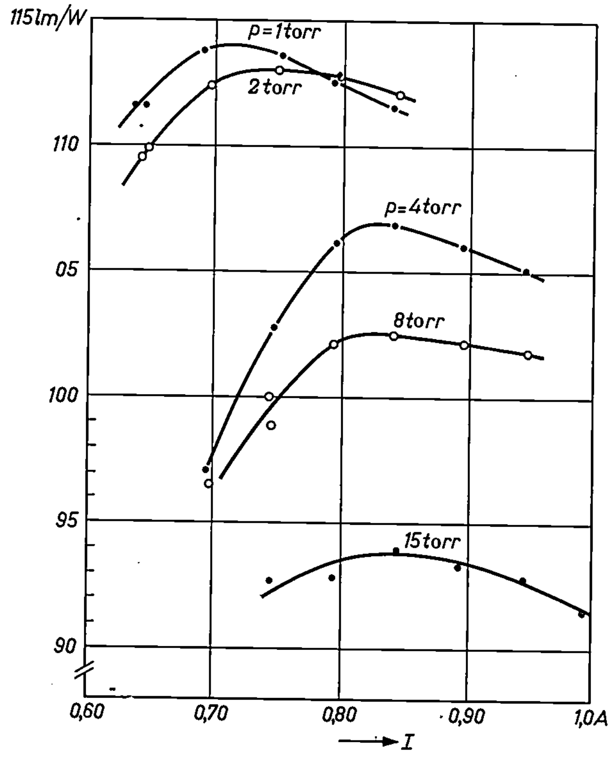


Fig. 11. Efficiency of SOI 140 W sodium lamps as a function of lamp current I, with the pressure p of the rare-gas filling (Ne + 1% A) as parameter. The maxima of the curves correspond to optimum loading conditions.

Sodium migration

Decline in light output as a consequence of sodium migration

Owing to the relatively high pressure of the rare gas in the lamp, each droplet of sodium only supplies sufficient vapour pressure for its own immediate surroundings. For this reason it has always been customary to distribute a fairly large surplus of sodium uniformly in drops over the whole surface of the discharge tube. Usually, however, the even distribution is not maintained. This is due to various causes, the main one being the differences in temperature along the wall of the tube, as a result of which the sodium tends to distil from hot to cold spots. A consequence of this migration is that the average vapour pressure of the sodium diminishes during the life of the lamp from its original, optimum value. Ultimately, all the sodium accumulates at the coldest spot in the lamp, and the vapour pressure drops to a minimum. A state may be reached where large parts of the discharge tube have hardly any sodium vapour, all the liquid sodium having accumulated at one point. The emission of sodium light from these parts is then virtually zero [10]).

Limitation of sodium migration by protuberances in the discharge tube

Sodium migration can be substantially reduced by applying the sodium droplets, during the manufacture of the lamp, to defined points along the tube that remain relatively cold. The points in question may be small protuberances in the tube wall, sufficient to hold a sodium droplet. This principle has been adopted by Philips in their more recent types of sodium lamps (figs 2 and 3).

A different method has been used in the "Linear Sodium Lamp", recently brought on to the market by the A.E.I. Lamp and Lighting Co. The tube wall in this lamp contains a number of depressions, at which positions the cross-section of the tube is not circular and has relatively cold spots where the sodium, once applied, is held in place [11]).

Suppression of sodium migration by the appropriate choice of rare-gas pressure

If for some reason or other there is a deficiency of sodium vapour in a part of the discharge tube,

[9]) First the optimum power loading was determined at every pressure by plotting efficiency versus lamp current (*fig. 11*)

[10]) Uyterhoeven, loc. cit. p. 251.

[11]) R. F. Weston, High-output sodium lamps, Electrical Times **135**, 719-722, 1959.

the discharge at that part is sustained almost entirely by the rare gas. At that part the voltage gradient is usually greater than in areas where the sodium participates properly in the discharge. Owing to the steeper gradient the part deficient in sodium consumes more power per unit length. Because of this the region where the sodium is lacking gets hotter than those well supplied with sodium, and this tends to encourage further migration. The migration thus has a cumulative character and the temperature equilibrium that initially existed becomes unstable.

It is possible, however, to choose the rare-gas filling in such a way that the power dissipated per unit length is lower in the parts deficient in sodium than in the parts well supplied. Contrary to the case just mentioned, the deficient parts then heat up less than the other parts, and sodium migrates back to where a shortage existed. The temperature equilibrium in such a tube is stable. The two cases are represented schematically in *fig. 12a* and *b*.

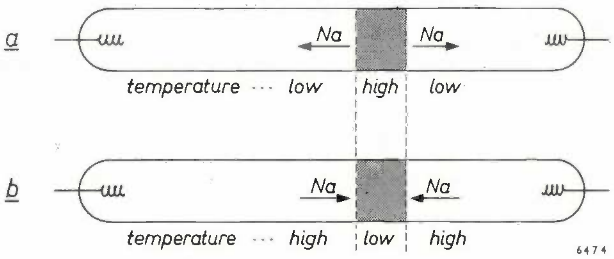


Fig. 12. Diagrammatic representation of sodium migration, (*a*) cumulative, (*b*) reduced. The shading in both cases denotes a zone deficient in sodium.
a) Classical case of relatively high rare-gas pressure. The low-sodium zone is hotter than the adjacent zones, which tends to promote migration.
b) When the rare-gas pressure is kept within specific limits, the low-sodium zone remains colder than the adjacent zones, which reduces or even eliminates migration.

This effect, in dependence on the rare-gas pressure, has been studied quantitatively on lamps containing neon plus 1% argon. The state where no sodium vapour takes part in the discharge was investigated by measuring, at room temperature and without giving the lamp a chance to get hot, the voltage gradient of the column and the power dissipated per centimetre length of column. The data thus obtained relate to the low-sodium discharge (rare-gas characteristic) [12]. The lamp was then allowed to heat up, and measurements were made of the discharge characteristic in the mixture of rare gas and sodium vapour. The arc voltage and the power consumption were determined in dependence on the rare-gas pressure for both the cold and the hot lamp.

The results of measurements on type SOI 140 W lamps are presented in *fig. 13a* and *b*. It can be seen from fig. 13*a* that at all rare-gas pressures the r.m.s. value of the arc voltage is higher in the pure rare gas than when sodium vapour is also present. Fig. 13*b* shows that there is a range of pressures in which, notwithstanding the higher arc voltage, the power consumed is lower than in the mixture of rare gas and sodium vapour. This is due to the fact that the wave-forms of the voltage in the two discharges are different (see the oscillograms in *fig. 14a* and *b*), so that the form factor — and hence the power factor α — is greater when the discharge

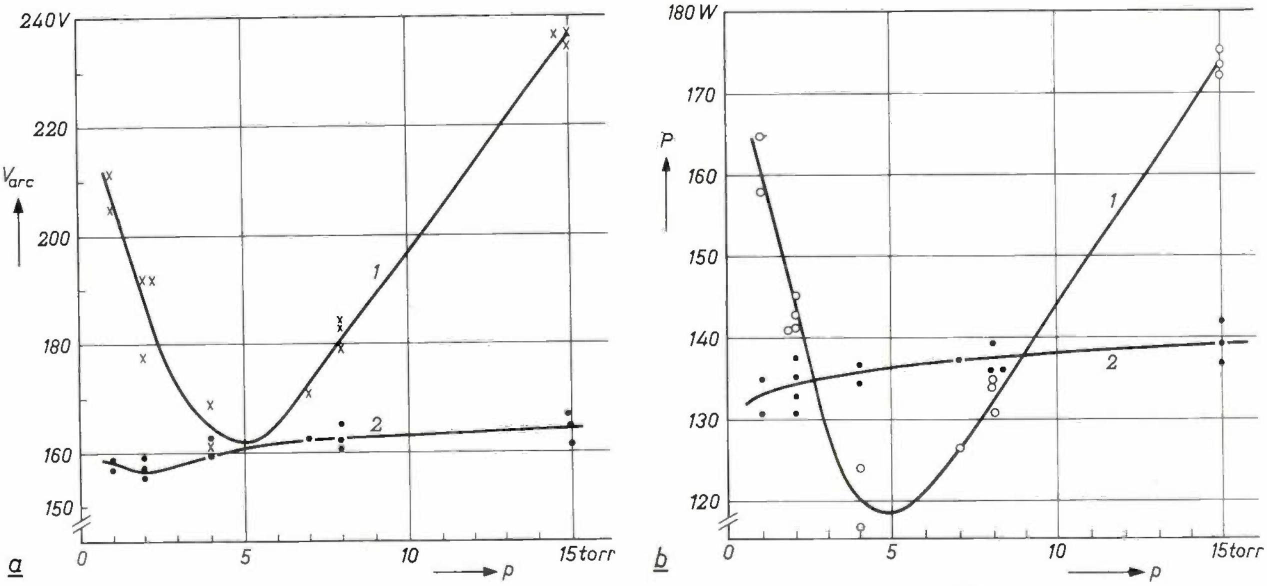


Fig. 13. *a*) Arc voltage V_{arc} (r.m.s. value) and *b*) lamp power P as a function of the pressure p of the rare-gas filling (Ne + 1% A); curves *1* without and curves *2* with sodium vapour. Measured on a standard SOI 140 W sodium lamp.

[12] A discharge tube containing no sodium was investigated to ascertain whether the rare-gas characteristics differed at 20 °C and 270 °C. This proved not to be the case.

takes place in a mixture of rare gas and sodium vapour than when no sodium vapour is present. The values measured on the above-mentioned lamps were $a = 0.83$ for the discharge in rare gas alone, and $a = 0.94$ for the discharge in rare gas with sodium vapour.

In the region from 3 to 8 torr it is reasonable, in view of the temperature differences caused, to expect self-stabilization of the sodium distribution. In this way, then, the decline in light output due to sodium migration is effectively reduced.

For simplicity we have disregarded here the effect of the rare-gas pressure on the optimum working point of the lamp. In fact, as indicated, the lamp currents for maximum efficiency must be smaller at lower rare-gas pressures. This has no influence, however, on the effect described, the pressure region concerned remaining virtually unchanged.

A life test on type SOI 140 W lamps, with rare-gas pressures of 6 and 9 torr, confirmed the stabilizing effect of the lower pressure. The measured efficiencies are collected in *Table IV*. After 5000 hours, marked migration was observed in the lamps with 9 torr, whereas in the lamps with 6 torr hardly any sodium was found outside the original points.

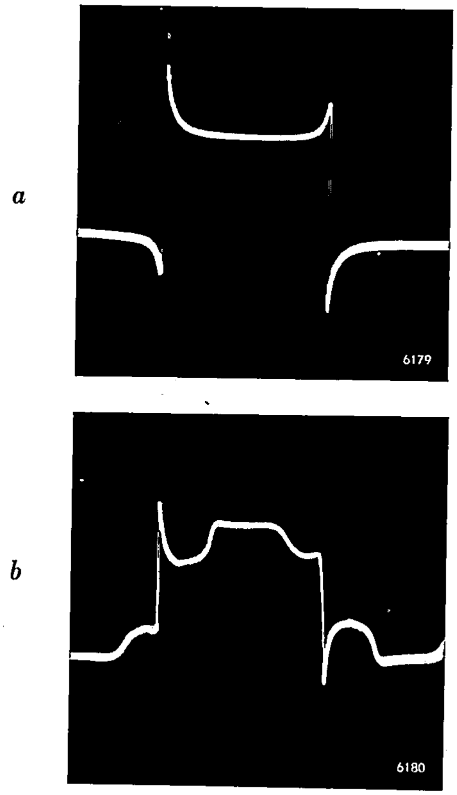


Fig. 14. Oscillograms of the arc voltage of a discharge in rare gas (Ne + 1% A, pressure 8 torr), *a*) without, *b*) with sodium vapour. Measured on a standard sodium lamp type SOI 140 W, lamp current 0.9 A. The difference in wave-form explains the smaller form factor in *a*) than in *b*).

Table IV. Efficiency of type SOI 140 W sodium lamps, during a life test with rare-gas pressures of 9 and 6 torr (Ne + 0.5% A).

Rare-gas pressure	efficiency in lm/W, measured after					
	100 h	1000 h	2000 h	3000 h	4000 h	5000 h
9 torr	97.7	94.6	91.9	88.2	81.7	81.4
6 torr	102.7	97.9	96.0	94.6	95.8	93.9

Decline in light output owing to glass discolouration

Sodium is chemically an extremely aggressive substance, particularly in vapour form, and quickly attacks all ordinary kinds of glass. The surface layer of the glass exposed to sodium vapour takes up a large quantity of sodium, and as a result turns brown. In certain cases concentrations of 2.1×10^{22} sodium atoms per cm^3 have been found in the brown layer; this implies that the average distance between the sodium atoms in the layer is less than twice that in metallic sodium [13]).

The discolouration of the glass obviously entails considerable absorption of light. Moreover, due to the uptake of sodium the composition of the glass changes and so too therefore does the coefficient of expansion; the consequent stresses may become so high as to crack the glass, prematurely ending the life of the lamp.

To avoid these effects, special kinds of glass have been developed to withstand the influence of sodium vapour under the conditions prevailing in a sodium lamp, without any significant discolouration. However, a serious drawback of these non-discolouring types of glass is the marked extent to which, in general, they adsorb argon.

Life of the gas filling

As stated, the rare-gas filling in present-day sodium lamps is neon with small admixtures of argon or xenon, or both. These admixtures, which are essential for lowering the ignition voltage [14]), steadily diminish in concentration during the life of the lamp, owing to adsorption, particularly at the glass wall. The gas filling will finally consist of almost entirely pure neon, resulting in such a high ignition voltage that the lamp can no longer be started by the available open-circuit voltage. The lamp is then said to have reached the end of its "gas life" (as opposed, for example, to the "cathode life").

[13]) J. W. Wheeldon, Absorption of sodium and argon by glass, Brit. J. appl. Phys. 10, 295-298, 1959.

[14]) F. M. Penning, Über Ionisation durch metastabile Atome, Naturwiss. 15, 818, 1927; Über den Einfluss sehr geringer Beimischungen auf die Zündspannung der Edelgase, Z. Phys. 46, 335-348, 1927/28.

The rate at which the glass wall adsorbs rare gases depends primarily on the following factors:

1) the pressure of the rare gas,
2) the voltage gradient in the column,
3) the composition of the glass used for the discharge tube.

When the lamp is burning, ions are formed not only from the sodium but also from the rare gas, in particular from the argon, which has a lower ionization voltage than neon. These (positive) rare-gas ions are attracted to the negatively charged glass wall, which they strike with a certain energy. If the energy upon impact is high enough, the ions — possibly after recombination with electrons — are trapped in the wall. The energy at which the ions impinge upon the wall increases with the field strength, i.e. with the voltage gradient in the column, and also with the free path, i.e. with decreasing rare-gas pressure. Whether the wall in fact continues to hold the ions depends further on the structure of the glass.

The rare-gas pressure needed for the lamp to attain a specific life is plotted in *fig. 15* as a function of the voltage gradient in the column, for two different compositions of glass. The lamps in question were experimental types filled with neon plus 0.5% argon. The curves show that the gas life is longest when the gas pressure is relatively high and the gradient in the column small; the composition of the glass is seen to have a considerable influence.

Of course it is a simple matter to employ a high gas pressure, but, as we have seen, this leads to low efficiency and a marked decline in light output during the life of the lamp. Both reasons are in fact a strong argument for a low pressure.

A low voltage gradient in the column can be obtained by using a wide discharge tube. If the sodium lamp is to have the correct operating temperature the choice of a wider tube must be associated with a higher current, i.e. with a lower lamp voltage at a specified power loading; in other words the lamp must be made shorter. A lower lamp voltage, however, entails relatively higher electrode losses and hence a lower efficiency (see Table III). The scope for lowering the voltage gradient is thus restricted.

The crux of the problem of achieving a satisfactory gas life is therefore the composition of the glass.

As remarked above, those types of glass that are not turned brown by sodium vapour generally have the disadvantage of strongly adsorbing rare gases. Recently, however, types of glass have been developed that show relatively favourable properties in both respects [15]). With glass of this kind it has

[15]) Due in particular to the work of C. M. La Grouw, Glass Development Centre, Eindhoven.

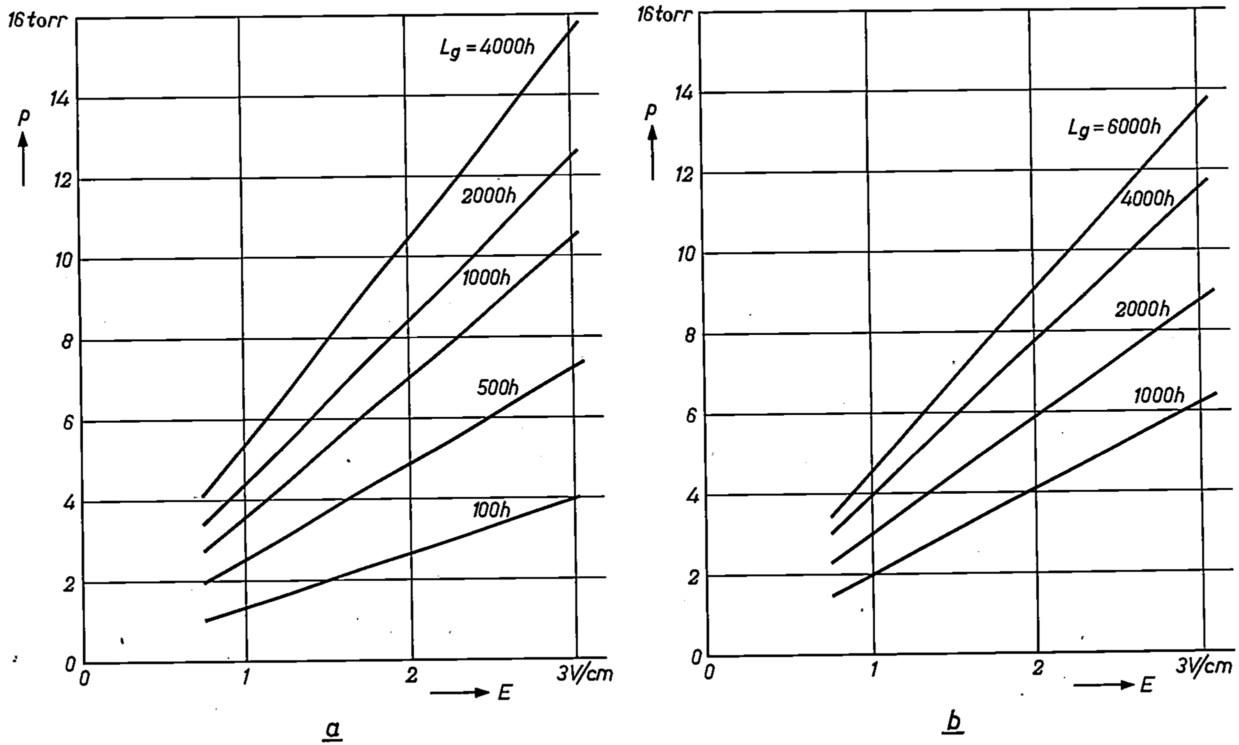


Fig. 15. The lowest rare-gas pressure p (Ne + 0.5% A) needed for a sodium lamp to attain a specific "gas life" L_g, as a function of the voltage gradient E in the column. The lines b relate to a better type of glass (adsorbing less argon) than lines a.

proved possible to reduce the pressure of the neon-argon mixture to 8 torr for a gradient of 1.8 V/cm in the column, ensuring a gas life of at least 6000 hours.

Future prospects

We shall now briefly examine the further evolution of sodium lamps that may be expected in the not too distant future.

Thermal insulation

From what has been said in the foregoing, it will be evident that there is still ample scope for refinements in the use of infrared-reflecting layers. It is by no means unlikely that advances in this direction will make it possible to produce sodium lamps giving an efficiency of 150 lm/W or even more. These lamps will be bulky, however, since if the thermal insulation is improved the optimum temperature of the lamp can only be maintained by reducing the power loading of the discharge tube per unit surface area.

Another potential development, running parallel with this, is therefore the construction of sodium lamps that combine a high lumen output with a relatively small size. The thermal insulation of such lamps would have to be deliberately poor in order to make a high power rating possible. For instance, a type SOI 140 W discharge tube, without an insulating jacket, has been found to give a luminous flux of more than 19 000 lm at 300 W, representing an efficiency of scarcely 60 lm/W. By means of a jacket offering very good thermal insulation and light transmission, the efficiency of the same discharge tube might be raised to perhaps 200 lm/W; the power consumed would then be only 35 W and the light output 7000 lumens. These widely divergent possibilities are probably both of importance, since lamps with the emphasis on efficiency can be developed at the same time as others with the emphasis on light output.

Rare-gas pressure

The use of lower filling pressures than have hitherto been feasible may lead to a further increase in efficiency, together with a luminous flux that will remain practically constant throughout the life of the lamp. These advantages might be achieved with a neon-argon pressure of 5 or even 4 torr. Such low pressures are not yet feasible because of the too rapid adsorption of argon by the glass in its present composition. It may be assumed, however, that the last word has not yet been said on the development of types of glass that are not turned brown by sodium vapour whilst at the same time adsorbing little rare gas.

If the filling pressure could be reduced to about 2 or 3 torr, the quantity of sodium used per lamp could also be considerably reduced, so that only a very small surplus of liquid sodium would be needed at the working temperature (just as the surplus of fluid mercury is very small in tubular fluorescent lamps). The present protuberances in the discharge tube (figs 2 and 3) could then be dispensed with, without the risk of any troublesome migration.

If, instead of a mixture of rare gases, one pure rare gas were to be used, it would already be possible to operate with a very low filling pressure, with all its attendant benefits. True, the lamps would then have a very high ignition voltage, but with a suitable ballast this need be no insuperable obstacle. The length to which one could go in this direction is limited by the life of the cathode, which is shortened when the filling pressure is reduced.

Use of other rare gases

After a thorough study has been made of the properties of gas discharges in mixtures of rare gases and sodium vapour, it may well be concluded that other rare gases offer more advantages as a filling than the classical neon. In cases of poor thermal insulation, for example, helium has proved to result in a higher efficiency than neon.

Renewed fundamental research into the sodium discharge should narrow the gap that still exists between the efficiency achieved in practice and the maximum efficiency theoretically possible, which is in the region of 450 lm/W. Notwithstanding the results achieved, the width of that gap makes it plain that we still have a long way to go.

Summary. Recent improvements in the sodium lamp concern in particular the luminous efficiency. With a lamp suitable for practical use the unprecedented efficiency of 100 lm/W has now been reached and even surpassed.

The author discusses the background of these advances and of others expected in the future, with particular reference to the influence of thermal insulation, rare-gas pressure and glass composition.

The use of infrared-reflecting layers on the glass — especially of transparent semiconductors, such as stannic oxide — is likely to result in considerably higher efficiencies, possibly up to 150 lm/W, though at the expense of making the lamps bulkier.

A promising trend is the development of types of glass that are not turned brown by sodium vapour and also adsorb little argon. Further progress in this field should lead to rare-gas pressures low enough for the light output of sodium lamps to remain almost constant throughout their life. The present protuberances in the discharge tube, which serve to hold the liquid sodium, can then be dispensed with.

Apart from the development of lamps of extremely high efficiency, lamps of lower efficiency but with smaller dimensions and a high light output are also to be expected.

LIGHTING OF TRAFFIC ROUTES

by J. B. de BOER *).

628.971.6

This article sets out to provide a survey of present-day views on the lighting of roads for motorized traffic. The first part deals with various conditions with which a road lighting installation should comply and discusses the more important experiments and considerations that have led to the drawing up of codes of practice or recommendations for meeting these conditions.

The light sources mainly used nowadays for traffic routes are incandescent (tungsten-filament) lamps, sodium lamps, high-pressure mercury-vapour lamps and fluorescent lamps. In the second part of this article these kinds of light source are compared at various points. The first point relates to the kind of light used, i.e. the spectral composition, and we shall see what bearing this has on the recommendations discussed. The second point concerns the significance of the dimensions, shape and luminance of the light sources, and the third their sensitivity to variations in temperature and voltage.

Recommendations on the lighting of traffic routes

The primary object of road lighting is to provide motorized traffic with suitable conditions for vision when daylight is no longer adequate. The lighting should preferably enable vehicles to proceed safely without the use of headlights. Experience and special experiments have shown that for this purpose there are three conditions to be fulfilled: the average road-surface luminance must be sufficiently high, the lanterns of the installation must not cause troublesome glare, and the road-surface luminance must be sufficiently uniform. We shall deal with these three conditions in turn. In doing so we shall distinguish between two aspects, namely "visual comfort" and "perceptibility". It has become increasingly clear that reasonably satisfactory perceptibility is no guarantee that the driver can also see comfortably. Visual comfort is not something that can be measured exactly. All one can do is to ask numerous observers for their appraisal of a given lighting situation, or allow each observer to alter the lighting situation to suit his own visual comfort. As may be expected, the appraisals of the same situation by different observers — or of a given situation by one observer at different times — differ considerably. Useful results can only be obtained by careful statistical analysis of a mass of data.

Average road-surface luminance

As regards the average road-surface luminance, we consider a value of about 2 cd/m^2 (0.6 footlambert) as a reasonable compromise between what is desirable and what is technically and economically feasible. Depending on the expected depreciation in light output due to dust and dirt, the design should then be based on 3 to 4 cd/m^2. Installations designed before 1955 seldom exceed 0.5 cd/m^2 under operating conditions, i.e. after a certain degree of dirt accumulation. From the results of a number of experiments we shall show that 2 cd/m^2 for traffic routes is certainly not too much to ask.

To obtain some idea of the road-surface luminance at which motorists feel the need to use their own lighting, i.e. to switch over from side lights to dipped headlights, observations were made at dusk of vehicle lighting as a function of the average luminance of the road surface. The results are presented in *fig. 1*. Curves *O*, *P*, *C* and *R* relate to observations on unlighted roads with little traffic, the latter to exclude as far as possible the interfering effects of oncoming traffic. Curve *C* (percentage with dipped headlights) shows that at 2 cd/m^2 only 10% thought the assistance of their own lights necessary.

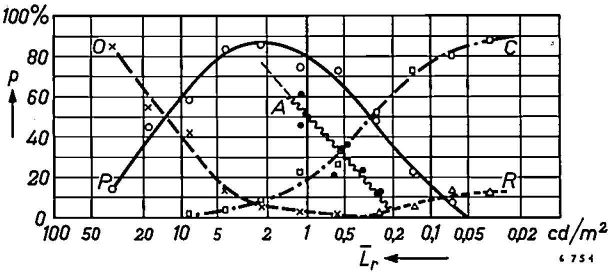


Fig. 1. Observed percentage *p* of cars using various kinds of lighting as a function of the average road-surface luminance $\bar{L}_r$. Curves *O*, *P*, *C* and *R* relate to unlighted roads at dusk: *O* = no lights, *P* = side lights, *C* = headlights on passing beam, *R* = headlights on driving beam. The wavy line *A* relates to cars driven at night with side lights on artificially lighted roads. (1 cd/m^2 = 0.292 footlambert.)

*) Philips Lighting Division, Lighting Laboratory, Eindhoven.

The problem was approached from the other side by making recordings of drivers travelling at night with side lights on artificially lighted roads. Extrapolation of the wavy line *A*, which represents the results of these recordings, shows that at 2 cd/m^2 about 75% of the drivers felt they were able to do without their headlights.

Further data on the desired average luminance of the road surface was obtained by asking a group of 16 observers (mostly engineers responsible for public lighting in Dutch towns) for their appraisal of the "lighting level" provided by the lighting installations in 70 streets [1]. The observers were also asked to assess the general impression, the non-uniformity of the luminance of the road surface and the glare caused. An appraisal of the lighting level amounts to assessing the road-surface luminance, while trying not to be unduly biased by unevenness in this luminance or by glare. The observers could express their assessment in the ratings bad, inadequate, fair, good or excellent, or with some intermediate qualification. At the same time the average road-surface luminance was measured. *Fig. 2* shows the result of the assessments of the lighting level. Each point represents the average of 16 assessments of one street. (For the purpose of averaging, the odd numbers from 1 to 9 were assigned to the successive ratings.) Through the 70 points thus obtained a curve was drawn by a common statistical method. The value of the average road-surface luminance corresponding, according to this curve, to the rating "good" (number 7) amounts to 1.5 cd/m^2 . Statistical analysis has demonstrated that this value lies between 1.3 cd/m^2 and 1.8 cd/m^2 with a reliability of 95%. It may be noted in this connection that observers very probably rate the lighting level in a street higher the less important the street is for traffic. Of the 70 streets in question 28 were of little importance to vehicular traffic, so that the figure of 1.5 cd/m^2 is probably on the low side. At all events, these experiments too show that 2 cd/m^2 is a reasonable value to aim at.

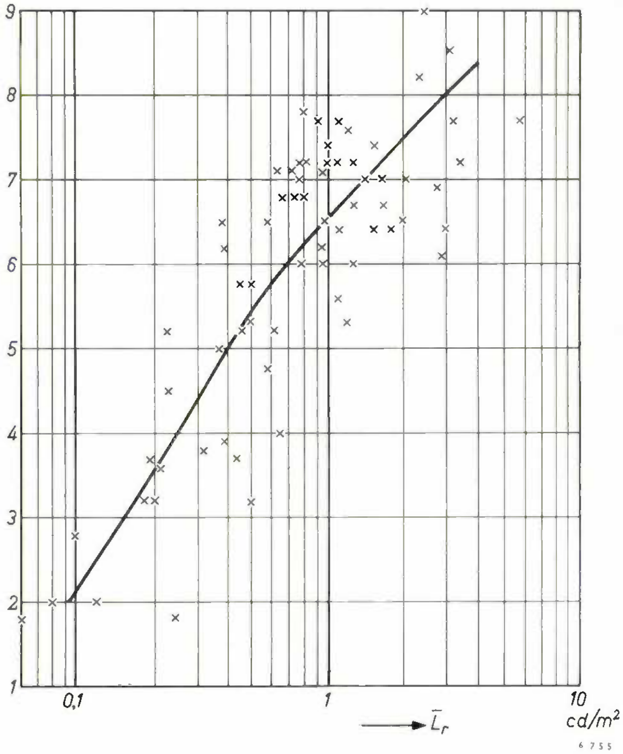


Fig. 2. Average assessment of the "lighting level" in 70 streets by a group of 16 experts. The average road-surface luminance $\bar{L}_r$ is plotted versus the following ratings: 1 = bad, 3 = inadequate, 5 = fair, 7 = good, 9 = excellent.

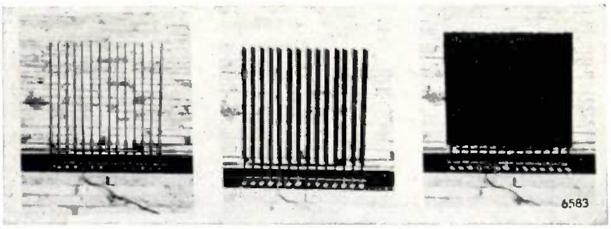


Fig. 3. Objects measuring 28 × 28 cm used in road perceptibility tests. The (average) luminance of the object was adjusted by means of the position of the vanes, varied by remote control. The object was not visible to the observer when all vanes were in line with him.

We have so far been concerned with visual comfort, and we now turn to the question of perceptibility. Numerous experiments have been carried out to investigate perceptibility as a function of average road-surface luminance. We have untertaken investigations of this kind on an experimental road equipped with a special lighting installation [2]. A number of observers were stationed at a fixed point along this road. At distances between 50 and 200 m from these observers, objects measuring 28 × 28 cm were made to appear and disappear on the road at places and times that were not known to the observers. These objects (*fig. 3*) were observed against the background of the road surface. The average luminance $\bar{L}_r$ of the surface and the ratio L_r/L_0 (L_r being the luminance of the section of road

[1] J. B. de Boer, F. Burghout and J. F. T. van Heemskerck Veeckens, Appraisal of the quality of public lighting based on road surface luminance and glare, Proc. Int. Comm. on Illumination, Brussels 1959.

[2] J. B. de Boer, Fundamental experiments on visibility and admissible glare in road lighting, Proc. Int. Comm. on Illumination, Stockholm 1951.

surface against which the object is seen, and L_0 the luminance of the object) could be independently varied. We determined the minimum value that L_r/L_0 must have in order to make the object visible. For objects of the size mentioned the distance from the observer was found to be of little influence. Curve *1* in *fig. 4* gives the average for numerous observations and observers. It can be seen that, for a road-surface luminance of 2 cd/m^2, the ratio L_r/L_0 must be at least 1.7 to make our objects visible. The value of 1.7 is fairly high: a good standard for safety is that an object measuring 20 × 20 cm should be clearly visible at a distance of 100 m when $L_r/L_0 = 1.5$ [3]). According to our results, in spite of using larger objects this standard is far from being satisfied at a road-surface luminance of 2 cd/m^2, and therefore as far as perceptibility is concerned the requirement of 2 cd/m^2 is in fact a modest one.

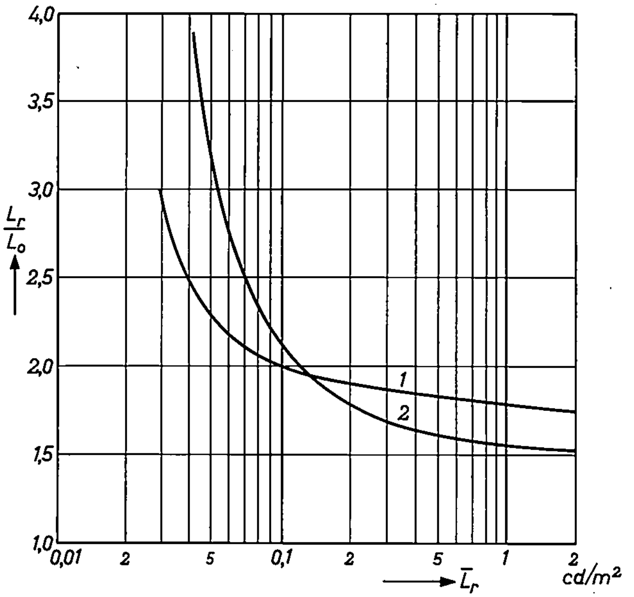


Fig. 4. Perceptibility of objects on the road as a function of average road-surface luminance $\bar{L}_r$. Along the ordinate is plotted the ratio L_r/L_0 of the luminance L_r of the part of the road surface against which the object is seen and the luminance L_0 of the object. Curve *1* (our own measurements) relates to objects measuring 28 × 28 cm (fig. 3) placed at distances between 50 and 200 m from observers stationed at a fixed point along the road. The objects appeared at places and times that were not known beforehand to the observers. Curve *2* (Dunbar [4])) relates to observations from a moving car.

For comparison we have included in fig. 4 a curve showing the results published by Dunbar as long ago as 1938 [4]). In this investigation observations were made from a car in motion of objects one and a half times larger than ours. His results also show however, that an object with a contrast of 1.5 is still not perceptible at a road surface luminance of 2 cd/m^2.

From experiments carried out by us in 1959, also with observations from a moving car [5]), it was concluded that objects of 20 × 20 cm are visible from a distance of 100 m at $L_r/L_0 = 1.5$ — even to observers unprepared for the object suddenly appearing in their visual field — provided the road-surface luminance is at least 2.2 cd/m^2.

Summarizing, it can be said that an average road-surface luminance of about 2 cd/m^2 meets reasonable requirements both as to visual comfort and perceptibility. It was of course not possible to predict that these two aspects would lead to the same recommendation, and in the two points now to be discussed, glare and uniformity of road-surface luminance, we shall see that this is by no means the case.

Glare

The light sources of a road-lighting installation can give rise to a certain amount of glare (we are not concerned here with the glare caused by oncoming traffic). Investigations in recent years have made it clear that the deterioration of perceptibility (disability glare) is not nearly as important in this respect as the deterioration of visual comfort (discomfort glare). Observations have shown that as long as the glare is not troublesome, i.e. is still acceptable from the point of view of visual comfort, there is scarcely any deterioration of perceptibility [6]). In a recommendation concerning admissible glare, visual comfort is therefore the decisive factor.

As a measure of the glare caused by a light source we can take the illumination E which the light source produces on the eye of the subject. Whether glare is experienced as a nuisance depends on other conditions, the more important of which are: the average road-surface luminance $\bar{L}_r$, the angle ϑ which the line from the observer's eye to the light source makes with the horizontal, and the solid angle ω subtended by the source at the observer's eye (see sketch in *fig. 5*). The particular importance of the angle ϑ is due to the fact that, when driving in traffic, a motorist usually keeps his line of sight almost horizontal, whilst he scans the road continuously from one side to the other. The maximum

[3]) The experiments discussed were carried out before this standard was proposed, hence the differing dimensions of the objects.

[4]) E. Dunbar, Necessary values of brightness contrasts in artificially lighted streets, Trans. Illum. Engng Soc. (London) 3, 187-195, 1938.

[5]) See the article cited under [1]), particularly pp. 9 and 10.

[6]) J. B. de Boer, Blendung beim nächtlichen Strassenverkehr, Zentralblatt Verkehrs-Medizin, Verkehrs-Psych. angr. Geb. 3, 185-203, 1957.

value E_b which the illumination E on the eye from a single light source should not exceed if visual comfort is to remain satisfactory is thus a function of $\bar{L}_r$, ϑ and ω. In fig. 5, which gives the average result of very numerous observations by a large number of observers, E_b is plotted as a function of $\bar{L}_r$ and ϑ for three values of ω. It can be seen that if E is greater than E_b there are three possible correctives: one can either increase $\bar{L}_r$, ϑ or ω. The latter can be done by, for example, modifying the lanterns.

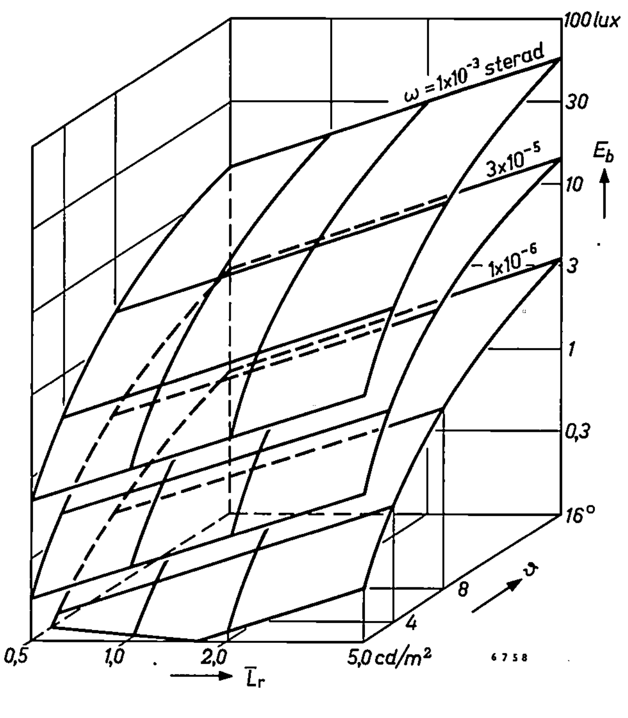
The results of fig. 5 can be expressed in the formula [6]):

$$E_b = 7.5\ \bar{L}_r^{2/3}\ \vartheta^{4/3}\ \omega^{2/5},$$

where E_b is in lux, $\bar{L}_r$ in cd/m², ϑ in degrees and ω in steradians.

In practice, glare is invariably caused by more than one light source at the same time. Observations have shown that in this case the question whether the degree of visual comfort is still satisfactory can be answered as follows. The quotient of the real illumination on the eye and the maximum admissible value calculated from the above formula is determined for each light source separately. If the sum of these quotients over all light sources is smaller than 1, the degree of visual comfort will be satisfactory from the point of view of glare. Given K light sources, this condition may be expressed mathematically as:

$$\sum_{k=1}^{K} E_k/E_{bk} \leq 1.$$



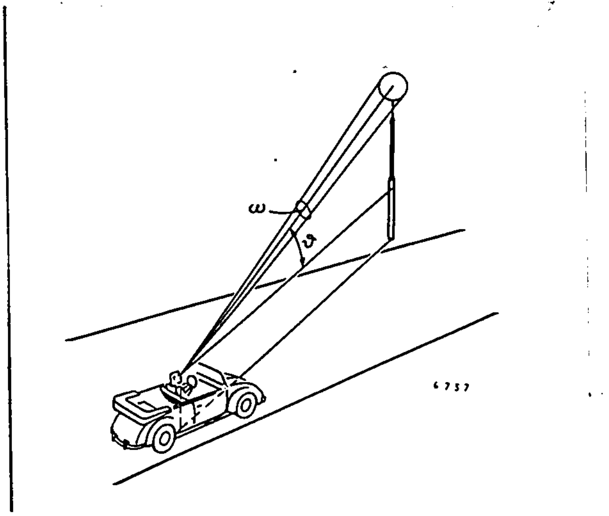


Fig. 5. The illumination E_b produced by a single light source on the eye of an observer, and corresponding to "satisfactory" visual comfort, as a function of the average road-surface luminance $\bar{L}_r$ and the angle ϑ, for three values of the solid angle ω (see sketch).

Here E_k and E_{bk} are respectively the real illumination and the maximum permissible illumination on the eye due to the kth light source. The recommendation, then, is that this condition must be fulfilled wherever the observer may be on the road. It appears that this condition is amply fulfilled in practice if the light emission at angles greater than 80° with the downward vertical is limited to a few tens of candela per thousand lumen, and where at the same time the direction in which the luminous intensity is maximum makes an angle of no more than 70 to 75° with this vertical.

Uniformity of road-surface luminance (patchiness)

Alternate bright and dark patches on the road surface are unavoidable to a certain extent. Like glare, a patchy luminance pattern adversely affects perceptibility, but even before its influence is measurable it has already become unacceptable from the point of view of visual comfort. Thus visual comfort is again the decisive factor. There has not yet been much research on this point. The recommendations on public lighting issued by the Nederlandse Stichting voor Verlichtingskunde (Netherlands Illuminating Engineering Society) [7]) state that the road-surface luminance in the transverse direction should not vary by more than a factor

[7]) Recommendations for public lighting, published in 1959 by the Nederlandse Stichting voor Verlichtingskunde, Arnhem, Netherlands.

of three, and that moreover the minimum value should not be smaller than one third of the average value. This is a provisional recommendation which at least provides some measure of certainty in regard to perceptibility, and hence to traffic safety. As far as visual comfort is concerned, a road showing the unevenness of surface luminance laid down as maximum in this recommendation would not be acceptable.

A particular problem is the considerable difference between the reflecting properties of a road surface in dry and wet condition. When the road surface is wet it is difficult even to meet the moderate requirements of the provisional recommendation.

Light sources employed

The kinds of light source nowadays used for road lighting — tungsten-filament lamps, sodium lamps, high-pressure mercury-vapour lamps and fluorescent lamps — comprise altogether more than fifty different types of lamp. (This does not include sporadically used types, and no distinction is made between tungsten lamps for different mains voltage but of equivalent wattage.)

Table I gives an idea of the relative extent to which the four kinds of light source are employed. Exact figures are rather difficult to come by. The data on which the table is based were provided by officials responsible for public lighting in various regions or large towns. The wide use still made of tungsten lamps is particularly striking. This appears not only from the preponderance in number and wattage, but also from the luminous-flux figures, which offer a better yardstick for comparison.

Table I. Data on the use of tungsten lamps (G), sodium lamps (S), high-pressure mercury-vapour lamps (M) and fluorescent tubular lamps (F) for road lighting in various countries.

		Netherlands	West Germany	France	Great Britain
Number (%)	G	57	24	77	65
	S	5	1	1	17
	M	9	21	15	15.5
	F	29	54	7	2.5
Luminous flux (%)	G	51	16	48	41
	S	17	2.5	3.5	37
	M	20	29	40	20
	F	12	52.5	8.5	2
Wattage (%)	G	78	38	73.5	63.5
	S	5	1.5	1.5	18
	M	10	27.5	21.5	17
	F	7	33	3.5	1.5

Influence of the kind of light used

The various kinds of light source differ considerably in the spectral composition of the light which they emit, as appears for example in the colour of the light and the colour rendering. As regards high-pressure mercury-vapour lamps a distinction must be made between those having a fluorescent bulb (referred to here as HPL lamps), and those having a clear bulb (referred to here as HP lamps). In Europe (apart from Great Britain) the latter are very seldom used for road lighting. In our view rightly so, since the colour rendering of HP lamps is poor. It was for this reason that HPL lamps were developed, the colour rendering being greatly improved by the light which the fluorescent bulb adds to the mercury light. Sodium lamps too have the disadvantage of poor colour rendering, but in their case, as opposed to HP lamps, this is offset by substantial advantages. We shall return to this point later. The high-pressure mercury lamps used in our investigation were exclusively HPL types.

As regards the tubular fluorescent lamps we have a type in mind which is suitable for public lighting and is known in professional parlance as "white". This type was developed primarily to give a high specific luminous flux rather than ideal colour rendering. In our experiments the main emphasis has been on sodium lamps and HPL lamps, which are the most important for traffic-route lighting. It is of interest to consider the extent to which the spectral composition of the light affects the requirements which, both as regards perceptibility and visual comfort, should be imposed on the average road-surface luminance, the admissible glare and the uniformity of the road-surface luminance. Our considerations are based partly on laboratory experiments carried out indoors under conditions differing considerably from those encountered in practice. Such experiments are nevertheless very useful for purposes of comparison; they have the great advantage that they can be done in the daytime, irrespective of weather conditions. It is also easier to obtain the required numbers of observers.

Kind of light and average road-surface luminance

As regards perceptibility in connection with the average road-surface luminance, it has already been frequently observed that sodium light is superior to tungsten (incandescent) light [8] [9]. The light from

[8] W. Arndt, Über das Sehen bei Natriumdampf- und Glühlampenlicht, Das Licht 3, 213-215, 1933.

[9] M. Luckiesh and F. Moss, Seeing in tungsten, mercury and sodium light, Trans. Illum. Engng Soc. (America) 1, 655-674, 1936.

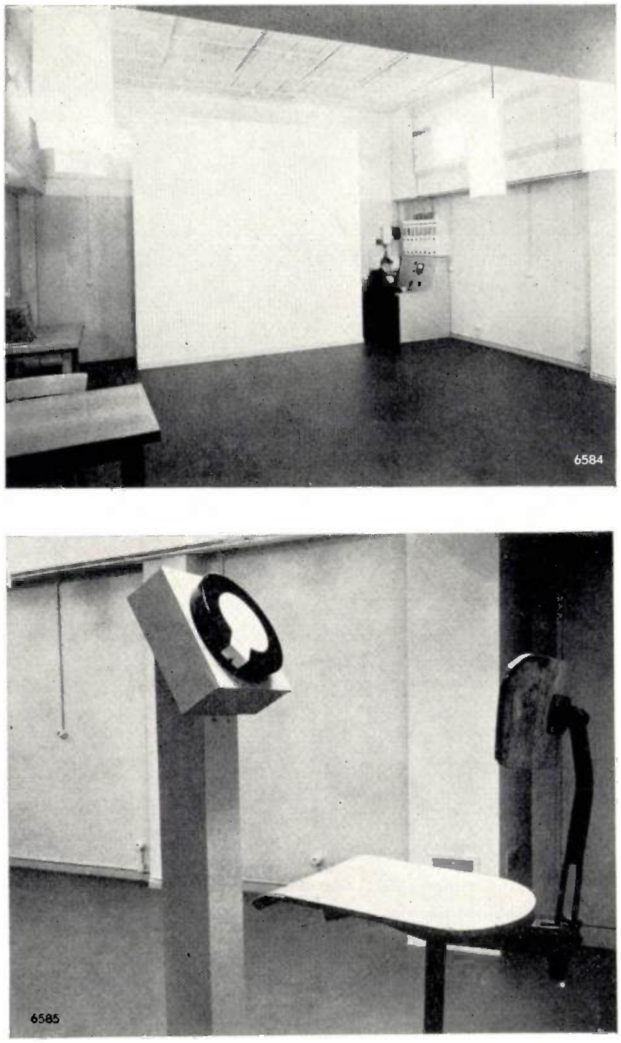


Fig. 6. Two views of the experimental arrangement for comparing the perceptibility of objects under light of different spectral composition. The photo above shows the uniformly illuminated screen, in the middle of which an interchangeable Landolt ring was presented to view for 0.1 sec. Rings of different size and different reflection coefficients were used. The observer had a Landolt ring in front of him (photo below) whose position he had to adjust in accordance with what he had seen or believed he had seen on the screen.

HP lamps was also compared years ago with sodium light. Bouma [10]) and Arndt [8]) found, for example, that visual acuity was somewhat greater in HP light than in sodium light. Weigel [11]) on the other hand found scarcely any difference between these two kinds of light. An investigation of this nature usually consists of experiments in which an object is silhouetted against a uniformly lighted background. In Philips Lighting Laboratory an extensive comparative study has been made of sodium and HPL light. *Fig. 6* gives an impression of the experimental set-up. The object was an interchangeable Landolt ring. Whether or not the ring is correctly perceived (i.e. whether or not the orientation of the gap in the ring is correctly recognized) depends on four quantities:

1) the background luminance L_a,
2) the contrast $C = (L_a - L_o)/L_a$, where L_o is the luminance of the object,
3) the size of the ring,
4) the time t available for perception (exposure time), i.e. the time during which the ring is exposed to view.

For the two kinds of light mentioned, threshold measurements were made with a group of 20 observers for a single exposure time of 0.1 sec. The results are presented in *fig. 7* in a three-dimensional graph, the coordinates of which are L_a, C (both logarithmic) and the distance l (linear) at which a Landolt ring of 160 mm diameter subtends the same angle as the actual ring. (The latter was situated 6 m away from the observer.) For a given kind of light the region where the (average) observer sees the object correctly when it is exposed to view for t seconds is separated from the region where this is not the case by a plane, called the threshold plane [12]).

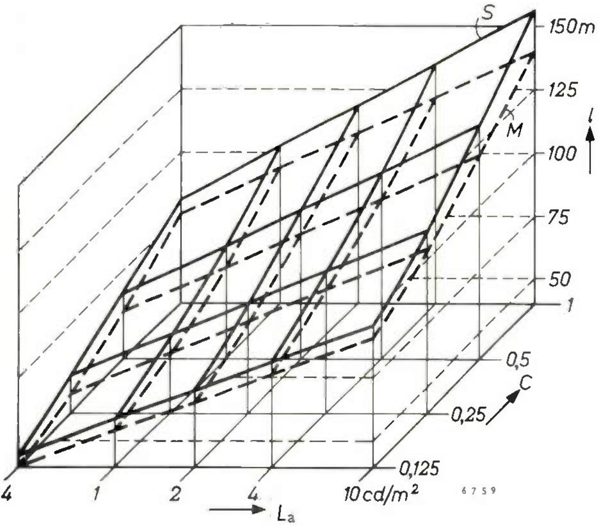


Fig. 7. Distance l at which a Landolt ring of 160 mm diameter is just correctly perceptible, as a function of background luminance L_a and contrast $C = (L_a - L_o)/L_a$ (where L_o is the luminance of the ring) at an exposure time of 0.1 sec, for sodium light (S) and HPL light (M). The distance l was calculated from observations made with the set-up shown in fig. 6, where the distance between observers and object was 6 m.

[10]) P. J. Bouma, Gezichtsscherptemetingen bij diverse lichtsoorten, Ingenieur **49**, A 243-A 246, 1934.

[11]) R. G. Weigel, Untersuchungen über Sehfähigkeit im Natrium- und Quecksilberlicht, insbesondere bei der Strassenbeleuchtung, Das Licht **5**, 211-216, 1935.

[12]) For a different exposure time, different planes are found. The experimental arrangement described was developed by Balder and Fortuin for investigating the influence of the time of observation on the visibility of stationary objects under tungsten light; see J. J. Balder and G. J. Fortuin, Proc. Int. Comm. on Illumination, Zürich 1955, I.

It is seen from the graph that sodium light is more favourable in this respect than HPL light. Analysis of the observations on which the graph is based show that, to find the same threshold value of l at a given contrast (this threshold value is a measure of visual acuity), the background luminance L_a for HPL light must be on an average 54% higher (standard deviation 12%) than for sodium light.

The reciprocal of the threshold value of C is called the contrast sensitivity. The more l decreases, i.e. the larger the object becomes (all details proportionately larger), the closer the threshold planes for the various kinds of light approach each other. For large objects, then, there is no difference between the various kinds of light, either in visual acuity or in contrast sensitivity. Numerous contrast-sensitivity measurements have been reported in the literature which indicate that the kind of light has scarcely any influence on contrast sensitivity. Measurements of this nature always relate to large objects.

Our finding that sodium light gives better perceptibility than HPL light is in good agreement with recent results obtained by Jainski [13]), likewise using Landolt rings but only one contrast value ($C = 0.96$). *Fig. 8* — a cross-section of fig. 7 perpendicular to the C axis at the point $C = 0.96$ — shows in addition to our own results the lines derived from Jainski's observations for sodium, HPL, fluorescent and tungsten light. It was not to be expected that Jainski's lines would coincide with ours, for they relate to another group of observers. His results too, however, reveal the difference between sodium and HPL light, indicating that the background luminance for HPL, fluorescent and tungsten light must be respectively 35, 65 an 100% higher than for sodium light in order to obtain the same visual acuity.

We have also compared sodium and HPL light in regard to perceptibility in actual road conditions. Landolt rings of 160 mm diameter were set up along the road, and the distance was determined at which the rings were still properly perceptible. The difficulty that this distance depends on the position on the road, owing to the road-surface luminance not being uniform, was overcome by distributing large numbers of rings systematically over the road (*fig. 9*). The experiments were done both on ordinary roads and in our outdoor laboratory for road lighting [14]) at Turnhout (where the photograph of fig. 9 was taken). The outdoor laboratory consists of a road provided with mobile trolley-mounted lighting masts on which the lanterns can be adjusted in height and are readily interchangeable. As was to be expected, the results of the tests on ordinary roads showed a greater spread than those in the outdoor laboratory, where the conditions can be much better controlled. In both cases, however, the results correspond to those of the indoor experiments; the differences were only greater [15]).

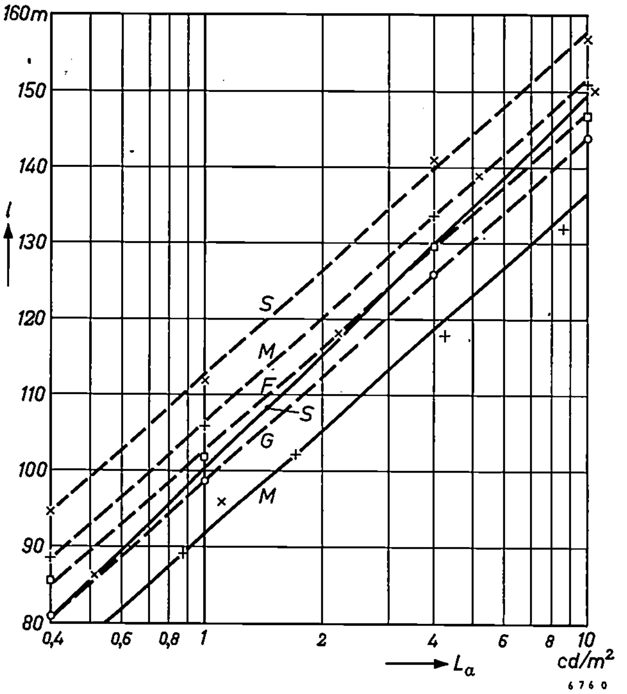


Fig. 8. Cross-section perpendicular to the C axis of fig. 7, at the point $C = 0.96$. The planes S and M from fig. 7 appear in this cross-section as the full lines S and M. For comparison, the broken lines represent the results of Jainski [13]), who compared tungsten light (G), sodium light (S), HPL light (M) and fluorescent light (F) at the same contrast value ($C = 0.96$).

In the following figures the letters G, S, M and F have the same meaning.

As stated on page 263, whether or not an object can be perceived also depends on the time available for perception, the exposure time. There is thus a threshold exposure time corresponding to any given combination of object size, contrast and background luminance. The reciprocal of this time is called the speed of perception. In *fig. 10* the speed of perception is represented as a function of object luminance for a given contrast and size of object, both for sodium and tungsten light [16]). *Fig. 11* gives the speed of

[13]) P. Jainski, Die Sehschärfe des menschlichen Auges bei verschiedenen Lichtarten, Lichttechnik **12**, 402-405, 1960.

[14]) J. Hamming and J. F. T. van Heemskerck Veeckens, An open-air laboratory for road lighting, Philips tech. Rev. **19**, 202-205, 1957/58.

[15]) J. B. de Boer, The application of sodium lamps to public lighting, Illum. Engng **56**, 293-301, 1961 (No. 4).

[16]) P. J. Bouma, Perception on the road when visibility is low, Philips tech. Rev. **9**, 149-157, 1947/48.

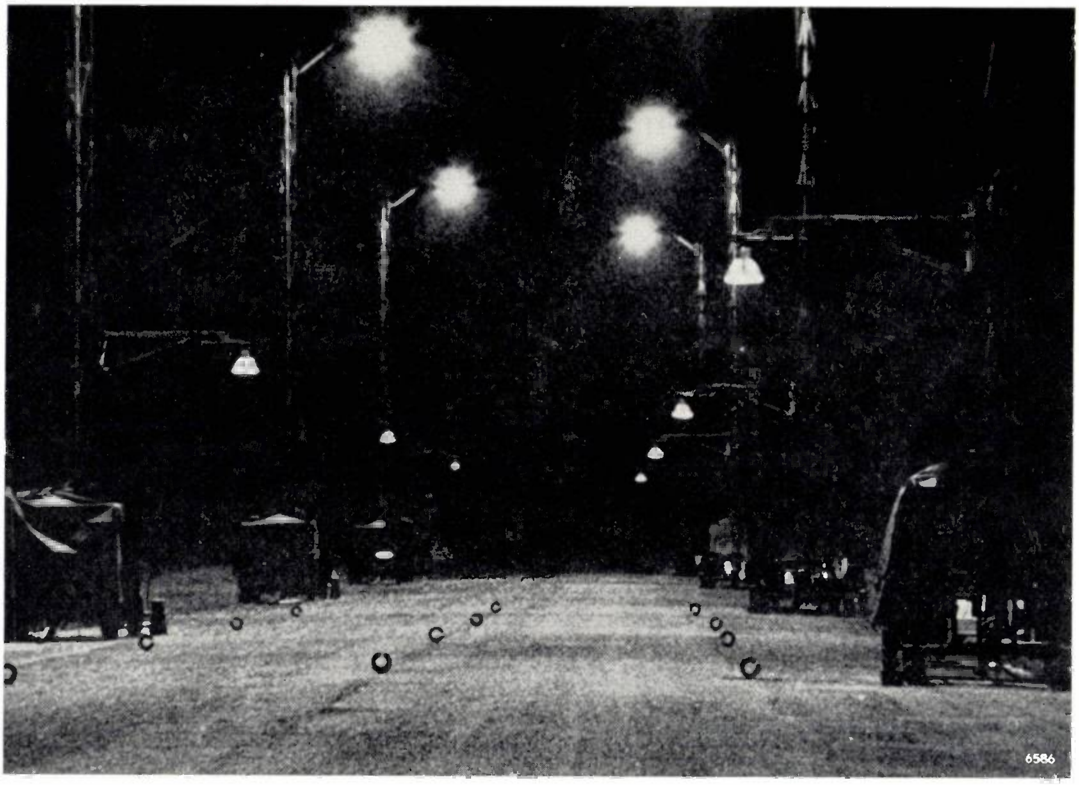


Fig. 9. View of Philips outdoor laboratory for road lighting at Turnhout [14]), showing Landolt rings systematically distributed for investigating the influence of the kind of light on perceptibility.

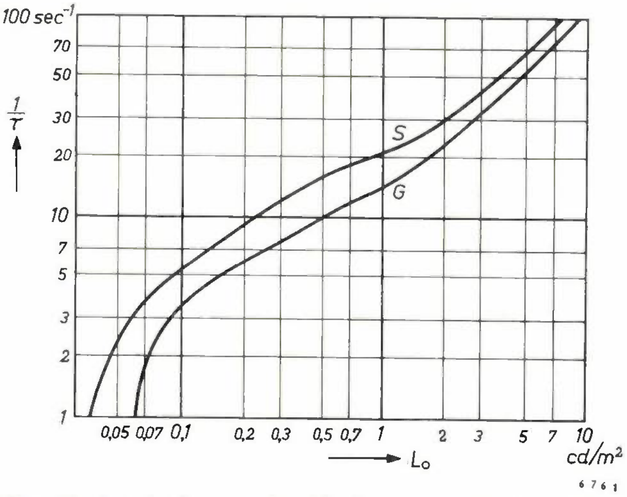


Fig. 10. Speed of perception $1/\tau$ (τ = exposure time) for a stationary object, at a fixed value of contrast, as a function of object luminance L_o for tungsten light and sodium light, after measurements by Bouma [16]).

perception for moving objects according to Weigel [11]) for sodium, tungsten and HP light. Similar measurements have also been done by Arndt [8]) and by Luckiesh and Moss [9]). They found that the values of background luminance needed under the different kinds of light in order to arrive at the same speed of perception stand to one another roughly in the ratio of the values needed to achieve the same visual acuity.

To obtain some idea of the influence which the kind of light has on the desired road-surface luminance from the point of view of visual comfort, the results collected in fig. 2 were analysed according to the kinds of light concerned. The considerable spread makes a significant comparison difficult. If

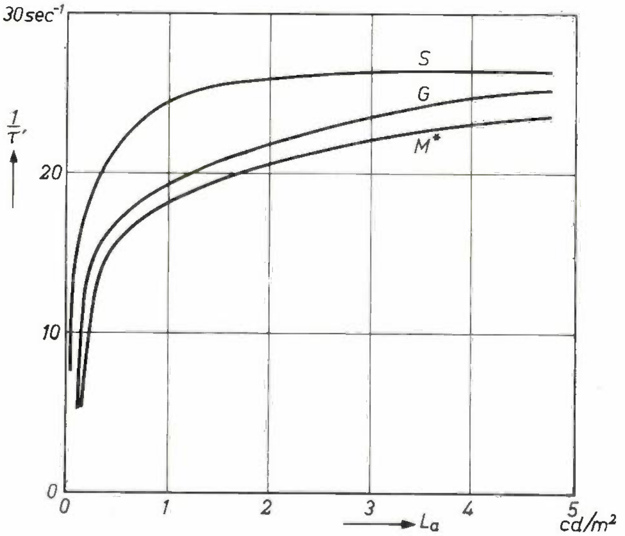


Fig. 11. Speed of perception $1/\tau'$ (τ' = exposure time) for a moving object, at a fixed value of contrast, as a function of background luminance L_a for tungsten light, sodium light and HP light (M*), after measurements by Weigel [11]).

we try to represent by straight lines the relation between the average road-surface luminance in the different kinds of light and the subjective evaluation of that luminance, we find lines of different slope. A statistical analysis has shown, however, that these differences of slope are not significant. Our procedure was therefore as follows. In *fig. 12* — which contains the same points as fig. 2 — we first drew the straight line which, disregarding the kind of light, gives the best approximation to the above-mentioned relation. Parallel with this we then drew lines through the centres of gravity of the clusters of points for the different kinds of light. Interpreting the results in this way it is seen that, in order to obtain the same average subjective evaluation of the road-surface luminance as for sodium light, the average luminance must be increased by 34% for HPL light, by 50% for tungsten light and by 125% for fluorescent light. In other words, given the same road-surface luminance, sodium light creates an impression of greater brightness. The standard deviations of the above percentages are 21, 28 and 45%, respectively.

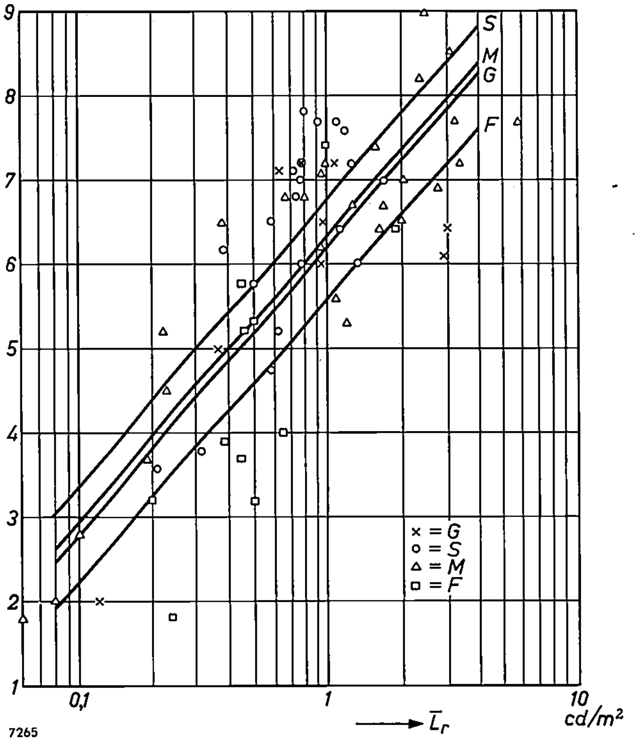


Fig. 12. Average assessments of the lighting level in 70 streets by 16 experts (cf. fig. 2), analysed according to the four kinds of light.

It should be added that the high value of 125% for fluorescent light is probably attributable in part to the lanterns used, which did not give a cut-off light distribution and therefore caused fairly strong glare. The lanterns for the other three kinds of light were all of the cut-off type. The result of this comparison of sodium and HPL light with respect to road-surface luminance shows the same trend as found by Stevens and Ferguson [17]) who compared sodium with HP light in a range of luminances usual for artificially lighted road surfaces.

Kind of light and glare

On the subject of glare we have already seen that the extent to which glare is admissible is determined by the visual discomfort it causes and not by perceptibility. For the purpose of comparing the discomfort glare caused by light of various kinds (tungsten light, sodium light, HPL light and fluorescent light) we used, among other things, a simulated street-lighting installation in our laboratory indoors [18]). This made it possible to vary separately the kind of light used, the brightness of the sources and the average road-surface luminance. All other factors that might affect the degree of discomfort, such as the size and situation of the light sources, and the luminance distribution of source and road surface, were kept constant.

For the four kinds of light concerned, *fig. 13* shows the luminance L_l of the light sources which is still just admissible, according to the average evaluation of six observers, as a function of the average road-surface luminance $\bar{L}_r$. It appears that, in the interval of $\bar{L}_r$ investigated (0.1 to 10 cd/m²),

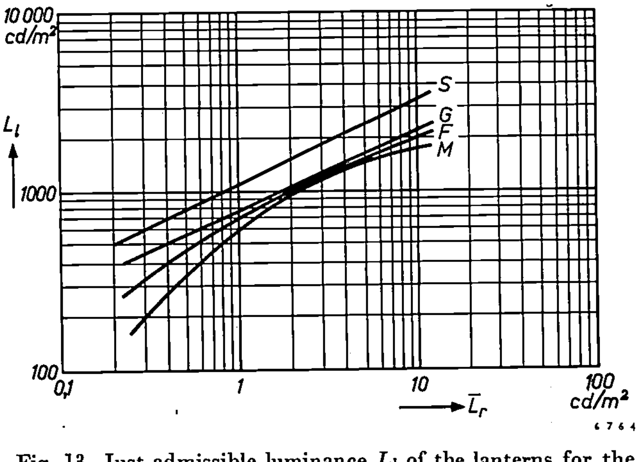


Fig. 13. Just-admissible luminance L_l of the lanterns for the four kinds of light source in a street-lighting installation, as a function of the average road-surface luminance $\bar{L}_r$, according to the average evaluation of a group of six observers.

[17]) H. M. Ferguson and W. R. Stevens, Relative brightness of coloured light sources, Trans. Illum. Engng Soc. (London) 21, 227-247, 1956.

[18]) For a description of the experimental arrangement, see J. B. de Boer and J. F. T. van Heemskerck Veeckens, Observations on discomfort glare in streetlighting. Influence of the colour of the light, Proc. Int. Comm. on Illumination, Zürich 1955.

visual comfort with sodium lighting is still found to be satisfactory at considerably higher light-source luminances than with the other three kinds of light.

Sodium, HPL and fluorescent light were again compared at one value of $\bar{L}_r$ (1 cd/m^2) by a larger group of observers (50 persons). The result found was that, for the same discomfort glare, the light-source luminance with sodium lighting may permissibly be 30% greater than with fluorescent lighting and 45% greater than with HPL lighting, in agreement with fig. 13. Ferguson, Reeves and Stevens, who compared sodium with HP lighting [17][19], also concluded that the light-source luminance may be greater with sodium light than with mercury light.

We tested these conclusions against the results of the evaluation by the 16 experienced observers of discomfort glare in the 70 streets mentioned earlier (page 259). When these evaluations are analysed according to the kind of light used, it is found here too that sodium light sources may permissibly have a considerably higher luminance than other kinds of light sources.

The influence which the colour of the light has on the extent to which glare impairs perceptibility is of less importance, since the admissible glare is governed by the visual discomfort caused. It may be added for completeness, however, that the colour of the light is found to have scarcely any influence in this connection.

A further point of importance is the effect of the colour of the light on the time taken to recover from glare. We have found that, given otherwise identical conditions, the recovery time in the case of sodium light is only three-quarters of that in the case of HPL light and two-thirds of that for tungsten light [20].

Kind of light and patchiness of luminance pattern

As regards patchy or irregular luminance patterns on the road surface the decisive factor, as with glare, is visual comfort. Investigations in this field are still in the initial stages: the following comparative experiments with sodium and HPL light, done in the outdoor laboratory, represent a first attempt to place the problem of non-uniform road luminance on a quantitative basis.

The average road-surface luminance $\bar{L}_r$ and the degree of patchiness were systematically varied. A group of 25 observers was asked to assess visual comfort in regard to the patchiness of the luminance pattern with one of the ratings: bad, inadequate, fair, good or excellent. For the purpose of averaging, the numbers 1, 3, 5, 7 and 9 were assigned to the respective ratings. The patchiness of the luminance pattern — the measure of which was taken to be the ratio R between the maximum and minimum road luminance — was varied by altering the height of the lanterns. Lanterns giving the same light distribution were used for both kinds of light so that the distribution of luminance on the road surface is in both cases identical for the same value of R. *Fig. 14* shows the average evaluation of the observers as a function of $\bar{L}_r$ for sodium and HPL light and for various values of R. The following two conclusions seem to be warranted:

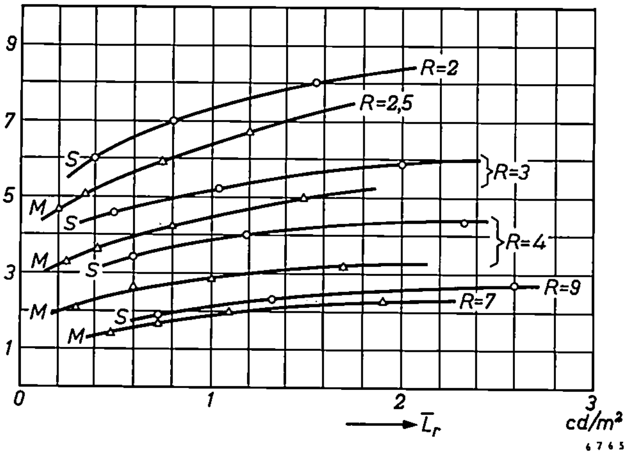


Fig. 14. Evaluation of the uniformity of road-surface luminance in outdoor laboratory experiments as a function of average road-surface luminance $\bar{L}_r$, for various values of the ratio R between the maximum and minimum road-surface luminance. The numbers 1, 3, 5, 7 and 9 correspond to the ratings bad, inadequate, fair, good and excellent.

1) Stronger local variations in road-surface luminance are acceptable the higher is the average value of that luminance.
2) Under sodium light, other conditions being identical, stronger local variations in road-surface luminance are permissible than under HPL light.

These conclusions have been confirmed in broad lines by subsequent experiments. Further investigation is needed before quantitative pronouncements can be made.

Summarizing, it can be said that investigations into the influence of the kind of light used have demonstrated that sodium light offers the following marked advantages over the other kinds of light considered.

[19] H. M. Ferguson, J. Reeves and W. R. Stevens, A note on the relative discomfort glare from mercury, sodium and tungsten light sources, G. E. C. Journal **20**, 184-187, 1953.

[20] J. B. de Boer, La couleur de la lumière dans l'éclairage pour la circulation routière, Lux, 1959, pp. 20-25 (No. 1), and pp. 46-50 (No. 2).

1) Greater visual acuity. With other kinds of light the average road-surface luminance must be about 1.5 times higher to give the same visual acuity. (This does not apply to a comparison with HP light; see page 263.)
2) For the same road-surface luminance, sodium light gives an impression of greater brightness. In that respect sodium light is more than 30% superior to HPL light, which takes second place.
3) Greater speed of perception. The data at present available suggest that, in this respect, the various kinds of light stand in roughly the same ratio to one another as in regard to visual acuity.
4) Less discomfort glare. As a result, the luminances of the light sources themselves can be about 1.4 times higher than for other kinds of light under otherwise identical conditions.
5) Patchiness of the road-surface luminance pattern is less disturbing under sodium lighting.

It follows from this that, for a lighting installation of a given quality, the use of sodium lamps requires only about 75% of the lumens needed from other light sources. This, combined with the very high luminous efficiency of sodium lamps, makes sodium lighting very attractive from the economic point of view.

Practical value of good visual acuity

It is sometimes doubted whether better visual acuity under a particular kind of light is an advantage for the purposes of road lighting. Visual acuity, it is argued, is of minor importance because perception in practice depends on seeing contrasts in the luminance of relatively large objects. It is therefore reasoned that perceptibility on the road is governed by the contrast sensitivity in respect of large objects, which is not, as we have seen (page 264), significantly affected by the kind of light used.

This opinion is understandable if related to the conditions that existed some twenty years ago, when levels of a few tenths of a cd/m^2 were common even on important traffic routes. At such levels it is indeed true that perception depends on seeing large objects as dark silhouettes against the road surface. Details of the objects are not then perceptible. In the busy traffic of today this is definitely an unsatisfactory standard of perceptibility. It is precisely the perception of details that leads to immediate recognition and to the decision as to whether the object calls for further attention or not. It is important, for example, that a motorist should be able, at night just as in the daytime, to see whether a pedestrian standing on the kerb has noticed him. If so, the motorist need pay no further special attention to that pedestrian. Every gain in visual acuity thus contributes to greater safety on the roads.

Importance of dimensions, shape and luminance of light sources

A high and uniform luminance on the road surface imposes certain demands on the distribution of light from a street lantern. The light distribution required in a vertical plane parallel to the axis of the road is usually of the type shown in *fig. 15*.

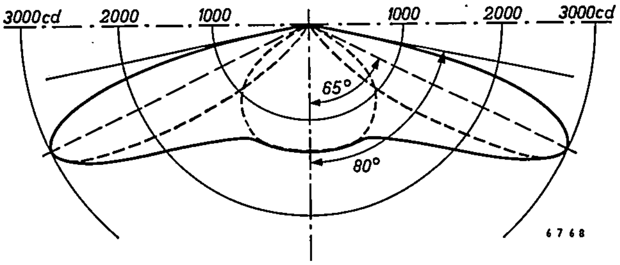


Fig. 15. Example of the distribution of light from a street lantern as required in the vertical plane parallel to the axis of the road. A distribution of this kind can be obtained with the aid of two mirrors, each of which produces one of the narrow beams represented by dotted lines. Together with the remainder of the light, including the direct light from the lamp — the third dotted curve — they provide the distribution required. The luminous intensities given in candelas are normal for a lantern fitted with lamps giving a total light output of about 10 000 lm.

The wide spread in this plane, i.e. in the longitudinal direction of the road, is needed in order to bridge the distance between two successive lanterns. A certain lateral spread is also needed in order to span the width of the road. Excessive lateral spread, however, means a loss of useful light on the road. The extent to which the requirements can be met by reasonable optical means depends on the dimensions, shape and luminance of the light sources. In these respects the four kinds of light source treated in this article show considerable differences. *Table II* gives a survey of these quantities for various types of lamp in each category. If we compare lamps in this table that have roughly the same luminous flux, we notice that tungsten lamps are the most concentrated light sources (highest luminance and hence smallest dimensions of radiating portion), whilst fluorescent lamps with their low luminance and great length represent the other extreme. HPL and sodium lamps have virtually the same luminance, but differ considerably in shape.

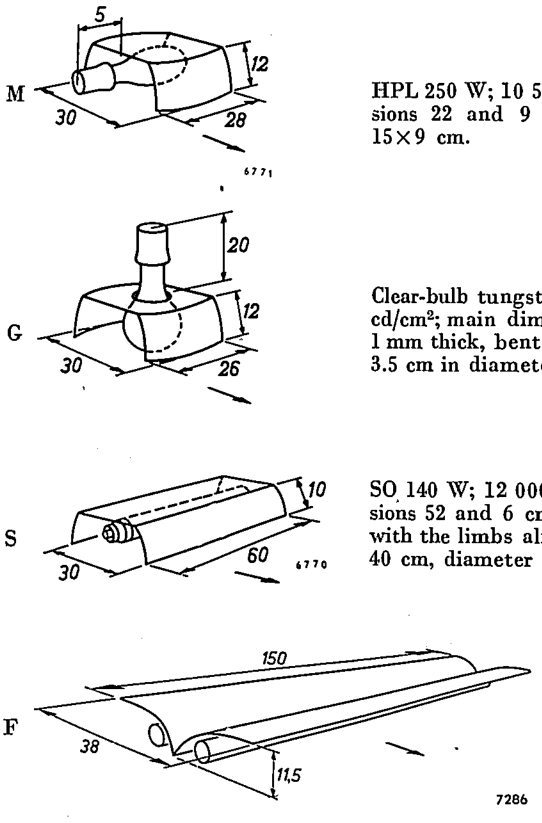
To give an example illustrating the significance of dimensions, shape and luminance, we assume that we have to design for each of the four kinds of light source a system of reflectors giving a longitudinal distribution as in fig. 15. The values of luminous

Table II. Luminous flux, luminance and dimensions of the radiating portion of various types of lamps in the four categories dealt with in this article. The depreciation (average luminous flux over whole life divided by the luminous flux after 100 hours) is also shown. The dimensions given for the radiating portion have the following meanings. For the tungsten lamps: length, breadth and height of an imaginary rectangular parallelepiped enclosing the filament; for the sodium lamps: length and diameter of the almost contiguous limbs of the U-shaped discharge tube; for the HPL lamps: length and diameter of the ovoid fluorescent part of the bulb; for the "TL" lamps: length and diameter of the fluorescent part of the tube. (1 cd/cm^2 = 2920 footlambert.)

Kind of light source	Type designation	Rated power	Luminous flux at half life	Luminance at half life	Dimensions of radiating portion	Depreciation
		watts	lm	cd/cm^2	cm	
Tungsten lamp (220 V)	GLS clear	100	1330	600	2.5 1 0.07	0.96
		500	9000	800	3.5 2.5 0.09	0.95
		1500	27600	1000	4 2 2.5	0.92
Sodium lamp	SOI	45	3100	10	14 1.2	0.94
		60	4700	10	20 1.2	0.94
		85	7500	10	30 1.3	0.94
		140	12000	12	40 1.6	0.94
		200	20200	14	66 1.6	0.94
High-pressure mercury-vapour lamp with fluorescent bulb	HPL	80	2700	10	11 7	0.90
		125	4850	10	12 7.5	0.90
		250	10500	15	15 9	0.90
		400	18000	15	19 12	0.90
		700	33000	15	23 14	0.87
		1000	45200	15	26 16.5	0.87
Tubular fluorescent lamp	"TL" 33	20	1080	0.60	57 3.8	0.87
		40	2430	0.65	117 3.8	0.87
		65	4400	0.8	148 3.8	0.80
		125	5500	1.1	148 3.5	0.80

intensity specified in fig. 15 are normal for a lantern containing a light source of about 10 000 lumens. (The value of 10 000 lumens was chosen because it can he achieved with all four kinds of light source.) Since the largest fluorescent lamp at present used for road lighting in Europe delivers only about 5500 lumens, two lamps have to be used in this case in each lantern. *Fig. 16* shows sketches of four designs that meet the requirements. For simplicity, the reflectors are considered to be symmetrical with respect to the vertical plane parallel to the axis of the road, i.e. the plane to which the specified light distribution applies. As we shall see below, the shape and size of the light source do play a considerable part in determining the design of the reflector.

The desired longitudinal distribution of the light can be obtained by means of two mirrors, each of which



HPL 250 W; 10 500 lm; 15 cd/cm^2; main dimensions 22 and 9 cm; ovoid radiating portion 15×9 cm.

Clear-bulb tungsten lamp 500 W; 9000 lm; 800 cd/cm^2; main dimensions 28 and 13 cm; filament 1 mm thick, bent to a not entirely closed circle 3.5 cm in diameter.

SO 140 W; 12 000 lm; 12 cd/cm^2; main dimensions 52 and 6 cm; radiating portion is a U-tube with the limbs almost touching; length of limbs 40 cm, diameter 1.6 cm.

"TL" 125 W; 2× 5500 lm; 1.1 cd/cm^2; main dimensions 151 and 3.5 cm; radiating portion cylindrical, length 148 cm, diameter 3.5 cm.

Fig. 16. Design sketches, for each of the four kinds of light source, of mirrors — and their position relative to the lamps — for producing in the vertical plane parallel to the axis of the road a light distribution approaching that shown in fig. 15. The flat surfaces above the lamps give diffuse reflection. The principal data of the lamps are given beside the sketches; the values of luminous flux and luminance mentioned hold at the half life of the lamp. The arrows indicate the direction of the road.

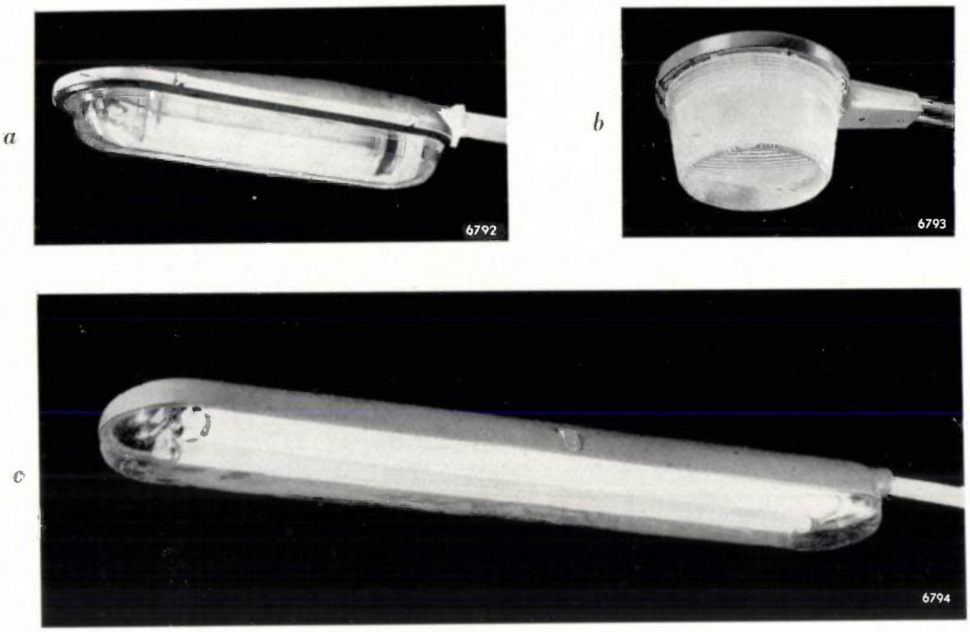


Fig. 17. Three examples of modern enclosed street lanterns; *a*) for a sodium lamp of 200 W (20 200 lm); *b*) for an HPL lamp of 400 W (18 000 lm); *c*) for two fluorescent "TL" lamps of 125 W (2 × 5500 lm). The lanterns are reproduced on the same scale; the overall length of the lantern for the "TL" lamps is 1.8 m.

gives a narrow beam at a wide angle with the vertical. With the direct light and the light reflected from the top of the lantern, the two beams produce roughly the distribution required (see the dotted curves in fig. 15). In the transverse direction of the road less pronounced beaming is required. It is a general rule that, in order to produce a specific beam in a given plane, a larger reflector is needed the larger are the dimensions of the light source in that plane. Since the strongest beaming is required in the vertical plane parallel to the axis of the road, the lamps are mounted so that their smallest dimensions fall in that plane. The smallest dimension of the HPL lamp is still a good 9 cm, and therefore this type of lamp constitutes the most difficult problem in regard to beaming the light in the vertical plane parallel to the axis of the road. It was found, however, that the requirement can be fulfilled by using a simple mirror (part of a paraboloid of revolution) of the dimensions given in fig. 16. Since the largest dimension is decisive in this arrangement as regards the transverse beaming, the lateral spread in this case is automatically greater; its value is found, however, to be acceptable.

The mirror for the tungsten lamp is designed so as to produce the same light distribution as the HPL lamp, both along the road and across it. The light source (filament) being so much smaller, the light distribution can be well controlled. The small dimensions of the light source are not, however, purely an advantage. In order to obtain the required beaming, longitudinally and laterally, it is necessary to give the mirror (fig. 16) a different radius of curvature in two mutually perpendicular directions. Moreover, high demands are made on the reflecting surface, since with a concentrated light source any irregularity in the mirror shows up as a patch of different brightness on the road. Where this is the case, other optical aids have to be enlisted, for example a simple mirror which gives good longitudinal beaming, combined with a ribbed cover of transparent material, which provides the necessary lateral spread and blurs any irregularities. A concentrated light source thus calls for a lantern which, although usually small, is fairly complicated in construction.

Fluorescent lamps create no difficulties as far as longitudinal beaming is concerned, but their length precludes any possibility of lateral beaming.

With sodium lamps the light distribution along the road is no problem either. Some lateral beaming is possible owing to the fact that the distance between the lamp and the mirror can be given a value comparable in some measure to the length of the lamp. Sodium and HPL lamps are roughly equivalent in regard to the control of the light distribution: with sodium lamps the lateral distribution is less easy to control, whereas with HPL lamps it is rather more difficult to achieve a completely satisfactory distribution in the longitudinal direction.

It is by no means the rule that street-lighting lanterns are equipped with mirrors. *Fig. 17* shows three examples where this is not the case. It is noticeable that the fluorescent-lamp lantern, although by far the biggest, gives the least light. Moreover it contains no provisions for beaming the light, a function which is fulfilled in the other lanterns by the prismatic ribbing in the transparent covers. To obtain effective beaming from the fluorescent-lamp lantern it would have to be made very much wider and larger. This indicates that fluorescent lamps are really not so suitable for lighting traffic routes. For HPL or sodium lamps, lanterns can be designed that are more easily manageable and more acceptable from the aesthetic point of view.

In this connection some remarks may be made on HP lamps, i.e. high-pressure mercury-vapour lamps of the clear-bulb type. These lamps have a concentrated radiating portion and are therefore in the same position as tungsten lamps in regard to the control of their light distribution. Good control is thus possible with lanterns of modest dimensions, though of somewhat complicated design. The lanterns for HP lamps can be designed more readily than for HPL and sodium lamps to give very wide-angled distribution, with the maximum luminous intensity almost horizontal (cf. fig. 15), and this has the advantage of allowing wider spacing between the poles. In countries where such light distributions are still preferred HP lamps are therefore in fairly common use. But apart from the growing recognition that such wide-angle lanterns are undesirable from the point of view of discomfort glare, there is another reason why HP lamps are more and more being superseded by sodium and HPL lamps. There is a trend in road engineering to make surfaces rougher, thereby giving more diffuse reflection. The aim is to reduce the difference in appearance between wet and dry surfaces, and also to reduce the hazard of skidding. On more diffusely reflecting road surfaces there is not much point in using these wide-angle lanterns, since in this case the light emitted at angles near the horizontal contributes very little to the luminance of the road. That being so, sodium lamps are preferable on traffic routes where colour rendering is a minor consideration, for HP light and sodium light — both of which give poor colour rendering — are only comparable as far as visual acuity is concerned. In other respects (glare, required surface luminance, speed of perception) HP light is inferior to sodium light. In cases where colour rendering *is* important, for instance on main roads in built-up areas, in shopping centres or city squares, HPL light is greatly to be preferred.

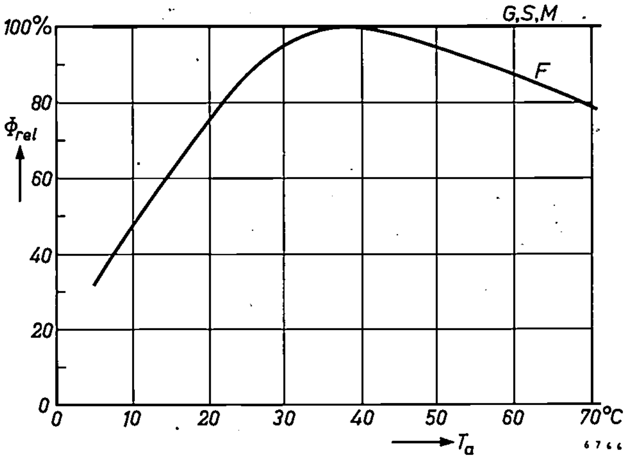


Fig. 18. Relative luminous flux $\Phi_{\rm rel}$ as a function of ambient temperature $T_{\rm a}$, for the four kinds of light source.

Temperature and voltage variations

If the luminous flux of the four kinds of light source be measured indoors as a function of the ambient temperature (the latter being measured at great distance from the lamp), curves roughly resembling that shown in *fig. 18* are found. In the case of fluorescent lamps this dependence is such that it must be taken into account in the design of the lantern. The design must provide for optimum light output in the most common range of night-time temperatures (in our latitudes between 0 and 10 °C). Under other conditions, e.g. during cold winter nights, the light output may then be appreciably lower. This is a disadvantage of fluorescent lamps compared with the three other light sources.

The light output from all four kinds of light source is dependent on the supply voltage (*fig. 19*). In that respect sodium lamps are the most favourable. This is fortunate, since the principal application of these

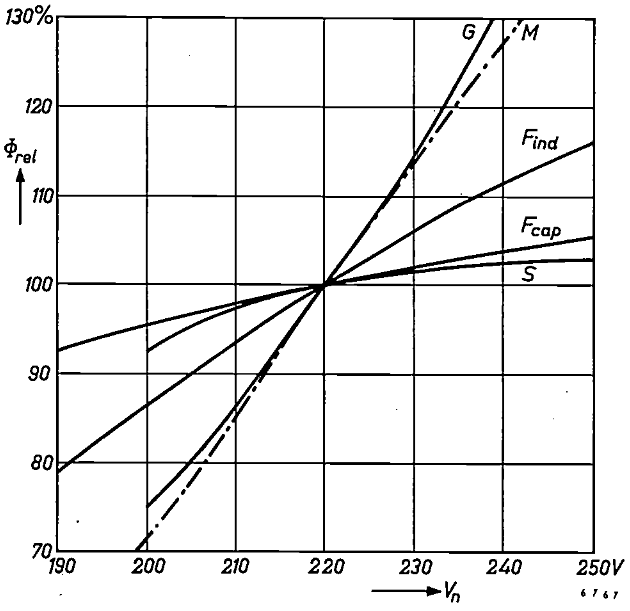


Fig. 19. Relative luminous flux $\Phi_{\rm rel}$ as a function of mains voltage $V_{\rm n}$ for the four kinds of light source. For the fluorescent lamps a distinction must be made between operation with a capacitive ballast ($F_{\rm cap}$) and with an inductive ballast ($F_{\rm ind}$).

lamps is on roads outside the built-up areas, where they are usually connected to branch lines of the electric mains and thus exposed to the worst voltage fluctuations. The relative insensitivity of sodium lamps to voltage fluctuations means that the voltage fluctuations along the supply cables may be larger while the difference in luminous flux between the various lamps fed by a given cable will still be within the prescribed limits. It is therefore possible to cut cable costs by reducing the copper crosssections of the cables.

It is evident from the foregoing that, for lighting traffic routes, sodium lamps deserve preference over other light sources in all cases where colour rendering is a minor consideration. Sodium light offers better visibility and more visual comfort than other kinds of light. Moreover sodium lamps possess properties that make them particularly attractive for large lighting installations: they have a high luminous efficiency, making economic operation possible; their shape and dimensions facilitate the optical design of street lanterns; and they are relatively little affected by temperature and voltage fluctuations. Where their poor colour rendering is an insuperable objection, high-pressure mercury lamps with fluorescent bulbs (HPL lamps) are preferable to fluorescent and tungsten lamps, chiefly because they can supply a much higher luminous flux, and also because — compared with fluorescent lamps — their shape and dimensions are more advantageous and they are less sensitive to temperature variations.

Summary. If a traffic route is to be lighted so that motorized vehicles can proceed safely without using headlights, the lighting installation must create conditions that offer acceptable visual comfort and perceptibility. The extent to which it does so is determined by the average luminance of the road surface, the glare caused by the lanterns of the installation and the uniformity of the surface luminance. Experiments in Philips' lighting laboratories, and elsewhere, indicate that 2 cd/m^2 (0.6 footlambert) is a reasonable average road-surface luminance. A recommendation is made as to what constitutes admissible glare; the effects of irregular surface luminance (bright and dark patches) have not yet been sufficiently investigated for a recommendation to be given.

Four commonly used kinds of light source — tungsten lamps, sodium lamps, high-pressure mercury lamps with fluorescent bulbs (HPL lamps) and tubular fluorescent lamps — are compared in regard to the influence of the kind of light on visual comfort and perceptibility, to the significance of their dimensions, shape and luminance, and to their sensitivity to temperature and voltage fluctuations. Sodium lamps compare favourably on all counts. The conclusion is that, where their poor colour rendering is a minor consideration, sodium lamps deserve preference for road lighting. Where colour discrimination is important, the use of HPL lamps is recommended.

DC/AC CONVERTERS USING SILICON CONTROLLED RECTIFIERS FOR FLUORESCENT LIGHTING

by J. J. WILTING *).

621.314.57:621.314.63:621.327.534.15

In lighting practice today there is a general trend towards higher levels of illumination. The lighting in public transport vehicles, such as railway carriages and buses, is no exception. The power in these vehicles is usually obtained from a dynamo of limited capacity, together with a buffer battery. It is therefore necessary to use very economical light sources, for which purpose tubular fluorescent lamps are particularly suitable.

Since the power source delivers direct current, two systems are possible:

1) Special fluorescent lamps may be used (e.g. Philips "TL" C type) which are fed in series with a ballast resistor directly from the DC source [1]). This system is restricted to voltages of at least 72 V.
2) Standard fluorescent lamps may be used in conjunction with a DC/AC converter.

The second system has the advantage over the first that the loss in the ballasts is very much lower. The converters originally used were almost invariably robust rotary types, like centrifugal converters employing a rotating mercury jet [1]). Nowadays electronic converters are gaining ground as a result of the development of semiconductor devices, such as the transistor [2]) and the silicon controlled rectifier.

The electrical behaviour of the silicon controlled rectifier resembles that of thyratrons and ignitrons, hence the name "solid-state thyratron" by which it is sometimes known [3]). Compared with these two

*) Philips Lighting Division, Eindhoven.

[1]) L. P. M. ten Dam and D. Kolkman, Lighting in trains and other transport vehicles with fluorescent lamps, Philips tech. Rev. 18, 11-18, 1956/57.

[2]) T. Hehenkamp and J. J. Wilting, Transistor D.C. converters for fluorescent-lamp power supplies, Philips tech. Rev. 20, 362-366, 1958/59.

[3]) F. W. Gutzwiller, Phase-controlling kilowatts with silicon semiconductors, Control Engng 6, 113-119, 1959. — This rectifier is available commercially under various names, such as SCR (silicon controlled rectifier), "Thyristor" and "Trinistor". The I.E.C. recently recommended the general name "pylistor" (Interlaken, June 1961; see Bull. Schweizer Elektrot. Ver. 52, 934, 1961). This name is derived from πυλη = gate, which is roughly synonymous with ϑυρα = door, from which the names thyratron and "Thyristor" were formed.

lamps is on roads outside the built-up areas, where they are usually connected to branch lines of the electric mains and thus exposed to the worst voltage fluctuations. The relative insensitivity of sodium lamps to voltage fluctuations means that the voltage fluctuations along the supply cables may be larger while the difference in luminous flux between the various lamps fed by a given cable will still be within the prescribed limits. It is therefore possible to cut cable costs by reducing the copper crosssections of the cables.

