THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 49 November 1970 Number 9

Device Photolithography

Foreword

The fabrication of semiconductor and thin-film integrated circuits requires the delineation of precisely defined patterns in various materials in order to obtain the required functional performance of the device. Photolithographic processing has primarily been used for this purpose, requiring that masks be generated as the basic "tool" for producing integrated circuits. This issue is devoted to a detailed description of a new mask-making system intended to satisfy the Bell System's requirements for increasing numbers of increasingly complex masks. The system features high precision and large throughput made possible by a specially designed family of machines linked together by a computer-controlled information system.

The heart of the system is the primary pattern generator (PPG) which produces the original artwork by scanning a tv-like raster pattern on a photographic plate with a focused laser beam. The horizontal deflection of the beam is provided by reflecting it off a spinning polygonal mirror while the vertical motion of the plate is provided by a precision stepping table. The laser beam is modulated by an acousto-optic element under the control of a digital data stream which contains the topographic information. The machine is capable of generating a 22-cm by 18-cm pattern with an address structure of 32,000 by 26,000 units in about 10 minutes. It provides a reproducibility of one part in 25,000 and an absolute accuracy greater than one part in 10,000. The reduction cameras and step-and-repeat camera that complete the system were designed to fully exploit the high-speed and accuracy capabilities of the PPG. Looking ahead to future device applications in which the higher resolution offered by an electron beam generator could be of importance, development work on such a unit is also described.

The articles in this issue discuss: (*i*) the overall system, including the engineering considerations that led to the choice of pattern generation; (*ii*) the computer programs required to transform topographic information into a digital data stream suitable for control of either the PPG or the electron beam pattern generator; (*iii*) the PPG, including the optical, mechanical and electrical design features; (*iv*) the electron beam pattern generator; (*v*) electron-sensitive materials for use with the electron beam machine; (*vi*) the design and characteristics of the lenses used in the mask-making system; (*vii*) the optical and mechanical design of the reduction cameras; (*viii*) the optical and mechanical design of the computer-controlled step-and-repeat camera; (*ix*) the thin photosensitive materials required for use in the above cameras; (*x*) the specially designed coordinate-measuring machine used to inspect masks and to maintain the mask-making system; and (*xi*) the information system which controls the flow of work through the mask laboratory.

Many people, too numerous to mention, throughout Bell Laboratories and Western Electric Company, have made significant contributions to the development of this mask-making system. Their efforts have led to the successful installation and operation of two mask laboratories, one at Murray Hill, New Jersey, and one at Allentown, Pennsylvania.

FRANKLIN H. BLECHER

Device Photolithography:

An Overview of the New Mask-Making System

By F. L. HOWLAND and K. M. POOLE

(Manuscript received July 9, 1970)

This paper reviews how photolithographic masks for silicon and thin-film integrated circuits are made. Increasing production and complexity of masks makes heavy demands on the operating time, reproducibility, and accuracy of the new mask-making system. The pattern generation step, in which the design is converted to a photographic image, is critical to the system. Advantages and disadvantages of other pattern-producing methods are discussed. The technique of producing patterns by optically scanning lines with a rotating mirror while mechanically stepping the photographic plate is described. This article develops the basic design parameters of address structure and operational speed for the primary pattern generator, and it defines the requirements for reduction cameras and the step-and-repeat camera for a system capable of meeting the needs for both thin-film and silicon integrated circuits. The article notes the system limitations imposed by optical generation of patterns and lens tolerances.

I. INTRODUCTION

The Electronic Materials and Components Development Area of Bell Telephone Laboratories has made the development of hybrid-integrated electronics, combining semiconductor and thin-film technologies, its major general field of activity for several years. Silicon integrated circuits provide the active elements for both digital and analog systems, and passive components can be incorporated if tolerances are not too tight. Thin-film circuits based on tantalum can provide stable resistors and capacitors which can be trimmed to precise values, while other thin-metal films can be used advantageously for conductors. Thus silicon and thin-film technologies together provide a sufficient set of elementary components for most systems functions. Equally important, the choice of silicon circuits made in the beamleaded, sealed-junction form and thin-film elements on ceramic or similar substrates give us complementary technologies which are physically compatible.

Both parts of this hybrid-device technology have come to depend primarily on photolithographic methods for delineating the areas in which material will be added, removed, or modified as the original substrate is successively transformed into the final circuit. Both parts of this technology have grown in volume of activity and in sophistication of technique. In doing so they have put increasing demands on mask-making laboratories for more masks per year and for more complex mask patterns.