It is evident from the foregoing that, for lighting traffic routes, sodium lamps deserve preference over other light sources in all cases where colour rendering is a minor consideration. Sodium light offers better visibility and more visual comfort than other kinds of light. Moreover sodium lamps possess properties that make them particularly attractive for large lighting installations: they have a high luminous efficiency, making economic operation possible; their shape and dimensions facilitate the optical design of street lanterns; and they are relatively little affected by temperature and voltage fluctuations. Where their poor colour rendering is an insuperable objection, high-pressure mercury lamps with fluorescent bulbs (HPL lamps) are preferable to fluorescent and tungsten lamps, chiefly because they can supply a much higher luminous flux, and also because — compared with fluorescent lamps — their shape and dimensions are more advantageous and they are less sensitive to temperature variations.

Summary. If a traffic route is to be lighted so that motorized vehicles can proceed safely without using headlights, the lighting installation must create conditions that offer acceptable visual comfort and perceptibility. The extent to which it does so is determined by the average luminance of the road surface, the glare caused by the lanterns of the installation and the uniformity of the surface luminance. Experiments in Philips' lighting laboratories, and elsewhere, indicate that 2 cd/m^2 (0.6 footlambert) is a reasonable average road-surface luminance. A recommendation is made as to what constitutes admissible glare; the effects of irregular surface luminance (bright and dark patches) have not yet been sufficiently investigated for a recommendation to be given.

Four commonly used kinds of light source — tungsten lamps, sodium lamps, high-pressure mercury lamps with fluorescent bulbs (HPL lamps) and tubular fluorescent lamps — are compared in regard to the influence of the kind of light on visual comfort and perceptibility, to the significance of their dimensions, shape and luminance, and to their sensitivity to temperature and voltage fluctuations. Sodium lamps compare favourably on all counts. The conclusion is that, where their poor colour rendering is a minor consideration, sodium lamps deserve preference for road lighting. Where colour discrimination is important, the use of HPL lamps is recommended.

DC/AC CONVERTERS USING SILICON CONTROLLED RECTIFIERS FOR FLUORESCENT LIGHTING

by J. J. WILTING *).

621.314.57:621.314.63:621.327.534.15

In lighting practice today there is a general trend towards higher levels of illumination. The lighting in public transport vehicles, such as railway carriages and buses, is no exception. The power in these vehicles is usually obtained from a dynamo of limited capacity, together with a buffer battery. It is therefore necessary to use very economical light sources, for which purpose tubular fluorescent lamps are particularly suitable.

Since the power source delivers direct current, two systems are possible:

1) Special fluorescent lamps may be used (e.g. Philips "TL" C type) which are fed in series with a ballast resistor directly from the DC source [1]). This system is restricted to voltages of at least 72 V.
2) Standard fluorescent lamps may be used in conjunction with a DC/AC converter.

The second system has the advantage over the first that the loss in the ballasts is very much lower. The converters originally used were almost invariably robust rotary types, like centrifugal converters employing a rotating mercury jet [1]). Nowadays electronic converters are gaining ground as a result of the development of semiconductor devices, such as the transistor [2]) and the silicon controlled rectifier.

The electrical behaviour of the silicon controlled rectifier resembles that of thyratrons and ignitrons, hence the name "solid-state thyratron" by which it is sometimes known [3]). Compared with these two

*) Philips Lighting Division, Eindhoven.

[1]) L. P. M. ten Dam and D. Kolkman, Lighting in trains and other transport vehicles with fluorescent lamps, Philips tech. Rev. 18, 11-18, 1956/57.

[2]) T. Hehenkamp and J. J. Wilting, Transistor D.C. converters for fluorescent-lamp power supplies, Philips tech. Rev. 20, 362-366, 1958/59.

[3]) F. W. Gutzwiller, Phase-controlling kilowatts with silicon semiconductors, Control Engng 6, 113-119, 1959. — This rectifier is available commercially under various names, such as SCR (silicon controlled rectifier), "Thyristor" and "Trinistor". The I.E.C. recently recommended the general name "pylistor" (Interlaken, June 1961; see Bull. Schweizer Elektrot. Ver. 52, 934, 1961). This name is derived from πυλη = gate, which is roughly synonymous with ϑυρα = door, from which the names thyratron and "Thyristor" were formed.

electronic switching or commutating devices — to which we may add the triode — the silicon controlled rectifier has the advantage of a much smaller voltage loss (only 1 to 1.5 V) and consequently a higher efficiency. It is particularly compact and rugged and requires no heating of the cathode. Although its switching speed is lower than that of a triode, it is greater than that of a thyratron or ignitron. The maximum permissible voltage is at present lower than for thyratrons and ignitrons but much higher than for transistors. Silicon controlled rectifiers are now commercially available [4]) for operation at a rated peak voltage of 400 V and a mean current of 70 A, and can be used for building converters for tens of kilowatts. It has been reported [5]) that further development is expected to raise these values to 1000 V and 1000 A, giving a power-handling rating of 1 MW per rectifier.

Silicon controlled rectifiers are still very expensive. Even so, as an investment for a lighting installation converters using these devices are already comparable with rotary converters, and are expected to compare even more favourably in the near future. Among their present advantages are their considerably higher efficiency, their smaller weight and volume, and the fact that they need no maintenance.

Advantages of operating fluorescent lamps from a converter

In the conversion of direct current into alternating current the choice of voltage and frequency is fairly wide. Where converters for fluorescent lighting are concerned, the voltage and frequency of the AC mains should preferably not be chosen, for neither have the optimum value for operating fluorescent lamps. A voltage of 220 V, for example, is inadequate for reliably starting long fluorescent lamps (types of 40 W and more) unless special measures have been taken, and at the usual low frequencies of 50 and 60 c/s the lamps are far from operating under the most favourable conditions. We shall now discuss both points briefly.

Most fluorescent lamps work in conjunction with a ballast and a starter, which delivers a high (and usually brief) voltage surge sufficient to initiate the discharge. Moreover, the electrodes of the lamp are preheated, which lowers the ignition voltage required. As a result, the ignition is always attended by some delay. In one type of lamp — Philips "TL" S lamp [6]) — a conducting strip on the inside of the tube has so reduced the ignition voltage as to make the above measures unnecessary. This lamp is of course somewhat dearer in construction than a conventional fluorescent lamp, and the losses in the strip make the efficiency somewhat lower.

Using a converter the voltage can be chosen high enough to ignite the longest lamps reliably and without delay with a simple ballast. Preheating is unnecessary (the electrodes must then be accordingly dimensioned) and no conducting strip is needed. To comply with safety regulations the secondary of the transformer in the converter can be centrally earthed, so that the AC cables carry half the full potential with respect to earth.

As regards the frequency, there are several advantages in raising its value to between 5 and 10 kc/s for fluorescent lamps:

1) The weight and size of the ballasts, and the ballast losses, can be considerably reduced. Light weight and low heat generation are sometimes of decisive importance.
2) The luminous efficiency of fluorescent lamps rises with increasing frequency [2]) [7]). The efficiency of a 40 W "TL" lamp, for example, is about 10% higher between 5 and 10 kc/s than at 50 c/s.
3) The light is steadier, without any stroboscopic effect.
4) The lamps cause less radio interference.

The latter two advantages are due to the fact that at high frequencies the gas discharge is much more regular than at the usual mains frequencies [8]): the ion concentration remains virtually constant, re-ignition effects are eliminated. This may also be assumed to benefit the life of the lamp, but life tests still have to confirm this assumption. (The choice of frequencies higher than 10 kc/s does not add much to the above-mentioned advantages, whereas it considerably increases the losses, including radiation losses, entailing the risk of interfering with communication channels.)

The conspicuous benefits of high frequencies are a strong argument in favour of using converters that deliver alternating current at these frequencies. The transistor converter is capable of doing so [2]), but because of its low power and the inability of transistors to withstand high voltages, its applications are very limited (lighting in buses etc.). The con-

[4]) See e.g. Controlled-rectifier manual of the (American) General Electric Co.

[5]) E. J. Duckett, DC to AC power conversion by semiconductor converters, Westinghouse Engr 20, 170-174, 1960 (No. 6).

[6]) W. Elenbaas and T. Holmes, An instant-starting fluorescent lamp in series with an incandescent lamp, Philips tech. Rev. 12, 129-135, 1950/51.

[7]) J. H. Campbell, New parameters for high-frequency lighting systems, Illum. Engng 55, 247-256, 1960 (No. 5).

[8]) Fluorescent lamps and lighting, edited by W. Elenbaas, Philips Technical Library 1959, p. 104.

verter equipped with silicon controlled rectifiers, on the other hand, which is also capable of producing alternating current at frequencies up to 10 kc/s, can operate on a DC input of the order of 100 V and deliver a power in the region of 1 to 10 kW. This converter is therefore eminently suited for powering fluorescent lighting systems in trains and ships.

The latter type of converter can also be useful for operating fluorescent lamps in office buildings and factories which have been specially wired for the purpose: the converters are then operated from the ordinary electricity mains via rectifiers. The energy saving due to the lower losses in the ballasts and the higher luminous efficiency more than offset the losses in the rectifiers and converters. Once the prices of silicon controlled rectifiers have dropped, the net saving in itself will justify the investment in an installation of this kind, which at present is still rather high. Plans for a test installation at Eindhoven are now being worked out.

Converters using silicon controlled rectifiers

We shall confine ourselves in this article to a concise description of a converter developed in the electrical laboratory of Philips' Lighting Division, primarily for train lighting. First of all we shall deal briefly with the properties of the silicon controlled rectifier [9] [3].

The silicon controlled rectifier

The silicon controlled rectifier consists of four alternate layers of P-type and N-type silicon (*fig. 1*). There are two principal electrodes: the anode a and the cathode b, and one control electrode c. Provided the voltage between a and b does not exceed a specific value, the rectifier in the quiescent state passes no current (apart from a leakage current of no more than a few milliamperes); this holds for both polarities of the voltage between a and b. To make the rectifier conductive — which, without causing damage, is possible only in the forward direction (a positive with respect to b) — a momentary current injection in the control electrode is sufficient. A very high current is permissible in the forward direction (up to several tens of amperes, depending on the type [4]). The voltage drop is only 1 to 1.5 V, i.e. an order of magnitude smaller than in gas discharge tubes such as thyratrons and ignitrons. The non-conductive state returns only when the main current has dropped below a certain lower limit, called the holding current, below which insufficient charge carriers are generated in the P-N junctions. For certain types of silicon controlled rectifiers a peak voltage of 400 V (500 V for short periods) of both polarities is permissible in the non-conductive state, but in the forward direction not before the concentration of residual charge carriers after the passage of current has dropped to a level such that the leakage current is lower than the minimum value of the control current. Nor must the voltage in the forward direction rise so rapidly as to cause the resultant capacitive current in the P-N junctions to exceed the minimum value of the control current. The limiting value of the various quantities is dependent on the temperature inside the rectifier and on the nature and dimensioning of the circuit. Breakdown in the reverse direction causes irreversible damage.

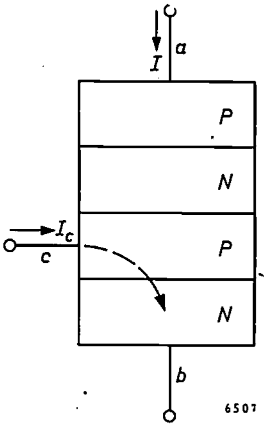


Fig. 1. A silicon controlled rectifier consists of four layers of silicon, alternately P type and N type. a anode, b cathode, c control electrode.

Fig. 2 shows a family of characteristics of a silicon controlled rectifier. It can be seen that the "breakdown voltage in the forward direction" decreases in value as the control current I_c increases.

As a commutating device the silicon controlled rectifier can often advantageously replace triodes,

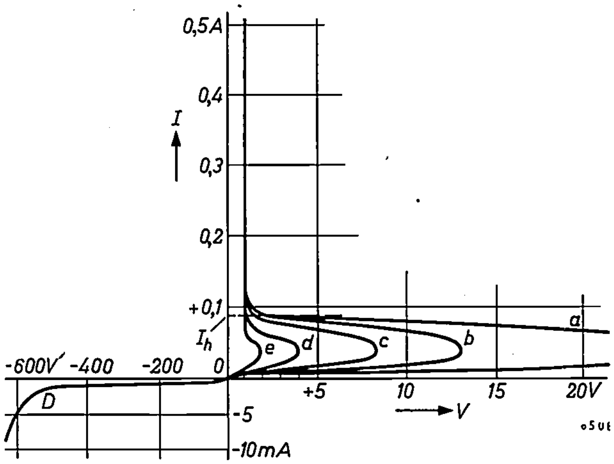


Fig. 2. Current-voltage characteristics of a silicon controlled rectifier. Top right: forward direction; bottom left: reverse direction (I and V are on different scales in the two different quadrants). The parameter is the control current I_c (= 0 in curve a, successively higher in curves b, c, d and e). I_h holding current. At D the rectifier breaks down, i.e. passes current in the reverse direction.

[9] J. L. Moll, M. Tanenbaum, J. M. Goldey and N. Holonyak, P-N-P-N transistor switches, Proc. Inst. Radio Engrs 44, 1174-1182, 1956.

thyratrons, ignitrons and transistors. In common with ignitrons and transistors this rectifier has no cathode that needs a certain time to heat up and that must be kept up to temperature. As mentioned, the voltage drop is lower than in thyratrons and ignitrons, and hence much lower than in triodes.

A drawback of converters using the new rectifiers, as opposed to mechanical converters, is their greater sensitivity to overloads. Overloading can easily cause permanent damage to the rectifying elements. To meet this difficulty the design should allow for a wide safety margin, and/or the converter should be provided with overload protection, preferably electronic.

A converter using silicon controlled rectifiers for train lighting

In order to bring silicon controlled rectifiers in a generator circuit recurrently from the non-conductive into the conductive state, periodic control pulses are needed. These are no problem to generate. A matter of more difficulty is the method of returning the rectifiers periodically to the non-conductive state at the high switching frequency which, for fluorescent lighting, is so advantageous.

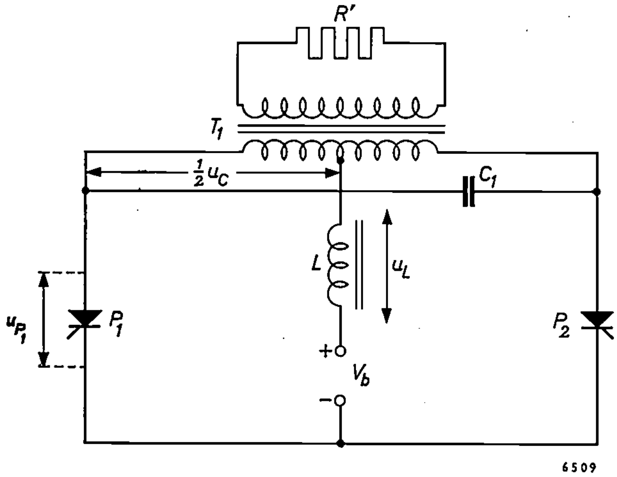


Fig. 3. Simplified circuit diagram (control circuit omitted) of a converter using silicon controlled rectifiers (P_1, P_2) in push-pull arrangement. T_1 transformer. C_1 commutating capacitor. L choke. V_b battery voltage. R' load resistance.

The simplest and most reliable method consists in using a series resonance circuit. The current then tends to have an oscillating character, but after the first half-cycle the rectifying element automatically becomes non-conductive.

Fig. 3 shows a simplified circuit diagram of a converter with silicon controlled rectifiers for train lighting. The two rectifying elements P_1 and P_2 are in a push-pull arrangement. The capacitor C_1 shunted across the primary of the transformer T_1 is called the commutating capacitor. The load on the transformer secondary is drawn here as a resistance R'. The situation when the first control pulse makes P_1 conductive can be represented by the equivalent circuit of *fig. 4*. Here R and C are the transformed values of R' and C_1 in fig. 3 in respect of one half of the transformer primary. The above-mentioned series resonant circuit is formed by C and the choke L. When P_1 becomes conductive, the battery voltage V_b suddenly appears across terminals *1* and *2*. If the circuit is properly dimensioned, the current i_1 which now starts flowing will tend to have the waveform represented in *fig. 5*, i.e. a damped oscillation superposed on an exponentially rising current. However, at the moment $t = t_1$, at which i_1 drops to zero, the rectifying element becomes non-conductive so that i_1 remains for a while at zero. If we now apply the next control pulse to P_2 — some time *after* t_1 to give the charge carriers in P_1 an opportunity to recombine sufficiently — the process will be repeated, except that some quantities change

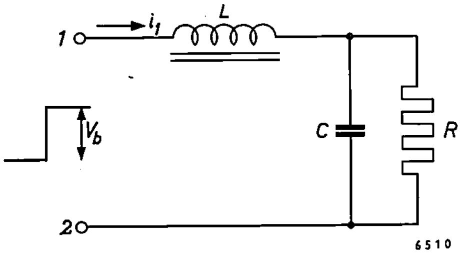


Fig. 4. Equivalent circuit for the moment at which the rectifying element P_1 in fig. 3 is made conductive. The DC voltage V_b then appears suddenly across terminals *1* and *2*. C and R are the transformed values of C_1 and R' in fig. 3 with respect to one half of the transformer primary.

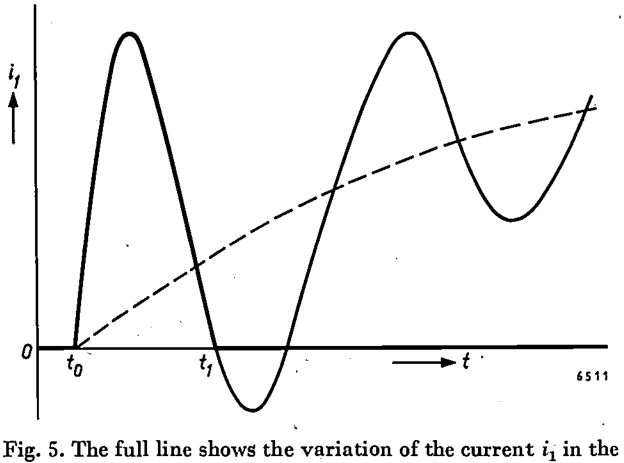


Fig. 5. The full line shows the variation of the current i_1 in the circuit of fig. 4: damped oscillation superposed on a current variation approximately of the form $V_b(1—e^{-at})/R$ (dashed line). Since the rectifying element prevents the current from changing direction, in the circuit of fig. 3 the current i_1 remains zero from point t_1 onwards, so that only the thickly drawn portion of the waveform is obtained.

their sign and that, since C retains a certain charge, the initial conditions are different. The current i_2, which now starts to flow through P_2, will also remain at zero after the first half cycle. A moment later, P_1 is again made conductive, and so on. Periodic repetition of the control pulses produces a certain steady state, resulting in an alternating voltage across the transformer (fig. 3).

Brief analysis

The voltages in fig. 3 satisfy the following equation:

$$u_{P_1} = V_b + u_L + \tfrac{1}{2}u_C \,. \quad \ldots\ldots \quad (1)$$

We distinguish two states of the circuit: 1) one of the two rectifying elements is conductive and the other not, and 2) neither is conductive.

1) *One rectifying element* (P_1) *is conductive and the other not*
As an approximation, then, $u_{P_1} = 0$, so that from (1):

$$\tfrac{1}{2}u_C = -(V_b + u_L) \,. \quad \ldots\ldots\ldots \quad (2)$$

The current i_1, which flows from t_0 to t_1 through P_1, can be shown to have the value:

$$i_1 = \frac{V_b}{R} - \frac{V_b}{R\sin\varphi}\,e^{-at}\sin(\omega t+\varphi) + \frac{V_b + \frac{1}{2}U_C}{\omega L}\,e^{-at}\sin\omega t, \quad \ldots\ldots \quad (3)$$

where $a = \dfrac{1}{2CR}$,

$$\omega = \sqrt{\frac{1}{LC} - a^2},$$

$$\varphi = \tan^{-1}\frac{\omega}{a},$$

U_C = voltage across C at the time $t = t_0$. (U_C can be found from a further calculation, not given here.)

For the voltage u_L during the interval t_0-t_1 we find:

$$u_L = -\frac{V_b}{\omega CR}\,e^{-at}\sin\omega t + \frac{V_b + \frac{1}{2}U_C}{\sin\varphi}\,e^{-at}\sin(\omega t - \varphi) \,. \quad (4)$$

The waveforms of i_1 and u_L during the interval t_0-t_1 are roughly illustrated in *fig. 6a* and *b*. Assuming that U_C is known, the waveforms of u_C and u_{P_1} can be constructed with the aid of (1) and (2) (see fig. 6c and *d*).

2) *Both rectifying elements are non-conductive.* After the moment t_1 no current passes through P_1. Up to the moment t_2 at which P_2 receives a control pulse, the R-C circuit is left to itself, so that C discharges exponentially with the time constant RC. Since i_1 and i_2 are now zero, so too is u_L, and the waveforms of the various voltages during the interval t_1-t_2 are easily found; see fig. 6*b*, *c* and *d*. The variation from the moment t_2 can also be found without much difficulty. For a more detailed analysis, reference may be made to the literature [10]).

It can be seen from fig. 6*d* that the peak voltage across the rectifying elements can considerably exceed V_b, the precise value depending on the dimensioning of the circuit and the load. This should be taken into account when choosing the value of V_b and the type of rectifying element.

It follows from (3) that the choice of L and C determines the duration t_0-t_1 of the conduction in every half cycle T. It is therefore possible after every pulse to give the rectifying elements a certain "recovery time", from t_1 to t_2 ($t_2 = t_0 + \frac{1}{2}T$) as needed to ensure proper operation.

The complete circuit diagram of the converter is shown in *fig. 7*. The resistive load of fig. 3 is now replaced by a number of "TL" S lamps (chosen because they operate without a starter). Half the

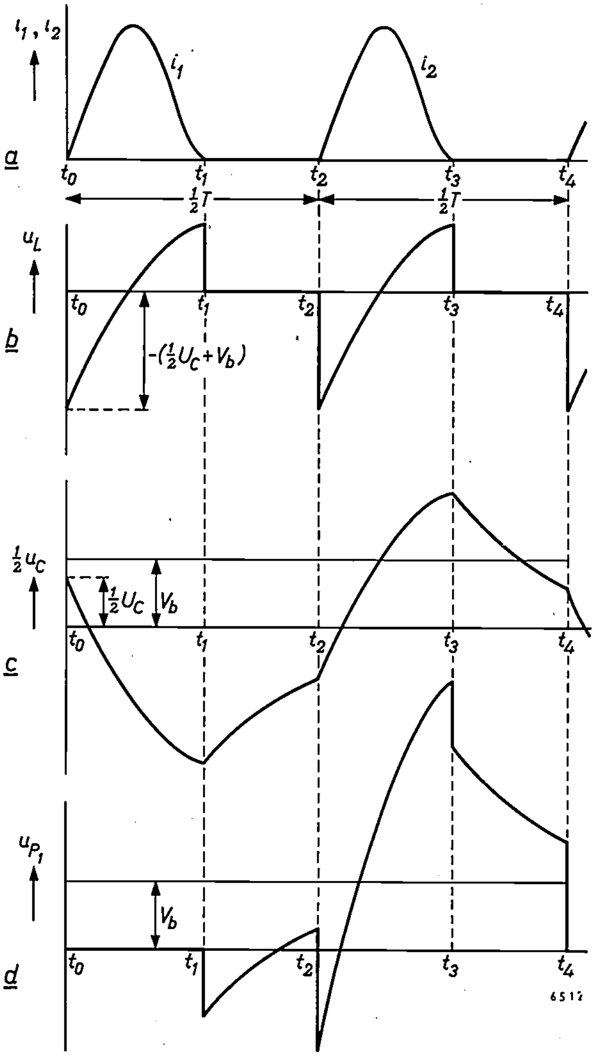


Fig. 6. Currents and voltages in the circuit of fig. 3. At the moments t_0 and t_4 the rectifying element P_1 receives a control pulse; at the moment t_2 a control pulse is received by P_2, and at t_1 and t_3 the currents i_1 and i_2 respectively are zero. t_0 and t_4 are separated by one period T.

a) Current i_1 through P_1 and current i_2 through P_2.
b) Voltage u_L across choke L (from t_0 to t_1 given by (4), from t_1 to t_2 equal to zero).
c) Voltage $\frac{1}{2}u_C$ over one half of the transformer primary (from t_0 to t_1 given by (2) and (4), from t_1 to t_2 varying as $-\exp(-t/RC)$).
d) Voltage u_{P_1} across P_1:
t_0 - t_1 . . . . $u_{P_1} \approx 0$,
t_1 - t_2 . . . . $u_{P_1} = V_b + \frac{1}{2}u_C$,
t_2 - t_3 . . . . $u_{P_1} = u_C$,
t_3 - t_4 . . . . $u_{P_1} = V_b + \frac{1}{2}u_C$.

[10]) W. Schilling, Berechnung des Parallelwechselrichters bei ohmscher Belastung, Arch. Elektrotechn. 29, 119-130, 1935. C. F. Wagner, Parallel inverter with inductive load, Trans. Amer. Inst. Electr. Engrs 55, 970-980, 1936.

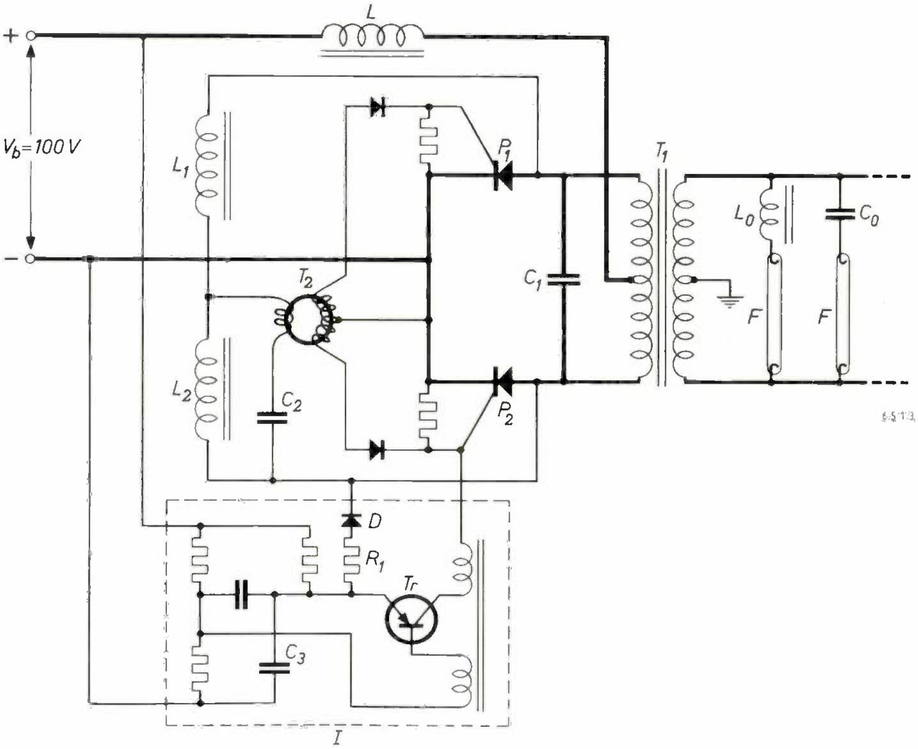


Fig. 7. Basic circuit of a converter using silicon controlled rectifiers, designed for train lighting. P_1, P_2, L, C_1, T_1 and V_b (here 100 V) as in fig. 3. F fluorescent lamps of type "TL" S (one half in series with chokes L_0, the other half in series with capacitors C_0). T_2 pulse transformer with ferrite core (rectangular hysteresis loop) fed by the converter itself via network L_1-L_2-C_2.

I starting-pulse generator. Capacitor C_3 discharges periodically through transistor Tr and the control electrode and cathode of P_2. Once the converter is oscillating, the anode of P_2 becomes negative with respect to the cathode during a part of each cycle (see fig. 6d); this negative voltage discharges capacitor C_3 through diode D and resistor R_1, which makes the starting-pulse generator inoperative.

number of lamps have a choke as ballast with an inductance L_0, the other half a small capacitor of capacitance C_0. The value L_0C_0 is chosen near to resonance at the frequency of the converter. Since the voltage is roughly sinusoidal (see fig. 6c), this load is not much different from a resistive load for the converter.

When the converter is switched on, it is set in operation by starting pulses from the transistorized pulse generator I. As soon as the converter starts oscillating, the pulse generator is automatically made inoperative (see caption to fig. 7). The control pulses are now delivered by the pulse transformer T_2, which has a ring-shaped ferrite core with a rectangular hysteresis loop. The primary of T_2 is fed from the converter itself via chokes L_1 and L_2 and the capacitor C_2. The pulse repetition frequency can be adjusted by varying L_2 and C_2.

Fig. 8 shows an experimental 1 kW converter designed on this principle, for operation from a 100 V DC supply. The load consists of 24 "TL" S 40 W lamps. *Fig. 9* shows a later version of this converter, intended for an experimental lighting installation in a train of the Netherlands Railways. Fig. 9

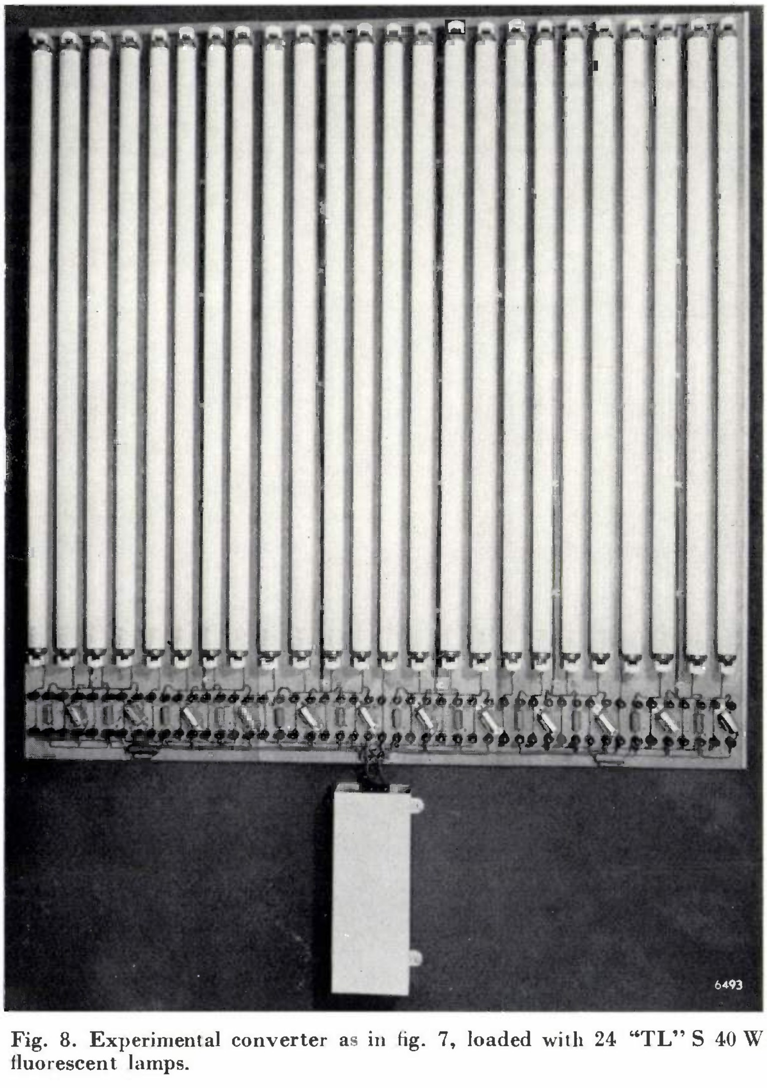


Fig. 8. Experimental converter as in fig. 7, loaded with 24 "TL" S 40 W fluorescent lamps.

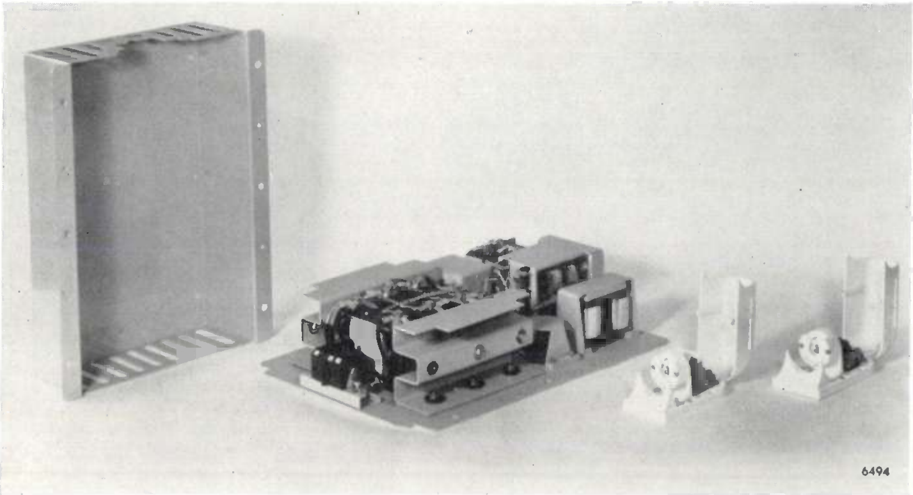


Fig. 9. Converter on the principle of fig. 7, designed for a fluorescent lighting system on trial in a train of the Netherlands Railways. Input voltage 100 V, power 1 kW, frequency 7 kc/s, efficiency better than 85%, weight approx. 10 kg. On the right, two standard lamp holders for a "TL" lamp, each fitted with a "ballast" (in one a choke, in the other a capacitor).

also gives an idea of how small the "ballasts" are, so small indeed that they can be accommodated in the standard covers of the lamp holders.

The frequency of the converter in fig. 9 is 7 kc/s, the efficiency more than 85%, and the weight about 10 kg.

Summary. For operating fluorescent lamps from a DC source (as in trains) use has mainly been made hitherto of rotary DC/AC converters. Electronic converters are now gaining ground, and a new type is discussed here which is equipped with *P-N-P-N* silicon controlled rectifiers. These rectifiers can handle a considerably higher power than transistors and have a voltage drop of only 1 to 1.5 V, which is an order of magnitude smaller than in thyratrons and ignitrons. Converters with silicon controlled rectifiers have a much wider field of application than rotary converters, which they will largely supersede in the near future. They can operate in the frequency range from 5 to 10 kc/s. This has particular advantages for fluorescent lighting, enabling small, light-weight ballasts to be used which have extremely low losses, and giving a luminous efficiency about 10% higher than at 50 c/s. A description is given of a converter using silicon controlled rectifiers which has been designed for train lighting; the input voltage is 100 V, the power 1 kW, the frequency 7 kc/s. In conjunction with mains rectifiers the new converters will be useful for fluorescent lighting in offices and factories. It is expected that the present fairly high costs of such installations will be reduced in the not too distant future, to such an extent that the energy savings will make them a profitable proposition.

AN INSTRUMENT FOR AUTOMATICALLY RECORDING ISOCANDELA DIAGRAMS OF BEAMED LIGHT SOURCES

by W. BÄHLER *).

535.247.4:628.971.85:629.113

One of the problems facing the designer of beamed light sources is to produce a beam pattern that meets specific requirements. For lighting airfields, for example, beams are required that are fairly broad in the horizontal plane but very narrow in the vertical. For beacon lights, signalling lamps and car headlamps the requirements are complicated, and where beacons are concerned they differ from case to case. For flood-lighting, too, beams are often needed that are not simply radially symmetrical.

As an example we shall briefly discuss, with reference to *fig. 1*, the present specifications applicable in many European countries to the dipped beam or passing light of car headlamps [1]). Broadly speaking, the beam should be such that a motorist driving on an unlighted road retains sufficient lighting to be able to see the road ahead without dazzling oncoming traffic. Fig. 1 gives a perspective drawing of a road 6 m wide as "seen" by a car headlamp at a height of 75 cm above the road surface, in the middle of the right half of the road. If we draw the system of lines in this figure, with the given dimensions, on a screen and set it up 25 m away from a headlamp, the light thrown by the headlamp on to the screen should meet the follow-

*) Philips Research Laboratories, Eindhoven.

[1]) A comparison of the properties of the (then) European and American dipped beam has been given by J. B. de Boer and D. Vermeulen, Philips tech. Rev. 12, 305, 1950/51.

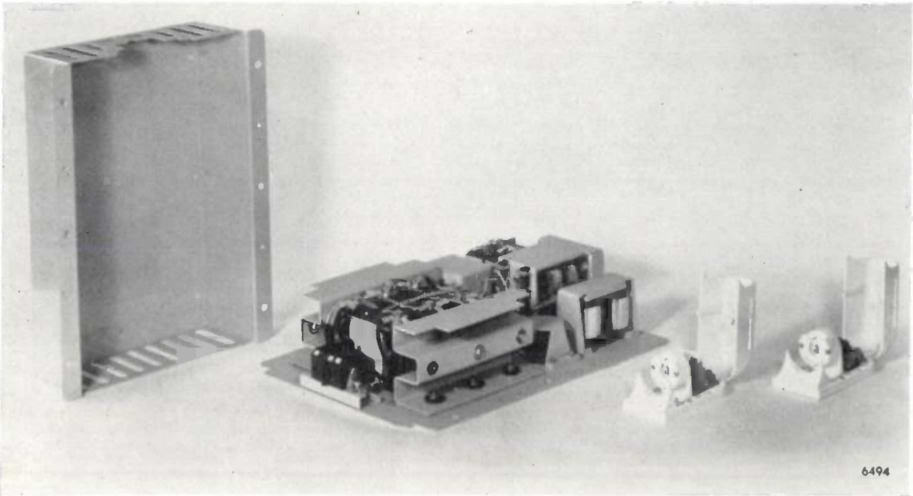


Fig. 9. Converter on the principle of fig. 7, designed for a fluorescent lighting system on trial in a train of the Netherlands Railways. Input voltage 100 V, power 1 kW, frequency 7 kc/s, efficiency better than 85%, weight approx. 10 kg. On the right, two standard lamp holders for a "TL" lamp, each fitted with a "ballast" (in one a choke, in the other a capacitor).

also gives an idea of how small the "ballasts" are, so small indeed that they can be accommodated in the standard covers of the lamp holders.

The frequency of the converter in fig. 9 is 7 kc/s, the efficiency more than 85%, and the weight about 10 kg.

Summary. For operating fluorescent lamps from a DC source (as in trains) use has mainly been made hitherto of rotary DC/AC converters. Electronic converters are now gaining ground, and a new type is discussed here which is equipped with *P-N-P-N* silicon controlled rectifiers. These rectifiers can handle a considerably higher power than transistors and have a voltage drop of only 1 to 1.5 V, which is an order of magnitude smaller than in thyratrons and ignitrons. Converters with silicon controlled rectifiers have a much wider field of application than rotary converters, which they will largely supersede in the near future. They can operate in the frequency range from 5 to 10 kc/s. This has particular advantages for fluorescent lighting, enabling small, light-weight ballasts to be used which have extremely low losses, and giving a luminous efficiency about 10% higher than at 50 c/s. A description is given of a converter using silicon controlled rectifiers which has been designed for train lighting; the input voltage is 100 V, the power 1 kW, the frequency 7 kc/s. In conjunction with mains rectifiers the new converters will be useful for fluorescent lighting in offices and factories. It is expected that the present fairly high costs of such installations will be reduced in the not too distant future, to such an extent that the energy savings will make them a profitable proposition.

AN INSTRUMENT FOR AUTOMATICALLY RECORDING ISOCANDELA DIAGRAMS OF BEAMED LIGHT SOURCES

by W. BÄHLER *).

535.247.4:628.971.85:629.113

One of the problems facing the designer of beamed light sources is to produce a beam pattern that meets specific requirements. For lighting airfields, for example, beams are required that are fairly broad in the horizontal plane but very narrow in the vertical. For beacon lights, signalling lamps and car headlamps the requirements are complicated, and where beacons are concerned they differ from case to case. For flood-lighting, too, beams are often needed that are not simply radially symmetrical.

As an example we shall briefly discuss, with reference to *fig. 1*, the present specifications applicable in many European countries to the dipped beam or passing light of car headlamps [1]). Broadly speaking, the beam should be such that a motorist driving on an unlighted road retains sufficient lighting to be able to see the road ahead without dazzling oncoming traffic. Fig. 1 gives a perspective drawing of a road 6 m wide as "seen" by a car headlamp at a height of 75 cm above the road surface, in the middle of the right half of the road. If we draw the system of lines in this figure, with the given dimensions, on a screen and set it up 25 m away from a headlamp, the light thrown by the headlamp on to the screen should meet the follow-

*) Philips Research Laboratories, Eindhoven.

[1]) A comparison of the properties of the (then) European and American dipped beam has been given by J. B. de Boer and D. Vermeulen, Philips tech. Rev. 12, 305, 1950/51.

ing requirements. In the first place the transition from light to dark (the cut-off) should be sharp enough for it to be used to adjust the headlamp. This is done in such a way that the cut-off falls on the left half of the screen on a horizontal line 25 cm below the line *h-h*; on the road this appears as a transverse line at a distance of 75 m in front of the car. Right of centre the light-dark cut-off should run upwards. When the lamp is thus adjusted, the illumination in the various zones and points of the screen should not exceed the values indicated in red. As can be seen, the lower boundary of zone III on the right side is inclined at 15° to the horizontal. The illumination of the right kerb must also be fairly high. A beam meeting these requirements will thus clearly be asymmetric [2]).

It will be evident that the designer of a beamed light source which must comply with these complicated requirements will want to measure the level of illumination at numerous points of the beam projected by the headlamp on to the screen. The usual practice is to determine a series of points where the illumination has a specific value and to draw through these points a closed curve, called an isolux curve. Since the headlamp produces the same luminous intensity in all directions corresponding to the points in such a curve (at least in the case of narrow beams) the curve may equally be regarded as an isocandela curve. A complete light-distribution diagram consists of ten or more isocandela curves.

In some cases the demands made on the accuracy of the isocandela curves are so high as to call for an exceptionally large number of measuring points. To eliminate the time-consuming measurements which this involves, an instrument has been designed at Eindhoven which is capable of tracing isocandela curves automatically. All that has to be done is to adjust on the instrument the luminous intensity for which an isocandela curve is required. In this way a complete isocandela diagram can be recorded in about 20 minutes.

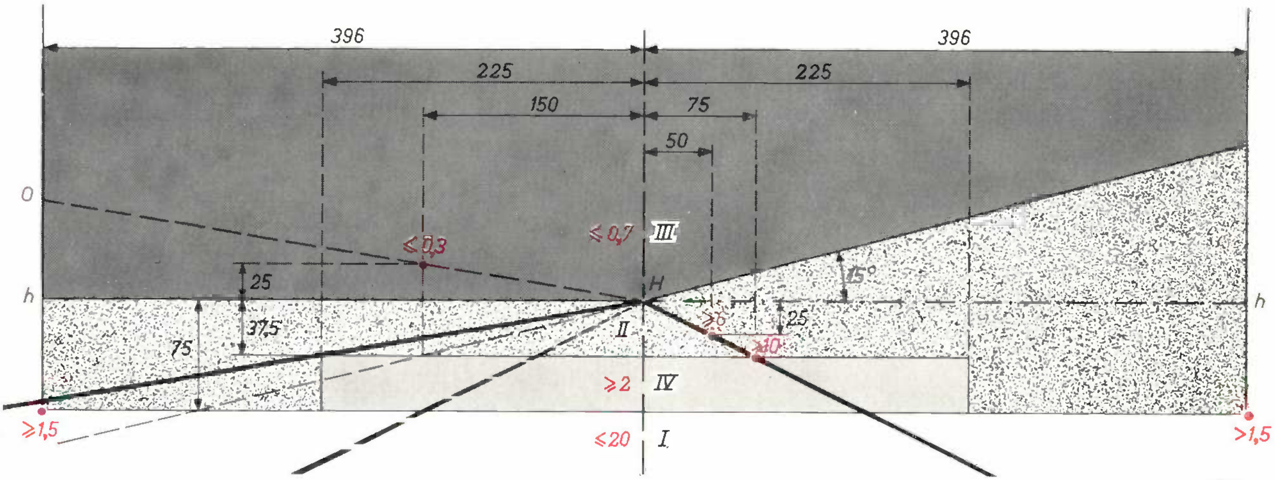


Fig. 1. Perspective sketch of a 6 metre wide road as seen by an eye 75 cm above the middle of the right lane. The dashed line *HO* is the line roughly followed by the eye of an oncoming motorist. To ascertain whether the passing light of a car headlamp complies with the requirements, this diagram, with the dimensions given, is drawn on a screen which is placed at a distance of 25 m from the lamp. The illumination at various points and in zones *I-IV* of the screen is required in most European countries to meet the standards indicated in red. The headlamp should be aligned so that the sharp light-to-dark cut-off required on the left lane coincides on the screen with a line 25 cm below the line *h-h* (that is 75 m ahead of the car on the road). After switching to the driving light the illumination at point *H* should not be less than 90% of the maximum.

In this article we shall describe the operation and design of this isocandela-diagram recorder [3]). We shall begin with the principle of its operation, and it will be shown that the instrument can be regarded as a closed control loop. After discussing the most important properties of the recorder, we shall describe its construction. In the last part of the article, we shall discuss the various parts of the recorder from the point of view of control theory.

Principle of operation

The principle underlying the operation of the isocandela-diagram recorder can be explained by considering the luminous intensity pattern as a

[2]) See e.g. J. B. de Boer, The "Duplo" car headlamp bulb with an asymmetric dipped beam, Philips tech. Rev. **16**, 351-352, 1954/55.

[3]) A brief description of a provisional model of the instrument was given in Philips tech. Rev. **20**, 288, 1958/59.

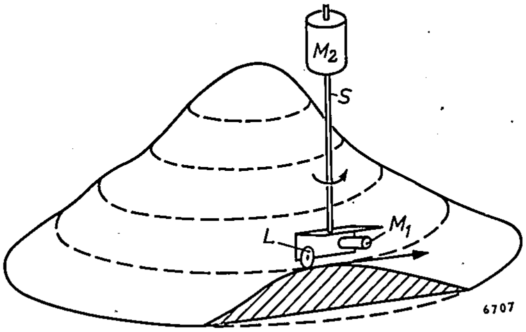


Fig. 2. Principle underlying the operation of the isocandela-diagram recorder, illustrated by considering an instrument that automatically draws contour lines on a mountain. The wheel L of the trolley is driven by a motor M_1. The trolley is fitted with an altimeter. If the altitude at which the wheel runs over the mountain slope differs from the preset value, the difference signal delivered by the altimeter actuates the steering motor M_2, producing a change of course that returns the trolley to the required contour.

mountain landscape (*fig. 2*). The height corresponds to the level of illumination on the screen, and the isolux lines are contour lines in the mountain landscape. The instrument which measures the illumination, the photocell, is in this case an altimeter. The latter is mounted on a single-wheeled trolley which is driven around the contours by a small motor. The trolley is fixed to the lower end of a vertical shaft, the steering rod, the position of which is controlled by a servo-motor.

From fig. 2 it can be seen that, if the steering rod is not turned, the trolley will not go on following the contour line on which it started. In the case sketched here it will run off the mountain, and the height indicated by the altimeter will be lower than that of the wanted contour line. The fact that any deviation from the desired value is immediately ascertained creates the possibility of making the servo-motor do its work correctly. For that purpose the deviation is converted into an electric signal, which is applied after amplification to the servo-motor. If the components are coupled with the right polarity, the servo-motor will now turn the steering rod in such a way as to correct the deviation from the proper course (in our case the contour line).

It may be concluded from the foregoing that the steepness of the slope along which the trolley moves must influence the behaviour of the instrument. On an almost *flat* part of the slope a given deviation from the correct course will produce only a slight change in the altimeter reading; in other words, the signal used to control the trolley will be small compared with that for a deviation of the same magnitude on a *steep* part of the slope. We mention this fact at the present stage because of its important bearing on the stability of the control loop.

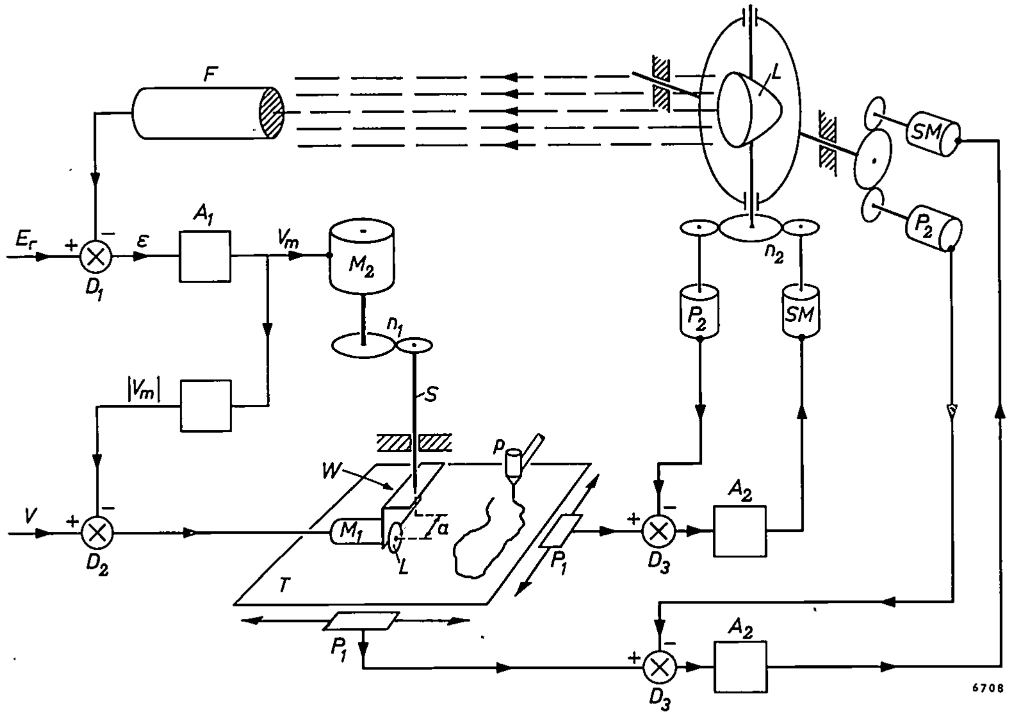


Fig. 3. Basic diagram of the isocandela-diagram recorder. Top right the lamp L, whose azimuth and elevation can be varied within defined limits. F calibrated photocell in fixed position. The output signal from F is compared in D_1 with the variable reference signal E_r. The difference signal ε is amplified in A_1 and energizes the steering motor M_2, which, via a reduction gear, steers the trolley W by turning the steering rod S, mounted in fixed bearings. The drive wheel L of the trolley is driven by motor M_1. T recording table which can be moved in its plane, by L, in two mutually perpendicular directions. The positional coordinates of T are each converted by separate servo-systems into azimuth and elevation of the lamp. Each servo-system consists of a position pick-off P_1 by the table, a servo-motor SM, a position pick-off P_2 by the lamp, a difference circuit D_3 and an amplifier A_2. The isocandela curve corresponding to the value of E_r is traced by a fixed stylus p on a sheet of paper fixed to the table T. The motor M_1 is supplied with the signal delivered by D_2, being the difference between the absolute value $|V_m|$ of the control-motor voltage and a constant voltage V. The point of contact of the wheel on the table is not in line with the steering rod but at a small off-set a from the axis of the rod.

The operation of the isocandela-diagram recorder itself is represented schematically in *fig. 3*. The difference compared with the contour recorder just described is that instead of the light-distribution pattern ("mountain") remaining stationary and the photocell (the "trolley") moving along the screen, the photocell remains stationary and the beamed light source is rotated. Middle left we again see the trolley at the lower end of the steering rod S, which here turns in fixed bearings. The wheel of the trolley is driven by the motor M_1. In its turn the wheel sets in motion the recording table T, which can move in its plane in two mutually perpendicular directions. The two coordinates that determine the position of the recording table are each converted by a positional servo system into an angle of rotation, namely the azimuth and the elevation of the lamp L. The illumination produced by the lamp in this position on the stationary photocell F is converted by the latter into an electrical signal. This is compared in the circuit D_1 with a reference signal E_r, which corresponds to the value of illumination for which a curve is required. The difference signal ε is amplified in A_1 and fed to the steering motor M_2. A fixed stylus maps out the curve on graph paper fixed on the recording table. In principle it is immaterial where the paper and stylus are situated; the positional coordinates of each point on the recording table always vary in the same way, and the curve traced is therefore in any case congruent with the path described by the wheel.

The diagram produced by the instrument is not an exact scaled-down copy of the diagram that might be drawn on the screen referred to (fig. 1). The difference is that the coordinates used are the azimuth and elevation of the given direction — i.e. the angles — and not the distance on the screen. As will be known, this makes no difference where small angles are concerned.

It can be seen that the essence of the method is the feedback brought about between photocell and lamp by D_1, the steering motor and trolley, the recording table and the positional servo systems, and which automatically ensures that the lamp points only in those directions where the illumination has the desired value.

As can be seen from the figure, the driving motor M_1 is not fed with a constant voltage but with the difference between the absolute value of the input signal of the steering motor M_2 and a constant voltage V. The significance of this will be discussed presently, as also will the fact that the point where the wheel touches the recording table T does not lie in line with the steering rod but at a distance a to one side.

The instrument as a control loop

Consideration of fig. 3 shows, as mentioned in the introduction, that the system can be regarded as a closed control loop. *Fig. 4* gives a schematic diagram of this loop.

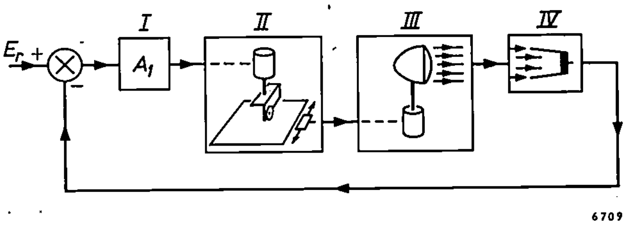


Fig. 4. The isocandela diagram recorder as a control loop. Block I represents the amplifier A_1 of fig. 3, block II the control motor with trolley, table and position pick-offs, block III the rest of the positional servo-systems and the lamp, and block IV the photocell whose output signal is compared with the reference signal E_r.

Although the characteristics of the components of the control loop will be dealt with at length at the end of this article, it will be useful here to discuss briefly the characteristics of the loop as a whole. In doing so we shall make use of a simplified formula for the transfer function $KG(j\omega)$ of the open loop — i.e. the complex ratio between the output signal of the photocell and the input signal of A_1 (fig. 4) for the case where the feedback is interrupted — and derive from this formula the Bode and Nyquist diagrams [4]. The formula reads:

$$KG(j\omega) \approx$$
$$\approx \text{const.}K_E R_p(E)\,\frac{1+j\omega\tau_3}{(j\omega\tau_0)^2(1+2\zeta j\omega\tau_6-\omega^2\tau_6{}^2)}, \quad (1)$$

where $\tau_3 = a/v$ is the quotient of the above-mentioned distance a (see fig. 3) and the velocity v of the trolley, i.e. τ_3 is the time taken by the wheel to cover the distance a. Further, τ_0 is the unit of time, τ_6 a time constant whose value is determined by the properties of the positional servo-systems, K_E the gradient of the illumination at the place which the photocell "sees" at the relevant moment — a quantity which varies during the tracing of the curve — and ζ is a damping constant connected with τ_6 and with various constants of the system which do not otherwise appear in (1) (see final section). Finally, the factor $R_p(E)$ is due to the photocell, its value varying with the illumination E.

The pertaining Bode diagram, approximated by three straight lines, is sketched in *fig. 5*, and the Nyquist diagram in *fig. 6*. The successive straight lines have the slopes —2, —1 and —3. The point of

[4] For the concepts, definitions and methods of calculation in control engineering, as used in this article, the reader may be referred to three articles on control-engineering subjects, which appeared a few months ago in this journal (pages 109, 151 and 167 of numbers 4, 5 and 6 respectively).

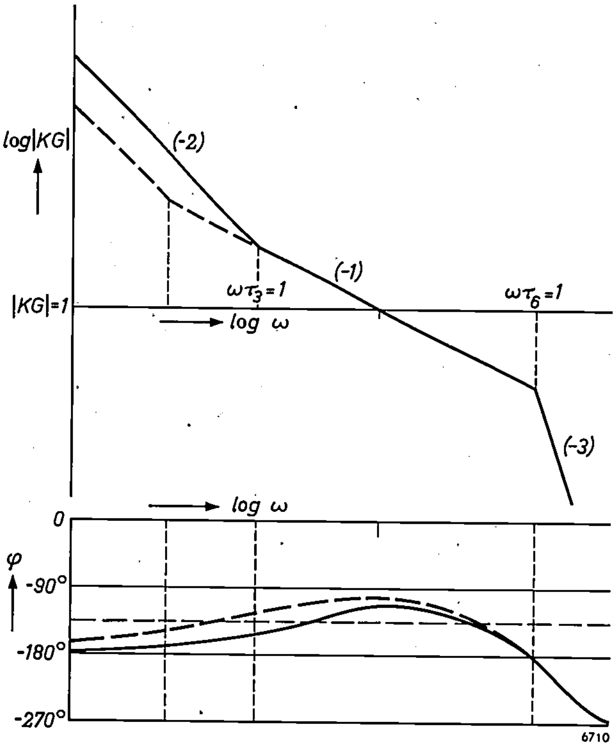


Fig. 5. Bode diagram (schematic) of the complete open loop. The amplitude characteristic, approximated by straight lines, consists of three branches of slope —2, —1 and —3. The frequency $\omega = 1/\tau_3$ at which the first break occurs varies with the peripheral velocity v of the wheel and the distance a (cf. fig. 3): $\tau_3 = a/v$. The phase shift φ at this frequency is more than 135°. The dashed lines show how the two characteristics change as v decreases. The gain then becomes lower and the phase margin (i.e. 180° — $|\varphi|$) wider.

intersection of the first two lines lies on the vertical $\omega = 1/\tau_3$. The phase shift at this frequency is more than 135°. The second break occurs at $\omega = 1/\tau_6$. The phase shift here is about —180°. (The phase shift at $\omega = 1/\tau_3$ is closer to —135° the greater is the difference between τ_3 and τ_6 and the smaller is ζ.)

The Nyquist diagram shows that the system is stable when the various constants have the values corresponding to the curves in figures 5 and 6. Going along the Nyquist curve from the point corresponding to $\omega = \infty$ (the origin) to the point corresponding to $\omega = 0$, we see that the point (—1, 0) is everywhere on our right — except where it is screened by another part of the curve [5]).

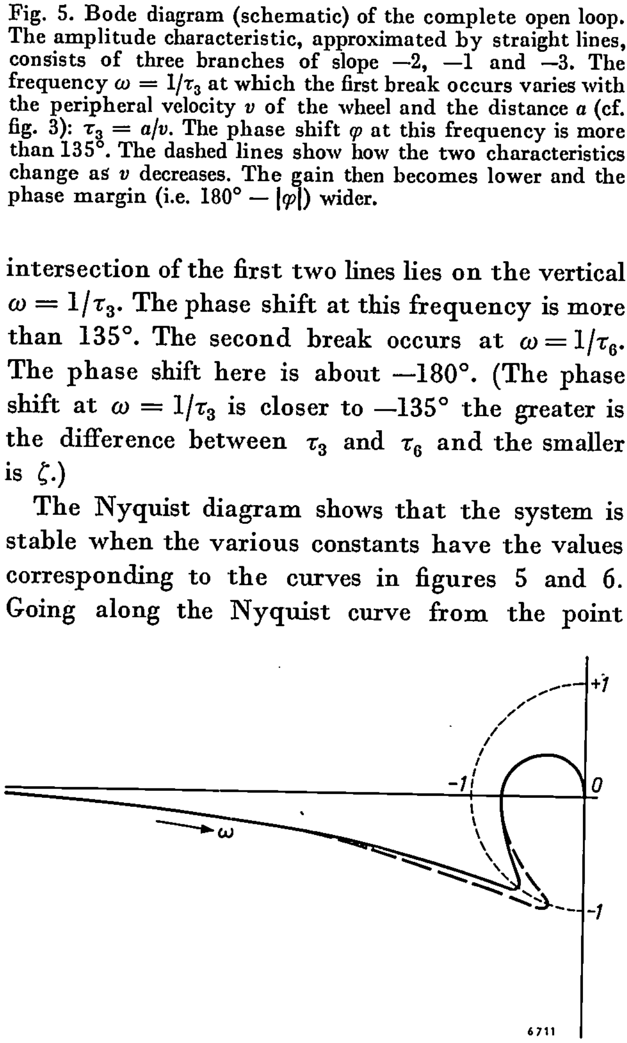


Fig. 6. Nyquist diagram derived from fig. 5. Here too, the dashed line relates to a lower value of v than the full line.

The stability characteristics differ, however, from those encountered in simple control systems. Here, too, of course, instability occurs if the gain is increased — the point (—1, 0) in fig. 6 then shifts to the right in relation to the curve — but there is also a lower limit to the gain. Although there is no instability when the gain is very small, the phase shift in such a case is so close to —180° — i.e. the phase margin is so small — that the damping of the system is too weak and oscillations last too long.

As indicated above, K in this control loop is not a constant that can be fixed at a desired value: K is proportional to the gradient K_E of the illumination pattern. These variations may be so considerable — as much as a factor of 20, see *fig. 7* — as to jeopardize the stability of the system. The inconstancy of R_p presents no difficulties. Along one and the same isolux contour E is of course constant and so too then is R_p. The change in R_p upon the transition to a succeeding value of E can be easily compensated by resetting the loop gain at the old value before tracing a further isolux curve.

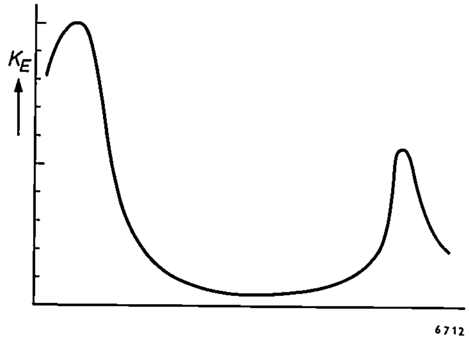


Fig. 7. Variation of the gradient K_E of the illumination along the isocandela curve for 4.0 lux in diagram A on p. 243 of this number. The extreme values of K_E are no less than a factor of 20 apart.

To explain the method of solving the difficulty presented by the variation of K_E, we shall first discuss the manner in which $G(j\omega)$ changes when τ_3 is varied. We begin with the extreme case where the distance a is 0 and thus τ_3 is also zero. In this case the Bode diagram, as far as amplitude is concerned, consists solely of two straight lines of slope —2 and —4, and the absolute value $|\varphi|$ of the phase shift is $\geqq$ 180° for all frequencies (= 180° for the low frequencies and > 180° for the high).

[5]) The simplified version of Nyquist's stability criterion used here is equivalent to that used in the first article mentioned in reference [4]).

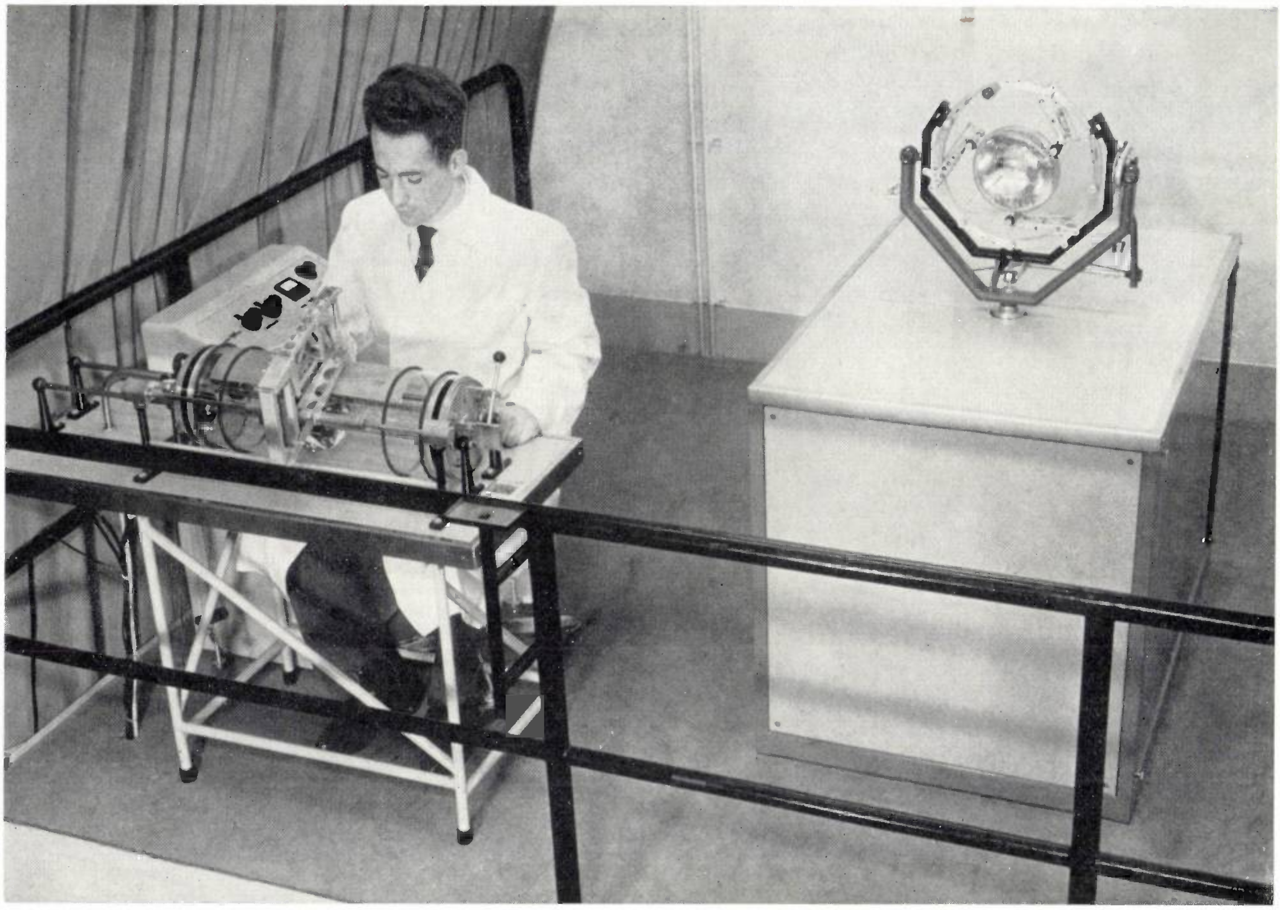


Fig. 8. View of the platform containing the major part of the equipment. The recording portion is mounted on the table, left (the recording table T of fig. 3 is here "rolled up" to form a drum). Behind the operator can be seen the control desk. The lamp holder, right, is mounted on a cabinet which provides a dust-free enclosure for the associated control mechanisms. The photocell is situated 25 metres from the lamp. Much of the electronic circuitry is contained in a rack under the platform.

For the low frequencies the Nyquist curve coincides with the negative real axis and is even above it for the high frequencies. At no value of amplification, then, can the system possibly be stable. *It is therefore an essential condition for the stability of the instrument that the distance "a" should not be equal to zero.*

The effect of limited variations of v on the Bode diagram is to move the first break along a line of slope —1. If τ_3 is increased the loop gain therefore decreases in the range of low frequencies. At the same time the phase margin widens (see dashed lines in fig. 5 and fig. 6). It is this effect that is used for more or less compensating the troublesome variation of K with K_E. Returning to fig. 3, we see that the motor M_1 that drives the wheel does not turn at a constant speed, but is controlled by the difference between the absolute value of the control signal, which is proportional to K_E, and a constant voltage V, which is higher. If K_E increases, then, the motor turns more slowly and τ_3 increases. The rise in K due to the increase of K_E is thus opposed by the associated increase in τ_3.

This rough compensation of the variations of K_E is sufficient for practical purposes, and is much simpler than a method involving measuring K_E and varying the gain factor of A_1 correspondingly. It should be noted in this connection that the velocity of the driving wheel varies not only with K_E but also with the radius of curvature of the isocandela curve. This is an additional advantage, since it means that sharp corners are carefully traced.

Particulars of construction

We shall now review the construction of the various components of the isocandela-diagram recorder [6]). It should be mentioned first of all that the mechanical operation of the instrument differs in one point from that implied in fig. 3. In that figure the steering rod and the stylus are represented for simplicity as being in fixed positions, and the recording table T as capable of moving in two direc-

[6]) The solution of the various problems of design was the work of L. de Wit of Philips Lighting Division.

tions at right angles to each other. The provisional version described earlier in this review [3]) already differed from this arrangement to the extent that the table moved in one direction and the trolley together with the stylus in the other. In the definitive version we have gone a step further and have turned the recording table into a drum which can rotate about its axis. During rotation the drum, like the earlier recording table, moves in its own plane in one coordinate direction. The paper is attached to the *outside* of the drum, and the driving wheel moves over the *inside.*

Fig. 8 shows the form and arrangement of the equipment. On the left can be seen the drum and the control desk, on the right the lamp-holder mounted on top of a cabinet which contains the positional servo-systems.

The drum with trolley and stylus

We shall now explain the construction of the drum and ancillary equipment with reference to *fig. 9.* On the table, at each end of the drum, are two supports *1*, which carry two parallel horizontal rods *2*. Attached to these rods are the two end plates *3* of the drum. Inside the drum itself the rods act as guide rails along which the carriage moves. The cylindrical surface of the drum, which is made of transparent plastic material, rotates about the two fixed end plates, which are provided with bearings in the form of steel end rings.

The trolley consists of two parts. The upper part (*D*), the "carriage", is equipped with three ball bushings for running on the guide rails. Also mounted on the carriage are the steering motor, the bearing for the steering rod and a reduction gear between the steering motor and the steering rod (see also *fig. 10*). The other part of the trolley (*W*) is at the bottom end of the steering rod and consists of a holder which carries the wheel and its driving motor.

By means of a lever system the wheel can be lifted from the drum surface, enabling drum and trolley to be moved independently of one another. The lever mechanism is operated by the handle *4*.

The stylus holder *5* can move parallel to the axis of the drum along the rod *6*. It is connected by cords to the trolley: if the trolley moves to the right, the stylus goes just as far to the left, and *vice versa.*

The paper on which the diagrams are drawn is fixed to the outside of the drum by rubber rings *7*.

The two positional coordinates, that is the angle of rotation of the drum and the position of the trolley on the rails, are transmitted electrically

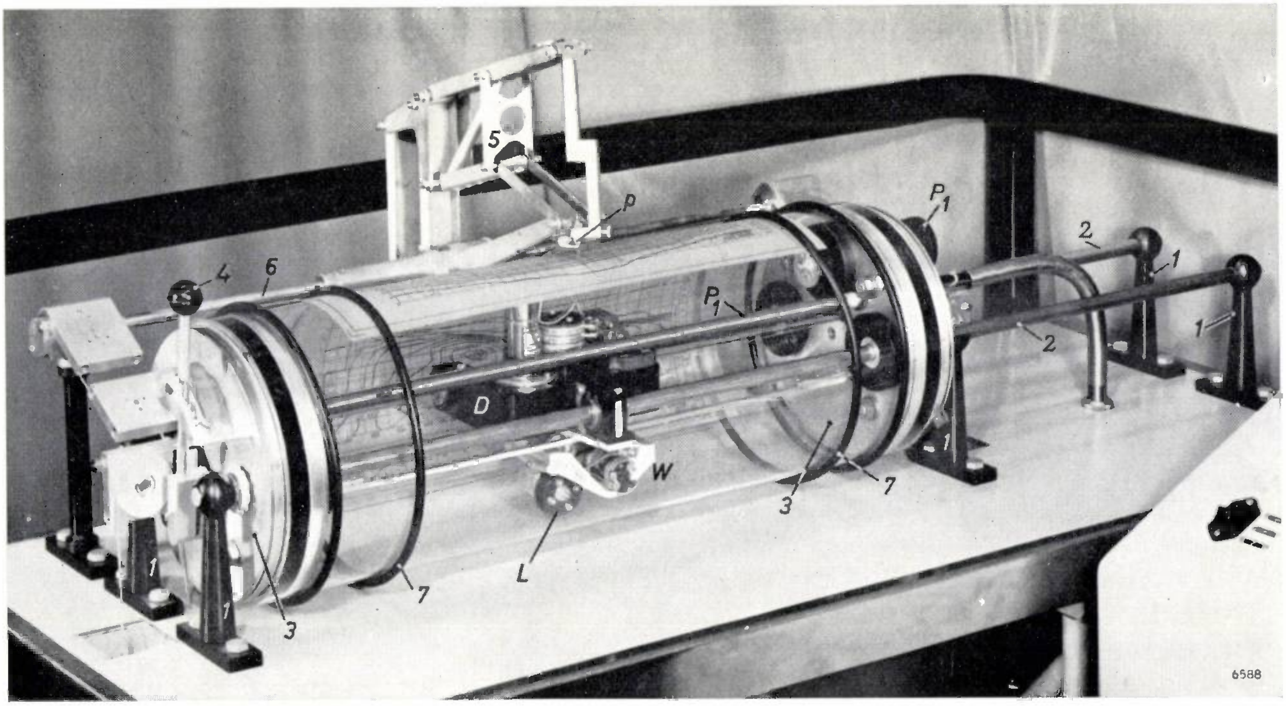


Fig. 9. The recording drum. *1* four supports for the horizontal rods *2* which act as guide rails inside the drum. *3* end plates fixed to *2*, also acting as drum bearings. *W* is the trolley connected to the bottom of the steering rod. *D* carriage, containing the steering-rod bearings and the steering motor. *4* handle for raising the trolley wheel *L* from the drum. *5* holder of stylus *p* on guide rail *6* and connected by cords to the trolley. *7* two rubber rings for clamping recording paper to the outer surface of the drum. P_1 positional potentiometers which follow the translational motion of the trolley and the rotation of the drum.

to the lamp, as mentioned in the introduction. For this purpose the drum is equipped with two potentiometers (P_1 in figs 9 and 10). One of them is a rotary potentiometer and is mounted on the outer face of one of the end plates. It is set in motion via a gear transmission by the steel ring at the relevant end of the drum, the ring being provided with teeth for this purpose. The other potentiometer is wound on a rod inside the drum and is secured, parallel with the guide rails, to the end plates. The pertaining slide contact is fixed to the trolley.

The construction of the recording part of the instrument described here is preferable for various reasons to a flat table. In the first place the dimensions of the apparatus are considerably reduced by "rolling up" the table. This is not only due to the drum form as such, but also to the fact that a flat surface would have to be appreciably larger than the "rolled out" drum, because the wheel, in order to avoid the risk of smudging, must not pass over a part of the paper that has already been traced. With the drum this is no problem because the paper and the wheel are on different sides of the wall. In the second place the drum construction makes it possible to accommodate part of the mechanism in a dust-free space. This is particularly important for the proper operation of the rod-potentiometer and for the guide rails along which the trolley moves.

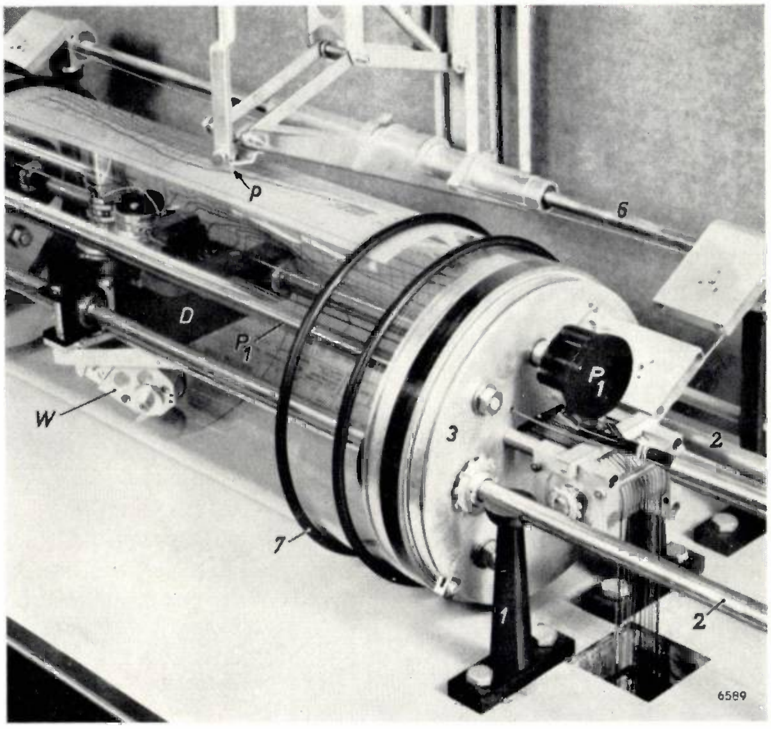


Fig. 10. The drum as seen from the control desk. The letters and figures have the same meaning as in fig. 9. Right, the trolley supply cables.

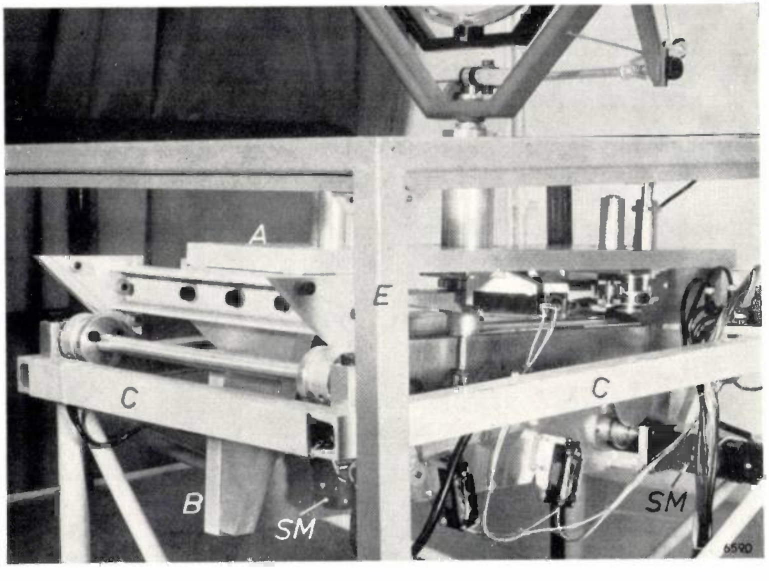


Fig. 11. The lamp holder and the two positional servo-systems. *A* horizontal plate carrying the lamp holder and the azimuth system. *B* vertical plate (joined rigidly to *A*) which carries the elevation system. The whole assembly can be turned through a small angle with respect to the frame *C*, without disturbing the lamp. *SM* positional servo-motors; the motor on the right is for the elevation system. *E* frame of dust-free cabinet.

The lamp holder and the positional servo-systems

The lamp holder and its positional servo-systems are mounted on two mutually perpendicular plates which are rigidly interconnected (*fig. 11*). The horizontal plate carries the lamp holder, and each of the plates carries a positional servo-system. The entire assembly is supported by a frame which has the form of a table without a top, and can rotate about an imaginary horizontal axis through the headlamp, at right angles to the line between headlamp and photocell. This makes it possible to adjust the lamp holder accurately with respect to the latter connecting line.

The construction of the lamp holder is shown in *fig. 12*. The lamp is fixed inside a ring by means of three self-centring clamps. The ring is mounted in a horseshoe-shaped bracket. The beam pattern can be adjusted "horizontally" by turning the ring in

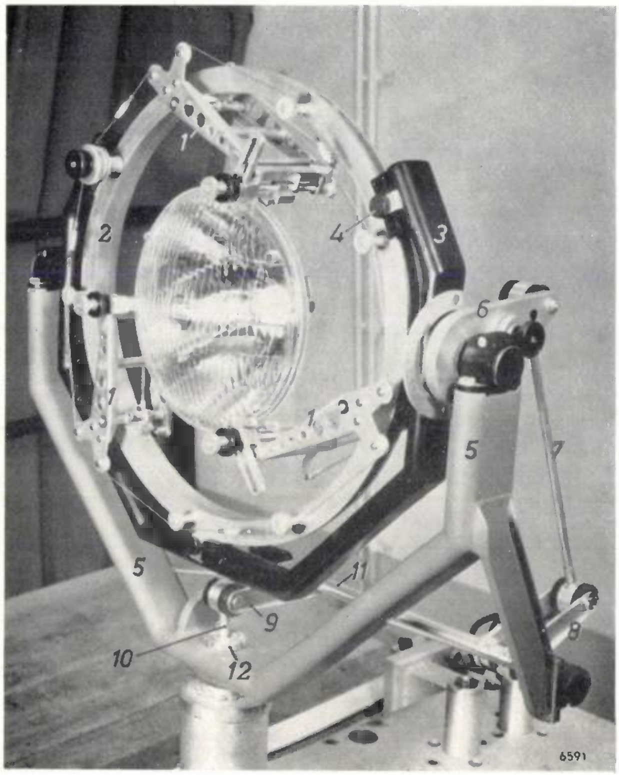


Fig. 12. The lamp holder. *1* self-centring clamps for holding the lamp in position. *2* ring which can turn in bracket *3*. After the appropriate alignment is found, the ring is locked with screws *4*. *5* fork in which *3* is mounted. *6-10* levers and rods which transmit the shaft rotation of the elevation servo-motor to the bracket, etc. Levers *8* and *9* are mounted on the horizontal spindle *11*. The rod *10* passes through the hollow shaft *12* of the fork (cf. fig. 13).

relation to the bracket. Once the right setting has been found the ring is locked in that position.

The bracket is in its turn mounted on bearings in a fork. The line through both bearings is horizontal and passes through the centre of the lamp. The servo-motor which ensures that the elevation of the lamp corresponds to the angle of rotation of the drum can turn the bracket in its bearings by means of a system of rods and levers (see *fig. 13*).

This transmission system is designed so that one of the rods is vertically in line with the centre of the lamp holder and fits into the hollow shaft of the fork. This rod, which moves vertically when the lamp is turned about its horizontal axis, is composed of two sections in line, connected by a coupling. The latter allows relative rotation of the two sections — and hence rotation of the lamp holder — but constitutes a rigid joint for translational movement.

The bottom of the rod is hinged to a lever, one end of which has the form of a toothed sector. The lever is actuated via a gear transmission by the above-mentioned servo-motor.

For varying the azimuth of the lamp, no such complicated mechanism is necessary. In this case a lever with toothed sector is mounted directly at the bottom of the fork. This lever is driven in the same way as the other by the azimuth servo-motor.

The positional indicators showing the state of adjustment of azimuth and elevation are potentiometers (P_2 in fig. 3), whose sliding contacts are driven via a separate transmission by the relevant lever with toothed sector.

For this purpose the levers have a sector without teeth at their other end, the movement of which is transmitted to a drum by means of two metal bands. The movement of this drum is transmitted with great precision to the potentiometer by a system of gears (*fig. 14*).

In this way the required accurate transmission between the elevation and azimuth shafts and the appertaining potentiometer is achieved with the minimum of precision gearing. High accuracy as regards angles is required in this transmission be-

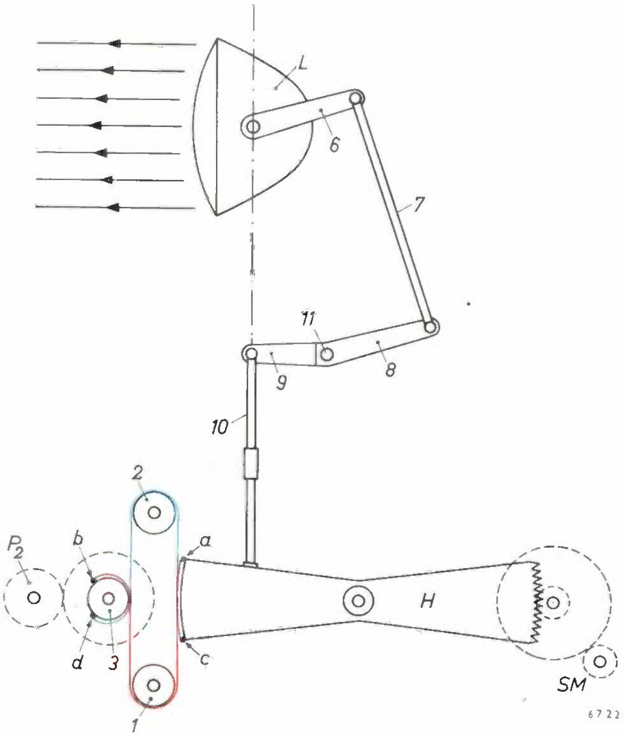


Fig. 13. Illustrating the method of transmitting the shaft rotation of the elevation servo-motor *SM* to the lamp. *L* lamp. *6*, *8* and *9* levers. *7* and *10* rods. *11* spindle on which *8* and *9* are mounted (cf. fig. 12). The rod *10* consists of two parts which, for the azimuthal movement, can rotate relative to each other about the common axis. *H* lever, having a toothed sector on the right and a plain sector on the left. The movement of *H* is transmitted to positonal potentiometers P_2 by two metal bands. One band (red) runs from point *a*, where it is fixed to the sector, over roller *1* to point *b* on drum *3*. The other (blue) runs from *c* over *2* to *d*. The spindle of *3* is fitted with a gear wheel which, via a second gear, rotates the shaft of potentiometer P_2.

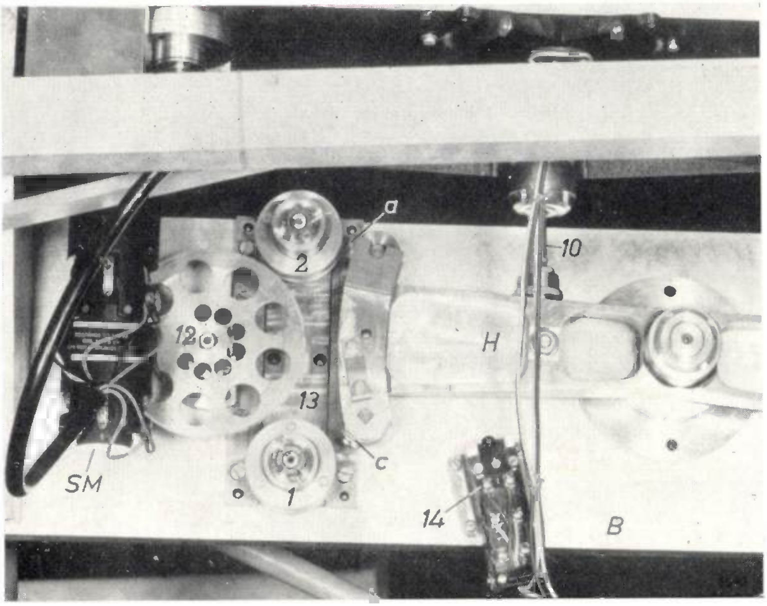


Fig. 14. Transmission of the elevation coordinate to the relevant positional potentiometer. The letters and figures have the same meaning as in fig. 13. The roller *3* in the latter figure is located behind the large gear wheel *12*. Rollers *1* and *2* are not mounted on the vertical plate *B* (cf. fig. 11) but on a brass strip *13*, secured only in the middle to *B*. The suitable choice of materials and dimensions ensures that the strip is not stressed too much or too little. If *H* is in danger of being turned through too wide an angle, end-switch *14* cuts out the elevation servo-motor. *SM* is the azimuth motor.

cause any error will not automatically be corrected. This precision is not demanded from the gears at the motor side, since they do not affect the positional accuracy.

The control desk

As may be seen in fig. 8, a control desk is situated beside the table on which the drum is mounted. The control desk comprises the following components.

a) The main switch, and the switches for the positional servo-motors and the driving motor (in fig. 3 SM and M_1 respectively).

b) Selector switches for presetting the reference voltage E_r, and hence the desired value of the luminous intensity.

c) Potentiometers for slightly modifying the alignment of the lamp holder in both coordinate directions without changing the position of the trolley relative to the recording drum.

   If the stylus is set on the recording paper at the zero point of the coordinate system, these potentiometers can be used to make the axial direction of the beam coincide with the line between photocell and lamp.

d) A switch for shifting the zero point of the elevation coordinate by a fixed amount, so that the distribution diagrams of the dipped beam and the main beam of a car headlamp can be recorded one below the other on a single sheet of paper, without having to reset the lantern.

e) A galvanometer for checking, before recording an isolux curve, whether the illumination on the photocell roughly corresponds to the value for which the curve is required. The trolley and the drum, and thus also the lamp coupled with them, must be set in such a way as to satisfy this condition. If the deviation is excessive, the apparatus will not automatically "home" on the required isolux curve.

f) Potentiometers for varying the gain in the control loop and the damping. This facility is necessary in view of the above-mentioned fact that the transfer function contains the factor $R_p(E)$, the value of which differs for each different isolux curve. Even if this were not so, it would still be desirable to have some means of varying the gain in connection with the automatic compensation of the variations in K_E: this is most effective when the range of loop gain has a specific position between the minimum and maximum limits of amplification.

The loop transfer function derived from the elements

The remainder of this article will now deal in control-engineering terms with the various components of the isocandela-diagram recorder, regarded as a control loop. We shall consider successively the following elements (cf. figs 3 and 4):

1) The photocell and associated circuit (including the circuit which compares the output signal with the reference signal).
2) The amplifier (A_1) and the control motor M_2.
3) The drive of the drum by the trolley wheel.
4) The positional servo-systems.
5) The lamp (beam pattern).

We shall then derive the formula for the loop transfer function from the transfer functions of these elements, and show that, as far as stability considerations are concerned, the formula can permissibly be reduced to the form of equation (1).

In deriving the transfer functions of the various components we shall treat the latter as linear elements, although some of them are by no means so. The approximation is entirely justified, however, as far as concerns the frequencies and amplitudes occurring in the loop when the instrument is in the process of recording an isocandela curve.

Photocell and associated circuit

The photocell used in the recorder is a selenium barrier-layer cell. Its spectral sensitivity is made equal to that of the human eye by means of filters. The spectral composition of the light can therefore be changed without having to make a separate calibration for finding the correct lux value. A drawback of barrier-layer cells is their rather low sensitivity (about 0.5 μA/lux in the present case), so that considerable amplification is necessary.

The circuit is represented schematically in *fig. 15.* The dotted square encloses the equivalent electrical circuit of the photocell. This comprises a current source, a capacitance C, representing the capacitance of the barrier layer, in parallel with a resistance R_p, which is the internal "leakage resistance" of the cell. The current i_c from the source is proportional to the illumination E. Since the value of the resistance R_p is not constant — it decreases with increasing E — neither the current in an external network nor the terminal voltage is in general proportional to E (see *fig. 16*). To eliminate the influence of R_p, steps must be taken to make the terminal voltage zero. For that purpose the cell is incorporated in a compensation circuit, consisting of the current source i_b and the resistors R_1 and R_2. G is the null instrument. When the value of i_b is adjusted so that the voltage between the terminals *1* and *2* of the photocell is zero, we can write:

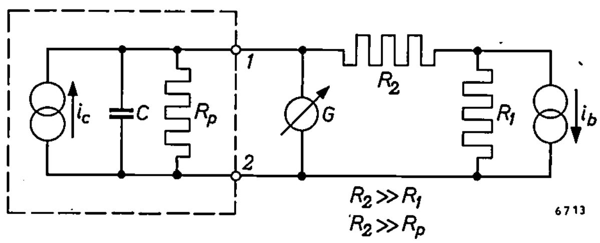


Fig. 15. Equivalent electrical circuit of the photocell (inside the dotted rectangle) and external circuit (compensation circuit). i_c photocurrent. C internal capacitance and R_p internal leak resistance of photocell. i_b compensation current. R_1 and R_2 resistors. G galvanometer. The current i_b is set to a value such that the voltage between the terminals *1* and *2* of the photocell is zero.

$$i_c = \frac{R_1}{R_1 + R_2} i_b . \quad \ldots \ldots \quad (2)$$

If we make $R_2 \gg R_1$, then $i_b \gg i_c$, and the value of i_b is high enough for easy measurement. At the same time i_b is proportional to E.

The current i_b is also very suitable as a reference signal; in the control loop, then, the compensation circuit functions at the same time as an error detector. Moreover, the current i_c is practically independent of temperature, so that there is no danger that the reference signal will no longer correspond to the desired illumination when the room temperature changes. (This does not apply at very low illuminations, owing to the fact that the highly temperature-dependent dark current then constitutes a significant fraction of the total current.)

For the purpose of calculating the transfer function K_1G_1 of the photocell and compensation circuit together, we should now consider how the output signal ε of the compensation circuit (cf. fig. 4) — i.e. the voltage appearing between the terminals of the photocell upon a change in the illumination — depends on the (small) difference ΔE between the actual illumination and that which corresponds to the preset value of i_b. We can at once put $G_1 = 1$, for at the low frequencies with which we are concerned the frequency dependence of the photocell (characteristic time R_pC) is of no consequence.

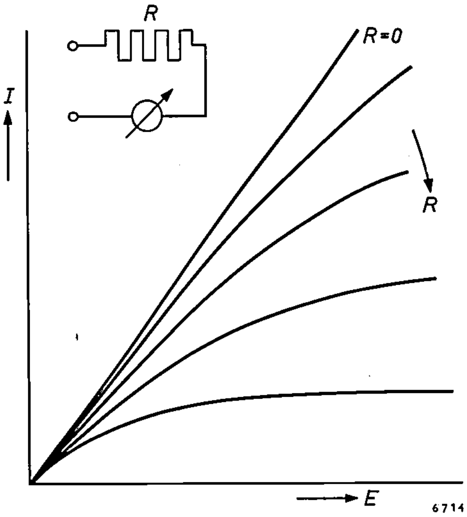


Fig. 16. The current I in the external circuit of a barrier-layer photocell (photovoltaic cell) as a function of the illumination E for various values of the resistance R in that circuit. If R is zero, I is equal to the photocurrent i_b (cf. fig. 15) and proportional to E. The deviation from linearity is greater the larger the value of R. The curve applicable to a very large R has of course virtually the same shape as the E-V characteristic (V is the open-circuit terminal voltage).

Assuming then that Δi_c is the change in i_c corresponding to ΔE, we have:

$$\varepsilon = \frac{(R_1 + R_2)R_p}{R_p + R_1 + R_2} \Delta i_c . \quad \ldots \ldots \quad (3)$$

Since R_2 is large not only compared with R_1 but also compared with R_p, the right-hand side can be approximated by $R_p\Delta i_c$. The desired transfer function $(\varepsilon/\Delta i_c)_{\omega=0}$ is therefore not constant but depends on the illumination: if E rises from 0.25 lux to 40 lux, K_1 is reduced by a factor of 2.

Referring to the V-I characteristics of the photocell, which are presented in *fig. 17*, we shall now analyse the properties of the compensation circuit (fig. 15) in somewhat more detail.

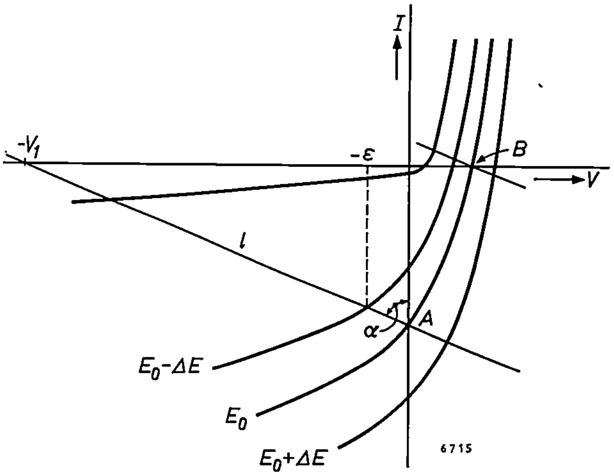


Fig. 17. Current-voltage characteristics of barrier-layer photocell in an arrangement as shown in fig. 15. V_1 is the voltage produced across the resistance R_1 by the current i_b. If i_b corresponds to the illumination E_0, the point A where the load line l intersects the relevant characteristic lies exactly on the I axis and the terminal voltage is zero. A drop in illumination by an amount ΔE gives rise to a terminal voltage $-\varepsilon$. The angle α between the load line and the I axis is determined by the value of $(R_1 + R_2)$. Using current compensation (load line through B) the ratio $\varepsilon/\Delta E$ is considerably smaller.

As can be seen, the *V-I* curves (each curve relates to one value of illumination) pass through three quadrants. Using the photocell without an external voltage source, the operating point is in the fourth quadrant (bottom right). Where, as in our case, the photocell is used in a compensation circuit with an auxiliary voltage V_1 — the voltage produced across the resistance R_1 by the current i_b — the operating point for an illumination E_0 is found at the point where the load line intersects the curve relating to the illumination concerned. The load line in this case is the straight line through the point $(-V_1, 0)$ which cuts the negative I axis at an angle $a = \tan^{-1}(R_1 + R_2)$. If E_0 and the current i_b correspond to one another, the operating point lies exactly on the negative I axis (point A).

When a small change $-\Delta E$ occurs in the illumination, the new operating point is found to be the point where the same load line intersects the characteristic for $E_0 - \Delta E$. The terminal voltage ε now appearing across the photocell is the abscissa of the new operating point. To obtain the maximum possible voltage ε for a given change ΔE in illumination, it is best to make R_2 large, in which case the angle a is large.

Finally, it can be seen from fig. 17 that, for a given slope of the load line, the ratio $\varepsilon/\Delta E$ with the compensation method employed here is appreciably larger than it would be with a method in which the current delivered by the cell were to be made equal to zero. In the latter case the operating point (for E_0) would not be at A but on the positive V axis at B. Here the curves are much closer together, and a change ΔE in the illumination therefore causes only a small change in the terminal voltage.

The amplifier A_1 and the steering motor

The voltage ε that appears at the terminals of the photocell when the illumination differs from the value corresponding to the reference signal is, as we have seen in fig. 3, considerably amplified so that it can be used to drive the steering motor. In order to minimize the influence of interfering signals, the amplifier is designed as a difference amplifier with a high rejection factor [7]. For all frequencies and amplitudes involved, the gain factor of this amplifier has the same value. The amplifier therefore contributes only one constant factor to the transfer function K_2G_2 of the part of the isocandela-diagram recorder that we are now about to discuss. We shall call that factor K_A.

In passing it may be noted that the signal ε is not supplied directly to A_1 but is first converted into a square-wave voltage by a chopper (frequency 400 c/s) to eliminate drift. The amplifier thus supplies an alternating voltage to the control motor. The latter is a two-phase asynchronous induction motor, and thus has two windings on the stator. The output signal from A_1 is supplied to the one winding. The other is supplied with a constant alternating voltage, also of 400 c/s, which differs in phase by 90° with respect to the output signal from A_1. This method of supply makes the torque of the rotor proportional to ε. For simplicity, these particulars have been omitted from fig. 3.

We shall now derive the transfer function of the steering motor from the equation describing the behaviour of the motor. The torque T supplied by the motor depends both on the supply voltage V_m and on the speed of the motor (the angular velocity Ω of the shaft). To a first approximation, then:

$$T(V_m,\Omega) = \left(\frac{\partial T}{\partial V_m}\right)_\Omega V_m + \left(\frac{\partial T}{\partial \Omega}\right)_{V_m} \Omega . \quad . \quad . \quad (4)$$

The value of the differential quotient $\partial T/\partial V_m$ for the case $\Omega = 0$ is the *torque constant* K_m. The differential quotient $\partial T/\partial \Omega$ is negative in the case of servo-motors: as the motor speed increases, the torque decreases in a manner that corresponds formally to the damping caused by a viscous liquid (opposing torque proportional to Ω). The negative value of $dT/d\Omega$ in the case $V_m = 0$ is therefore called the *damping coefficient* and is denoted by the symbol f_m. Let the total moment of inertia around the motor shaft be J_1 and the total viscous damping f_1 (both of motor and load together), then:

$$K_mV_m - f_1\Omega = J_1\, d\Omega/dt . \quad . \quad . \quad . \quad . \quad (5)$$

If we substitute the complex quantities $\overline{V}_m$ and $\overline{\Omega}$ for V_m and Ω following the practice in control-system analysis [8], equation (5) becomes:

$$K_m\overline{V}_m - f_1\overline{\Omega} = J_1\, \mathrm{j}\omega\overline{\Omega} ,$$

so that we find for the transfer function of the motor:

$$\frac{\overline{\Omega}}{\overline{V}_m} = \frac{K_m}{f_1 + J_1\mathrm{j}\omega} = \frac{K'}{1 + \mathrm{j}\omega\tau_1} , \quad . \quad . \quad . \quad . \quad (6)$$

where $\tau_1 = J_1/f_1$ and $K' = K_m/f_1$.

In arriving at this formula, however, we have not yet entirely reached our objective. In order to be able to adjust the damping to a suitable value, the motor shaft is coupled to the shaft of a tacho-generator, which generates a voltage proportional to the speed. This output voltage, amplified if necessary, is fed back to the terminals of the motor. This produces an opposing torque which is proportional to the speed, so that here too we can speak of viscous damping. The transfer function of the tacho-generator is simply equal to the constant ratio K_t between the (amplified) output voltage and Ω (*fig. 18*). The transfer function of the motor and tacho-generator combined is therefore:

$$\frac{\overline{\Omega}}{\overline{V}_m} = \frac{K'/(1 + \mathrm{j}\omega\tau_1)}{1 + K'K_t/(1 + \mathrm{j}\omega\tau_1)} . \quad . \quad . \quad . \quad (7)$$

After manipulation and introduction of the factor K_A, the transfer function of the amplifier A_1, this expression reduces to:

$$K_2G_2(\mathrm{j}\omega) = \frac{K_2}{1 + \mathrm{j}\omega\tau_2} , \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where $\tau_2 = \tau_1/(1 + K'K_t)$ and $K_2 = K_AK'/(1 + K'K_t) \approx 1/K_t$. The characteristics of elements having a transfer function of the form $K/(1 + \mathrm{j}\omega\tau)$ were discussed at some length in the articles quoted above [4].

The drive of the drum

We shall now consider how the position of the steering rod varies with respect to the surface of the drum when the rod is

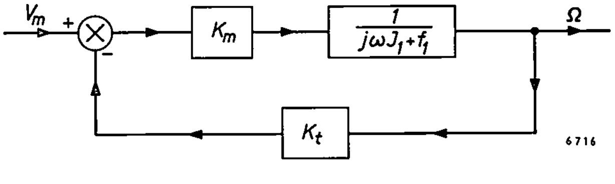


Fig. 18. Block diagram of steering motor with tacho-generator feedback circuit. $\overline{V}_m$ motor supply voltage. Ω angular velocity of motor shaft. K_m torque constant of motor. f_1 total viscous damping. J_1 moment of inertia of motor and load together. K_t ratio between Ω and the voltage delivered by the tacho-generator.

[7] See e.g. G. Klein and J. J. Zaalberg van Zelst, General considerations on difference amplifiers, Philips tech. Rev. **22**, 345-351, 1960/61 (No. 11).

[8] See e.g. M. van Tol, Philips tech. Rev. **23**, 1961/62, No. 4, pages 109 and 112.

turned through an angle Θ_s (for simplicity we can just as well treat the drum as the flat table in fig. 3). We shall characterize this position as the distance x from the rod to the original line of travel. This distance consists of two components. In the first place there is a component x_1, the time derivative of which is proportional to the peripheral velocity v of the wheel rolling over the surface and to the sine of the angle Θ_s between the old and the new line of travel. For small values of Θ_s we can thus write:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = v\Theta_s. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

The second component arises from the above-mentioned fact that the wheel is in contact with the drum not in line with the steering rod but at a distance a behind it. A displacement Θ_s of the steering rod causes in the x direction a displacement x_2 of magnitude $a \sin \Theta_s$ (see *fig. 19*). For small Θ_s we therefore have $x_2 = a\Theta_s$. The displacement $a(1 - \cos\Theta_s)$ in the direction perpendicular to x can be neglected for small values of Θ_s.

The equation that gives the relation between $x (= x_1 + x_2)$ and Θ_s is thus

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v\Theta_s + a\,\frac{\mathrm{d}\Theta_s}{\mathrm{d}t}. \quad . \quad . \quad . \quad . \quad . \quad (10)$$

Replacing x and Θ_s by the complex quantities [8] $\overline{x}$ and $\overline{\Theta}_s$, we find:

$$\mathrm{j}\omega\,\overline{x} = v\overline{\Theta}_s + a\,\mathrm{j}\omega\,\overline{\Theta}_s\,, \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

so that

$$\frac{\overline{x}}{\overline{\Theta}_s} = \frac{v + \mathrm{j}\omega a}{\mathrm{j}\omega} = v\tau_0\,\frac{1 + \mathrm{j}\omega\tau_3}{\mathrm{j}\omega\tau_0}, \quad . \quad . \quad . \quad . \quad (12)$$

where $\tau_3 = a/v$ (cf. p. 281) and τ_0 is again the unit of time.

We have still not found, however, the transfer function K_3G_3 of the relevant part of the recorder. We have to relate x not to the angular position of the steering rod but to the angular velocity Ω of the shaft of the steering motor (see eq. 7). This means first of all that we have to replace Θ_s by $\mathrm{d}\Theta_s/\mathrm{d}t$, and thus the quantity $\overline{\Theta}_s$ in (12) by $\mathrm{j}\omega\overline{\Theta}_s$, and further that we must take into account the gear transmission ratio n_1 between the two shafts. We then find:

$$K_3G_3(\mathrm{j}\omega) = \frac{v\tau_0^2}{n_1}\,\frac{1 + \mathrm{j}\omega\tau_3}{(\mathrm{j}\omega\tau_0)^2}. \quad . \quad . \quad . \quad . \quad (13)$$

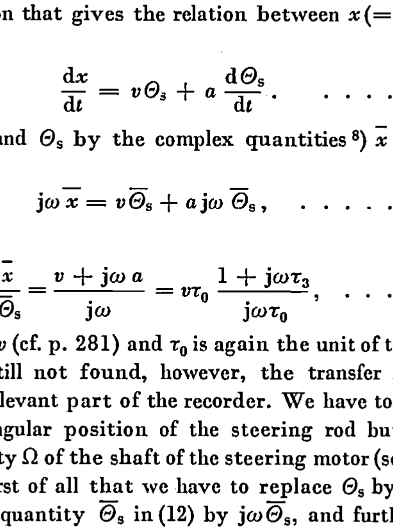


Fig. 19. The point of contact of the drive wheel L on the drum is not in line with the steering rod S but at a small distance a behind the axis of the rod. Consequently, when S is turned through the angle Θ_s the distance between the rod and the old line of travel l_1 consists not only of a component x_1 increasing proportionally with time but also of a constant component x_2 of magnitude $a \sin \Theta_s$.

Fig. 20 shows the relevant Bode diagram, approximated by straight lines. The diagram consists of two parts of slope —2 and —1. The break lies at the frequency $\omega_3 = 1/\tau_3$. The phase shift at this frequency is —135°. Since $\omega_3 = v/a$, we can therefore shift the break towards a lower frequency by increasing the distance a or by reducing the peripheral velocity v of the wheel. It can be derived from (12) that, if v is varied, the break moves along the line having the slope —1, and if a is varied the break moves along the line with the slope —2.

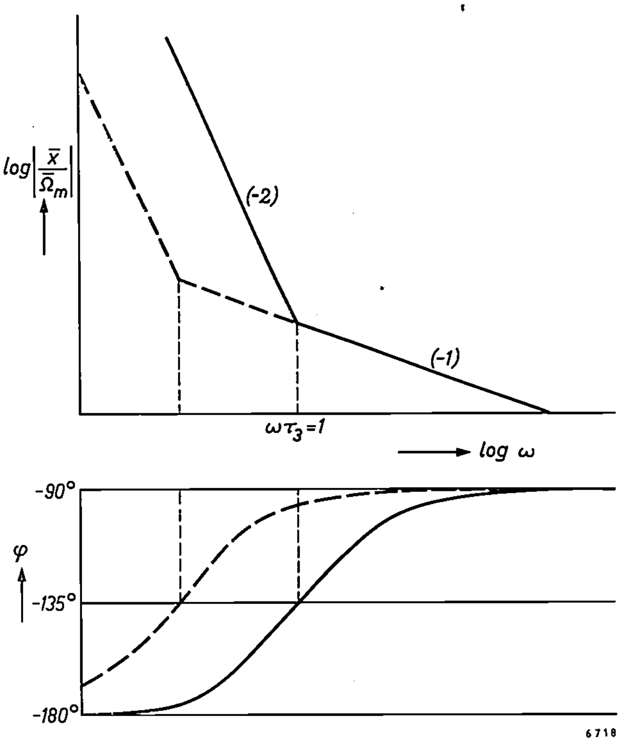


Fig. 20. Bode diagram of the transfer function that describes the relation between the angular velocity $\overline{\Omega}$ of the steering-motor shaft and the change $\overline{x}$ in the position of the steering rod with respect to the drum (eq. 13). The amplitude characteristic can be approximated by two straight lines, of slope —2 and —1, which intersect at the frequency $\omega_3 = 1/\tau_3$ ($=v/a$). The dashed lines relate to a smaller value of v. As can be seen, when v is varied the break in the curve shifts along a line of slope —1.

The positional servo-systems

The most complicated part of the isocandela-diagram recorder from the control-engineering point of view is that formed by the two positional servo-systems, which transmit the movements of the drum to the lamp. Since the systems are virtually identical, it will be sufficient to discuss only one of them, and for deriving the transfer function of the whole instrument it will be permissible to assume that there is only one positional servo-system.

A block diagram of one of the systems is given in *fig. 21*. It can be seen that, apart from the feedback from output to input, which makes it possible for the output signal to follow the input signal faithfully, there is also an inner loop. The latter relates to the servo-motor (*SM* in fig. 3) which, like the steering motor, is provided with a tacho-generator feedback circuit (cf. fig. 18). In this case, however, the output voltage from the tacho-generator is not directly returned to the motor but first passes through a high-pass filter. The transfer function of this filter is $\mathrm{j}\omega\tau_4/(1 + \mathrm{j}\omega\tau_4)$, where τ_4 is equal to the product

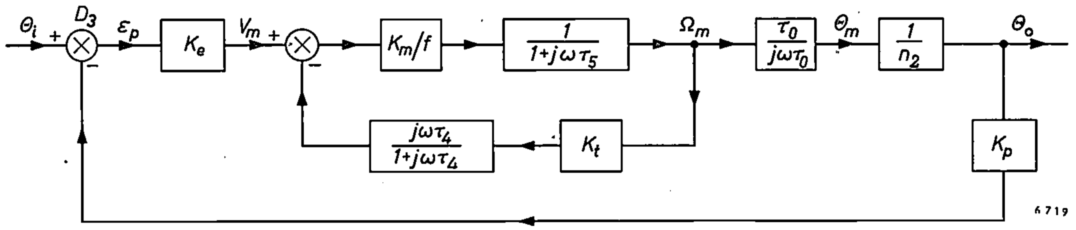


Fig. 21. Block diagram of each of the servo-mechanisms which transmit the positional changes of the drum to the lamp (cf. fig. 3). Θ_i electric input signal corresponding to the relevant coordinate of the drum. Θ_o output signal, one of the coordinates of the lamp. K_p transfer function of the position pick-off P_2. D_3 difference circuit. ε_p difference signal. K_e gain factor of amplifier A_2. V_m servo-motor supply voltage. Apart from their subscripts, K_m, f, J, K_t and Ω_m have the same meaning as in fig. 18, but now relate to the servo-motor. For τ_5 see eq. (14). n_2 ratio of the rotation of the servo-motor shaft to the rotation of the lamp. τ_4 time constant of high-pass filter in the inner control loop (tacho-generator feedback).

of the resistance and capacitance which are the main components of the circuit concerned. The reason for the presence of this filter will be made clear after we have arrived at the transfer function of the complete open loop of fig. 21.

In the same way as we found the transfer function of the closed steering-motor loop (eq. 7) from that of the open loop (eq. 6), we find for the transfer function $\overline{\Omega}_m/\overline{V}_m$ of the closed inner loop:

$$\frac{\overline{\Omega}_m}{\overline{V}_m} = \frac{K_m}{f}\,\frac{1 + j\omega\tau_4}{1 + j\omega(\tau_4 + \tau_5 + K''\tau_4) - \omega^2\tau_4\tau_5}, \qquad (14)$$

where $\tau_5 = J/f$ and K'' is the loop gain K_mK_t/f of the inner loop.

The open-loop transfer function of the whole positional servo-system is thus:

$$\frac{K_p\overline{\Theta}_o}{\overline{\varepsilon}_p} = \frac{K_pK_mK_e\tau_0}{n_2 f}\,\frac{1 + j\omega\tau_4}{j\omega\tau_0\{1 + j\omega(\tau_4 + \tau_5 + K''\tau_4) - \omega^2\tau_4\tau_5\}}. \qquad \ldots \ (15)$$

Since K'' has been chosen to be large, in the discriminant of the quadratic form occurring in the denominator of (15) we have:

$$(\tau_4 + \tau_5 + K''\tau_4)^2 \gg 4\tau_4\tau_5\,.$$

This form can therefore be resolved to a good approximation into the factors $(1 + j\omega K''\tau_4)$ and $(1 + j\omega\tau_5/K'')$, so that:

$$\frac{K_p\overline{\Theta}_o}{\overline{\varepsilon}_p} = K'''\tau_0\,\frac{1 + j\omega\tau_4}{j\omega\tau_0(1 + j\omega K''\tau_4)\,(1 + j\omega\tau_5/K'')}, \qquad (16)$$

where $K''' = K_pK_mK_e/n_2f$. The Bode diagram, in a linear approximation, is given in *fig. 22*. The slope of successive sections is —1, —2, —1 and again —2. The breaks lie at the frequencies at which $\omega K''\tau_4 = 1$, $\omega\tau_4 = 1$ and $\omega\tau_5 = K''$. Choosing the gain so that it is equal to unity at a frequency ω_0, a value between the two last-mentioned breaks, the phase margin is then $\geqq 45°$, which means that the closed loop is sufficiently stable.

The significance of the above-mentioned filter in the tacho-generator feedback circuit is now clearly apparent from fig. 22. Without that filter the Bode diagram would have the form shown by the dashed lines (together with the full lines joining them). The filter evidently leaves the characteristics unchanged at high frequencies, but at low frequencies it considerably increases the gain. The reason for this may be inferred from the block diagram: for high frequencies the filter may be regarded as not present, i.e. as a closed switch, and for low frequencies as an open switch. In the latter case there is therefore no feedback and the function $\overline{\Omega}_m/\overline{V}_m$ is obviously larger.

The inclusion of the high-pass filter in the inner loop thus has the effect of increasing the value of the open-loop transfer function $K_p\overline{\Theta}_o/\overline{\varepsilon}_p$ for the low frequencies and of correspondingly reducing the velocity error of the systems [9]) without enlarging the bandwidth, i.e. without increasing the influence of disturbances.

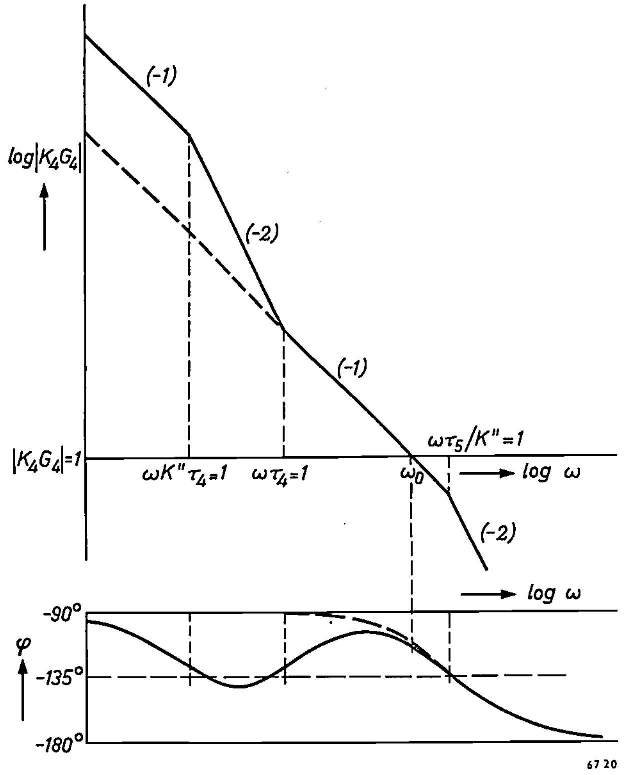


Fig. 22. Bode diagram of each of the positional servo-systems. The dashed lines would apply if the high-pass filter were omitted from the inner loop (cf. fig. 21). With the filter the gain at low frequencies is greater for the same bandwidth, and hence the velocity error smaller.

[9]) Just as in the case of a proportional controller (cf. p. 110 of the first and p. 151 of the second article cited under [4])) the static error (offset) decreases as $KG(0)$ increases, in the case of the servo-system as considered here the steady-state velocity error is smaller the higher is the gain. This is an example of one of the methods for improving a control system, discussed in the second article mentioned in footnote [4]). It can be shown that the high-pass filter in the feedback circuit has the same effect as an integrating element.

Taking into account that $K''\tau_4 \gg \tau_5/K''$ and $K'''\tau_4 \gg 1$, we can write the formula derived from (16) for the closed-loop transfer function K_4G_4 of the positional servo-system as

$$K_4G_4 = \frac{\overline{\Theta}_o}{\overline{\Theta}_r} = \text{const.}\ \frac{1 + j\omega\tau_4}{1 + j\omega\tau_4 - \omega^2\tau_4 K''/K''' - j\omega^3\tau_4\tau_5/K'''}. \quad \ldots \quad (17)$$

If we now choose the gain not merely so that ω_0 lies between the breaks in the curve at the frequencies for which $\omega\tau_4 = 1$ and $\omega\tau_5 = K''$, but moreover so that ω_0 lies close to the latter frequency, and if we have satisfied the requirement that $\omega_0\tau_4 \gg 1$, the denominator of (17) can be represented approximately as the product of the factors: $[1 + j\omega\tau_4(1 - K''/K'''\tau_4)]$ and $[1 + j\omega K''/K''' - (1 + K''/K'''\tau_4)\tau_5\omega^2/K''']$. We can then simplify (17) to:

$$K_4G_4 = \text{const.}\ \frac{1 + j\omega\tau_4}{[1 + j\omega\tau_4\ \tau_5/K'''\tau_6^2]\ [1 + 2\ \zeta\ j\omega\tau_6 - \omega^2\tau_6^2]}, \quad \ldots \quad (18)$$

where $K'''\tau_6^2 = \tau_5(1 + K''/K'''\tau_4)$ and $2\zeta = K''/K'''\tau_6$, and furthermore $\tau_6 \approx \omega_0^{-1}$.

From the assumptions $\tau_6 \approx \tau_5/K''$ and $\tau_4 \gg \tau_6$ it follows that $\tau_5/K'''\tau_6^2 \approx 1$, so that in our case, at least as far as stability is concerned, the first factor of the denominator can be cancelled against the numerator. The formula for G_4 then becomes:

$$G_4 = \frac{1}{1 + 2\zeta\ j\omega\tau_6 - \omega^2\tau_6^2}. \quad \ldots \quad (19)$$

(In practice the case with the two complex conjugate roots is preferred to the other because then the system is not aperiodically damped but exhibits oscillation. If the overshoot can be kept within bounds, a faster response can be obtained.)

The beam pattern of the lamp

The transfer function (1) of the whole recorder contains the factor K_E, the gradient of the illumination at the point in the patch of light on the screen where the photocell is situated. The reason for this is readily apparent. The signal delivered by the photocell upon a slight shift $\Delta\Theta_o$ (eq. 17) in the angular position of the lamp is proportional to the change ΔE of the illumination, i.e. proportional to the product of the positional change and the gradient K_E. The ratio $\Delta E/\Delta\Theta_o$, the only transfer function not yet obtained, is thus equal to K_E. In fig. 7 we saw how the value of K_E may fluctuate along one isolux contour.

The transfer function of the whole loop

Now that we have found the transfer functions of all the elements of the isocandela-diagram recorder, we can write the transfer function KG of the whole control loop. The factors that are not dependent on ω will not, in so far as they are constant, be explicitly mentioned, but two other factors will be, namely K_E and R_p (the latter being the internal leakage resistance of the photocell). We thus find:

$$KG = \text{const.}\ K_E R_p\ \frac{1}{1 + j\omega\tau_2}\ \frac{1 + j\omega\tau_3}{(j\omega\tau_0)^2}\ \frac{1}{1 + 2\zeta j\omega\tau_6 - \omega^2\tau_6^2}. \quad \ldots \quad (20)$$

In this expression, apart from the quantities already mentioned,

τ_2 is the time constant of the control motor (with tacho-generator feedback): $\tau_2 = J_m/(f_1 + K_mK_t)$,

τ_3 the quotient a/v of the distance between axis and point of contact and the peripheral velocity of the wheel,

τ_6 the reciprocal of the cross-over frequency ω_0, above which the gain of the positional servo-system falls below unity.

As stated, for considerations of stability the complete formula for KG can be considerably simplified. The reduction of (17) to (19) was already a substantial simplification, but a further one is possible. The time constant τ_2 is so small that the frequency ω_2 $(= 1/\tau_2)$ is appreciably larger than ω_0. Its effect on the part of the Bode diagram of interest from the point of view of stability can therefore be disregarded. (Since τ_2 is very small its contribution to a transient also dies out very rapidly.)

The simplified form of (20), which is sufficient for a stability analysis and has already been encountered as equation (1), is thus:

$$KG \approx \text{const.}\ K_E R_p\ \frac{1 + j\omega\tau_3}{(j\omega\tau_0)^2\{1 + 2\zeta j\omega\tau_6 - \omega^2\tau_6^2\}}.$$

The stability considerations themselves were dealt with when the latter formula was first presented.

Summary. The luminous-intensity pattern (an array of isocandela curves) of a beamed light-source is determined in principle by finding curves of constant luminous intensity (isolux contours) on a screen set up in front of the lamp. Such a curve joins all the points corresponding to those directions in which the luminous intensity is the same. An instrument is described which traces the isolux curves automatically, reduced in scale by about 30×. The light detector (a barrier-layer photocell) remains in a fixed position while the angular setting of the light source is varied. The entire instrument can be regarded as a closed control loop. The driving element is a single-wheel "trolley", which causes the drum to which the recording paper is fixed to turn about its axis and propels a stylus in the direction of that axis, the position of the stylus on the paper being transmitted by servo-systems to the lamp. The difference between the photocell signal and a reference signal is used, after amplification, to actuate the motor which steers the trolley. The feedback which, via the steering motor, the trolley and the drum, exists between the photocell and the lamp automatically ensures that the lamp, and therefore the stylus on the paper, can only take up those positions that correspond to the preset value of luminous intensity. The stylus thus traces the isolux curve for that value.

VOLUME 23, 1961/62, No. 10 pp. 293-324 Published 25th. July 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

STANDARD NOISE SOURCES

by P. A. H. HART *).

621.391.822.3:621.385.2

A standard noise source can be a resistor, a saturated diode or a gas discharge. These three types are dealt with in the article below, and it is shown which type is to be preferred in the various frequency ranges. Some standard noise sources specially designed in the Philips Laboratories are discussed.

Introduction

If signals are to be made perceptible — we are concerned here primarily with radio and radar signals — the signal strength must exceed a certain minimum, irrespective of the amplification of the receiving system. The reason for this is the presence of noise. The minimum referred to depends on the particular technique of "information processing" used, and is low in some special techniques such as single-sideband systems and the methods used in radio astronomy.

The noise comes from sources that can be divided into two categories. The first comprises the *external* sources. It is these that cause the receiving aerial to pick up noise in addition to the desired signal. External kinds of noise include atmospherics, thermal noise from the earth and cosmic noise.

The second category comprises the *internal* sources of noise, inside the receiver itself. Their contribution makes the signal-to-noise ratio at the output of the receiver worse than at the input. The noise added by internal sources can be minimized by careful circuitry and the suitable choice of components, but it cannot be entirely eliminated; some noise from resistors, valves and other circuit elements always remains.

The strength of the internal noise is usually measured in a relatively narrow band of frequencies; the average frequency of this band is called "the" frequency at which the noise is measured. The measurement can be made by comparison with a known noise power delivered by a *standard noise source.* There are various types of standard noise source. Which type is used depends among other things on the frequency at which the measurement is to be made.

A standard noise source that delivers an accurately known noise power is a *resistor of known resistance and temperature.* Requiring no calibration, this noise source is an *absolute* standard.

It is often more convenient to use a *noise diode,* i.e. a diode operated at the saturation current. Because of the shot effect the current fluctuates. A noise diode is a *noise-current generator.* The value of the noise current can be calculated from the dircct current flowing in the diode; the noise diode too is therefore an absolute standard, but only in a limited (though wide) range of frequencies. As will presently be shown, this range has both a lower and an upper limit.

At frequencies higher than those at which the noise diode is effective, use can be made of a *gas-discharge noise source.* Noise generators of this type are sub-standards, in the sense that the noise power delivered cannot be exactly calculated but must be determined by calibration. They can be made with highly stable characteristics and are relatively insensitive to fluctuations in supply voltage and ambient temperature. For these reasons it is not necessary to calibrate them individually, unless

*) Philips Research Laboratories, Eindhoven.

extreme precision is required, as for instance in radio-astronomic measurements. Compared with a resistor, gas-discharge noise sources deliver a high noise power. Various designs are possible. As we shall see, in decimetre-wave equipment the gas discharge is coupled to a helix or a Lecher line, and in equipment operating on centimetre or millimetre wavelengths the discharge tube is mounted in a waveguide. In other designs the gas discharge is in a resonant cavity or in a horn antenna.

Fig. 1 gives a broad indication of the operating ranges of the Philips noise diodes K 81A and 10 P, and of gas-discharge tubes of varying constructions; the boundaries are in fact not as sharp as are shown here.

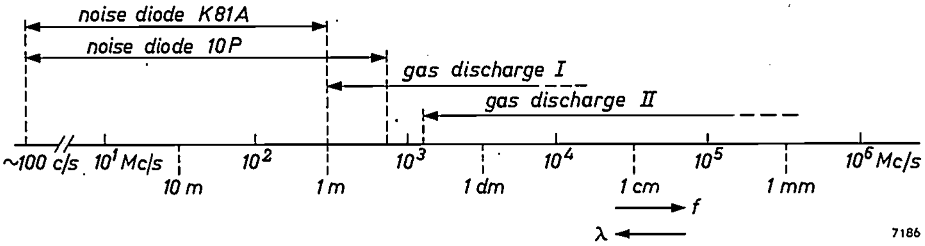


Fig. 1. Rough indication of the frequency and wavelength ranges in which noise diodes and gas discharges can be used as standard noise sources. *Gas discharge I* relates to a discharge tube coupled to a helix (fig. 17*a*), *gas discharge II* to a discharge tube mounted in a waveguide (fig. 17*b* and *c*).

In this article the three main types of noise source — resistor, diode and gas-discharge — will be discussed. We shall consider some special designs, the corrections necessary in certain applications of the noise diode, and gas-discharge noise sources for various wave ranges, including the millimetre band.

The resistor as a noise source

The thermal noise of a resistance R in a relatively narrow frequency band Δf around the frequency f is given by Nyquist's noise theorem:

$$\overline{u^2} = \frac{4\,hf}{e^{hf/kT}-1}\,R\,\Delta f\,. \quad . \; . \; . \quad (1)$$

Here $\overline{u^2}$ is the mean square noise voltage, k is Boltzmann's constant ($= 1.38\times10^{-23}$ joule/°K), T is the absolute temperature of the resistor, and h is Planck's constant ($= 6.6\times10^{-34}$ joule-second). If hf is small compared with kT, i.e. if

$$f \ll \frac{k}{h}\,T \approx 2\times10^{10}\,T \text{ c/s}\,,$$

formula (1) can be simplified to

$$\overline{u^2} = 4\,k\,T\,R\,\Delta f\,. \quad . \; . \; . \; . \quad (2)$$

Measurement of the noise factor with standard noise resistors

An example of the use of resistors as standard noise sources is the measurement of the noise factor of a circuit, an amplifier for instance, that can be treated as a linear four-terminal network.

The noise factor F as defined by the American standard (53 I.R.E. 7 S1) is [1]:

$$F = \frac{N_0 + N_{\text{extra}}}{N_0}\,. \quad . \; . \; . \quad (3)$$

Here $N_0 + N_{\text{extra}}$ is the total noise power at the output in the narrow frequency band Δf, and N_0 is the share contributed by the thermal noise of an impedance Z_i which is connected externally across the input terminals and is equal to the impedance of the signal source to which the four-terminal network is normally connected (e.g. an antenna); the temperature of Z_i must be 290 °K. Therefore N_{extra} is the portion of the output noise added by the four-terminal network [2].

The noise factor is determined by measuring the output noise power (still in the narrow band Δf) when Z_i is successively at the temperatures T_1 and T_2, which must be known; see *fig. 2*. If only the temperature of Z_i were to change between the two measurements (Z_i itself thus remaining constant) the output noise power when Z_i has the temperature T_1 would be:

$$P_1 = CT_1 + N_{\text{extra}}\,,$$

and when Z_i has the temperature T_2:

$$P_2 = CT_2 + N_{\text{extra}}\,.$$

[1] See also F. L. H. M. Stumpers and N. van Hurck, An automatic noise figure indicator, Philips tech. Rev. **18**, 141-144, 1956/57.

[2] It would be going too far to deal here with the way in which the noise factor as defined depends on Z_i. In this connection reference may be made to A. G. T. Becking, H. Groendijk and K. S. Knol, The noise factor of four-terminal networks, Philips Res. Repts **10**, 349-357, 1955.

In these two expressions C is a constant. Let the temperature 290 °K, mentioned in the definition, be denoted by T_0; then $N_0 = CT_0$. If we put $P_2/P_1 = a$, it follows from (3) that:

$$F = \frac{\frac{T_2}{T_0} + a - 1 - a\frac{T_1}{T_0}}{a-1}. \quad \quad (4)$$

For $T_1 = T_0$ this reduces to

$$F = \frac{\frac{T_2}{T_0} - 1}{a-1}. \quad \quad (5)$$

The latter expression is still valid to a good approximation when the difference between T_1 and T_0 is small.

In accordance with the definition of the noise factor we have spoken above of an *impedance* Z_i. Every impedance can be treated as composed of a resistance and a reactance connected in parallel (or in series). As the reactance causes no noise and is not changed during the measurement, it can be regarded as belonging to the four-terminal network. The noise-factor measurement therefore amounts to determining the ratio a of the output noise powers when a *resistor* R of temperature T_1 ($\approx T_0$) and T_2 respectively is connected across the input.

Measurements using a resistor as standard noise source are in principle possible at any frequency. In the range of ultra-high frequencies (decimetre wavelengths and shorter) a resistor is in fact the only absolute standard noise source. Nevertheless, in all frequency ranges — with the sole exception of the audio frequencies [3]) — the resistor has been superseded as a noise source for routine measurements by the noise diode or the gas discharge. The reasons are of a practical nature:

1) To bring a resistor to the temperature T_2, an oven or a temperature bath is needed, which is a complication.
2) An error is caused by the fact that the resistance changes as a rule with temperature. Two resistors are therefore needed, one of which must have the same resistance at the temperature T_1 as the other at the temperature T_2, which is much higher or lower than T_1.
3) If T_2 is higher than T_1, the temperature difference T_2—T_1 cannot be made very large without damaging the hot resistor. The highest temperature can be achieved with a tungsten filament ($T_2 = 2700$ °K); $T_2 - T_1$ is then about 2400 °K. This means that large noise factors cannot be accurately measured with a hot resistor: N_{extra} is then large compared with CT_1 and CT_2, and therefore $P_2 \approx P_1$, i.e. $a \approx 1$, so that $a - 1$ cannot be determined with great precision.

Apart from its use at audio frequencies, the resistor as a noise source is now mainly used for *calibrating* noise diodes in the decimetre bands and gas-discharge tubes in the centimetre and millimetre bands. An example will be given at the end of this article.

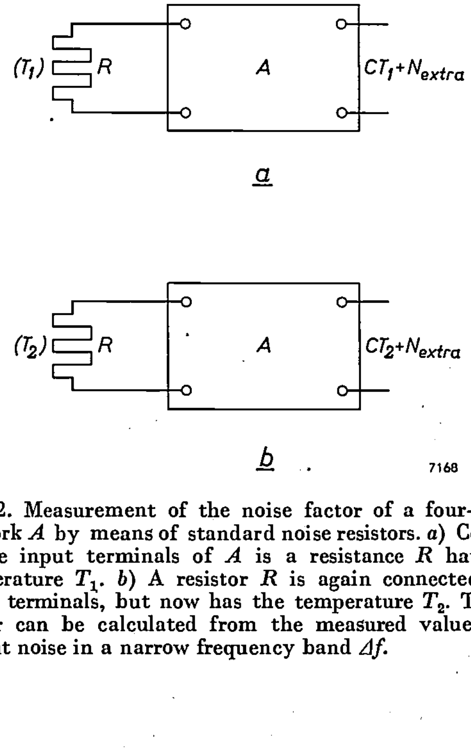


Fig. 2. Measurement of the noise factor of a four-terminal network A by means of standard noise resistors. *a*) Connected to the input terminals of A is a resistance R having the temperature T_1. *b*) A resistor R is again connected to the input terminals, but now has the temperature T_2. The noise factor can be calculated from the measured values of the output noise in a narrow frequency band Δf.

Special designs of resistors as standard noise sources

In noise measurements at very short waves the stray inductance and capacitance of conventional resistors cause an impermissible error. Special resistors are therefore needed, and we shall describe here two that have been designed and constructed in the Philips Research Laboratories at Eindhoven. One serves for calibrating noise diodes in the decimetre bands and is a *cold* resistor ($T_2 = 77$ °K)[4]); the other is used for calibrating gas discharges in the centimetre and millimetre bands, and is a *hot* resistor ($T_2 = 1336$ °K).

a) A cold resistor for decimetre waves

The resistor proper (*fig. 3*) consists of a very thin layer of platinum on a hard-glass tube, located at the end of an impedance transformer fitted with shorting plungers. The layer has a resistance of roughly 50 ohm. The resistance, as "seen" from the input of the impedance transformer, is accurately set to 50 ohm by adjusting the plungers until a standing-wave detector, having a 50-ohm characteristic

[3]) See e.g. A. van der Ziel, Noise, Prentice Hall, New York 1954, p. 31.

[4]) Another cold resistor is described in Philips tech. Rev. **21**, 327 (fig. 13), 1959/60.

impedance, gives a standing-wave ratio of 1. Both the resistor itself and the impedance transformer are immersed in liquid nitrogen (temperature 77 °K). As regards its noise contribution, therefore, the dissipative resistance of the transformer also behaves as a resistor having a temperature of 77 °K. *Fig. 4* shows a photograph of the transformer with the resistor inside.

The second resistor in the measurement can be a conventional type of 50 ohm, kept at room temperature.

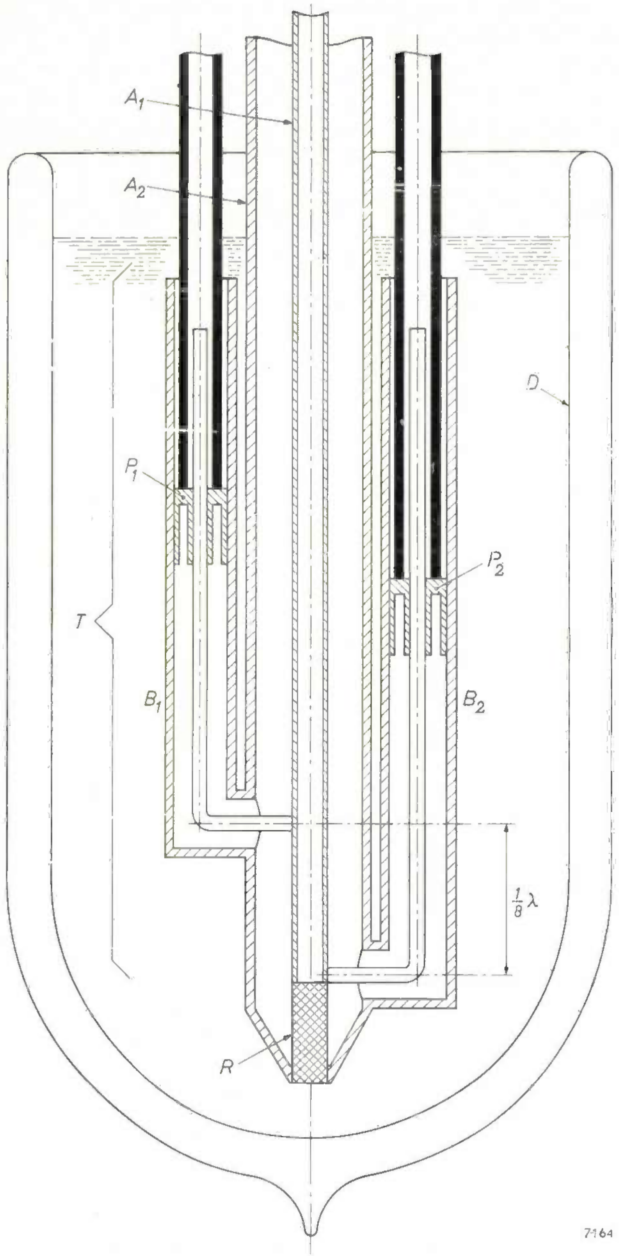


Fig. 3. Construction of a cold standard noise resistor for decimetre waves (not to scale). R is the actual resistor (approx. 50 ohm), being a layer of platinum on glass. T coaxial impedance transformer consisting of an inner conductor A_1 and an outer conductor A_2 with coaxial side branches B_1 and B_2, which are terminated by shorting plungers P_1 and P_2. At the top a coaxial plug (N plug) can be connected. D Dewar vessel filled with liquid nitrogen (temperature $T_2 = 77$ °K).

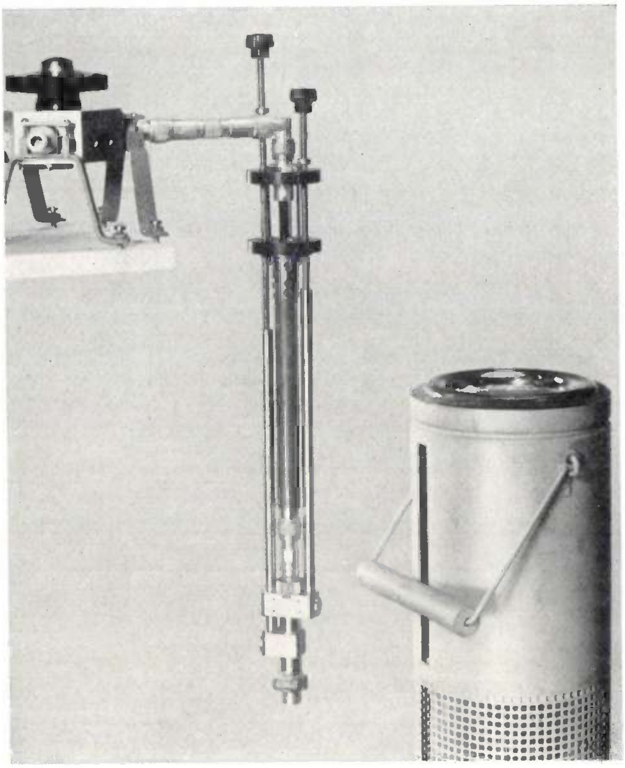


Fig. 4. The impedance transformer represented in fig. 3 (with the noise resistor inside) removed from the nitrogen bath.

b) A hot "resistor" for centimetre and millimetre waves

A waveguide (*fig. 5*) is provided with a matched termination at one end in the form of an absorption wedge, made in this case of the ceramic material "Caslode". The part of the waveguide with the wedge is uniformly heated in an oven to a well-defined high temperature. The temperature is measured with a thermocouple or a pyrometer, which need only be calibrated for this one temperature — which has been chosen as the melting point of gold (1336 °K) [5].

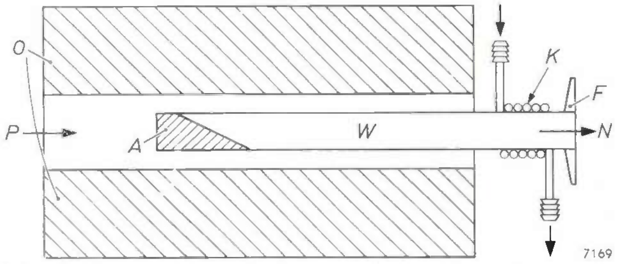


Fig. 5. Cross-section of a hot standard noise resistance for centimetre or millimetre waves. W platinum waveguide. A absorption wedge (the noise source proper). F connecting flange. O electric oven. K pipe carrying cooling water. Arrow N indicates the direction in which the noise leaves the waveguide, arrow P the direction in which the optical pyrometer faces; the latter measures the temperature T_2 of the wedge A.

[5] The method of calibration has been described in: K. S. Knol, A thermal noise standard for microwaves, Philips Res. Repts 12, 123-126, 1957.

To prevent oxidation, the waveguide is made of (thin) platinum. The part outside the oven is water-cooled to avoid overheating the coupling flange and the components connected to it. *Fig. 6* shows waveguides of this type for wavelengths of 3 cm and 8 mm.

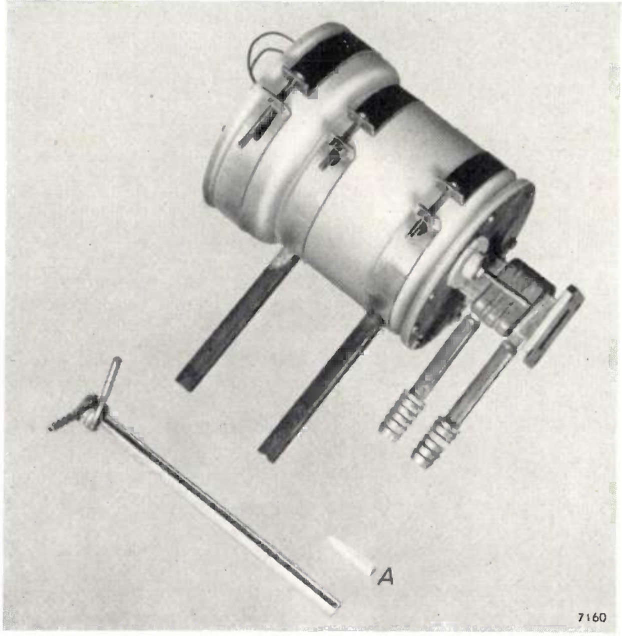


Fig. 6. Hot standard noise sources as in fig. 5, for wavelengths of 3 cm (with oven) and 8 mm (without oven). At *A* can be seen the "Caslode" absorption wedge for the 8 mm waveguide.

The second waveguide used in the measurement contains a matched termination made of wood, at room temperature.

Before concluding this account of standard noise resistors, there are two further points to be noted. One concerns the choice between cold and hot resistors, the other a correction required in certain cases.

We have just discussed a cold resistor for decimetre waves and a hot resistor for centimetre and millimetre waves. In principle the converse is also possible. Compared with a cold resistor a hot resistor has the advantage of delivering a higher noise power, which can increase the accuracy of the measurement (unless the noise figure is low). This is an argument in favour of choosing a hot resistor. Decimetre-wave techniques using Lecher wires or coaxial lines are not so suitable at high temperatures, however, as microwave techniques using waveguides. For this technological reason we decided on a cold resistor for the decimetre bands.

The standard noise resistor which is at the high or low temperature T_2 gives rise to a temperature gradient in the supply line (coaxial cable or waveguide). This must be taken into account in two respects. Firstly, the materials employed must be capable of withstanding the temperature gradient. Secondly, the calculated value of the noise generated by the resistor R needs to be corrected if the losses in the supply line are at all significant (e.g. a few percents of those in R), for part of the noise produced by R is lost in the dissipative resistance of the supply line, whilst the latter resistance itself contributes a certain amount of noise.

The diode as a noise source

A diode operated at saturation behaves in a wide range of frequencies like a noise-current generator having an infinite internal resistance. In a relatively narrow band Δf within that range the mean square noise current is given by Schottky's formula:

$$\overline{i^2} = 2\, qI_s\Delta f\,, \qquad \qquad (6)$$

where q is the charge on the electron ($= 1.60 \times 10^{-19}$ C) and I_s is the saturation current flowing in the diode. I_s depends on the temperature of the filament and therefore the noise current can be given a different value by changing the filament current.

The usual practice for measurements is to connect the noise diode in parallel with a resistor R (*fig. 7a*). In fig. 7*b* the diode is represented by a noise-current source I, and the resistance R by an equal but hypothetically noiseless resistance R^* in series with a noise-voltage source U which accounts for the thermal noise of R. In addition to this thermal noise, given by $4kT_1R\Delta f$ (T_1 being the temperature of R), the resistance R in fig. 7*a* carries the noise voltage generated by the noise current of the diode; this noise is given by $2qI_sR\Delta f$. Since the two noise sources are independent of each other, the total noise is found by simply adding the two contributions mentioned. We now want to find the temperature T_{eq} which a resistor R must have if its thermal noise is to be equal to this total:

$$4\, kT_{\mathrm{eq}}R\Delta f = 4kT_1R\Delta f + 2qI_sR^2\Delta f\,.$$

Fig. 7. *a*) Noise diode D in parallel with resistor R. *b*) Equivalent circuit of (*a*) consisting of a noise-current source I, representing the diode noise, shunted across a noiseless resistor R^* which is in series with the noise-voltage source U representing the noise from R. *c*) Equivalent circuit of (*b*) — likewise of (*a*) — consisting of a noiseless resistor R^* in series with a noise-voltage source U_{eq} which delivers the same noise as a resistor R at the temperature T_{eq}.

We can solve this expression for T_{eq}:

$$T_{\mathrm{eq}} = \frac{q}{2k} I_s R + T_1 .$$

Insertion of the numerical values of q and k gives $q/2k = 5800$ °K/V $= 20\times290$ °K/V $= 20\ T_0$ °K/V, so that the formula for T_{eq} can also be written:

$$T_{\mathrm{eq}} = 20\ I_s R T_0 + T_1 . \quad . \quad . \quad . \quad . \quad (7)$$

A noise diode in parallel with a resistor R (fig. 7*a*) thus constitutes a noise source which is equivalent to a resistor R at the temperature T_{eq} (represented in fig. 7*c* as a noiseless resistor R^* in series with a noise-voltage source U_{eq}); as can be seen from eq. (7), T_{eq} can be varied by changing the current I_s by means of the filament current of the diode.

Determining the noise factor with a noise diode

The noise factor of a four-terminal network can be determined with a noise diode as follows. The diode with a resistor in parallel is connected to the input terminals of the network, and the noise power P_1 is measured at the output, in the small frequency band Δf, with the diode passing no current; the noise power P_2 is then measured with the diode in operation. Again putting $P_2/P_1 = a$, we find from the definition of the noise factor, using (6) and (7):

$$F = \frac{20\ I_s R - 1 + a + (1-a)\dfrac{T_1}{T_0}}{a-1} . \qquad (8)$$

This formula is very much simpler if the temperature T_1 of the resistor is roughly equal to T_0 ($= 290$ °K) and if a is given the value 2 (i.e. if the diode current I_s is adjusted so that $P_2 = 2P_1$; see [1])). In that case:

$$F = 20\ I_s R . \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

Formulae (8) and (9) are valid in a wide frequency range. In this range the diode is an absolute noise standard, since the formulae contain no empirical constants. We shall now consider the limits of this frequency range. We shall see that they are partly of a fundamental nature and partly due to the finite dimensions of the diode.

Lower limit of frequency range

Below a certain frequency a diode shows in addition to the above-mentioned noise, which is due to the shot effect, a kind of noise known as flicker noise [6]). According to one theory, flicker noise is bound up with the "scintillating" character of the emission: after a certain spot on the cathode has emitted a batch of electrons, some time elapses before the same spot can emit electrons again. Investigations have shown [7]) that a tungsten cathode, as used in noise diodes, exhibits distinct flicker noise only at frequencies of 10 c/s and lower. Impurities in the cathode and traces of gas may raise this limit, however, so that it is safer not to use a diode for noise measurements below, say, 100 or 1000 c/s.

In practice, this limitation is of little significance, noise diodes seldom being used in the audio-frequency range. Noise measurements at audio frequencies are usually done with two different resistors, both at room temperature [8]).

Upper limit of frequency range

At very high frequencies there are two causes of deviations from the Schottky formula (eq. 6): one we shall call the "transformation error" and the other the "transit-time error". The transformation error is attributable to stray capacitances and inductances; the transit-time error is significant at frequencies which are so high that the period of oscillation is not long compared to the time taken by the electrons to travel from cathode to anode. We shall now consider the magnitude of these errors.

The transformation error

Some years ago a short article appeared in this journal [9]) describing a new noise diode (the 10 P type already mentioned) and two circuits making use of this diode for noise measurements in the decimetre wave range; one circuit employed a Lecher system and the other a coaxial system (*fig. 8a*). To calculate the transformation error of such a circuit, we use the equivalent circuit shown in *fig. 9a*. Here the current source i represents the noise diode, C_0 the capacitance of the diode, L_0 the inductance of the lead-in wires, Z_1 the impedance of the coupling capacitors, and Z the resistance R and the section of transmission line connected in parallel with it. The other end of this line is fitted with a shorting plunger, which is so adjusted that, at the operating frequency f, the impedance Z is equal to R (i.e. such that the shorted section of line exactly compensates the influence of the various reactances at this frequency). To examine the influence of L_0, C_0 and Z_1 on the noise, we transform this circuit into one in which Z is shunted

[6]) W. Schottky, Small-shot effect and flicker effect, Phys. Rev. 28, 74-103, 1926.

[7]) J. G. van Wijngaarden, K. M. van Vliet and C. J. van Leeuwen, Low-frequency noise in electron tubes, Physica 18, 689-704, 1952.

[8]) See page 75 *et seq.* of the book by Van der Ziel mentioned in footnote [3]).

[9]) H. Groendijk, A noise diode for ultra-high frequencies, Philips tech. Rev. 20, 108-110, 1958/59.

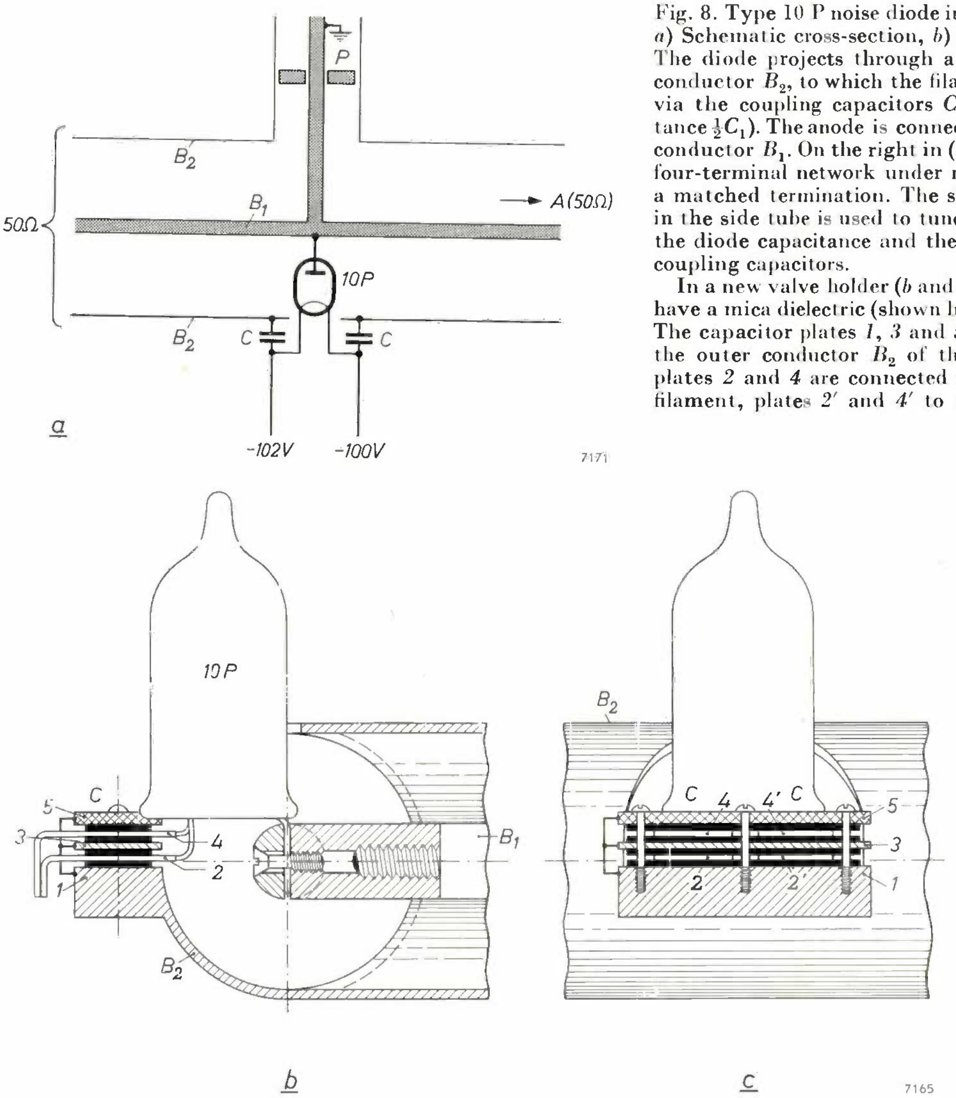


Fig. 8. Type 10 P noise diode in a coaxial system. *a*) Schematic cross-section, *b*) and *c*) side views. The diode projects through a hole in the outer conductor B_2, to which the filament is connected via the coupling capacitors C (each of capacitance $\frac{1}{2}C_1$). The anode is connected to the central conductor B_1. On the right in (*a*) is connected the four-terminal network under measurement; left, a matched termination. The shorting plunger P in the side tube is used to tune out the effects of the diode capacitance and the impedance of the coupling capacitors.

In a new valve holder (*b* and *c*) the capacitors C have a mica dielectric (shown black in the figure). The capacitor plates *1*, *3* and *5* are connected to the outer conductor B_2 of the coaxial system, plates *2* and *4* are connected to one side of the filament, plates *2'* and *4'* to the other side.

across the current source i' (fig. 9*b*); i' is the current which produces across the noiseless resistance R^* (fig. 9*c*) a noise voltage equal to the output noise voltage in fig. 9*a*. It can be calculated that the two circuits are equivalent at the frequency f if the following relation exists between the currents i' and i:

$$i' = \frac{i}{1-(2\pi f)^2 L_0 C_0 + \mathrm{j}\times 2\pi f C_0 Z_1} \quad . \quad . \quad (10)$$

The impedance Z_1 of the two coupling capacitors in parallel can be reasonably approximated by the impedance of a capacitance C_1 and an inductance L_1 in series. Introducing a frequency f_t, defined by $(L_0 + L_1)C_0 = (2\pi f_t)^2$, we find from (10), after taking the mean square, the following expression for the *transformation factor* γ_t:

$$\gamma_t = \frac{\overline{i'^2}}{\overline{i^2}} = \left[1-\left(\frac{f}{f_t}\right)^2+\frac{C_0}{C_1}\right]^{-2}.$$

$1-\gamma_t$ is the *transformation error* we wish to find.

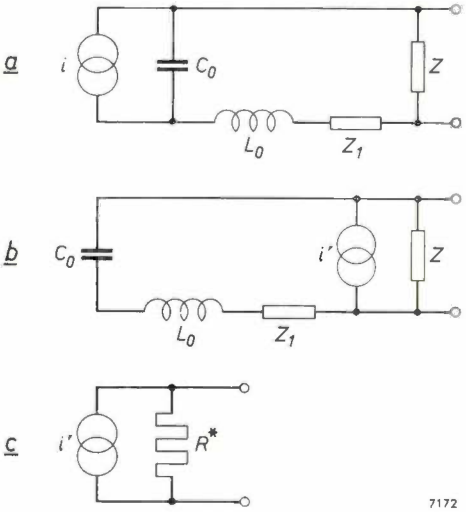


Fig. 9. *a*) Equivalent circuit of a noise diode (noise-current generator i), the capacitance of which is C_0; L_0 is the inductance of the lead-in wires, Z_1 the impedance of the coupling capacitors, and Z the impedance of the resistance R in parallel with a section of transmission line.
b) The current source i in (*a*) has been replaced by a current source i' in parallel with Z.
c) Equivalent circuit of (*b*), consisting of the current source i' in parallel with the noiseless resistance R^*.

There is no objection to raising the capacitance of the coupling capacitors sufficiently for C_1 to be large compared with the capacitance C_0 of the diode. In that case, at least at frequencies f not too close to f_t, we can disregard C_0/C_1, giving:

$$\gamma_t \approx \left[1 - \left(\frac{f}{f_t}\right)^2\right]^{-2} . \quad . \quad . \quad . \quad . \quad (11)$$

Fig. 10 shows a plot of γ_t versus f in accordance with (11), with f_t as parameter. At frequencies a great deal lower than f_t, the transformation factor γ_t is roughly 1, and the error therefore about zero. The value of f_t at which γ_t becomes 1.10 ($f = 0.22\, f_t$) is roughly the upper limit of the frequency range in which the diode can be regarded as an absolute noise standard; at higher frequencies the correction $1 - \gamma_t$ is too inaccurate.

From fig. 10 it can be seen that $1 - \gamma_t$ increases markedly with increasing f_t. The aim is therefore to make $f_t = 1/[2\pi\sqrt{(L_0 + L_1)C_0}]$ as high as possible and thus to minimize L_0, L_1 and C_0. Careful assembly of the diode and the proper choice of coupling capacitors are therefore of considerable importance in this respect. In the 10 P noise diode C_0 has the very low value of 1.8 pF. The values of L_0 and L_1 depend closely on the construction of the valve holder and the coupling capacitors. In the article cited [9]) a valve holder was described using ceramic coupling capacitors for coaxial systems. In a later model (fig. 8*b* and *c*) these capacitors were replaced by mica capacitors (of 230 pF each, so that $C_1 =$ 460 pF). From impedance measurements a value of 2900 Mc/s was found for the frequency f_t of the 10 P diode in this holder; noise measurements [10]) at 500 to 1500 Mc/s yielded values from 2600 to 3400 Mc/s, depending on the frequency. Comparison of these results with the resonance frequency of the diode itself: $1/(2\pi\sqrt{L_0C_0}) = 3500$ Mc/s, shows that the stray inductance L_1 is satisfactorily low.

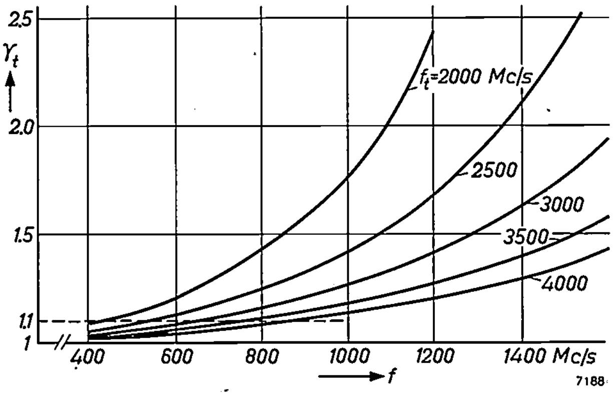


Fig. 10. The transformation factor γ_t, given by (11), as a function of frequency f, with the frequency f_t as parameter.

The transit-time error

In a diode the electrons take a finite time to travel from the cathode to the anode. In diodes such as the 10 P type, of coaxial cylindrical construction with a cathode diameter of 0.1 mm and an anode diameter of 1 mm, and with an anode voltage V_a which is high enough for operation well within the saturation region, the electron transit time τ is:

$$\tau = \frac{1.08 \times 10^{-9}}{\sqrt{V_a}} \text{ second} \quad . \quad . \quad (12)$$

(with V_a in volts). During this time the electron induces a current pulse in the external circuit. Assuming that the current pulses of the different travelling electrons are mutually independent (i.e. that the electron emission is random and that there is no space charge), and moreover that the electrons leave the cathode radially and without an initial velocity, the noise current can be calculated by a suitable summation of the current pulses [11]). We then find that we must add to Schottky's formula (6) a transit-time factor γ_τ:

$$\overline{i^2} = 2\gamma_\tau q I_s \Delta f . \quad . \quad . \quad . \quad . \quad . \quad (13)$$

For the 10 P diode operating well within the saturation region this factor is given by:

$$\gamma_\tau = 1 - 2.67\,(f\tau)^2 ,$$

in which the terms in the fourth and higher powers of $f\tau$ are neglected. Using (12) we can then write:

$$\gamma_\tau = 1 - 3.13\,\frac{f^2}{V_a} \times 10^{-18} . \quad . \quad . \quad (14)$$

It can be seen from (14) that at low frequencies γ_τ approaches unity, in which case (13) reduces to (6), the original Schottky equation. The use of (6) at higher frequencies gives rise to an error of $1 - \gamma_\tau$, the *transit-time error*.

It should be noted that the error is greater than follows from (14) if the anode voltage is so low that the diode is only barely saturated. One reason is that the assumption of radial emission without initial velocity is then no longer correct (an electron that does not leave the cathode radially is a longer time in transit than one that does, and consequently

[10]) See p. 302 under "Experimental determination of $\gamma_t\gamma_\tau$".

[11]) E. Spenke, Die Frequenzabhängigkeit des Schroteffektes, Wiss. Veröff. Siemens-Werke 16, No. 3, 127-136, 1937. G. Diemer and K. S. Knol, The noise of electronic valves at very high frequencies, I. The diode, Philips tech. Rev. 14, 153-164, 1952/53.

induces a pulse of different shape). Another reason is that the space charge affects the potential gradient and so changes the transit time of the electrons. The result is that equations (12) and (14) are not valid if V_a is too low.

Some means is therefore needed of ascertaining whether the diode is saturated or not. This cannot be seen clearly enough from a plot of the diode current I_d ($\leqq I_s$) versus V_a. A sharper criterion can be derived from the variation with V_a of the factor Γ^2, which is a measure of the suppression of the space charge. This factor occurs in the formula for the shot noise of a diode at frequencies up to about 100 Mc/s:

$$\overline{i^2} = 2\Gamma^2 q I_d \Delta f.$$

The factor Γ^2, which is equal to unity at saturation, quickly drops to a low value if the diode becomes less saturated, making the space-charge effect perceptible. For a given value of V_a a maximum value of I_d can be indicated at which Γ^2 deviates from unity and saturation is thus no longer present.

Fig. 11 shows Γ^2 as a function of V_a, with I_d as parameter, for the 10 P diode at 30 Mc/s. From this we can determine what the minimum V_a must be, at given values of I_d, if Γ^2 is not to differ by more than 1%, 2% or 3% from unity; see *fig. 12*. The values of V_a at which the anode dissipation $I_d V_a$ reaches 2 W, which is the maximum permissible value for the 10 P diode, are also shown in this figure; operation in the shaded region is thus not permissible.

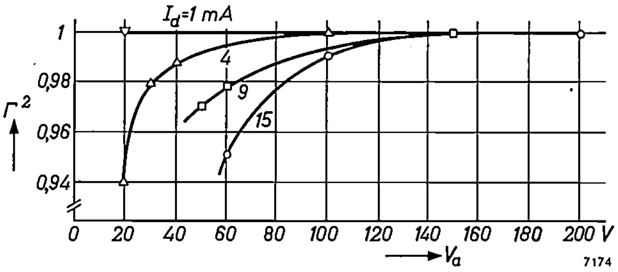


Fig. 11. The factor Γ^2 of the type 10 P noise diode at 30 Mc/s as a function of anode voltage V_a, for various values of the diode current I_d.

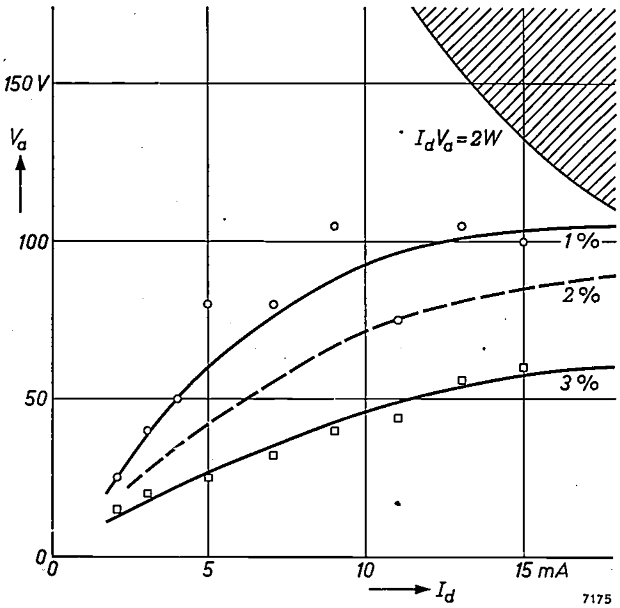


Fig. 12. Anode voltage V_a of type 10 P noise diode as a function of diode current I_d, with Γ^2 differing from unity by 1%, 2% and 3%, respectively. Curve $I_d V_a = 2$ W indicates the values of V_a at which $I_d V_a = 2$ W.

Transformation and transit-time errors combined

In general, both errors are present and must be taken into account:

$$\overline{i'^2} = 2\gamma_t \gamma_\tau q I_s \Delta f.$$

The highest frequency, then, at which the Schottky formula is still applicable is that where $\gamma_t\gamma_\tau$ only just permissibly deviates from unity. Having regard to the accuracy required for most noise measurcments, the deviation is usually fixed at 10%. If $\gamma_t\gamma_\tau$ is known in a particular frequency range, the Schottky formula can be corrected to enable the diode to be used in that range.

From (11) and (14) it appears that, where f is smaller than f_t, the two γ factors differ from 1 in opposite senses: γ_t is greater than 1 and γ_τ is smaller. The two errors, then, compensate one another to some extent. In usual conditions γ_t differs more from 1 than γ_τ. The compensation can therefore be improved, i.e. the product $\gamma_t\gamma_\tau$ brought closer to unity, by increasing the electron transit time. This can be done by lowering the anode voltage (*fig. 13a*), but not so far that the diode ceases to be saturated; otherwise a space charge would form and γ_τ would then depend on the diode current. The space-charge effect may be greater at high frequencies (of the order of 1 Gc/s) than at lower.

The magnitude of the effect of lowering the anode voltage can only be roughly estimated, because the diode then enters a region where (14) is no longer valid. Calibration is therefore necessary. We can see this plainly by comparing fig. 13*a* with fig. 13*b*. Both graphs give γ_τ as a function of frequency for various values of the anode voltage; fig. 13*a* gives values calculated from (14), and fig. 13*b* measured values. It is assumed that γ_τ for $V_a = 300$ V varies in accordance with (14); theory and measurements indicate that this will be correct to within about 2%.

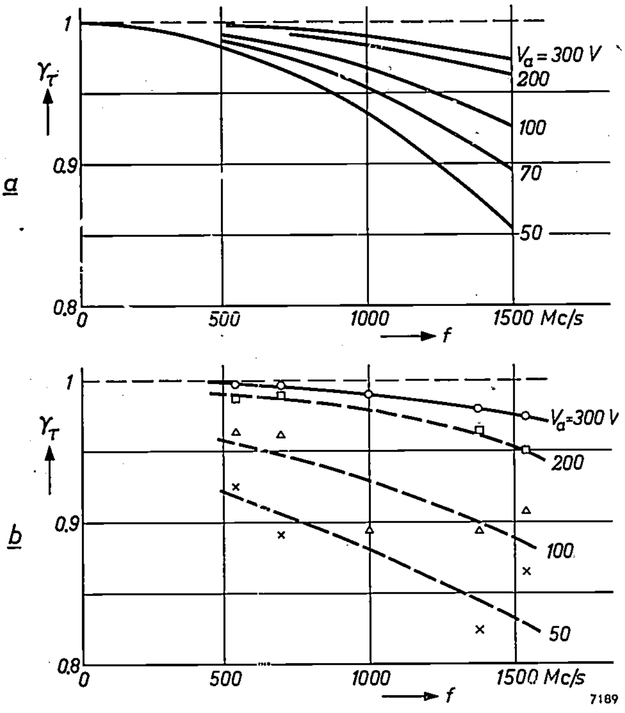


Fig. 13. *a) Calculated* values of the transit-time factor γ_τ of a 10 P noise diode as a function of frequency *f*, with the anode voltage V_a as parameter.
b) Measured values of γ_τ for V_a = 200, 100 and 50 V, in relation to the calculated values for V_a = 300 V.

The values of γ_τ measured at V_a = 200, 100 and 50 V are plotted in relation to the calculated values for V_a = 300 V. (As will appear in the next section, γ_τ by itself cannot be measured but the product $\gamma_t\gamma_\tau$ can.) We see from fig. 13*a* and *b* that the measured γ_τ differs increasingly from the calculated values as the anode voltage is reduced: the average discrepancy is about 1% at V_a = 200 V, about 4% at 100 V, and about 5% at 50 V.

In *fig. 14* the calculated and measured values of $\gamma_t\gamma_\tau$ for the 10 P diode are plotted versus frequency, with V_a as parameter. Lowering V_a from 100 to 50 V widens the frequency range from about 730 to about 980 Mc/s, a gain of 35%. Set against this is the great disadvantage that the diode current at V_a = 50 V should not be more than 4.5 mA, as otherwise the factor Γ^2 will deviate by more than 1% from unity; see fig. 12. At such a low diode current only small noise factors can be measured satisfactorily, large ones not, at least not without correction.

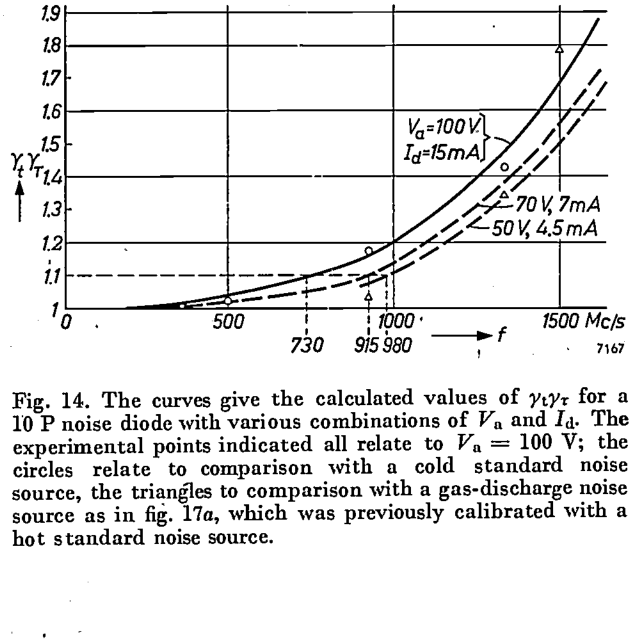


Fig. 14. The curves give the calculated values of $\gamma_t\gamma_\tau$ for a 10 P noise diode with various combinations of V_a and I_d. The experimental points indicated all relate to V_a = 100 V; the circles relate to comparison with a cold standard noise source, the triangles to comparison with a gas-discharge noise source as in fig. 17*a*, which was previously calibrated with a hot standard noise source.

Experimental determination of $\gamma_t\gamma_\tau$

The quantity $\gamma_t\gamma_\tau$ has been determined experimentally [12]) in the following way. A resistor $R(T_1)$ is held at room temperature, and a resistor $R(T_2)$ of roughly the same value at 77 °K (in liquid nitrogen). Both are provided with a variable impedance transformer (e.g. of the type sketched in fig. 3), which has the same temperature as the appertaining resistor. First of all, the transformed resistance of one of the resistors is made as nearly as possible equal (to within about 5%) to 50 ohm, by means of a standing-wave detector. This resistor is then connected via a coaxial switch to a bridge circuit (which need not be calibrated), after which the bridge is balanced. The second resistor is connected to the bridge by turning the switch, and the relevant impedance transformer is adjusted until the bridge is again balanced. This substitutional method makes it possible to equalize the two resistances with an accuracy up to 0.1%.

The next step is to assemble the circuit indicated in *fig. 15*: behind the switch *S* come a 10 P noise diode in a coaxial holder [13]), a superheterodyne

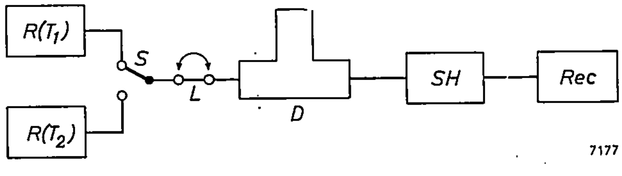


Fig. 15. Measurement of $\gamma_t\gamma_\tau$. The resistors $R(T_1)$ and $R(T_2)$ have as nearly as possible the same value R (≈ 50 ohm) at the respective temperatures T_1 (≈ room temperature) and T_2 (= 77 °K). *D* noise diode type 10 P in coaxial holder. *SH* superheterodyne receiver. *Rec* recorder. When switch *S* is in the position as drawn and the diode is without filament current, the recorder gives a certain deflection. When *S* is in the other position, the filament current is adjusted so as to produce the same deflection on the recorder. Using eq. (15) it is then possible to calculate $\gamma_t\gamma_\tau$.

At *L* a coaxial line of 50 ohm characteristic impedance and a quarter wavelength long can be inserted for a second measurement, to reduce the error caused by the fact that $R(T_1)$ and $R(T_2)$ do not have exactly the same values.

[12]) By W. E. C. Dijkstra of this laboratory.

[13]) The switch with resistors thus takes the place of the anode resistor of 50 ohm described in the article mentioned in footnote [9]).

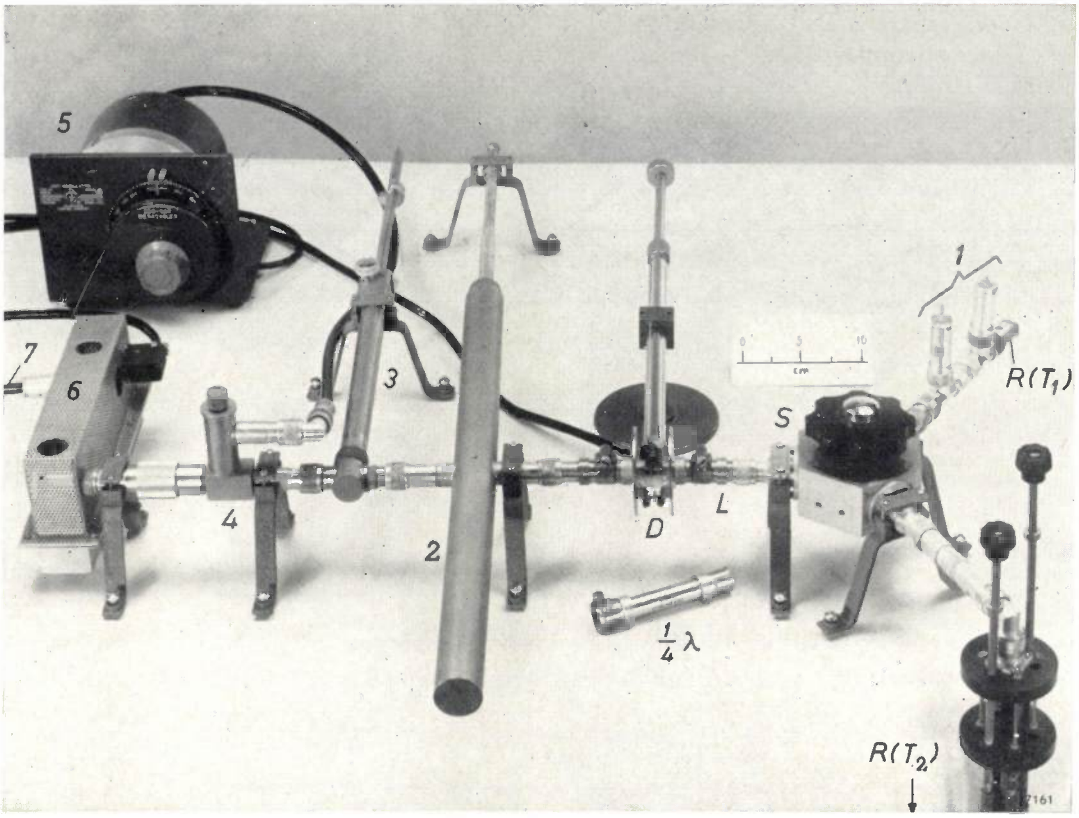


Fig. 16. Part of the equipment for measuring $\gamma_t\gamma_\tau$ at decimetre wavelengths by the method illustrated in fig. 15. $R(T_1)$ standard noise resistor at room temperature, with matching transformer *1*. $R(T_2)$ standard noise resistor at 77 °K as in figs 3 and 4 (only the upper part of the impedance transformer can be seen). *S* coaxial switch. *D* noise diode type 10 P in holder as shown in fig. 8. At *L* a section of waveguide a quarter wave in length can be inserted.

The equipment to the left of *D* is part of the superhet receiver: *2* coaxial resonant cavity, functioning as a filter that only passes a narrow band of frequencies around the measuring frequency *f* (it therefore does not pass the frequency of the local oscillator, nor the image frequency); *3* impedance transformer for matching the crystal mixer *4* to the resonant cavity *2*; *5* local oscillator; *6* first and second stages of the IF amplifier (30 Mc/s); *7* cable to the remaining part of the IF amplifier.

receiver and a recorder. Some of the equipment can be seen in *fig. 16*.

The resistor $R(T_1)$ is now switched in, the diode not yet passing any filament current. The noise of $R(T_1)$, which reaches the receiver through the switch and the diode holder, produces a certain deflection on the recorder. After switching over to $R(T_2)$ the filament current of the diode is adjusted until the recorder shows the same deflection as before. The noise produced by the resistor at 77 °K together with the diode is then equal to the noise of the resistor alone at the temperature T_1. For this condition we easily arrive at the following formula for $\gamma_t\gamma_\tau$:

$$\gamma_t\gamma_\tau = \frac{T_1 - 77}{RT_0I_s \times 20}. \qquad (15)$$

An important point is that the receiver requires no calibration. Since R occurs in the formula, however, its value must be accurately known.

The effect of a small error ΔR in R can be reduced by a simple expedient: *two* measurements are done as described above, but in the second we insert between the switch and the diode holder, at *L* in fig. 15, a coaxial line whose characteristic impedance is 50 ohm and which is a quarter wavelength long. In (15) we must then replace R by

$$R' = \frac{50^2}{50 + \Delta R} \approx 50 - \Delta R\,.$$

The error in R', then, is just as great as in R, but of opposite sign; the average of $\gamma_t\gamma_\tau$ from the two measurements is therefore more accurate than the result of one measurement. A condition, however, is that the characteristic impedance of the inserted line must be more exactly equal to 50 ohms than R, otherwise the improvement is illusory.

Summarizing, it can be said that the 10 P diode, with $V_a = 100$ V and $I_d = 15$ mA, can be used without correction as a standard noise source from about 100 c/s to about 730 Mc/s, the maximum error being 10%. At higher frequencies either correction or compensation is necessary. With wide-band circuits, compensation has the advantage that the noise does not depend on the frequency.

It should also be noted that in certain applications, where for instance the equipment has to be adjusted for minimum noise, only relative noise differences are important; what is needed here, then, is a constant source, which need not be a standard one. For example, the automatic noise-figure indicator mentioned [1]) in combination with a 10 P diode has proved very useful for adjusting 4 Gc/s equipment for minimum noise, and this is certainly not the highest frequency at which this is possible.

The gas discharge as a noise source

Mechanism of noise generation by a gas discharge

The positive column of a gas discharge emits electromagnetic radiation having the character of noise. In the centimetre and millimetre wavebands the column of an inert-gas discharge, which is both long and strongly coupled to the measuring circuit, is a particularly suitable noise source for the purposes of measurement.

The positive column consists of ions, electrons and neutral particles. There are roughly just as many positive elementary charges per unit volume as there are negative ones. Such a quasi-neutral mixture is called a *plasma.* The electrons in the plasma have a much higher average velocity than the ions and the neutral particles. This is due to their very low mass (compared with the other particles) as a result of which they are much more accelerated in the electric field and moreover lose very little energy upon elastic collisions with heavy particles.

Owing to the deceleration which an electron suffers upon a collision, a small fraction of its energy is converted into electromagnetic radiation ("Bremsstrahlung"). Since the velocities of the electrons show a random distribution, botb in magnitude and direction, the emitted radiation has the character of noise. The power of this radiation depends on the average kinetic energy of the electrons upon collision, that is on the "electron temperature" T_{el}. Let the mass of an electron be m and the mean square velocity be $\overline{v^2}$, then T_{el} is defined as:

$$\tfrac{1}{2}m\,\overline{v^2} = \tfrac{3}{2}\,k\,T_{\text{el}}\,.$$

Because the average energy of an electron is much greater than that of the other particles, T_{el} is a high temperature (of the order of 10^4 °K), much higher than the temperature of the gas, which as a rule is not much above room temperature.

It may be asked how the power radiated by the plasma depends on T_{el}. If either the electron velocities show a Maxwell distribution, or the collision frequency is constant, the radiant power of the plasma is equal to the noise power available from a resistor at the temperature T_{el} [14]), in other words in a small frequency band Δf it is equal to $kT_{\text{el}}\,\Delta f$. Experiments have shown [15]) that even though the above conditions are not entirely fulfilled, $kT_{\text{el}}\,\Delta f$ can be a good approximation for the power radiated by a plasma in the band Δf. In most cases, then, the power can be assumed to have this value.

Various forms of gas-discharge noise sources

A simple form of gas-discharge noise source consists of a gas-discharge tube mounted in a resonant cavity, with an impedance-matching transformer. This construction is useful for physical research on plasmas, but does not constitute a noise source suitable for a wide range of frequencies. Such a source can be realized, however, in another way. Three examples are given in *fig. 17*; the frequency ranges for which they are suitable will be found in fig. 1 (lines *I* and *II*).

The constructions in fig. 17, although differing considerably from one another, are all based on the same principle. We can make this clear with the aid of the diagram in *fig. 18.* This figure represents the longitudinal cross-section of a waveguide, which has a matched termination at one end in the form of an absorption wedge A, whose temperature is T_1. From B to C extends a homogeneous plasma which, we shall assume, fills the entire cross-section of the waveguide.

Both the wedge A and the plasma emit noise waves. The *wedge* sends out noise waves (power P_A, temperature T_1) from A to D. The waves, after passing through the plasma, where they are attenuated, have the lower power P_A' corresponding to a temperature lower than T_1. The *plasma* sends out noise waves to left and right, having powers P_B and P_C, which can be calculated by integrating the emission over the whole plasma column. The wave P_B is completely absorbed in the wedge A, and may therefore be left out of further consideration. The wave leaving the noise generator at D has an equivalent noise temperature T_{eq}, which is the sum of the temperatures corresponding to P_A' and P_C (this summation is permissible owing to the fact that

[14]) G. Bekefi and S. C. Brown, Microwave measurements of the radiation temperature of plasmas, J. appl. Phys. **32**, 25-30, 1961 (No. 1).
G. H. Plantinga, The noise temperature of a plasma, Philips Res. Repts **16**, 462-468, 1961 (No. 5).

[15]) K. S. Knol, Determination of the electron temperature in gas discharges by noise measurements, Philips Res. Repts **6**, 288-302, 1951.

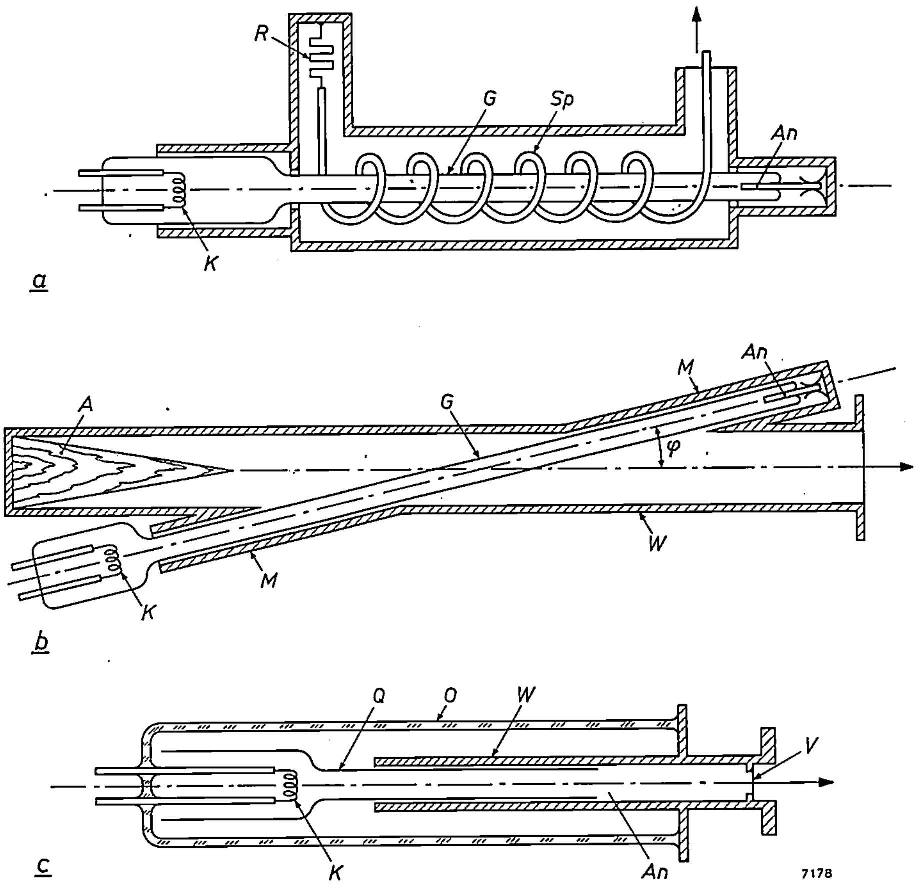


Fig. 17. Three constructions of gas-discharge noise sources. G gas discharge tube with cathode K and anode An.

a) Construction for decimetre waves. The discharge tube is coupled via the helix Sp to a system which has a matched termination on the left in the form of a resistor R, and on the right goes over into a coaxial system.

b) Construction for centimetre waves, the discharge tube passing obliquely through the waveguide W with side arms M. A is an absorption wedge providing a matched termination.

c) Construction with axial gas discharge in a circular waveguide W, for millimetre waves. Q thin-walled tube of quartz glass, open at both ends. The glass envelope O is filled with neon. V mica window. The inside wall of the waveguide W just past one end of the tube Q serves as anode (An).

the two noise waves are mutually independent). Thus,

$$T_{\text{eq}} = T_A' + T_C\,.$$

If L is the attenuation suffered by the power of the waves in passing through the entire plasma column, we can write

$$T_{\text{eq}} = \frac{1}{L}\,T_1 + \left(1 - \frac{1}{L}\right) T_{\text{el}}\,. \qquad (16)$$

It is assumed here that the reflections at the boundary planes B and C are negligible. (Reflection from C lowers the equivalent temperature T_{eq}; reflection from B lowers or raises T_{eq}, depending on whether the waves are in phase or in anti-phase.) If L is sufficiently large, T_{eq} differs very little from T_{el}, and reflections from B have no further perceptible influence, no more than has the temperature T_1.

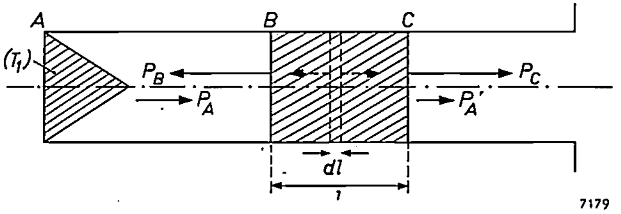


Fig. 18. Illustrating the principle shared by the three designs in fig. 17. AD waveguide terminated at the left by an absorption wedge, with a plasma between B and C. The plasma sends out noise waves to left and right, having a power P_B and P_C respectively; the wedge emits noise waves with a power P_A, which are attenuated in the plasma to the power P_A'.

Equation (16) can be derived as follows. We imagine that the wedge A is heated to a temperature equal to T_{el}. According to the Nyquist theorem, the total noise power in the band Δf which goes to D must be equal to $kT_{\text{el}}\,\Delta f$, for to the left of C everything is at the temperature T_{el} (the plasma being equivalent to an absorption medium at the temperature T_{el}), and we have assumed that there are no reflections. To the power passing from C to D the wedge A contributes $L^{-1}\,kT_{\text{el}}\,\Delta f$. The plasma contribution is therefore $(1 - L^{-1})kT_{\text{el}}\Delta f$. When the wedge is now cooled to the temperature T_1, its contribution drops to $L^{-1}kT_1\Delta f$, whereas that of the plasma shows no change. The total is thus:

$$kT_{\text{eq}}\Delta f = L^{-1}\,kT_1\Delta f + (1 - L^{-1})kT_{\text{el}}\Delta f\,,$$

which leads directly to eq. (16).

Some requirements which must be met by a gas-discharge noise source

In order to build a gas-discharge noise source that will deliver a high and constant noise power in a broad range of frequencies, it is necessary, as we have seen, to start with a plasma of high electron temperature T_{el}, and to introduce this in the circuit in such a way that the attenuation L is high and there are no reflections at the boundary plane C (fig. 18).

The electron temperature depends in a complicated manner on the kind of gas used, the gas pressure, the temperature of the column and the discharge current. In inert-gas discharges, T_{el} increases with a decrease in the atomic weight of the gas or its pressure, or the column diameter or the current density [15][16]. Under otherwise identical conditions, therefore, helium gives the highest electron temperature. For various reasons, however — including the short life obtained with helium and the considerable heat generated in a helium discharge — the lightest but one inert gas is preferred, namely neon.

The requirement of a high attenuation L was difficult to fulfil in the millimetre wavebands, and necessitated a new design. The elimination of reflections at C gave the greatest difficulties in the decimetre range. To provide some insight into these problems, we shall consider the complex relative dielectric constant ε_{r} that can be assigned to a plasma. The imaginary part of ε_{r} relates to the dissipation, and the ratio of the imaginary to the real part relates to the phase shift:

$$\varepsilon_{\mathrm{r}} = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \nu^2} + \mathrm{j}\,\frac{\omega_{\mathrm{p}}^2\nu}{\omega(\omega^2 + \nu^2)} \,. \qquad . (17)$$

Here ω_{p} is the plasma frequency, given by $\omega_{\mathrm{p}} = qN^2/\varepsilon_0 m$ (where N is the electron density and ε_0 the dielectric constant of free space), ω is the angular frequency of the wave, and ν the average collision frequency; we assume that ν does not depend on the electron velocity. It may be concluded from (17) that for any given plasma (given ω_{p} and ν) ε_{r} is closer to unity the higher is the angular frequency ω, i.e. the shorter the wave. If the remainder of the waveguide in fig. 18 is filled with a non-ionized gas (e.g. air) for which $\varepsilon_{\mathrm{r}} = 1$, there will be no reflection at the boundary planes. However, the absorption in the gas — and hence the attenuation L — shows a marked decrease, for as ω increases, the imaginary part of ε_{r} approaches zero. A high attenuation L can therefore only be obtained by filling the waveguide with plasma *over a considerable length.*

The construction that most closely approaches this schematic picture is that shown in fig. 17*c*. Before discussing this, we shall consider the more conventional designs sketched in fig. 17*a* and *b*.

A gas-discharge noise source for decimetre waves

Since waveguides for decimetre waves would have to be very large, *coaxial* systems are generally used in this wave range. A suitable noise source can be seen in fig. 17*a*. A gas-discharge tube G, e.g. a type K 50 A tube, is placed inside a silver-plated helix Sp, which effects the coupling with the gas discharge and at each end passes into the central conductor of a coaxial system. The coupling takes place over the entire length of the helix, and thus has the gradual nature that enables the reflection to be kept at a very low value. One end of the helix is connected to a matched termination (the resistance R of temperature T_1). The helix is dimensioned so that it has a characteristic impedance of 50 ohm, equal to that of the output plug.

Further particulars of the dimensioning will be found in the literature [17].