The system described in this issue of *The Bell System Technical Journal* provides for both semiconductor and thin-film integrated circuits using facilities that are coupled by an information system. The mask-making system is designed to have the capability of meeting the demands for larger numbers of increasingly complex masks with a known time interval between the receipt of design information and the delivery of a complete set of masks.

II. HISTORICAL BACKGROUND

All mask-making systems can be described schematically as shown in Fig. 1. Two streams of information, one topographic and the second descriptive, must be provided.

The topographic stream starts with the designer who generates the input information on the topography for each mask level and stores the information using a program such as XYMASK. The information thus generated is not suitable for direct use in making artwork, so a post-processor is used to modify data and make it compatible with a specific artwork-generating system. After the processing and, if necessary, recycling to eliminate errors, the output data can be used to drive the artwork-generating equipment.

After the artwork is generated, a series of photo-reductions are performed and, if required, an array of images is produced using a step-and-repeat camera to produce the master photo mask. From this master, working copies are generated, the specific process depending on the ultimate need. Working copies can be emulsion or chrome on glass for semiconductor circuits, or emulsion on glass or transparent plastic for thin-film applications.

In parallel with the topographic information, descriptive information is also required. The descriptive information includes the tone of

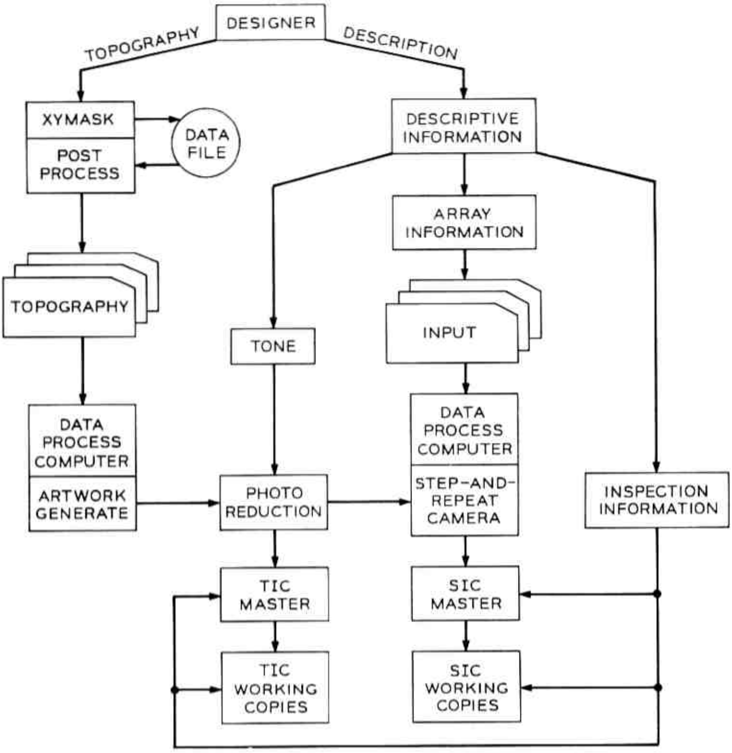


Fig. 1—Schematic of the mask-making process.

the mask; that is, are there clear features on an opaque background or are there opaque features on a clear background? The tone is established by the specific process to be used for delineating the pattern in the final product. For masks requiring the step-and-repeat operation to generate the array, information concerning the specific pattern of images must be defined and the necessary data generated for producing the array. Finally, the descriptive information must include drawing numbers, tolerances, and critical features to be used as inspection points; this information relates to the final inspection of the master and working copies. The descriptive information is as critical in mask making as the topographic information. Because of the combined topographic and descriptive information paths and the complex of processes, management of a mask-making laboratory is a very important part of the system.

As device complexity has increased, with a consequent increase in the amount of data required to describe the topography of an image, computer-controlled artwork generators have been developed. Two distinct types of artwork-generating equipment have evolved. The first are mechanical systems, such as a coordinatograph, the Gerber,* or a mechanical reticle generator which operates by moving a generating head on a mechanical *XY* stage or moving the recording medium past a fixed optical head. The second type uses an electron beam and camera to generate the artwork.

The mechanical systems which generate the artwork feature-by-feature have a potential address structure that is not fully utilizable because of errors in the mechanical systems. In general, however, they can be operated reproducibly with 6000 addresses in the *X* and *Y* directions. Because of the nature of the mechanical motion, the time required to produce a given piece of artwork is sensitive to both the complexity and the size of the feature.