A gas-discharge noise source for centimetre waves

Fig. 17*b* shows the commonly used system designed by Johnson and Deremer for waveguides of not all too small dimensions [18]. The gradual transition between the column and the waveguide — necessary for minimizing reflections from the column — is obtained here by passing the gas-discharge tube obliquely through the waveguide (at an angle φ). The tube current (on which ω_{p} depends, see eq. 17) and the angle φ can be chosen in such a way that the reflection from the column is practically zero. At one end the waveguide has a matched termination in the form of an absorption wedge.

Since the cathode and anode of the tube are outside the waveguide, there is a danger that a considerable part of the noise power will be lost through the side arms M which contain the discharge tube. To avoid such losses, the arms are made so narrow that their lowest cut-off frequency is higher than the operating frequency of the noise generator. This means, of course, that the cross-section of the column must be quite a bit smaller than that of the waveguide; as a result the attenuation L is not maximum, but this is not a serious objection in the centimetre waveband, where the attenuation is still amply sufficient for the purposes for which it is used.

Table I gives some equivalent noise temperatures obtained in our laboratory with noise generators of

[16] A. von Engel and M. Steenbeck, Elektrische Gasentladungen, Part II, Springer, Berlin 1934, p. 85 *et seq.*
F. M. Penning, Electrical discharges in gases, Philips Technical Library 1957, p. 58 *et seq.*

[17] H. Schnittger and D. Weber, Über einen Gasentladungs-Rauschgenerator mit Verzögerungsleitung, Nachr.-techn. Fachber. 2, 118-120, 1955.

[18] H. Johnson and K. R. Deremer, Gaseous discharge super-high-frequency noise sources, Proc. Inst. Radio Engrs 39, 908-914, 1951.

Table I. Equivalent noise temperatures obtained with various types of discharge tube in the noise generator shown in fig. 17*b* (except for the last line, which relates to the construction in fig. 17*c*).

Freq.	Wave-length	Gas-discharge tubes				Equivalent noise temperature
		type	gas	pres-sure	current	
Gc/s	cm			torr	mA	°K
4	7.5	K 51 A	Ne		200	23 800
		exper.	Xe	10	150	9 550
6	5	exper.	Ne	8	125	23 400
		exper.	Ar	8	125	14 000
10	3	exper.	He	10	125	28 000
		K 50 A	Ne		125	21 700
		exper.	Ar	8	125	14 000
		exper.	Xe	5.5	125	9 400
34	0.88	exper.	He	40	100	22 700
		exper.	Ne	90	100	20 600
		exper.	Ar	40	75	13 400
		exper.	Kr	20	95	11 200
		exper.	Xe	8	75	9 200
75	0,40	exper.	Ne	100	75	21 000

this type [19]); a few of the discharge tubes employed are shown in *fig. 19*. It can be seen from the table that high noise temperatures T_{eq} are achieved with helium, in accordance with the expected high electron temperature T_{el}. We have already mentioned, however, some of the reasons why this gas is nevertheless unsuitable (short life, excessive heat generation). A further reason is that, with helium, T_{el} (and hence T_{eq}) is highly dependent on impurities in the gas. Neon does not have these drawbacks and the noise temperatures obtained are only in a few cases lower than those achieved with helium; this is attributable to the high attenuation L which is possible in neon [19]). Neon is therefore the gas preferred in practice.

It is also important to note that the values of T_{eq} obtained with neon are fairly close to one another, in spite of the markedly divergent conditions (gas pressure, current and frequency), which promises well for its usefulness at even higher frequencies.

A gas-discharge noise source for millimetre waves

The small dimensions of waveguides for waves shorter than about 8 mm make the construction in fig. 17*b* difficult for practical reasons. For millimetre waves the construction shown in fig. 17*c* is more suitable; this was described some time ago in this review [20]), and can therefore be dealt with here very briefly.

In a circular waveguide W (fig. 17*c*) a thin-walled tube Q of quartz glass is introduced. A flared part of the tube contains the oxide cathode K for the gas discharge. The neon is contained (at a pressure of 100 torr) in the tube Q, and also in the waveguide, which is closed at V by a mica window. Since the mean free path in the gas is very short (approx. 6.5 μ), the plasma ends fairly abruptly where the quartz-glass tube ends, and at that position, at An, the anode is formed by the inside of the waveguide; beyond that point, then, the neon is not ionized. The last line of Table I relates to a tube of this type.

From equation (16) we saw that a high attenua-

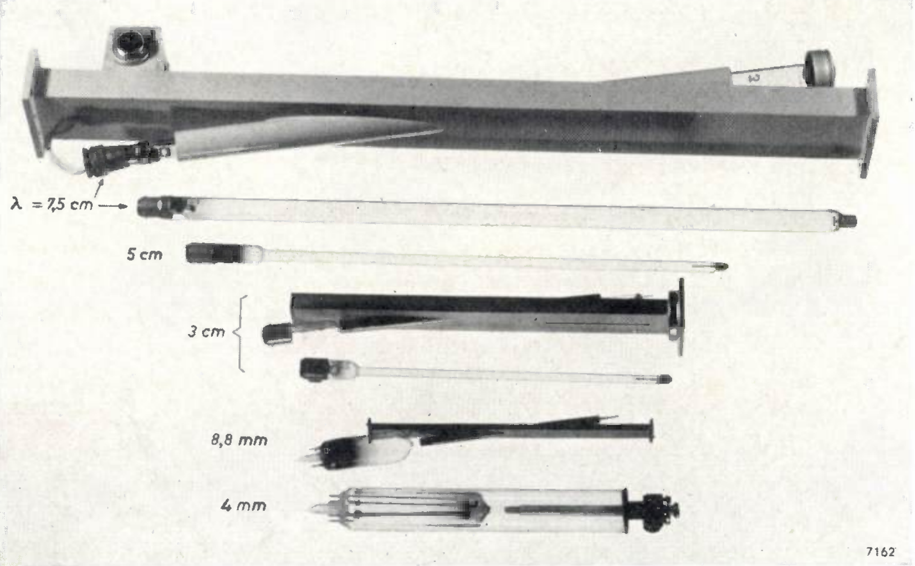


Fig. 19. Experimental gas-discharge noise sources. From top to bottom:
a 7.5 cm waveguide with discharge tube passing obliquely through it,
the 7.5 cm tube separately,
a tube for 5 cm wavelength,
a 3 cm waveguide with tube,
the 3 cm tube separately,
an 8.8 mm waveguide with tube,
a 4 mm tube as described in reference [20]).
The tubes are all shown with the cathode on the left and the anode on the right, the waveguides with the output on the right.

[19]) For further details of experiments at frequencies from 10 to 75 Gc/s see: P. A. H. Hart and G. H. Plantinga, Millimetrewave noise of a plasma, Proc. 5th Internat. Conf. on ionization phenomena in gases, Munich 1961, pp. 492-499 (North-Holland Publishing Co., Amsterdam 1962).

[20]) P. A. H. Hart and G. H. Plantinga, An experimental noise generator for millimetre waves, Philips tech. Rev. **22**, 391-392, 1960/61 (No. 12).

tion L is necessary for matching in a wide range of frequencies. In this construction, the attenuation is obtained through the column, which is here in the axial direction of the waveguide. Since a column can be made arbitrarily long (provided the applied voltage is high enough), the attenuation can in principle be made as large as required by simply making the quartz-glass tube and the waveguide long enough.

The wall of the quartz-glass tube is made very thin (about 0.1 mm); this gives the filling factor of the plasma in the waveguide a high value and limits the losses resulting from noise power leaking away along the tube.

As indicated in the last line of Table I, a noise temperature of 21 000 °K has been achieved at a wavelength of 4 mm.

We shall now briefly consider what the minimum and maximum frequencics are for the noise generator of fig. 17*c*.

Minimum frequency

The lower the frequency, the larger must be the diameter of the waveguide. This has adverse consequences for the gas discharge. The diameter of a positive column cannot be widened indefinitely without giving rise to effects such as constriction and striations. The positive column then gradually loses its character and the electron temperature drops. Moreover, as seen from equation (17), as ω decreases the relative dielectric constant differs more and more from unity; this makes it evident that the abrupt ending of the column at the anode will increasingly give rise to reflection.

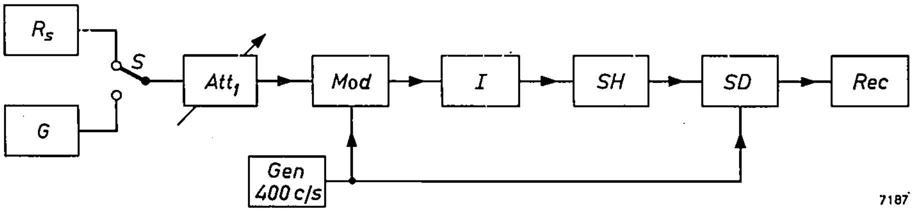


Fig. 20. Calibration of a gas-discharge tube G with a standard noise resistor R_s. S switch. Att_1 calibrated attenuator. *Mod* modulator (directional isolator) which modulates the noise with a 400 c/s signal delivered by the generator *Gen*. *SH* superhet receiver. *SD* synchronous detector. *Rec* recorder. *I* directional isolator which prevents the noise from *SH* from reaching the modulator (this noise would otherwise be modulated and cause errors).

For these reasons the lowest frequency at which the noise generator in fig. 17*c* can be used with advantage is in the region of 35 Gc/s.

Maximum frequency

With increasing frequency the attenuation of the column per unit length decreases. To keep L large enough, it is therefore necessary to make the column and the waveguide longer. Lengthening the waveguide increases the losses in the waveguide wall. There is thus a danger that these losses will finally predominate. The noise would then largely be due to the guidc wall, whosc temperature is relatively low (400 to 500 °K).

One way of getting around this difficulty is to make the waveguide relatively wide. For a wavelength of 4 mm, for example, we have made the inside diameter of the waveguide 4 mm. Outside the noise generator a gradual transition is then needed to the usual rectangular waveguide.

The upper limiting frequency is probably higher than 300 Gc/s.

Calibration of gas-discharge noise sources

A resistor, and also a noise diode (at least in a wide range of frequencies) can be regarded as absolute noise standards; a gas discharge, however, cannot be so regarded and must therefore be calibrated. To conclude this article, we shall briefly consider this process of calibration.

Our gas-discharge noise sources were calibrated by comparing them with a standard noise source. This consisted of a resistance in the form of an absorption wedge (the matched termination in fig. 17) whose tcmperature was adjusted as accurately as possible to 1336 °K, the melting point of gold[5]).

Fig. 20 shows a simplified block diagram of the set-up; the major part of the equipment can be seen in *fig. 21*. The part on the right of the switch is a somewhat simplified version of Dicke's noise receiver, as used in radio astronomy [21]). This consists essentially of a modulator *Mod*, which modulates the incoming noise with an audio signal (400 c/s), followed by a superheterodyne system *SH*, a synchronous detector *SD* and a recorder *Rec*. In the synchronous detector the output signal from the superhet receiver is compared with the 400 c/s signal, with the result that the recorder responds only to a signal modulated with 400 c/s, i.e. to the incoming noise.

The modulator is a modified version of a directional isolator using Faraday rotation [22]). The modi-

[21]) R. H. Dicke, The measurement of thermal radiation at microwave frequencies, Rev. sci. Instr. **17**, 268-275, 1946. See also: C. A. Muller, A receiver for the radio waves from interstellar hydrogen, II. Design of the receiver, Philips tech. Rev. **17**, 351-361, 1955/56.

[22]) H. G. Beljers, The application of ferroxcube in unidirectional waveguides and its bearing on the principle of reciprocity, Philips tech. Rev. **18**, 158-166, 1956/57.

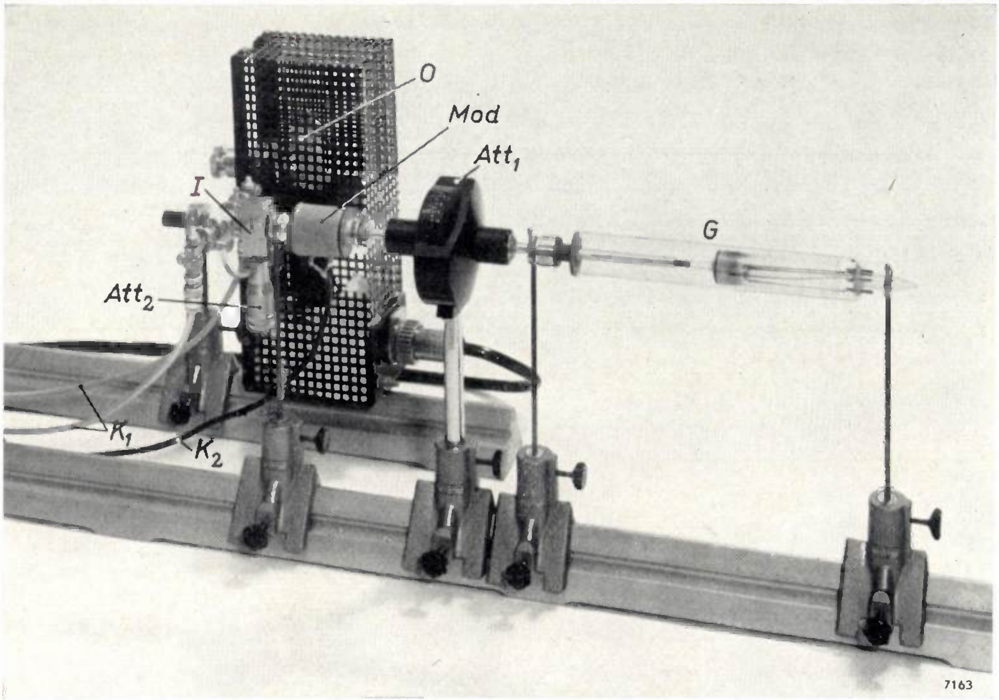


Fig. 21. Part of the equipment for calibrating the gas-discharge tube *G* (here a 4 mm tube). For Att_1, *Mod* and *I*, see fig. 20. *O* local oscillator (4 mm klystron) of the superhet receiver. Att_2 variable attenuator for adjusting the signal from *O* to the correct value. The cables K_1 go to the IF amplifier (with push-pull input); cable K_2 comes from the 400 c/s oscillator.

fication consists in the permanent magnet having been replaced by an electromagnet, which is energized by the 400 c/s signal. To prevent the inherent noise of the receiver also being modulated via reflections from the modulator, a second directional isolator is inserted between the modulator and the superhet receiver.

The procedure of calibration is as follows. First of all, the modulator is connected to the standard noise source R_s via a variable attenuator Att_1 set at minimum attenuation; the recorder shows a certain deflection. Next, the attenuator is connected to the gas-discharge noise source *G*, and the attenuation is adjusted until the recorder shows the same deflection as before. The gas-discharge noise source together with the attenuator now delivers just as much noise as the standard source.

Now the equivalent temperature T_{eq} of the gas-discharge noise source is given by

$$T_{eq} = B\,T_s' - (B-1)\,T_1\,,$$

where *B* is the inserted attenuation (including the attenuation due to losses in the line), T_1 the temperature of the attenuator, and T_s' the temperature (corrected for line losses) of the standard noise source.

As mentioned at the beginning of this article, gas-discharge noise sources are relatively insensitivc to fluctuations in supply voltages and in ambient temperature, so that after calibration they can serve as sub-standards. Because of their reproducibility, it is not necessary to calibrate them individually; it is sufficient to calibrate a few samples.

Summary. Survey of the three main types of standard noise source: resistors, saturated diodes and gas discharges.

Resistors can be used as noise standards at frequencies ranging from the lowest to the ultra-high, in the order of 100 Gc/s (mm waves). For measuring the noise factor of a four-terminal network, two noise resistors are needed, one of which has the same resistance at the temperature T_1 as the other at the temperature T_2. Usually T_1 is made roughly equal to room temperature, and T_2 much higher or much lower. Examples discussed are a cold resistor for a coaxial decimetre-wave system ($T_2 = 77$ °K, bath of liquid nitrogen) and a hot resistor, mounted in a waveguide, for centimetre or millimetre waves ($T_2 = 1336$ °K, in an oven).

Dealing with the *saturated diode*, the author examines the limits of the useful frequency range. The lower limit is set by flicker noise and lies between 10 and 1000 c/s. The upper limit depends on the extent to which two errors occur: the transformation error and the transit-time error. The correction and mutual compensation of these errors are discussed. The type 10 P noise diode can be used in a new type of holder, without correction, up to 730 Mc/s.

On the subject of *gas-discharge* noise sources, the mechanism of the noise generation is examined. Three noise generators of this kind are reviewed: one for the decimetre band, one for the centimetre band and one for the millimetre band; they differ in the method of coupling the plasma with the waveguide. The article ends with a description of the calibration of this type of noise source.

A PHOTORECTIFYING LAYER FOR A READING MATRIX

by J. G. van SANTEN *) and G. DIEMER *).

621.383.44:621.319.2

Photoconductive materials have already proved their usefulness in many fields; we may mention the use of photocells for flame monitoring in oil-heating systems and for crackle-free volume control in radio receivers, and the use of photocathodes for pick-up tubes in television cameras. The use of photoconductive cadmium sulphide for a "reading matrix" involves a special problem, which has been elegantly solved by the authors.

Automatic read-out with the aid of photoresistors

With the continued advances in the field of electronic computers and their uses, it is frequently desirable to have some means by which an optical image can automatically be recognized or "read". For this purpose the image can be projected on to a plate which is prepared in such a way that the quantity of light incident on it can be determined point by point by electrical means. In cases of importance in practice, only the contour of the figure needs to be recognized, so that it is sufficient to determine which points are illuminated and which are not.

To this end, use can be made of photoresistors, e.g. of cadmium sulphide, whose electrical resistance is dependent on the intensity of illumination. A number of these photoresistors are arranged in the manner of the elements in a matrix, the image to be identified is projected on to this array, and a voltage is applied successively to each of the photoresistors. Depending on whether a large or small current begins to flow, one can conclude whether the photoresistor is illuminated or not. In order for the "reading matrix" to "see" the contour in sufficient detail, the dimensions of the picture elements, i.e. of the photoresistors employed, must be as small as possible (*fig. 1*).

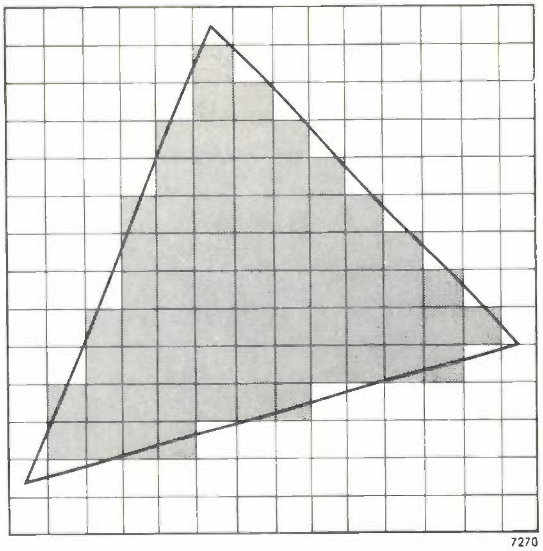


Fig. 1. How a matrix, with which an optical image can be automatically "read" (identified), "sees" the figure projected upon it.

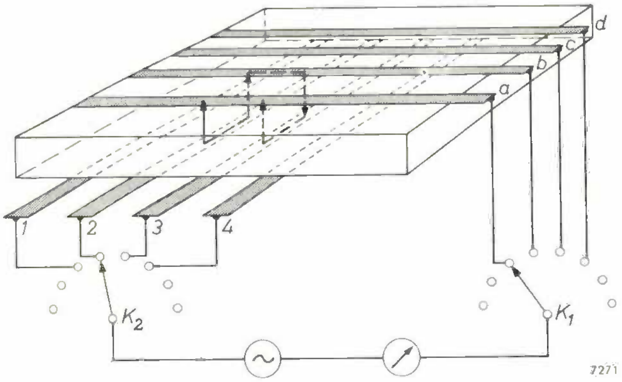


Fig. 2. Principle of automatic read-out with the aid of a matrix. On the two sides of a layer of photoconductive material, strips *a*, *b*, *c*, *d*, *1*, *2*, *3*, *4* are applied. The strips are scanned by selectors K_1 and K_2. Whenever K_1 is in one of the positions *a*, *b* . . ., K_2 traverses the positions *1*, *2*, ... When the voltage is on strips *a* and *2* and the cross-point of the strips is illuminated, a current *I* begins to flow (solid arrow). Since, as a rule, other cross points are illuminated too, stray currents flow along all sorts of paths. One of these paths is indicated in the figure by a dashed line.

The reading matrix is most simply constructed by applying conductive strips to both sides of a flat, thin layer of cadmium sulphide, as illustrated in *fig. 2*. A picture element consists of that part of the layer where two strips cross each other. (As will be explained at the end of this article, the construction of the actual array (*fig. 3*) is slightly different.) The figure to be examined is projected on to the matrix, and the strips are scanned with the aid of two selectors K_1 and K_2. Every time K_1 is in one of the positions *a*, *b*, *c*, . . . , K_2 traverses the positions *1*, *2*, *3*, ... When the position of the

*) Philips Research Laboratories, Eindhoven.

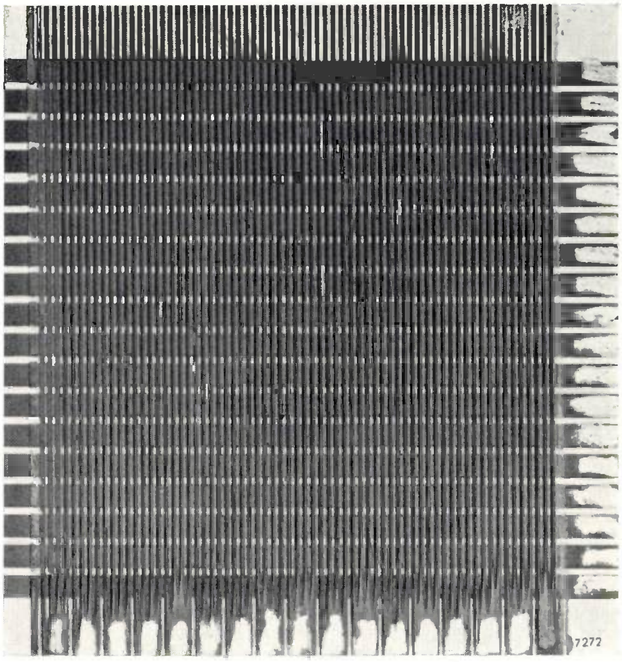


Fig. 3. A matrix for reading optical images.

selectors is such that the voltage appears on the element $a2$ (between strip a and strip 2), for example, a current I_{a2} starts to flow, which is higher the more light falls on $a2$. If this current is higher than a preset value, $a2$ is counted as being inside the contour. The total information about the elements which is thus obtained can be compared in a computer with that for a series of contours stored in the computer memory. If the computer finds a correspondence, it reports this and the figure is recognized.

The matrix made in the manner described is not readily usable, however, for the following reason. When the voltage is across strips a and 2, a current flows not only through the photoresistor $a2$ but also along all sorts of other paths, e.g. through the series-connected photoresistors $b2$, $b3$ and $a3$; see fig. 2. *Fig. 4* shows the equivalent circuit for a matrix containing two sets of five strips; all other resistors are seen to be in certain combinations parallel with $a2$. If there are a number of illuminated photoresistors in the parallel circuits, it is possible that stray currents will flow which will make a contribution I' to the total current I_{a2}. Thus, even though $a2$ is not illuminated, the total current I_{a2} may be greater than the preset value referred to, and as a result the element $a2$ is interpreted as being illuminated.

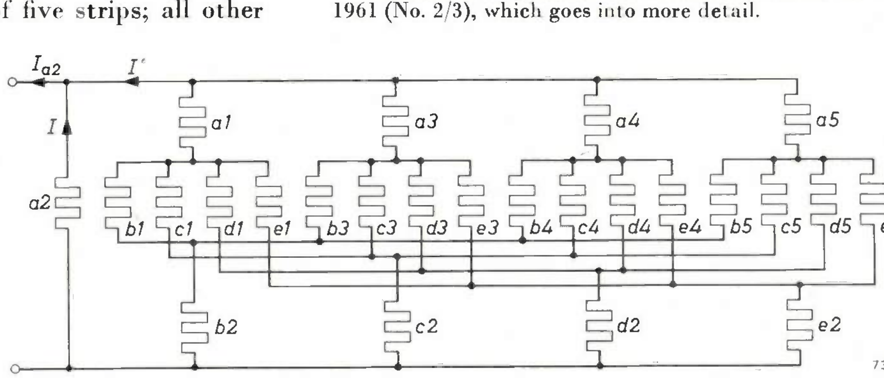


Fig. 4. When the voltage is on strips a and 2, all other resistors are connected in certain combinations parallel to $a2$, as shown here for a matrix with two sets of five strips. The current I' through the parallel circuits can make such a large contribution to the total current I_{a2} as to give the impression that the element $a2$ is illuminated, although it is not.

This difficulty can be overcome by using a material whose resistance depends not only on the light intensity but also on the direction in which the voltage is applied. In such a case we speak of a *photorectifier*. When the matrix is made using a material of this kind, we have what amounts to a rectifier in series with each resistor. The path for a stray current will always contain at least one rectifier in the reverse direction, so that the total contribution from the stray currents remains sufficiently low.

Although there are already a few types of photorectifiers on the market, they cannot readily be used in a matrix. This is due for one thing to their size, and for another to the fact that the matrix would have to be built up from a large number of individual elements, which would be cumbersome work compared with the application of strips to a layer. It has now proved possible, on the basis of the photoconductive material cadmium sulphide[1]), to produce a new type of photorectifier with which the matrix can easily be formed with strips in the manner described.

The new photorectifier [2])

The new photorectifier is prepared by mixing CdS powder with a small quantity of powdered

[1]) The properties and possible applications of CdS have been dealt with at length in: N. A. de Gier, W. van Gool and J. G. van Santen, Photo-resistors made of compressed and sintered cadmium sulphide, Philips tech. Rev. **20**, 277-287, 1958/59. It may be recalled here that the best photoconductive properties are obtained not with pure CdS but with powdered CdS which contains suitable additives in carefully controlled amounts. This mixture is compressed into pellets and sintered. Extremely sensitive photoresistors can be made in this way, e.g. types ORP 30, ORP 90 and LDR.B8.73104.

[2]) This photorectifier is described in: J. G. van Santen and G. Diemer, Photorectifier based on a combination of a photoconductor and an electret, Solid-state electronics **2**, 149-156, 1961 (No. 2/3), which goes into more detail.

glass enamel [3]), and firing the mixture at the melting point of the enamel (about 600 °C). As a result of the adhesive forces, the enamel spreads in the form of a thin layer (thickness a few Å) between the grains (diameter a few μ), thus, as it were, cementing the grains together (*fig.* 5). Although, owing to the insulating nature of the enamel, the resistance upon illumination is considerably more than that of the sintered CdS powder, the essentially new feature is that we can now, by a special treatment, make from the enamel an *electret*, i.e. a permanently polarized dielectric, which gives the required rectification. The treatment consists in again firing the CdS and glass enamel, this time at about 200 °C, and cooling it in a strong electric field. The polarization produced in the enamel by the action of the field is thus "frozen in".

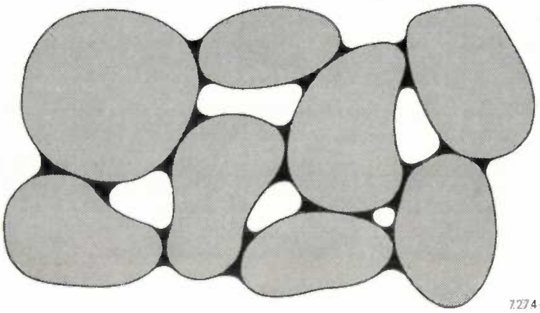


Fig. 5. Schematic cross-section of the new photorectifying material, consisting of a layer of CdS grains with glass enamel between them.

We shall now consider in more detail the explanation of the rectifying action of the combination described, and discuss its application in a matrix.

Conduction mechanism of photoconductor-electret combination

We shall first consider the case of two ordinary conductors (e.g. of copper) separated by an insulator, and discuss the conduction mechanism in terms of the band scheme as represented in *fig. 6a*, and explained in the caption to this figure *).

When a voltage is applied between the two conductors, the band schemes are mutually displaced (fig. 6*b*). Although there is now a potential difference across the insulator, no perceptible current flows, provided the insulator is not too thin. There is little chance that an electron will "tunnel" through the potential barrier *klmn*, i.e. pass from one conductor to the other without having sufficient energy to surmount the potential barrier; the probability is in fact smaller the greater are the "cross-section" *klmn* and the height *kl*. If we raise the voltage, more and more electrons will tunnel through the barrier, so that the current increases. This does not change the picture in fig. 6*b*, however, since the potential drop in the conductors is negligible. We note that the application of the voltage causes polarization in the insulator (displacement of the positive and negative charges). The accompanying additional surface charges at the contact faces between insulator and conductors are compensated by the supply and removal of the free negative charge carriers abundantly present in a conductor.

We shall now turn to the case of two grains of photoconductive CdS between which there is a very thin layer of insulating glass enamel (*fig. 7a*). When a voltage is applied to the CdS grains, the potential barrier due to the presence of the enamel undergoes a change similar to that in fig. 6*b*. Because the enamel layer is very thin, the potential barrier has a relatively small cross-section even before the

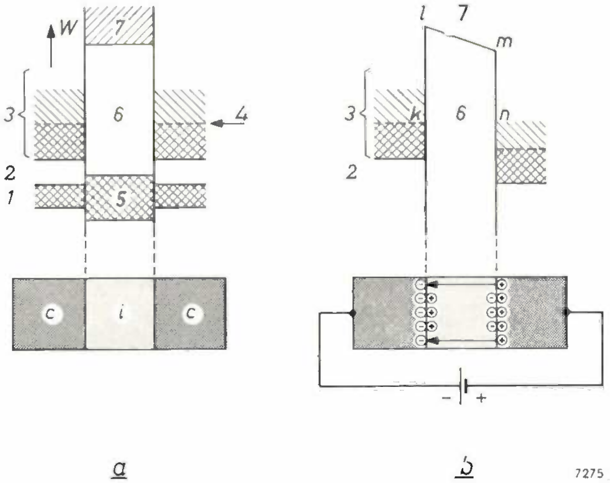


Fig. 6. *a*) Energy-level diagram showing the allowed and forbidden bands of two conductors *c* (of the same material) separated by an insulator *i*. The single hatching relates to bands of allowed levels, the cross-hatching indicates that the levels are occupied by electrons. *1* is a completely occupied band, *2* a forbidden band, *3* is the partly occupied band typical of a conductor. Some of the electrons which cannot move in the occupied levels below *4* are raised, upon the application of a voltage, to the allowed empty levels above *4*, where they can take part in the conduction. *5* and *7* are respectively a fully occupied and an empty band of the insulator, *6* is the broad forbidden band typical of an insulator: at voltages lower than the breakdown voltage the electrons cannot pass from *5* to *7*, and conduction is precluded.

Bands *1* and *5* are not drawn in *b*.

b) The application of a voltage to the conductors causes relative displacement of the band schemes. A surface charge is produced in the conductors, and partly serves to compensate the insulator surface charge, which is due to the polarization produced in the insulator by the voltage. The cross-section of the potential barrier *klmn* and the height *kl* are so large that there is hardly any chance for an electron to "tunnel" through the barrier.

[3] Glass enamel is a type of glass which has a very short melting range (at about 600 °C), unlike normal kinds of glass, whose melting range is very long.

*) *Editor's note:* This subject will be dealt with in detail in a forthcoming article in this journal, dealing with the principles of photoconduction.

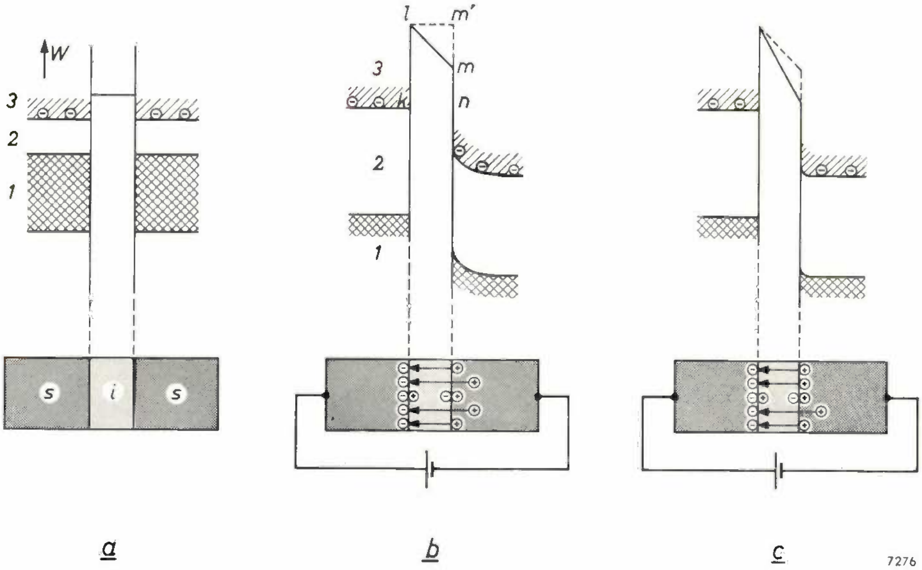


Fig. 7. *a*) The conductors in fig. 6 are replaced here by two grains *s* of photoconductive CdS; the insulator *i* is a very thin layer of glass enamel. Characteristic of a semiconductor like CdS is the forbidden band *2*, which separates the valence band *1* from the conduction band *3*. Upon absorption of a photon of sufficient energy, the electrons from the valence band can be raised to the conduction band. The relatively few electrons contained in the conduction band upon illumination are denoted here by a small number of free charge carriers and not by cross-hatching. Upon the transition of electrons from *1* to *3*, holes appear in the valence band which behave like positive charges, but which in CdS can only move slowly through the crystal lattice.
b) The already small cross-section of the potential barrier (due to the extreme thinness of the glass enamel layer) is narrowed still further when a voltage is applied. The electrons now have a greater chance to tunnel through the potential barrier, and a perceptible current flows. The holes, which are attracted towards the barrier under the action of the field, move slowly. They give rise to a space charge, and hence to a potential distribution which corresponds to a curvature of the energy levels.
c) After some time, most of the holes have reached the surface. Consequently, the potential drop is much less localized in the CdS grain than in (*b*). There is now a greater voltage across the barrier, whose cross-section is therefore still smaller. More electrons now tunnel through the barrier, as a result of which the electric current increases until a saturation value is reached. Holes constantly disappear as a result of recombination and are replaced by new holes, so that the space charge is never entirely zero.

voltage is applied. The application of the voltage narrows it still further from *klm'n* to *klmn* (fig. 7*b*). So many electrons can now tunnel through the barrier that a noticeable current starts to flow. During the first ten seconds after switching on, this current is found to be time-dependent (*fig. 8*). The reason is that the conduction mechanism in this case differs in one point from the previous one. CdS contains relatively few charge carriers, some positive and some negative. The positive charge carriers are the "holes" that are formed in the valence band when an electron jumps to the conduction band upon the absorption of a photon. These holes are not only few in number, they are also restricted in their freedom of movement by being relatively strongly bound to certain lattice sites. In the case under consideration, the holes in the positive grain also take part in the charge transport. Owing to their low mobility the holes take some time to travel to the insulator, so that there arises in the positive grain, in addition to the surface charge, a space charge near the insulator. The electrons move very fast and cause no perceptible space charge in the negative grain. Owing to the presence of the space charge, part of the potential drop is localized inside the positive grain, and this is represented by a curvature in the energy levels (fig. 7*b*). This state is not, however, stationary; more and more holes reach the surface, where they add to the surface charge. As a result the space charge decreases and the potential drop in the grain is reduced, while the potential difference across the insulator increases. The barrier becomes narrower and the current higher. After about 10 seconds a stationary state is reached. A small space charge remains (fig. 7*c*) as a result of recombination (electrons dropping back to the valence band): holes keep disappearing and fresh holes are supplied to replace them.

Summarizing then, a voltage supplied to two CdS grains gives rise to a current which rises in a time of 10 seconds to a stationary value. Its magnitude depends on the thickness of the insulator. It should be borne in mind, incidentally, that CdS is itself an insulator in the dark, when it has no electrons in the

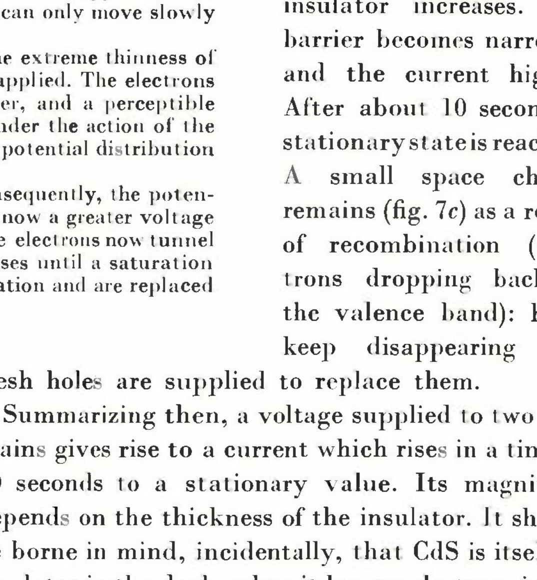


Fig. 8. The current *I* through the CdS with glass enamel, as a function of time *t* for constant voltage and illumination. The voltage is switched on at the moment $t = 0$. The curve consists of two portions: *1* immediate response, *2* slow rise to a saturation value.

conduction band. The currents mentioned are thus dependent on the illumination.

We now consider a third case, again with the two CdS grains, which are now however embedded in an electret, i.e. a thin layer of glass enamel having the above-mentioned permanent polarization. *Fig. 9a* shows the relevant band scheme. The permanent polarization in the insulator corresponds to an electric field, denoted by the arrow in the figure. The form of the potential barrier is now asymmetrical. An external voltage, applied in such a direction that the field between the grains acts in conjunction with that of the electret (fig. 9c), reduces the width of the barrier much more than a voltage of the same magnitude but in the opposite direction (fig. 9b). In the first case a current starts to flow immediately, whilst in the second case the current immediately after switching on is still almost zero.

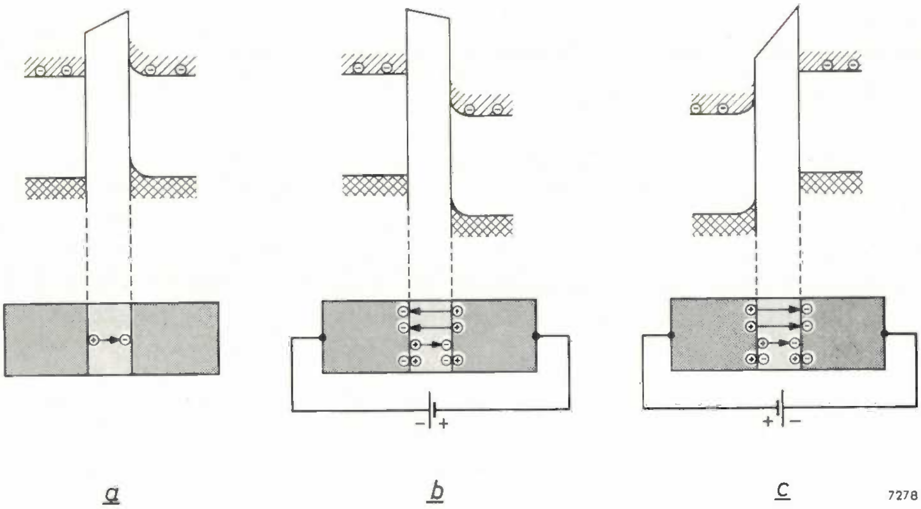


Fig. 9. *a*) The glass enamel is made into an electret by a special treatment, i.e. part of the polarization remains after the external field is removed.
b) An external voltage of the opposite polarity to that of the field of the electret is opposed by the latter field, so that the potential barrier is not significantly reduced.
c) An external voltage in the direction of the electret field makes the barrier very narrow. Immediately after application of the voltage, therefore, a much higher current flows than in case (*b*).

The above applies only *immediately after the application of the voltage*. After some time, sufficient charge carriers have arrived at the surfaces to neutralize the surface charge of the electret. The potential barrier in fig. 9c is then not quite so sharp, and that in fig. 9b is somewhat sharper, so that the two figures are each other's mirror image. The current from that moment onwards is of equal magnitude in both cases. *Fig. 10* shows the difference between the currents in the two directions as a function of time.

Now it follows from fig. 10 that the resistance of illuminated CdS with glass enamel is equal in both directions upon the application of direct voltage (at least after 10 seconds), *but upon the application of alternating voltage it is greater in one direction than in the other*. For alternating currents, the system is always in the state prevailing immediately after switching on. The combination CdS-electret-CdS is thus seen to have a rectifying action for alternating current. A current-voltage characteristic of this new rectifier can be seen in *fig. 11*.

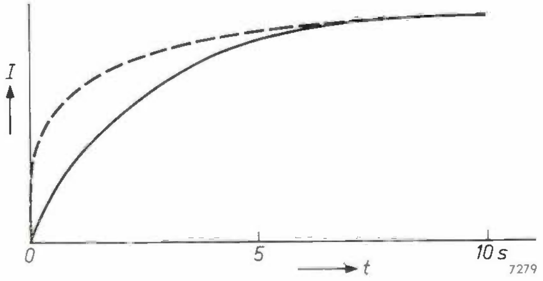


Fig. 10. Photocurrent *I* in CdS with polarized enamel as a function of time *t*, in the forward direction (broken curve) and in the reverse direction (solid curve). In the forward direction part of the response is immediate, but not in the reverse direction.

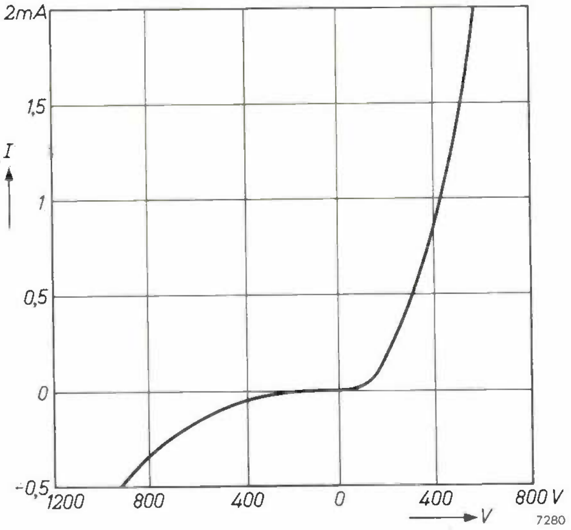


Fig. 11. Current-voltage characteristic of CdS with polarized enamel, for alternating voltage. All frequencies above about 1 c/s give the same curve.

Application of the photorectifier material in a reading matrix

To give an impression of the effect of illumination, *fig. 12* shows the relation between current and voltage for CdS with unpolarized enamel, subjected to various intensities of illumination. For comparison, the figure includes a curve relating to sintered CdS powder [1]. It should be remembered in this connection that, apart from the photoconductive properties of the two materials, a part is also played by other factors such as the interelectrode spacing and the surface area exposed to the light [2]. Even so, it is reasonable to deduce from fig. 12 that the sintered CdS powder has a *small* and *constant* resistance, whereas the resistance of CdS with enamel is much *higher* for the same illumination, and is moreover *dependent on the voltage*. This is bound up with the fact that the resistance is primarily governed by the form of the potential barrier and not by the CdS.

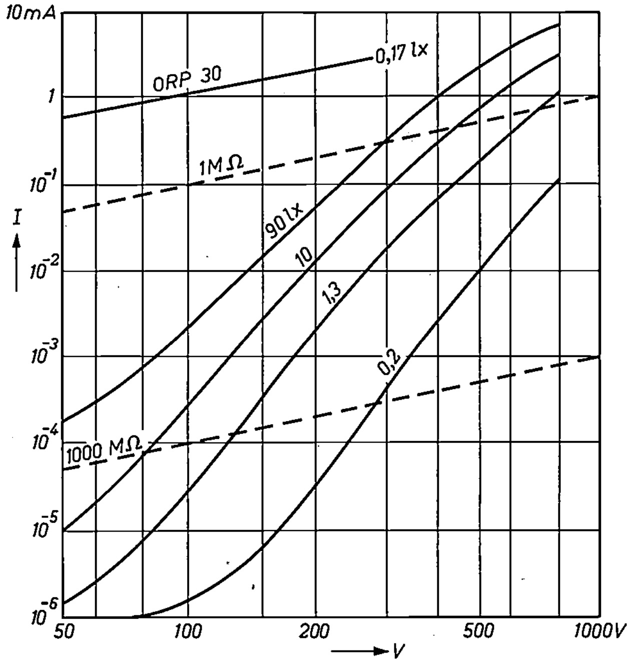


Fig. 12. Current-voltage characteristics of the photoconductive CdS layer with unpolarized enamel, illuminated with 0.2, 1.3, 10 and 90 lux respectively, and of sintered CdS powder (ORP 30) illuminated with 0.17 lux. The dashed lines apply to normal resistors of the value indicated. The relation between I and V for the CdS layer with enamel is approximately given by $I \propto V^5$, and its resistance is roughly a factor of 10^6 greater than that of sintered CdS.

Fig. 12 shows that the illumination can easily reduce the resistance of CdS with glass enamel by a factor of 1000. The material can thus be used for a reading matrix as described in the introduction. The obstacle to the discrimination between light and dark due to stray currents (fig. 2) is now overcome by the rectifying effect: when the voltage is applied in the forward direction to one particular photoresistor, there will then, as explained above, be rectifiers in the reverse direction in all other possible paths.

Fig. 13 represents a cross-section of a matrix as actually made. The grooves between the upper strips are applied for the following reason. The photons cannot penetrate deep into the material, so that the photoconduction takes place primarily near the surface. When the photoconductive material is grooved in the manner illustrated, the current is able to flow from an upper strip to a lower one via the surface layer. Without this grooving, the resistance in the forward direction would be too high.

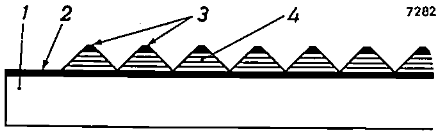


Fig. 13. Cross-section of an actual matrix. *1* ceramic base. *2*, *3* electrode strips. The photoconducting layer *4* of CdS embedded in glass enamel is deeply grooved.

One last comment on this resistance. For practical applications, the relatively high resistance is a drawback. Because of the voltage dependence mentioned (fig. 12), however, we can make the resistance quite a lot smaller by choosing a higher voltage. If the matrix is used in conjunction with equipment containing electron tubes, this presents no difficulty. Since the advent of the transistor, however, there has been a tendency to work with low voltages (below 40 V), and in that case the resistance of each element of the matrix would be too high. This difficulty can be met by choosing the highest feasible value of illumination and by designing the matrix so that every picture element always contains a number of photorectifiers, e.g. four, in parallel, thus reducing the resistance per element by a factor of 4. (The matrix in fig. 3 was designed in this way, the upper strips being divided into three.) This conflicts of course with the requirement that the elements should be as small as possible in order to allow accurate read-out, so on this point a compromise has to be accepted.

Summary. The simplest way of making a matrix with which an optical image can automatically be read (identified) is by applying parallel conducting strips to both sides of a thin layer of photoconductive cadmium sulphide, in such a way that the strips on one side cross those on the other side at right angles. This "cross-bar" construction entails stray currents, which hinder the discrimination between light and dark. A new photorectifying material is described, which is based on a combination of a photoconductor (CdS) and an electret (i.e. a permanently polarized glass enamel). When the ordinary CdS is replaced by this photorectifying material, the stray currents are largely suppressed. The rectifying effect in this material can be satisfactorily explained with the aid of the energy-band model.

DISPENSER CATHODES FOR MAGNETRONS

by G. A. ESPERSEN *).

621.3.032.213.2:621.385.16

The use of dispenser cathodes such as the L cathode or the impregnated cathode offers great advantages in magnetrons, klystrons and other tubes. Millimetre-wave magnetrons and high-power continuous-wave magnetrons for centimetre waves using these cathodes can attain a life of 1000-3000 hours. The operation is moreover very stable, and long pulses can be used.

Comparison of dispenser cathodes with oxide-coated cathodes

Most of the *oxide-coated cathodes* used in magnetrons have a porous nickel matrix. In the manufacture of these cathodes, a layer of nickel powder a few tenths of a millimetre thick is sintered on to a nickel cylinder. This porous layer is then covered with a paste of the emitting material, usually a mixture of $BaCO_3$ and $SrCO_3$, which is later converted to the oxides by heating. The object of this porous layer of metal is to reduce the voltage drop across the oxide layer, which may occur when large pulsed currents are used. If this voltage is high, breakdown may occur locally, leading to the emission of incandescent oxide particles (sparking) or the production of a visible gas discharge (arcing).

Oxide-coated nickel-matrix cathodes give excellent performance in most pulse-type magnetrons if the cathode temperature does not exceed 900 °C. It has however been found that oxide-coated cathodes have very short lives when used in millimetre-wave magnetrons, where high cathode current densities are required, and in high-power continuous-wave magnetrons, where the cathode operates at temperatures in excess of 900 °C.

It has therefore become the practice in recent years to replace the oxide-coated cathode by a *dispenser cathode* in such magnetrons. This dispenser cathode may be an L cathode [1]) or an impregnated cathode [2]). The smooth surface of the dispenser cathode ensures its ability to operate at high temperatures with relatively few arc breakdowns. This may be clearly seen from *fig. 1*, which shows the number of discharges per minute as a function of time for a 22-kW pulsed-type magnetron for the 3-cm band (type PAX-6) with an oxide-coated cathode (*a*) and for one with an L cathode (*b*) [3]). Each breakdown is moreover much less dangerous in the dispenser cathode than in the oxide-coated cathode, since the body of the former is composed entirely of tungsten and molybdenum, which have very high melting points (above 2600 °C). It is noted that the cathode temperature of both the oxide-coated matrix and the L-type cathodes in these arcing studies was approximately 950 °C and it is believed that under these conditions considerable nickel vaporization of the oxide-coated matrix cathode took place, resulting in the excessive number of arc counts.

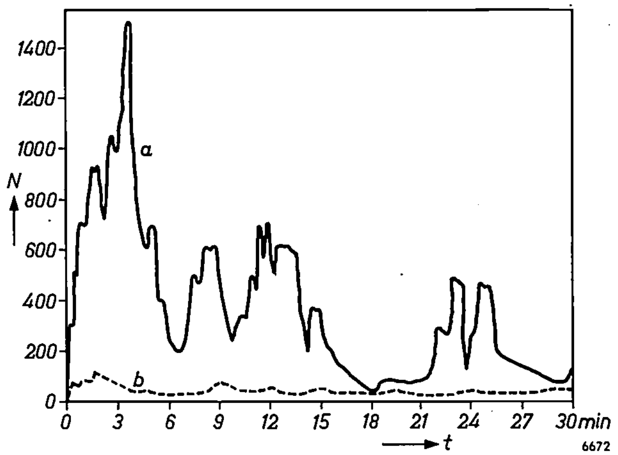


Fig. 1. Number of discharges per minute, N, as a function of the time t in a PAX-6 tube (200 pulses per sec of 5 μsec, 15 kV and 16 A), *a*) with oxide-coated cathode, *b*) with L cathode.

Another advantage of the use of the dispenser cathode in magnetrons is that for a given repetition rate it can be operated at much longer pulse lengths than the oxide-coated cathode. This has been investigated using 70-kW pulsed-type magnetrons for the 3-cm band (type 6507). In operating these tubes at the rated value of duty factor (0.001) the cathode

*) Philips Laboratories, Irvington-on-Hudson, N.Y., U.S.A.

[1]) H. J. Lemmens, M. J. Jansen and R. Loosjes, A new thermionic cathode for heavy loads, Philips tech. Rev. 11, 341-350, 1949/50.

[2]) R. Levi, Philips tech. Rev. 19, 186-190, 1957/58.

[3]) Most of the experiments mentioned here have already been described in: G. A. Espersen, Dispenser cathode magnetrons, I.R.E. Trans. on Electron Devices ED-6, 115-118, 1959.

temperature is 775 °C with the heater turned off. Under these conditions, tubes with oxide-coated matrix cathodes and with dispenser cathodes have comparable performance. If the pulse length (and hence the duty factor) is increased about threefold, the cathode temperature increases to 950 °C and the tubes with oxide-coated cathodes show considerable plate-current instability (*fig. 2a*), while the tubes with dispenser cathodes show little or no instability (fig. 2*b*). This difference was also found with tubes of other types.

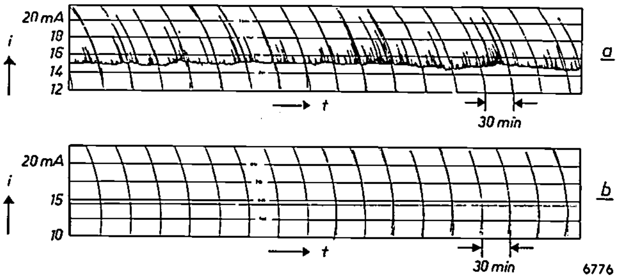


Fig. 2. Anode current *i* recorded as a function of the time *t* for a 6507 tube, *a*) with oxide-coated cathode, *b*) with impregnated cathode.

The construction of the cathode

The cathodes of magnetrons are nearly always cylindrical. *Fig. 3* shows typical constructions for the L cathode (*a*) and the impregnated cathode (*b*). The impregnated cathode has various advantages over the L cathode, e.g. a simpler structure with better thermal efficiency, more homogeneous temperature distribution along the emitting area, no need for a gas-tight weld, and more rapid evacuation. Moreover, when the external dimensions of the two types are the same, the impregnated cathode has more room for the heater (*2* in fig. 3). A heater having a larger surface area may be used and consequently it can be operated at a lower temperature to obtain the same temperature of the emitting body. This is an important advantage, especially as the temperature of the emitting body of a dispenser cathode must be higher than that of an oxide cathode in order to give the same emission.

Breakage of the heater filament is in fact the main cause of the failure of dispenser cathodes in a magnetron; the filament must therefore be designed with the greatest care. The insulating layer of aluminium oxide which is found on the heaters of oxide-coated cathodes is often omitted in dispenser cathodes because aluminium oxide reacts with the tungsten at high temperatures. The heater must then not touch the inside of the cathode at any point. As may be seen in fig. 3, a ring of ceramic aluminium oxide (*3*) is used, on one end of the cathode, for centering and insulation; but this is not in direct thermal contact with the heater.

The holder for the porous tungsten cylinder is made of molybdenum, and is brazed to the tungsten with a eutectic mixture of nickel and molybdenum at 1400 °C in an atmosphere of hydrogen.

The cathode is "seasoned" during the evacuation of the tube, i.e. the heater voltage and anode voltage are applied, so that the cathode passes current. This process is completed more quickly with dispenser cathodes than with oxide-coated cathodes, because the latter have more sharp points and other irregularities on the surface which must be burnt away. Care must be taken that the temperature of the dispenser cathode remains as low as possible, preferably below 1150 °C, in order to achieve optimum life, ensure coverage of the cathode surface by a full barium monolayer and to prevent "end emission". This very undesirable effect will now be discussed.

End emission

The magnetron has a high efficiency because the electrons can reach the anode only when their potential energy is mainly used for maintaining the

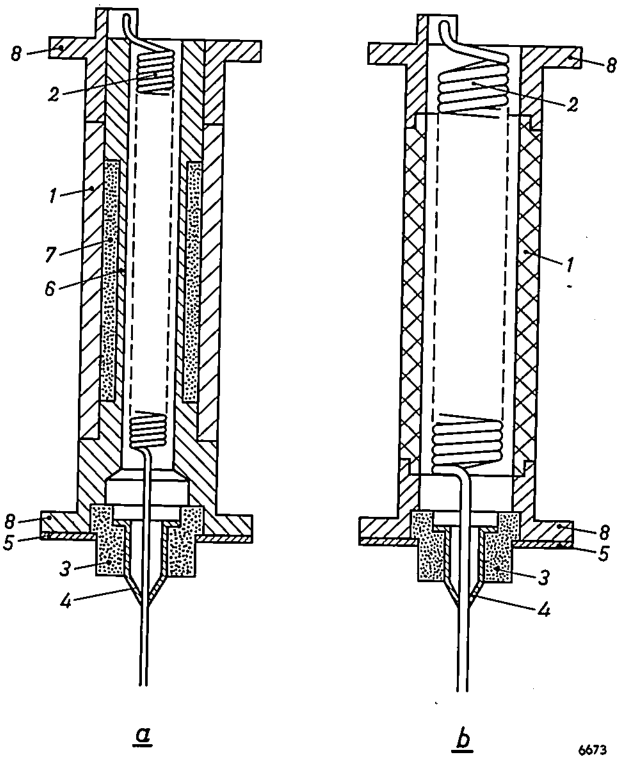


Fig. 3. Sketch of *a*) L cathode, *b*) impregnated cathode. In both figures, *1* is the porous tungsten cylinder, *2* the heater, *3* the aluminium-oxide insulation for the heater, *4* a tantalum eyelet and *5* a tantalum washer. *6* is the molybdenum inner cylinder in *a*), *7* is the emitting material in *a*), and *8* are the molybdenum end shields. In *b*) the tungsten cylinder is impregnated with the emitting material.

HF electric field. The axial component of the velocity of the electrons could however lead to a considerable current loss between the cathode and the anode. The cathode is therefore provided with end shields (fig. 3) which give rise to a field which forces the electrons back into the space between the cathode and the anode. If however the end shields themselves emit, these electrons can proceed in an axial direction, e.g. to the pole pieces, causing a decrease in the efficiency.

End emission can be caused by evaporation and/or migration of the emitting substance from the cathode proper. This can be prevented in a number of ways, in the first place by keeping the cathode temperature as low as possible. This is especially important during the seasoning of the cathode, as has already been mentioned above. It is however never possible to prevent migration and evaporation of the emitting substance at the temperatures needed for emission, at least with the dispenser cathode. We are thus left with the possibility of choosing the material and the construction of the end shields so that their temperature is kept as low as possible and so that they emit as little as possible when the emitting substance is deposited on them.

Fig. 4 shows a construction which considerably lowers the temperature of the ends by surrounding them with cylindrical screens about 1 cm long, and with a wall thickness of 0.25 mm. If the cathode temperature is 1185 °C at point *C*, the temperature at the points *B* and *D* on the screens is 1025 °C if molybdenum is used for making the latter and only 900 °C if tantalum is used. The end shields of the cathode shown in fig. 3 (which are 0.5 mm thick) had a temperature of 1110 °C under the same conditions. Titanium is also a suitable material for the screens: it has a low thermal conductivity and low thermal emission [4]), and has a gettering effect, but it can only be used if its temperature does not exceed 800 °C. Zirconium likewise has low emission and a gettering effect. Tantalum and titanium, however, present a difficult brazing problem; the same is true for zirconium, which moreover recrystallizes under the operating conditions as may be clearly seen from *fig. 5*. Zirconium-coated molybdenum screens were therefore adopted for the final design of the 5780 A 3-cm magnetron [5]).

A measure of the end emission is given by the anode current of the magnetron with a much reduced voltage, in the presence of the magnetic field. The tube does not oscillate under these conditions and the current then comes mainly from the ends of the cathode. A tube which normally operates at an anode voltage of 32 kV and an anode current of 40 A

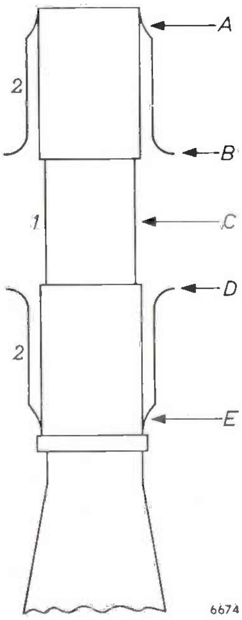


Fig. 4. Special construction of magnetron cathode to reduce end emission. *1* cathode, 7.6 mm in diameter and 8.2 mm long. *2* peripheral construction, 9.4 mm long and 0.25 mm thick.

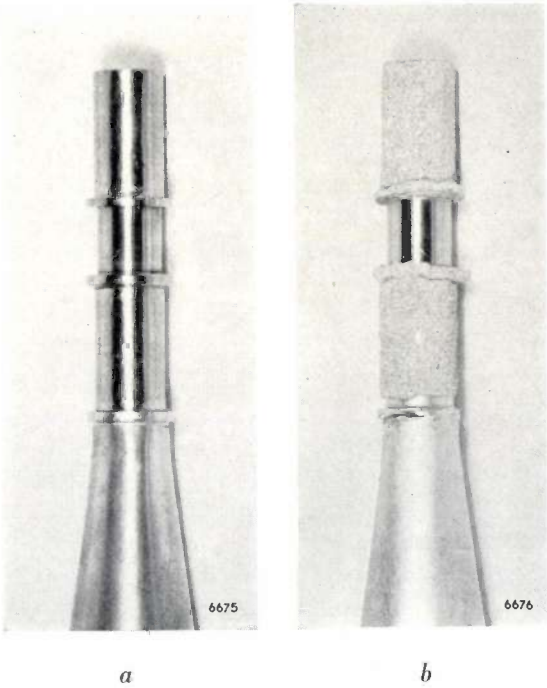


Fig. 5. Cathode with zirconium screens, *a*) before seasoning, *b*) after 280 hours' life testing.

[4]) G. A. Espersen and J. W. Rogers, Philips tech. Rev. **20**, 269-274, 1958/59.

[5]) A new technique for suppressing the electron emission from portions of tungsten dispenser cathodes consists in carburizing the areas to be rendered non-emitting by completely covering them with finely divided graphite and heating at a temperature of 1500 to 1800 °C in an atmosphere of hydrogen for a period of 5 to 15 minutes. See R. Levi and E. S. Rittner, Proc. Inst. Radio Engrs **49**, 1323, 1961 (No. 8).

Table I. Life data of magnetrons with dispenser cathodes.

Type	Sort of cathode	Frequency	Operating conditions						Mean life
			pulse or continuous	anode voltage (kV)	anode current (A)	power (kW)	duty factor	pulse length (μs)	
		(Gc/s)							(h)
DX 107 A	imp.	3.4	pulse	28	50	500	0.0005	4.0	960
PAX-6	L	9.4	pulse	15	7	22	0.0023	4.65	650
PAX-6	imp.	9.4	pulse	15	7	22	0.0023	4.65	670
6507	imp.	9.4	pulse	16	15	70	0.001	1.0 and 5.0	690
6507	imp.	9.4	pulse	16	15	70	0.001	15.0	1080
7093	imp.	35.0	pulse	15	16	32	0.0003	0.5	1500
DX 164	imp.	75.0	pulse	13	10	25	0.0002	0.1	200
7091	imp.	2.4	cont.	4.5	0.75	2.0	—	—	3000
7292	imp.	2.4	cont.	4.5	0.75	2.0	—	—	3000

may e.g. be given an anode voltage of 10 kV for this measurement. The anode current measured on one particular tube in this way was 1.0 mA for the unused tube, 2.5 mA after 1000 hours and 5.0 mA after 5000 hours.

Secondary emission

It is known that a large part of the current in oscillating magnetrons may be supplied by secondary emission, due to electrons which are accelerated by the RF field in the cathode-anode space and return to the cathode along a curved path. For example, an oscillating magnetron can give a peak anode current of 16 A and at a cathode temperature corresponding to a (primary) saturation current of 100 mA. Such an extreme case is not favourable, because the current can easily become unstable, but a certain supplementation of the primary emission by secondary emission is desirable. The secondary-emission coefficient, i.e. the number of secondary electrons produced per incident electron, is greatest for an oxide-coated cathode, other things being equal, and it is greater for an impregnated cathode than for an L cathode. Even an L cathode, however, has a secondary-emission coefficient greater than unity under normal conditions [6]), so use can still be made of the secondary emission.

Life

Life tests were carried out on a number of magnetrons of four different types. The results are summarized in *Table I*. The reason for failure was nearly always a broken heater. Types 7091 and 7292 differ from each other only in the anode cooling system. The 3-cm types mentioned in this table have recently been superseded by the types 7008, 7110, 7111 and 7112, which have a mean life exceeding 1000 hours and are considerably more stable, showing less jitter for short pulses.

It may be seen from the table that the life of these magnetrons can be much longer than that of magnetrons with oxide-coated cathodes, which usually lies between 50 and 500 hours. The long life of types 7091 and 7292 is especially important, as they are used in cooking ovens [7]), where the price of the magnetron is an important part of the cost of the whole installation.

[6]) I. Brodie and R. O. Jenkins, Brit. J. appl. Phys. 8, 202-204, 1957.

[7]) W. Schmidt, The heating of food in a microwave cooker, Philips tech. Rev. 22, 89-102, 1960/61 (No. 3).

Summary. The use of dispenser cathodes in magnetrons of a number of types is described. A cylindrical cathode offers only limited room for the heater, which necessitates special heater constructions. End emission, which is very undesirable in magnetrons, can be reduced by giving the edges of the cathode a low thermal conductivity and low emission. The results of a number of life tests on magnetrons indicate that those with dispenser cathodes last much longer than those with oxide-coated cathodes. Moreover, tubes with dispenser cathodes are much more stable and can withstand longer pulses.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of those papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2896*: J. D. Fast: Entropie. Die Bedeutung des Entropiebegriffes und seine Anwendung in Wissenschaft und Technik (Philips Technical Library, 1960, XII+328 pp., 68 figures, 26 tables). (Entropy. The significance of the concept of entropy and its applications in science and technology; in German.)

A German translation of a Dutch book. The English translation will be published in the course of this year, and will be more fully reviewed.

2897*: H. G. van Bueren: Imperfections in crystals (North-Holland Publishing Co., Amsterdam 1960, XVIII+676 pp.).

This book is written mainly for solid-state physicists who are interested in the effect of imperfections in the crystal lattice on the properties of the material they are investigating. A large number of properties of the solid state are discussed, together with the role imperfections play in determining these properties. Considerable attention is paid to the theoretical interpretation of the phenomena in question, and their interrelation. The book is divided into three parts, which deal with: the general properties of imperfections in crystals (4 chapters, 110 pp.), the metals (19 chapters, 400 pp.) and the non-metals (8 chapters, 143 pp.). The lists of references at the end of the various chapters contain in all about 1000 items.

2898: E. E. Havinga: Contribution to the theory of the dielectric properties of the alkali halides (Phys. Rev. **119**, 1193-1198, 1960, No. 4).

In this paper some relations between dielectric properties of diagonal cubic ionic crystals are derived on the basis of the Dick and Overhauser shell model for ions. The relations, which contain no model constants, are in good agreement with experimental data. Also it is shown that Dick and Overhauser overestimated the number of electrons in the shells of the ions, which accounts for the failure of their quantitative treatment. In an appendix the paper of Hanlon and Lawson on the same subject is discussed.

2899: H. Bremmer: On the theory of wave propagation through a concentrically stratified troposphere with a smooth profile (J. Res. Nat. Bur. Standards **64 D**, 467-482, 1960, No. 5).

The Wentzel-Kramers-Brillouin approximation for the solution of the height-gain differential equation for a curved stratified troposphere is discussed in detail. The approximation depends mainly on a variable $u_l(r)$ which can be interpreted as the height-dependent contribution of the phase for a field solution obtained by separation of variables. An expansion of $u_l(r)$ with the aid of partial integration leads to further approximations which facilitate the determination of the eigenvalues, and of the amplitudes of the modes connected with the propagation problem. The influence of the refractive-index profile, if assumed as smooth, then appears to be restricted to a dependence on the surface values of this index and of its gradient insofar as propagation over the ground is concerned. Further, all height effects of elevated antennas can be expressed in terms of the distance to the corresponding radio horizon. This results in simple relations between the fields connected with two different refractive-index profiles, provided both profiles coincide near the earth's surface.

2900: W. Black, J. G. V. de Jongh, J. Th. G. Overbeek and M. J. Sparnaay: Measurements of retarded Van der Waals' forces (Trans. Faraday Soc. **56**, 1597-1608, 1960, No. 11).

Molecular attractions are measured between pairs of flat quartz plates and between a flat and a spherically curved plate. Considerable precautions are taken against spurious electric charges, dust and gel particles which might interfere with the measurements. Silicone oil is used for damping. Distances between the flat plates have been varied from 5000-9500 Å, and from 940-5000 Å for the plate and sphere combination. Attraction forces varied from 0.002-0.3 dyne. The results agree with the presence of retarded Van der Waals' forces (Casimir and Polder, Lifshitz). If the force per unit area between flat plates is represented by $F = -B/d^4$, the value of B, which is $1\text{-}2 \times 10^{-19}$ erg cm, is in good agreement with existing theories, and with the previous experimental results obtained by Derjaguin and Abrikosova and by Kitchener and Prosser. An explanation is suggested why earlier measurements by Overbeek and Sparnaay using a similar method led to much stronger attractions.

2901: S. Duinker and B. van Ommen: The scansor, a new multi-aperture rectangular-loop ferrite device (Solid-State Electronics **1**, 176-182, 1960, No. 3).

The scansor consists of a multi-aperture plate of rectangular-loop ferrite material which is provided with a large number (e.g. 10-20) of separate output windings across which consecutive output pulses can be developed by driving the plate from one remanence position into the other by triangular-shaped pulses. The various output pulses are of rather uniform height but mutually delayed. Depending on the geometry of the scansor and on the slope of the driving current, the delay between pulses corresponding to any two adjacent single-turn output windings can be varied from 0.05 to 2 μs with corresponding pulse heights of 10 to 0.2 V. Experiments are described showing that the response of certain output windings can be either suppressed or shifted in time by applying appropriate additional pulses. A few possible applications of scansors such as rapid scanning devices and code converters are briefly discussed.

2902: C. M. van der Burgt: Piezomagnetic ferrites — applications in filters and ultrasonics (Electronic Technol. **37**, 330-341, 1960).

A survey of the properties and possible applications of ferrites with a high magnetostrictive effect, in particular of the recently developed Ni-Cu-Co ferrites. The conditions which these materials must satisfy for use in high-power transducers (devices for transforming electrical energy into mechanical and *vice versa*) are discussed, as are the conditions for use in electro-mechanical band-filters. Some details of such band-filters are also given. Special attention is paid to the use of such transducers for cleaning small objects by means of ultrasonic vibrations in a liquid. A self-oscillating ultrasonic drill for use with brittle materials is also described; with a few modifications, this could also be used as a self-oscillating ultrasonic welder. All these piezomagnetic transducers can incorporate feed-back elements made of piezo-electric ceramic materials.

2903: U. Enz: Spin configuration and magnetization process in dysprosium (Physica **26**, 698-699, 1960, No. 9).

Dysprosium is ferromagnetic below 85 °K and has some type of antiferromagnetic structure between 85 °K and 178.5 °K, the Néel temperature. It is shown in this paper that the magnetic properties of Dy in the antiferromagnetic region can be completely understood by assuming a helical spin configuration. The period of the helix is a function of the temperature. The behaviour of a helical spin configuration in a magnetic field is calculated.

2904: H. C. Hamaker: Le contrôle qualitatif sur échantillon (Rev. Statistique appl. **8**, No. 2, 5-40, 1960). (Quality control by sampling; in French.)

The fundamental principles forming the basis of sampling inspection are reviewed in this article. The theory of operating-characteristic curves is first examined briefly. The aims of sampling inspection procedures and the various factors to be considered are then enumerated and discussed. In the following sections economic theories, the distribution of percentage rejects in inspected batches and the relation between batch size and sample size are studied in some detail. In particular, emphasis is laid on the importance in many industrial problems of the use of samples of constant size, independent of the batch size.

In the second part of the article the special characteristics are examined of the various sampling tables in common use on both sides of the Atlantic. They are compared from various points of view: specifications of the quality, relations between size of sample and quality and between sample size and batch size, stringency of inspection, simple and double sampling, range of scales and general aspects. In a final section the author presents the characteristics that he considers desirable in a sampling procedure.

2905: B. G. van den Bos: Investigations on pesticidal phosphorus compounds, II. On the structure of phosphorus compounds derived from 3-amino-1,2,4-triazole (Rec. Trav. chim. Pays-Bas **79**, 836-842, 1960, No. 8).

A previous publication (No. **2883**) described pesticides obtained e.g. by reaction of 3-amino-1,2,4-triazole (or a 5-substituted derivative thereof) with *bis*(dimethylamido) phosphoryl chloride. It is shown in the present publication that the phosphoryl group in these compounds is attached to the triazole ring. The preparation of a number of compounds involved in this investigation (in collaboration with A. J. Visser and K. Wellinga) is described in the experimental part.

2906: H. J. L. Trap and J. M. Stevels: Conventional and invert glasses containing titania. Part 1 (Phys. Chem. Glasses 1, 107-118, 1960, No. 4).

An investigation of some properties of conventional and invert glasses containing titania. The viscosity and thermal expansion of such glasses do not differ from those of glasses which do not contain titania, but their electrical properties do. The phenomena observed are related to two effects: 1) reinforcement of the structure of the glass by Ti^{4+} ions (mainly in conventional glasses), and 2) the formation of sub-microscopic regions rich in TiO_2 (mainly in invert glasses). The Ti^{4+} ions act as network-modifying ions, not as network-forming ions. They give rise to closer packing, and thus to anomalous variation of the dielectric constant. The effect of these ions is reduced by the presence of Pb^{2+} ions. The presence of Ti^{4+} ions gives rise to attractive properties in both conventional and invert glasses. See also Philips tech. Rev. **22**, 300, 1960/61 (No. 9/10).

2907: C. Jouwersma: On the theory of peeling (J. Polymer Sci. 45, 253-255, 1960, No. 145).

A critical discussion of an equation derived by Bikerman for the force needed to pull a glued layer loose from a flat surface. With the aid of improved boundary conditions, an equation which agrees better with the experimental results is derived. The degree to which the glue may be regarded as following Hooke's law is also discussed.

2908: F. J. Janssen: An electronic spirometer (Proc. 2nd int. Conf. on medical electronics, Paris, June 24-27, 1959, pp. 339-340, Iliffe, London 1960).

Brief extract from an address delivered at the above-mentioned congress, dealing with an electronic method of measuring the volume of air displaced during respiration.

2909: W. J. Oosterkamp: Nieuwe mogelijkheden voor de röntgendiagnostiek (J. belge Radiol. 43, 379-385, 1960, No. 4). (Advances in X-ray diagnostic techniques; in Dutch.)

The main idea behind the development of new X-ray equipment in the Philips laboratories is how to get more information with smaller doses. With this aim in mind, the author discusses: an X-ray image intensifier with a screen 23 cm (9 inches) in diameter, a closed-circuit television system for observing the image formed by the image intensifier, and a magnetic memory for recording X-ray images. Although economy in the use of radiation is recognized to be important, this economy must not be at the expense of image quality. Even in the future, full-size pictures will be indispensible where the utmost image quality is required.

2910: H. F. L. Schöler: Biological properties of 9,10-isomeric steroids, I. Progestational activity of $9\beta,10\alpha$-steroids (Acta endocrinol. **35**, 188-196, 1960, No. 2).

The progestative effect of four steroids with the $9\beta,10\alpha$ structure was compared with that of progesterone by means of the Clauberg test. The four compounds investigated, retro-progesterone, 6-dehydro-retro-progesterone, 17α-acetoxy-retro-progesterone and 6-dehydro-17α-acetoxy-retro-progesterone were effective whether administered orally or subcutaneously. The maximum effect (compared with 10 mg progesterone subcutaneously) was obtained by subcutaneous administration of 5, 1, 1 and 0.5 mg respectively, or by oral administration of 50, 10, 2.5 and 1.25 mg respectively. See also these abstracts, Nos **2881** and **2882**.

2911: G. W. van Oosterhout: Morphology of synthetic submicroscopic crystals of α and γFeOOH and of γFe_2O_3 prepared from FeOOH (Acta crystallogr. **13**, 932-935, 1960, No. 11).

The orientation of the needle axis of synthetic acicular crystals of α and γFeOOH with respect to the unit cell has been determined by selected-area electron diffraction. The needle axis is [001] for αFeOOH ($c = 3.03$ Å) and γFeOOH ($c = 3.06$ Å) and [110] for γFe_2O_3 prepared either by dehydration of γFeOOH or by reduction of αFeOOH or γFeOOH followed by oxidation.

The results are compared with previous work on this subject and the possible causes of the discrepancies between the results of Osmond and of Campbell and those obtained in the present paper are discussed.

2912: J. M. Stevels: Network defects in non-crystalline solids (Conf. non-crystalline solids, Alfred (N.Y., U.S.A.), Sept. 1958, pp. 412-448, Wiley, New York 1960).

After a survey of the present theories about the structure of glass, in particular quartz glass, the author discusses the information that can be gained about the structure from measurements of dielectric losses at high and low temperatures and from the effect on the optical absorption spectrum of exposure to ultraviolet light, X-rays, γ-rays and neutrons. The measurements of dielectric losses show promise of allowing us to distinguish between five ways in which ions and electrons can move in glass, quartz glass and quartz. Combination of these measurements, particularly those at low temperature, with measurements of the optical absorption and paramagnetic resonance allow a number of network defects in glass and quartz to be recognized and identified. These methods can also sometimes yield information about local crystallization in glass or local vitrification in quartz crystals. In a number of cases the effect of irradiation can be described by a sort of chemical equation ("operational equation").

2913: J. Rodrigues de Miranda: Photo-sensitive resistor in an overload-preventing arrangement (IRE Trans. on Audio **AU-8**, 137-139, 1960, No. 4).

Description of a device for preventing overloading of an amplifier. The pre-amplified signal is fed via a voltage divider to the input of a high-power amplifier. The voltage divider contains a photo-sensitive resistor (CdS). The output of the amplifier feeds a neon lamp, which is situated together with the photo-resistor in a light-proof box. The neon lamp begins to burn as soon as the voltage exceeds a certain limiting value. The input signal is thus attenuated, making overloading impossible.

2914: G. Diemer: Nature of an ohmic metal-semiconductor contact (Physica **26**, 889, 1960, No. 11).

The author clarifies some assumptions made in a previous publication about the nature of the contact between an indium electrode and a crystal of *N*-type CdS. The indium probably diffuses along dislocations etc. into the CdS, forming "spikes" of much lower resistivity than the surrounding crystal. The current will then spread from these spikes throughout the crystal, giving rise to an ohmic voltage drop near the electrodes.

2915: H. G. van Bueren: The flow stress of germanium crystals (Physica **26**, 997-999, 1960, No. 11).

The author has previously published an expression for the creep velocity of germanium crystals which implicitly involves the velocity *v* of dislocations. The exponential relationship between the creep velocity on the one hand and the tensile stress and temperature on the other can be explained from the way *v* varies with these quantities. This view contradicts the results of Haasen and Alexander to a certain extent. An attempt is made in the present publication to extend the theory so as to eliminate this contradiction.

2916: M. J. Sparnaaij: Electro-osmosis experiments at the germanium/electrolyte interface (Rec. Trav. chim. Pays-Bas **79**, 950-961, 1960, No. 9/10).

Measurements of the electro-osmotic fluid transport in a mixture of Ge powder and an electrolyte. In analogy with a theory of Verwey and Payens (originally developed for surface layers of weakly ionized fatty acids) it is assumed that the surface-oxidized Ge is surrounded by a layer of H_2GeO_3 or some other acid when it is in contact with an aqueous solution. This acid yields up hydrogen ions to the solution, while the $HGeO_3^-$ residues are adsorbed on the semiconductor, giving rise to a surface charge which is responsible for the observed phenomena. The difference in behaviour between *P*-type and *N*-type germanium is mentioned.

2917: M. P. Rappoldt: Studies on vitamin D and related compounds, XIV; investigations on sterols, XVII. The photoisomerization of pre-ergocalciferol and tachysterol$_2$ (Rec. Trav. chim. Pays-Bas **79**, 1012-1021, 1960, No. 9/10).

The quantum yields of the isomerization reactions of pre-ergocalciferol (P) and tachysterol$_2$ (T) in ether at 20 °C under the influence of ultraviolet light of 2537 Å were found to be 0.49 and 0.11 respectively. It had previously been found that irradiation of P gives rise to small quantities of ergosterol (E) and lumisterol$_2$ (L). It is now shown that E is formed directly from P, but not from T, while L is probably formed from T produced by the isomerization of P.

A 32: H. G. Grimmeiss, W. Kischio and A. Rabenau: Über das AlP; Darstellung, elektrische und optische Eigenschaften (Phys. Chem. Solids **16**, 302-309, 1960, No. 3/4). (Preparation and electrical and optical properties of AlP; in German.)

Methods of preparing and doping AlP are described. Reflectivity and transmission measurements indicate that the band gap of this compound is 2.42 eV at 20 °C. Undoped crystals show electroluminescence with maxima at 5550 and 6150 Å. This effect is shown to be due to the recombination of charge carriers injected from *P-N* junctions. The maximum photoluminescence is found at 6100 Å. These results together with conductivity data allow the band structure of undoped AlP to be deduced. AlP crystals also show point-contact rectification and photovoltaic effects. The maximum photoconductivity is found between 5000 and 5150 Å.

A 33: A. Stegherr, F. Wald and P. Eckerlin: Über eine ternäre Phase im System Ag-Sb-Te (Z. Naturf. **16a,** 130-131, 1961, No. 1). (A ternary phase in the system Ag-Sb-Te; in German.)

A ternary phase with the NaCl structure found in the system Ag-Sb-Te is briefly described. The homogeneity range varies considerably with temperature, but it is obvious that this phase cannot be described as "the compound $AgSbTe_2$", as other workers have done. The composition $Ag_{19}Sb_{29}Te_{52}$ seems much nearer the truth. This corresponds very nearly to the formula 2 Ag_2Te.3 Sb_2Te_3; the crystallographic formula can be written $Ag_{0.366}Sb_{0.558}\square_{0.077}Te$.

A 34: A. Rabenau and P. Eckerlin: Compounds in the system Be_3N_2-Si_3N_4 (Special Ceramics, Proc. Symp. Brit. Ceramic Res. Ass., Ed. P. Popper, pp. 136-143, Heywood, London 1960).

The preparation of the substances mentioned in this article requires a special technique which is described and which has a wide range of applications to other investigations. The system Be_3N_2-Si_3N_4 was investigated between 1600 and 2000 °C. There exists a new hexagonal modification of Be_3N_2 which dissolves up to 7 mol. per cent Si_3N_4. Further compounds in the system are Be_4SiN_4 and the wurtzite-type compound $BeSiN_2$.

A 35: R. Groth: Über Ultrarotempfänger auf der Basis von Phosphoren (Z. Naturf. **16a,** 169-172, 1961, No. 2). (Infrared detectors based on phosphors; in German.)

An infrared detector based on infrared-sensitive phosphors is proposed. The principle of such a detector is explained, and experimental data obtained with a SrS-Ce-Sm phosphor are given. It was possible to obtain a sensitivity of 6×10^{-11} W with this phosphor, for wavelengths around 1 micron.

A 36: I. Maak, P. Eckerlin and A. Rabenau: Über $GaF_3.3H_2O$ und andere Trifluoridtrihydrate (Naturwiss. **48,** 218, 1961, No. 7). (GaF_3. $3H_2O$ and other trifluoride-trihydrates; in German.)

A preliminary report of an investigation of the structure of $GaF_3.3H_2O$. This compound gives the same X-ray powder diffraction pattern as α-$AlF_3.3H_2O$, and forms mixed crystals with the latter in all proportions. The lattice constants of $GaF_3.3H_2O$ and of other compounds isomorphous with α-$AlF_3.3H_2O$ and with β-$AlF_3.3H_2O$ are given.

A 37: K. Weiss, P. Fielding and F. A. Kröger: Untersuchungen an kupferdotiertem Wismuttellurid Bi_2Te_3 (Z. phys. Chem. Neue Folge (Frankfurt a.M.) **26,** 145-158, 1960, No. 3/4). (Investigations on copper-doped bismuth telluride; in German.)

An investigation of the influence of copper on the electrical and mechanical properties of P-Bi_2Te_3. The experimental results cannot be satisfactorily explained on the assumption that all the copper occupies interstitial sites. The behaviour of the dissolved copper can better be understood if the layer structure of Bi_2Te_3 and the resulting anisotropy of the host lattice are taken into account.

A 38: A. Klopfer and W. Schmidt: An omegatron mass spectrometer and its characteristics (Vacuum **10,** 363-372, 1960, No. 5).

An omegatron with noble-metal electrodes is described, which can be given a constant sensitivity to within 10 per cent by the application of a suitable electrostatic field. The sensitivity is not even affected by the action of gases and vapours, such as H_2O, CO_2 and CH_4, over long periods, and is reproducible from omegatron to omegatron as long as the dimensions are kept constant. A comparison of the ionization probability given in the literature with that calculated from the calibration curve of the omegatron shows that nearly all the resonance ions formed by the electron beam reach the ion collector. The adjustment of the operating data needed to achieve this condition is in general independent of the mass. Some characteristics of the omegatron are described. (See also Philips tech. Rev. **22,** 195, 1960/61 (No. 6) and **23,** 122, 1961/62 (No. 4).)

°A 39: G. Schuchhardt: Ion movements in an omegatron (Vacuum **10,** 373-376, 1960, No. 5).

The motion of resonant and non-resonant ions in the omegatron described in **A 38** is investigated. It is shown that the auxiliary electrostatic field in this type of omegatron creates conditions which make for optimum collection of the resonant ions and ensure reproducible results. Space-charge effects are estimated, and an expression is derived for the resolving power.

VOLUME 23, 1961/62, No. 11 pp. 325-364 Published 12th September 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

THE NEUTRON TUBE, A SIMPLE AND COMPACT NEUTRON SOURCE

by O. REIFENSCHWEILER *) and K. NIENHUIS **).

621.387:539.125.5

In recent years the rapid expansion of nuclear physics and its practical applications, and the central rôle of the neutron in that branch of physics, have occasioned a wide demand for neutron sources. A great variety of neutron sources and generators exists, but no simple, compact, transportable type generating fast, monochromatic neutrons has been available until now. The present article deals with a sealed neutron tube that fills this gap. In some details of design and operation it resembles an X-ray tube, and it is capable of providing a continuous output of at least 10^8 neutrons per second. It is interesting to note that F. M. Penning, whose name is familiar in connection with the vacuum gauge he invented, built an experimental neutron tube as long ago as 1937, but did not follow up this line of development.

The rapid expansion of nuclear physics is due in great measure to the research into the properties of the neutron and into its interaction with matter that has been conducted all over the world, with more and more elaborate equipment, ever since this elementary particle was first discovered in 1932. So important are the applications of nuclear physics developed in recent decades that a new discipline, nuclear engineering, has come upon the scene; here too, the neutron may be said to take a central position. In a nuclear reactor the neutron is the agent that causes fission and transmits the fission reaction from one nucleus to another and, in virtue of this chain of reactions, liberates energy. Neutrons can be used to produce many of the radioactive isotopes needed as tracer elements in the study of industrial, chemical or biological processes. Crystal structures and the properties of materials can be investigated by means of neutrons, and there are many further uses for these particles.

Inquiry into the properties of neutrons, and experimental work with the aid of neutrons, are therefore important aspects nowadays of research in many fields; this kind of work invariably calls for neutron sources. The present article describes the general make-up and functioning of a sealed-off neutron source which in construction and operation resembles an X-ray tube. The development of the neutron tube in the Philips Research Laboratories [1]) can be regarded as a resumption of the work of F. M. Penning [2]) in 1937; in the last few years the development has reached the stage at which, in co-operation with the Electron Tubes Division, it was ripe for quantity production. Its potential applications are reviewed at the end of the article.

Nuclear reactions yielding neutrons; neutron sources

Since, in general, neutrons exist only as component parts of atomic nuclei they can only be freed by way of nuclear reactions. Essentially, then, neutron sources are devices for staging suitable nuclear reactions. A beam of fast, lightweight nuclei (protons, deuterons or α particles) can be used for this purpose. Nuclei must be brought into close proximity if they are to react. This means overcoming the repulsion between the nuclei due to their electrical charges

*) Philips Research Laboratories, Eindhoven.
**) Electron Tubes Division, Philips, Eindhoven.

[1]) An earlier stage of this work was reported on at the Deutsche Physikertagung 1957, held in Heidelberg. See O. Reifenschweiler and A. C. van Dorsten, Eine abgeschmolzene. Neutronenröhre, Phys. Verh. (Mosbach) 8, 163, 1957. For *in extenso* treatment of all the problems and their solution, see O. Reifenschweiler, Philips Res. Repts 16, 401-418, 1961 (No. 5). See also Nucleonics 18, No. 12, 69-71, 1960.

[2]) F. M. Penning and J. H. A. Moubis, Eine Neutronenröhre ohne Pumpvorrichtung, Physica 4, 1190-1199, 1937.

(Coulomb field); their velocity relative to each other must therefore be high. The relative velocity and hence the kinetic energy the particles must have are dependent on the reaction it is desired to bring about.

Three of the many types of neutron-yielding nuclear reactions may be cited here. Firstly, beryllium enters with an α particle into the reaction ^{9}Be(α, n)^{12}C or, in the more detailed notation:

$$^{9}_{4}\mathrm{Be} + ^{4}_{2}\mathrm{He} \rightarrow ^{12}_{6}\mathrm{C} + ^{1}_{0}\mathrm{n} + 5.65\ \mathrm{MeV}\,. \quad . \quad (1)$$

Secondly, deuterium reacts with deuterium; known as the D-D reaction, this is written D(d,n)^{3}He, or, in the extended notation,

$$^{2}_{1}\mathrm{H} + ^{2}_{1}\mathrm{H} \rightarrow ^{3}_{2}\mathrm{He} + ^{1}_{0}\mathrm{n} + 3.26\ \mathrm{MeV}\,. \quad . \quad (2)$$

Thirdly, the so-called D-T reaction, between deuterium and tritium, is written T(d,n)^{4}He or

$$^{3}_{1}\mathrm{H} + ^{2}_{1}\mathrm{H} \rightarrow ^{4}_{2}\mathrm{He} + ^{1}_{0}\mathrm{n} + 17.6\ \mathrm{MeV}\,. \quad . \quad (3)$$

In order to bring about any of these reactions it is necessary to give one of the partners a high kinetic energy and cause it to collide with the other partner. The energy of the incident particle plus the energy liberated by the reaction, as specified in the above equations, is shared between the two products of the reaction in accordance with the principles of conservation of energy and momentum. The neutron acquires most energy when it is emitted in the forward direction; when emitted in any other direction it acquires a smaller amount dependent on the angle it makes with the direction of motion of the incident particle, and on the initial energy of this particle. It is possible by collimation to obtain a beam of monoenergetic neutrons.

Reaction (1) above can easily be induced: all that is necessary is to bring a radioactive substance emitting alpha particles, radium or polonium for example, into proximity with beryllium. It was in this way that the neutron was discovered in the first place. Accordingly, the simplest neutron sources, which are also the oldest, consist of a mixture of powdered beryllium and (say) radium enclosed in a capsule. Various other combinations of α-emitters and neutron-yielding elements are in use.

Neutron sources embodying radioactive substances are simple in design and small in size, the rate at which they emit neutrons is virtually constant, and they are usually long-lived; nevertheless they have certain drawbacks as compared with other kinds of source. For one thing the neutrons supplied have a complex energy spectrum, and this is often undesirable in experimental work. Some types of radioactive neutron source are associated with a strong background of γ-radiation; a Ra-Be source, for example, emits about 10^4 γ-quanta per neutron. The neutron flux cannot be varied and it is therefore impossible to pulse the neutrons, though in many experiments a pulsed neutron supply is desirable. The source cannot be switched off, and consequently careful shielding is necessary even when it is not in use. The financial outlay climbs so steeply with the power of the source that for fluxes exceeding 10^7 neutrons/sec, other types of source are preferable.

There is no natural particulate radiation suitable for inducing reactions (2) and (3) or other nuclear reactions falling into the same category; the particles constituting one of the two partners in the reaction must therefore be artificially accelerated up to a high velocity. It is on this principle that the tube described here is based. Since 1932, the year in which Cockcroft and Walton did their pioneering experiments with artificially accelerated particles, the technique of using such beams to induce nuclear reactions has been developed to a high level of efficiency. Though exhibiting certain essential

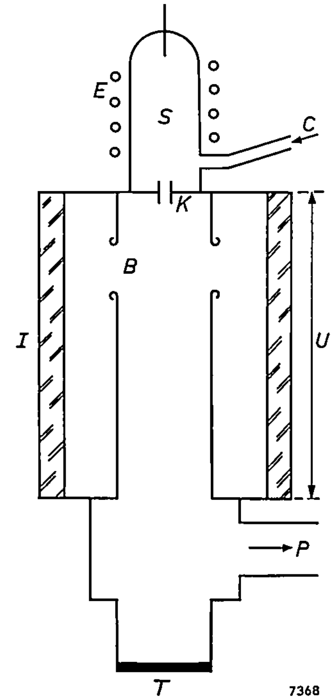


Fig. 1. Principle of a simple particle accelerator. Gas atoms are ionized in the discharge space *S* of an ion source (the source shown here is of the high-frequency type; *E* represents the *RF* coil). The ions leave the source via a small duct *K* and, in the evacuated cylindrical acceleration chamber *B*, they pass through a large difference of potential *U*; with the high energy thus acquired, they strike target *T*. *I* is a high-tension insulator. *P* is the tube through which the whole system is continuously evacuated; a stream of fresh gas is fed to the ion source via intake *C*. By choosing a suitable gas for ionization, and a suitable target material, neutron-yielding nuclear reactions can be caused to take place.

differences, our neutron tube shares many features with a particle accelerator and for that reason it will be as well to devote a little time to the principle of the latter (*fig. 1*). Gas molecules are converted into ions in an ion source *S*. Via *K*, a narrow cylindrical duct, some of the ions enter the evacuated cylindrical acceleration chamber *B* where they fall through a large potential difference *U*; having thus acquired a high kinetic energy, they strike an electrode of suitable material, the target *T* (usually earthed), with which they enter into a neutron-yielding nuclear reaction. The ancillary apparatus includes, apart from a high-tension generator, various devices omitted from fig. 1 for the sake of simplicity — namely, an HF generator for the ion source, with power supply unit, a complete vacuum system embodying backing and high-vacuum pumps and cold traps, and a gas storage and supply system, with a sensitive control valve, for the ion source. Though free of the drawbacks of the type containing radioactive substances, accelerator-type neutron sources are generally very large and expensive installations that often have to be housed in a separate building [3]) (see also the photograph on page 341 of this issue).

In recent years there have been various attempts to build compact and, if possible, transportable versions of these accelerator-type neutron-generating equipments. On the basis of reaction (3) above, these compact generators have become a practical possibility mainly in consequence of the availability of tritium in adequate quantities. Nowadays this isotope of hydrogen can be produced from lithium in nuclear reactors, using the reaction ^{6}Li(n,α)^{3}H. Reaction (3) shows that with tritium as the target nucleus, a fairly high neutron yield is obtained even when the voltage accelerating the bombarding deuterons (i.e. deuterium ions) is relatively low (60 kV for example). The effective cross-section for the D-T reaction exhibits a peak at the unusually low energy of 107 keV. In virtue of all this it has been possible to build neutron generators which, while still rather complicated in design, and necessarily including a complete vacuum system, can be mounted on trolleys and so moved from place to place.

While enumerating the various types of neutron source and generator we must not neglect to mention that nuclear reactors are nowadays the biggest-scale producers of neutrons [4]). The reaction taking place in these is of a different nature from those of reactions (1) - (3), namely nuclear *fission*. As a result of bombardment with neutrons, and also to some extent as a result of slow neutron capture, certain nuclei (^{233}U, ^{235}U and ^{230}Pu) can be caused to split and to release further neutrons, a series of these neutron-yielding reactions thus being initiated. A small proportion of the neutrons can be extracted via a duct in the shield of the reactor without prejudice to the continuity of the process. For many applications requiring a high neutron flux, a better source than a nuclear reactor is hardly likely to be found. Yet such big, complicated and costly installations are of course available only in a limited number; in practice, therefore, their usefulness as neutron sources is limited. Also a transportable neutron source in the form of a nuclear reactor is probably not a practical possibility.

There is a third category of neutron-yielding nuclear reactions which should be mentioned here for the sake of completeness; "photonuclear" reactions occur when nuclei are exposed to high-energy γ-radiation. In neutron sources working on this principle, the γ-radiation is either obtained from a radioactive substance — this type has the same disadvantages as the radioactive sources discussed above — or produced in a large particle accelerator with the aforementioned drawbacks of complicated design and high cost.

It may be noted in passing that scientific resources are currently being lavished on attempts to realize a further theoretical possibility, namely that of using the thermal energy of gases and plasmas at extremely high temperatures in order to induce nuclear *fusion*. A "fusion reactor", if it were ever built, would be a particularly rich source of fast neutrons.

A big advance in the direction of a compact, straightforward, monoenergetic neutron source was made with the development of an accelerator-type neutron generator in the form of a sealed tube; the need for a pumping arrangement and a continuous supply of gas for the ion source thus being eliminated. The Philips neutron tube [5]), based on this principle and under development since 1955, is shown in *fig. 2*; its general construction is shown in *fig. 3*. The tube bears a certain family resemblance to an X-ray tube; it has been possible in virtue of this to make use of accumulated experience in X-ray tube design, particularly in the matter of high-tension insulation and the prevention of arcing. However, in regard to the principle on which it works, the neutron tube is a direct descendant of the conventional particle accelerator (fig. 1). It accordingly displays the same essential structural features — an ion source, accelerating electrodes and a target that, bombarded by the accelerated ions, emits neutrons in virtue of the D-T reaction (3). The yield is in excess of 10^8 neutrons/sec.

[3]) See for example A. C. van Dorsten and J. H. Spaa, A high output D-D neutron generator for biological research, Nucl. Instr. **1**, 259-267, 1957, or Philips tech. Rev. **17**, 109-111, 1955/56.

[4]) A reference to the use of a nuclear reactor as a neutron source may be found in J. Goedkoop, Neutron diffraction, Philips tech. Rev. **23**, 69-82, 1961/62 (No. 3).

[5]) In the meantime neutron generator tubes have been described by other authors: J. D. Gow and H. C. Pollock, Rev. sci. Instr. **31**, 235-240, 1960. P. O. Hawkins and R. W. Sutton, Rev. sci. Instr. **31**, 241-248, 1960. A. H. Frentrop and H. Sherman, Nucleonics **18**, No. 12, 72-74, 1960. B. J. Carr, Nucleonics **18**, No. 12, 75-76, 1960.

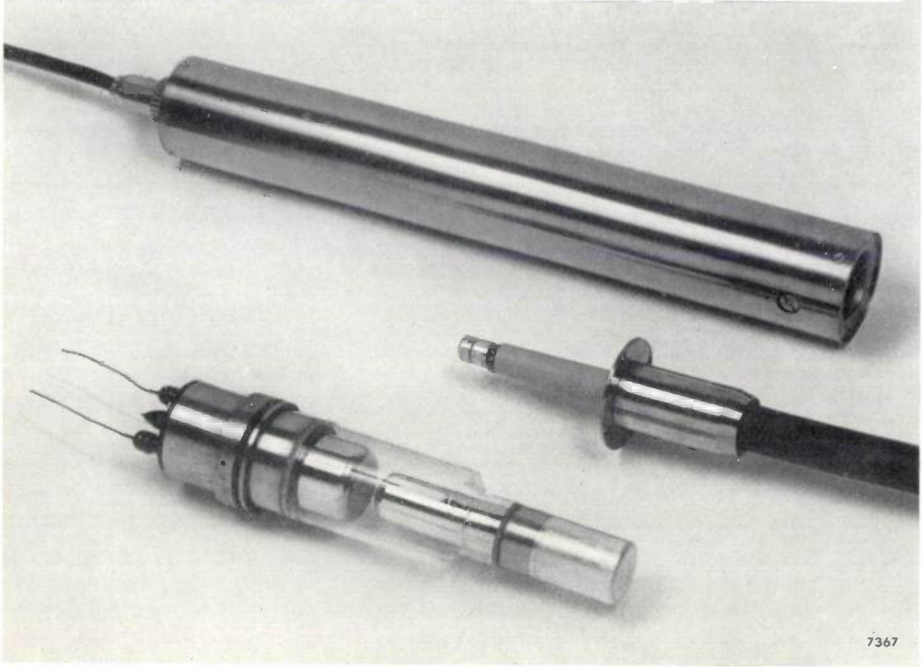


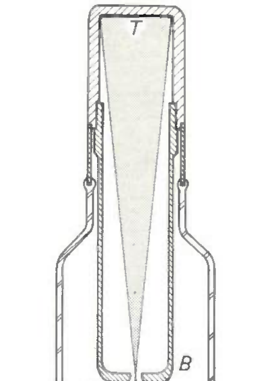
Fig. 2. The Philips neutron tube (at bottom in photo). The left-hand end of the tube, containing the ion source and a hydrogen pressure regulator, is at earth potential. High tension is applied to the right-hand end, via a HT cable and plug (centre). The tube as delivered (top) is enclosed in a metal cylinder with an outer diameter of 7 cm. This cylindrical sheath, which is earthed, affords mechanical protection and holds the HT insulation in place around the tube; it does not present any appreciable obstacle to the neutrons issuing from the target. These are expelled in all directions, but normal practice is to use only those with directions perpendicular to the tube axis.

One minor difference from a particle accelerator is that the high tension is applied to the target and the ion source is earthed; this arrangement offers certain operational advantages.

In the next section we shall deal with three main problems presented by the design of the neutron tube and explain how we have solved them.

The neutron tube: fundamental problems and their solution

A characteristic of accelerator-type neutron sources of the type shown in fig. 1 is that the particles to be accelerated are produced in a chamber in which a relatively high pressure prevails (of the



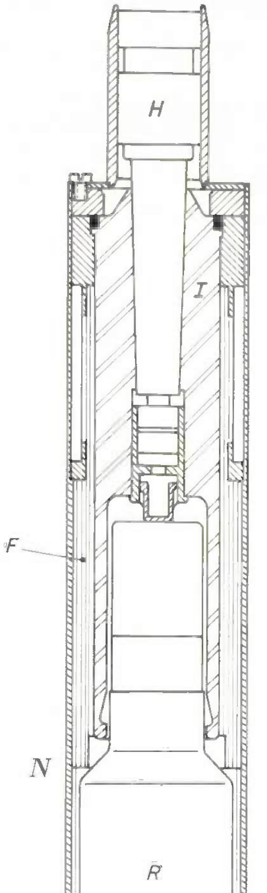


Fig. 3. *a*) Constructional details of the Philips neutron tube. *S* Penning ion source, at earth potential, also acting as one electrode of the accelerating stage. *B* high-tension electrode, to which an accelerating voltage of 125 kV is applied. *D* feed-through insulators for supply voltages to ion source (2 to 3 kV) and hydrogen pressure regulator *P* (0 to 2 V). *G* soft iron walls of ion source chamber. *M* permanent magnet. K_1, K_2 are the disc cathodes and *A* the cylindrical anode of the Penning ion source. *T* target. The ion beam is shaded; its shape is determined by the properties of the electron-optical lens formed by *G* and *B*.
b) Section through the neutron tube *R* enclosed in its earthed metal sheath *N*. *I* Araldite insulator. *F* oil-impregnated insutaing foil. *H* plug of HT cable.

order of 10^{-2} torr) whereas acceleration takes place in a space under relatively low pressure (of the order of 10^{-5} torr). The pressure difference between ion source and acceleration space is maintained by continuously feeding gas into the former (via inlet C) and continuously evacuating the latter by means of a high-vacuum pump. Now, in a *sealed* neutron tube this continuous gas flow must be dispensed with, and so in consequence must the pressure difference between the ion source and the accelerating space.

Thus the first problem involved by a sealed neutron tube is that of designing an ion source and an accelerating system that will work at the same gas pressure. If an intermediate pressure value is decided on, then the gas discharge in the ion source must sustain itself at a relatively low pressure and at a voltage that is not unduly high; the implication for the accelerating stage is that *no* discharge or flashover must take place between the accelerating electrodes despite the high voltages these are required to carry, and despite the relatively high pressure prevailing.

A second problem, likewise connected with the absence of a stream of gas through the tube, is that of gas clean-up as a result of the discharge. In a sealed tube this effect involves a rapid fall-off in pressure. Some expedient must therefore be found for keeping the gas pressure during operation constant throughout its useful life.

A third problem inherent in a sealed neutron tube is that of target life. Tritium targets of the type commonly used hitherto (in tubes exploiting the D-T reaction) have only a limited life, shorter than is generally desirable in a sealed tube. Obviously there can be no question, in such a tube, of replacing the target at intervals, as is possible in pumped neutron generators.

The ion source

The first of the problems referred to above — that of finding an ion source that will work at low pressures — can be solved by recourse to a device known for twenty-five years. It was precisely the fact that the Penning ion source could function at a relatively low pressure that led its inventor to design a sealed neutron tube [2]). As in other ion sources, the ions are produced in a gas discharge, i.e. mainly by the impact of electrons. For this it is necessary that the mean free path of the electrons should be smaller than the distance available for them to travel through. Now, in other ion sources this distance is of the same order as the dimensions of the discharge chamber, and consequently the mean free path must be small and, generally speaking, the pressure must be in excess of 10^{-2} torr; the Penning ion source is so designed that the average distance travelled by the electrons is many hundred times greater than the electrode separation. This is achieved by using a special electrode layout and introducing a magnetic field. In these circumstances, then, there is still plenty of chance of ion formation by electrons with a large mean free path; it is therefore possible for the source to work at a lower pressure (10^{-3} to 10^{-4} torr).

A further virtue of the Penning ion source is its simplicity. No heated filamentary cathode is used: the discharge is sustained by electrons released from cathode plates by ion impact and by the photo-electric effect; accordingly, only one DC supply of about 2 kV is required for the ion source.

Fig. 4 is a schematic cross-section through the Penning type ion source incorporated in our neutron tube [1]). A cylindrical anode A lies between two disc-shaped cathodes K_1 and K_2. A direct voltage of 2 to 3 kV is applied between the anode and the cathodes. In the resulting discharge, electrons oscillate between the cathodes many times before finally striking the anode, being forced into helical paths by the magnetic field. The ions created by collisions between electrons and gas atoms move towards the cathodes. Some are able to quit the ion source chamber through an opening in cathode K_1. The mag-

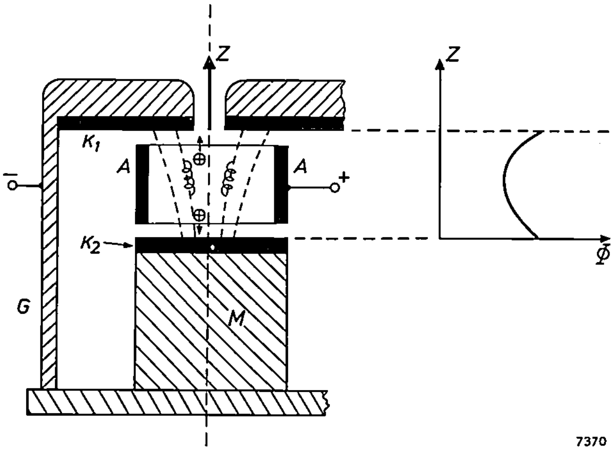


Fig. 4. Schematic cross-section of the ion source incorporated in the Philips neutron tube. G walls of ion source chamber. A is the anode, K_1 and K_2 the cathodes of the Penning electrode system. M permanent magnet. The broken lines represent magnetic lines of force; the magnetic circuit is completed by the soft iron walls G of the ion source chamber. Z direction of ion beam leaving the chamber. Φ is the electrical potential as a function of the distance z along the axis of the tube; the electrons sustaining the discharge swing back and forth in the resulting potential well between the cathodes (see diagram) and are forced by the magnetic field into helical paths. The distance travelled by the electrons is thus many times greater than the geometrical electrode separation. It is for this reason that the discharge can sustain itself at gas pressures as low as 10^{-3} torr.

netic field is brought about by a small permanent magnet M which is accommodated inside the chamber. The walls of the chamber G are of soft iron, and so serve to complete the magnetic circuit and to screen the neighbouring acceleration space from the magnetic field in the ion source. This is important from the standpoint of freedom from high-voltage breakdown in the acceleration space, since a magnetic field increases the chance of ionization by electrons and so lowers the striking voltage for a self-sustaining discharge. This fact is exploited in the ion source itself, but just the opposite desiderata are relevant to the acceleration space.

The Penning ion source has a certain drawback in that it mainly supplies singly-charged molecular ions (H_2^+ with hydrogen as filler gas, D_2^+ with deuterium) in the range of pressures of interest. This we discovered at the outset, in 1956, when the ion beam was submitted to mass-spectrometric analysis in our Research Laboratories (*fig.* 5). For a given discharge current, the number of atoms in a beam of singly-ionized diatomic molecules is twice as large as that in a beam of monatomic ions. Moreover, each of the two atoms in the ionized molecule acquires only half of the total energy, so that in a neutron tube operated at 125 kV, only 62.5 keV of primary energy is available for the nuclear reaction. Since in this range of energies the probability of the reaction taking place diminishes sharply with decreasing voltage, the final result is that the neutron yield is lower than it would be with a beam of monatomic ions. This drawback of the Penning ion source has been accepted in view of the far greater importance attached to its advantages — long life, robustness, and simplicity of construction and operation.

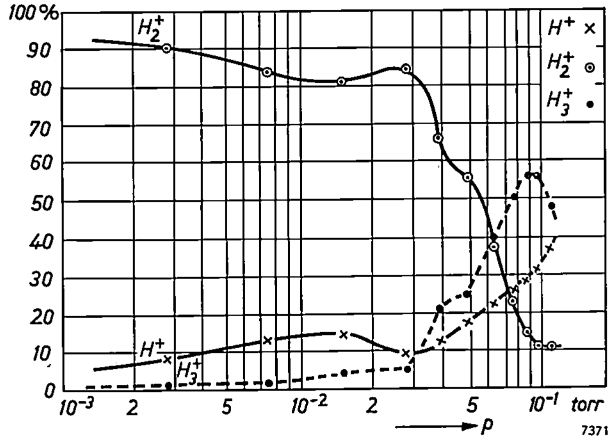


Fig. 5. Composition of the beam of hydrogen ions produced by a Penning ion source, as a function of pressure. In the range between 10^{-3} and 3×10^{-2} torr the beam mainly consists of molecular (H_2^+) ions.

Arrangements for feeding in the direct-voltage supply are greatly simplified by the fact that the ion source is at earth potential.

The 125-kV acceleration stage

Conventional accelerators working in conjunction with a pump operate at pressures lower than 10^{-4} torr, and acceleration generally takes place in several stages. Preliminary experiments [1]) revealed that a 200 kV acceleration system, to operate at pressures up to 2×10^{-2} torr, could be designed as a single stage. The electrode separation in these experiments was about 1 cm. Accordingly, no design difficulties would be involved by a neutron tube operating at a pressure of 10^{-3} torr and an accelerating voltage of 125 kV. It would, however, be necessary to give the electrodes a high polish, since otherwise there would be a risk of field emission from the negative electrode. Accelerating electrodes for the neutron tube described here are made of chrome steel which, as experience with X-ray tubes has shown, exhibits very little liability to field emission.

As will be clear from the sketches in fig. 3, accelerating electrode B has an opening through which the ion beam enters the field-free space inside the electrode. The beam is conical in shape, its apex angle being determined by the properties of the electron-optical lens formed by the ion source and accelerating electrode. *Fig. 6* is a photograph of the ion beam in an experimental tube. In the final version of the neutron tube, as developed on the basis of findings from these experiments, the target is set up at a distance from the ion source such that its whole surface is just covered by the ion beam (fig. 3*a*). This results in efficient loading of the target and hence in a better neutron yield. If the target were placed only 1 cm in front of the ion source, the narrow ion beam would strike only a restricted area of the target surface, with undesirable local overheating as a result. The tubular form of the accelerating electrode has the further advantage of allowing it to catch most of the secondary electrons leaving the target, thereby preventing their acceleration back into the ion source. If allowed to travel through the whole acceleration space the secondary electrons would lower the breakdown voltage in consequence of the ionization they would cause; they would also be responsible for unnecessary heating of the walls of the ion source, and the unwanted heat would be conveyed by conduction to other parts of the tube; finally, the secondary electrons would generate X-radiation of the bremsstrahlung type, which would be undesirable in many applications of the neutron tube.

The hydrogen pressure regulator

The second problem, that of regulating and stabilizing the gas pressure inside the tube, calls for a device whereby additional filler gas can be supplied and, if necessary, withdrawn again.

If in the course of the manufacturing process the tube were filled with gas at the desired pressure and then sealed, this pressure would be liable to fall off rapidly when the tube was taken into operation, in consequence of gas clean-up in the discharge. This loss is mainly a result of gas ions being shot into the electrodes and the walls of the discharge chamber.

For replenishing the gas supply to the ion source, conventional accelerator-type neutron generators embody a valve, e.g. a small electrically-heated tube of palladium or nickel, through which hydrogen isotopes readily diffuse at higher temperatures; such devices are unsuitable for a sealed tube because they can only be used to bring about a *rise* in pressure. It would accordingly be impossible — by this means alone — to compensate the spontaneous pressure rises that occur, for example, if previously occluded filler gas is released in consequence of normal heating of the tube in operation.

An efficient hydrogen pressure regulating device (hydrogen replenisher) can be based on the fact that certain metals absorb hydrogen exothermically. Amongst these are zirconium and titanium, which are able to take up large quantities of hydrogen or hydrogen isotopes, evolving heat at the same time, and release them when heat is applied from the outside. Care must however be taken to ensure that the surface of the metal offers no obstruction to the passage of the gas, i.e. that it is physically and chemically clean.

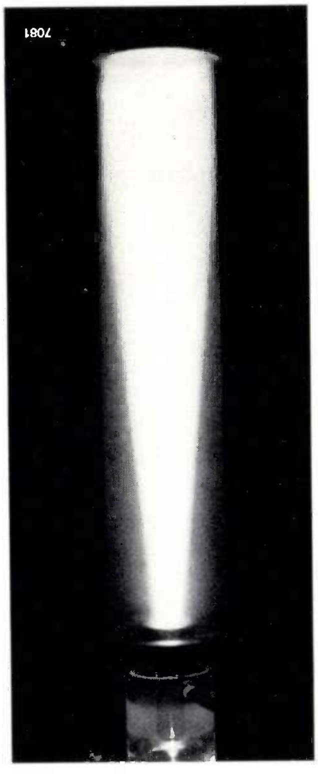


Fig. 6. Photograph of the ion beam in an experimental version of the neutron tube, which was expressly built up mainly of glass in order that the beam could be studied.

If a metal-hydrogen system of this kind is placed in a vacuum-tight vessel, the gas pressure will attain an equilibrium value p which is related to the absolute temperature T by the Clausius-Clapeyron formula:

$$\ln p = -\frac{\Delta H}{RT} + C$$

where R is the gas constant, ΔH is the heat of reaction, and C is a constant of integration. Thus the equilibrium pressure increases with temperature and, given a metal-hydrogen system whose temperature can be varied, we shall have, in principle, a device for regulating the pressure of the hydrogen. Generally speaking, an increase in the number of atoms per metal atom, such as may be occasioned by release of gas from other parts of the tube, will result in a rise in pressure that has to be compensated by lowering the temperature of the system. However, there is a range within which the equilibrium pressure is *not* dependent on the ratio between the number of hydrogen atoms and the number of metal atoms present. This useful fact can be explained in terms of a phase transformation, and in our tube advantage is taken of it to keep the gas pressure constant.

In its simplest form, the pressure regulator consists of a zirconium wire wound for the sake of mechanical stability around a tungsten wire [6]), the whole being built into the neutron tube (P in fig. 3). The zirconium wire is charged with gas before the tube is sealed, and thereafter its temperature can be conveniently and reproducibly adjusted by passing an electric current through it. In operation, then, the pressure inside the tube can be adjusted to the desired value by altering the voltage across the ends of the wire (see *fig.* 7).

Like the ion source, the pressure regulator is at earth potential, so that no difficulty is involved by feed-in arrangements for the heater current.

[6]) K. Nienhuis, Hydrogen-filled electric discharge devices, U.S. Patent 2 766 397. This pressure regulator was first used in hydrogen thyratrons, and allowed their useful life to be increased to several thousands of hours. See also Philips tech. Rev. **20**, 102, 1958/59 (fig. 2).

A drawback of this design is that the rate at which hydrogen is absorbed is rather slow, particularly at low pressures, in consequence of the inevitable poisoning of the zirconium wire, the superficial area of which is comparatively small. To get a system with a faster response it is necessary to increase the area available for absorbing hydrogen. This has been done by employing the absorbing metal in the form of an extremely fine powder deposited by evaporation on the surface of a different metal [7]). Instead of zirconium, titanium is used as absorbing metal. The pressure regulator in our neutron tube consists of a cylinder made of thin nickel sheet with a heater wire mounted inside it. Around the heater is wound a titanium wire which is evaporated, not in vacuum, but in argon at a pressure of about 10 torr, in consequence of which it deposits itself on the inside of the cylinder as a kind of soot. The deposit is composed of spherical grains which are about 100 Å in diameter, and therefore much smaller than could ever be obtained by mechanical crushing of the titanium. This fine powder will absorb large quantities of hydrogen and its isotopes very quickly [8]), and quickly release the gas when heated. In operation, the wire originally used for evaporating the titanium serves as a heating element for the pressure regulator. For example, to obtain a deuterium pressure of 10^{-3} torr inside the tube, the temperature of the titanium deuteride must be maintained at about 260 °C.

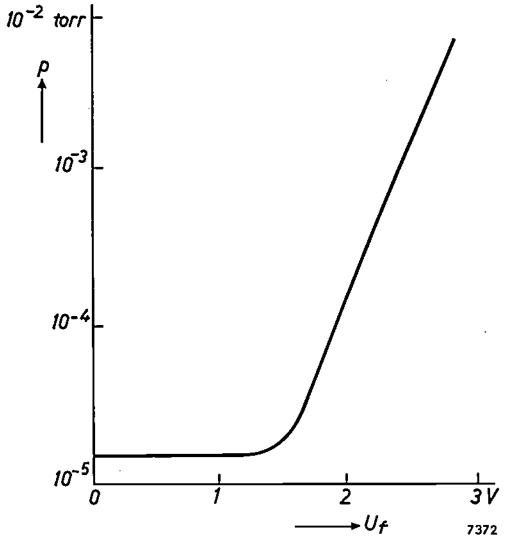


Fig. 7. Pressure p in the neutron tube, as a function of the voltage U_f across the filament of the hydrogen pressure regulator.

The target

Only one main component of the neutron tube remains to be dealt with, namely the target. The output and useful life of the tube are mainly governed by the target quality. Our 'neutron tube exploits the D-T reaction, of which mention has several times been made in the foregoing and which was numbered (3) above; under the primary energy conditions designed into the tube, this reaction yields neutrons with an energy of about 14 MeV at an angle of 90° to the primary beam. To bring about the D-T reaction, deuterons must be shot into a *tritium* target, and since tritium is a gas, some way must be found of occluding it or anchoring it to a solid body. Tritium targets suitable for conventional accelerator-type neutron generators are commercially available, and consist for example of a metal disc 0.2 mm thick on to which has been evaporated a film of zirconium or titanium about 1 μm thick saturated with tritium. Our experiments showed that titanium-coated targets had a higher neutron yield than zirconium-coated ones, and were better able to stand up to high temperatures. We found that they could be freed from impurities by degassing at temperatures as high as 200 °C and brazed on to the target support without any loss of tritium. However, these commercial targets have two drawbacks that render them less suitable for use in a sealed neutron tube: firstly degassing cannot be done at temperatures above 200 °C, and secondly, experience with conventional neutron generators indicates that the targets may have a life of only a few hundred hours. Ways and means were therefore sought of circumventing this limitation on the life of the tube [1]).

One way out of the difficulty was offered by the "drive-in" types of self-replenishing target that have several times been described in the literature [9]). If a plate of metal such as gold is bombarded with deuterons with energies of several hundreds of keV, the particles penetrate the metal and spread through it by diffusion. In this way a deuterium target is gradually formed; deuterons subsequently striking the plate enter into the D-D reaction, producing neutrons. In an experiment of this kind, then, it is

[7]) O. Reifenschweiler, Ein Druckregler für Wasserstoffisotope mit grosser Einstellgeschwindigkeit des Gleichgewichtsdrucks, Phys. Verh. (Mosbach) Verbandsausgabe, 1961, 181 (No. 9).

[8]) For many purposes the gaseous radioactive hydrogen isotope tritium can be handled more conveniently if it is absorbed in this way and the titanium powder is prepared as a suspension. See. O. Reifenschweiler, A suitable tritium carrier for gas discharge tubes, Proc. 2nd United Nations Intern. Conf. Peaceful Uses Atomic Energy **19**, 360-362, Sept. 1958, Geneva.

[9]) Self-replenishing targets exploiting the D-D reaction were first investigated thoroughly by K.Fiebiger, Die Bildung von „Selbsttargets" für die Kernreaktion D(d,n)^{3}He und ihr Zusammenhang mit dem Problem der Wasserstoffdiffusion in Metallen, Z. angew. Physik **9**, 213-223, 1957.

very soon found that neutrons are being emitted, and the rate of emission increases as time goes on. Generally, after a certain lapse of time, the yield attains a saturation value. The reason is that, by this time, a certain proportion of the deuterons shot into the metal are diffusing back to the surface and out of the plate. A steady state of saturation is attained when the deuterons diffuse out of the target at the same rate as they are shot into it.

The proportion of deuterons "used up" by nuclear reactions is extremely small, so small that it has no effect on the constitution of a self-replenishing target. Fiebiger's experiments [9]) indicated that the highest saturation concentration of deuterons and the greatest neutron yield was obtainable with a gold self-replenishing target.

So far we have dealt only with the self-replenishing target for the D-D reaction. One exploiting the D-T reaction can be produced by filling the neutron tube with a mixture of tritium and deuterium and using it to bombard the target. As a result, tritium as well as deuterium penetrates the target material, and enters into reaction with the incident deuterons. True, it also happens under these circumstances that deuterium and tritium ions strike deuterium atoms and that tritium ions strike tritium atoms, in consequence of which the neutron yield is only a third of that available when a pure tritium target is bombarded with deuterons. Measurements and approximate calculations have shown that, of the above competing reactions, only the D-T and T-D reactions — which naturally yield neutrons of the same energy — make any appreciable contribution to the output of the tube.

The reasoning underlying the above statements is as follows. Let us assume that, in a tube filled with deuterium and tritium in equal proportions, the ion beam and the gas absorbed into the target are also 1:1 mixtures of the two isotopes. The chances of a D-T, a T-D, a D-D and a T-T impact are in each case 25% of the probability that a D-T impact will take place in a tube with a pure tritium target and a beam composed exclusively of deuterons, other conditions (beam current, gas content of target etc.) being the same. The output from a tube of the latter kind may amount to 5×10^8 neutrons/sec; if this is regarded as a 100% yield, then the yield of 14-MeV neutrons from the D-T reaction in the tube filled with a gas mixture will only be 25%.

In regard to the contribution of the T-D reaction to the neutron yield of the tube, it must be remembered that the tritium ion has a mass 50% greater than that of a deuteron; consequently, the kinetic energy of the tritium ions in the beam, measured in relation to the centre of gravity of the system formed by the two reacting particles, is only two-thirds that of deuterons accelerated through the same potential difference.

The energy available for a nuclear reaction is equal to the kinetic energy of the reacting particles due to their motion relative to their common centre of gravity; their movement relative to an outside frame of reference is irrelevant. Consequently 60-keV tritium ions are only equivalent to 40-keV deuterons. On account of the smaller effective cross-section of the D-T reaction at 40 keV, its neutron yield at this primary energy is only a quarter of what it would be at a primary energy of 60 keV. This latter, as we have already established, would be 25% of the yield from a tritium-target deuteron-beam tube. The actual yield of 14 MeV neutrons from the T-D reaction is therefore about 6%.

The contribution of the D-D reaction can be worked out from the measured yield from tubes filled with pure deuterium. These have been found, under the same operating conditions, to produce 5×10^5 neutrons/sec. In the mixture-filled tube only a quarter of the reactions are D-D reactions. Accordingly, these will contribute 1.25×10^5 neutrons (of 2.5 MeV) per second, which represents less than 0.1% of the overall yield from a "pure" D-T tube.

Finally, in regard to impacts of tritium ions on tritium atoms, the chance of a nuclear reaction resulting from these is so slight at the accelerating voltages employed that it can be neglected altogether.

On adding the above contributions together we arrive at a figure of about 31%. Accordingly, a neutron tube filled with a deuterium-tritium mixture has a yield about one-third of that from a tube with a pure tritium target and a beam composed exclusively of deuterons. Of this overall yield, the D-T reaction is responsible for four-fifths and the T-D reaction for one-fifth, both these reactions yielding 14-MeV neutrons. The D-D reaction, yielding 2.5-MeV neutrons, contributes less than 0.1% to the tube output.

The above considerations still fail to take full account of actual conditions in the tube, since the particles accelerated are mainly molecular ions, not monatomic ones. During absorption into and desorption from the hydrogen pressure regulator, the formation of D-T molecules may occur, so that we really have three kinds of molecule to deal with; in the ideal case these would be present in the proportions 1:2:1. However, the result of a more elaborate study taking account of all such factors would not differ appreciably from the figure just arrived at.

The use of drive-in targets removes a limitation on the tube life, which ceases to be dependent on that of the target. A further advantage is that the temperature at which the tube can be degassed and evacuated ceases to be governed by the thermal properties of the target, since this is not charged with gas until after the tube has been sealed. Still, as has already been shown, the usual type of drive-in target has a smaller neutron yield than the targets commercially available, and for that reason we cast about for a better type. As will now be explained, we have succeeded in designing drive-in targets that have an appreciably higher neutron yield.

Gold, which Fiebiger's experiments indicated was the best material for drive-in targets, does not naturally occlude hydrogen isotopes. In fact, in the target experiments, it is *forced* to absorb the isotopes, these being shot into it at high velocity;

thereafter, they distribute themselves through the metal by diffusion. The diffusion coefficient of hydrogen isotopes in gold is small but, of course, increases with temperature. Consequently it is only in a superficial layer of the target heated by ion bombardment (the thickness of which corresponds to the maximum range of the ions penetrating the surface) that distribution of the hydrogen isotopes is at all efficient, by reason of the rather higher diffusion coefficient of this heated layer. Hence the highest concentration obtainable within the layer penetrated by incident ions — the only part of the target where nuclear reactions are likely to take place — is governed by the rate at which the ions are diffusing out of the layer via the target surface; diffusion into the depth of the target, *away* from the surface, can be neglected. This theory of Fiebiger's was confirmed by his experiments on gold targets [9]). The above is outlined in visual terms in *fig. 8a*.

Unlike gold, the metals zirconium and titanium are able to go on absorbing hydrogen isotopes exothermically right up to the point of hydride formation, a property exploited in the pressure regulator. This means that throughout the charging process, i.e. until the target has become saturated, there is a potential barrier at its surface, hindering the escape of hydrogen isotopes from the metal. It is therefore easier to achieve and maintain a high concentration of hydrogen isotopes in titanium or zirconium than in gold, provided only that some means can be found of preventing too many of the hydrogen atoms from escaping at the *other* side, beyond the limit of the reaction zone, as they are liable to do in consequence of the high diffusion coefficients of zirconium and titanium. This is made clear in fig. 8*b*. It is obvious that in Fiebiger's experiments on drive-in targets, zirconium and titanium only gave poorer neutron yields than gold because the ions were escaping at the back door, so to speak.

Diffusion out of the reaction zone can be prevented by using the target metal in the form of a thin layer applied to a base with a low diffusion coefficient for hydrogen. The thickness of the layer must more or less correspond to the range of penetration of the

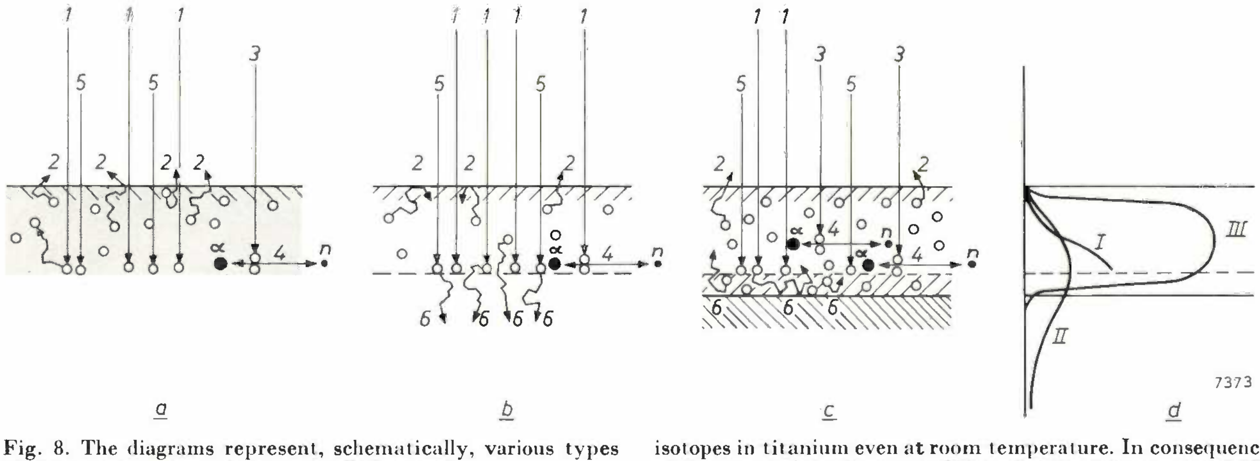


Fig. 8. The diagrams represent, schematically, various types of drive-in targets exploiting the D-D reaction.
a) Fiebiger's gold target [9]). Gold is a non-occluder of hydrogen isotopes: it does however absorb deuterium when bombarded with deuterons (*1*) of sufficient energy to effect penetration to some depth. In a thin layer of the target (grey area in the diagram) the deuteron concentration gradually attains a steady state as the nett result of a number of processes — penetration of the deuterons, their diffusion through the metal (only at high temperature), heating of the metal by the deuteron bombardment, and escape of the particles by diffusion out of the target (*2*). Deuterons present in the target may enter into neutron-yielding nuclear reactions (*4*) with newly arriving deuterons (*3*) or, alternatively, these latter may merely contribute to maintaining the concentration in the target (*5*). Curve *I* in diagram *d* shows the presumed concentration distribution of deuterium in a saturated drive-in target of gold [9]).
b) Solid titanium target. The numerals *1* to *5* have the same significance as in *a*. Unlike gold, titanium is an exothermic absorber of hydrogen isotopes, absorbing deuterium very readily up to the point of TiD_2 formation; a high saturation concentration ought therefore to be attainable in a titanium target. However, the incident deuterons cannot build up this high concentration because of the high diffusion coefficient of hydrogen isotopes in titanium even at room temperature. In consequence of this most of the deuterons diffuse further into the metal (*6*), out of the reaction zone. (The particles have to overcome a potential barrier before emerging from the target surface (*2*) and this effect is therefore unimportant.) Curve *II* in diagram *d* shows the presumed concentration distribution of deuterium in a titanium target; the neutron yield from such a target is accordingly smaller than in case *a*.
c) This diagram represents the backed titanium-film drive-in target employed in the neutron tube [1]). The thickness of the film is rather greater than the maximum range of the incident ions (*1*) in titanium and the backing metal is one that has a low diffusion coefficient for hydrogen. It is impossible for the particles to diffuse further into the depth of this target (*6*) and consequently the concentration of deuterons obtainable in the titanium film is very high. Curve *III* of diagram *d* shows the presumed distribution of deuterium densities. The neutron yield is three times as great as in case *a*.
Diagrams *a*, *b* and *c* are of course applicable to the D-T reaction as well as to the D-D reaction. In fact only a proportion of the small circles are to be regarded as deuterons; the others represent tritons. In addition to neutrons, *α*-particles are produced as a result of the D-T reaction (as shown); in the D-D reaction, however, ^{3}He nuclei are produced instead of the *α*-particles.

ions. On bombardment, the layer becomes charged up to saturation in a comparatively short time (about 15 hours at a beam current of 100 μA); the saturation concentration of hydrogen isotope is high (see fig. 8*c*). The present neutron tube is equipped with a target of this kind, consisting of an approximately 1 μm film of titanium evaporated on to a silver base. For an acceleration voltage of 125 kV and an ion current of 100 μA, a yield of over 10^8 neutrons/sec is obtained. Measured yields from the best of these targets have attained the figure of 2.4×10^8 neutrons/sec.

Apart from providing a roughly three times higher neutron yield, the titanium-film target has the advantage over the gold drive-in target of being better able to stand up to elevated temperatures, remaining stable up to 200 °C. Fiebiger found that the neutron yield from gold targets fell off steeply above 120 °C.

The neutron tube in operation

The necessary ancillary equipment includes a high tension supply (e.g. a Greinacher cascade generator) supplying 125 kV AC at currents ranging from 100 to 200 μA, and a power unit supplying 2 to 3 kV DC at 0.3 to 1.0 mA for the ion source, and 0 to 2 V AC at currents up to 5 A for the hydrogen pressure regulator (*fig. 9*). The current through the neutron tube is measured with a microammeter inserted in the HT lead, for example. Of course, part of this current is due to electrons liberated from the target and accelerated in the reverse direction, back to the ion source. The true ion current can be determined by calorimetric methods. An alternative method is to measure the X-radiation generated by the backward-accelerated electrons. Measurements using both methods revealed that ions are responsible for about 80% of the tube current.

The neutron yield can be measured with any kind of radiation detector responsive to neutrons; some detectors will require preliminary calibration. We have carried out an exact measurement by an activation method in which calibration is not necessary. The reaction ^{63}Cu(n,2n)^{62}Cu was induced by neutrons from the tube in a copper disc set up in the vicinity of the tube target. The ^{62}Cu thus produced is a positron emitter with a half-life of about 10 minutes. Consequently only a limited amount of ^{62}Cu is formed, determined by the balance between the rates of formation and decay of the isotope. Having measured the β^+-activity of the irradiated target (with an end-window counter tube), and knowing the effective cross-section of the (n,2n) reaction with the ^{63}Cu, it is possible to work out the neutron yield of the tube. The activation method was checked by the method of associated particles: In the D-T reaction an α-particle is liberated at the same time as the neutron and is emitted in the opposite direction (see reaction (3) above). The number of α-particles produced, as registered by an end-window counter tube, for example, is the same as the number of neutrons striking and activating a copper disc set up opposite the counter, provided the geometry of the arrangement is exactly symmetrical. There was good agreement between our results from the two methods — calculation on the basis of the effective cross-section and measurement of the associated α-particle production.

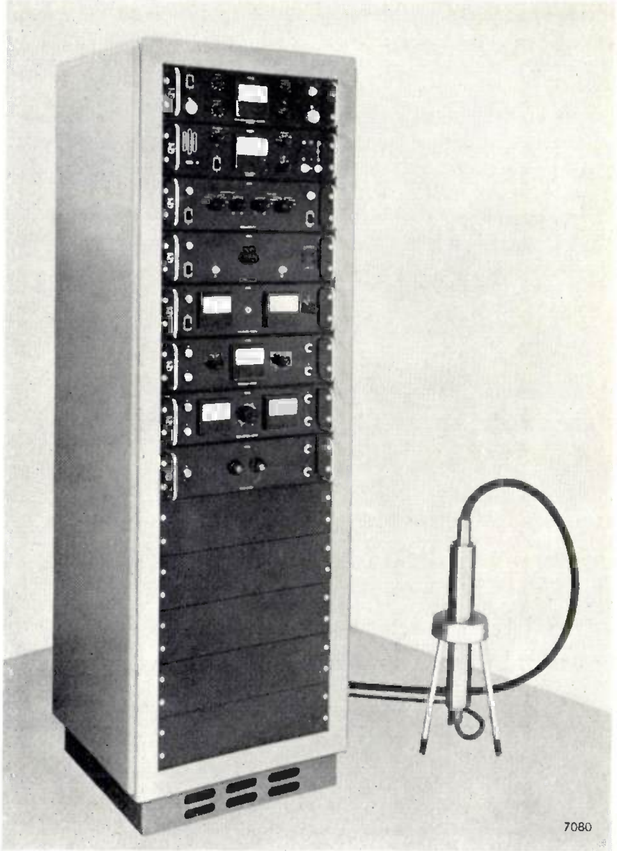


Fig. 9. The Philips neutron tube in its cylindrical metal sheath, complete with power supply. The rear of the lower part of the rack is occupied by the high-tension generator; there is room in front of this for five units for measuring and other purposes. Immediately above are the four units necessary to the operation of the neutron tube; the lowest is a power pack and the other three, taken in ascending order, contain control and monitoring circuits for the HT supply, the hydrogen pressure regulator and the ion source. Accommodated in the rack above these are the pulse generator necessary for pulsed operation of the tube (two units) and counter circuits embodying a count-rate meter, for measuring and checking the neutron flux (two units).

Further experiments were done to determine the neutron yield from a saturated Ti-film drive-in target as a function of accelerating voltage. In addition, neutron yields under various conditions were calculated from the effective cross-section of the D-T reaction and the atomic stopping power of titanium tritide, for different ion beam compositions and for two different targets, one with a tritium-titanium ratio of 1:1 and the other with a deuterium-tritium-titanium ratio of 1:1:2. The results are displayed in *fig. 10*. It will be noted that the measured curve lies very close to the relevant calculated one. At an accelerating voltage of 125 kV and an ion current of 100 μA — under which operating conditions the tube is very stable — a yield of more than 10^8 neutrons/sec is obtained. Fig. 10 also shows that the employment of a pure tritium target and a pure deuterium atmosphere would increase the neutron yield by a factor of 3; however, as has already been explained, the life of the tube would then be limited to a few hundred hours.

The drive-in target actually embodied in the tube has an almost unlimited life. Several tubes taken from the production line have been life-tested under the above conditions (accelerating voltage of 125 kV at ion currents of 100 or 200 μA, corresponding to neutron yields exceeding 10^8 and 2×10^8 neutrons/sec respectively) for periods exceeding 1000 hours; one tube has been run for 4000 hours. No fall-off in reliability of operation or in neutron yield could be detected either during or after these tests.

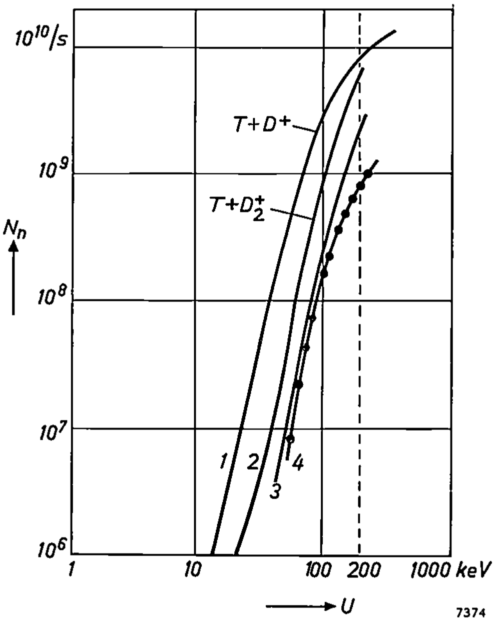


Fig. 10. The yield N_n in neutrons per second at a tube current of 100 μA, as a function of the accelerating voltage U. Curves *1*, *2* and *3* are theoretical, and curve *4* is experimental. Curve *1* has been plotted from calculated yields based on the effective cross-section of the D-T reaction and on the atomic stopping power of the titanium tritride in a target in which titanium and tritium are present in the ratio of 1:1. Furthermore, the ion beam has been assumed to be composed exclusively of monatomic D^+ ions (deuterons). Curve *2* relates to the case where the ion beam is mainly composed of molecular D_2^+ ions; neutron yields are only one-third of those obtained in the case to which curve *1* relates. Curve *3* shows calculated neutron yields in the case where the filler gas, and hence the ion beam, is a 1:1 mixture of deuterium and tritium (molecular ions) and the target is a titanium-film drive-in type, all other conditions remaining the same. These are the conditions appropriate to the final version of the neutron tube, and there is good agreement between this theoretical curve and curve *4*, which is a plot of values obtained from actual measurements.

The neutron yield can be adjusted to any desired value between zero and maximum by varying the accelerating voltage; an alternative method of controlling the tube output is by adjusting the gas pressure in the tube or the voltage across the ion source. It is also possible to stabilize the neutron yield; generally it suffices to keep the ion source current constant, and this can be done automatically by a feedback arrangement whereby this current controls the voltage across the hydrogen pressure regulator. About 10^{-3} torr has been found to be the most favourable gas pressure.

A pulsed neutron output is desirable in many applications of the tube; this can easily be effected by pulsing the ion source voltage. The minimum usable pulse duration is 5 μs. During a pulse, the neutron yield is about 10 times greater than it is under continuous operation; hence, if the pulse occupies 10% of the cycle (i.e. intervals 9 times the pulse length) the mean neutron output will be roughly the same as under continuous operation.

Applications of sealed-off neutron tubes

Finally, let us take a look at a few typical applications of neutron tubes. It must be made clear from the outset that these sources are *not* designed to produce the less common elementary particles that are the speciality of extremely large accelerators. The usfulness of the neutron tube described here lies rather in the fact that it is a particularly simple, convenient and comparatively cheap source of fast, monoenergetic neutrons, its main advantage over conventional accelerators being that it is readily transportable.

Fundamental research in nuclear physics, and particularly in neutron physics, immediately suggests itself as the most obvious field of application for neutron tubes in those cases where a yield 10^8 neutrons/sec is sufficient. The tubes are accordingly being used for investigating the elastic and inelastic scattering of fast neutrons, for studying slow neutron capture, for producing radioactive isotopes and investigating their properties, and for calibrating neutron spectrometers [10]), Wilson chambers, bubble chambers and nuclear emulsions.

[10]) L. J. de Vries and F. Udo, A fast pulse-shape discriminator with applications as a spectrometer and sensitive monitor for 1-30 MeV neutrons, Nucl. Instr. and Meth. **13**, 153-160, 1961 (No. 2).

A further broad field of application for neutron tubes arises out of the employment of the neutron in fields outside nuclear physics for the solution of both purely scientific and more immediately practical problems. For example, the tubes are convenient for producing small quantities of short-lived artificial radioactive isotopes on the spot. This facility is of value in chemical and biological research, and it is relevant to many of the problems with which engineers are confronted. It is particularly in connection with the production of *short-lived* isotopes that the advantages of the neutron tube become evident: normally a nuclear reactor is not available at the place where the isotopes are to be used, and the time required for transportation prohibits irradiation of the material in a reactor some distance away.

A further, related field of application is that of non-destructive analysis. Here a distinction must be made between two methods — activation analysis, and analysis by means of directly induced γ-emission. In activation analysis the sample is irradiated with fast or slow neutrons (cf. the method for measuring neutron yield described above). By investigating the resulting radioactivity of the sample — the half-life and/or the nature of the emitted radiation — the presence of various chemical elements can be determined quantitatively often with great accuracy. Almost all the elements can be identified in this way, though in some cases it may be necessary to undertake separation first or treat the sample afterwards, using ordinary chemical methods.

In the second method of non-destructive analysis, a γ-spectrometer is employed to investigate the γ-radiation generated during neutron bombardment of the sample by inelastic scattering of fast neutrons or by slow neutron capture (n,γ reaction). Analysis of the resulting γ-spectrum is thus analogous to X-ray spectrometry.

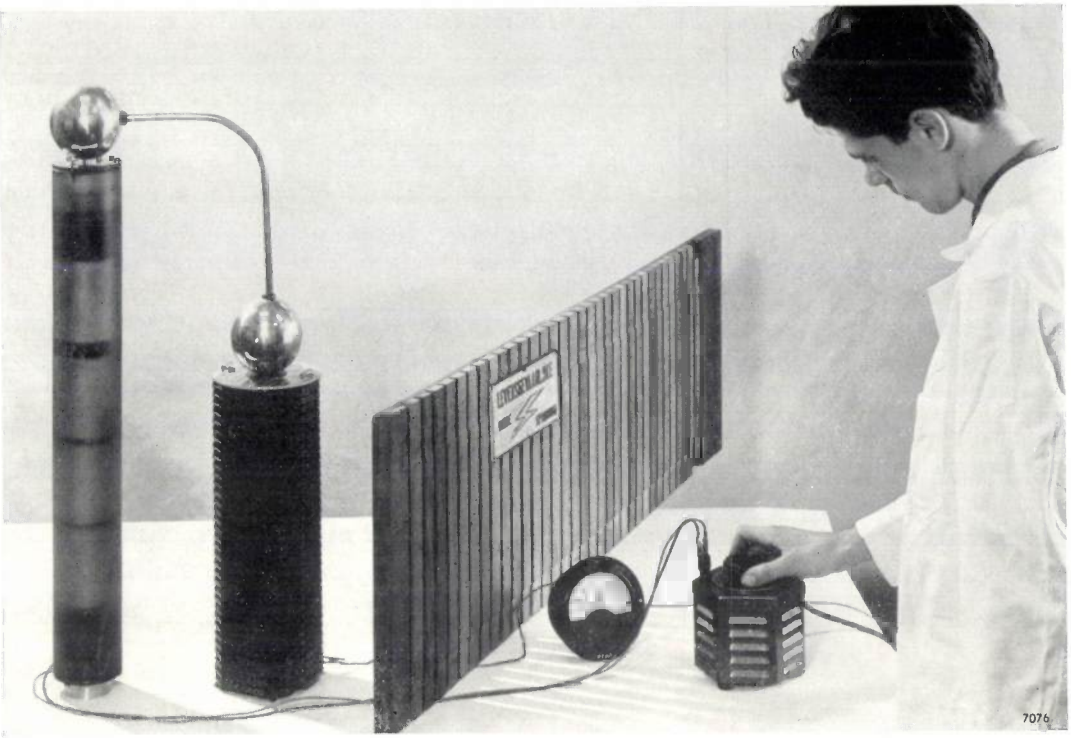
Both methods of analysis are highly suitable for geological investigations, the mobility of the neutron tube making it possible to assay mineral samples in the field. A typical application of great practical importance is the detection of petroleum in the vicinity of drillings ("oil-well logging"). Together with a suitable detector, the neutron source is lowered into the drill hole. In principle the activation method of analysis just described allows all geologically important elements to be identified. Methods based on the recording of γ-spectra seem to offer even better prospects, in view of the fact that γ-radiation can be detected more easily and more efficiently. A particularly useful aspect of these methods is the possibility of exploiting the delay between the γ-radiation generated by fast neutrons (i.e. due to inelastic scattering of the 14 MeV neutrons coming direct from the tube) and the γ-radiation generated by thermal neutrons (those slowed down by elastic scattering). This enables the elements carbon and oxygen to be differentiated more clearly from hydrogen, silicon and chlorine and, accordingly, salt or fresh water to be readily distinguished from mineral oil. Methods exploiting these time differences naturally call for a pulse-operated source of fast neutrons [11]).

The study of reactor materials represents another important field of application for neutron tubes with facilities for pulsed operation. Certain properties of reactor materials, important from the viewpoint of neutron physics, can be measured in sub-critical rigs by injecting the materials with a brief, intense burst of neutrons. The tubes are also convenient for testing neutron-shielding materials.

Of no less importance is the employment of the neutron tube in education and technical training. The tube is simple to operate, it can be adjusted to supply any desired flux from zero to the maximum, and it should therefore represent an ideal neutron source for demonstrations in the lecture theatre and for practical work in the laboratory, opening up the interesting field of artificial radioactivity as well as that of neutron physics proper.

[11]) R. L. Caldwell and W. R. Mills Jr., Nucl. Instr. andMeth. **5**, 312, 1959; J. Tittman and W. B. Nelligan, Amer. Inst. Mining Engrs Casper (Wyoming) meeting, April 1959, Paper 1227 G; B. G. Erozolimskii, A. S. Shkol'nikov and A. I. Isakov, Atomic Energy (Moscow) **9**, 144, 1960.

Summary. After a brief review of current types of neutron sources, the authors describe a neutron source tube which is about the size of an X-ray tube, and is operated in a similar way. Supplied with a direct voltage of 125 kV, this tube is capable of generating neutrons with an energy of 14 MeV at rates exceeding 10^8 neutrons/sec. The neutrons arise out of a nuclear reaction between tritium and deuterium: T(d,n)^{4}He. The deuterons are produced in a Penning ion source and accelerated up to 125 keV in a single-stage accelerating system. They strike a target consisting of a 1 μm film of titanium which has been evaporated on to a silver base, and which contains tritium. The tube is filled with a deuterium-tritium mixture at a pressure of about 10^{-3} torr; consequently the target is bombarded with tritium ions as well as deuterium ions, and in this way its charge of tritium is kept at saturation more or less indefinitely. The life of the tube is not therefore limited by that of the target. The pressure inside the tube is adjusted by means of a built-in replenisher containing a large reserve of D-T mixture, so enabling gas clean-up to be compensated and removing a further restriction on the life of the tube. In life tests, a number of tubes have worked for periods exceeding 1000 hours (one for 4000 hours) without any fall-off in operating reliability or neutron yield. The yield from the tube can attain 10^9 neutrons/sec when it is pulse-operated (the shortest pulse duration is 5μs). In conclusion, some typical applications for this type of neutron tube are touched upon.



AN EXPERIMENTAL HIGH-TENSION GENERATOR OF VERY SMALL DIMENSIONS

by H. P. J. BREKOO *) and A. VERHOEFF **).

621.314.54

The high DC voltages needed e.g. in nuclear physics for particle accelerators or for the supply of the neutron tubes described in the previous article are produced with the aid of either a Van de Graaff generator or a Greinacher cascade generator. In the latter, the voltage delivered by a high-tension transformer is rectified and multiplied by means of a suitable arrangement of rectifiers and capacitors in series and in parallel with one another [1]) (*fig. 1a*). The cascade generator can deliver considerably higher currents than the Van de Graaff generator, and moreover has the advantage of containing no moving parts which are subject to mechanical wear.

The conventional construction of cascade generators follows directly from the circuit of fig. 1*a*, and consists of two columns containing high-tension capacitors, interconnected as shown by rectifiers (tubes or semiconductor diodes). The spacing between the two columns and between successive stages of the cascade must be sufficient to preclude flashover and to keep brush-discharge losses within reasonable limits. As a consequence, the dimensions of a cascade generator for e.g. 150 kV are so large that as a rule a separate room is needed for it. In order to increase the attainable voltage or to reduce the dimensions, modern high-tension generators are frequently built in a pressurized vessel filled with a gas which diminishes the brush-discharge losses [2]).

A very compact cascade generator can be built if use is made of modern insulating materials and of the technique of "potting" or encapsulation in resin. The construction of the experimental cascade generator for 100 kV and 200 μA shown above becomes clear when the Greinacher circuit of fig. 1*a* is redrawn as in fig. 1*b*. The generator components are then distributed over *three* columns [2]). Space is saved

*) Philips Research Laboratories, Eindhoven.

**) Formerly of Philips Research Laboratories, Eindhoven.

[1]) See also A. Kuntke, Philips tech. Rev. **2**, 161-164, 1937; T. Douma and H. P. J. Brekoo, ibid. **11**, 123-128, 1949/50 and A. C. van Dorsten, ibid. **17**, 109-111, 1955/56.

[2]) See also the third of the articles cited under [1]).

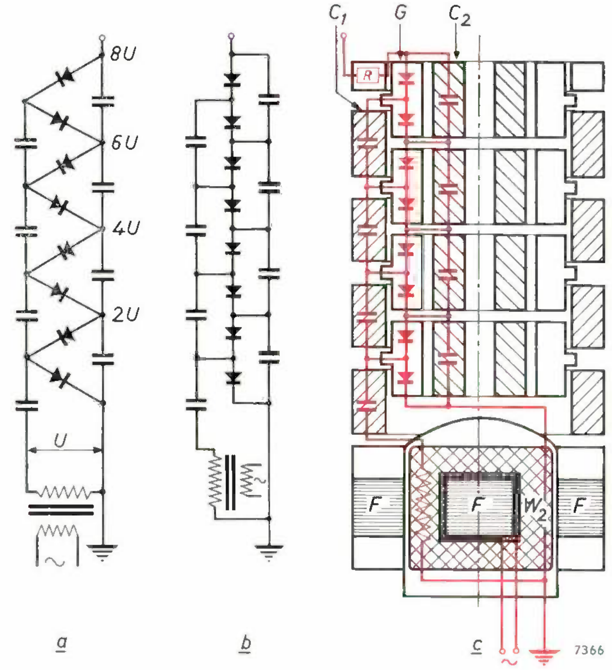


Fig. 1. *a*) Greinacher circuit for a cascade generator. The alternating voltage (of peak value U) is rectified and multiplied by a suitable combination of capacitors and rectifiers (tubes or semiconductor diodes); in the case shown here, the DC output voltage has a value of $8U$.
b) The Greinacher circuit drawn in a somewhat different way; the components are now distributed over *three* columns.
c) Sketch showing the construction of the miniature cascade generator shown in the title picture. Two tubular columns containing capacitors (C_1 and C_2) and one containing rectifiers (G) are fitted coaxially into one another. W_2 high-tension winding of the transformer. F iron core. R damping resistor. In order to make it easier to recognize the individual components, the Greinacher circuit as drawn in fig. 1*b* is also shown here, in red. The insulation is provided by insulating foil and by encapsulation in cast resin; the whole generator is moreover built into a tube filled with transformer oil.

by building these three columns as hollow cylinders and fitting them coaxially, as shown in the sketch of fig. 1*c* and the photo of *fig. 2*.

The capacitors consist of rolled-up sheets of aluminium foil and high-tension insulation foil, and are impregnated with transformer oil during the winding process. The rectifier unit (shown more clearly in *fig. 3*) consists of 2×175 silicon diodes of type OA 202 in a helical series arrangement, embedded in resin. The high-tension transformer is adapted to the cylindrical form of the other components, its laminations being circular. The windings of this transformer are also encapsulated in resin. The entire four-stage generator is finally placed in a PVC tube filled with transformer oil. The annular space left above the outer capacitor of the last stage of the cascade contains a damping resistor of about 100 kΩ (R in fig. 1*c*). This resistor is built up of a helical arrangement of smaller resistors encapsulated in cast resin, like the rectifier unit shown in fig. 3. The entire cascade generator without its earthed metal sheath is 80 cm long and has a diameter of 9.6 cm. It is designed for operation on a mains frequency of 50 c/s. The same principle can of course be used for generators for higher frequencies; the dimensions can then be even smaller, because of the smaller capacitances needed.

Most casting resins on the market have excellent insulating properties, but poor thermal conductivity. The permissible loading of high-tension generators insulated with casting resin is limited by this factor.

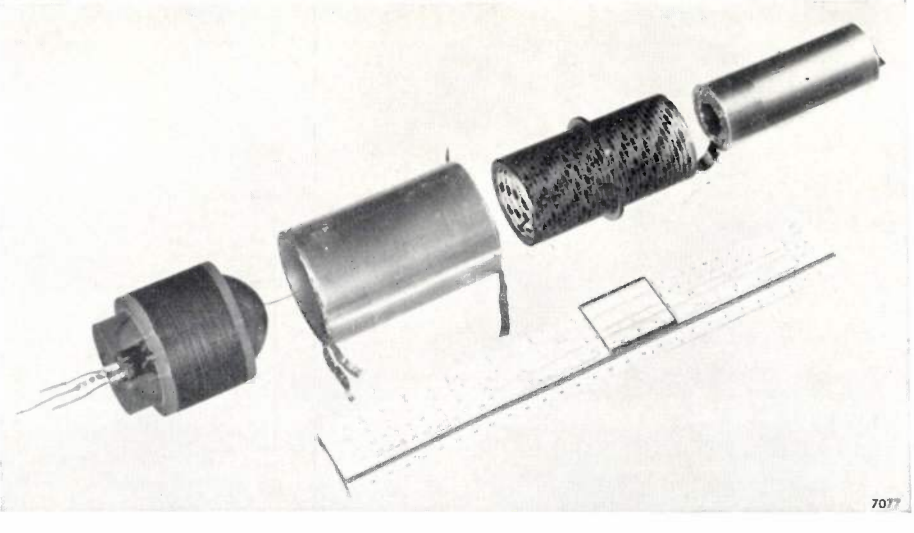


Fig. 2. The first stage of the high-tension cascade generator before assembly. The transformer, which is encapsulated in resin, is designed for 12.5 kV and 50 c/s; the output voltage of the first stage is then 25 kV DC.

A continuous load of 20 W, i.e. 200 μA at 100 kV, is however permissible in the present case.

The generators we have built deliver a high tension of positive polarity. A negative polarity can however easily be obtained by placing the transformer and the first of the outer capacitors at the other end of the column. (The damping resistor must then also be shifted.)

The title picture shows the generator in use, connected to a load resistor.

Miniature high-tension generators of this type may be used for many purposes in research and industry. A particularly important application is in combination with the neutron tube described in the previous article for oil-well logging (testing for the presence of oil-bearing strata in deep borings). The diameter of the high-tension generator must then be matched to that of the neutron tube so that they can both be lowered into the bore-hole. The use of long high-tension cables would then be obviated. We have so far reached an external diameter of 10 cm (*fig. 4*); the object is to reduce the diameter to 7 cm, i.e. that of the neutron tube in its definitive form.

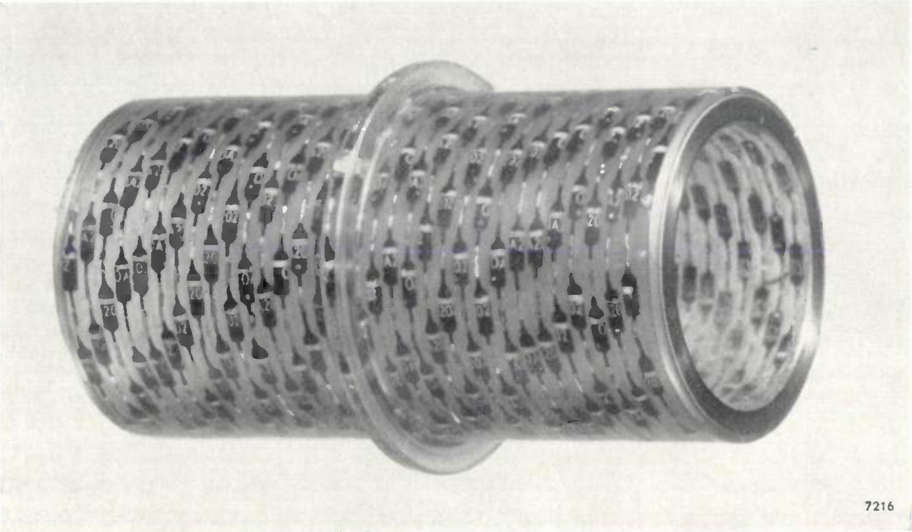


Fig. 3. A rectifier unit of the cascade generator. 2× 175 silicon diodes type OA 202 are arranged along a helix and are encapsulated in resin for high-tension insulation.

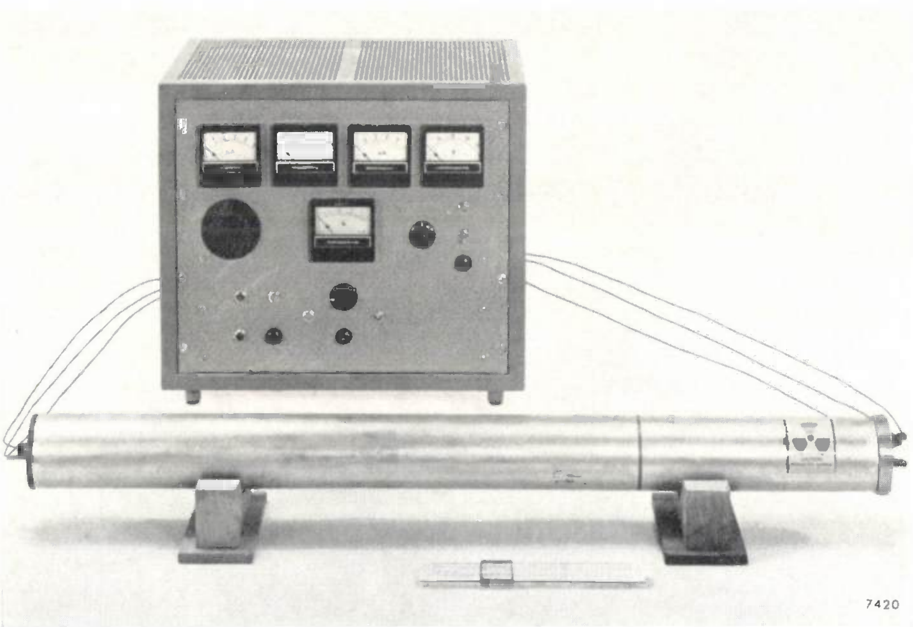


Fig. 4. Combination of the experimental high-tension generator and the neutron tube described in the previous article by Reifenschweiler and Nienhuis; both are built into an insulating tube with an earthed sheath (diameter 10 cm), the cascade generator on the left and the neutron tube on the right. The line drawn on the tube shows the position of the target; the radiation-hazard sign refers only to the tritium in the neutron tube. The ancillary current supply and control equipment may be seen in the background.

Such a combination would be very useful when the neutron tube is used for oil-well logging, since it would obviate the use of a high-tension cable.

Summary. Description of a four-stage Greinacher cascade generator, for 100 kV and 200 μA, built in the shape of a cylinder. It has a diameter of 9.6 cm (with earthed sheath, 10 cm), and a length of 80 cm. The compact construction is made possible by dividing the generator into *three* columns, which are given a tubular form and fitted coaxially. The insulation is provided partly by insulating foil, partly by encapsulation in resin. The generator is designed for use with a mains frequency of 50 c/s; for higher frequencies, an even smaller generator can be constructed on the same principle. Attempts are being made to reduce the diameter still further, to that of the Philips neutron tube (7 cm), so that the two together can be lowered into bore-holes for the purposes of oil-well logging.

ADJUSTING THE ION BEAM IN A PARTICLE ACCELERATOR

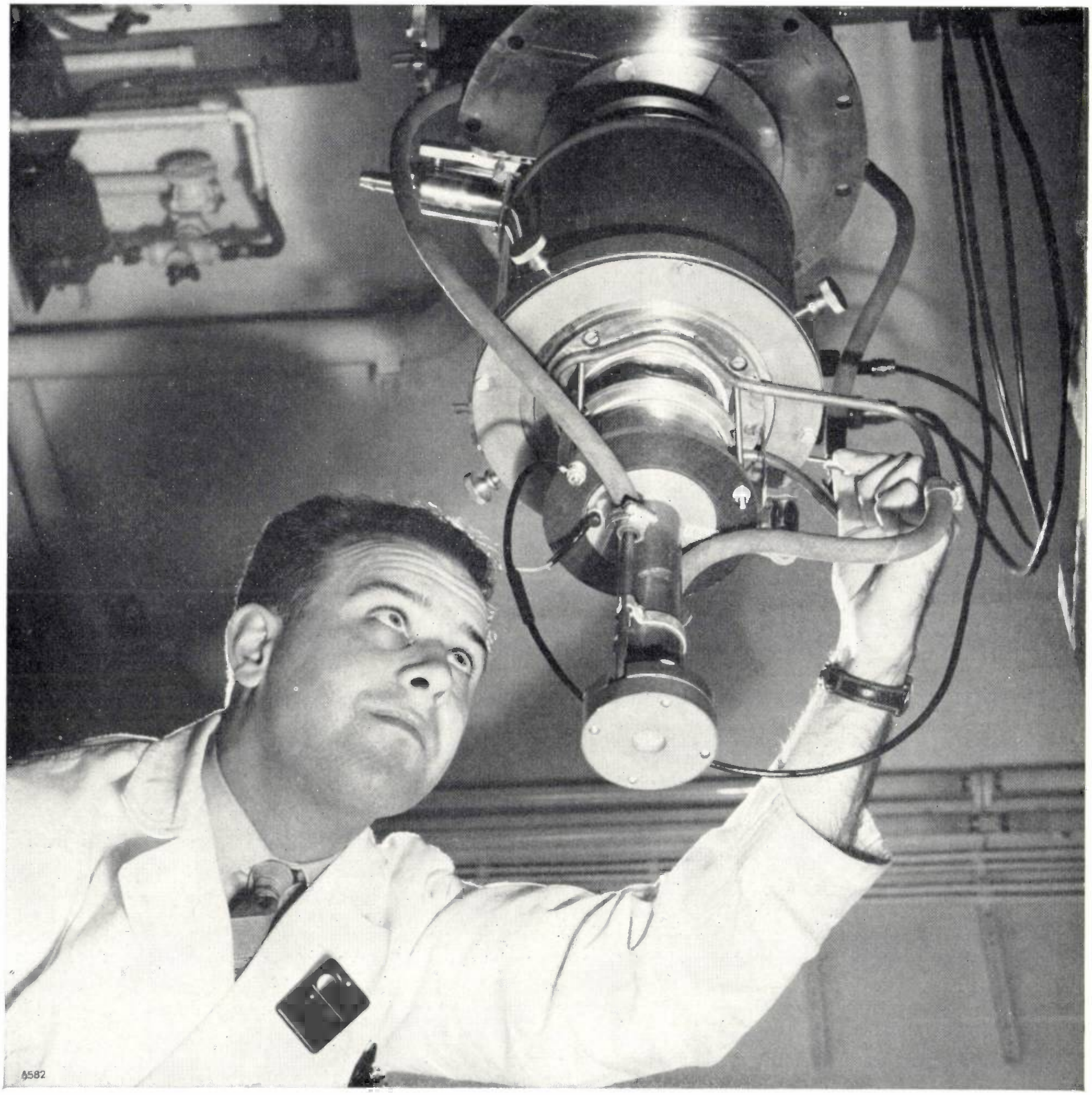


Photo Maurice Broomfield

In nuclear research "large" accelerators are used, such as cyclotrons and synchrocyclotrons, and also various "small" types, capable of accelerating particles up to an energy of 1 MeV. In these small accelerators the target section, which should be readily accessible, is usually located in a room below that containing the high-tension generator. The photograph shows the target section of an accelerator of this kind (cascade generator in a pressure tank), made by Philips for an institute of nuclear physics. If the ions are accelerated with a relatively low voltage, it is permissible to remain near the target, and the luminous beam of ions can then be seen through a glass insulator. By introducing into the beam a quartz plate attached to a spindle, as is being done here, the area struck by the beam is brilliantly illuminated. The target can then be accurately positioned with respect to the ion beam by means of four adjusting screws (three of which are visible on the photograph). A vacuum lock above the adjustable target section makes it possible to admit air into this section and to change the target whilst maintaining the vacuum in the accelerator itself.

The simple water-cooled target shown here, which serves for measuring the target current, can be replaced by, for example, a rotating target of heavy ice, cooled with liquid air. When bombarded with deuterons of 800 keV this target is able to deliver up to 10^{10} neutrons per second [1]), produced by the reaction D(d,n)^{3}He.

[1]) A. C. van Dorsten and J. H. Spaa, A high-output D-D neutron generator for biological research, Nucl. Instr. 1, 259-267, 1957.

MANUFACTURE AND TESTING OF ENAMELLED WIRE

by R. J. H. ALINK *), H. J. PEL *) and B. W. SPEEKMAN **).

621.315.337

Enamelled wire, which is used for winding coils, is at present one of the basic materials of the electrical industry. The present article gives a short survey of its manufacture, methods of testing and of the problems connected with the chemical and physical structure of the insulating lacquer.

One of the basic materials of the electrical industry is insulated copper wire, used for winding coils, transformers, rotors, stators etc. Originally copper wire surrounded by e.g. rubber or cotton was used for this purpose, but nowadays enamelled copper wire (more accurately "enamel lacquered wire") is used almost exclusively. The great advantage of insulating with enamel is that a very thin layer of the insulator is sufficient, so that coils with a relatively high "space factor" (i.e. coils in which a relatively large part of the volume is taken up by copper) can be obtained.

The variety of demands made on enamelled wire in the electrical industry has led to the development of a number of types differing in the chemical composition of the enamel layer. Pope's factory in Venlo (Netherlands) makes mainly the following four types:

a) Oil-laquer enamelled wire, the oldest type of enamelled wire [1]);
b) "Povin" wire, which is very strong;
c) "Posyn" wire, which can be easily soldered: the layer of enamel is readily removed when the wire is dipped into molten solder;
d) "Potermo" wire, which is resistant to quite high temperatures (up to about 155 °C).

All the above-mentioned types of wire are manufactured in more or less the same way, starting from chemically different lacquers. This article starts with a brief description of the manufacturing process. We shall then consider the demands which are made on enamelled wire, and the methods of testing whether these demands are met with. After this section, which deals with the practical side of the matter, follows a theoretical discussion of the relationship between the most important properties of the various types of wire and the structure of their insulating layer. Finally we shall pay special attention to a recent investigation which illustrates the presence and the danger of microscopic "flaws" in the enamel layer.

Manufacture

The manufacture of good enamelled wire entails applying the lacquer as *thinly* as possible to the wire, as otherwise it collects in drops and is thus unevenly distributed over the wire. It is therefore necessary to apply a number of coats of lacquer one after the other to get the required insulating thickness, which may be e.g. 16 μ for a copper wire 0.1 mm in diameter. The usual procedure is to pass the wire a number of times alternately through a lacquer reservoir and a muffle furnace: when it passes through the lacquer reservoir, the wire is covered with a thin coat of liquid lacquer; it is then dried ("muffled") in the furnace, returned to the reservoir for a new coat of lacquer, and so on, until the enamel layer has the desired thickness.

The machines which are used for this process may mainly be divided into two types: those in which the wire is led *vertically* through the furnace, and those in which it is led *horizontally*. Machines of these two types are shown in *figs 1* and *2* respectively.

Certain characteristic differences between the two types may be seen from these figures. The following will help to clarify these differences.

The force of gravity makes a horizontal wire lie in a slight curve (a "catenary"), which has undesirable consequences for the manufacture. The furnace of the horizontal machine must therefore be relatively short. Vertical machines, where this difficulty does not arise, can use furnaces 10 metres or more in length. The short horizontal machines have the advantage that the wire can be threaded through by one man, while the attractive point of the vertical machines is that, for

*) Philips Research Laboratory, Eindhoven. — Our late colleague Dr. Alink, who died in 1959, had as part of his work carried out a series of investigations on enamelled wire. The results of these investigations form part of the material of this article.

**) Pope's Wire and Lamp Factory, Venlo, Netherlands.

[1]) See e.g. J. Hoekstra, Philips tech. Rev. 3, 40, 1938.

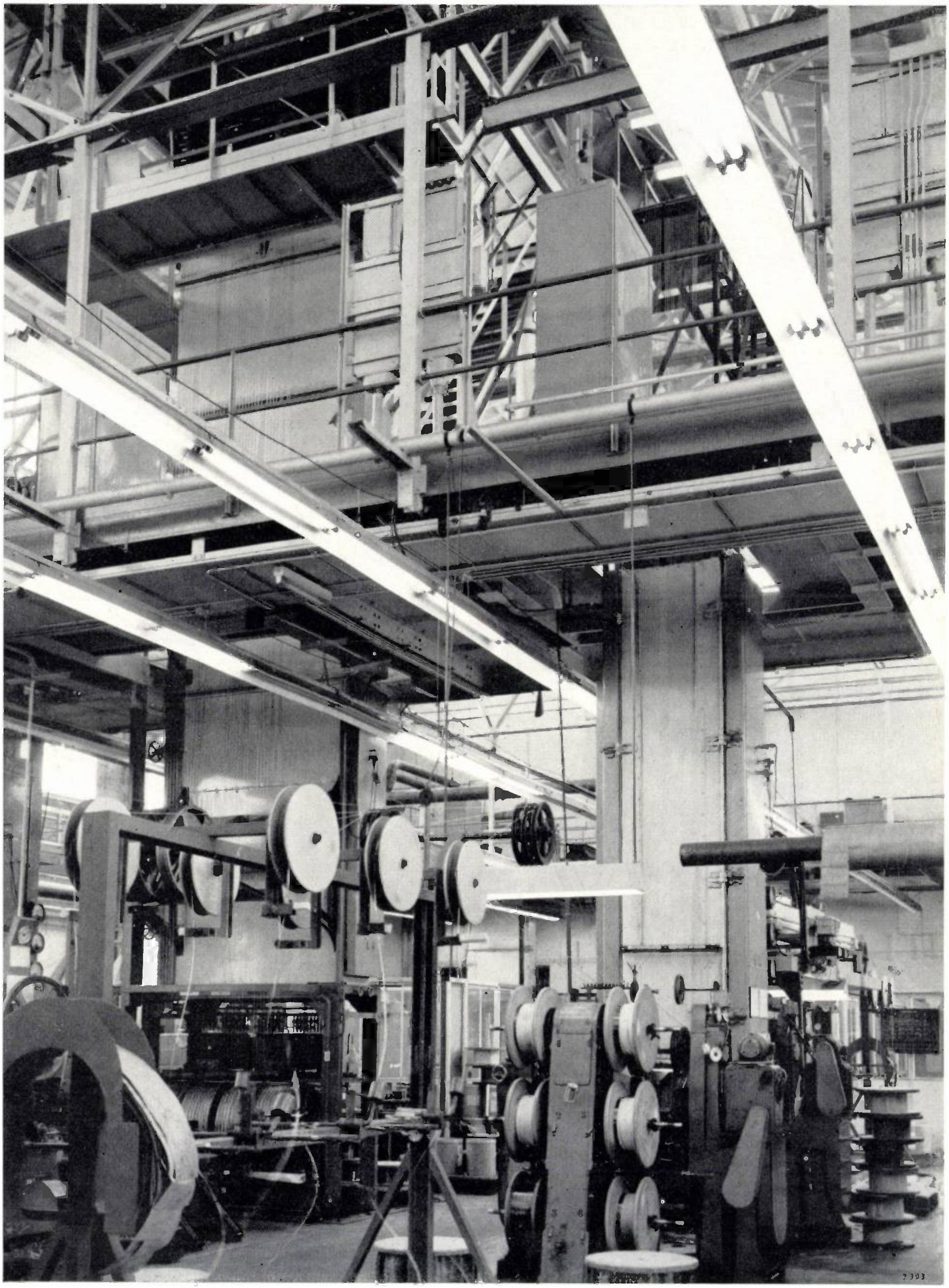


Fig. 1. Two *vertical* machines for the manufacture of enamelled wire. The one on the left is 14 m high. Each of these machines processes a number of wires at a time, each wire passing a number of times alternately through a lacquer reservoir and a muffle furnace. See also figs. 3 and 4.

a given drawing rate, the wire remains longer in the furnace. This is especially important with *thick* types of wire, where the lacquer is applied in relatively thick coats, and which therefore need rather a long drying time.

In general it is desirable for economic reasons to let one machine process a number of wires at a time. The advantage of this will however be lost if the wire breaks often, and if the whole machine has to be stopped each time a break occurs. Thick wire does not break very often, but with thin wire this possibility must indeed be kept in mind. As a logical conclusion, the manufacture of enamelled wire has tended towards the use of horizontal machines for thin wires, with one wire per machine (fig. 2), and multiple vertical machines for thick wire, taking 8 or more wires at a time (fig. 1).

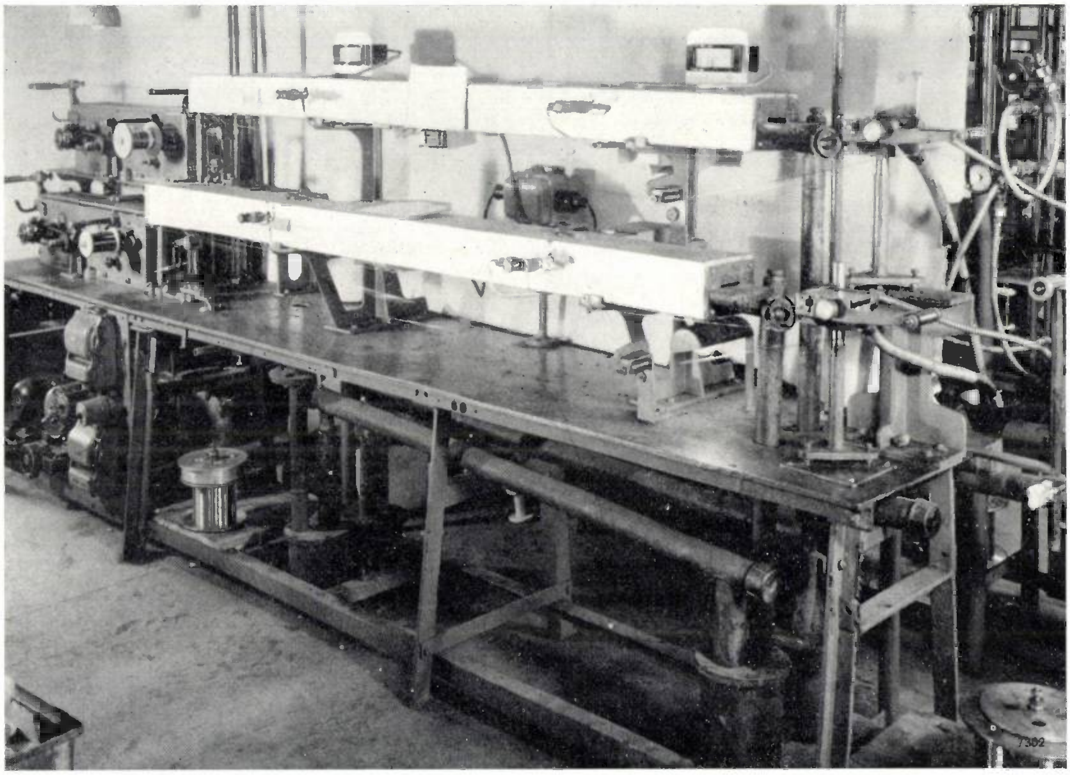


Fig. 2. Two *horizontal* machines for the manufacture of enamelled wire, placed one above the other. In each machine, the wire is passed a number of times alternately through a lacquer reservoir and an oven.

Figures 3 and *4* show some details of a vertical machine which enamels 16 wires at a time, applying 6 coats of lacquer to each wire. Fig. 3 shows the lacquer reservoir at the bottom of the machine. We see here 6×16 = 96 slits in the slanting wall through which the wires are drawn up via a roller to the furnace. It is inevitable that a small stream of lacquer will flow out through these slits; this ends up in a gutter, whence it is pumped back into the reservoir.

Fig. 4 shows the part of the machine where the 16 finished wires are "coiled".

Demands made on the wire and methods of testing

The finished wire is tested as regards a considerable number of properties. The *uniformity of the enamel layer* is of prime importance for wire to be used for winding purposes: the enamel must not show any inhomogeneities or pits, and must be applied equally thickly all round the copper. The mechanical, electrical and chemical properties of the enamel layer are also of great importance. We shall mention a few of the methods described in standard specifications for the testing of enamelled wire in these respects.

Holes and other large flaws in the insulation are detected by passing the wire through a conducting liquid (mercury or a solution of common salt), while a voltage is applied between the core of the wire and the liquid. The circuit also contains a resistor, a neon lamp and a counter. If the resistance of the enamel layer in contact with the liquid is less than e.g.

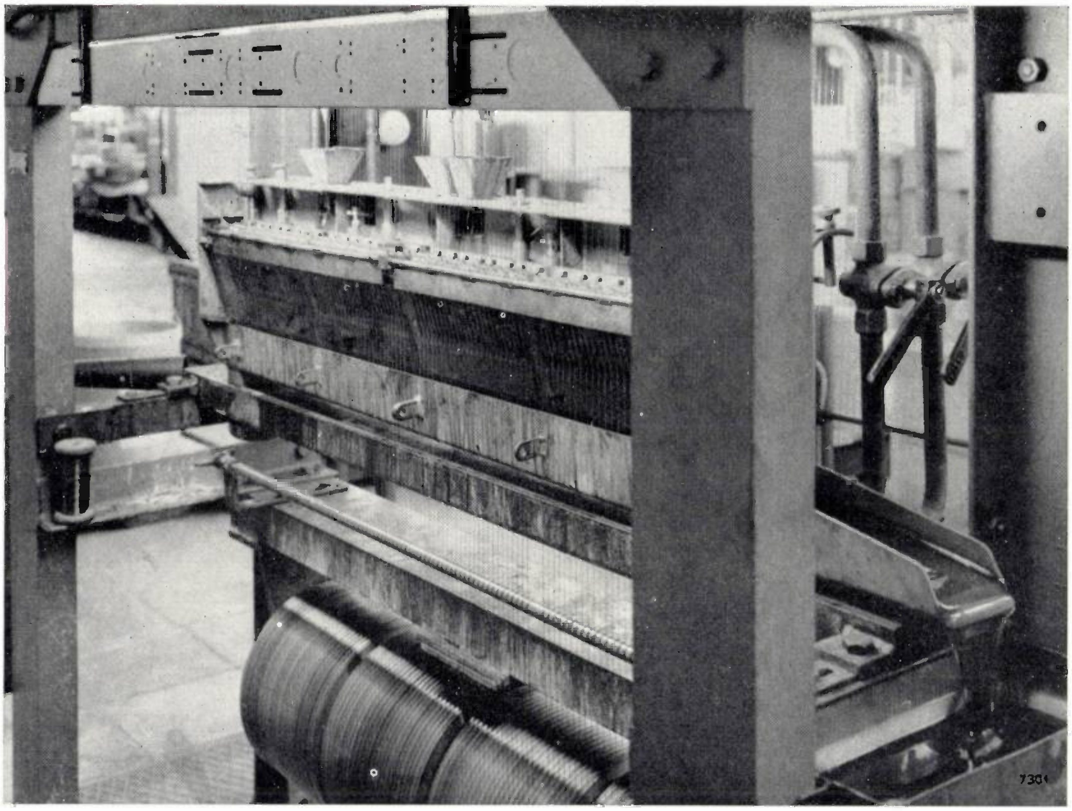


Fig. 3. The lacquer reservoir for a vertical machine. This machine supplies 16 wires at a time with a six-coat insulating layer. We see here how the wire is drawn up through 96 slits in the wall to the muffle furnace above. The thin stream of lacquer which flows out of these slits is caught in a gutter and pumped back to the lacquer reservoir. The degree of "muffling" is judged with the aid of a colour sample (the reel on the left): the colour of the enamelled wire is darker if the wire is muffled at a higher temperature or for a longer time. If the colour is lighter or darker than that of the sample, the rate at which the wire is drawn through the machine is decreased or increased, respectively.

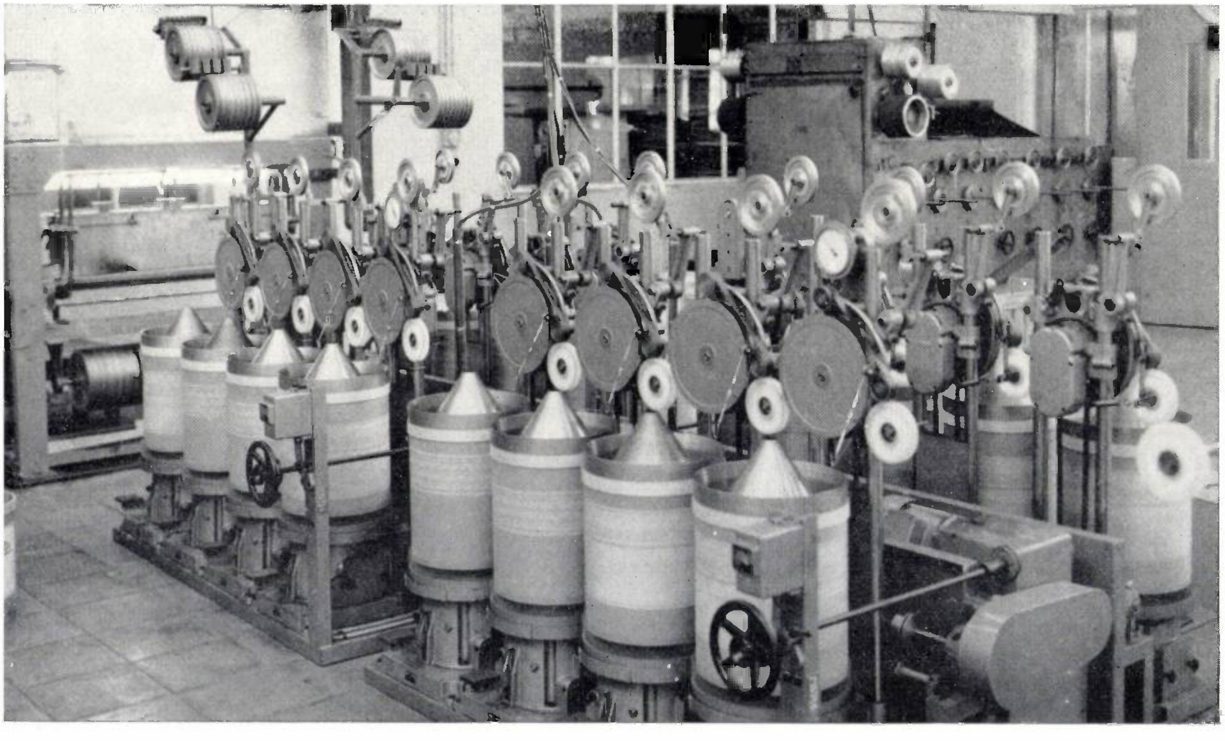


Fig. 4. Rotating containers in which the 16 finished wires coming from the machine in the background are "coiled".

10 000 Ω anywhere, a current flows through the circuit which lights up the neon lamp and sets the counter in operation (see *fig*. 5).

A uniform thickness of the enamel layer is of importance not only in connection with the insulation resistance, but also because a very constant wire diameter (standard deviation of less than 2%) is demanded for certain winding methods. We shall not discuss here the fairly simple methods used for testing the thickness of the wire.

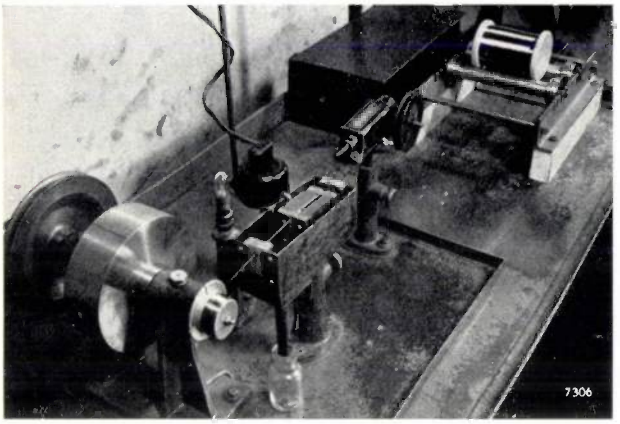


Fig. 5. An apparatus for detecting (fairly big) flaws in the insulation of enamelled wire. The motor on the left draws the wire under investigation through a tray containing mercury. The mercury and the core of the wire are connected in a circuit together with a neon lamp and a counter. Flaws in the wire passing through the mercury, i.e. places where the insulation resistance is less than 10 000 Ω, make the lamp light up and actuate the counter.

In order to check the layer structure of the enamel insulation simply and rapidly, the wire in question is placed for about 1 minute at an angle of 45° in a reagent which is capable of dissolving the enamel. A slanting cross-section of the enamel is then to be seen on the wire, and the various layers can be clearly seen with the aid of a magnifying glass (*fig*. 6).

The *mechanical* tests cover such diverse properties as resistance to wear, flexibility, etc. There are, for example, wear tests in which the number of strokes of a loaded reciprocating needle which the insulation can stand is determined. The flexibility and "windability" are tested by winding the wire round pins of roughly the same diameter as the wire itself. It is also stipulated that the wire after being wound in this way remains undamaged if the temperature is suddenly increased. This demand, which is one of the most stringent with which the wire must comply, is mainly of importance in connection with treatment which a coil may undergo after it is wound ("compounding" or impregnating).

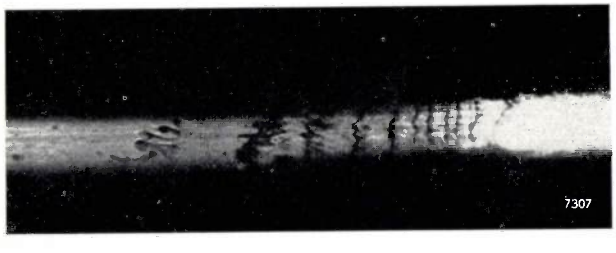


Fig. 6. A piece of enamelled wire in which the insulating layer has been etched off at an angle. The different coats in which the lacquer was applied can be clearly seen.

As far as the *electrical* properties of the enamel are concerned, the breakdown voltage in particular is of importance. This can be determined e.g. by winding a piece of wire round a polished steel cylinder, and applying a gradually increasing voltage between the core of the wire and the cylinder. Other electrical properties which it is desirable to know are the insulation resistance and the dielectric losses. Measurements of the insulation resistance carried out as an aid in the investigation of the physical structure of the enamel layer are discussed on p. 349.

The *chemical* properties of the enamel which are of special interest to the electrical engineer are the thermal aging and the resistance to solvents and impregnating agents. We shall mention thermal aging again briefly towards the end of the article, but for the rest we will leave this subject here: any further details would carry us too far into the practical aspect of the investigations. We shall now consider enamelled wire from a theoretical point of view.

Relationship between the chemical structure and the properties of the enamel layer

It is a difficult but rewarding task to try to find some relationship between the observable properties of enamelled wire and the chemical composition of the enamels used. We are far from having a complete insight into this matter, but what has already been reached is worthy of mention.

General principles of the structure of lacquer enamels

Lacquers may be divided into two groups: *physically drying* and *chemically drying*.

A physically drying lacquer consists of a volatile solvent with a macromolecular substance, e.g. nitrocellulose, dissolved in it. When such a lacquer is applied to an object, the solvent evaporates and the residue forms a layer on the object in question. This process takes place without any chemical change, i.e. the molecules do not react with one another. The enamel layer thus formed can be redissolved in the original solvent, and also has a tendency to soften when heated. These lacquers are therefore also sometimes called *thermoplastic* lacquers.

An enamel layer with entirely different properties is obtained if the molecules of the lacquer in question do react with each other, forming a three-dimensional network, as is the case with the chemically drying lacquers. The layer thus formed will not have the tendency to soften on heating, and is resistant to the usual solvents (though it may swell to a certain extent in some cases).

It will be clear that only the second type of lacquers, which are also known as *thermosetting* lacquers, come into consideration for use as insulation for wire. Thermoplastic lacquers are sometimes used as an *extra layer* on a wire which is already insulated with a thermosetting lacquer ("Thermoplac" wire). When a coil wound with this wire is heated, the outer layer will soften and thus cause the windings to stick together. This is useful e.g. in coils which must be given a rigid form without a support, such as the deflection coils in television picture tubes. In what follows, however, we shall forget about such thermoplastic lacquers, and only discuss the real insulating layer, which must thus comply with the above-mentioned condition that the molecules form a three-dimensional network.

A second necessary condition is that the network in question must adhere firmly to the copper. This can be explained as follows. It is of course essential that the enamel layer on wire used for winding coils should be able to accommodate the *strain* induced by the winding process without rupture. The copper core itself can meet this demand, as long as it is made in the right way. Now experiments have shown that a "loose" film of enamel can accomodate hardly any strain; instead, local constrictions are produced in the enamel, leading to breakage. If however the molecular network adheres rigidly to the copper everywhere, the enamel layer can easily be stretched with the copper without breaking.

We shall discuss a third general "structural principle" of lacquer enamels with reference to the oldest type of lacquer, oil lacquer.

Oil lacquers are made according to a principle going back to the last century for the preparation of e.g. carriage lacquer. The main ingredients of such lacquers were naturally occurring *resins* (rosin, copals) and *drying oils* (linseed oil, tung oil). Modern wire enamels with an oil base still use natural drying oils. The natural resins have however been replaced by synthetic resins, which gives the advantages of a more constant composition and a wider possible range of properties.

In the new oil lacquers as well as in the old ones we find an important structural principle, which is manifest to some degree in all types: there is a certain balance between the ingredients of a *flexible* nature and those of a more *rigid* nature. The first are here represented by the drying oils with their long linear carbon chain, the second by the resins, whose molecules contain a large number of rings (see *fig. 7*).