An example of the use of an electron beam and camera system is the SC 4020.† This system is capable of generating a pattern at electronic speeds by moving an electron beam over a cathode ray tube and photographing the image. It produces a mask rapidly but the address structure is limited and, as a consequence, it can only be used for low-precision artwork generation.

After the artwork is generated it is, in general, reduced in size. Typical reduction cameras for both silicon and thin-film circuitry produce images that are reduced by a factor of from 10 to 30 from the original artwork. These cameras are all physically large and require high-quality lenses to minimize distortion. At this step the master mask for thin-film applications is produced. Working copies for device processing are generated by contact printing.

For silicon integrated circuits the image produced by the reduction camera is typically ten times the final size. The final reduction and the fabrication of the circuit array is done on a step-and-repeat camera. Because of the complexity of the array, in terms of the variety of images to be produced, the cameras are computer controlled. For a typical mask the primary interest is, of course, the formation of an array of precisely placed images of the primary pattern that is required for the fabrication of the working device. In addition, however, special patterns such as test patterns for checking processing and alignment features are also

* Gerber Scientific Instruments Company, South Windsor, Connecticut.

† Stromberg Data Graphics, San Diego; California.

required. Since a typical semiconductor integrated circuit requires from nine to twelve mask levels to complete the device fabrication, the step-and-repeat camera must provide not only for the final optical reduction but also for the precisely controlled and reproducible positioning of the images so that registration from one mask to another in the set is achieved. In the past step-and-repeat cameras could place an image with a reproducibility of ±1.5 μm. However, the errors in the mechanical drive and position-sensing systems made absolute positioning considerably less accurate.

III. MASK-MAKING PRECISION, STANDARDS, AND CAPACITY

With this background of the mask-making process and the then-available equipment to produce the mask, the changing complexity, as measured by the number of coordinates required to describe the image of the masks for both silicon and thin-film circuits, has had a major impact on the capability of mask-making systems to meet the demands. Projection of our future needs for integrated-circuit masks suggested that we will have to provide for: (*i*) a minimum feature size five thousand times smaller in linear dimension than the over-all size of the circuit pattern; (*ii*) incremental sizes of about one-fifth of this minimum feature size; (*iii*) reproducibility of about one part in 25,000; and (*iv*) absolute accuracy of about one part in 10,000 (both reproducibility and accuracy being referred to the over-all size of the pattern). Examination of the state of the art of lens design suggested that cameras could be built to be consistent with these needs, provided that we adopted a set of standard mask formats and that we designed lenses and cameras for each standard field size and reduction ratio.[1]

Such a set of standards has been chosen (Table I). They provide for large thin-film circuits with a nominal field size of 12.5 cm and a smaller format, 5 cm, which both provides for medium-sized thin-

TABLE I—STANDARD MASK SIZES

Principal Function	Field Size	Minimum Line Width	Address Size
Thin-film circuits	12.5 cm 5.0 cm 2.5 cm	25 μm 10 μm 5 μm	5 μm 2 μm 1 μm
Semiconductor circuits	5.0 mm	1 μm	0.2 μm

film circuits and serves as an intermediate step in semiconductor-mask fabrication. A third standard may become necessary for small, fine-lined, thin-film masks and appropriate values are listed in Table I. Semiconductor integrated circuits seem likely to remain under 5 mm square, and a single standard field for a step-and-repeat camera is sufficient. This set of standards embodies (*i*) a decision to "go metric" in device design, (*ii*) a compromise between design flexibility and the capital cost of equipment, and (*iii*) a preference that the address units, which quantize internal device dimensions, be such that large integral multiples be immediately identifiable.

In the same period of time in which the growth in the complexity of mask patterns has occurred there has been a parallel increase in the demand for numbers of masks. This growth has been the direct result of a need for larger numbers of masks to fabricate a given device coupled with an increase in the number of designs. To illustrate this growth of demand, information has been collected from a variety of Bell Laboratories groups covering the period from 1966 to the present and estimating the needs for the early 1970s. The results are shown in Fig. 2.

The growth in demand for silicon integrated circuits, SIC, from 1966 through 1969 has been nearly exponential and has been in part inhibited by our inability to produce sufficient quantities of masks. Because of the increased numbers of people designing integrated circuits, the growth will continue to be slightly greater than linear during the early 1970s. Thus, somewhere between 7,500 and 8,000 pieces of artwork per year will be required by 1972 or 1973.