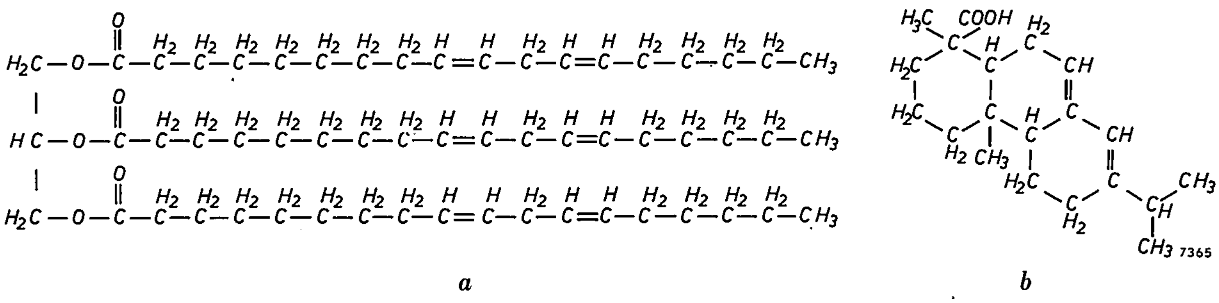


Fig. 7. The chemical structure of *a*) an oil ingredient of oil lacquer (the glycerol ester of 9,12-linoleic acid) and *b*) of a resin ingredient (abietic acid).

If the molecular network in question is formed only of the flexible components, as in e.g. rubber, we obtain a very elastic product, which however has the disadvantage of being soft and liable to swell in many solvents: molecules of the latter can easily penetrate the "open" network and cause it to expand. On the other hand, if the network is formed only of rigid components (as e.g. in certain synthetic resins) the result is a hard but brittle substance.

The combination of resins and drying oils actually used is just what is needed to give enamelled wire its excellent properties for use in winding coils: the enamel combines good flexibility with a considerable resistance to mechanical and chemical effects.

Now that we have learnt some general structural principles, we may show how the individual properties of certain types of enamelled wire can be understood against their structural background.

"Povin" wire

As our first example we shall consider the entirely synthetic lacquer for "Povin" wire. This lacquer also has two main ingredients, which show the above-mentioned balance of properties. The first ingredient is polyvinyl formal, whose molecule is characterized by *a*) a mainly linear structure, *b*) the occurrence of ring systems of the 1,3-dioxane type, and *c*) a large number of hydroxyl and acetate groups (see *fig. 8*). The second ingredient is a synthetic resin marketed under the name "Novolak". Of this substance we need only mention that its molecule is mainly made up of rigid components and also contains a number of hydroxyl groups.

It is not yet quite clear what chemical reaction takes place between these two ingredients during the manufacturing process. It is assumed that the hydroxyl groups of the "Novolak" as well as those of the polyvinyl formal play a part in the formation of the three-dimensional network. This will not, however, contain a large number of chemical bonds, as in the case of oil lacquers: the chemical structure of the ingredients simply does not allow this.

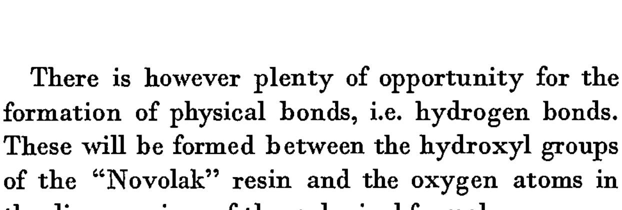


Fig. 8. The chemical structure of polyvinyl formal, an important ingredient of the insulation of "Povin" wire.

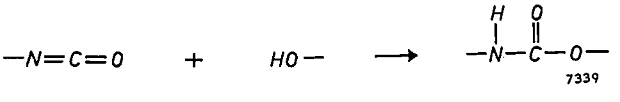
There is however plenty of opportunity for the formation of physical bonds, i.e. hydrogen bonds. These will be formed between the hydroxyl groups of the "Novolak" resin and the oxygen atoms in the dioxane rings of the polyvinyl formal.

Apart from the wide-meshed "loose" network of chemical bonds, a close-meshed network of hydrogen bonds will thus be formed. Each individual hydrogen bond is weaker than a chemical bond, but because of their great number a very strong structure results. Such a structure is just what is needed in a wire enamel: it may be expected that when the wire is bent, during which process the enamel layer is subjected to considerable strain, many of the quite weak hydrogen bonds will be broken. They will thus offer very little resistance to bending, while the breaking of the hydrogen bonds will lead to very little permanent damage: if the strain is relieved e.g. by heating for a short time, a new network of hydrogen bonds will immediately be formed, so that the old strength is regained. This is the reason for the excellent mechanical properties of "Povin" wire to which we alluded at the beginning of this article.

"Posyn" wire

As mentioned, the particular property of "Posyn" wire is that its enamel layer can be removed by dipping in molten solder. Here too, the connection between this property and the chemical structure of the enamel layer may be clearly seen.

The chemical bonds which hold much of the network together are formed by the reaction:



i.e. the reaction between an isocyanate group and a hydroxyl group. Now this reaction is a reversible reaction, which proceeds in the opposite direction at high temperatures. When the wire is dipped in the molten solder, the three-dimensional network breaks up into fragments of low molecular weight as a result of the reversal of this reaction. These fragments evaporate, and the copper which is thus exposed comes out of the solder bath covered with solder.

"Potermo" wire

The enamel of the recently developed "Potermo" wire has one of the least complicated network structures: the network is formed by the esterification of terephthalic acid with simple polyalcohols such as ethylene glycol and glycerol. The way in which the network is built up of these components is shown in *fig. 9*.

The excellent resistance to high temperatures of "Potermo" wire has been found by experiment to be due to the fact that oxygen from the air has very little effect on it. It is not entirely clear why a network like this should be so insensitive to oxygen, but on the other hand it is not altogether surprising: the network does after all consist practically entirely of benzene rings and ester groups, which are known not to react easily with oxygen. The other groups are probably sufficiently well shielded by the oxygen-resistant groups.

It is clear from the above what a complicated interplay of factors determines the quality of enamelled wire. In fact, the situation is even more complicated, because not only the structure itself, but also "flaws" in the structure have an influence on the quality. We shall devote the last section of this article to this important subject, in particular to an investigation in this field carried out in the Research Laboratories in Eindhoven.

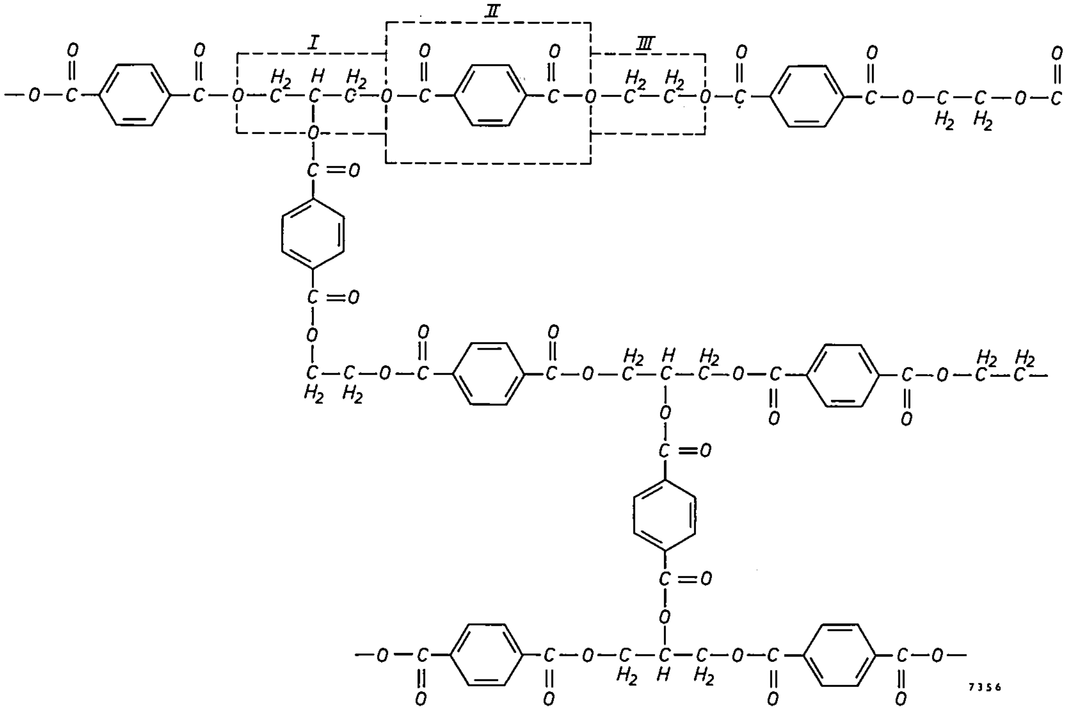


Fig. 9. The network structure of the insulation of "Potermo" wire. *I*: glycerol; *II*: terephthalic acid; *III*: ethylene glycol. These three components occur in various sequence in the network.

Investigation of "flaws" caused by mechanical stresses

It has been indicated in the foregoing how rather large flaws in the insulation are detected during the routine inspection of the wire. Work has been going on in the Research Laboratories over the past few years with a view to developing more refined methods for studying enamelled wire under mechanical stress. It has been found that measurement of the insulation resistance of the enamel layer, if carried out by sensitive methods, shows up the presence even of microscopically small cracks produced by such stress. Although this investigation has certainly not yet reached its final stage, the preliminary results fully justify great attention to this subject.

The first method used is based on measurement of the discharge time of a capacitor formed by clamping a metre of enamelled wire between tin foil. The tin foil is earthed, and the core of the wire connected to an electrometer. The capacitor so formed is charged to an known voltage, and the time needed for it to discharge to a certain lower voltage is measured. This time depends in a simple way on the unknown insulation resistance and on the capacitance, which is also unknown. In order to eliminate the latter unknown, a second measurement is carried out with a calibrated resistor included in the circuit in parallel with the capacitor.

Measurements on *dry* (and unstressed) enamelled wire gives a high value of the insulation resistance (e.g. 10^{13}-10^{14} Ω) for all types of wire, as long as there are no real holes in the enamel layer. Measurements on *damp* enamelled wire normally indicate a relatively small reduction in this resistance, the precise value of the reduction showing characteristic variations with the type of wire. The reduction increases in the order: oil-lacquer enamelled wire — "Posyn" wire — "Povin" wire.

This picture alters considerably as soon as the wire is subjected to bending tests, so that stresses are produced in the enamel layer. Even a slight bend (radius of curvature e.g. 1 m) in the wire produces a considerable reduction in the insulation resistance (e.g. by a factor 10-100) when the measurements are made on *damp* wire. If the wire is bent rapidly, or bent around a pin or a sharp corner, the effect is increased so much that the method is no longer applicable, because the capacitor discharges too fast.

The following method is used for further study of the problem. The core of the strongly bent wire is connected to the negative terminal of a battery, and the wire is dipped into a solution of common salt containing a little phenolphthalein. The other terminal of the battery is connected to a nickel rod which is placed in the same solution. Electrolytic effects are immediately produced: bubbles of hydrogen gas, together with OH^- ions, are produced at the bends in the immersed wire. The latter can be observed from the red colour assumed by the phenolphthalein at the *outer* side of each bend. This is what would be expected if one assumed that small cracks can be produced by tensile stress.

This method gives a negative result if the wire is only *slightly* bent, even after long periods of time, which surprised us at first.

Careful microscopic investigation finally produced a fairly complete picture of the nature of these flaws. When the wire is quickly bent round a pin, one observes cracks suddenly forming in the enamel layer after some time. Even if the wire is bent slowly and slightly, very fine cracks are observed after enough time has elapsed. These hair cracks, however, do *not* extend right through to the core, but are restricted to a few outer layers of the insulation. It may be assumed that such cracks do produce a considerable decrease in the resistance, but do not allow electrolytic effects to occur. If the enamelled wire is warmed, the hair cracks often close up again: they can then no longer be observed under the microscope, and the insulation resistance is observed to have increased again. The hair cracks can be regarded as fairly "harmless" precursors of the cracks produced by e.g. bending round a sharp angle.

Cracks of the latter type are most undesirable in a coil wound with enamel wire: if the coil is internally damp, they allow undesirable small currents to flow — i.e. they can cause a certain measure of short-circuiting. If such currents contain a DC component, they can also cause severe corrosion of the copper wire as a result of electrolysis. If the corrosion is allowed to proceed far enough, the copper wire may even break.

So much for this investigation. Its results throw an interesting light on various practical developments of recent times. For instance, for the winding of coils the *rectangular* cores and formers which were formerly common are being replaced more and more by ones with *rounded* corners. Moreover, attempts are being made to improve the quality of the coils by using *waterproof* materials for separating the windings.

As far as understanding of the above-mentioned phenomena is concerned, it will be clear that science still has a lot of ground to win from empirical knowledge in this field; and this is just what makes this subject so fascinating. There is thus no better way of closing this short survey of the work on enamelled wire than with the mention of a couple of examples which are still not easy to understand. *Fig. 10* illustrates the curious complications which are found in the thermal aging of enamelled wire. Quite un-

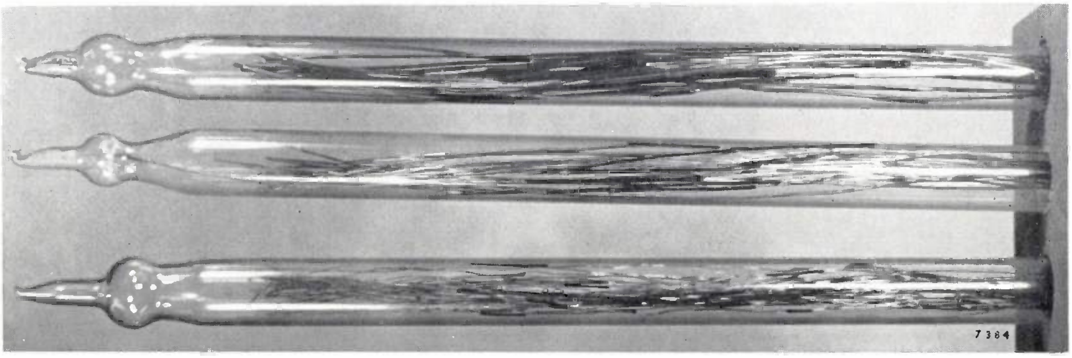


Fig. 10. Thermal aging of "Posyn" wire in a sealed tube for 48 hours at 180 °C. Surprisingly enough, the insulation is damaged more by thermal aging under these conditions than when it is exposed to the same temperature for the same time in air.

expectedly, "Posyn" wire in a sealed-off tube shows a more rapid deterioration of the insulation than the same wire which is simply exposed to the air. This phenomenon is found to a greater or lesser extent with all "solderable" wires. It is not so pronounced with "Posyn" wire, but it is still not entirely suppressed. *Fig. 11* shows a series of ring-shaped cracks which are produced when a bent piece of "Povin" or "Posyn" wire is dipped in a solvent. This phenomenon, which is called "solvent cracking", is very closely connected with the mechanical stresses present in the enamel layer: if the wire is heated to 100-120 °C after bending, so that the mechanical stresses in the enamel are relieved, solvent cracking does not occur. It is also very strange that such cracks are quite absent when oil-lacquer enamelled wire is used.

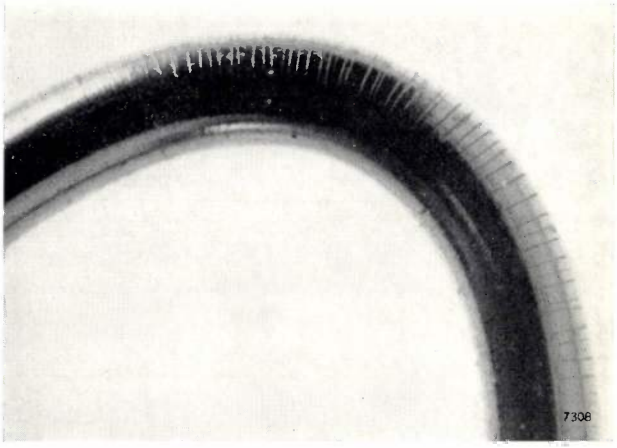


Fig. 11. A series of cracks (magnification 20×) in a piece of "Povin" wire, produced by bending the wire and dipping it into methanol. No really satisfactory explanation has yet been found for this "solvent cracking".

Summary. Among the properties required of enamelled wire used for the winding of coils are flexibility and resistance to mechanical and chemical effects. The "enamel" insulation consists of "chemically drying" lacquers, which are applied in several coats by passing the wire a number of times alternately through a lacquer reservoir and a muffle furnace. This process is carried out horizontally for thin wire but vertically for thick wire. To obtain satisfactory enamel layers the following conditions must be satisfied: *a*) the formation of a three-dimensional network of atoms which *b*) is firmly attached to the copper and which *c*) contains both flexible and rigid elements. The last requirement is illustrated for the case of oil-lacquer enamelled wire. Further, a theoretical explanation is given for the great mechanical strength of "Povin" wire, for the remarkable fact that "Posyn" wire can be soldered through the enamel layer, and for the thermal resistance of "Potermo" wire. Finally, a recent investigation of microscopic "flaws" produced by mechanical stress is described, and a number of phenomena which are as yet rather difficult to understand are mentioned.

THE PLOTTING OF ELECTRON TRAJECTORIES WITH THE AID OF A RESISTANCE NETWORK AND AN ANALOGUE COMPUTER

by A. J. F. de BEER *), H. GROENDIJK *) and J. L. VERSTER *).

537.533.3:621.517.729:681.14.001.57

Controlling the form of an electron beam is a problem frequently encountered in electronic engineering, for example in designing the focusing system of an electron microscope, an X-ray tube or a television picture tube. With the aid of a resistance network and an analogue computer the manner in which a very narrow electron beam is focused by an electrostatic lens can be ascertained quickly and accurately.

Determining the electron trajectories in the electric field of a given electrode configuration is a problem encountered in the design of many types of electron tube. The calculation of the fields and of the electron trajectories is very time-consuming work, and therefore various methods have been developed in the course of the years for speeding up the calculations involved. An elegant method of determining electron trajectories consists in rolling steel balls over a rubber membrane [1] [2]. This method can only be used for "two-dimensional" fields and is therefore not suitable for most modern types of tubes. The electrolytic tank with automatic plotting mechanism, described some time ago in this journal [3], does not suffer from this limitation and is therefore generally useful for systems possessing rotational symmetry, such as the focusing lens of television picture tubes for example (*fig. 1*). It is less suitable, however, where the trajectories considered are close to the axis of such a system (paraxial trajectories), for the principle underlying the apparatus in question consists in calculating the curvature of the trajectory. The latter is always very small in the paraxial case, so that the relative error in the result is too great.

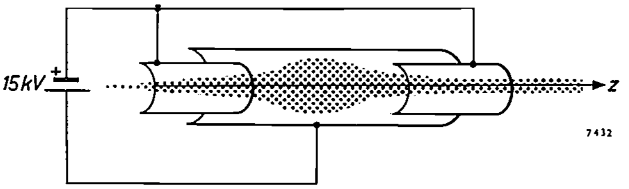


Fig. 1. Cross-section of an electrostatic focusing lens in a television picture tube. The lens consists of three coaxial cylinders at different potentials. The width of the electron beam is here exaggerated.

We shall now describe another method by which precisely the paraxial trajectories can be determined automatically and with sufficient accuracy. This method makes use of a differential equation which has the paraxial trajectories as solutions. The coefficients of the differential equation are functions of the potential Φ on the axis of the electrode system. The variation of Φ is found with the aid of a *resistance network* [4]. A PACE *analogue computer* [5] produces the solutions of the differential equation. The solutions are traced by an x-y recorder. *Fig. 2* gives a view of the equipment employed.

Principle of calculating the trajectories

If z is the coordinate along the symmetry axis of the electrode system and r is the distance to this axis, then the differential equation for the trajectories $r(z)$ is:

$$4\,\Phi r'' + 2\,\Phi' r' + \Phi'' r = 0. \quad . \quad . \quad . \quad (1)$$

Here Φ is the potential at the axis, and the primes represent derivatives with respect to z. This equation, which contains as coefficients only the function Φ *on the axis* and derivatives of Φ *in the direction of the axis*, is derived from the general differential equation for the path of an electron for the special case of an electric field possessing rotational symmetry, and for r and r' both small (paraxial rays) [6]. By substituting [7])

$$r = R\,\Phi^{-1/4} \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

*) Philips Research Laboratories, Eindhoven.

[1] P. H. J. A. Kleynen, The motion of an electron in two-dimensional electrostatic fields, Philips tech. Rev. 2, 338-345, 1937.

[2] G. Alma, G. Diemer and H. Groendijk, A rubber membrane model for tracing electron paths in space charge fields, Philips tech. Rev. 14, 336-344, 1952/1953.

[3] J. L. Verster, An apparatus for automatically plotting electron trajectories, Philips tech. Rev. 22, 245-259, 1960/61.

[4] J. C. Francken, The resistance network, a simple and accurate aid to the solution of potential problems, Philips tech. Rev. 21, 10-23, 1959/60.

[5] This computer, made by Electronic Associates, Long Branche, N.J., has been installed in Philips Computing Centre at Eindhoven. W. P. J. Fontein and L. G. J. ter Haar of the Computing Centre helped to design the circuits for solving our problem and assisted in the experiments.

[6] V. E. Cosslet, Introduction to electron optics, Clarendon Press, Oxford 1950, p. 36.

[7] The letter R will also be used in this article to denote electrical resistance. As there was little risk of confusion, there was no reason to depart from this internationally accepted symbol in electron optics.

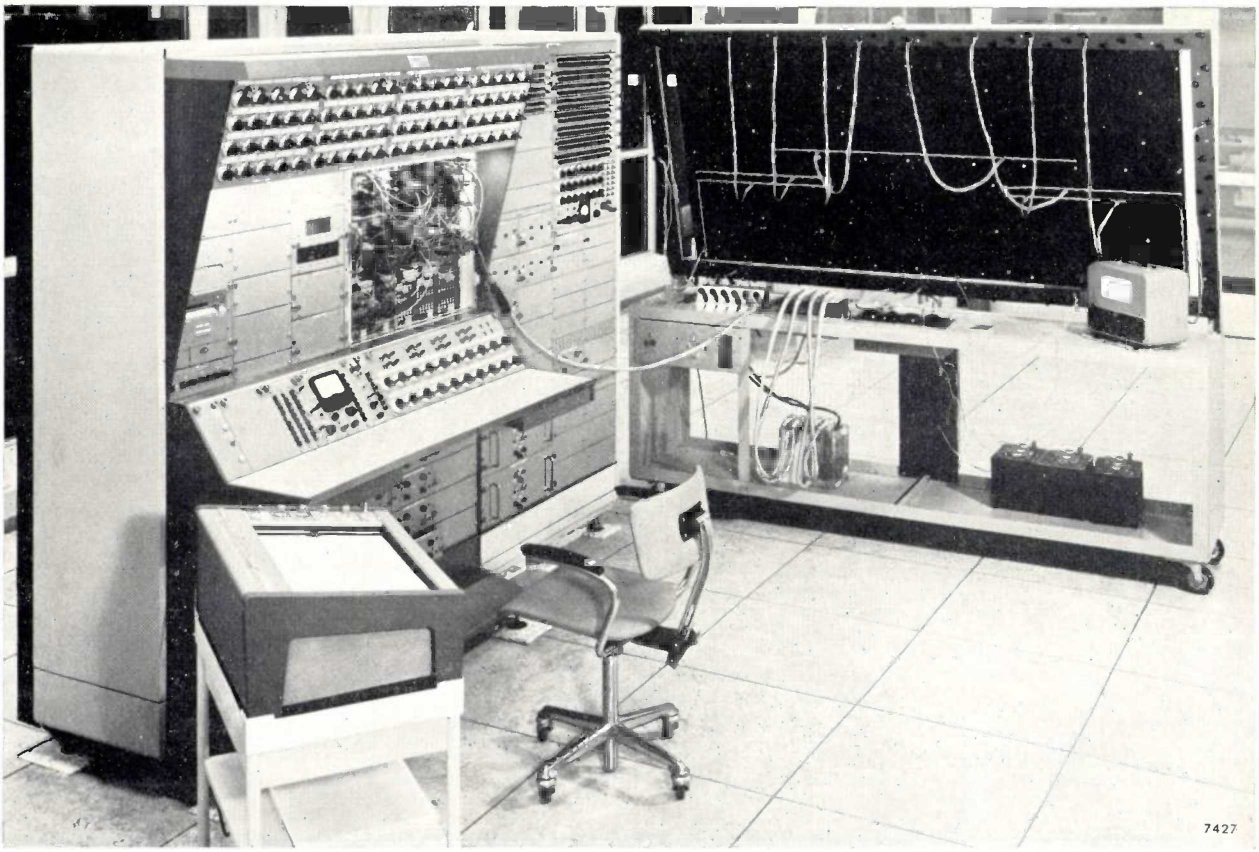


Fig. 2. View of the equipment for the automatic tracing of paraxial electron trajectories. On the right is the resistance network, on which the model of the investigated electrode system is marked, for clarity, with white cord. The cord is wound around the contacts of the network points, so that the mesh spacing is visible. The contacts themselves are too small to be seen on the photograph. On the left is the PACE analogue computer. The trajectories calculated by the computer are traced on the *x-y* recorder in the foreground.

the equation can be further simplified to Picht's equation:

$$R'' + \frac{3}{16}\left(\frac{\Phi'}{\Phi}\right)^2 R = 0. \quad \ldots\ldots \quad (3)$$

This equation no longer contains the second derivative of Φ. Since, as a rule, second derivatives cannot be determined so accurately as first derivatives, the latter equation is to be preferred.

In order first to describe broadly the process by which the analogue computer solves this differential equation, we assume for the moment that $(\Phi'/\Phi)^2$ has a constant value K. The reader may accept provisionally that the machine can generate a voltage V which, as a function of time t, satisfies the equation analogous to eq. (3):

$$\frac{\mathrm{d}^2 V}{\mathrm{d}t^2} + \frac{3}{16} KV = 0. \quad \ldots\ldots \quad (4)$$

This voltage is traced on an *x-y* recorder, the stylus moving at a constant velocity in the t direction. If we replace the V and the t axis of the graph thus produced by an R and a z axis respectively, we have one of the solutions of Picht's equation (3), i.e. that with the initial values R and R' corresponding to the preset initial values of V and $\mathrm{d}V/\mathrm{d}t$.

Since this is important to the understanding of what follows, it should be emphasized that in the machine the varying of the z coordinate is substituted by the passage of time, appearing in the fact that the stylus of the recorder moves at a constant speed in the z direction.

Now, in the case considered here the coefficient $(\Phi'/\Phi)^2$ is not constant but is a function of z. In order to construct the solutions in the same way with the analogue computer, the variation of $(\Phi'/\Phi)^2$ must be supplied to the machine in such a way that, at the moment during the solving process when an arbitrary point z_1 is reached on the z axis, the machine is supplied with a voltage whose magnitude is proportional to the value of $(\Phi'/\Phi)^2$ at z_1.

As mentioned, $\Phi(z)$ — from which $(\Phi'/\Phi)^2$ is calculated — is determined with the aid of a resistance network. For our present purposes we shall disregard in the first instance the fact that the resistance network only gives the potential at a number of discrete points along the axis; we assume, then, that the potential Φ can be *continuously* scanned by means of a sliding contact (*fig. 3*). The derivative Φ' along the axis can be approximated by the quotient ($\Delta\Phi/\Delta z$) of the difference between the potentials at two closely adjacent points, z_1 and z_2, and the distance between these two points. For this purpose we must use *two* sliding contacts, S_1 and S_2, moving along the same axis. A circuit in the computer itself divides by Φ the value of Φ' thus obtained, and squares the result. At the same moment at which the two contacts sliding along the z axis pass the points z_1 and z_2, the value of $(\Phi'/\Phi)^2$ pertaining to z_1 appears at the output of the circuit mentioned. (The "calculation" in an analogue computer takes no computing time.)

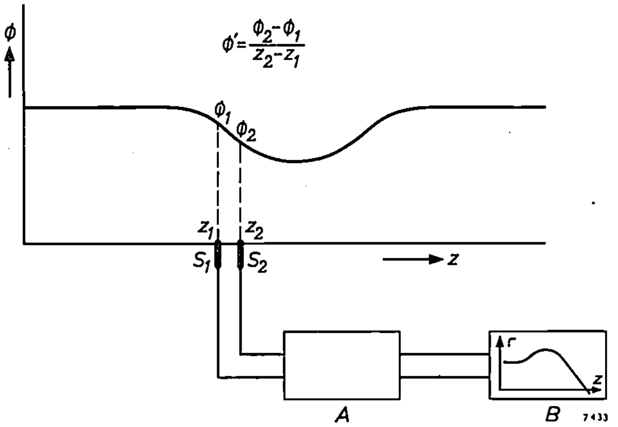


Fig. 3. Diagram illustrating the principle of the trajectory calculation. The potential along the z axis and its derivative are scanned with two sliding contacts S_1 and S_2. From the result the computer A calculates the variable coefficient $(\Phi'/\Phi)^2$ which occurs in Picht's equation, and solves that equation. B is the x-y recorder.

To ensure the above-mentioned time relation between $(\Phi'/\Phi)^2$ and $r(z)$, the contacts must move along the z axis at the same speed as the z coordinate is swept when finding the solution. A different speed means that, when the quotient ($\Delta\Phi/\Delta z$) is computed, Φ is not differentiated with respect to z but with respect to, for example, az. This amounts to substituting $3/(16a)^2$ for the factor 3/16 in eq. (3), so that the wrong differential equation would be solved. In the following we shall consider how the required time relation is obtained. We shall also explain how, using the resistance network (with which, as mentioned, the function Φ cannot be continuously scanned) an approximate value of Φ and Φ' is determined by interpolation. An incidental problem, which will also be dealt with, concerns the method of ensuring that the coupling of the computer to the resistance network does not distort the correct potential distribution over the network.

The solution of these problems was well worth while. Hitherto it was necessary to measure the relevant voltages separately on the network and to compute each trajectory numerically; with some training, this took about two hours. The computer takes only a minute to do the same work.

Before dealing with the above questions, we shall discuss the principal elements from which the circuits in an analogue computer are built up.

Basic elements of an analogue computer

The adder

Fig. 4a shows the symbol used for an "adder" in analogue computer technique, and fig. 4*b* gives the circuit diagram. The triangle with curved base A represents an "operational amplifier", whose main features are a high gain (approx. 10^8) and the fact that it reverses the polarity of the voltage. Because of this there is virtually no voltage across the input i of the amplifier (potential of p equal to zero). If the input voltage were to differ slightly from this value, the result at the output u would be a very high potential of opposite polarity which, via the feedback resistor R_u, would again immediately reduce the input voltage to zero. Since the potential of p remains constant at zero, the currents V_1/R_1 and V_2/R_2, which arise when the voltages V_1 and V_2 are applied to resistors R_1 and R_2, cannot flow to earth. If they were to do so, they would produce across the input resistance of A a voltage that would make the potential of p differ from zero. The currents mentioned must therefore pass through the negative feedback resistance R_u. Let the output voltage be V_u, then:

$$\frac{V_u}{R_u} = -\left(\frac{V_1}{R_1}+\frac{V_2}{R_2}\right),$$

or

$$V_u = -\left(\frac{R_u}{R_1}\,V_1+\frac{R_u}{R_2}\,V_2\right).$$

The result, then, is that the circuit multiplies V_1 and V_2 by the negative constants $-R_u/R_1$ and $-R_u/R_2$, respectively, and adds the products. Since the amplifier is used for this and other elementary mathematical operations, it has been given the name "operational amplifier".

If only one voltage V_1 is applied to the input, and moreover $R_1 = R_u$, what appears at the output is simply the same voltage of opposite polarity, $-V_1$. (The symbol in fig. 4*a* is then, of course, drawn with

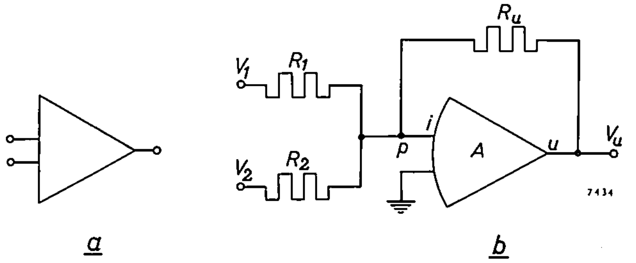


Fig. 4. *a*) Symbol and *b*) circuit diagram of adder in an analogue computer. As a result of the high negative feedback of the operational amplifier A, the point p remains at zero potential. Currents V_1/R_1 and V_2/R_2 have to flow together through the feedback resistor R_u because no perceptible current flows through the input resistance of A. When $R_u = R_1 = R_2$, then $V_u = -(V_1+V_2)$.

only one input.) If V_1 and V_2 are to be subtracted from one another, two operational amplifiers are needed: one to produce $-V_1$, and the other for the addition of the V_2.

The integrator

Fig. 5a represents the symbol of an integrator, and fig. 5*b* the relevant circuit diagram. The current V_1/R_1 must again flow entirely through the feedback circuit, which in this case contains the capacitor C. For the output voltage V_u we now have:

$$\frac{V_1}{R_1} = -C\frac{\mathrm{d}V_u}{\mathrm{d}t}$$

or

$$V_u = -\frac{1}{R_1C}\int\limits_0 V_1\,\mathrm{d}t + V_u(0).$$

In the latter expression $V_u(0)$ is the output voltage at the moment $t = 0$. By giving this voltage a specific value we can introduce the initial conditions of a differential equation. Choosing $R_1C = 1$ second, we find for V_u precisely the value of the integral of $-V_1$. By taking another value for RC we can multiply the value of the integral by a specific constant.

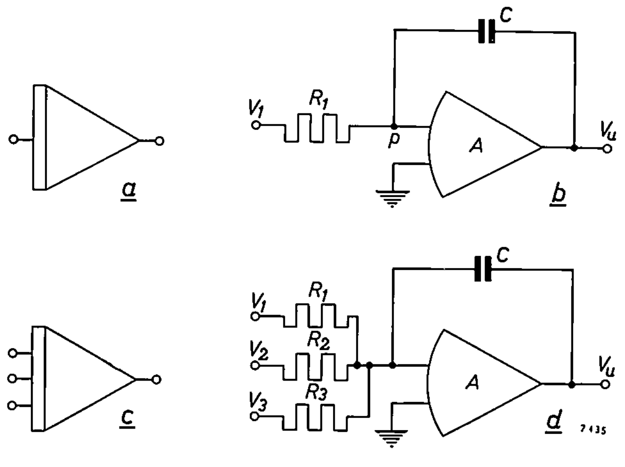


Fig. 5. *a*) Symbol and *b*) circuit diagram of an integrator. Point p again has zero potential. As in the case of the adding circuit, the current V_1/R_1 flows through the feedback circuit, which here contains the capacitor C. When $R_1C = 1$, then $V_u = -\int V_1\mathrm{d}t$. *c*) Symbol and *d*) circuit diagram of an integrator which simultaneously adds and integrates: $V_u = -\int(V_1 + V_2 + V_3)\,\mathrm{d}t$.

In practice an integrator always has various resistors R_1, R_2, R_3, ... at the input. If we apply to three of these resistors, for example, the voltages V_1, V_2, V_3 (fig. 5*d*, with the appertaining symbol in fig. 5*c*) and if we choose $R_1C = R_2C = R_3C = 1$ sec, we obtain at the output:

$$V = -\int(V_1 + V_2 + V_3)\mathrm{d}t.$$

The multiplier

Of the various kinds of multipliers used in analogue computers, we shall treat here only the kind known as *servo-multipliers*. The symbol and the diagram can be seen in *fig. 6*. Along the potentiometer P_1 the potential rises from zero to the "reference voltage" V_r, usually chosen as 100 V. The operational amplifier A_1 supplies an energizing voltage to the servo-motor M until the voltages at the two inputs i' and i'' are equal to one another, that is until the sliding contact of P_1 is at the point where the voltage is equal to V_1. The voltage-division ratio a on the potentiometer P_1 is then equal to V_1/V_r. Since the sliding contact of potentiometer P_2 is mechanically coupled to that of P_1, the voltage-division ratio on that potentiometer is now also equal to V_1/V_r. When a voltage V_2 is applied to P_2, the voltage on its sliding contact is $V_u = aV_2 = (V_1\times V_2)/V_r$, that is to say, except for a constant factor, the product of V_1 and V_2.

To prevent this constant factor always appearing in the formulae, the actual computing voltage V[8]) is not used when designing the circuits but the quotient V/V_r. In this way, all computing voltages are normalized with respect to the reference voltage. In the following this normalization will be tacitly applied, and this should be borne in mind when considering the various formulae, since the normalization transforms the computing voltages into dimensionless quantities.

A multiplier usually contains several potentiometers, e.g. five, all sliding contacts of which are mechanically coupled to that of the first. It is then possible simultaneously to produce the products V_1V_2, V_1V_3, V_1V_4, and so on. If, for instance, we take for V_3 the product V_1V_2 already formed, we then obtain $V_1^2V_2$.

[8]) The term "computing voltage" is used to distinguish it from other voltages in the computer that have nothing to do with the computation.

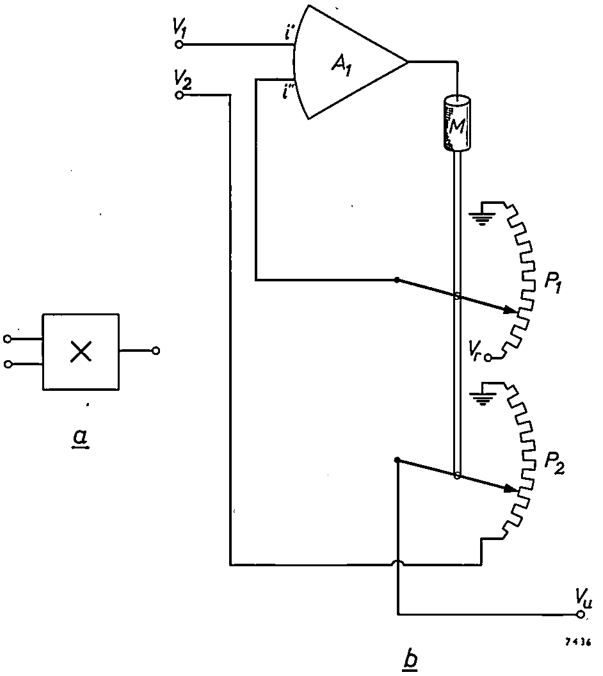


Fig. 6. *a*) Symbol and *b*) circuit diagram of a multiplier. M is a servo-motor which simultaneously drives the sliding contacts of potentiometers P_1 and P_2. The amplifier A_1 energizes the servo-motor until the voltages at both inputs i' and i'' are identical, that is until the sliding contact P_1 is at the point where the voltage is equal to V_1. The voltage-division ratio of both potentiometers is then equal to V_1/V_r, where V_r is a reference voltage. When a voltage V_2 is applied to the second potentiometer, the voltage on its sliding contact is $V_u = V_1V_2/V_r$, that is, except for a constant factor, the product of V_1 and V_2.

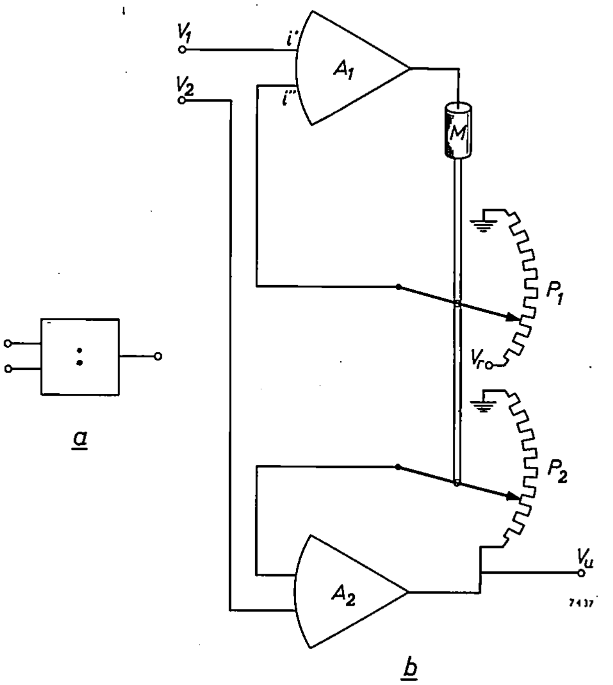


Fig. 7. *a*) Symbol and *b*) circuit diagram of a divider, which produces the quotient V_2/V_1. The position of the two sliding contacts is regulated in the same way as for the multiplier. The normalization applied makes the voltage division ratio on potentiometers P_1 and P_2 both equal to V_1. The operational amplifier A_2 ensures that V_u acquires a value at which zero difference exists between the voltages at its inputs. In that case $V_2 = V_1V_u$, and thus: $V_u = V_2/V_1$.

The divider

By means of the circuit in *fig. 7* the quotient V_2/V_1 can be obtained. The operation of the circuit is explained in the caption to the figure.

All the above-mentioned basic elements are electronic. It is also possible to use other basic elements, for example mechanical elements, as done for the first time by Bush in a computer for solving differential equations [9]). Mechanical analogue computers are often accurate enough, but they are relatively slow in operation.

Solving Picht's equation

To solve Picht's equation with the analogue computer, the equation is written in the form:

$$R'' = -\frac{3}{16}\left(\frac{\Phi'}{\Phi}\right)^2 R.$$

Two integrators I_1 and I_2 (*fig. 8*) are now connected in series, and R'' (in the form of a voltage varying with time) is applied to I_1. Between I_1 and I_2 one then finds $-R'$, and I_2 delivers the function R itself, which represents the solution. For the purpose of applying R'' to the integrator I_1, it is computed with the aid of the basic elements described from R and $(\Phi'/\Phi)^2$ in accordance with the equation in the above form.

It may be asked how this is possible, for R has only just been produced by integration from R''. We are not concerned here, however, with the performance of successive mathematical operations, but with a feedback circuit (a "loop") in which the currents flowing, governed by the basic elements, are such that the variation of the voltage V at the output gives a solution of the differential equation.

Since the circuit in fig. 8 satisfies the same differential equation (in the "computer variables" V and t) as applies to the electron trajectories $R(z)$ (with the "problem variables" R and z), it can be said that the circuit is an analogue for the behaviour of the electrons in the electric field. The name analogue computer expresses this kind of relationship. We have attempted above to prepare the reader for the perhaps somewhat confusing interplay between the computer variables V and t and the problem variables R and z (we shall henceforth largely be concerned with the latter).

[9]) V. Bush and H. L. Hazen, Integraph solutions of differential equations, J. Franklin Inst. 204, 575-615, 1927. See also A. S. Jackson, Analog computation, McGraw-Hill, New York 1960. The latter book is recommended for a further study of the theory and application of analogue computers.

The computer contains a separate circuit with which the transformed trajectory coordinate R is retransformed in accordance with (2) into the actual trajectory coordinate r, thus immediately giving the solution in the required form.

When the capacitors in the integrators are charged to the voltages corresponding to the chosen initial values, the computer is set in operation by the simultaneous closing of a number of switches. The sliding contacts then immediately begin to sweep the z axis of the network, and the recording stylus begins to trace the solution. At every point z_1 the deflection of the stylus in the y direction is proportional to the value of the trajectory coordinate r in z_1. After roughly one minute, the contacts reach the end point; the recorder has then traced one of the solutions of the differential equation (1).

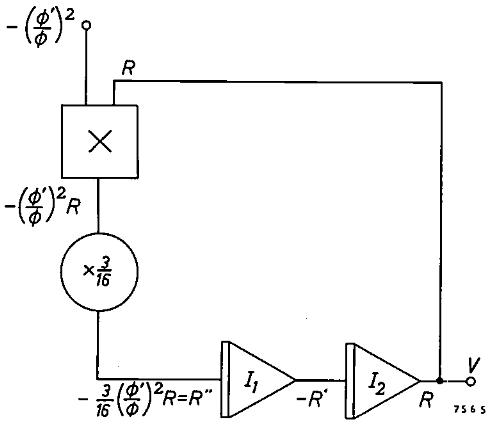

$-(\frac{\Phi'}{\Phi})^2$
R
×
$-(\frac{\Phi'}{\Phi})^2 R$
$\times\frac{3}{16}$
$-\frac{3}{16}(\frac{\Phi'}{\Phi})^2 R = R''$
I_1
$-R'$
I_2
R
V
7565

Fig. 8. Circuit for solving Picht's equation. When the computer is started, R and R' must be given the values $R(0)$ and $R'(0)$. The solution $R(z)$ corresponding to these intial conditions is supplied by the circuit in the form of the varying output voltage $V(t)$.

The sliding contacts and stylus are returned to their starting point, the capacitors in the integrators are discharged, and the machine is ready to produce another solution. For this purpose, different initial values are assigned to R and R'.

Two trajectories plotted in this way can be seen in *fig. 9*.

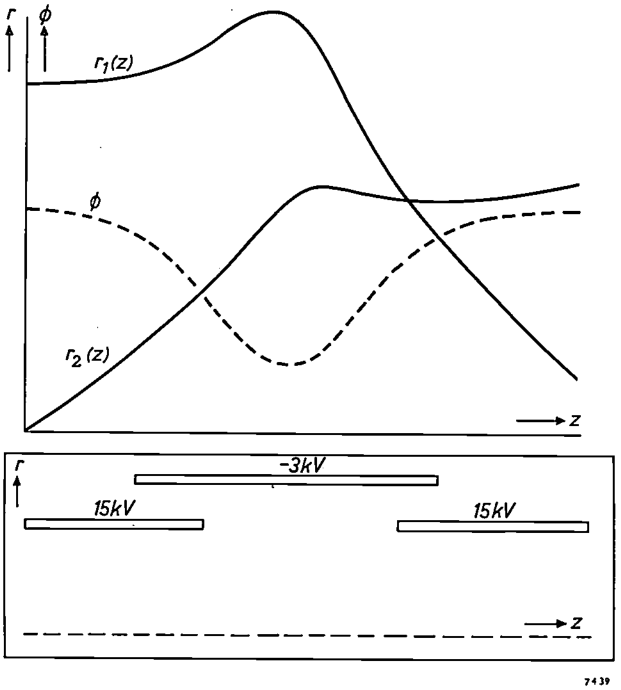


Fig. 9. Two paraxial trajectories $r_1(z)$ and $r_2(z)$ with different initial conditions. Dotted line: variation of the potential Φ on the axis of the electrode system. A half cross-section of the electrode configuration is shown below the graph: the z axes are on the same scale.

Calculation of $(\Phi'/\Phi)^2$

We have already mentioned that the variation of the potential Φ along the axis is found with the aid of a resistance network. In such a network the resistances between all neighbouring network points are dimensioned in such a way that the potential distribution over these points is a good approximation of the potential distribution in an axial cross-section of a rotational-symmetric electrode system. This electrode system is simulated in the network by electrically interconnecting a number of network points and giving them the appropriate potential [10]).

The fact that the network does not give a continuous curve for Φ, but only the values of Φ at discrete points, has its consequences. For solving Picht's equation it is, of course, necessary to provide a value of Φ for every value of z. To this end, interpolation is carried out (with the aid of the computer) between the discrete points. This is done by simultaneously determining the values of Φ and Φ' at two successive network points n and $n + 1$, while the computer ensures that Φ and Φ' increase linearly with time (and hence with z) from the initial values to the final values in the interval concerned.

This linear interpolation of Φ and Φ' between the points n and $n + 1$ (*fig. 10*) is effected in accordance with the formulae:

$$U_n(x) = \Phi_n + (\Phi_{n+1} - \Phi_n)\frac{x}{a} \quad . \; . \quad (5a)$$

and

$$W_n(x) = \frac{\Phi_{n+1} - \Phi_{n-1}}{2a} + \left(\frac{\Phi_{n+2} - \Phi_n}{2a} - \frac{\Phi_{n+1} - \Phi_{n\;1}}{2a}\right)\frac{x}{a} \quad . \; . \; . \quad (5b)$$

[10]) For further particulars, see the article under reference [4]). The fact that the "model" of the investigated electrode system can so readily be obtained on the resistance network is one of the great advantages of the network compared with the electrolytic tank: with the latter it takes quite a lot of time to make the electrode models.

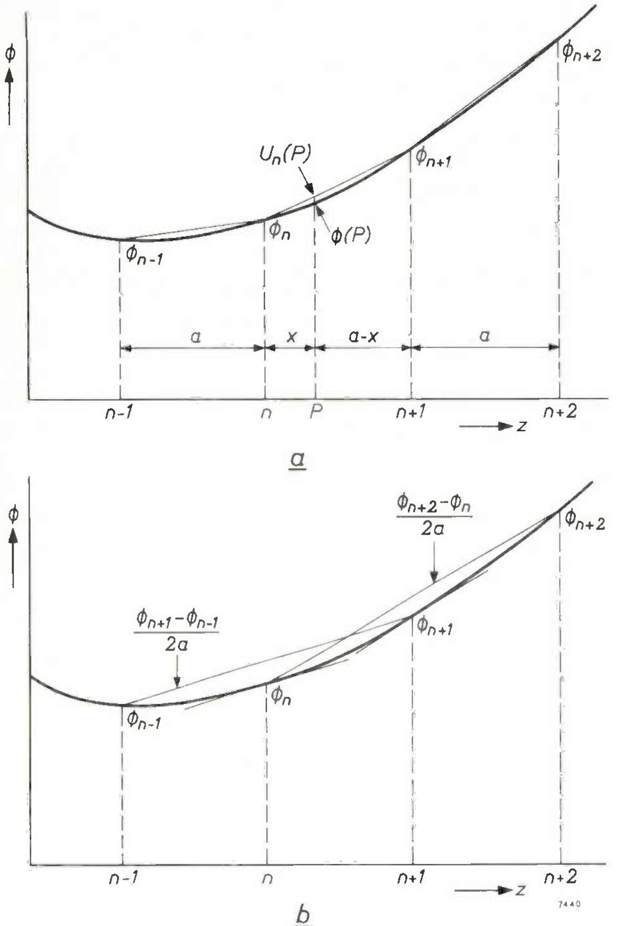


Fig. 10. *a*) Linear approximation of Φ and *b*) of Φ'. Between the points n and $n+1$ we approximate to Φ' by letting the slope of the line $(n-1, n+1)$ increase linearly to the slope of $(n, n+2)$.

In these expressions Φ_n is the value of $\Phi(z)$ at the point n, $U_n(x)$ is the approximation of $\Phi(z)$ between the points n and $n+1$, $W_n(x)$ is the approximation of $\Phi'(z)$ between the points n and $n+1$, a is the mesh width of the network, and x is the abscissa between the points n and $n+1$, which varies from 0 to a. The derivative Φ' at the point n is approximated, according to eq. (5*b*), by the slope of the line connecting Φ_{n+1} and Φ_{n-1}, as illustrated in fig. 10*b* for an arbitrary curve and an exaggeratedly large mesh width.

In order to show how U_n and W_n are obtained with the computer we write the equations (5) in the following form:

$$U_n = \Phi_n + \int_0^{x/a} (\Phi_{n+1} - \Phi_n)\,\mathrm{d}z,$$

$$W_n = \frac{\Phi_{n+1}-\Phi_{n-1}}{2a} + \int_0^{x/a} \frac{\Phi_{n+2}-\Phi_{n+1}-\Phi_n+\Phi_{n-1}}{2a}\,\mathrm{d}z.$$

The circuit with which U_n is computed will be dealt with separately under another heading.

From the formula for W_n it can be seen that the voltages Φ_{n-1}, Φ_n, Φ_{n+1} and Φ_{n+2} have to be simultaneously fed to the computer during the time that the z coordinate is swept between the points n and $n+1$. This simultaneous feed of the four successive voltages is effected with the aid of a telephone selector, shown in *fig. 11*; the circuit diagram is given in *fig. 12*. The wiper arms are all mounted on one shaft, so that they change position simultaneously. In the computer a sinusoidal voltage with a period of one second is generated, and a separate circuit delivers after each period a voltage pulse for the stepwise actuation of the selector. In this way the axis on the network is swept at the rate of one mesh spacing per second. Now the circuit in fig. 8 is dimensioned in such a way that, during the solving of the differential equation, the z axis is swept at a speed which also corresponds to one mesh spacing per second. The above-mentioned time relation between the scanning of the network and the sweeping of the z axis is brought about by the sinusoidal voltage.

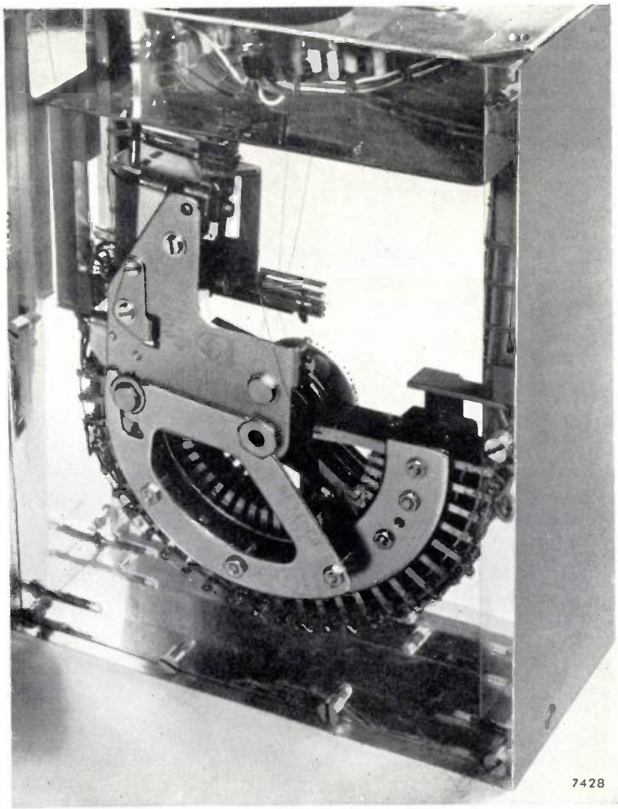


Fig. 11. The four-armed telephone selector (it has five wiper arms but only four are used). There are 26 contacts available on each arc, of which there are two sets of four, mounted one above the other. The four wiper arms, which are mounted on the same shaft, are designed so as to sweep 52 points in each revolution. The selector is mounted in an oil-filled container of transparent plastic, measuring 12×19×23 cm.

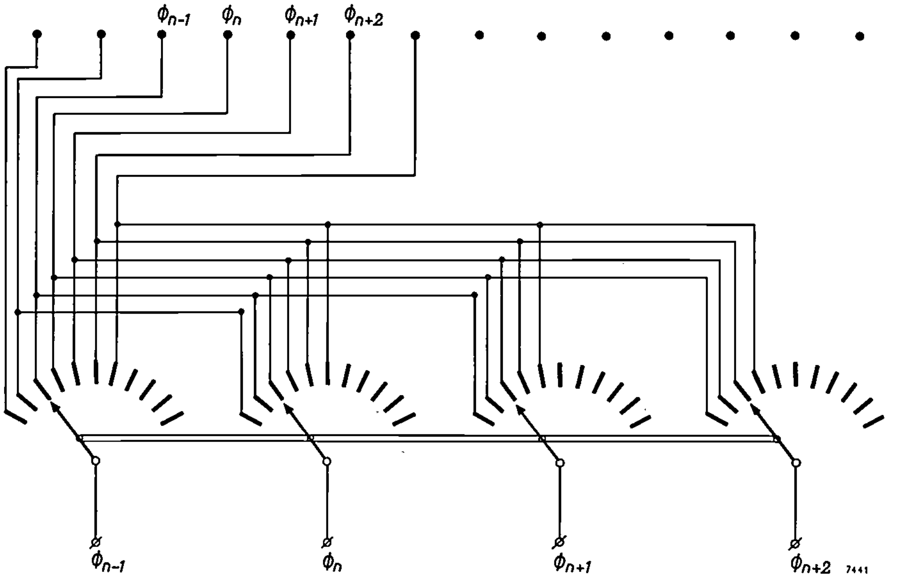


Fig. 12. Circuit diagram of the selector. For clarity, not all contacts are shown. The upper row of points represents a number of axial points of the network.

The four voltages scanned by the selector are fed to the circuit of *fig. 13*, which carries out the interpolation of Φ' between the points n and $n+1$.

At the moment when the x-y recorder begins to trace the trajectory between $n+1$ and $n+2$, the selector jumps to the next position, so that the voltages Φ_n, Φ_{n+1}, Φ_{n+2} and Φ_{n+3} are supplied to the circuit and $W_{n+1}(x)$ is computed. At the same moment as the selector changes position, the capacitor in the integrator (fig. 13) discharges, so that the circuit can begin with a clean slate on the interpolation between $n+1$ and $n+2$.

Network feedback

Fig. 13 shows that the network points switched in by the telephone selector are connected to adders. In fig. 4 it can be seen that, as a result of this, these network points are connected via resistors (R_1, R_2) of about 1 MΩ to a point (p) which has a fixed potential, viz earth potential. The potential Φ in Picht's equation is computed with respect to the potential at the point where the electrons have zero velocity. If we also made this "cathode potential" equal to earth potential, a point of the network having the voltage Φ would be loaded with a current Φ/R_1, which can have a marked effect on the voltage at that point.

To avoid this, the network with its voltage source is made "floating", and negative feedback is applied in such a way that the four points whose potentials are measured, and not the cathode, are on average at earth potential. This is done with the aid of the circuit of *fig. 14*. The operational amplifier A is provided with negative feedback via the resistors of the network, as a result of which the voltage at point *1* is virtually zero. At the moment when the selector has just changed position, $x = 0$, and therefore the voltage at point *2* is zero. It follows, then, that the voltage on *3* must also be zero. This is arranged by means of the amplifier A, which delivers a negative feedback voltage U_t, to be added to all network voltages, such that $\Phi_n + U_t = 0$. In the time between the successive changes of the selector,

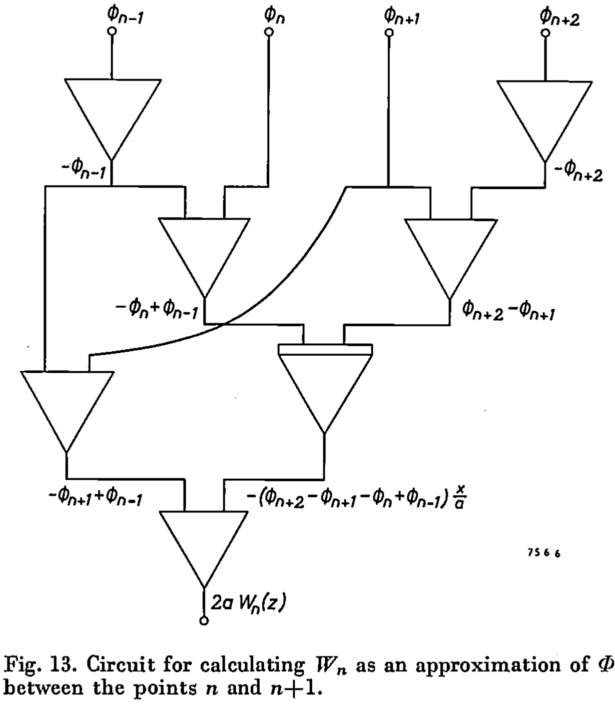


Fig. 13. Circuit for calculating W_n as an approximation of Φ' between the points n and $n+1$.

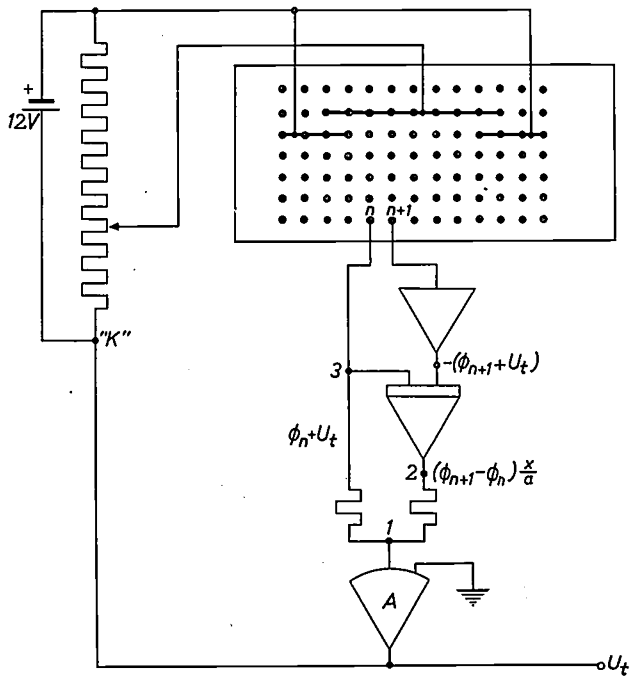


Fig. 14. Negative feedback is applied to the operational amplifier A via the network resistances. The amplifier delivers a negative feedback voltage U_t such that the points between which the potential is measured, n and $n+1$, and not the "cathode" of the network (point "K") are on an average at earth potential. This prevents the computer drawing currents from the network that would disturb the potentials at the network points.

The circuit is so designed that $-U_t$ is exactly equal to U_n, which approximates to $\Phi(z)$ between n and $n+1$.

The voltages at points n and $n+1$ are also conducted to the relevant points in fig. 13. The selector is not shown.

the voltage at *2* increases and, owing to the change of U_t, the voltage at *3* decreases by the same amount. In this way the values of Φ_n and Φ_{n+1} remain near earth potential during a given position of the selector, and unduly large currents cannot be drawn from the network.

The circuit is now so chosen (in particular by including in it the adder and the integrator) as to make the negative feedback voltage U_t exactly the opposite of U_n, which approximates to $\Phi(z)$ between n and $n+1$. This is evident, for when the voltage at *1* is zero,

$$\Phi_n + U_t + (\Phi_{n+1}-\Phi_n)\, x/a = 0,$$

whence

$$-U_t = \Phi_n + (\Phi_{n+1}-\Phi_n)\, x/a,$$

and this, according to (5*a*), is equal to $U_n(x)$.

This negative feedback has no influence on W_n, the determination of which involves only potential differences: these are independent of the potential level of the network.

The circuits in figs 13 and 14 thus deliver the values U_n and W_n, with which Φ and Φ' can be approximated by continuous functions consisting of straight line sections. From these the computer calculates $(W_n/U_n)^2$ as an approximation of $(\Phi'/\Phi)^2$, which function can be fed to the multiplier in fig. 8.

Accuracy of the potential measurement

When the quotient W_n/U_n is computed with the aid of the "divider" (fig. 7) the voltage-division ratio on both potentiometers is governed by U_n. In order to utilize the whole range of the potentiometers, U_n is first multiplied by a number which makes its maximum value only slightly less than that of the reference voltage V_r. (At another point of the circuit it is then necessary to divide by that number.) If now the minimum value of U_n is very much smaller than its maximum value the divider will function very inaccurately at the moment at which this minimum value occurs, for only a few windings of the potentiometer will then be used. Reasonable accuracy is still achieved when the ratio between maximum and minimum values is 20 : 1.

We shall now consider the error caused by approximating Φ and Φ' by U_n and W_n.

In a particular position of the telephone selector the voltages Φ_{n-1} up to Φ_{n+2} are fed to the computer. We wish to know by how much the voltage $U_n(P)$ applied for the point P (fig. 10*a*) differs from the voltage $\Phi(P)$ which follows from the third-degree curve that can be drawn through the four points Φ_{n-1} up to Φ_{n+2}. Although the latter value is not equal to the actual value of $\Phi(z)$ at point P either, it is in any case a better approximation. For calculating the errors in $U_n(P)$ and $W_n(P)$ we therefore assume that the third-degree curve represents the true variation of $\Phi(z)$. This amounts to saying that $\Phi(z)$ can be represented by the first four terms of the Taylor series in the neighbourhood of P (coordinate z_p):

$$\Phi(z) = \Phi(P) + (z-z_p)\Phi'(P) + \tfrac{1}{2}(z-z_p)^2\Phi''(P) + \tfrac{1}{6}(z-z_p)^3\Phi'''(P).$$

Applying this formula to the points $n-1$, n, $n+1$, $n+2$, and writing $z_p - z_n = x$, we find:

$$\begin{aligned}
\Phi_{n-1} &= \Phi(P) - (a+x)\Phi'(P) + \tfrac{1}{2}(a+x)^2\Phi''(P) - \tfrac{1}{6}(a+x)^3\Phi'''(P),\\
\Phi_n &= \Phi(P) - x\Phi'(P) + \tfrac{1}{2}x^2\Phi''(P) - \tfrac{1}{6}x^3\Phi'''(P),\\
\Phi_{n+1} &= \Phi(P) + (a-x)\Phi'(P) + \tfrac{1}{2}(a-x)^2\Phi''(P) + \tfrac{1}{6}(a-x)^3\Phi'''(P),\\
\Phi_{n+2} &= \Phi(P) + (2a-x)\Phi'(P) + \tfrac{1}{2}(2a-x)^2\Phi''(P) + \tfrac{1}{6}(2a-x)^3\Phi'''(P).
\end{aligned}$$

Substituting the expressions thus found for Φ_{n-1}, Φ_n, Φ_{n+1} and Φ_{n+2} in equations (5a) and (5b) we find a relation between the approximations $U_n(P)$, $W_n(P)$ and the "true" values $\Phi(P)$ and $\Phi'(P)$:

$$\begin{aligned}
U_n(P) &= \Phi(P) + C_1(x),\\
W_n(P) &= \Phi'(P) + C_2(x).
\end{aligned}$$

The two correction terms, which are functions of x, are given by:

$$\begin{aligned}
C_1(x) &= \tfrac{1}{2}x(a-x)\Phi''(P) + \tfrac{1}{6}x(a-2x)(a-x)\Phi'''(P),\\
C_2(x) &= \tfrac{1}{6}\,(a^2+3ax-3x^2)\Phi'''(P).
\end{aligned}$$

We have calculated the values of C_1/Φ and C_2/Φ' for a particular electrode configuration [11]). The maximum values found were:

$$C_1/\Phi = 0.002 \quad \text{and} \quad C_2/\Phi' = 0.003,$$

representing a maximum error of 0.2% in U_n and 0.3% in W_n.

[11]) Two identical cylinders in line, and a very small distance apart, their voltage ratio being 20 : 1. As explained above, this is the maximum ratio at which the computer is sufficiently accurate.

These errors should be compared with the errors in the potential due to the finite mesh-width of the network. This error has been determined by Francken [2]) for a given electrode configuration produced on three different scales, and thus for three different mesh numbers [12]). As regards the mesh number, our electrode model on the network corresponds to the above configuration on scale I, for which the error found in the potential is roughly 1%. We see then, that the interpolation of Φ and Φ' is sufficiently accurate. The accuracy can be improved by adopting the imaging mentioned by Francken. Since this involves depicting only half the configuration, the size of the model can be doubled. In that case our electrode model would correspond with Francken's model on scale III, for which an error of 0.5% is given. The errors C_1/Φ and C_2/Φ' are then also smaller, and the linear approximation is accordingly still justified.

Spherical aberration

When a number of paraxial electron trajectories in the field of an electrostatic lens are determined with the set-up described, one can derive from these the positions of the foci and of the principal planes, thus defining the behaviour of the lens as far as focusing is concerned.

Like an optical lens, an electron lens also has its imaging errors (aberrations). It is therefore of considerable importance to be able to determine such aberrations with this equipment. The most serious aberration found in electron beams in the guns of television picture tubes is spherical aberration (*fig. 15*). This is in fact the only aberration that occurs when imaging a point situated on the axis of the lens. The size of the circle of confusion is given by [13]):

$$\Delta r_0 = \frac{5}{64} f r^3 \sqrt{\Phi_0} \int\limits_{-\infty}^{z(F)} \frac{R_n^4}{\sqrt{\Phi}} \left[\left(\frac{\Phi''}{\Phi}\right)^2 - \frac{101}{120}\left(\frac{\Phi'}{\Phi}\right)^4 + \right.$$
$$\left. + \frac{13}{3}\left(\frac{\Phi'}{\Phi}\right)^3 \frac{R_n'}{R_n} - \frac{6}{5}\left(\frac{\Phi'}{\Phi}\right)^2 \left(\frac{R_n'}{R_n}\right)^2\right] \mathrm{d}z. \qquad (6)$$

Here Φ_0 is the potential of the equipotential space from which the parallel beam proceeds, $z(F)$ is the coordinate of the focus and f the pertaining focal length, and $R_n(z)$ is a normalized paraxial trajectory, i.e. that which comes parallel to the axis from $-\infty$ with $R = 1$. (The normalization, as in the case of V, makes R_n into a dimensionless quantity.)

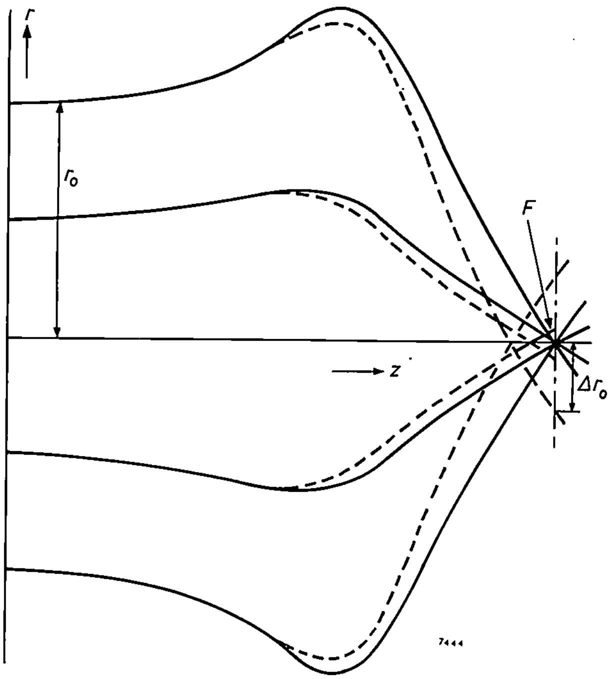


Fig. 15. Spherical aberration in the focusing of electrons. The solid lines represent the theoretical paraxial trajectories which all intersect the axis at the focus. The dashed curves are the actual trajectories when spherical aberration occurs. The outermost trajectories cut the axis nearest to the lens. At the focus a circle of confusion of radius Δr_0 is produced.

To determine Δr_0 with the computer we also need Φ'', for which we make use of the difference between $(\Phi_{n+2} - \Phi_{n+1})$ and $(\Phi_n - \Phi_{n-1})$. The value calculated in this way for each interval n to $n+1$ is practically equal to the mean value of Φ'' in that interval. For calculating the integral the network is again scanned in the same manner. The result of the integration, which is accurate to within 10%, can be read on a voltmeter.

Further applications

To conclude, we shall touch briefly on some other possible applications.

In the case of higher beam currents the mutual repulsion of the electrons has a perceptible influence on the electron trajectories. It has been shown [14]) that the term

$$\frac{1}{4\pi\varepsilon_0 \sqrt{2e/m}} \times \frac{I}{\Phi R}$$

should be added in such cases to the left-hand side of Picht's equation, I being the beam current, e the charge on the electron and m its mass. The circuit in the computer must be correspondingly modified.

[12]) The "mesh number" is defined as one of the dimensions of the electrode system, divided by the mesh width. Which dimension is taken is immaterial, provided it is consistently adhered to when comparing models on different scales.

[13]) See W. Glaser, Grundlagen der Elektronenoptik, Springer, Vienna 1952, pp. 371 and 675. Formula (6) is obtained by applying the substitution of eq. (2) in the formula given there. The formula is applicable when object and image are located in a field-free space and when the object distance is much greater than the image distance. If the former is much smaller than the latter, as it is in picture tubes (fig. 1) the integral remains the same, but the factor in front of the integral changes.

[14]) See the book mentioned in reference [13]), p. 142.

Where the beam currents are even higher, the potential distribution is influenced by the space charge caused by the electrons. This can be allowed for by the appropriate application of currents to the network points.

The lens action of a magnetic field with rotational symmetry can also be computed with this set-up, provided the field can be defined with a potential function. In this case, of course, the paraxial ray equation differs from that for electrostatic fields [14]) and the circuit in the computer must be adapted correspondingly.

It is also possible to compute the path of an electron moving in both an electrostatic field and a rotational-symmetric magnetic field. For this purpose *two* networks are needed, supplying respectively the electrostatic and magnetic potential distribution to two telephone selectors, one for each network, which simultaneously scan the axial potentials. The path is again calculated with an analogue computer on the basis of the paraxial-ray equation [14]) for combined electric and magnetic fields.

Finally, there is the case where electron trajectories are to be determined in potential fields which vary in the time during which the electron traverses the field. With our set-up it is necessary for this purpose to make the voltages on the electrodes vary with time. The best method is to supply the network with a voltage generated by the computer itself and which varies in accordance with a specific function.

Summary. The effective design of certain electrode systems possessing rotational symmetry (like the focusing lens of a television picture tube) calls for a detailed knowledge of the paraxial trajectories of the electrons. A PACE analogue computer can compute the trajectories as the solutions of a differential equation, Picht's paraxial-ray equation. To make the calculation entirely automatic, the potential variation along the axis of the electrode configuration, as obtained on a resistance network, is scanned by means of a four-arm telephone selector and fed directly to the computer. After discussing the principal component elements of an analogue computer, the authors explain how the computer derives a continuous function by interpolation from the discrete potential values obtained from the resistance network, and describes the process of solving the potential equation. The accuracy of the interpolation is also examined. The same set-up can be used to determine the spherical aberration of the electron lens.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of those papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, The Netherlands, where a limited number of reprints are available for distribution.

2918: P. Westerhof and E. H. Reerink: Investigations on sterols, XVIII. The synthesis and properties of some hydroxylated 9β,10α-pregnanes (Rec. Trav. chim. Pays-Bas **79**, 1118-1125, 1960, No. 9/10).

The synthesis of a number of 9β,10α-pregnanes and 9β,10α-androstanes has already been described (these abstracts, Nos **2881** and **2882**). The synthesis of the 17α-hydroxy derivatives of 9β,10α-progesterone and of 6-dehydro-9β,10α-progesterone, and some acylates of these compounds, is now described in the present publication. Tests on rats, mice and rabbits show that the acylates have a strong progesterone-like effect when administered subcutaneously, while they show the same effect on rabbits when orally administered.

2919: H. B. G. Casimir: A note on multipole radiation (Helvetica Phys. Acta **33**, 849-854, 1960, No. 8).

The expansion of a radiation field in a series of electric and magnetic multipole fields has already been investigated by several authors. Bouwkamp and Casimir (1954) showed that this expansion can conveniently be obtained by means of "Debye potentials". The use of group theory was avoided in reaching this result. It is now shown that the use of group theory leads to the same result, in a logical and straightforward way.

2920: C. Haas: The diffusion of oxygen in silicon and germanium (Phys. Chem. Solids **15**, 108-111, 1960, No. 1/2).

Starting from a simple model for the location of O atoms in crystals of Si and Ge, and from the assumption that internal friction and diffusion are both governed by the same relaxation process, the diffusion coefficient for oxygen in these elements, $D = D_0 \exp -(U/kT)$, is calculated from experimental data on internal friction. The results are for

Si: $D_0 = 0.21$ cm^2/sec, $U = 2.55$ eV; for Ge: $D_0 = 0.17$ cm^2/sec, $U = 2.02$ eV. The calculated values are in satisfactory agreement with the available experimental data.

2921: M. J. Sparnaaij: Gas adsorption and surface charge density of germanium surfaces (Phys. Chem. Solids **14**, 111-116, 1960).

Investigation of the effect of the presence of discrete charges on the surface of germanium crystals on the adsorption of gases such as argon. This effect is to be expected at charge densities of 10^{13}-10^{15} elementary charges per cm^2. Below this range, the effect will be too slight, and above the individual fields all merge into one. The Van der Waals forces which hold the gas atoms on the surface are compared with the polarization energies calculated in this publication for various charge densities. The theoretical results are compared with experimental determinations of the difference in the chemical potentials of argon adsorbed on non-oxidized and oxidized Ge. The difference is of the expected order of magnitude, and the charge density of 10^{14}-10^{15}/cm^2 thus found also agrees with the results obtained by Green on the change in the contact potential on oxidation.

2922: W. Albers, C. Haas and F. van der Maesen: The preparation and the electrical and optical properties of SnS crystals (Phys. Chem. Solids **15**, 306-310, 1960, No. 3/4).

Single crystals of SnS were prepared by heating a stoichiometric mixture of the components in vacuo to about 900 °C. The melting point is 880 ± 5 °C. Rhombic single crystals were prepared by passing a powdered sample at 420 °C in vacuo slowly through a melting zone (900 °C). Platelets measuring e.g. $2 \times 0.5 \times 0.3$ cm can be obtained by cleaving such crystals perpendicular to the *c*-axis. It follows from measurements of the Hall effect that the crystals are *P*-type semiconductors with 10^{17}-10^{18} holes per cm^3. The mobility μ_p of the holes depends on the temperature, with a maximum at about 200 °K. At room temperature, $\mu^p = 65$ cm^2/V sec. Preliminary measurements show that the specific conductance parallel to the *c*-axis is about 6 times smaller than that in a direction perpendicular to this. The crystals are completely opaque for radiation with a wavelength of less than 1 μm, which indicates that the band gap is 1.07 ± 0.04 eV. For wavelengths greater than 1 μm the absorption coefficient is proportional to λ^2. Application of Drude's equation gives an effective mass of the holes equal to 0.4 of the mass of an electron.

2923: J. Hornstra: Dislocations, stacking faults and twins in the spinel structure (Phys. Chem. Solids **15**, 311-323, 1960, No. 3/4).

The slip plane in crystals with the spinel structure is probably the (111) plane, in analogy with the (0001) slip plane in corundum-type crystals. This is confirmed by an observation of twinning during deformation, which can be regarded as a process in which partial dislocations with a (111) slip plane take part. Dislocations with this slip plane probably consist of 4 partial dislocations separated by 3 regions of stacking faults. The structure of these stacking faults is compared with that of the perfect spinel lattice. During slip, the cations move in a direction different from that of the oxygen ions. This process is discussed in some detail for the case of 4-coordinated cations. Another type of stacking fault shows similarities with the olivine lattice. Two possible configurations of the (111) twin boundary are discussed.

2924: J. Th. G. Erhardt: Technische beschermingsmaatregelen tegen warmtestraling (Ingenieur **72**, G.78 - G.88, 1960, No. 51). (Protective measures against thermal radiation; in Dutch.)

Personnel (and equipment) in the vicinity of e.g. a furnace can be protected against the heat by means of radiation screens. The temperature of and thermal radiation from the outer layer of a number of types of reflecting screens have been calculated as functions of the temperature of the body to be screened off. In order to check the practical utility of these calculations, the results were compared with experiment. The agreement was good; moreover, the calculated values were on the high (i.e. the safe) side. Finally, a few remarks are made about the use of radiation screens in practice.

2925: L. J. van der Pauw: An analysis of the circuit of Dauphinee and Mooser for measuring resistivity and Hall constant (Rev. sci. Instr. **31**, 1189-1192, 1960, No. 11).

Measurements of the resistivity and Hall coefficient of semiconductors with the aid of the circuit of Dauphinee and Mooser can show certain systematic errors due to capacitive effects in the switch. The magnitude of these errors is estimated, and it is shown how they can be eliminated with the aid of trimming capacitors in the right places and a suitably chosen switching programme. The sensitivity of the circuit is also calculated.

2926: H. J. Eichhoff and N. W. H. Addink: Untersuchungen zur Übertragbarkeit spektrochemischer Verfahren mit vollständiger Verdampfung (VIII. Colloquium Spectroscopicum Internationale, Lucerne 14th-18th Sept. 1959, Ed. H. Guyer, pp. 89-92; Sauerländer, Aarau (Switzerland) 1960). (Investigations of the transferability of spectrochemical methods based on complete evaporation; in German.)

5-10 mg of a material can be accurately spectrochemically analysed by completely evaporating the sample in a DC carbon arc. Careful preparatory work is needed to work out the exact method of analysis for the sixty or so elements for which it can be used. If however the method is once worked out for one type of spectrograph, it is relatively easy to change over to another type of instrument: all that needs to be done is to determine the conversion factors. Values of these factors are given for various spectrographs in the wavelength region 2500-3500 Å. For prism instruments, the factor is equal to 1 ± 0.1 over the whole range of wavelengths, but with grating instruments it depends strongly on the wavelength, and on the "blaze" of the grating. Analytical results are given.

2927: H. Bienfait: Management policy of the Philips Research Laboratory (Spring Meeting of the Industrial Research Institute, Virginia Beach, 8th-11th May 1960, pp. 201-215).

A survey of the history of the development and the organization of the Philips Research Laboratories, and also of the various steps taken to ensure a good scientific atmosphere. The relationship between the Research Laboratories and the development laboratories is discussed. The article ends with a look into the future.

2928: J. C. Francken and J. van der Waal: Residual gases in picture tubes (Vacuum **10**, 22-26, 1960, No. 1/2).

See Philips tech. Rev. **23**, 122, 1961/62 (No. 4).

2929: J. H. N. van Vucht: The Ceto getter — its chemical structure and hydrogen gettering properties (Vacuum **10**, 170-177, 1960, No. 1/2).

An investigation of the absorption of hydrogen by Th_2Al and of the reaction products formed. The equilibrium pressure of hydrogen above Th_2Al is measured as a function of the temperature and of the amount of hydrogen absorbed. The structure of the reaction products is studied by means of X-ray diffraction, neutron diffraction and nuclear magnetic resonance. The kinetics of the absorption are also studied, for the case where the surface of the Th_2Al is covered with oxygen (inactivated surface). Autocatalysis is observed to occur; a simple explanation is given for this effect. This investigation arose out of the study of the non-evaporating getter "Ceto", which has important uses in industry. This getter has the same structure as Th_2Al, but some of the thorium atoms are replaced by cerium and lanthanum. See also **R 411.**

2930: W. K. Westmijze: Eigenschaften und Anwendungen von Magnetkernen in der Messwertverarbeitung (Elektrotechn. Z. **A 81**, 779-783, 1960, No. 22). (Properties and use of magnetic cores in data processing; in German.)

The processing of data obtained as the results of measurements plays an important role in modern control engineering. It is often necessary to store the data, or quantities calculated therefrom, in a memory until they are needed again. For many applications it must be possible to place the data in the memory, and to retrieve them, very rapidly. The properties of magnetic cores in this connection are discussed in this article, which ends with a simple example of the uses to which a magnetic-core memory can be put.

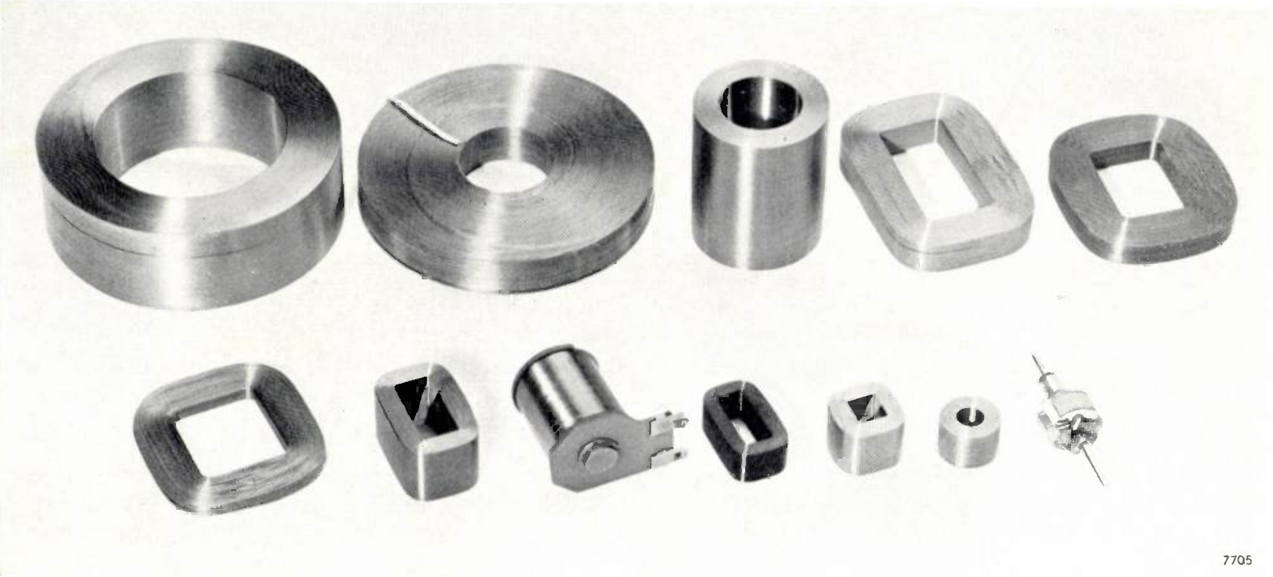
2931: J. A. M. Dikhoff: Cross-sectional resistivity variations in germanium single crystals (Solid-state electronics **1**, 202-210, 1960, No. 3).

Germanium single crystals often show an undesired variation in the dope concentration. It has been found that flat facets can develop on an otherwise curved growth interface during the growth of the crystals. These facets coincide with {111} planes. The reason for the above-mentioned variation in the dope concentration is that the segregation constants of various elements differ for the flat and the curved parts of the growth interface. Methods of producing more uniform crystals are discussed.

VOLUME 23, 1961/62, No. 12 pp. 365-404 Published 16th October 1962

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES



THE ORTHOCYCLIC METHOD OF COIL WINDING

by W. L. L. LENDERS *).

621.3.045.16

Orthocyclic winding is a method developed by Philips for obtaining coils whose turns are stacked in the most compact fashion possible. Such coils have certain particularly good properties including good heat conduction, even distribution of electric field strengths, and the highest possible space factor (which means that the coil dimensions, and hence also the statistical spread therein, are reduced to a minimum).

Any device exploiting electromagnetism contains one or more coils or windings made of insulated copper wire. The performance of the device and certain of its more important characteristics, such as size, weight and price, are to some extent dependent on the success with which the basic ingredients of coil design have been mixed. We are referring to the number of turns, the gauge of the wire, the shape and volume of the coil and the method employed for winding it. This hetrogeneous collection of variables opens the door to a great variety of coil properties. The range of possibilities becomes even greater where, as is usually the case, the magnetic flux produced by the coil passes through a ferromagnetic material, since the magnetic properties of these materials differ considerably and there is further a wide choice in the dimensioning of the magnetic circuit.

One aim that is generally striven for in the coil design is to get the highest possible *space factor*. In this context the space factor F means the ratio between the aggregate cross-sectional area of the turns inclusive of insulation and the area of the rectangle enclosing them ($ABCD$ in *fig. 1*). F is dependent on the shape of the wire (though we shall only concern ourselves here with *round* wire) and on the way in which it is wound. Inevitably there are interstices between the turns, and these represent wasted space, so that F is less than unity even

*) Radio, Television and Record-playing Apparatus Division, Philips, Eindhoven.

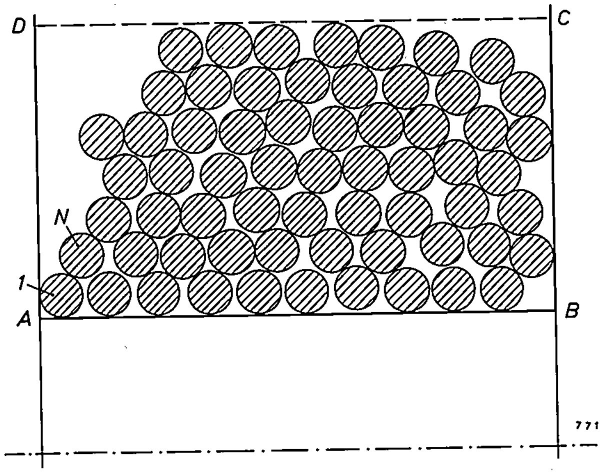


Fig. 1. Cross-section of irregularly wound coil.

in an ideal coil, though it may be only very little less. In a non-ideal coil, in other words in an irregularly wound one (fig. 1), F is considerably smaller than unity. The theoretical maximum space factor can actually be attained with a winding method which has been worked out at Philips and which we call the *orthocyclic* method. The method and its applications form the subject of the present article.

But first of all the importance of a high space factor will be demonstrated: let us take as example a telephone relay or selector embodying a coil which has not been wound by the orthocyclic method, and which provides a required number of ampere-turns at a certain rated voltage. Now let us suppose it proves possible, by adopting a better winding method to accommodate (say) 1.7 times as many turns of the same wire in the same space; the adoption of orthocyclic in place of "wild" winding does in fact result in an improvement of this order. The resistance of the new coil is 1.7 times that of the old; at the same voltage, the current is $1/1.7 \approx 0.6$ times what it was before. The number of ampere-turns remains the same, so that this design requirement continues to be satisfied; but the power consumption of the new coil is only 0.6 times that of the old one. In a telephone exchange, where the number of relays and selectors is large, the resulting economies are appreciable: less power-supply capacity is required, lighter-gauge wiring and cables can be used, and less energy is dissipated as heat.

Coils wound "wild" and coils with interleaved insulation

We shall first discuss two common ways of winding coils. *Fig. 2* shows a coil with the first layer completed: the turns cross the axis of the coil at an angle $90° - \alpha$. They have been wound from left to right, and form a left-handed helix. The second layer will run from right to left; it cannot be wound with the same neatness as the first. It ought to form a right-handed helix. In practice, however, its regularity will continually be disturbed because the wire has a tendency to lie in the grooves between the turns of the first layer, but since the two helices do not interlock, the wire of the second layer can only follow the grooves for short distances. Hence the irregularities will become ever greater as the winding of the coil proceeds, to such a point that it is soon no longer possible to distinguish one layer from the other. It is for that reason that we call this "wild winding".

The advantages of this primitive coil-winding method are that it is quick, that it can be done on a very simple machine, and that little time need be spent on the training of the operator. But examination of fig. 1, which represents a section through a coil wound by the "wild" method, will soon make it plain that this procedure has the following drawbacks:

1) The coil has a much smaller space factor than a neatly wound coil.
2) There will be a considerable random variation in the space factors of such coils. In order to reduce the number of rejects, the design will have to be based on almost the lowest space factor likely to be obtained. The variation in the space factor is associated with a variation in the average length of wire constituting a turn, and hence in the resistance of a coil.
3) Once the winding process starts to "run wild", some turns, particularly those at the sides, will inevitably end up nearer the axis than they

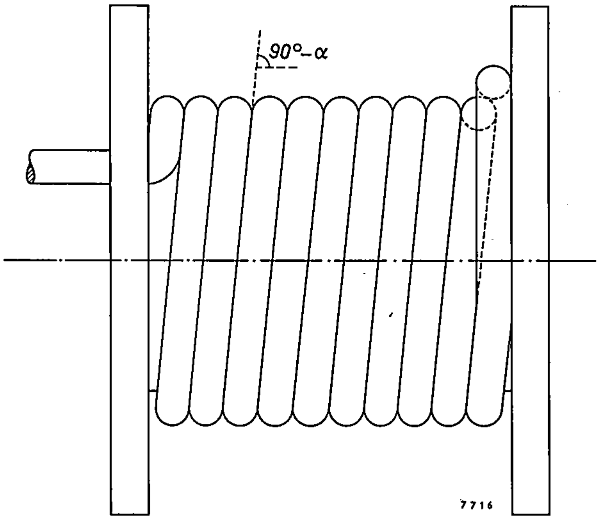


Fig. 2. Single-layer coil wound from left to right and having the form of a left-handed helix. The turns cross the axis of the coil at an angle of 90° — α. Irregularities will develop when the second layer is wound from right to left, and it will not be long before the winding process "runs wild".

ought to be, so that there will be relatively high voltages between turns lying in close proximity.

In consequence of (3) there is no way of knowing U_n, the highest voltage arising between contiguous turns of an irregularly wound coil with a voltage U across its ends; nor, therefore, is there any way of assessing the risk of breakdown in a coil carrying AC and of corrosion in a coil carrying DC. If the coil were regularly wound, all this would be known: for a winding consisting of l layers,

$$U_n = 2\,U/l\,. \qquad (1)$$

(U_n is the voltage between the first turn of one layer and the last turn of the next.) Fig. 1 illustrates the worst case that can arise, that being when the last turn (N) lies on top of the first one (*1*).

To overcome these difficulties, it is a common practice to lay a strip of paper or other insulating material on each layer as it is completed; sometimes the insulation is inserted every two layers. This can be done automatically, so no loss of time is entailed. The great advantage of using interleaved insulation is that even and uniform layers are obtained; the paper prevents the layer which is being wound from being disturbed by the pattern of the layer underneath. The effect is as if a new start is made with each layer. Eq. (1) holds in this case and, moreover, the interleaved paper improves the breakdown voltage. Yet the space factor remains small: the gain obtained by regularity is cancelled out because the paper takes up some space, and because spaces also have to be left on either side of the coil (*fig. 3*). If the layers were wound over the whole breadth of the paper, the end turns would be liable to slip off.

Orthocyclic coils

We have seen that in "wild" coil winding it is impossible for regularity to be maintained after the first layer. The reason is that the tendency of the wire to lie in grooves formed by the preceding layer conflicts with the need for adjacent layers to have opposite helical configurations. The condition relating to orientation can be expressed in another way: if the turns of the first layer lie at an angle of 90° $- \alpha$ to the coil axis (fig. 2), then those of the second layer should lie at an angle of 90° $+ \alpha$. It follows that the difficulty disappears if α can be reduced to zero, in other words, if the first layer can be wound in such a way that each turn, or at least the greater part of each turn, crosses the axis *orthogonally*.

Fig. 4 shows the pattern the first layer would have to assume in this case. Now, with a little care

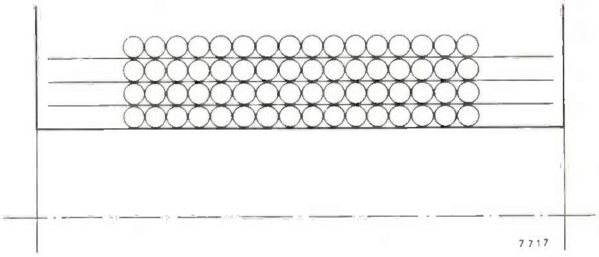


Fig. 3. Cross-section of coil with interleaved insulation. The turns are arranged in an orderly manner but the space factor is still small because the insulation takes up some space, and because some space has to be left free on either side of the layers.

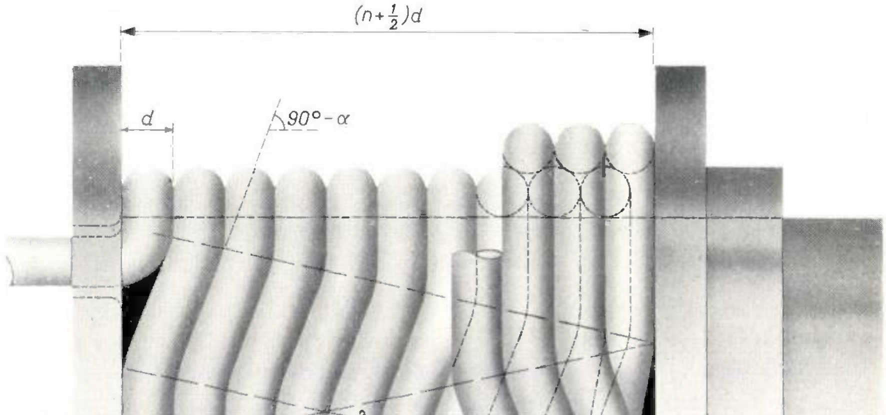


Fig. 4. A round orthocyclic coil at the stage where the first layer and a few turns of the second layer have been wound. Over about 90% of its length, each turn crosses the axis orthogonally. The remaining, inclined portion is known as the crossover; together the crossovers of one layer form a crossover line. Except for their crossover portions, the turns of the second layer lie neatly in the grooves formed by those of the first layer.

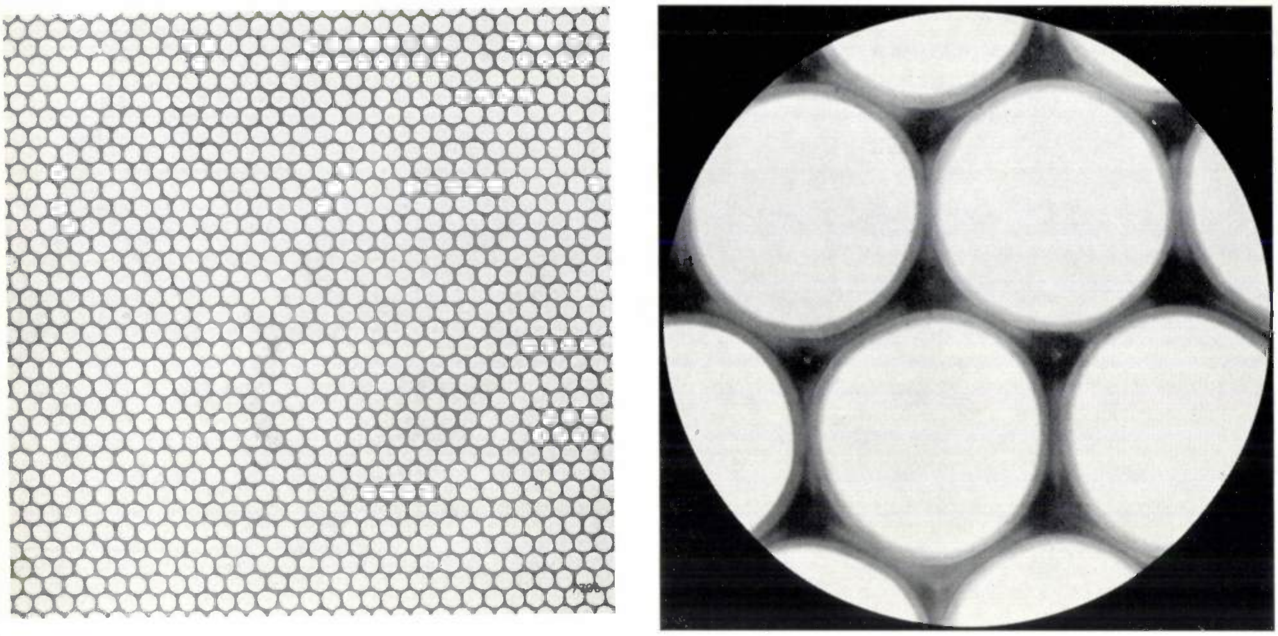


Fig. 5. Section revealed by sawing through an orthocyclic coil wound with copper wire of diameter 50 μm; the photograph on the left is 48 times, that on the right 600 times actual size. In virtue of the perfect orderliness and extreme compactness of the stacking pattern, the highest possible space factor has been achieved.

and ingenuity it is in fact possible to wind the first layer in this fashion, and it has been found that thereafter no difficulty is encountered in stacking succeeding layers evenly; this optimum stacking is the characteristic feature of what we call *orthocyclic winding*, by reason of the fact that the turns must lie orthogonal to the axis. The remarkable regularity and compactness of orthocyclic coils will be evident from the photographs (*fig. 5*) of a section revealed by sawing through a coil of this kind.

Crossovers

Obviously, the orthogonality requirement cannot be satisfied over the whole length of the turn. We use the term *crossover* to denote that section of each turn which lies out of the orthogonal (see fig. 4). Together, the crossovers of one layer form a *crossover line*.

In coils with a circular cross-section the crossover occupies about 10% of the total turn length. The crossover lines run through the coil in a zigzag fashion and their ends, as seen in the cheek of a finished coil, form an Archimedes' spiral (*fig. 6*). In coils with rectangular cross-section, crossovers in the lowest layers occupy one side of the rectangle (*fig. 7*).

Coil thickness and space factor

The thickness of orthocyclic coils is slightly greater in places where crossovers have been made. Let us first consider the places where no crossovers are present. Here it is only the first layer that contributes the full diameter of the wire d to the thickness of the coil. Because they lie in the grooves of the preceding layer, all layers other than the innermost contribute only $\frac{1}{2}\sqrt{3}d$ (*fig. 8*). Accordingly, the crossover-free portion of an orthocyclic coil comprising l layers has a depth of

$$[1+\tfrac{1}{2}\sqrt{3}(l-1)]d\,,$$

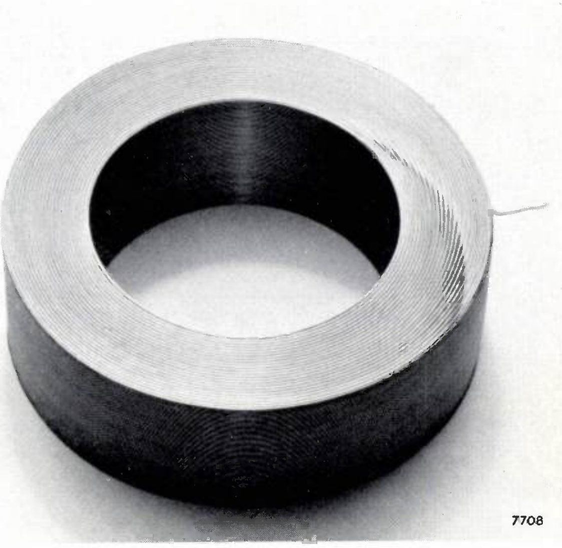


Fig. 6. Round orthocyclic coil. The ends of the crossover lines form an Archimedes' spiral in the cheek of the coil.

which, for a large number of layers, is roughly $\frac{1}{2}\sqrt{3}ld = 0.87ld$.

In considering the portion in which crossovers are present, round and rectangular coils must be dealt with separately. In *round* coils, the number of successive layers whose crossovers overlap does not exceed three; the local increase in coil thickness due to crossovers does not therefore exceed $3(1 - \frac{1}{2}\sqrt{3})d = 0.4d$. Whether this is important or not depends entirely on circumstances. In *rectangular* coils the length of the crossovers, those of the lower layers at least, is the same at that of the side on which they lie (fig. 7*b*). Here, then, the crossovers are stacked one above the other, and the increase in thickness they occasion will be greater than the corresponding increase in a round coil with the same number of layers.

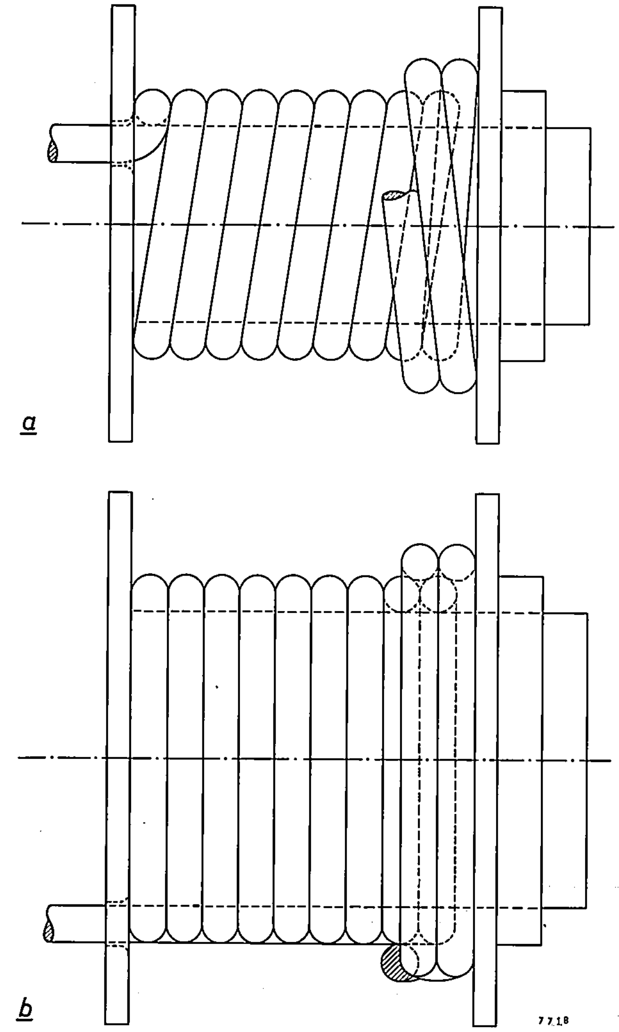


Fig. 7. Two faces of a rectangular orthocyclic coil, (*a*) that on which the crossovers are made, and (*b*) one of three faces on which the turns lie orthogonally.

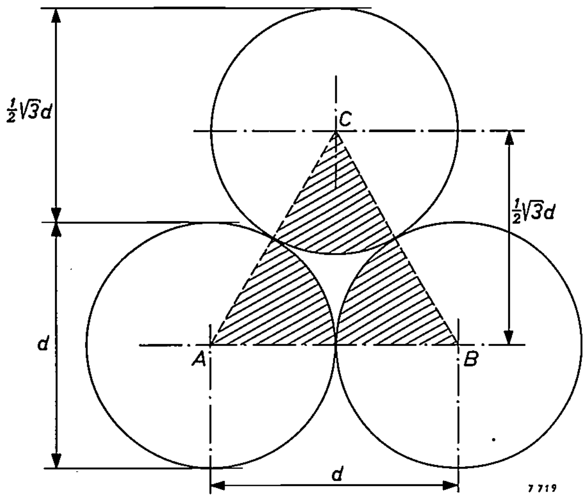


Fig. 8. Section through three contiguous turns in an orthocyclic coil wound with wire having an outside diameter *d*. If the crossovers and the "edge effect" be neglected, the space factor is the proportion of triangle *ABC* occupied by the hatched areas, i.e. 91%.

We shall take the symbol F_0 to denote the "pure" space factor of an orthocyclic coil, in which no account is taken of the crossovers and of a second effect that will be dealt with below. F_0 is easily calculated: it is the ratio of the total area of the three hatched sectors in fig. 8 to the area of the equilateral triangle *ABC*:

$$\frac{\frac{3}{6}\times\frac{1}{4}\pi\, d^2}{\frac{1}{4}\sqrt{3}\, d^2} = \frac{\pi}{2\sqrt{3}} = 0.91.$$

But so far we have ignored the fact that certain turns of the coil are not completely surrounded by other turns; the grooves in the innermost and outermost layers and in the cheeks of the coil are not utilized. The magnitude of this "edge effect" — to which, of course, non-orthocyclic coils are also subject — depends on the shape of the coil's cross-section. It may be important in the case of long, thin coils (in which the number of turns per layer is large relative to the number of layers) and still more so for flat, disc-shaped coils (in which the number of layers is large relative to the number of turns per layer). Still, even in the latter case, the true space factor F is unlikely to fall short of F_0 by more than one or two percent.

Because the space factor remains much the same from one coil to another, orthocyclic coils exhibit a particularly small random variation in external dimensions. The size of the coils is determined entirely by the outer diameter of the wire, the number of turns per layer and the number of layers.

The coil mandrel

Generally coils are wound on a former or bobbin made of impregnated paper or plastic; the bobbin affords mechanical protection and additional electrical insulation. Such bobbins are not as a rule up to the standard of dimensional exactitude required for orthocyclic winding. That greater precision is called for by this coil-winding method will be clear from the fact that each layer must consist of exactly the same number of turns n; the distance between the cheeks of the coil former must therefore be $(n + \frac{1}{2})d$, and experience has shown that deviations from this dimension must not exceed $\pm 0.2d$. This requirement is a severe one, particularly where it is a matter of winding very thin wire (with a diameter of 50 μm, for example).

Orthocyclic coils are not wound on a bobbin, then, but on a rigid metal mandrel that has been carefully machined to close tolerances. It is only after removal from the mandrel that the completed coil is mounted on a bobbin and provided with additional insulation.

The wire

In fig. 4 it has been tacitly assumed that in orthocyclic winding no clearance is left between turns of the same layer. This tight packing means that d must remain within particularly close limits. For example, if at a certain point in the wire its outside diameter should increase by 1%, only 50 turns later it would be riding on top of a turn in the preceding layer, instead of lying in the groove; the next turn would slip off the turn beneath, leaving a gap of $\frac{1}{2}d$. Thus disturbed, the regularity of the coil would be irremediably lost, and the disorder would worsen progressively.

The above tolerance of $\pm 0.2d$ for the layer breadth imposes a tolerance of $\pm 0.2d/n$ for the outside wire diameter. If for example d is nominally 100 μm and there are $n = 120$ turns per layer, then $0.2d/n = 0.17$ μm. In wire manufacture, the bare copper is reduced to the desired gauge by drawing through a die. Consequently the spread in the diameter of bare copper wire on a given reel is negligibly small [1]). An "enamel" covering is the only kind of insulation suitable for wire that is to be used for orthocyclic coil winding. The deposition of the "enamel" (which is really a type of varnish) is an easily disrupted process which must take place under strictly defined electrical, chemical and mechanical conditions [2]). It may be regarded as something of an achievement that ways and means have been found of keeping the spread in d to within ± 1% for a given reel of enamelled wire. Yet this is still six times more than the tolerance required in the above example.

There are two expedients whereby enamelled wire with an excessive spread in d can nevertheless be utilized for orthocyclic winding: a clearance can be left between turns, or the wire can be redrawn.

1) The variation in d can be allowed for by winding the first layer with a small clearance between the turns. It has been found in practice that the clearance must not exceed $0.03d$, otherwise the turns of the second layer will push those of the first layer too far apart. As can be imagined, the maintenance of an inter-turn clearance of a few microns calls for a rather complicated winding machine which accurately controls the pitch of the winding. We shall return to this problem later.
2) Wire exhibiting an excessive spread in d can be rendered suitable for orthocyclic winding *without* inter-turn clearances by redrawing it, i.e. by passing the enamelled wire through another die. The aperture of the die used must have a diameter equal to the minimum value of d, so that all thicker portions of the wire are reduced to this particular outside diameter.

Redrawing involves a perhaps somewhat surprising phenomenon: the variations in d are transferred, with a change of sign, to the copper core; in places, then, the diameter of the core is reduced in the same proportion as the outside diameter of the enamel covering (*fig. 9*), and this has two undesirable consequences:

a) The copper hardens on account of plastic deformation and so loses some of its pliability; consequently the crossovers tend to lengthen, and it is less easy to control their positioning.

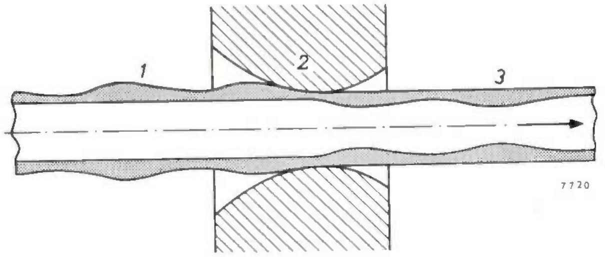


Fig. 9. Schematic representation of enameled wire undergoing the redrawing process. The aim is to get insulated wire whose outer diameter is everywhere the same. Besides producing the desired effect, redrawing transfers to the copper core the variations that were originally present in the outer diameter. *1* represents the wire still to be redrawn, which has a constant core diameter but exhibits variations in outer diameter, *2* is the drawing die, and *3* is the redrawn wire, which has a constant outer diameter but exhibits variations in core diameter. The thickness of the insulation and the variations therein have been exaggerated for the sake of clarity.

[1]) Slightly bigger differences of diameter can naturally be expected in wire from different reels, which has been drawn through different dies.

[2]) R. J. H. Alink, H. J. Pel and B. W. Speekman, Manufacture and testing of enamelled wire, Philips tech. Rev. **23**, 342-351, 1961/62 (No. 11).

b) Since the diameter of the copper has been reduced in all the places where d was in excess, the electrical resistance per metre will have increased; if d was on average $a\%$ too large, the resistance increase will be about $2a\%$.

Attention is being devoted, in wire manufacture, to making the spread in d even smaller than it is at present, so that redrawing will involve less deformation, the above-mentioned disadvantages of the treatment thus being minimized.

In consequence of the variation in the core diameter of redrawn enamelled wire, orthocyclic coils wound from such wire exhibit a certain variation in resistance values. This, however, is generally smaller than the variation in the resistance of coils wound by the "wild" method. In the latter case the lack of uniformity in coil resistance is due, as stated above, to the considerable spread in the space factor and hence in the total length of wire that goes into the coil.

Orthogonal wire feed arrangements in coil winding

As we have seen, the characteristic feature of an orthocyclic coil is that the greater part of each turn lies in a plane at right angles to the axis. Accordingly, feed arrangements for the winding machine must be such that the wire coming off the reel likewise remains at right angles to the coil axis, at any rate within close limits. This aim might be achieved by dispensing the wire via a sheave wheel set up at a sufficient distance from the machine; in practice, the wheel would have to be a good many metres away. Apart from the inconvenience of the set-up, vibration in the long stretch of wire between the sheave and the machine could interfere with the regularity of the winding process. For that reason we make use of a guiding sheave which is free to move laterally, and which maintains the orthogonal feed direction automatically without having to be set up at a considerable distance from the machine. The device is shown schematically in *fig. 10*.

"Thermoplac" wire

The finished orthocyclic coil would fall apart on removal from the winding mandrel if it were not for the fact that "Thermoplac" wire [3]) is used in orthocyclic winding. The insulation of this wire has a thin thermoplastic coating. Before the coil is taken off the mandrel, it is warmed in order to fuse and bond the coatings on contiguous turns. Cooling results in a sturdy self-supporting coil. The thermoplastic film need only be about 2 μm thick. It can withstand the drawing process.

All the following methods are suitable and are in fact employed for heating the coil while still on the mandrel.

a) Baking in an oven.

b) Induction heating in an HF magnetic field (most of the heat is generated in the mandrel and passes to the coil by conduction).

c) Heating the coil by passing through it a current from an external source.

d) Short-circuiting the coil and inducing in it an alternating current (normally of mains frequency).

In methods (c) and (d), as against (b), heat is developed in the coil itself and the mandrel acts as a sink. It is therefore necessary, if either of these

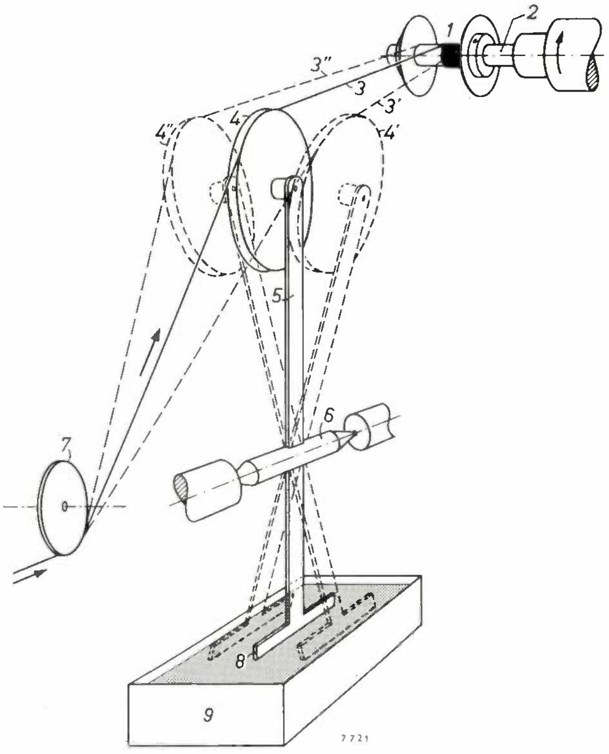


Fig. 10. Automatic arrangement to ensure that wire coming off a reel (not shown) will at all times be orthogonal to the coil axis, irrespective of the position of the turn being wound. The coil *1* is being wound on a mandrel turning about axis *2* and fed with wire *3* via a sheave wheel *4* on an arm *5* pivoting on a pin *6* that is orthogonal to *2*. Wire from the reel goes to *4* via a fixed sheave *7*. The point at which it leaves *7* is in line with *6*. By reason of the tension in the wire, sheave *4* automatically takes up position *4'* when a new layer is started, and moves to position *4''* as the layer is completed. If the small deviation due to the finite length of arm *5* be discounted, the wire dispensed to the mandrel remains orthogonal to axis *2* as it repeatedly passes from position *3'* via position *3* to position *3''*.

An extension to arm *5* carries a vane *8* which is immersed in an oil bath *9*. The oil provides enough damping to prevent transverse oscillations of the arm.

[3]) See the article cited in footnote [2]), p. 347.

two methods is used, to ensure adequate heating of those parts of the coil which are in contact with the mandrel while guarding against overheating of the remainder. In practice this means that the desired temperature must be attained within a few seconds. This is quite feasible with method (c) provided the resistance of the coil is low enough; fast heating of a coil having a high resistance would require a voltage involving the risk of breakdown between the first and last turns via the metal mandrel. Method (d) does not endanger the insulation of the coil because the terminals are shorted. It does, however, necessitate a core carrying a primary winding marrying with the type of coil being manufactured, and the cost of this will only be acceptable if the coils are to be turned out in sufficient quantity.

A closer look at certain aspects of orthocyclic winding

The exact manner in which orthocyclic winding can best be done depends on various circumstances, among these being the shape of the coil (round or rectangular) and the batch size (a small number of coils or enough to justify continuous production). Here we shall divide orthocyclic coil-winding procedures into two classes, those in which the first layer is wound and the pattern established by hand, and those in which this initial phase is done by machine. The two classes correspond more or less to the two cases of production on a small and on a large scale respectively.

Patterning the first layer by hand

If done by hand, orthocyclic patterning of the first layer calls for a certain amount of skill on the part of the operator. This is not a serious drawback so long as the number of coils to be made is small. However, only redrawn wire can be used, since it would be impossible, by hand, to maintain the small inter-turn clearance that is necessary when winding with wire whose outside diameter has not been corrected.

In these circumstances a simple winding machine will suffice, consisting of a spindle running easily in good bearings, a source of mechanical power and a revolution counter. The mandrel must have one fixed and one movable flange, the position of which can be accurately adjusted; the whole must have been carefully machined. Orthogonal wire feed arrangements are necessary (fig. 10). Redrawing of the wire can be combined with the winding operation.

Once the first layer has been wound and the movable flange correctly set, winding can be continued mechanically; it has been found in practice that the subsequent turns automatically adopt the correct "lie", the same regular pattern repeating itself from layer to layer. But care must be taken to see that the regularity of the winding process is not disturbed by incidental troubles such as vibration in the wire coming off the reel and — when very thin wire is being wound — the intrusion of dust particles. Insofar as it is not limited by such interfering effects, the winding speed can be stepped up to about 10 000 turns per minute from the second layer onward.

Forming the pattern mechanically

Mass-production requirements have led to the mechanization of the process whereby the orthocyclic pattern is established in the first and repeated in subsequent layers of the coil. The resulting winding machine is considerably more complicated, as will be clear from the account that now follows.

The separation of the mandrel flanges is so adjusted that there is just enough room for the prescribed number of turns per layer when the wire has its maximum diameter. By leaving a small clearance between turns it is possible to achieve even distribution, over the same breadth, of wire which has not been redrawn and whose diameter is (say) 3% below the maximum value. Even spacing of the turns is effected automatically by a shuttle arrangement that moves intermittently along a line parallel to the axis of the coil.

During the winding of the first layer, the shuttle is motionless for part of each revolution of the mandrel; it is during this interval that the orthogonal portion of the turn is produced. Oversimplifying for a moment, we can say that immediately a crossover is due to begin, the shuttle moves abruptly over a distance s equal to the wire diameter d plus the inter-turn clearance t.

This simplified description only holds good for a shuttle that is close to the mandrel — between its flanges, in fact. But if this was so in practice it would be impossible for the layer to extend from flange to flange, since the breadth of the greave must necessarily be greater than the diameter of the wire it guides on to the mandrel. Accordingly, the distance between the shuttle and the mandrel axis must be at least half the mandrel flange diameter. But if it is placed further away from the mandrel, each time the shuttle advances it must temporarily overstep its cumulating displacement $s = d + t$ by an amount s_e; see *fig. 11*, which shows the crossovers on a rectangular coil. If the shuttle did not travel this extra distance it would pull the wire against the preceding turn, the inter-turn clearance would be locally reduced to zero (along a line running transversely over the layer), and this part of the layer would be "waisted", the desired regularity thus being lost. The result, in a round coil, would be that the crossover became longer with every successive winding; in a rectangular coil with all the crossovers on one side of the rectangle, the layers would be narrower along one edge than along the other three.

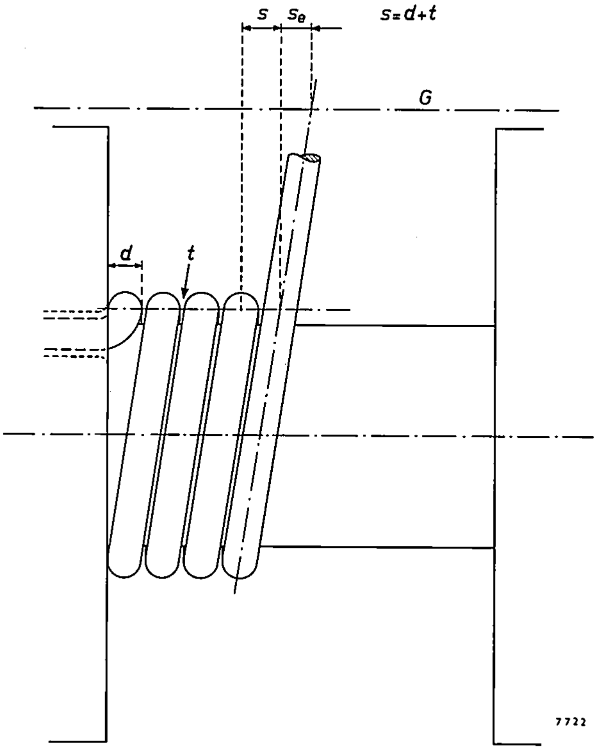


Fig. 11. Winding of the first layer of a coil with an inter-turn clearance t. Shuttle movement takes place along line G. Each time a crossover is begun, the shuttle, having undergone displacement through an interval $s = d + t$, must continue its forward movement, advancing over a further distance s_e, through which it will return as soon as the crossover is completed.

On completion of each crossover the shuttle must move back over the distance s_e.

There is no need for this extra shuttle displacement during the winding of the second and subsequent layers, the "lie" of their turns being determined by the grooves in the preceding layers.

While winding each turn of the innermost layer, then, the shuttle has to perform a somewhat complicated movement. The time available for its performance is the time taken by the mandrel to turn through the angle within which the crossover is made. This sets a limit to the winding speed. The crossover angle is 90° for square coils, and rather more or rather less for rectangular coils, depending on whether the crossovers are being made on the long or short side of the rectangle; for round coils the crossover angle is only 30° to 40°. It follows that the first layer of rectangular coils can be wound more quickly than the first layer of round ones.

Reasons why crossovers in rectangular coils must be confined to one side of the rectangle

Fig. 4 shows how the crossover lines run in zigzag fashion over the surface of a round coil. In rectangular coils the crossover lines have a tendency to do the same thing; but it is necessary to confine them to one side of the rectangle, as in fig. 7*a*, for otherwise the coil cannot remain orthocyclic. This is due to the fact that the length of the crossover, and hence also the inclination of the crossover line with respect to the coil axis, is dependent on the radius of curvature of that part of the turn which constitutes the crossover. If the crossover line should turn the corner, as it were, invading an adjoining face of the coil, it would do so at a point where the radius of curvature of the wire changes abruptly, and the angle of the crossover line with respect to the axis would also alter. The effect is as if the crossover line is bent by the corner of the rectangle, with the result that one turn is missed out. Starting at this point, the irregularity worsens progressively as further layers are added. Finally the winding process "runs wild".

There is a second reason why it is desirable to keep all crossovers on one face of a rectangular coil. Take for example the winding of a transformer with a core made up of E and I-shaped laminations: two sides of the rectangular winding, which we may call A and C, have to be accommodated inside the openings in the core, and the winding depth should therefore be as small as possible on these two sides. The endeavour will accordingly be to confine the crossovers to B or D, the sides of the rectangle lying outside the core.

A special winding mandrel, designed to provide effective control of crossover lines, enables all crossovers to be confined to one side of rectangular coils. The ends of the crossover lines form a pattern in the cheek of the coil, revealing whether the crossovers do in fact lie on one side of the rectangle, or whether they have invaded an adjoining side. They have done so in coil *a* in *fig. 12*, but not in coil *b*.

On sides A and C the winding — we still have a transformer coil in mind — will bulge slightly, to an extent dependent on the tension in the wire coming off the reel, and will therefore take up more space than necessary in the core openings. The bulge can be evened out by compressing the sides in question either during or after the heat treatment. There is no risk of damaging the insulation because

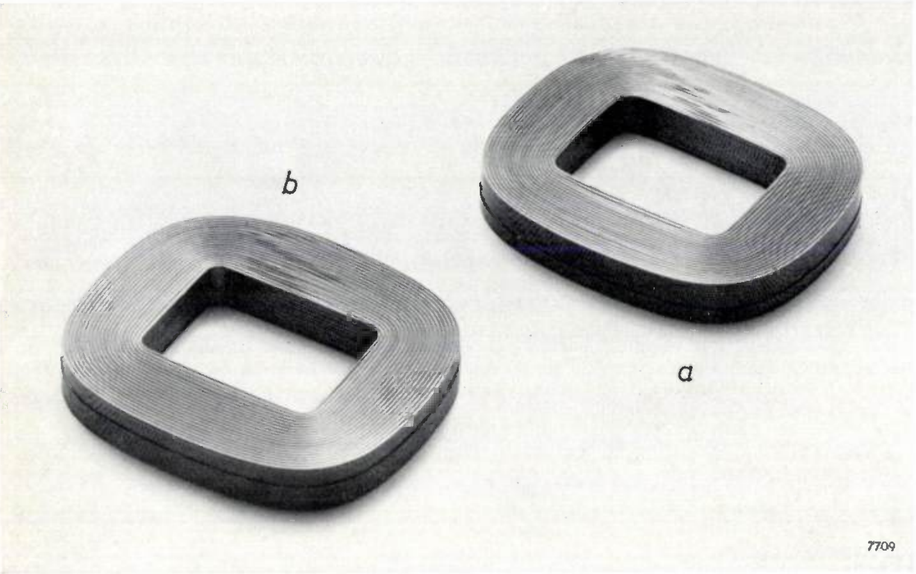


Fig. 12. Rectangular orthocyclic coils. The patterns traced in the cheeks of the coils by the ends of crossover lines reveal that in coil *a* the crossovers have "turned the corner", but in the case of coil *b* have been confined to one side.

sides *A* and *C* are free of crossovers. Coil *a* in *fig. 13* has not been compressed; coil *b* has been squared up on three sides.

Orthocyclic coils compared with types wound by other methods

The foregoing will have made it clear that orthocyclic coils possess two big advantages as compared with other kinds:

1) In orthocyclic coils the space factor attains the theoretical maximum value. In consequence, the random variation in the space factor, and hence in the overall dimensions of the coil, is extremely small.
2) Since each turn of an orthocyclic coil lies in its appointed place, the voltages between given pairs of turns are well defined; moreover, the greatest voltage between contiguous turns is as given by eq. (1). It is therefore possible to ascertain in advance whether $U_{\rm n}$ will be too high for the wire-covering and whether, accordingly, interleaved insulation will have to be used (in which case there will not of course be any point in winding the coil orthocyclically).

Further important properties which distinguish orthocyclic coils from other kinds will now be enumerated.

3) *Permanence of shape*

Thanks to the thermoplastic bonding agent the turns of an orthocyclic coil coalesce into a very sturdy unit, much stronger than a coil that has been wound "wild" out of wire having a thermoplastic coating of the same thickness. This greater robustness is due to the fact that the turns bind one another over their whole length, and it enables orthocyclic coils to withstand considerable mechanical stresses without loosening of the turns. In virtue of this property orthocyclic coils are particularly suitable for applications in which they are subject to severe vibration, as is the case with the rotor windings of electric motors and moving coils for loudspeakers. We shall revert to the subject of orthocyclic loudspeaker coils below, when dealing with individual applications.

4) *Good heat conduction*

The heat developed in a coil has to be dissipated via its outside surfaces. There is therefore a temperature gradient in the coil. With an eye to the life of the insulation, a ceiling must be fixed for temperatures arising in the coil [4]); but normally the point where the highest temperature prevails is

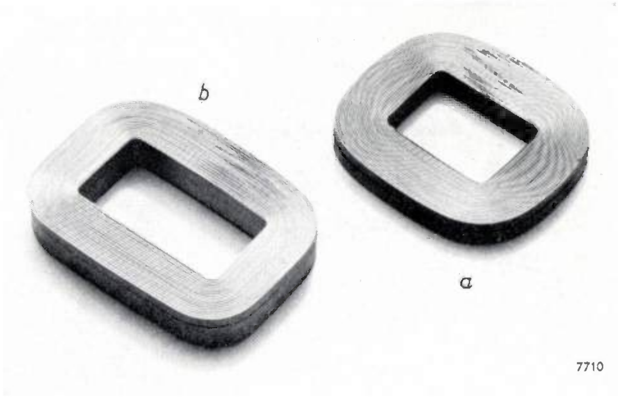


Fig. 13. Rectangular orthocyclic coils in which crossovers have been confined to one side of the rectangle. The faces of coil *a* bulge slightly. The three faces of coil *b* not containing crossovers have been pressed flat, with some saving of space as a result.

[4]) In regard to the connection between life and temperature, see for example T. Hehenkamp, The life of ballasts for gas-discharge lamps, I. Transformers and chokes, Philips tech. Rev. **20**, 59-68, 1958/59.

inacessible for measurement, and in practice, by way of compromise, the average temperature is determined, this being worked out from the easily measured increase in electrical resistance resulting from the warming-up of the coil. The highest temperature arising in the interior of a non-orthocyclic coil may easily be 10 °C above the temperature of the outside surface and 5 °C above the average temperature.

Orthocyclic coils conduct heat so well that in normal use no perceptible difference of temperature arises between the interior and the outside. The thermal conductivity of orthocyclic coils, measured perpendicular to the wire, was found to be 1 W/m^2 per °C/m (or 0.85 kcal/mh°C), which is 20 times the value for coils with interleaved paper insulation. Hence, if a non-orthocyclic coil exhibiting the temperature differences just cited is replaced by an orthocyclic coil, the insulation of the latter will not be endangered if its temperature is allowed to rise 5 °C above the average temperature of the first coil. That higher temperatures are permissible in orthocyclic coils is fortunate in view of the fact that, on account of their better space factor, they occupy a smaller volume and are thus likely to have a smaller surface area than non-orthocyclic coils. By reason of better heat conduction an orthocyclic coil is warmer on the outside than a non-orthocyclic coil with the same internal temperature, and is thus able to dissipate heat at a given rate through a smaller surface area.

5) *High quality factor*

A direct consequence of the high space factor of orthocyclic coils is that less wire of a given gauge is needed to wind a coil with a given number of turns. On account of the smaller winding length the ratio of resistance R to inductance L is smaller than in non-orthocyclic coils; the quality factor $Q = \omega L/R$ is therefore relatively high.

6) *Small variation in inductance and self-capacitance*

In orthocyclic coils each turn has its exactly defined position, and in consequence of this the inductance and self-capacitance differ very little from the nominal values. This is of importance in connection with centre-tapped transformers, which, in telephony and other fields, have to satisfy high standards of electrical symmetry. If such transformers are wound by ordinary methods, exact balance can only be achieved by having recourse to trimming capacitors. These are generally superfluous if the transformer has been wound by the orthocyclic method.

Disc-shaped coils, consisting of narrow layers stacked to a considerable height, are called for in cases where very low self-capacitance is essential.

The random variation in the resistance of the coils has already been mentioned (page 371).

Applications of orthocyclic coils

Our factories have now been employing the orthocyclic winding method for several years. In the early days, of course, knowledge of certain factors affecting the product — tolerances for wire thickness and mandrel dimensions, the correct pattern for the first layer, and the like — was still incomplete. Because these matters had not yet been fully investigated, more had to be asked of the coilwinder in the way of skill and insight into the process than was required by the old, familiar method. For that reason a long training period was necessary, output was inclined to be small, and the rejection rate was sometimes high. As a consequence, at that time orthocyclic coils were comparatively expensive to make, and there could be no question of producing them in quantity. Initially, then, applications of the technique were limited to a few products in the professional class. At present all factors influencing the orthocyclic winding process are well under control, so much so that the method is being adopted in mass production.

A few examples of applications in both categories will now be given.

Professional applications

1) The 600 MeV CERN synchrocyclotron at Geneva is equipped with a modulator that causes the frequency of the accelerating voltage to sweep periodically (55 times per second) over a range extending from about 29 to 16.5 Mc/s. The frequency-modulating device is a gigantic "tuning fork" that acts as a vibrating capacitor [5]). The fork is excited by the field of a set of orthocyclic coils carrying current with a frequency of 27.5 c/s. The whole assembly is in a vacuum.

The main reasons for choosing orthocyclic coils were their high space factor (space was very limited), good heat dissipation and ability to stand up to vibration. During evacuation of the vacuum chamber it was found that the coils released very little gas, the pumping time thus being particularly short.

2) In incandescent lamp manufacture, the bulbs are filled with gas via an electromagnetic valve.

[5]) B. Bollée and F. Krienen, The CERN 600 MeV synchrocyclotron at Geneva, III. The tuning-fork modulator, Philips tech. Rev. 22, 162-180, 1960/61 (No. 5). — The excitation system is shown in figs 11 and 12 of that article.

For constructional reasons it was highly desirable that the magnet dimensions be reduced to a minimum; an orthocyclic solenoid was therefore indicated.

3) For monitoring steam turbines, electronic equipment has been developed which keeps a continuous record of vibration amplitudes and the dimensional changes affecting various parts of the turbine [6][7]). For example, the eccentricity of the shaft is measured with the aid of a displacement pick-up working on the inductive principle, as illustrated in *fig. 14.* The pick-up windings form part of an AC bridge circuit. For zero eccentricity the bridge should be in balance; any deviation from dead true running should give rise to a signal proportional to the eccentricity of the shaft. It often happens in practice, however, that the temperatures and hence also the resistances of the different windings differ considerably. Consequently the bridge circuit is not always exactly in balance for zero eccentricity, and the signal it yields is not dependent on shaft eccentricity alone. The lower the resistance of coils having a given inductance value — the higher their Q factor, in other words — the smaller the error. In this respect, as we have seen, orthocyclic coils and windings have the edge over non-orthocyclic ones.

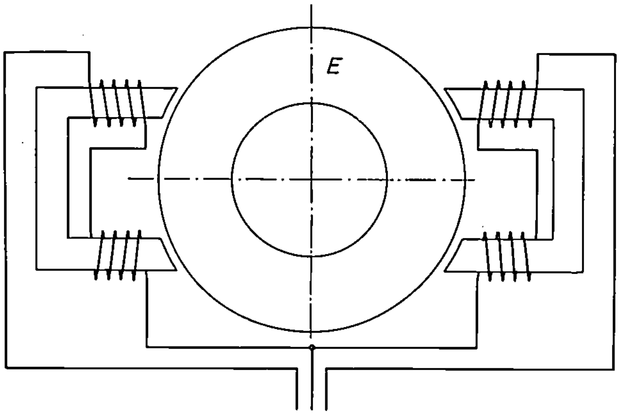


Fig. 14. Inductive pick-up for the continuous detection of eccentricity in a turbine shaft [6]). A disc E of nickel-iron alloy is mounted on the shaft. On either side of the disc is an iron yoke carrying windings that form part of an AC bridge circuit. Provided the two pairs of windings have the same resistance, the signal originating in the bridge will be proportional to the eccentricity of the shaft. A disparity in their resistances, due to temperature differences, gives rise to a measuring error. The higher the Q factor of the windings, the smaller this error is, and that is one of the reasons why the windings are made by the orthocyclic method.

The same considerations apply to other inductive types of displacement pick-ups used on turbines (see figs 9 and 10 of the above-mentioned article [6])).

4) Focusing coils for travelling-wave tubes. In this form of electron tube — one of the kinds used to generate or amplify centimetre waves — an electromagnetic wave propagates itself along a metal helix, through the centre of which passes an electron beam focused by a magnetic field [8]). The field must possess a certain distribution in the axial direction, and it must satisfy stringent requirements in regard to rotational symmetry. In most travelling-wave tubes the field is supplied by permanent magnets. However, in some high-power types a very strong field is required, extending over a considerable distance along the axis, so that an electromagnet is the only practicable means of focusing the beam. The desired distribution of field strength along the axis can be obtained by setting up three differently shaped coils side by side, two disc types being placed at either end of a long tubular coil. Besides offering the familiar advantages of small dimensions with a negligible statistical spread, round orthocyclic coils, by reason of their regular structure, satisfy the rotational symmetry requirement much better than do coils wound by the "wild" method.

Mention may be made here of one of the few cases in which orthocyclic winding was not a success. During design studies for a large X-ray apparatus in which the HT transformer was to be combined with the tube in a single block, it became evident that a smaller and lighter transformer would be desirable. Accordingly, the secondary winding with its interleaved paper insulation was replaced by an orthocyclic winding. However, this secondary had the drawback of excessive self-capacitance. It is true that this difficulty could be got over by substituting for the single secondary a set of flat orthocyclic disc windings connected in series, but the voltage between neighbouring windings was then so high (of the order of 10 kV) that corona and flashover gave rise to further difficulties. Adequate insulation between the disc sections would have meant spacing these so far apart as to cancel out completely the gain obtained by winding them orthocyclically.

Applications in mass production

Over recent years our coil winders have acquired such a degree of expertise in the orthocyclic tech-

[6]) C. von Basel, H. J. Lindenhovius and G. W. van Santen, Electronic equipment for the continuous monitoring of turbines, Philips tech. Rev. **17**, 59-66, 1955/56.

[7]) C. von Basel, Messanlage zur Überwachung von Dampfturbinen, Arch. techn. Messen V 8232-2, 221-224, October 1956.

[8]) J. G. van Wijngaarden, A travelling-wave tube for the frequency band of 3800 to 5000 Mc/s, Le Vide **15**, 36-40, 1960.

nique that there is a growing tendency for it to be employed on a large scale in mass production. A few examples will now be given.

1) Selectors for telephone exchanges. The type U 45a high-speed uniselector (*fig. 15*), made by N.V. Philips' Telecommunicatie-Industrie, Hilversum [9]), embodies one electromagnet which couples the contact wipers to a continuously rotating shaft, and another which arrests them at the desired place. Both the coupling and the stopping magnets are energized by orthocyclic windings, which, as we saw at the beginning of this article, have the great advantage of saving power and so reducing the amount of heat developed, the numerous selectors and relays usually being the main source of unwanted heat in telephone exchanges.

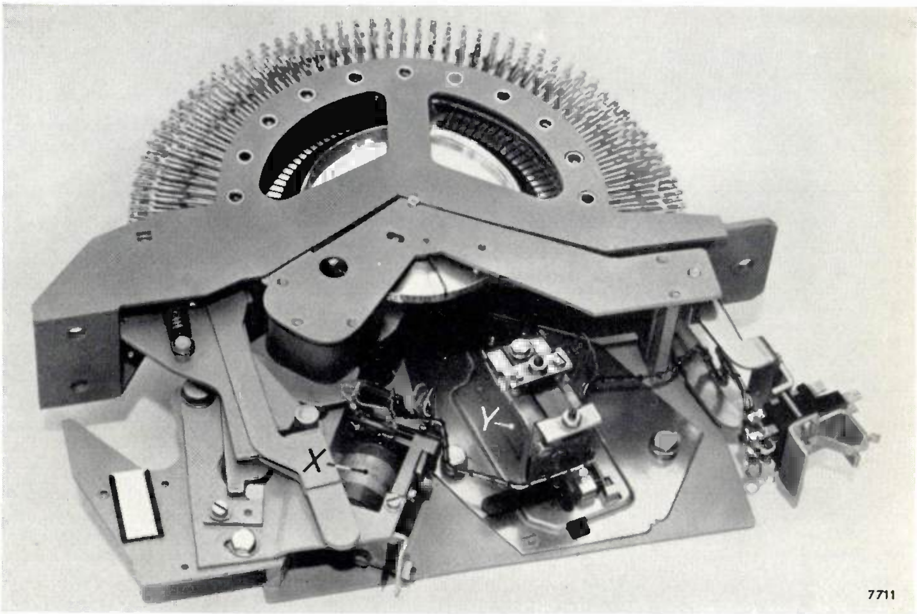


Fig. 15. Type U 45a high-speed uniselector [9]). Both coupling magnet X and stopping magnet Y embody orthocyclic windings.

2) Moving coils for loudspeakers are wound on a miniature bobbin that is subsequently glued to the cone. Thin-walled as the bobbin may be, it still takes up a certain amount of space in the air gap of the magnet, where even fractions of a millimetre count. Furthermore, the dimensioning of the gap has to allow for the deviations from exact circular shape which are exhibited by most bobbins in consequence of their lack of rigidity in the radial direction.

Thanks to a method which has been developed for attaching self-supporting coils to loudspeaker cones, and which provides a particularly strong bond between the two, an immediate advantage of winding the moving coils by the orthocyclic technique is that the bobbin can be done away with. If the coil consists of several layers, as it is bound to do if destined for a high-impedance speaker [10]), then orthocyclic winding will also reduce its thickness appreciably. In addition, the radial rigidity of orthocyclic coils is much greater than that of coils of older design. For various reasons, then, a narrower air gap suffices for an orthocyclic moving coil, and this pays off either in enhanced efficiency or in reduced magnet size. Another advantage is that orthocyclic coils are better able to stand up to vibration.

3) Self-supporting (but non-orthocyclic) moving coils have long been used in electrodynamic microphones. If the moving coil is wound orthocyclically the air gap can be narrowed, with a resulting increase in the sensitivity of the microphone.

4) In the winding of field coils for moving-iron voltmeters and ammeters, the change-over to the orthocyclic technique has allowed so many additional turns to be accommodated that the sensitivity of the meters is 2 to 2.5 times as great as before.

5) Being battery-fed, portable electronic products must be small in size and have the lowest possible power consumption. When an electric motor for battery gramophones and tape recorders was being designed, it was found that an orthocyclic rotor winding saved so much space that a much lighter rotor could be made if the "wild" method were abandoned in favour of the orthocyclic one. Also, because less iron was present, losses due to eddy currents were smaller, resulting in a good 25% saving in power consumption. A second advantage of orthocyclic rotor windings is the negligible variation in their weights, in virtue of which they are unlikely to be responsible for unbalanced distribution of rotor masses. *Fig. 16* shows the rotor with its three orthocyclic windings.

6) Ballasts for "TL" lamps in aircraft. By reason

[9]) J. M. Unk, A high-speed uniselector for automatic telephone exchanges, Philips tech. Rev. **18**, 349-357, 1956/57.

[10]) J. Rodrigues de Miranda, Audio amplifiers with single-ended push-pull output, Philips tech. Rev. **19**, 41-49, 1957/58.

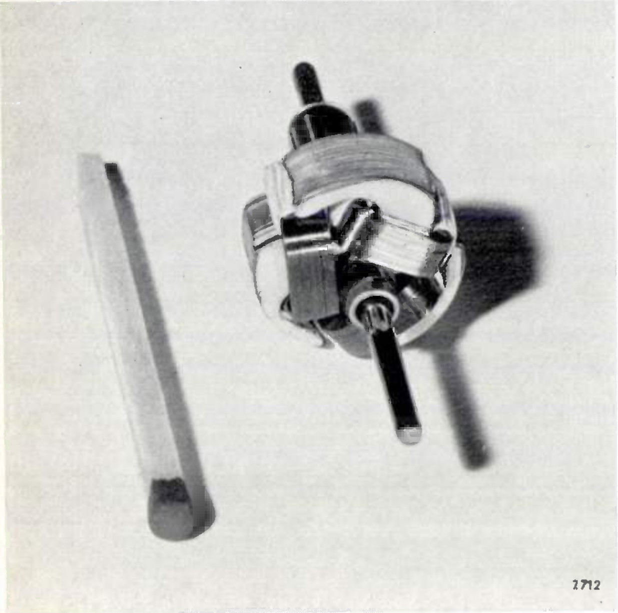


Fig. 16. Rotor of the electric motor incorporated in battery-operated portable gramophones and tape recorders. Employment of the orthocyclic technique for the three armature windings has allowed the size of the motor and its power consumption to be reduced.

of their high efficiency tubular fluorescent lamps are being employed more and more in the public transport sector [11]), including aviation. Passenger-carrying aircraft can now be equipped with (say) twenty 40 W "TL" lamps. In common with everything else that goes to equip aircraft, the weight of the ballasts must be cut to an absolute minimum. Employment of the orthocyclic technique for winding the choke coil has made it possible to bring down to 195 grams the weight of the ballast (consisting of a choke coil and a capacitor) for a 40 W lamp, rated for 115 V at 400 c/s, the standard aircraft electricity supply. The ballast in question may be seen in *fig. 17*.

7) Electromagnetically-controlled air valves. A

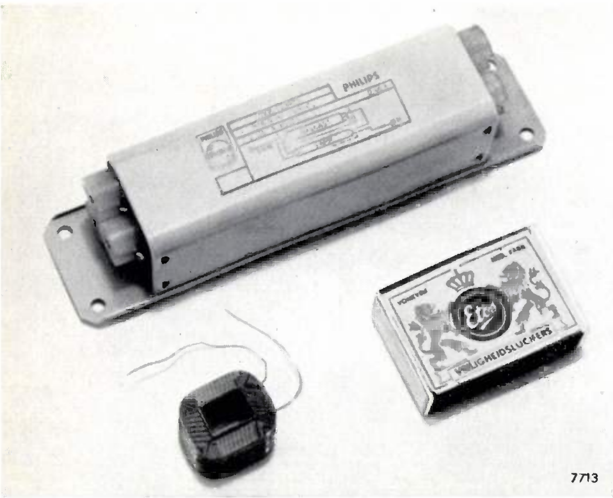


Fig. 17. Ballast for a 40 W "TL" lamp to work on the standard aircraft supply of 115 V at 400 c/s. Altogether, the choke coil and associated capacitor weigh only 195 g. Below, left, the orthocyclic choke coil.

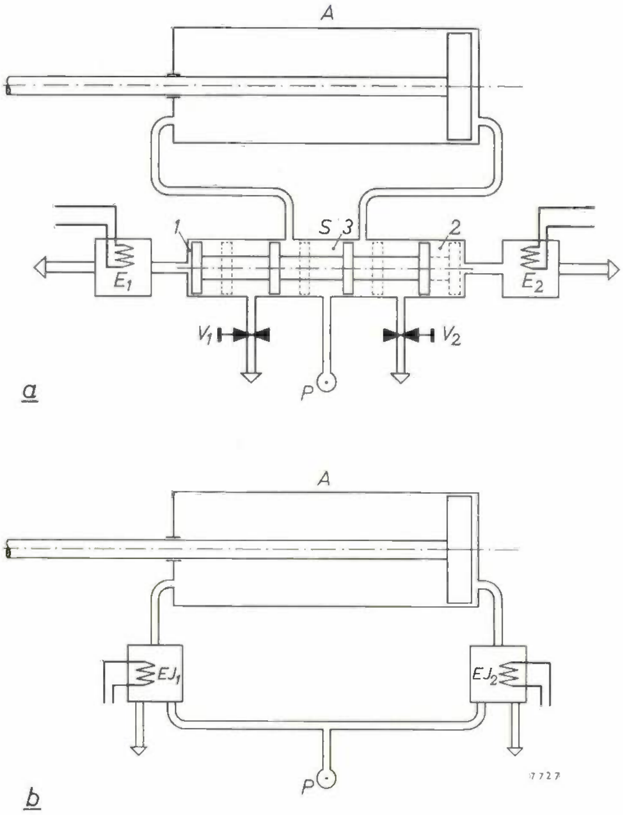


Fig. 18. *a*) Air cylinder *A* and associated pneumatic equipment. *S* piston valve, E_1-E_2 solenoid-operated air-control valves, V_1-V_2 adjustable bleed chokes. Compressed air enters the system from pipe *P*. When the pistons in valve *S* occupy the positions indicated by fully-drawn rectangles, the space in *A* to the left of the main piston is connected to *P*, and the air to the right of the main piston can escape via V_2. In these circumstances the cylinder piston is at the extreme right of its stroke. Spaces *1* and *2* in *S* are connected (by ducts not shown in the diagram) to space *3*, the pressure in which is at all times that of the compressed-air supply. So long as control valves E_1 and E_2 are closed, then, the air in *1* and *2* is at the same high pressure as that in *3*.

To shift the cylinder piston to the left, control valve E_2 is opened, whereupon air escapes from *2*, the valve pistons move into the positions indicated by the broken-line rectangles, compressed air from *P* is free to pass via *3* into the space on the right of the cylinder piston and the air in the left-hand side of *A* exhausts through V_1. The speed with which the cylinder piston travels to the left can be regulated by adjusting V_1. To return the cylinder piston to the right, E_2 is closed and E_1 is opened.

b) In the new system, control of air cylinder *A* is effected solely by means of EJ_1 and EJ_2, two "Electrojet" valves made by the Martonair Company. These supersede air control valves E_1 and E_2 as well as piston valve *S*, and accordingly their bore must be at least as great as that of *S* in (*a*). Having adjustable outlets, the "Electrojet" valves are further able to take over the function of bleed chokes V_1 and V_2. They embody orthocyclic coils, these having been chosen on account of their small size and good heat dissipation.

[11] L. P. M. ten Dam and D. Kolkman, Lighting in trains and other transport vehicles with fluorescent lamps, Philips tech. Rev. **18**, 11-18, 1956/57.

favourite practice in machine building is to fit an air cylinder to supply mechanical power for certain movements which, as a rule, are initiated by switching an electric circuit. Until recently the air cylinder, essentially a simple device, had to be used in conjunction with a pneumatic circuit comprising, in addition to a filter and lubricating unit, a piston valve controlled by one or two solenoid-operated air control valves, one or two adjustable bleed chokes and the associated piping and fittings. An example of a circuit of this kind (many variations are possible) is given in *fig. 18a.*

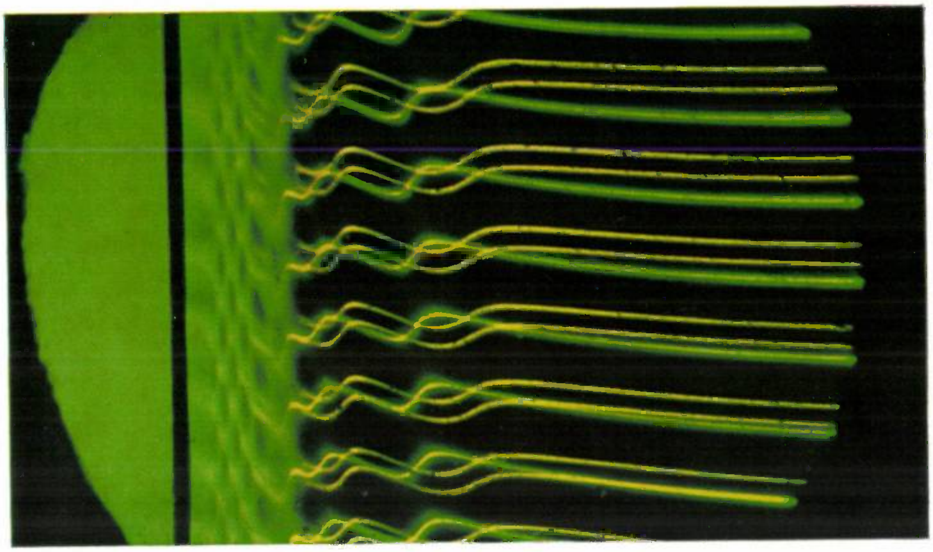
These arrangements can be greatly simplified by using the "Electrojet" system designed and marketed by the Martonair Company of Amsterdam. Fig. 18*b* shows the "Electrojet" version of the circuit in fig. 18*a*: it will be noted that in addition to the air cylinder itself, only two electromagnetic valves of the new type are necessary, these having taken over the functions of the piston valve, the two air control valves and the two bleed chokes. The overall dimensions of the "Electrojet" valve are no greater than those of the air control valve in the original circuit. The new valve must however have a much bigger bore in order to do the work of the piston valve, and the electromagnet actuating it must therefore develop a stronger pull. One of the ways in which increased pull has been obtained is by using an orthocyclic solenoid which, thanks to good heat conduction, can be heavily loaded without overheating, and is thus able to supply the required higher number of ampere-turns.

The bore of the "Electrojet" valve is large enough for most of the air cylinders on machines used at Philips. It passes enough air at 5 atm to displace a 4″ diameter cylinder piston at a speed of about 17 cm/s against a force of 300 kg, or to enable a 1¼″ piston to travel at about 170 cm/s while overcoming a force of 30 kg.

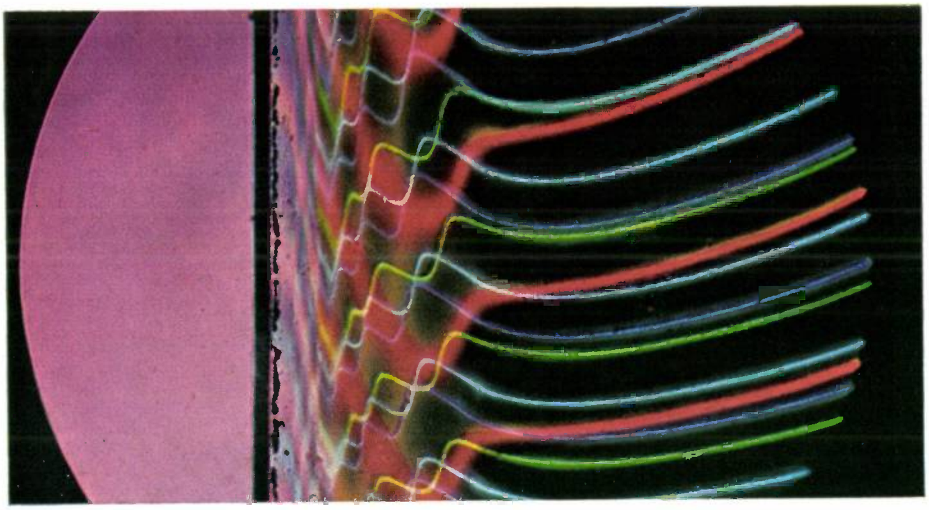
The above examples will doubtless have sufficed to demonstrate the advantages of the orthocyclic winding method, and to illustrate the variety of its applications. It need hardly be added that the method can profitably be employed in many other cases.

Summary. The term "orthocyclic" is applied to a coil-winding method whereby the greater part of each turn is made to lie orthogonally with respect to the coil axis. Provided the correct pattern is established in the first layer and certain precautions are taken, subsequent layers will automatically fall into the same regular pattern, each turn fitting into its appointed place, a groove in the layer beneath. The turns are stacked in the most compact fashion possible, in consequence of which (1) the space factor attains its theoretical maximum value, (2) the highest voltage existing between contiguous turns is fully defined, and (3) the dimensions, inductance and self-capacitance of individual coils of the same type fall within very close limits. It is possible by winding with "Thermoplac" wire to obtain strong self-supporting coils which have invariable dimensions and stand up well to vibration. The thermal conductivity of orthocyclic coils is so high (1 W/m °C, or 20 times that of coils with interleaved paper insulation) that they display no appreciable temperature gradient, and their temperature can safely rise to a value higher than the average permissible in non-orthocyclic coils. In conclusion, various applications of orthocyclic coils (to focusing coils for travelling-wave tubes, electromagnet windings in uniselectors, field coils in moving-iron meters, rotor windings in small motors, etc.) are discussed and the resulting benefits enumerated.

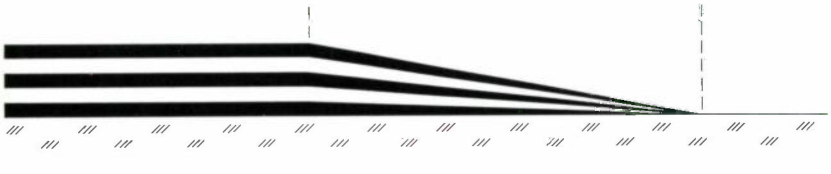
FIZEAU FRINGES OBSERVED WITH AN INTERFERENCE MIRROR



a



b

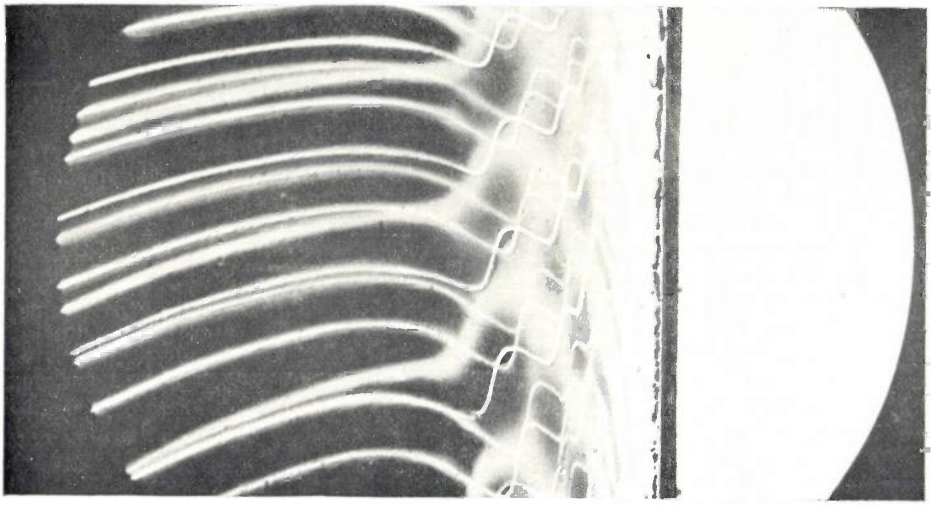


When a monochromatic beam of light is directed upon an interference mirror (consisting of thin dielectric layers of alternate high and low refractive index) which is placed a short distance from a half-silvered mirror, a pattern of relatively sharp interference fringes (Fizeau fringes) can be seen in transmission. These fringes represent lines of constant (optical) thickness of the space between the mirrors, and can therefore in general never intersect, even when several wavelengths are used.

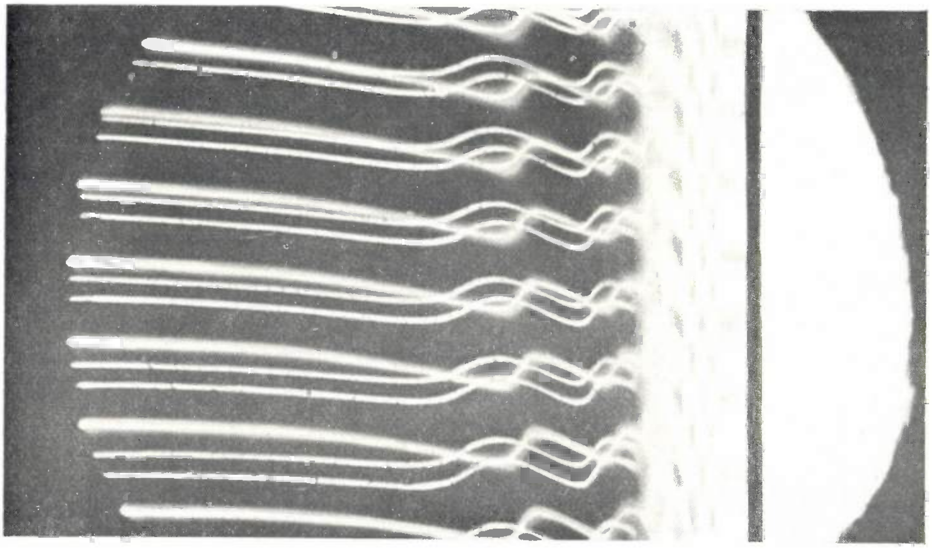
The above photographs show the fringes for several different wavelengths at the same time (*a*: cadmium lines 4678-4800-5086-6438 Å; *b*: green mercury line 5461 Å, and yellow mercury doublet 5770-5791 Å) in the neighbourhood of the edge of an interference mirror, where the layers gradually taper to zero thickness (see drawing). It can be seen that here the fringes for different wavelengths intersect. This remarkable effect can be explained in terms of the wavelengths dependence of the phase with which a multilayer mirror reflects the light (phase dispersion). For further particulars see: G. Bouwhuis, A dispersion phenomenon observable on dielectric multilayer mirrors, Philips Res. Repts **17**, 130-132, 1962 (No. 2).

ERRATUM

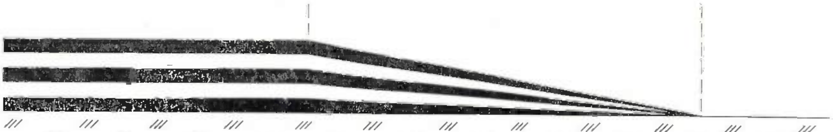
The colour photographs on page 380 have been printed in a wrong position. The correct position is indicated below.



a



b



THE ORSAY 160 MeV SYNCHROCYCLOTRON WITH BEAM-EXTRACTION SYSTEM

by G. T. de KRUIFF *) and N. F. VERSTER **).

621.384.611.2

I. GENERAL DESCRIPTION

II. THE BEAM-EXTRACTION SYSTEM

III. CALCULATION AND DESIGN OF THE BEAM-EXTRACTION SYSTEM

The article below describes the 160 MeV synchrocyclotron built by Philips for the University of Paris. Special emphasis is placed on the design of a suitable beam-extraction system. After a general description of the machine in Part I, Part II deals with the operation, construction and alignment of this beam-extraction system. In Part III details are given of the method of calculation, devised by one of the authors, which enabled the builders to design the extraction system largely on paper at a time when the machine was not yet delivering an ion beam. The same method of calculation was used for the CERN synchrocyclotron at Geneva.

Introduction

Since the beginning of 1959 a synchrocyclotron built by Philips has been in operation at the Laboratoire de Physique Nucléaire at Orsay (near Paris), an institute belonging to the Faculté des Sciences of the University of Paris. The machine delivers protons of about 160 MeV or deuterons of about 80 MeV, which makes it the largest cyclotron yet built by Philips [1]) and the largest but four in Western Europe [2]).

For many investigations in the field of nuclear physics a cyclotron is a useful instrument only when the beam of fast ions which it produces can be extracted from the accelerating chamber. Although a cyclotron having only internal targets is eminently suitable for the production of neutrons and radioactive isotopes, the nuclear reactions themselves cannot properly be studied inside the accelerating chamber, where as a rule the required measuring instruments cannot be set up, where the magnetic field causes interference, and the stray radiation too — caused for example by ions incident on one of the walls — is often an insuperable obstacle. Finally, since the increase in the orbit radius per revolution is very slight, the beam can only strike the *edge* of an internal target.

The Paris cyclotron was primarily intended as a tool of pure scientific research, and therefore one of the specifications was that it should be equipped with a device to enable an ion beam to leave the accelerating chamber — a beam-extraction system. Apart from the production of an intense beam of ions, the construction of this system was the main problem to be solved. The realization of the extraction system differed from that of others operating on the same principle, inasmuch as it had to be designed and mounted before the cyclotron was delivering an ion beam with which it would have been possible to experiment. The existing theory and method of treatment therefore had to be refined. The refined treatment was first applied to the 600 MeV synchrocyclotron of the CERN at Geneva [3]); some of the necessary preliminary calculations were carried out at Philips.

In the first part of this article we give a short, general description of the cyclotron. In the second part the operation and design of the extraction system are discussed, and the third part deals with the refined mathematical treatment on which the design of the extraction system was based.

*) Industrial Equipment Division, Philips, Eindhoven.

**) Philips Research Laboratories, Eindhoven.

[1]) The others are all roughly the same size as the Amsterdam cyclotron (see Philips tech. Rev. **12**, 241, 247 and 349, 1950/51, and **14**, 263, 1952/53).

[2]) The four larger machines are at Geneva (CERN) 600 MeV, Liverpool 400 MeV, Uppsala 200 MeV and Harwell 175 MeV.

[3]) N. F. Verster, Symposium CERN, 1956, p. 153. See also Philips tech. Rev. **22**, 141-180, 1960/61 (No. 5), in particular note [3]) on p. 146.

I. GENERAL DESCRIPTION

A plan and side view of the whole machine (excluding ancillary equipment) can be seen in *fig. 1*. In the middle is the magnet with the accelerating chamber, on the left the radio-frequency system and the high-vacuum pumps, and on the right are the magnets for focusing and deflecting the external beam. In the plan view, the beam-extraction system can be seen at the top of the accelerating chamber, and bottom right of the chamber are the mechanisms for introducing the ion source and an internal target. The ions are made to rotate clockwise to obtain an external beam, and anti-clockwise to produce neutrons with an internal target. The change of direction is simply produced by reversing the direction of the magnetic field. The neutrons produced — most of which leave the target in the direction of the incident ions — thus move towards a part of the experiment room, adjoining the cyclotron bay, which is not occupied by equipment for proton (or deuteron) experiments. The energy of the neutrons can be varied by moving the target towards or away from the centre of the magnet. Some photographs of the equipment are shown in *figs. 2* to *5*. The principal dimensions and other data are collected in *Table I*.

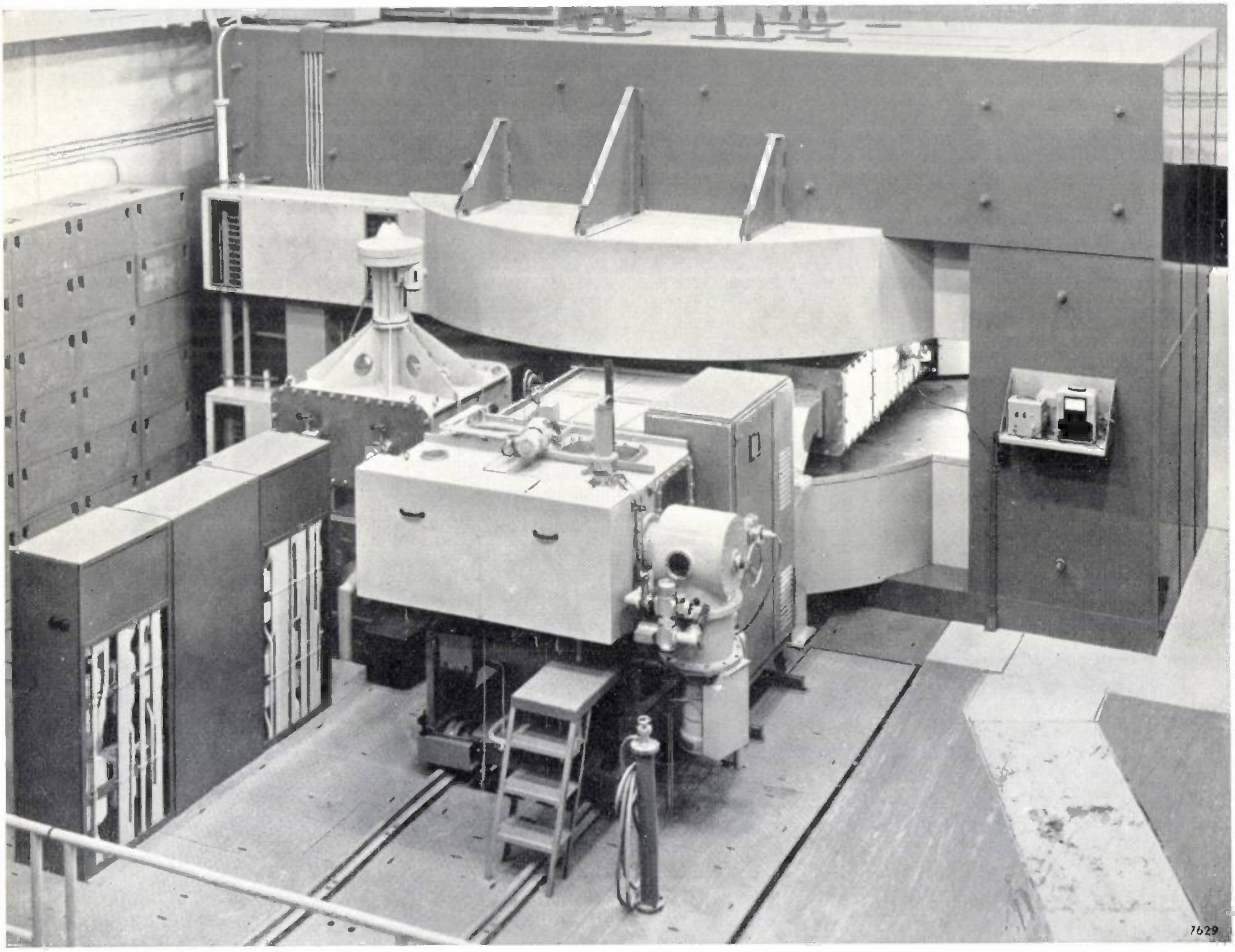


Fig. 2. The cyclotron seen from the side where the RF system is situated. The steel floor (middle foreground and left) is the roof of the large trolley (4 in fig. 1) which carries the whole system. On the rails visible in the photograph runs the small trolley (5 in fig. 1) which carries the modulator (centre). Right of the modulator can be seen the appertaining high-vacuum pump (Vac_3 in fig. 1) and behind it the oscillator. On the other side of the transmission line is the high-vacuum pump for the accelerating chamber (Vac_1). Behind this, visible through windows and accessible through doors, are the coil connections to the cooling-water pipe (eleven connections per coil). The three cabinets in the left foreground contain the panels for operating and regulating the water cooling. The pillar (with hose) in the foreground is connected to the general-purpose vacuum line (7 in fig. 1).

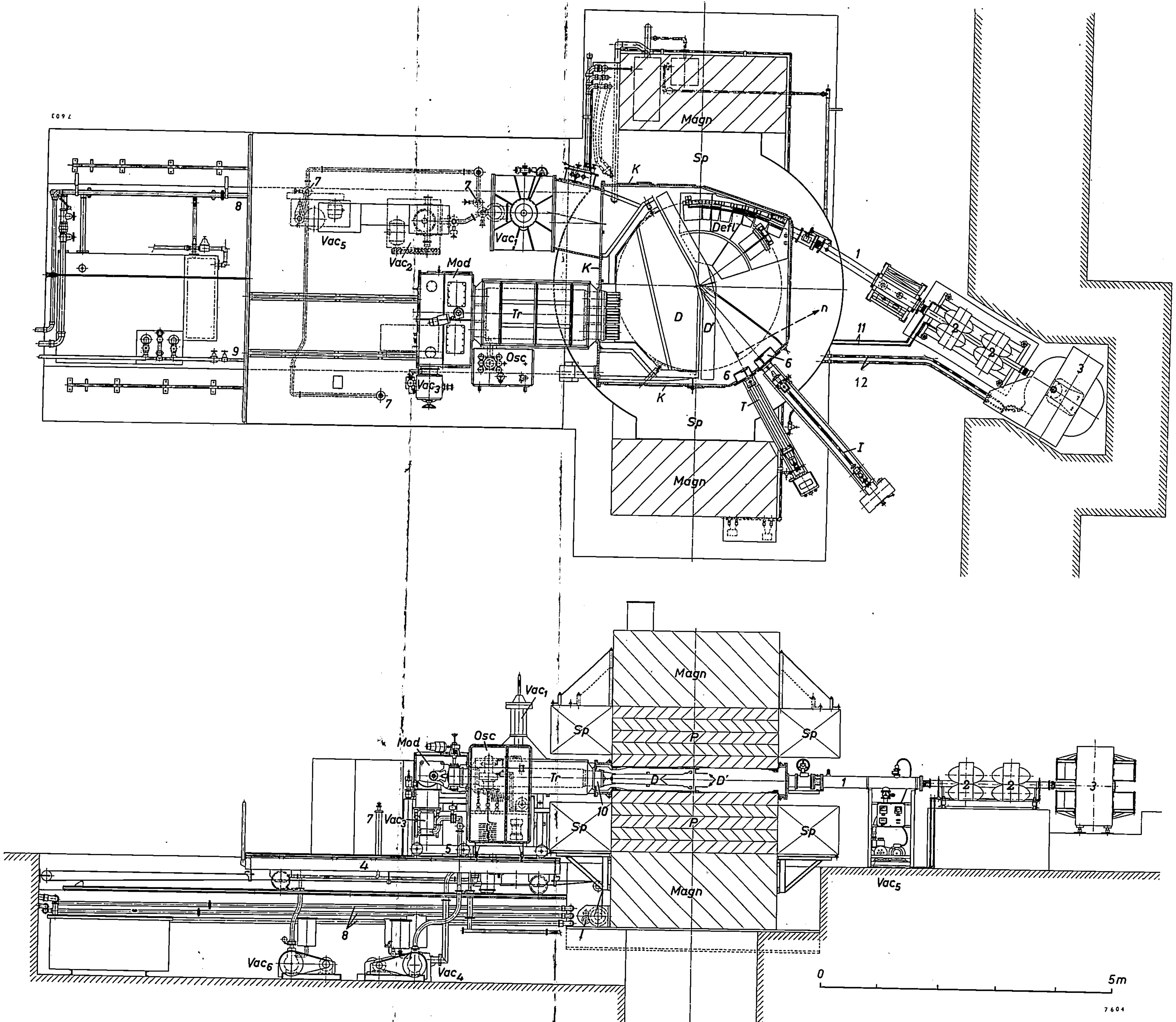


Fig. 1. Plan and side view of the synchrocyclotron with ancillary equipment, in the Laboratoire de Physique Nucléaire at Orsay. *Magn* yoke of magnet. *P* the two poles, each consisting of five steel discs 20 cm thick. *Sp* the two coils. *K* wall of accelerating chamber. *D* dee. *D'* dummy dee. *I* carriage for introducing the ion source. *T* target carriage. *n* neutron beam. Vac_1 high-vacuum pump connected to accelerating chamber. Vac_2 appertaining fore pump. *Tr* transmission line (for proton acceleration) of RF system with lead-in insulators *10*. *Mod* modulator, in separate vacuum chamber; the motor above the chamber drives the trimming capacitor. Vac_3 appertaining high-vacuum pump (the plan view indicates the vacuum valve mounted above the pump). Vac_4 appertaining fore pump. *Osc* oscillator; the oscillator valve is mounted top left in the cabinet. *Defl* beam-extraction system. *1* evacuated tube for external beam. Vac_5 appertaining vacuum pumps. *2* two magnetic quadrupole lenses for focusing external beam. *3* deflection magnet. *4* trolley on rails on which, after removing the relevant wall of the accelerating chamber, the entire RF system can be wheeled out. *5* trolley on rails, mounted on trolley *4*, with which the modulator plus transmission line are wheeled out for inserting the extra line section required when deuterons are to be accelerated. *6* vacuum locks. Vac_6 vacuum pump for general purposes (including evacuation of the locks *6*) with connections *7*. *8* cooling-water pipes for the coils *Sp* (one feed pipe for each coil and a common drainage pipe). *9* the same for the ion source. *11* the same for the quadrupole lenses. *12* the same for the deflection magnet. In the side view, *1*, *2* and *3* are seen perpendicular to their long axis.

Table I. Principal data of the 160 MeV synchrocyclotron at Orsay.

General	
Maximum beam radius	123 cm
Proton energy in internal beam at maximum radius	164 MeV
Proton energy in external beam (deflection radius 120 cm)	157 MeV
Energy gain per orbit	approx. 12 keV
Current in external beam, at the internal beam currents mentioned between brackets	(20 μA) 0.7 μA (8 μA) 0.4 μA (2 μA) 0.24 μA
Magnet	
Weight	650 tons
Pole diameter	2.80 m
Height of air gap	30-40 cm
Induction $(r = 0)$	16.30 Wb/m³
Induction $(r = 123$ cm$)$	15.65 Wb/m³
Number of turns per coil	616
Resistance of coils (in series) at 40 °C	1.05 Ω
Current	max. 690 A
RF system	
Oscillator power	$\leqslant$ 30 kW
Repetition frequency	450 c/s
Dee: capacitance	700 pF
maximum radius	137 cm
aperture at the centre	20 cm
aperture at $r = 132$ cm	10.4 cm
aperture at $r = 137$ cm	8.5 cm
Frequency for protons	25.0 to 20.2 Mc/s
Frequency for deuterons	12.5 to 11.0 Mc/s

The magnet

The poles of the magnet (diameter 280 cm) are made of circular sections 20 cm thick; to ensure a high degree of homogeneity, they were cast in vacuo, then forged and finally annealed. The yoke is assembled from 6 cm thick plates of mild steel. For our purpose the magnetic properties of this material are only slightly inferior to those of special types of steel. Except at the positions where the correction rings (shims) are mounted, the air gap is 40 cm high; at the edge, where the shims are thickest, the height is 30 cm. The induction in the air gap is about 1.6 Wb/m^2 (16 000 gauss).

The two energizing coils — one around each pole — each consist of 616 turns of aluminium tubing (24×24 mm; inside diameter 13 mm) through which cooling water flows. The cooling circuit is a closed system and contains deionized water. The heat is dissipated by mains water via a heat exchanger. The coils, which are connected in series, carry a current of 635 A (constant to within 0.01%). The total resistance of the coils is 1.05 Ω (at an average coil temperature of 40 °C).

Particular care was paid to the uniformity of the magnetic field. Nowhere does it differ from the required value by more than 0.0015 Wb/m^2 (15 gauss), i.e. about 0.1%, and at most places the discrepancy is in fact considerably less. The magnetic plane of symmetry coincides accurately with the geometric symmetry plane of the air gap. The relevant measurements will be discussed later in this article.

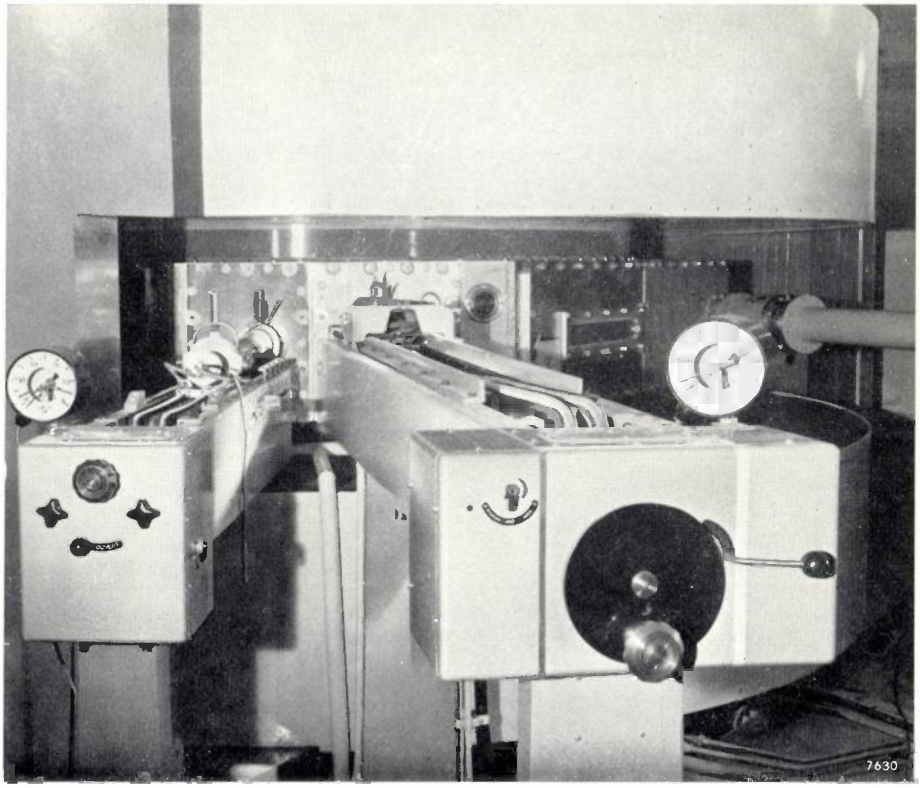


Fig. 3. Detail of the other side of the cyclotron. From left to right: the carriage for introducing a target, the ion-source carriage, window for neutron beam (in the side wall of the accelerating chamber), and the pipe for the extracted ion beam. Ion source and target are introduced in the usual way through vacuum locks. The two cabinets in the foreground contain, in addition to operating mechanisms, various locking devices. The pressure gauges indicate the pressure in the relevant vacuum lock. The vertical pipe between the two carriages is the vacuum line for the locks. The ion source is in principle identical with that in the Amsterdam cyclotron [1]).

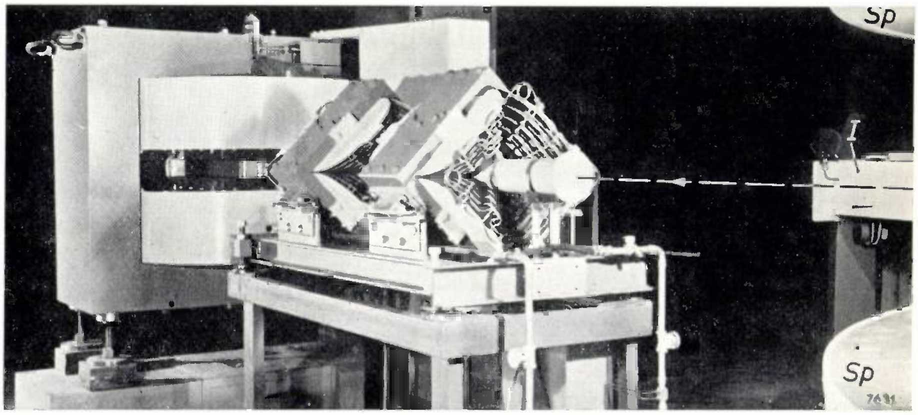


Fig. 4. The two quadrupole lenses (*2* in fig. 1) which focus the external beam, and behind them the deflection magnet (*3* in fig. 1) with which the beam can be directed to various points of the experiment room. On the extreme right are the magnet coils *Sp* of the cyclotron, and between them the ion-source carriage *1*. The communicating pipe (*1* in fig. 1) between the accelerating chamber and the lenses is missing. The centre line of the beam is denoted by dashes.

The radio-frequency system

As far as the resonator is concerned, the radio-frequency system is basically the same as in the other Philips cyclotrons. Here too the dee, the transmission line and the modulator form a coaxial resonator about half a wave in length, whose resonant frequency is determined by the capacitance of the

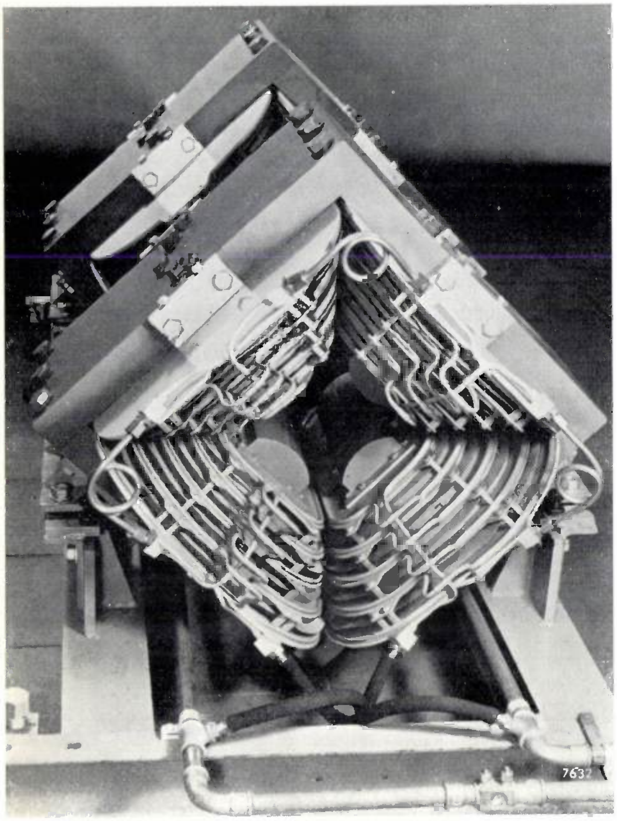


Fig. 5. The quadrupole lenses, photographed before the vacuum tube for the beam was introduced.

modulator. The dee and the modulator are at the ends of the line — at the voltage antinodes — and the coupling to the RF oscillator is somewhere in the middle. With this construction it was again possible to house the modulator in its own vacuum chamber, and as before the coupling with the oscillator could be effected at that part of the transmission line which is outside the vacuum. The modulator is of the conventional rotating type with a capacitively earthed rotor.

The diagram in *fig. 6a* represents the RF system for proton acceleration, and that in 6*b* the deuteron acceleration system. The latter can be produced from the former by wheeling the modulator back a certain distance and inserting an extra length of line between the dee and the transmission line. The modulator capacitance then has to be increased slightly with a trimming capacitor. It is not necessary to move the oscillator, steps having been taken to ensure that the distance AA' between the points where the oscillator has to be coupled to the resonator to obtain the requisite dee voltage (25 kV) in the two cases is equal to the length of the extra line section. It is thus a particularly simple matter to convert the machine from a proton cyclotron into a deuteron cyclotron.

The circuitry of the oscillator and the method of coupling the oscillator to the resonator are the same as in the CERN 600 MeV cyclotron at Geneva ("flywheel" circuit) [4]). This circuit makes it unnecessary to change the coupling in any way when switching from protons to deuterons. With an inductive coupling, as used for example in the Amsterdam cyclotron, separate loops are needed for the two frequency ranges, which makes conversion much more cumbersome.

Since, owing to the complicated form of the dee and of the feed-through insulators in the transmission line, neither the dee capacitance nor the characteristic impedance of the feed-through insulators can be exactly calculated, it is not possible to design an RF system as required in a cyclotron completely on paper. The approximate calculations were checked on a

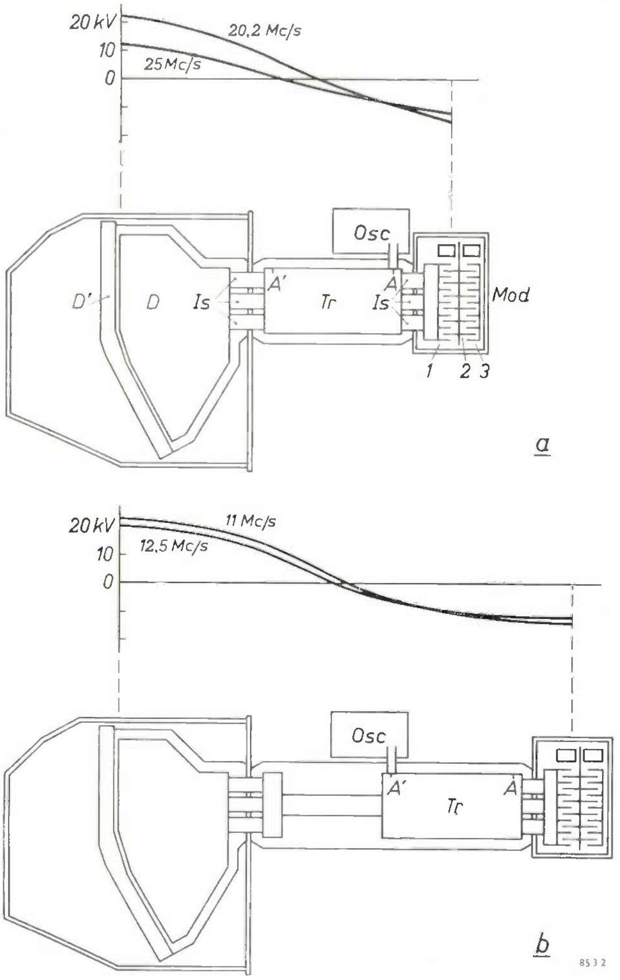


Fig. 6. Diagram of the RF system, indicating the voltage waveform across the resonator at the extreme frequency values; *a*) for proton acceleration, *b*) for deuterons. *D* dee. *D'* dummy dee. *Tr* transmission line. ***Mod*** modulator (*1* fixed teeth, *2* rotor, *3* earthing plates). *Osc* oscillator. *Is* feed-through insulators. The oscillator must be coupled to the transmission line at point *A* for accelerating protons, and at point *A'* for accelerating deuterons. In the conversion the oscillator can therefore stay in position. Points *A* and *A'* are situated so that at the lowest frequency a voltage multiplication of about 2× is obtained at the mouth of the dee.

[4]) See Philips tech. Rev. **22**, 159, 1960/61.

model (scale 1:5) and the first design improved where necessary. In order to obtain a transmission line long enough to make the above-described positioning and oscillator coupling possible, the characteristic impedance of the line had to be chosen as low as 12 Ω. The characteristic impedance of the feed-through section was kept low by using no less than three feed-through insulators in parallel at both ends.

The operation of the oscillator, in particular the coupling to the resonator, was also studied on a full-scale model. Since the maximum anode voltage of the oscillator valve (a Philips water-cooled triode type TAW 12/35) is about 12 kV, and the aim is a maximum dee voltage of about 25 kV, the oscillator is coupled to the resonator at a point where the voltage multiplication at the minimum frequency is roughly 2 (see fig. 6*a* and *b*).

The joule heat generated in the RF system is removed from the dee and the transmission line by water cooling and in the rotor of the modulator almost exclusively by radiation. To prevent the rotor getting too hot — with a view to the necessary limitation of thermal expansion — the rotor is covered with a thin layer of copper oxide. The emission coefficient is thereby increased from 0.08 to 0.65 and the operating temperature consequently lowered from about 200 °C to about 60 °C [5]).

II. THE BEAM-EXTRACTION SYSTEM

The fact that a beam of charged particles cannot usually be extracted from the cyclotron — i.e. from the magnetic field — without taking special measures, is due to the occurrence of certain instability effects. By way of explanation, and also as an introduction to the method of designing the extraction system (see Part III) we shall begin this part of the article by briefly discussing the properties of the orbit which a charged particle describes in a cyclotron.

A charged particle, of mass m and charge e, which moves in a uniform magnetic field B with a velocity v normal to the lines of force, describes a circular orbit whose radius r is proportional to the momentum mv of the particle:

$$\frac{mv}{e} = Br. \quad \ldots\ldots \quad (1)$$

The angular velocity ω in the orbit is therefore:

$$\omega = v/r = eB/m. \quad \ldots\ldots \quad (2)$$

If the velocity also has a component in the direction of B, the orbit is a helix.

In a cyclotron the magnetic field is not perfectly uniform but somewhat barrel-shaped (*fig. 7*). This means firstly that in the plane of symmetry the value of B decreases slightly with increasing r, and secondly that outside the plane of symmetry (assumed to be horizontal) B has a *horizontal* component B_r as well as the vertical component B_z. A particle outside the plane of symmetry is subjected to a vertical restoring force (Lorentz force) due to this horizontal component of the field. Consequently a charged particle in the cyclotron field describes vertical oscillations about the circular orbit. This vertical motion is given by:

$$m\frac{d^2z}{dt^2} = e\,v_h\,B_r(z)\,, \quad \ldots\ldots \quad (3)$$

where z is the distance to the symmetry plane and v_h is the horizontal component of the velocity v of the ion. In the vicinity of the symmetry plane the radial component of the magnetic induction is approximately given by

$$B_r(r,z) = z\left(\frac{\partial B_r}{\partial z}\right)_{z=0}. \quad \ldots\ldots \quad (4)$$

Further, in the air gap between the poles of the cyclotron magnet, we have curl $B = 0$. Hence:

$$\frac{\partial B_z}{\partial r} = \frac{\partial B_r}{\partial z}\,,$$

and therefore

$$B_r(r,z) \approx z\frac{dB(r)}{dr}\,.$$

(In the latter expression the abbreviated notation $B(r)$ is used for $B_z(r, 0)$. Outside the symmetry plane we continue, of course, to distinguish B_z from B.) Substituting this in (3) we obtain:

$$m\frac{d^2z}{dt^2} = e\,v_h\,z\frac{dB(r)}{dr}\,.$$

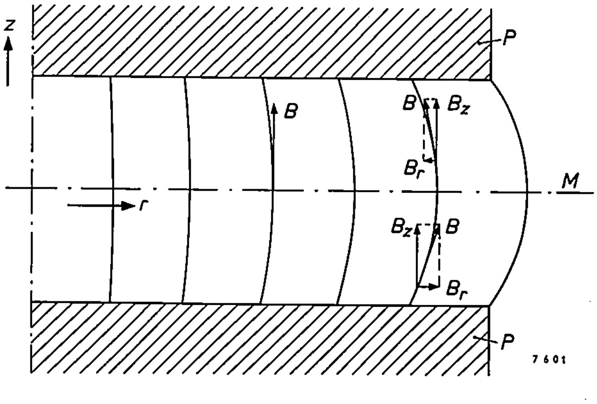


Fig. 7. Schematic representation of the cyclotron field in half the cross-section of the air gap. P magnetic poles. Outside the plane of symmetry M the horizontal component B_r of the induction is not equal to zero. In the plane M, the gradient dB/dr is negative (cf. fig. 8).

[5]) This subject has been extensively investigated by G. Henneberke of Philips Industrial Equipment Division.

Provisionally we approximate to $\mathrm{d}B(r)/\mathrm{d}r$ by the value which this differential quotient has in the plane of symmetry on the circle whose radius r_0 corresponds to the momentum of the particle (cf. equation (1)). We call this circle the equilibrium orbit. Denoting the quantity $-(r/B)(\mathrm{d}B/\mathrm{d}r)$ (which is known as the field index) by n, introducing the azimuthal coordinate $\Theta = \omega t$ and disregarding the difference between v and v_h, we can reduce this equation with the aid of (1) to the very simple form:

$$\frac{\mathrm{d}^2 z}{\mathrm{d}\Theta^2} + nz = 0 . \quad \ldots\ldots \quad (5)$$

In the radial direction too, a charged particle is found to oscillate around the equilibrium orbit. The radial deviation $\varrho = r - r_0$ from the equilibrium orbit is given by:

$$\frac{\mathrm{d}^2 \varrho}{\mathrm{d}\Theta^2} + (1 - n)\varrho = 0 . \quad \ldots\ldots \quad (6)$$

The method of deriving this equation resembles that used for (5).

Equation (5) and (6) have harmonic solutions, that is to say the radial and vertical oscillations of the ion are stable if $0 < n < 1$. The oscillation frequencies are respectively $\omega\sqrt{n}$ and $\omega\sqrt{1-n}$. The field index in the magnetic field of a synchrocyclotron is always positive. It increases with r but remains fairly small up to high values of r. In the fringing field of a magnet there is a marked increase in n and the critical value of unity is exceeded (*fig. 8*). This means that the ion stops orbiting and leaves the cyclotron field. The reason for this is that on the radius at which $n = 1$ the product Br has reached its maximum value, and starts to decrease with further increase in r. An ion which has been accelerated to a momentum greater than that corresponding to the maximum value of Br cannot therefore continue in a circular orbit.

This could be the basis of a very simple method of beam extraction, if in fact it were possible to accelerate an ion to the momentum mentioned. Unfortunately that cannot be done. Although, owing to our approximations, it does not appear from equations (5) and (6), the radial and vertical oscillations are not entirely independent of one another. This implies that at certain frequency ratios the one oscillation, e.g. the radial, can lose energy to the other, giving the vertical oscillation such a large amplitude that the ions are lost by collision with the dee. The smallest radius at which this effect can become serious is that corresponding to $n = 0.2$; the frequency of the radial oscillation ($\omega\sqrt{1-n}$) is then exactly twice that of the vertical ($\omega\sqrt{n}$). The next dangerous radius corresponds to $n = 0.25$. It is evident, then, that the orbit radius should not exceed the value r_c at which $n = 0.2$, for the energy gain per revolution is always fairly low in a synchrocyclotron. The orbit radius of each particle therefore increases only very slightly in each revolution (in the Orsay cyclotron about 0.1 mm at $r = 120$ cm) and thus remains very near to the dangerous value for a fairly large number of revolutions.

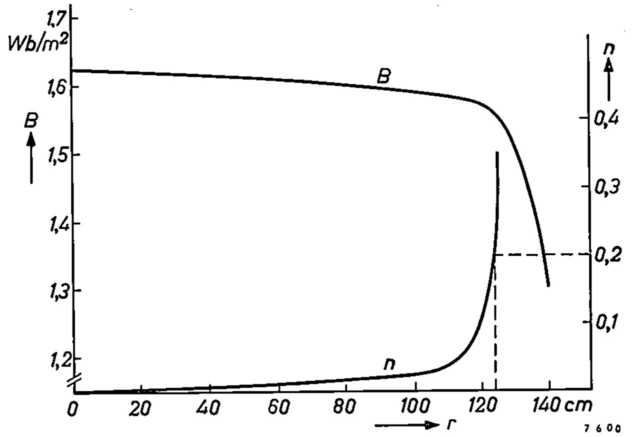


Fig. 8. Variation of induction B and of the field index n in the plane of symmetry as a funtion of distance r to the centre line of the magnetic poles. The field index reaches the critical value 0.2 at $r = 123$ cm.

Operation of the extraction system

A cyclotron beam can in principle be extracted either by means of a transverse electrostatic field or by local attenuation of the magnetic field. The first method can be used only when the energy of the ions is not too high; an impracticably high electric field strength would be required to deflect ions of 150 MeV.

In the other method of beam extraction, use is made of what is known as a magnetic channel. This is an oblong enclosure in which iron walls are used to attenuate the cyclotron field. The mouth of the channel is located between the poles of the magnet. Here the channel is perpendicular to the line connecting the mouth to the centre of the air gap. The outlet of the channel is situated outside the air gap. When an ion enters this channel, the decrease of the field causes the orbit radius to increase to a high value and the ion leaves the cyclotron field. At the outlet of the channel the ion already travels in a more or less straight line.

The difficulty of magnetic extraction is the problem of how to get the ions to enter the channel. In a synchrocyclotron it is generally speaking not possible for the channel simply to "peel off" the outermost ions from an orbiting group. The iron wall of the magnetic channel has to be many times thicker than the increase in the radius per orbit, so that the ions would land on the wall before reaching the mouth of the channel. (In the electrostatic system, on the other hand, the electrode need be no thicker than e.g. 50 μm.)

In a *regenerative extraction system* as used at Orsay, above a certain radius — which is fairly large though smaller than r_c — a forced radial oscillation is generated whose frequency gradually becomes equal to the rotational frequency of the ions. The amplitude of this oscillation increases exponentially with time, and the difference in amplitude between the ultimate and penultimate orbit is no less than, for example, 5 cm. The ions can thus enter the channel freely, provided the mouth is not too close to one of the nodes of the oscillation.

The orbit described by an ion under the influence of the extraction system can be regarded as a circle with a moving centre. At the beginning this centre precesses around that of the equilibrium orbit. Later, when the frequency of the forced oscillation is exactly equal to the frequency of rotation, it moves away quickly to one side. To a first approximation we can disregard the increase in the radius of the orbit caused by the fact that the ions continue to be accelerated during the extraction process. In the cyclotron at Orsay the circle where $n = 0.2$ has a radius of 123 cm, and that where $n = 1$ has a radius of 131 cm. The artificial generation of the required radial oscillations begins when the ion orbits have a radius of 120 cm, the channel mouth being at a distance of 133 cm from the centre.

In a regenerative extraction system the radial oscillations are generated by means of a *local* field disturbance having a marked positive gradient, i.e. in which the field increases with r; this disturbance is produced by the *regenerator.* Both radially and in azimuth the dimensions of the regenerator are small. When the orbit of an ion has a sufficiently large radius, the ion passes the regenerator each revolution. Since not only the induction in the regenerator but also the radial field gradient differs from that which the ion encounters in its orbit, the regenerator influences the state of oscillation of the ion in two ways: it alters the orbit equation (6) not only by introducing a periodic function into the right-hand side but also by causing a periodic variation in n. It can be shown that this is the reason why the frequency of the forced oscillation ultimately becomes *exactly* equal to the orbital frequency of the ions [6]).

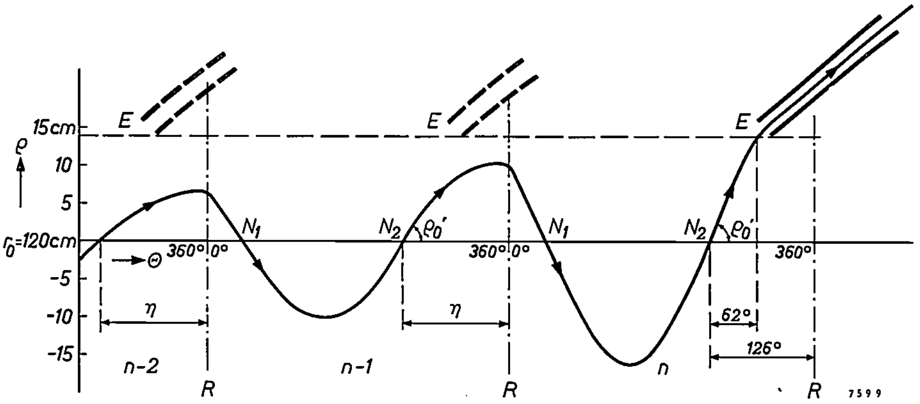


Fig. 9. Principle of beam extraction with a regenerative extraction system. The radial deflection ϱ of an ion, measured from the equilibrium orbit, is plotted as a function of azimuth Θ for the last revolutions of the ion (orbit number n, $n-1$, $n-2$). The regenerator is located at R, where it produces a field with a steep positive radial gradient; its azimuth is taken as zero. E mouth of magnetic channel. N_1 and N_2 penultimate and ultimate passage of the ion through zero in each revolution before reaching the regenerator. In the extraction system as designed, the azimuth difference η between N_2 and R has the same value in each revolution. The amplitude of the oscillation increases to such an extent during the last revolutions that the ion in orbit $n-1$ still misses the channel by a considerable distance. The azimuth differences indicated refer to the actual extraction system.

The amplitude of the vertical oscillations is kept within limits with a second local field disturbance, in which the field has a strong *negative* gradient. In older cyclotrons using magnetic beam extraction, this field disturbance was produced with the aid of a steel construction known as a peeler [7]). In the cyclotron under consideration the region of negative gradient is not obtained artificially but by using the fringing field. Although this is not in fact a local field disturbance, it amounts to that as far as radially oscillating ions are concerned: these enter the fringing field only in a limited part of their orbit.

Since the behaviour of an ion in a magnetic field is governed solely by its momentum, i.e. by the product Br, ions of all kinds behave in the same way in the cyclotron field on a circle having a specific radius; protons and deuterons can therefore be extracted by one and the same system.

Fig. 9 gives a qualitative illustration of the

[6]) This method was first successfully applied in the Liverpool cyclotron. We acknowledge here our indebtedness to the late Professor Skinner and to Dr. V. Moore for the hospitality they showed us on our visits to the Nuclear Physics Research Laboratory of Liverpool University, to Dr. K. J. Le Couteur, who allowed us to see one of his articles before publication, and to Dr. A. V. Crewe for many useful discussions.

[7]) Although the names regenerator and peeler are not a fortunate choice, we shall conform to general usage in this matter.

positioning of regenerator and channel and the path described by an ion on its last orbit. Depending on the current, the intensity of the external beam is 5 to 12% of that of the internal beam (Table I). The remainder of the internal beam falls on a water-cooled diaphragm mounted at the mouth of the channel; this diaphragm also serves as a target for the production of radioactive isotopes with a long half life.

This form of the regenerative method of extraction, which was proposed and treated theoretically by Le Couteur and Lipton [8]) in 1955, has two advantages over the system using a peeler [9]). In the first place a peeler is more difficult to make than a regenerator. Moreover, when a peeler is used the deflection has to be effected in a region where the field gradient is practically constant, i.e. at an appreciably smaller distance from the centre than half the pole diameter of the magnet. The ions can therefore not be accelerated to the maximum possible energy (i.e. the energy at which the orbit coincides with the circle where $n = 0.2$). It might be regarded as a disadvantage of the method not using the peeler that its mathematical treatment is more difficult and leads to equations that can only be solved numerically (see Part III). It is equally true, however, that excellent insight is obtained in this way into the behaviour of the ions.

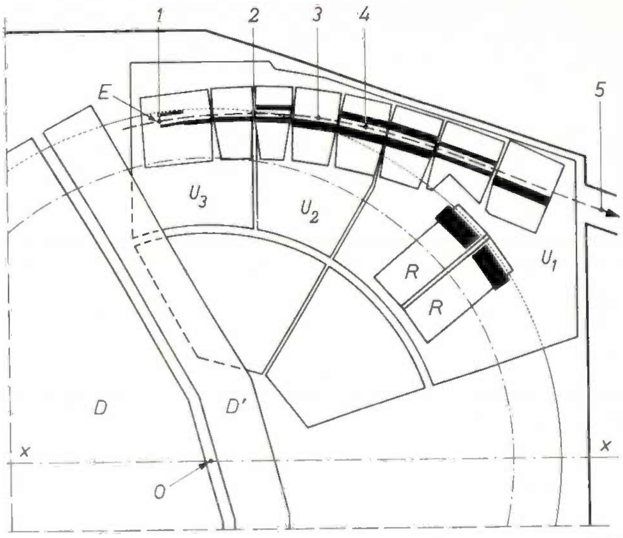


Fig. 10. Positioning in the cyclotron of the regenerator and of the nine segments that together form the magnetic channel (see fig. 1). O centre point of magnet poles. The outer circle represents the edge of the magnet poles. x-x principal axis of the cyclotron. D dee. D' dummy dee. U_1, U_2 and U_3 aluminium plates on which the regenerator and the channel segments are mounted. R regenerator, consisting of two parts. E inlet of first channel segment. The dot-dash circle is the equilibrium orbit on which the deflection begins ($r = 120$ cm). The external beam makes an angle of less than 30° with the x-x axis. The significance of the numbers *1* to *5* is explained in fig. 12.

Construction of regenerator and magnetic channel

Owing to the manner in which, for various reasons, the cyclotron had to be sited in the space available, a restriction was imposed on the direction of the external beam: the angle between that direction and the principal axis of the cyclotron could not be greater than 30° (cf. fig. 1). This requirement could only be satisfied by having a relatively long magnetic channel which would quickly remove the beam from the influence of the magnetic field. The form and situation of this channel have already been broadly indicated in fig. 1. A more detailed plan view being given in *fig. 10*.

The channel consists of nine segments, the length of which increases as the curvature of the orbit decreases: the segment at the mouth is the shortest and that at the outlet the longest. Most of the segments consist of two thick vertical iron plates, mounted in the plane of symmetry of the cyclotron at either side of the path which the beam has to follow. Some segments consist of only one such plate.

Since of course the field is reduced on *both* sides of the plates, they also disturb the cyclotron field outside the channel. This disturbance imposes certain limitations on the field reduction that can be achieved in the first segments, and measures are needed to correct for it as far as possible. This is done with the aid of iron correction plates which are mounted in pairs in the accelerating chamber — this time above and below the plane of symmetry and far enough away from it to avoid obstructing the ions. The field

Fig. 11. Method of assembling the steel walls W of one segment of the magnetic channel, here the first segment, with the appertaining steel correction plates C. P the two magnet poles. M plane of symmetry. S shims. U aluminium plates, fixed to P and S, to which the U-shaped aluminium frame F is secured. This frame carries both the channel walls and the correction plates.

[8]) K. J. Le Couteur and S. Lipton, Phil. Mag. **46**, 1265, 1956.

[9]) J. L. Tuck and L. C. Teng, Phys. Rev. **81**, 305, 1951. A mathematical treatment will be found in: K. J. Le Couteur, Proc. Roy. Soc. A **232**, 236, 1955.

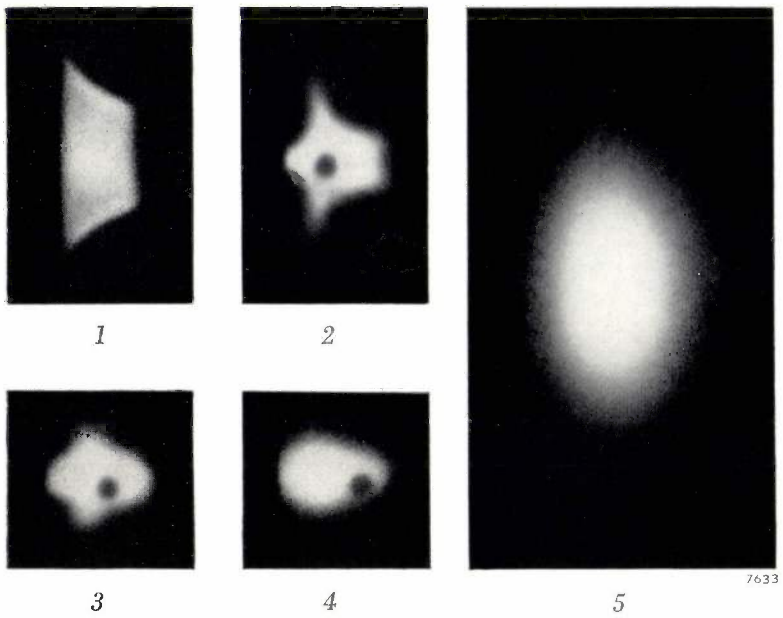


Fig. 12. Cross-section (true size) of the proton beam in the magnetic channel at the points indicated by *1* to *5* in fig. 10. The figures are autoradiographs, obtained by keeping a copper plate at the places mentioned for a few seconds until it became radioactive and then laying it on photographic paper. The spots visible in *2*, *3* and *4* indicate the middle of the channel.

reduction brought about by the segment plates is relatively greater nearby and relatively smaller at a considerable distance away, as the height of the segment plate decreases. The segment plates should therefore not be made too high.

The plates of each segment and the pertaining correction plates are mounted on a U-shaped aluminium frame as illustrated in *fig. 11*. The frames themselves are mounted on three horizontal aluminium plates — the first two carry three segments and the last one three segments plus the regenerator — which are secured to the bottom pole face of the magnet: on the outside they rest directly against the shims, which give the fringing field the appropriate pattern, and on the inside they rest on studs. Similar plates (ceiling plates) are also mounted above the frames on the upper pole face of the magnet; the frames are fixed to these plates so that they can better withstand the large radial forces acting upon them in the strong cyclotron field. The whole assembly is designed in such a way that during check measurements (see following section) changes can quickly be made to the channel mouth and the correction plates.

Just as the channel walls disturb the cyclotron field, so too do the correction plates, which are intended to eliminate this effect, disturb the field in the channel: this causes a further increase in the (radial) field gradient, which is already quite large here. The consequence in the channel is a focusing effect in the vertical plane and a defocusing effect in the horizontal plane. To eliminate the latter effect, four segments have only a single wall, the outer wall being omitted. The field gradient in these segments is considerably less negative or even positive. The result is a channel that can be regarded as a system of magnetic lenses which alternately focus and defocus in the horizontal plane, and do just the reverse in the vertical plane. The field reduction is of course appreciably lower in the single-walled segments than in the others. The principal data relating to the nine segments are summarized in *Table II*. *Fig. 12* shows the cross-section of the beam at various points in the channel.

It will be clear, having regard to the complexity of the whole assembly, that the final form of the channel (including the correction plates) was not achieved entirely without trial and error. We shall return to this point presently.

A cross-section of the regenerator can be seen in *fig. 13*. The regenerator field is produced with the aid of the steel blocks *R*, the plates *C* on the right again

Table II. Principal data of magnetic channel. The segments are numbered in the sequence in which they are traversed by the ions.

Segment	Length	Thickness inside wall	Thickness outside wall	Width	Height	Field drop	Gradient	Type of steel
						10^{-4}	10^{-6}	
	mm	mm	mm	mm	mm	Wb/m^2	Wb/m^3	
I	100	7.5	7.5	12	35	—3100	—1490	cobalt steel
II	100	10.0	—	—	37.5	— 800	— 186	,,
III	150	12.5	—	—	40	—1375	— 62	,,
IV	150	15.0	15.0	15	40	—5450	—1140	,,
V	200	25.0	—	—	40	—3760	+1211	,,
VI	200	30.0	30.0	20	60	—6040	— 727	mild steel
VII	200	30.0	30.0	20	60	—	— 640	,,
VIII	250	25.0	15.0	30	60	—	+ 60	,,
IX	250	35.0	—	43	80	—	+ 380	,,

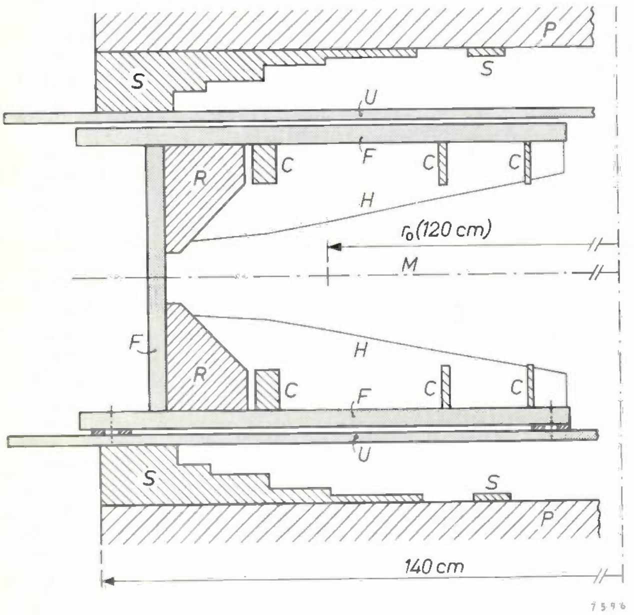


Fig. 13. Schematic cross-section of one of the regenerator halves. *P* and *S* are again respectively the magnet poles and shims. The regenerator field is produced by means of the steel blocks *R*. The blocks and plates *C* again have a corrective function. *R* and *C* are mounted in the same way as the channel segments (see fig. 11), except that the construction is reinforced by two ribs *H*. Dimensions of the four blocks *R* are: 200 mm long, 75 mm wide, 97.5 mm high outside and 37.5 mm high inside.

being correction plates. The mechanical design resembles that of the channel segments. Two reinforcement ribs *H* are fitted to the centre of the frame, one above and one below, to enable the system to withstand the powerful radial forces which act on the steel blocks in the very inhomogeneous boundary field. A set of regenerator blocks and correction plates as shown in fig. 13 is mounted on each side. The two sets are at an angle of 7° to one another in the horizontal plane, so that the shape of the regenerator is adapted somewhat to that of the ion orbits. The whole assembly can be displaced radially. A general view of the channel and regenerator is to be seen in *fig. 14*.

The field disturbances encountered by an ion passing the regenerator in an orbit of radius 127 cm are represented in *fig. 15*.

Field correction and alignment of the extraction system

For applying the field corrections presently to be discussed, and in aligning the channel and regenerator, use was made of a light, flexible copper wire which acted as an "orbit simulator" [10]). With refer-

[10]) See G. R. Lambertson, UCRL 33-66.

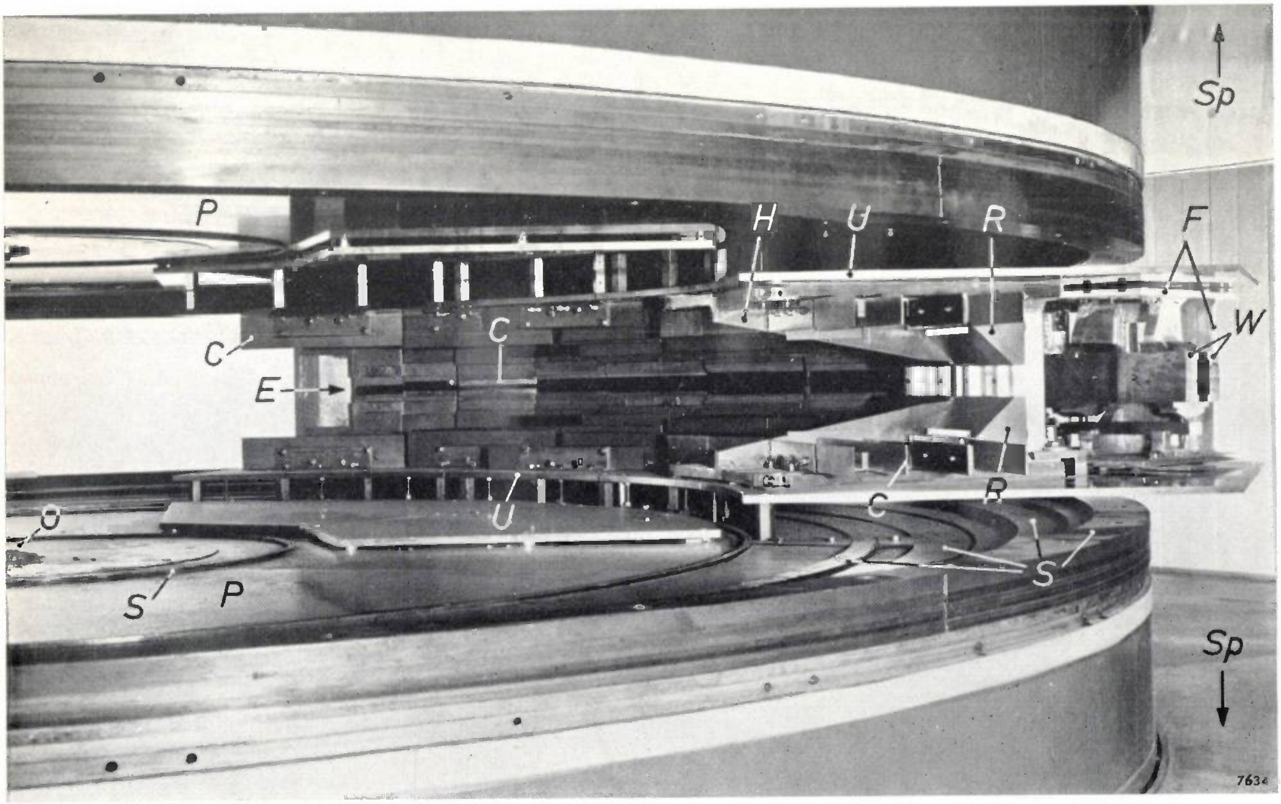


Fig. 14. The extraction system, photographed when the cyclotron was under construction. The letters have the same meaning as in figs 10, 11 and 13. On the far right can be seen the walls *W* of the last channel segment (in the final alignment the outer wall of this segment was omitted, see Table II). The walls of the segments *I* to *VII* can be seen on the left of the regenerator *R* as dark patches between the correction plates *C*.

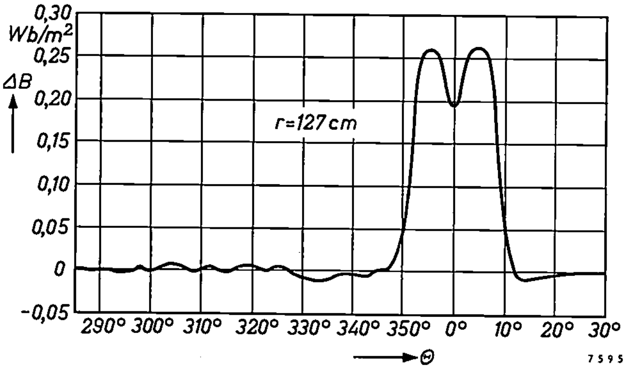


Fig. 15. The difference ΔB between the induction B in (and around) the regenerator and the average value of B outside the regenerator on a circle of 127 cm radius. The abscissa is the azimuth Θ of the points of the circle.

ence to *fig. 16* we shall first briefly examine the principle of this method.

In this figure the curve represents the conductor, assumed to be perfectly flexible. It is rigidly secured at points X and Y. The lines of force of the field B are perpendicular to the plane of the drawing. In this plane each element Δs of the wire is subjected to a Lorentz force of magnitude $IB\Delta s$ perpendicular to the direction of I. In the equilibrium state the shape of the wire is such that this force on each element is equal and opposite to the radial component $K\Delta a$ of the mechanical strain K in the wire:

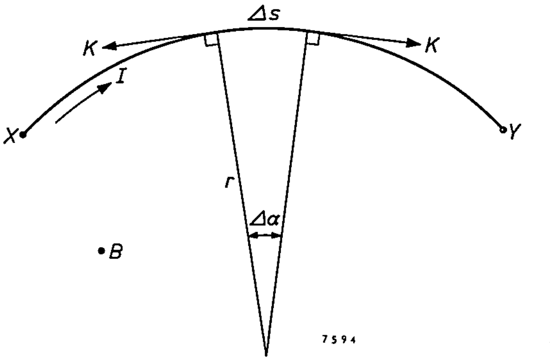


Fig. 16. Illustrating the use as an "orbit simulator" of a copper wire fixed between two points X and Y.

$$K\Delta a = -IB\Delta s\,.$$

Now $\Delta s = r\Delta a$, so that this formula can be reduced to $K/I = -Br$. It follows, then, that the shape of the wire is such that the product Br has the same value at every point, for K and I have the same value at every point.

The same also applies, however, to the orbit of an ion (eq. 1), for the product Br is proportional to the momentum. The wire thus gives us at once the form on any particular ion orbit. If *three* points of an orbit are known, I and K need not be measured at all; it is then only necessary to make the wire long enough to pass the third point. The two extreme points are here again the fixed points, and the length is adjusted until the wire also passes through the third. If a sufficiently large current is chosen, the weight of the wire can be disregarded.

We used a hard-drawn, enamelled copper wire of 0.25 mm diameter and a current of 3 A. If the orbit of a 157 MeV proton in the cyclotron field is simulated in this way (at a B value of 1.57 Wb/m^2 the radius is 120 cm), the tension in the wire is about 6 newton.

It should be noted that the Lorentz force acting on an element of the wire is opposed to that which acts on a moving ion. Outside the plane of symmetry the vertical component of this force repels the wire away from this plane instead of attracting the wire towards it, as in the case of an ion. A freely suspended wire would thus be vertically unstable. We shall return to this point when discussing the various applications.

Field corrections

The field corrections made using the orbit simulator method were 1) a symmetry-plane correction and 2) a centre-point correction.

In designing the extraction system it was assumed that the magnetic plane of symmetry, in which the equilibrium orbits of the ions lie, coincides with the geometric plane of symmetry. In fact, however, no matter how carefully the magnet and coils are made, this may not be the case. The manner in which this was investigated in the present cyclotron is sketched in *fig. 17*. The ends of the wire are fixed to a vertical pin attached to a float. The wire can thus move freely in the horizontal plane. If necessary, the level of the water in which the assembly floats can be altered slightly. To prevent

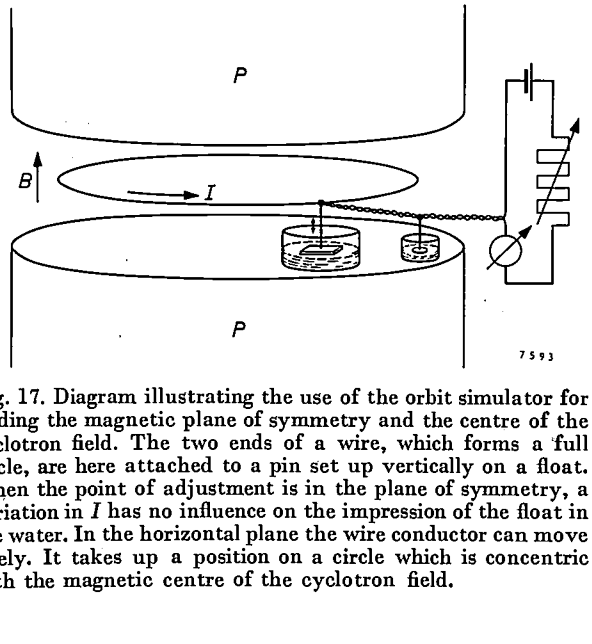


Fig. 17. Diagram illustrating the use of the orbit simulator for finding the magnetic plane of symmetry and the centre of the cyclotron field. The two ends of a wire, which forms a full circle, are here attached to a pin set up vertically on a float. When the point of adjustment is in the plane of symmetry, a variation in I has no influence on the impression of the float in the water. In the horizontal plane the wire conductor can move freely. It takes up a position on a circle which is concentric with the magnetic centre of the cyclotron field.

the wire being influenced by radial forces that do not in reality act on an ion, the supply line is twisted and attached to a second float.

By mounting the wire on a float we have eliminated the vertical instability which, as just noted, would attach to a freely suspended wire, while at the same time retaining some vertical mobility. The position of the symmetry plane at the point of suspension of the wire is now found by determining that level of water in the vessel at which a change in the current I through the wire does not influence the vertical position of the float. If the wire is outside the symmetry plane, an increase of I will shift the equilibrium position to a point still further removed from this symmetry plane. The entire plane of symmetry can be charted by performing measurements of this kind at numerous points.

This procedure revealed that the magnetic plane of symmetry was initially somewhat curved. It coincided with the geometric symmetry plane near the centre, but at a radius of 120 cm it was 12 mm above it. Especially in view of the limited height of the channel mouth (about 40 mm) this was unacceptable. The trouble was corrected by compensating the asymmetry in the iron core of the magnet with an opposite asymmetry in the coils. For that purpose the lower coil was shunted with a resistance of 15.5 Ω, causing a current of 635 A to flow through the upper coil and 615 A through the lower coil. In this way the discrepancy was reduced to less than 2 mm.

The centre-point check, which we shall now discuss, shows whether the centre of an ion orbit coincides with the geometric centre of the magnet. Plainly, this can also be investigated using the set-up in fig. 17. Discrepancies occur when a cyclotron field is not azimuthally homogeneous. In a cyclotron with a regenerative extraction system, inhomogeneities are due to regenerator and the channel: before correction, their effects are perceptible down to quite a small radius. The presence of such inhomogeneities is undesirable. They tend to give rise to radial oscillations — as in the regenerator — and therefore have an adverse effect on the intensity and energy distribution of the external beam. Ions that describe large radial oscillations may well come under the influence of the regenerator too soon and be extracted before they have reached the energy that corresponds to an orbit radius of 120 cm.

The check was carried out by successively determining the radial position of four points in pairs diametrically opposite to one another. The discrepancies were corrected by locally increasing the induction with the aid of thin plates (1 mm thick) secured symmetrically with respect to the plane of symmetry to the upper and lower poles. The magnetic centres of orbits having a radius between 30 cm and 115 cm are now less than 1 cm from the geometrical centre of the magnet.

Alignment of magnetic channel and regenerator

To find the definitive position of the magnetic channel (*fig. 18*) one end of the wire conductor was fixed to the circle with $r = 120$ cm at the calculated penultimate node N_1 (cf. fig. 9) of the orbit terminating at the channel. The other end was passed through the provisionally mounted channel and attached some distance outside it to a point which we shall call Z. The length was adjusted so that the wire passed the calculated ultimate node N_2. The position of Z was chosen so that the line between Z and the outlet of the channel satisfied the abovementioned requirement that the external beam should make an angle of at most 30° with the principal axis of the cyclotron.

Next, the position of the channel was adjusted until the wire passed precisely through its middle. By shifting the point Z towards either side so that the wire only just failed to make contact with the walls of the magnetic channel (see dashed lines in fig. 18) it was possible to determine which part of the ion beam entering the mouth of the channel

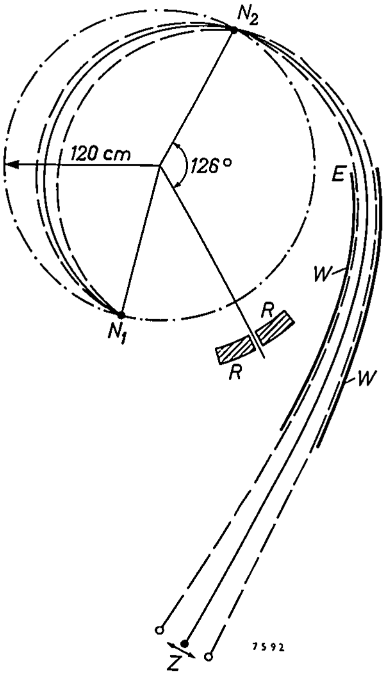


Fig. 18. Finding the optimum position of the channel W by means of the orbit simulator. The ends of the wire are secured at the calculated penultimate node N_1 of the orbit described by the ions, and at an external point Z. The length was adjusted so that the wire passed through the calculated last node N_2.

reaches the outlet, and also the horizontal divergence of the external beam. Originally only the fifth segment of the magnetic channel was single-walled. Calculations — confirmed by the measurements with the orbit simulator — showed that with this configuration only a 1.5 mm wide part of the beam entering the mouth of the channel would reach the outlet. As mentioned, the radial defocusing in the magnetic channel, which was responsible for the poor result, is substantially reduced by also making segments *II*, *III* and *IX* single-walled (see Table II). The consequent decrease in the deflection made it necessary to situate the mouth of the channel 2 cm further outwards. The simulator measurements now showed that an 8 mm-wide portion of the beam incident on the channel mouth would reach the outlet of the channel. Considering that the channel mouth itself is only 12 mm wide, this is a highly satisfactory result.

Two different check measurements were made with the orbit simulator to determine the optimum position of the regenerator. The first was intended to investigate whether the generation of the forced oscillations with increasing amplitude in fact began at a radius of 120 cm. To this end, use was again made of a closed wire loop, length 2π×120 cm, mounted on a float (cf. fig. 17). The wire was set up in the cyclotron field and the regenerator moved as far as possible away from the middle. The regenerator was then moved (radially) inwards and the position was accurately determined at which the wire became unstable. We call this position R_1.

The second measurement was made with the wire arrangement illustrated in *fig. 19*. This differs from fig. 18 only inasmuch as the "internal" fixed end had been shifted towards a point L in front of the regenerator, situated on the calculated orbit of an ion which describes the path (through N_1) shown as a solid curve in fig. 18 ($\Theta = -27.5°$, $r = 128.8$ cm; the azimuth angle Θ was measured from the regenerator). To start with, the regenerator was moved as far as possible outwards, and the length of the wire adjusted so that the nodal points N_1 and N_2 were passed at equal distances (on the inside and outside respectively). Next, the regenerator was moved inwards and the position was determined at which the wire cut the equilibrium orbit $r = 120$ cm exactly at N_1 and N_2. Since the channel position corresponding to N_1 and N_2 had already been found (see above), the wire then runs automatically through the middle of the channel. We call this position of the regenerator R_2

It was found that the regenerator positions R_1 and R_2, and the previously calculated position used for applying the correction plates, differed by less than 1 millimetre.

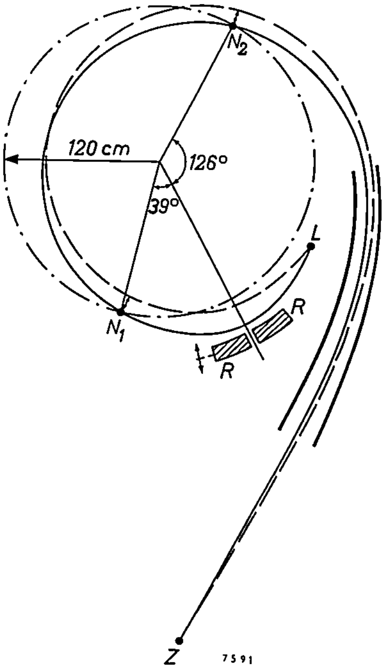


Fig. 19. To find the correct radial position of the regenerator R, the internal fixed end of the wire was set at point L ($\Theta = -27.5°$, $r = 128.8$ cm) situated on the calculated orbit of an ion that follows, starting from N_1, the path shown as a full line in fig. 18. The regenerator was moved from outside inwards until the wire passed through the nodes N_1 and N_2.

Finally, *fig. 20* shows an autoradiograph of the cross-section of the external beam (in this case a proton beam) 3 metres behind the outlet of the extraction system.

Nature of the external beam

If no special measures are taken, the current of the accelerated ions delivered by a synchrocyclotron is not continuous but consists of short pulses, the interval between which is equal to the oscillation period of the modulator. In the present cyclotron the length of these pulses is about 25 μsec and the interval between them is about 2200 μsec; this means that the current in the pulses is about 100 times stronger than the average current.

After completion of this cyclotron, Dr. A. Cabrespine (University of Paris) fitted an auxiliary device which "spreads" the pulses so as to produce a more or less continuous ion current [11]). In this way the maximum count rate to be handled by the counting equipment used in the investigations is appreciably reduced, causing a decrease in the fraction of particles (or quanta) *not* counted and

[11]) See A. Cabrespine, J. Phys. Radium 21, 332, 1960. This method of spreading out the pulses is based on an idea of R. Keller (Int. Conf. on high energy accelerators and instrumentation, CERN, 1959, p. 187).

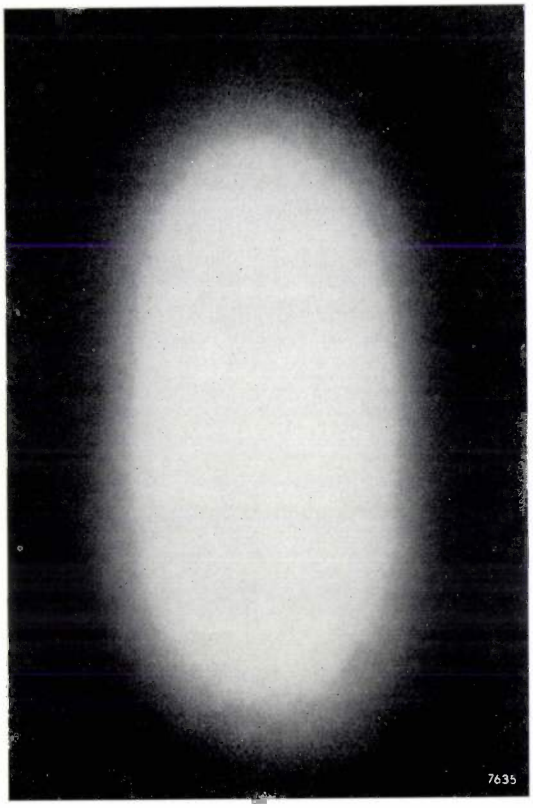


Fig. 20. Autoradiograph (true size) of the cross-section of the external proton beam, three meters behind the exit of the extraction system. The quadrupole lenses (2 in fig. 1) were not in operation when this autoradiograph was made.

also in the number of random coincidences in coincidence measurements. As a result the precision of the measurements is increased and the duration of the measurements can be considerably shortened. The method of spreading out the pulses is as follows.

The oscillator is switched off during that part of each modulation period in which the frequency would pass through the range of minimum values, and in this way the ions are accelerated to an energy which is not quite sufficient to bring them under the influence of the regenerator. Instead of being extracted, they thus continue revolving in the last orbit reached, and after a short time they populate the entire periphery. When the accelerating voltage is switched off, there is no longer any phase stability and the effect of small differences in energy is soon felt. The energy which the ions still lack in order to be extracted is imparted to them by means of an auxiliary electrode. (In figs. 1 and 10 this should be drawn near the right side wall of the accelerating chamber between the regenerator and the ion-source carriage.) This electrode is connected to a low-power RF generator whose average frequency is equal to the average value of the frequencies which the large generator would otherwise have swept during the quiescent period. The frequency of this small generator is very rapidly modulated (for example, modulation frequency 30 000 c/s and frequency sweep 300 kc/s). As a result of this very rapid frequency modulation, combined with the fact that the ions no longer revolve as a group but are uniformly distributed over the entire periphery, during every frequency sweep (of the small generator) only a small fraction of the ions which together would otherwise have delivered a single pulse is accelerated to the energy at which extraction is possible. Remarkably enough, this process can be continued for some time after the main generator has again been switched on and has started to accelerate a succeeding group of ions: the pulse can thus in fact be spread out. Depending on the circumstances, the average intensity of the beam in this process is 20 to 25% of what would otherwise be obtained.

III. CALCULATION AND DESIGN OF THE BEAM-EXTRACTION SYSTEM

In view of the large radial amplitudes that occur in the last phase of the extraction, we used in our design of the beam-extraction system not only the equations (5) and (6), mentioned in Part II, but also, where necessary, the more accurate equations:

$$r'' = -r\frac{rB - r_0B_0}{r_0B_0} = -\frac{r}{r_0}\left(r - r_0 + r\frac{B-B_0}{B_0}\right), \qquad (7)$$

$$z'' = z\frac{r^2}{r_0B_0}\frac{\mathrm{d}B}{\mathrm{d}r}. \qquad (8)$$

Here B_0 is an abbreviated notation for $B(r_0)$ and $r'' = \mathrm{d}^2r/\mathrm{d}\Theta^2$ etc. The main difference compared with equations (5) and (6) is that B in this case is not approximated by $B_0 + (\mathrm{d}B/\mathrm{d}r)(r - r_0)$. If the amplitude is allowed to approach zero, equations (7) and (8) reduce to (5) and (6).

Equations (7) and (8) are entirely adapted to the situation that arises during the extraction: the amplitude of the radial oscillations then increases considerably whilst that of the vertical oscillations remains small. As can be seen, the coupling between the radial and vertical oscillations, which was not included in (5) and (6), is now taken into account in the equation for the vertical oscillation — r appears in (8) and cannot be approximated by r_0 — but not in that for the horizontal oscillation. The neglect of the influence of the vertical oscillation on the horizontal is an essential feature of the method of calculation discussed in this part of the article. Numerical computations which did allow for the latter influence have shown that this neglect is permissible.

Since the dimensions of a regenerator in the orbital direction are small compared with the wavelength of the oscillations described by the ions, the regenerator in our calculations is replaced by an equivalent infinitely thin regenerator. Its strength can be characterized by a function I which depends solely upon r:

$$I(r) = \int \frac{\Delta B}{B_0} r_0 \mathrm{d}\Theta \,. \quad \ldots\ldots \quad (9)$$

Here ΔB is the difference between the field inside the regenerator and that outside it, both at a radius r. Assigning to the regenerator the azimuth $\Theta = 0°$ and giving symbolically the quantities directly in front of the regenerator the coordinate $-0°$ and those behind it $+0°$, we can find by integration from (7) and (8) that the radial motion for every passage of an ion is given by

$$r(+0) = r(-0)\,, \quad \ldots\ldots\ldots \quad (10a)$$

$$r'(+0) = r'(-0) - \left(\frac{r}{r_0}\right)^2 I(r)\,, \quad \ldots \quad (10b)$$

and the vertical motion by:

$$z(+0) = z(-0)\,, \quad \ldots\ldots\ldots \quad (11a)$$

$$z'(+0) = z'(-0) + z\left(\frac{r}{r_0}\right)^2 \frac{\mathrm{d}I}{\mathrm{d}r}\,. \quad . \quad (11b)$$

We shall first briefly illustrate how the orbit of an ion can be found when the function $I(r)$ is known, after which we shall discuss how, conversely, starting from a required form of orbit, we designed a suitable regenerator.

In computing the behaviour of an ion in a cyclotron with a regenerative beam-extraction system, we divide the history of the ion into four periods:

a) The period in which the ion is accelerated in the normal way.

b) The period in which, under the action of the regenerator, the free oscillations become forced oscillations. The amplitudes are still small. This period has an important bearing on the spread in the energy of the extracted ions [12]).

c) A period in which the amplitude of the forced oscillations, while still being small, is nevertheless already too large for one to be able to use equations (5) and (6). Periods *b* and *c* together comprise a large number of revolutions.

d) The last period before the ion enters the magnetic channel. In this period, which comprises only a few revolutions, the amplitude rapidly increases (cf. fig. 9). Accurate analysis of this period is particularly necessary to the design of an effective regenerator and the magnetic channel.

We shall now examine the behaviour of an ion during periods c and d, dealing separately with the radial and the vertical oscillations. We begin with period d, and assume for the moment that we know the state of oscillation at the end of period c.

We describe the radial behaviour (outside the regenerator) of an ion on an orbit with "equilibrium radius" r_0 by the variation of the deviation $\varrho = r - r_0$ along the orbit. For period d we find this, using an electronic computer, by solving equation (7) for the measured variation of $B(r)$ around $r = r_0$ and for a chosen value of the amplitude $\varrho_{\max}$. It is simpler here to take as our starting point not the amplitude $\varrho_{\max}$ itself but the slope ϱ_0' of the orbit in relation to the equilibrium orbit at the point where these intersect. Examples of such solutions for this cyclotron are given in *fig. 21*. It can clearly be seen that the small-amplitude oscillation is almost sinusoidal, but that the large-amplitude one is not. The outward deflections, bringing the ion into the fringing field, are larger and last longer than the inward deflections. In fig. 21 the azimuth coordinate used is not the angle Θ with respect to the regenerator but the angle ψ with respect to the last point N_2 where the ion passes through zero before reaching the regenerator (cf. fig. 9). We shall use this azimuthal quantity in all our subsequent calculations. The ψ values of the regenerator at the beginning and end of a revolution are denoted as η and $\eta + 360°$. The angular position of the oscillation with respect to the regenerator will be characterized by this angle η, which for convenience we shall refer to as the "phase" of the oscillation (cf. fig. 9).

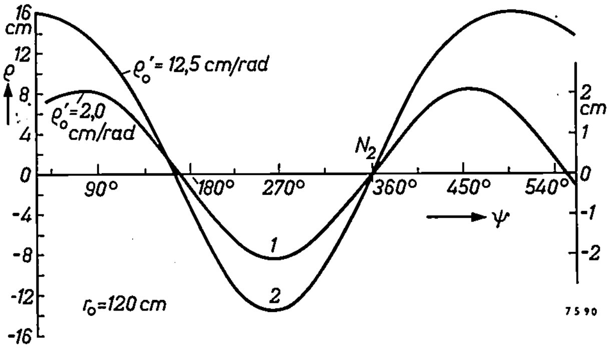


Fig. 21. Solutions of the radial orbit equation for orbits having an equilibrium radius of 120 cm. In orbit *1*, ϱ_0' is small (2 cm/rad), in orbit *2* it is large (12.5 cm/rad). In orbit *1* the ion does not enter the fringing field; the oscillation is almost sinusoidal. In orbit *2* there is marked asymmetry: the deflection outwards is greater than that inwards and also lasts longer. The solutions were obtained by going from N_2 (cf. fig. 9) first towards smaller and then towards larger azimuth values. The azimuth measured with respect to N_2 is denoted by ψ.

[12]) See the article by N. F. Verster in Sector-focused cyclotrons (Proc. Conf. Sea Island, Georgia, 2-4 February 1959), p. 199-202.

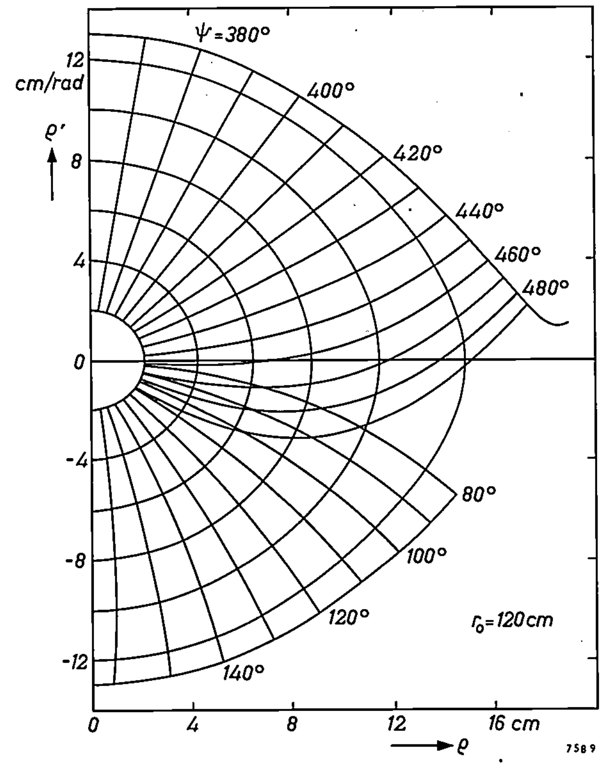


Fig. 22. Phase plot characterizing the nature of the free radial oscillation of an ion in a cyclotron. The deflection ϱ is plotted versus its derivative ϱ' ($= \mathrm{d}\varrho/\mathrm{d}\Theta$). The left half of the figure (ϱ negative) is omitted. A stable (undamped) oscillation of specific amplitude is represented by a closed curve, a harmonic oscillation by an ellipse. (In the scale on which the figure is drawn the ellipses appear as circles.) In the case of large-amplitude oscillations the influence of the boundary field is clearly visible: the curves become ovoid. The lines through the origin are lines of equal "phase" η (see fig. 9). Since the oscillation frequency is not as a rule identical with the cyclotron frequency, the isophase lines for 80° and 440°, for example, do not coincide. The figure relates to oscillations of the ions in the Orsay cyclotron about an equilibrium orbit of 120 cm radius.

Apart from in a figure of the type of fig. 9, the nature of the oscillation can also be represented graphically in a "phase plot", which shows how the deflection ϱ of an oscillation having a specific amplitude varies with the radial "velocity" ϱ'. For every stable oscillation a closed curve is produced, and for a sinusoidal oscillation an ellipse (*fig. 22*). This is a particularly clear method of presentation.

In the regenerator the ion undergoes a change in direction, but r remains unchanged (see eq. 10a) and therefore the same applies to ϱ. In the phase plot the point that descibes the situation thus drops a distance $(r/r_0)^2 I(r)$ vertically downwards at a certain moment (see eq. 10b). Generally speaking the phase η of the radial oscillation has a different value in every successive revolution. As can be seen, an oscillation of $\Theta = +0$ to $\Theta = 2\pi - 0$ (eq. 10) is entirely governed by ϱ_0' and η.

Let us now examine the vertical behaviour. We again consider one revolution from $\Theta = +0$ to $\Theta = 2\pi - 0$, of an ion whose radial oscillation is ϱ_0', η. If we describe the behaviour by means of the vector **z**, whose components are z and z' — just as we described the radial oscillation by the phase plot of ϱ against ϱ' — then since the vertical orbit equation (8) is linear in z, we can find the components of $\mathbf{z}(2\pi - 0)$ from those of $\mathbf{z}(+0)$ via a set of two linear equations. Putting the four coefficients of these equations into matrix form $||P(\varrho_0',\eta)||$, we then have

$$\mathbf{z}(2\pi - 0) = ||P(\varrho_0',\eta)||\, \mathbf{z}(+0)\,. \quad . \quad . \quad (12)$$

The elements of the matrix $||P||$ — i.e. the four coefficients — can be found, by the method indicated in *fig. 23*, from the two independent solutions of eq. (8) using respectively the initial conditions $z = 0$, $z' = 1$ cm/rad and $z = 1$ cm, $z' = 0$.

We write equations (11), which describe the influence of the regenerator on z and z', likewise in matrix form:

$$||R|| = \left\| \begin{matrix} 1 & 0 \\ \left(\dfrac{r}{r_0}\right)^2 \dfrac{\mathrm{d}I}{\mathrm{d}r} & 1 \end{matrix} \right\|,$$

and we thus have for a complete revolution:

$$\mathbf{z}(2\pi + 0) = ||R|| \times ||P||\, \mathbf{z}(+0) = ||A||\, \mathbf{z}(+0). \quad (13)$$

Eq. (6) still holds for small ϱ_0' and we have

$$||P|| = \left\| \begin{matrix} \cos 2\pi\omega & \dfrac{1}{\omega} \sin 2\pi\omega \\ -\omega \sin 2\pi\omega & \cos 2\pi\omega \end{matrix} \right\|,$$

so that the determinant $|A|$ is equal to 1. It can be proved that this also holds for large ϱ_0' (provided we use for the orbit equation a form in which — as in (8) — there is no term with z'). If we examine a

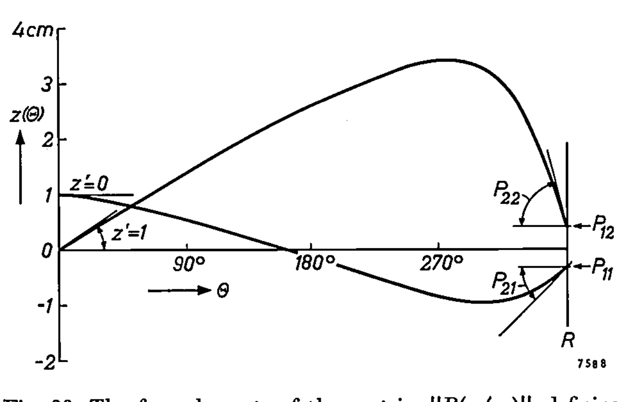


Fig. 23. The four elements of the matrix $||P(\varrho_0',\eta)||$ defining the change which the vertical oscillation state (z,z') of an ion undergoes outside the regenerator R in one revolution, are found by solving for the revolution concerned the vertical orbit equation with the initial conditions $z = 0$, $z' = 1$ and $z = 1$, $z' = 0$, also taking into account the radial oscillation. The two values of z and the appertaining slopes of the curves at the end of the revolution give the four elements in question. The curves shown here hold for the case $r_0 = 120$ cm, $\varrho_0' = 13.0$ cm/rad and $\eta = 110°$.

group of ions of identical amplitude h but different phases, we find that the points in the z,z' plane which characterize the state of oscillation of these ions together constitute an ellipse of surface area $\pi\omega h^2$. From the fact that $|A| = 1$, it may be inferred that, however the shape of this ellipse may vary from revolution to revolution, the enclosed area remains constant.

During the period of the extraction in which the amplitude is small (period c) the form of $||A||$ changes very little per revolution ("adiabatic process"). During the period d, however, there is such a marked variation in $||A||$ that it is necessary to calculate the elements of the matrix separately for each new revolution. Because the period c (small amplitude) comprises a very large number of revolutions, it is not possible for this period to follow the change of state per revolution with the matrix method. On the other hand, in period d this is the only possible method.

A practical method of mathematically treating the behaviour during period c, both for radial and vertical oscillations, has been given by Le Couteur [8]. We shall mention only the results.

In the first place it is found, as mentioned in Part II, that the phase η, which at the beginning of the deflection was still entirely governed by the free oscillation originally present, converges towards a certain fixed value. This value is determined solely by the cyclotron and regenerator field and is independent of the initial phase. In the second place this theory yields the z,z' ellipse. (We are again concerned with a group of ions of initially identical amplitude and arbitrary phase.) Thirdly, the theory indicates at what point the "deflection" begins, namely on the radius where

$$\mathrm{d}I/\mathrm{d}r = 2\pi(1 - \sqrt{1-n})\,\sqrt{1-n}\,. \quad . \quad . \quad (14)$$

Further, with a view to vertical stability, the value of $\mathrm{d}I/\mathrm{d}r$ must lie within specific limits in the entire regenerator field.

Starting from the value of η and from the z,z' ellipse as yielded by Le Couteur's theory for a given variation of $B(r)$ and $I(r)$, the orbit described by an ion during the last period of the extraction can be found using the ϱ, ϱ' diagram (fig. 22) and the form of the matrix $||A||$ calculated per revolution.

When the extraction system was designed the variation of $B(r)$ was already established, and it was therefore necessary to look for a matching variation of $I(r)$. Since, at $r_0 = 120$ cm, the radius at which we wanted the deflection to begin, the field index n has the value 0.11, we had to ensure that $\mathrm{d}I/\mathrm{d}r$ on this radius would acquire the value 0.34; see eq. (14). Further, the given azimuthal position of the channel mouth was 60° in front of the regenerator ($\Theta = 300°$). The most suitable variation of $I(r)$ was now found by working from back to front, i.e. by beginning with the last period of the extraction.

The first step was to solve eq. (7) for various values of ϱ_0' and the chosen r_0 value of 120 cm [13]. The ϱ, ϱ' phase plot obtained in this way has already been shown in fig. 22.

Next, eq. (8) for the vertical motion was solved in the manner earlier described for each of the orbits of fig. 21 and for a number of chosen values of η between 100° and 150°. This was done in all cases over an interval from $\psi = \eta$ to $\psi = \eta + 360°$. Fig. 23 gave an example of such a solution. The matrix $||P||$ was found in all cases by interpolation from these data (see fig. 23). As regards the form of $z(\Theta)$ in this figure, it should be noted that the maximum value, measured in centimetres, is roughly 3.5 times greater than the initial slope z' measured in cm/rad. It was found that this relationship holds approximately in all instances, and we shall use this rule of thumb in our following considerations. The value of Θ at which the maximum occurs (about 270°) is also in all cases roughly the same.

To work as systematically as possible in finding a suitable $I(r)$, we looked only for functions for which the phase remained constant during the last revolutions. These can easily be derived from the phase plot (fig. 22) by inserting the isophase lines $\psi = \eta$ and $\psi = \eta + 360°$ and then measuring the vertical distance between these lines for various values of ϱ. In this way one finds the variation of $-(r/r_0)^2 I(r)$ with ϱ and hence the variation of $I(r)$. In the region of small ϱ values the curves were extrapolated so as to make them tangent to the line $I = 0.34\ \varrho$ at the origin (see above). The result is presented in *fig. 24.* It was found that the curves can be represented to a good approximation by the formula $I = 0.34\varrho + \beta\varrho^2$.

In order to construct the ion orbits in period d for each of the regenerator functions thus found, we first computed the z, z' ellipse at the end of the penultimate extraction period, and also the (radial) phase η. As we have seen, the latter converges to a constant value, i.e. to the value at which it remains fixed in the last extraction period. (It should be noted here that it is not so important to know η for describing the penultimate extraction period itself; it is only in the period where the amplitude is large that we need to know η.) With the aid of these data we then

[13] The calculations were carried out using the small experimental computer PETER, made in the Philips Research Laboratories and which was available at the time.

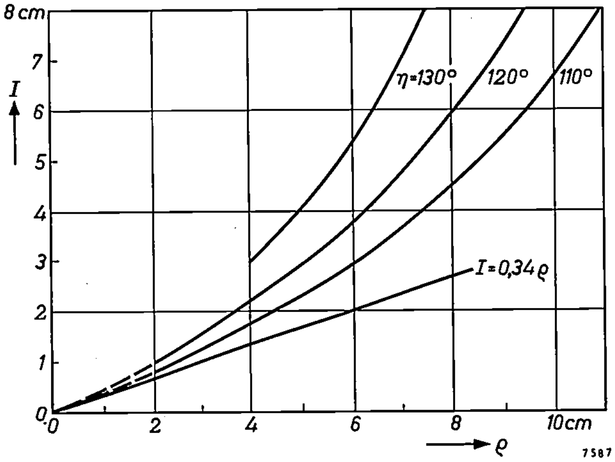


Fig. 24. The required form of the regenerator field I as a function of r, for regenerators that keep the phase η of the oscillations constant. The curves are derived from the phase plot in fig. 22. The quantity ϱ ($= r - r_0$) is plotted as abscissa. The curves hold for the indicated values of η. In the origin they are all tangent to the straight line $I = 0.34\ \varrho$.

constructed the further course of the orbits for various initial conditions: the radial oscillation from the phase plot and the two isophase lines referred to, and the vertical oscillation from the matrix $||A||$, the elements of which can be calculated (eq. 13) now that $||P||$ and $I(r)$ are known. As mentioned, the latter calculation had to be repeated for every revolution.

The radial distance between the ultimate and penultimate revolutions of the ions arriving in the channel mouth, and also the position of the (virtual) source of the ion beam entering the channel, can directly be found from these calculated orbits. (The positions of the virtual source derived from the vertical divergence will not necessarily coincide with that found from the horizontal.)

Now it was found that, to obtain a suitable radial increase of ϱ combined with a not unduly large vertical amplitude, the value of η had to be chosen between 110° and 130°. This may be inferred from *fig. 25.*

This figure shows the upper and lower limits of $\mathrm{d}I/\mathrm{d}r$ which must not be exceeded if vertical "stability" is to be ensured, plotted against ϱ for

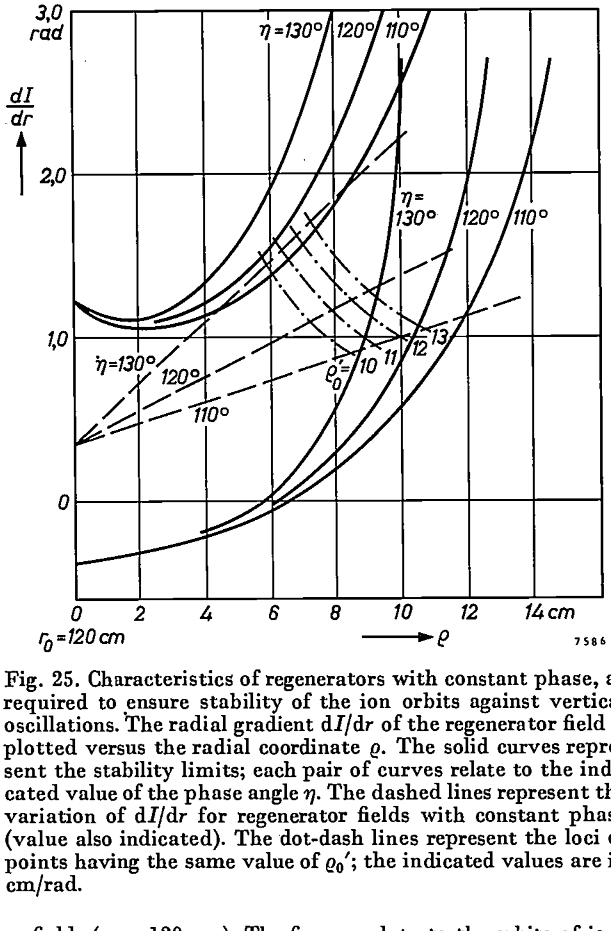


Fig. 25. Characteristics of regenerators with constant phase, as required to ensure stability of the ion orbits against vertical oscillations. The radial gradient $\mathrm{d}I/\mathrm{d}r$ of the regenerator field is plotted versus the radial coordinate ϱ. The solid curves represent the stability limits; each pair of curves relate to the indicated value of the phase angle η. The dashed lines represent the variation of $\mathrm{d}I/\mathrm{d}r$ for regenerator fields with constant phase (value also indicated). The dot-dash lines represent the loci of points having the same value of ϱ_0'; the indicated values are in cm/rad.

Table III. Principal data of the three carefully analysed regenerator fields (r_0 = 120 cm). The figures relate to the orbits of ions which, in their last revolution, pass the equilibrium orbit with a slope of 13 cm/rad.

Regenerator field	Characteristic quantities of ion revolution			In the regenerator (Θ = 360°)			At the channel mouth (Θ = 300°)		Line 1 to 4: the matrix $\|\|A\|\|$ Line 5: $\|\|P(\Theta = 300°)\|\|$				$z_{\max}/h$ for three azimuth values		
	No.	ϱ_0'	η	ϱ	I	$\left(\frac{r}{r_0}\right)^2\frac{\mathrm{d}I}{\mathrm{d}r}$	ϱ	ϱ'	A_{11}	A_{12}	A_{21}	A_{22}	Θ = 0°	Θ = 270°	Θ = 300°
I_1	n—4	3.4	121°	3.4	1.72	0.68			—0.38	2.78	—0.49	0.96	0.98	1.5	
	n—3	4.6	121°	4.5	2.54	0.82			—0.38	2.65	—0.52	0.99	0.98	1.5	
	n—2	6.1	121°	6.2	4.01	1.01	5.5	3.3	—0.39	2.33	—0.54	0.65	0.93	1.5	
	n—1	8.5	121°	9.1	7.28	1.35	7.6	4.9	—0.38	1.36	—0.39	—1.25	0.88	1.5	
	n	**13.0**	120°				11.7	8.1	—0.48	3.27	—0.11	—1.37	0.52	2.6	2.6
I_2	n—4	3.0	124°	2.8	1.50	0.75			—0.37	2.77	—0.52	1.20	1.00	1.5	
	n—3	4.0	124°	3.8	2.32	0.90			—0.36	2.67	—0.55	1.33	1.00	1.5	
	n—2	5.4	124°	5.3	3.79	1.12	5.1	2.7	—0.36	2.42	—0.60	1.25	1.00	1.6	
	n—1	7.8	125°	8.1	7.38	1.52	7.3	4.1	—0.35	1.73	—0.63	0.23	0.95	1.6	
	n	**13.0**	125.5°				12.3	7.5	—0.28	3.16	—0.16	—1.74	0.62	1.8	1.7
I_3	n—4	2.5	130°	2.1	1.16	0.77			—0.36	2.78	—0.54	1.36	1.02	1.6	
	n—3	3.3	130°	3.0	1.95	0.97			—0.34	2.68	—0.58	1.63	1.02	1.6	
	n—2	4.8	129°	4.5	3.87	1.27	4.6	2.0	—0.33	2.50	—0.66	1.88	1.07	1.7	
	n—1	7.3	130°	7.2	7.64	1.88	7.1	3.2	—0.32	1.76	—0.77	1.09	1.13	2.0	
	n	**13.0**	130.2°				13.0	6.9	—0.16	3.02	—0.22	—2.09	0.73	1.6	1.4

three values of η, and also the variation of $\mathrm{d}I/\mathrm{d}r$ with ϱ for the regenerator fields corresponding to these η values (dashed lines of slope 2β). The dot-dash lines in this figure are curves of constant ϱ_0'. From the latter one can find the maximum value of ϱ_0', i.e. the maximum useful amplitude corresponding to the relevant value of η. As can be seen, this increases with η.

It is not in the least necessary, however, to try to make ϱ'_0 as large as possible. At the position of the channel mouth ($\Theta = 300°$) the value of $\varrho_n - \varrho_{n-1}$ (which increases with ϱ_0' and η) need not be made larger than is necessary for "peeling off" the ions (e.g. 5 cm). By making the value larger we only reduce the fraction of the ions entering the channel mouth. A reasonable compromise is obtained with a value of about 120° for η.

On the basis of these results we made a closer analysis of three regenerator fields, having thc β values 0.05, 0.07 and 0.09. We call these fields respectively I_1, I_2 and I_3. The value of η for these fields is not exactly constant — the formula $I = 0.34\varrho + \beta\varrho^2$ was no more than a good approximation — and it therefore had to be determined for each revolution. As initial values we chose those underlying the construction of the curves in fig. 24. The values for revolutions with greater amplitudes were found by drawing the relevant orbit in the ϱ, ϱ' plane. From a point having the initial η value we draw a line on the ϱ,ϱ' curve to the point with phase $\eta + 360°$, and then go vertically downwards over the distance corresponding to $I(r)$. In this way we reach the isophase line corresponding to the new phase, and so on. In addition we constructed for all three fields the orbit ending with an oscillation for which $\varrho_0' = 13$ cm/rad. From this the components of $||P||$ were computed in the manner described, and then those of $||A||$, using $I(r)$. The results are collected in *Table III*. It can be seen that the elements of $||A||$ only begin to change markedly during the last two revolutions. Up to the orbit $n - 2$ we therefore used the z,z' ellipse applicable, according to Le Couteur's theory, to the first period. (It will be recalled that this ellipse describes the state of oscillation of a group of ions of initially identical amplitude and arbitrary phase after leaving the regenerator ($\Theta = +0$).) This then, gives the vertical structure of the ion beam at the beginning of revolution $n - 1$. From this point onwards the structure was calculated using the relevant form of the matrices $||P||$ and $||R||$. The result is shown in *fig. 26*. The five ellipses, drawn in each figure, apply respectively to the end of revolution ($n - 2$), i.e. just in front of the regenerator, to the beginning of revolution ($n - 1$), i.e. just behind the regenerator, and so on. The last one applies at the location of the channel mouth. In fig. 26 ellipse *2* is thus the one which was computed by Le Couteur's

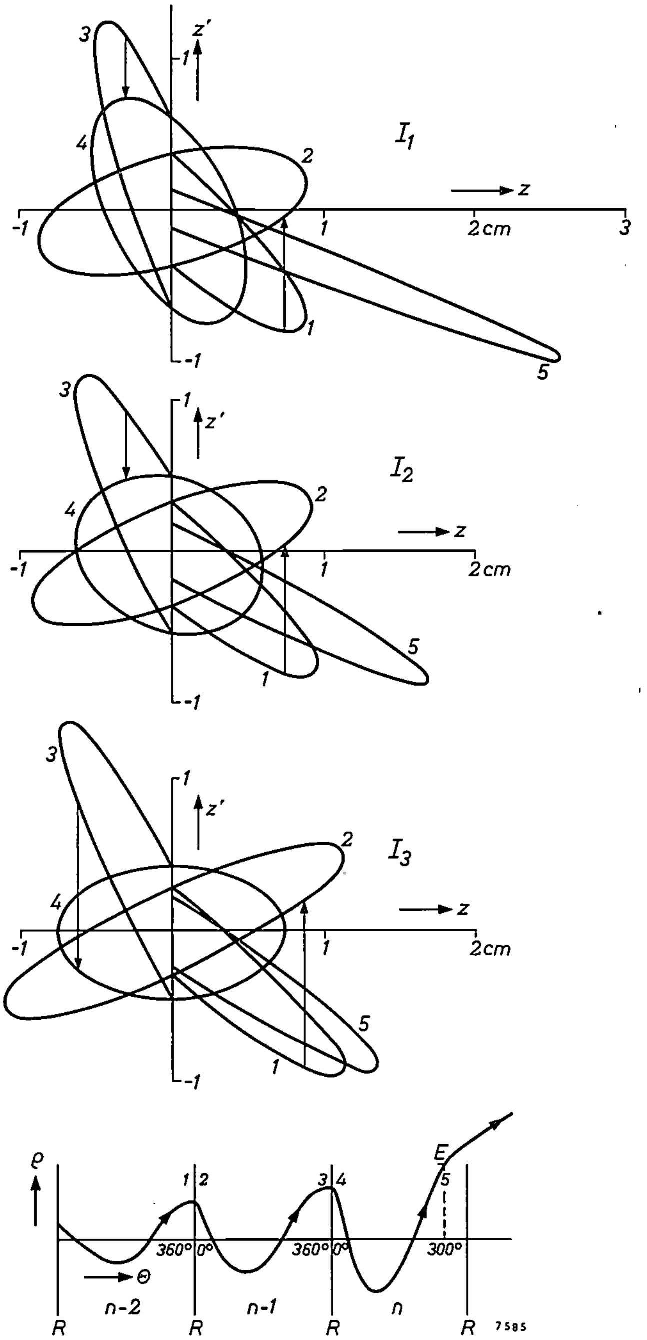


Fig. 26. The vertical structure of the ion beam during the last two revolutions, for the three completely calculated regenerator fields I_1, I_2 and I_3, and relating to a group of ions which, before the beginning of the extraction process, had the same vertical amplitude and arbitrary phase. This structure is described by an ellipse in the z,z' plane, which gradually changes its shape and position with increasing ϱ, but not its surface area. For clarity, only the halves of some ellipses are shown. Ellipses *1* to *5* relate to the following points (see bottom waveform and also fig. 9): *1* at the end of revolution ($n - 2$), just before the regenerator; *2* at the beginning of revolution ($n - 1$), just behind the regenerator; *3* at the end of revolution ($n - 1$); *4* at the beginning of revolution n; *5* at the channel mouth. The vertical arrows represent the change in z' produced in an arbitrary ion when it passes the regenerator. The arrow between the ellipses *1* and *2* and that between *3* and *4* in each diagram do not refer to the sameion.

method. Ellipse *1* was found by working backwards.

Since, as we have seen, the maximum deflection in the z direction is roughly $3.5z'_{\max}$ $(+\ 0)$ in the last revolution (see fig. 23), we must ensure that ellipse *4* in fig. 26 is not too high. From this standpoint, the regenerator fields I_2 and I_3 are preferable to I_1. From these two we made our final choice of I_2, because the relevant regenerator is somewhat easier to realize, and also because the difference $\varrho_n - \varrho_{n-1}$ at $\Theta = 300°$ was 5 cm in this case, which is sufficient, as against 6 cm in the case of field I_3. *Fig. 27* gives a representation in the ϱ,ϱ' plane of the orbit of an ion which, under the influence of the regenerator, finally arrives in the channel. For constructing the orbit curve in this figure, use was made of the $I(r)$ function of the actual regenerator; this I function is practically identical with I_2.

Summarizing the principal characteristics of the regenerator, and of the method of treatment described in Part III of this article, we note first of all that we have obtained a regenerator which, at the end of the last extraction period, causes a sufficiently large increase in the radial coordinate of the ions per revolution at $\Theta = 300°$ without substantially increasing the height of the beam. The method of computation outlined also makes it possible to investigate quickly, during the construction of an extraction system, the influence of discrepancies between the regenerator field chosen on theoretical grounds and that which is present at a given moment. It takes a great deal of time to obtain a desired field exactly, and it is therefore of great importance to be able to investigate what deviations are tolerable. Fortunately, these were found to be fairly considerable in the region of the last revolution. Turning again, for instance, to the z,z' ellipses for I_2 (fig. 26), we see that the arrow which describes events during the last passage of the regenerator can comfortably be about 20% shorter or about 40% longer: the shape of ellipse *4* — which cuts ellipse *3* on the z' axis — is not changed thereby sufficiently to seriously affect the maximum value of z', which of course determines the maximum value of z. Since the length of this arrow is proportional to the gradient $\mathrm{d}I/\mathrm{d}r$ of the regenerator field (see eq. 11b), the tolerable variation mentioned also applies to the latter quantity.

The effect of a change in I for large ϱ on the radial behaviour can be seen from fig. 27. In this figure the above variation of I results in a proportional variation of the largest of the vertical lines. As regards the orbit drawn in the figure, this means that the bottom of the line comes to lie somewhat higher or lower. The phase η for large ϱ does not therefore remain constant but either increases or decreases slightly. If we construct for the new phase line the orbit that ends at E, we find that, as far as the increase of ϱ during the last revolutions is concerned, this orbit is hardly less suitable than the drawn one.

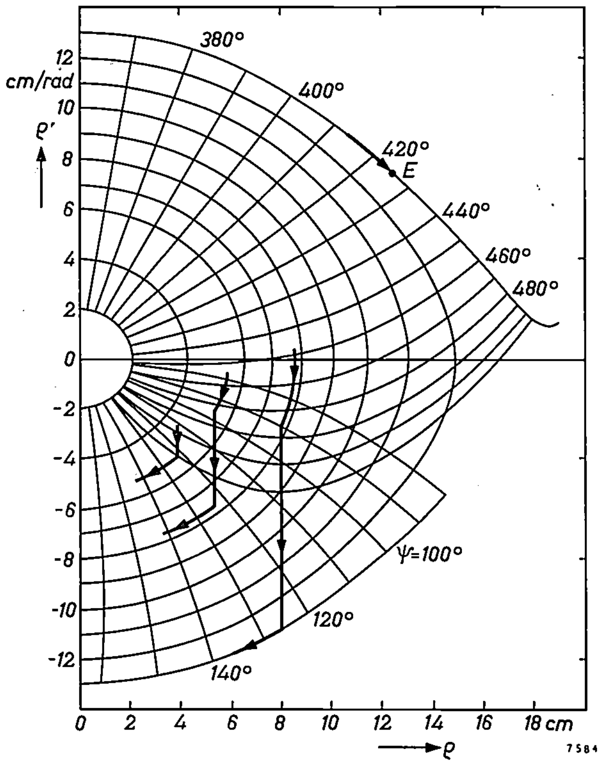


Fig. 27. Phase plot of the Orsay cyclotron (cf. fig. 2) showing the last three revolutions of an ion which finally arrives in the channel. The point E represents the state upon entering the channel.

Summary. The synchrocyclotron built by Philips at Orsay for the University of Paris is capable of accelerating protons up to about 160 MeV, and deuterons up to about 80 MeV. The RF system, with flywheel circuit, is designed to enable a change-over to be made from proton to deuteron frequencies by means of a simple structural alteration. A radially adjustable target makes it possible to produce a neutron beam of variable energy. By means of a beam-extraction system an external beam (proton energy 157 MeV) is obtained whose intensity, depending on the current, is 5 to 12% of that of the internal beam. The extraction system induces a radial oscillation of the ions around their equilibrium orbit, enabling the ions to enter a magnetic channel without being lost by collision with the channel wall, and without the radius having to be so large as to make the equilibrium orbit coincide with the circle on which the field index has the dangerous value of 0.2. The ions are then extracted through the magnetic channel. The oscillations begin when the orbit radius is 120 cm and are generated by a local field disturbance having a positive radial gradient, produced by a device referred to as a regenerator. The amplitude of the vertical oscillations remains small because the ions in part of their orbit enter the fringing field (where the gradient is strongly negative). A detailed treatment is given of the calculations used to find a suitable form of regenerator. Also discussed are the construction of regenerator and channel, an investigation of the magnetic field using a copper wire to simulate the ion orbits and further the adjustment of regenerator and channel, likewise using the orbit simulator.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS BY THE STAFF OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an asterisk * can be obtained free of charge upon application to the Philips Research Laboratories, Eindhoven, Netherlands, where a limited number of reprints are available for distribution.

2932: B. G. van den Bos: Investigations on pesticidal phosphorus compounds, III. The structure of the reaction product of 3-amino-5-phenyl-1,2,4-triazole and bis(dimethylamido)phosphoryl chloride (WP 155) (Rec. Trav. chim. Pays-Bas **79**, 1129-1136, 1960, No. 11).

The reaction of 3-amino-5-phenyl-1,2,4-triazole with *bis*(dimethylamido)phosphoryl chloride gives a compound (WP 155) with a strong anti-mildew effect. Comparison of the UV absorption spectra of a bromine compound derived from WP 155 and the three N-ethyl-3-bromo-5-phenyl-1,2,4-triazoles with those of the three known N-methyl-3-methyl-5-phenyl-1,2,4-triazoles indicates that WP 155 is 5-amino-1-[*bis*(dimethylamido)phosphoryl]-3-phenyl-1,2,4-triazole.

2933: J. H. Uhlenbroek and J. D. Bijloo: Investigations on nematicides, III. Polythienyls and related compounds (Rec. Trav. chim. Pays-Bas **79**, 1181-1196, 1960, No. 11).

Part of an investigation of the possibilities of using polythienyls as nematicides (i.e. for control of eel-worms; see also these abstracts, Nos **2692**, **2743** and **2786**). A number of isomeric polythienyls and some mixed polyphenylthienyls were tested as to their nematicidal effect. All compounds which gave good results proved to be derivatives of 2,2′-bithienyl. Some new compounds are described, e.g. 2,3′-4′,2″-terthienyl, 1,4-di(2-thienyl)-benzene and various methyl polythienyls.

2934: J. R. Roborgh and Th. J. de Man: The hypercalcemic activity of dihydrotachysterol$_2$ and dihydrotachysterol$_3$ and of the vitamins D$_2$ and D$_3$ after intravenous injection of the aqueous preparations, II. Comparative experiments on rats (Biochem. Pharmacol. **3**, 277-282, 1960, No. 6).

Vitamin D$_2$, vitamin D$_3$ and the dihydrotachysterols derived from them have a hypercalcemic effect (i.e. they can increase the calcium content of the blood serum). In an earlier investigation (these abstracts No. **2785**) it was shown that, when orally administered as a solution in oil, dihydrotachysterol$_3$ had the greatest effect, followed by dihydrotachysterol$_2$, vitamin D$_3$ and vitamin D$_2$. The same order of activity is found when these compounds are injected intravenously as an aqueous dispersion. The blood serum shows the highest calcium content two to four days after administration. The relative activities of the various compounds are found to depend strongly on the length of time which elapses between the administration of the preparations and the sampling of the blood sera, and are in reasonably good agreement with the results of the previous investigation.

2935: W. Duyfjes: Interfacial phenomena in pesticide application (Pest Technol. **2**, 239-243, 1960, No. 11).

Many pesticides are applied in aqueous media. Since most of them are not soluble in water, they must be used as an emulsion or suspension. A number of physical problems which arise in this connection in the formulation of pesticides are discussed in this publication. See also Philips tech. Rev. **19**, 165, 1957/58.

2936: E. W. Gorter: Some structural relationships of ternary transition metal oxides (XVIIth International Congress of Pure and Applied Chemistry, Munich 1959, Part I, pp. 303-328, Verlag Chemie, Weinheim 1960).

The oxides of metals of the first transition group have been much studied, because they can be put to many uses. This article deals only with compounds of the type B_nO_p and $A_mB_nO_p$, where B is a cation which can fit in the interstices of a close-packed array of oxygen ions, e.g. a transition-metal ion, and A is a cation which is too large for this. A new method of describing crystal structures is discussed, where the crystal is imagined to be cut up into a number of blocks by planes mid-way between crystallographically equivalent or chemically identical atoms. This method is illustrated by listing all the ways in which an ion of a close-packed array can be surrounded by six-coordinated cations, and discussing all the structures thus obtained.

The different ways of modifying simple oxides by substitution and the formation of superstructures are discussed for the cases of the NaCl, spinel, corundum and rutile structures. A number of existing compounds of the type $A_mB_nO_p$ are shown in an A-B-O diagram. The structural relationships between a number of compounds near the point $AB\mathrm{O}_3$, and between others near the points $B_2\mathrm{O}_3$ and $B_3\mathrm{O}_4$, are discussed.

2937: J. H. Spaa: A rapid indicating instrument for the stepwise measurement of airdust radioactivity (Progr. nucl. Energy, series 12 (Health Physics) **4**, 25-31, 1960).

Description of an instrument for the rapid detection of radioactive dust in the air, which has already been described in abstract No. **2823a.** The present article also discusses the experience gained with the instrument.

2938: F. K. Lotgering: Topotactical reactions with ferrimagnetic oxides having hexagonal crystal structures, II (J. inorg. nucl. Chem. **16**, 100-108, 1960, No. 1/2).

If a mixture of powdered oxides of the system BaO-MeO-$\mathrm{Fe_2O_3}$ (where Me = Co, Ni, Zn, etc.) consists *in part* of an oxide which can be magnetically oriented, while the rest of the sample cannot, then if the mixture is heated after magnetic orientation, the reaction product is often found to consist of a single-phase oxide which is *highly* oriented. Such reactions have been given the name "topotactical reactions" (see these abstracts No. **2707**). The present publication deals with an investigation of the mechanism of various topotactical reactions, and of the effect of sintering on the degree of orientation of the reaction products. See also Philips tech. Rev. **20**, 354, 1958/59.

2939: J. N. Walop, Th. A. C. Boschman and J. Jacobs: Affinity of N-acetylneuraminic acid for influenza virus neuraminidase (Biochim. biophys. Acta **44**, 185-186, 1960, No. 1).

Preliminary report of a new method for the determination of neuraminidase activity. The system of influenza virus and N-acetylneuraminyllactose is found to follow Michaelis-Menten kinetics. It is shown that N-acetylneuraminic acid (NANA) is a competitive inhibitor of this reaction. Comparison of the inhibitor constant with the Michaelis constant for the substrate shows that the affinity of NANA for the enzyme molecule is strengthened 8-fold by the ketosidic linkage of the NANA moiety to lactose.

2940: W. Albers and J. Th. G. Overbeek: Stability of emulsions of water in oil, III. Flocculation and redispersion of water droplets covered by amphipolar monolayers (J. Colloid Sci. **15**, 489-502, 1960, No. 6).

It is shown by simple calculations that emulsions of water in oil cannot be sufficiently protected against flocculation by an adsorbed layer of amphipolar molecules with an oleophilic chain of about 20 Å at the surface of the water droplets. Neither does the combination of electrical charge and adsorbed layer prevent flocculation. It might however be expected that the flocculated system could be redispersed by means of not too large shearing forces. This is confirmed by viscosity measurements. Non-Newtonian behaviour found at low rates of shear is a consequence of flocculation. From the minimum rate of shear required to reach the Newtonian region (where the rate of shear is a linear function of the shear stress), i.e. to cause complete redispersion, the effective Van der Waals constant A between the water droplets can be estimated. The value of $A = 4 \times 10^{-15}$ erg is found.

2941: J. L. Meijering: Hardening by internal oxidation as a function of velocity of the oxidation boundary (Trans. Metall. Soc. AIME **218**, 968-971, 1960, No. 6).

The hardness of small metal cylinders or spheres which have been hardened by internal oxidation first decreases with increasing distance from the surface, but then increases again. This is in agreement with the theory that the change in hardness is mainly determined by the change in the rate of advance of the oxidation boundary, which velocity determines the dispersion of the oxide.

2942: H. J. L. Trap and J. M. Stevels: Conventional and invert glasses containing titania, Part 2 (Phys. Chem. Glasses **1**, 181-188, 1960, No. 6).

Investigation of the variation in the dielectric properties of silicate glasses containing PbO as the various components are gradually replaced by $\mathrm{TiO_2}$. The observed effects can be satisfactorily explained on the basis of considerations on the behaviour of titanium ions in glass given in a previous publication (these abstracts, No. **2906**). The addition of the titanium ion usually improves the dielectric properties. The infrared spectra of the investigated glasses are also in agreement with the above-mentioned considerations. See also Philips tech. Rev. **22**, 300, 1960/61 (No. 9/10).

2943: E. Havinga, R. J. de Kock and M. P. Rappoldt: The photochemical interconversions of provitamin D, lumisterol, previtamin D and tachysterol (Tetrahedron **11**, 276-284, 1960, No. 4).

A summary of the results of recent investigations into the photochemical isomerizations in the vitamin D field which have been the subject of research during the last decade. A new scheme is proposed showing the various reactions occurring during irradiation of a provitamin D. The quantum yields of these reactions at 2537 Å were determined. On the basis of these data the effect of the wavelength of the light used on the yields of products is explained. Emission spectra of ergosterol and its photoisomers were measured at 80 °K. No phosphorescence was observed. Some aspects of the mechanism of the photochemical cyclizations, ring openings and the cis/trans isomerization are discussed.

2944: G. Hardeman: La polarisation dynamique nucléaire dans du polytétrafluoréthylène irradié (Bull. Group. Inf. mut. AMPERE **9**, No. spéc. (compte rendu 9e Coll., Pisa, 12-16 Sept. 1960), 669-673, 1960). (Nuclear dynamic polarization in irradiated polytetrafluorethylene; in French.)

See these abstracts No. **R 410.**

2945: S. van Houten: Semiconduction in $Li_xNi_{(1-x)}O$ (Phys. Chem. Solids **17**, 7-17, 1960, No. 1/2).

NiO is an insulator, which may be made conducting by the addition of lithium oxide. This behaviour can be explained in terms of an energy-level scheme consisting of full, localized Ni^{2+} levels with empty Ni^+ levels approximately 5 eV above them. The consequence of introducing lithium into the lattice is that the Li^+ ions are compensated by Ni^{3+} ions, giving $(Li^+.Ni^{2+})$ acceptor levels at approximately 0.03 eV above the Ni^{2+} levels. Electrical conduction, which is always *p*-type, may be described in terms of a thermally activated diffusion of holes from one nickel ion to another. The activation energy is connected with self-trapping by the polarization induced by the hopping hole itself. A detailed account is given of the calculation of the energy levels, starting from the ionization energies of the free ions and combining them with Madelung potentials. Corrections are made for the polarization of the lattice and for differences in crystal field stabilization between the Ni ions of different valencies. Measurements of Seebeck effect and electrical resistance as a function of temperature and lithium concentration are discussed in terms of this model. It is shown that the oxygen band, lying much lower than the Ni^{2+} levels, does not give any contribution to the electrical conduction.

2946: H. Koelmans: Association and dissociation of centres in luminescent ZnS-In (Phys. Chem. Solids **17**, 69-79, 1960, No. 1/2).

ZnS-In phosphors, which are efficient upon excitation with $\lambda = 3650$ Å, were prepared by firing in H_2S at 1200 °C and quick cooling to room temperature. The emission spectrum depends on the In concentration and consists of three bands at 6200 Å, 5350 Å and 4700 Å. X-ray analysis shows that the maximum amount of In incorporated is about 10^{-2} gram-atoms In per mole ZnS. Refiring the phosphors at 600 °C kills the luminescence at room temperature, an effect shown to be due to association; the associated centres act as killer centres. The low-temperature phosphorescence of the ZnS-In phosphors is strongly stimulated by irradiation into bands at $\lambda = 1.8\ \mu$ and $\lambda < 1.2\ \mu$.

2947: W. Nijenhuis: Benaderingsmethode van overdrachtsfuncties, waarbij een rimpel zowel in het doorlaatgebied als in het dempingsgebied wordt voorgeschreven (T. Ned. Radiogenootschap **25**, 297-306, 1960, No. 5/6). (Method of approximating frequency-response curves in which limits are set on the ripple in the pass-band as well as in the damping region; in Dutch.)

For telephony and other purposes, filters are needed with response curves in which the ripple in the pass-band and in the damping region, and the width of the transition region, are kept within (narrow) specified limits. As is known, the frequency-response curve of a filter composed of resistances, self-inductances and capacitances can always be expressed as the ratio of two rational functions. This publication describes a method, originally developed by Klinkhamer, for finding a function of this type which fulfils the above-mentioned conditions. Use is made of conformal mapping of the complex frequency plane by means of elliptical functions. The problem can be visualized by consideration of a membrane model of the response function.

2948: M. J. Koopmans: Systemic fungicidal action of some 5-amino,1-bis(dimethylamido)phosphoryl triazoles 1,2,4 (Meded. Landbouwhogesch. Opzoekingsstat. Gent **25**, 1221-1226, 1960, No. 3/4).

A discussion of the fungicidal action of the 3-pentyl and 3-phenyl derivatives of 5-amino,1-*bis*-(dimethylamido)phosphoryl 1,2,4-triazole. (The 3-phenyl derivative is on the market under the name "Wepsyn".) This group of compounds is very effective against the powdery mildew species which occur on plants such as apple trees, roses, and barley. With apple trees and barley, this effect is found on spraying of the leaves. The effect is comparable with that of 2,4-dinitro-6-capryl-phenyl crotonate ("Dinocap"). These compounds also protect the above-mentioned plants when administered to the roots. Their *in vitro* activity is much less than that of "Dinocap"; the resistance is thus apparently induced in the host plant, the effect being produced either via the leaves or via the roots.

2949: W. J. Oosterkamp and C. Albrecht: Methods of evaluating the new instrumental systems for diagnostic radiology (IXth Int. Congress of Radiology, Munich, 23-30 July 1959, Ed. B. Rajewsky, pp. 1451-1460, Thieme, Stuttgart 1960).

A short description of various new instrumental aids to X-ray diagnosis (image intensification, X-ray television and various combinations of these) is followed by a discussion of the effect of radiation contrast, noise and blurring on the perception of small details in the final picture.

2950: O. Reifenschweiler: Neutrons from small tubes, I. Philips tube: continuous or pulsed operation (Nucleonics **18**, No. 12, 69-71, 1960).

Description of a sealed neutron-generator tube, developed over the past years to meet the growing needs of neutron physics in science and engineering. The tube uses the D-T reaction to produce 3×10^8 n/sec during continuous operation. When adapted for pulsed operation, it yields a peak output of 3×10^{10} n/sec. The tube contains a Penning ion source, a 100-200-kV accelerating system, a replenisher that keeps the pressure within the tube constant, and a titanium-tritium target that is self-replenishing and, therefore, does not limit the lifetime of the tube. See also Philips tech. Rev. **23**, 325, 1961/62 (No. 11).

2951: F. L. H. M. Stumpers: Balth. van der Pol's work on nonlinear circuits (IRE Trans. on circuit theory **CT-7**, 366-367, 1960, No. 4).

A survey of the pioneering work of Van der Pol in the field of nonlinear networks (Van der Pol's equation, relaxation oscillations, etc.) in the special number on this subject which he was to edit, and which is now dedicated to his memory. (For a more complete survey of the work of Van der Pol see Philips tech. Rev. **22**, 36, 1960/61, No. 2.)

2952: W. van Gool and J. G. van Santen: Sintered cadmium sulphide, a photoconductive ceramic (Special ceramics, Proc. Symp. Brit. Ceramic Res. Ass., Ed. P. Popper, pp. 252-264, Heywood, London 1960).

A description of the preparation of photoconductive cadmium-sulphide powders and the making of compressed and sintered discs from these powders. The mechanism of photoconduction is discussed and it is shown that measurement of the properties of photoconductive ceramics could give additional information about the sintering process. The controlled preparation following the methods described makes possible a reproducible large-scale production of highly sensitive photoresistors. See also Philips tech. Rev. **20**, 277, 1958/59.

2953: M. P. Rappoldt and P. Westerhof: Investigations on sterols, XIX. 6-dehydro-9β,10α-progesterone from pregnenolone (Rec. Trav. chim. Pays-Bas **80**, 43-46, 1961, No. 1).

6-dehydro-9β,10α-progesterone is a very effective oral progestative. It was originally made from lumisterol$_2$. It is however also possible to start from pregnenolone (pregn-4-en-3β-ol-20-one), which is more readily available. Bromination of this compound followed by dehydrobromination gives pregna-5,7-dien-3β-ol-20-one, which yields the 9β,10α isomer when irradiated with ultraviolet light. Oppenauer oxidation followed by isomerization by means of HCl in isopropanol gives the desired end product.

Now available

H. Zijl: Large size perfect diffusors (Philips Technical Library, 1960, 196 pp., 120 figures, 49 graphs).

Second edition (with minor alterations) of the book which has already been reviewed in Philips tech. Rev. **15**, 262, 1953/54.