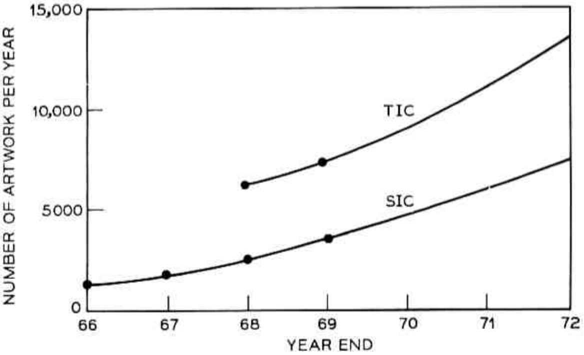


Fig. 2—Growth in demand for artwork for silicon and thin-film integrated circuits.

Because the silicon integrated circuit and thin-film circuits are intimately connected in design, it can be expected that the need for thin-film masks, TIC, will also rise during the early 1970s as shown in Fig. 2. In part, this growth represents the need for increasing numbers of masks for crossovers and tantalum circuits that are combinations of resistors, capacitors, and crossovers.

If we take the composite of these two trends, we find that development activities will require that approximately 14,000 pieces of artwork be generated per year by 1972. To meet this demand, it was decided to build two mask-making laboratories, one at the Murray Hill, New Jersey, location and one at the Allentown, Pennsylvania, location. Each laboratory was to have a master mask capacity of 10,000 per year.

IV. CHOICE OF PATTERN GENERATOR

Pattern generation is a key element in the total process of mask-making in the sense that the difficulty of meeting the many demands placed on this step is so great that the adjacent steps of the process must largely be tailored to the choice of pattern generator. The overall process resulting from each plausible choice of pattern generator design must then be evaluated before a final system choice is made.

The nature of the problem logically requires relative motion in two dimensions between a writing element and a recording medium. The functional requirements which have been discussed in the previous section suggest a digitally controlled plotter having resolution corresponding to 25,000 by 25,000 address points in the pattern field and a plotting time for the more complex patterns of about 10 minutes.

Reviewing the pattern generators which have previously been used, we first have machines such as automatic coordinatographs and automatic drafting machines with optical exposure heads. A machine of this type could be designed to give the desired resolution. The plotting time for complex patterns on such machines has already exceeded ten hours. Another approach is the reticle generator which makes a set of elementary figures available from which every mask will be assembled. We have not found any set of figures which offer sufficient speed and flexibility.

The following three approaches to pattern generation appear to have sufficient resolution, accuracy and speed to meet our requirements: drum recording, electron-beam recording and light deflection. Each is discussed in turn.

4.1 *Drum Recording*

In the drum recorder the recording medium is wrapped around a cylinder as shown in Fig. 3. The two dimensions of motion are now achieved by synchronizing the rotation of the drum and translation of either the drum or writing head parallel to the axis of rotation. If we insist on a system capable of writing on various areas of the recording medium in an arbitrary sequence (random access), this system offers no advantage over a flat-bed plotter; however, it does make it possible to create any pattern by continuous rotation of the drum and a synchronized translation. After unwrapping the recording medium, the image would appear as though it had been created by a TV-like raster. It is this concept of a uniformly swept raster which makes a mechanically scanned system feasible.

This pattern generator could be engineered within a relatively wide range of sizes, tolerances on the precision of the translational mechanism, on the concentricity of the drum, and on the thickness of the recording medium becoming increasingly tight in smaller machine sizes. A 12.5-cm pattern size would be possible, while a 25-cm size unit would be relatively simple to develop. The primary problem in

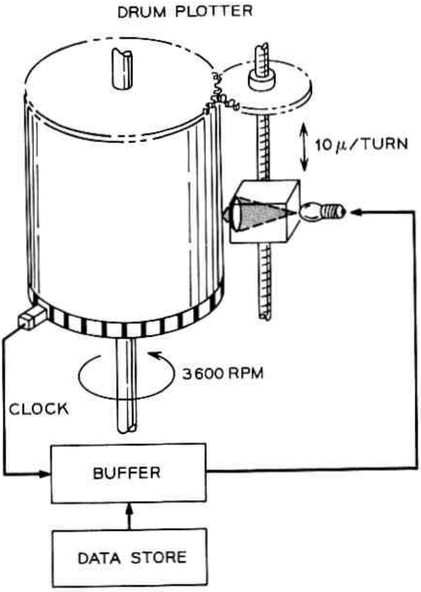


Fig. 3—Schematic of a drum plotter.

this approach is that the recording medium must be flexible. The combination of a silver halide emulsion on a film base does not have sufficient dimensional stability for our purposes. An alternative which was considered was laser machining some appropriate coating from a metal based multi-layer medium. Brief experiments suggested that such a medium would not be easy to handle and, being opaque, would have to be used in front-lighted reduction cameras. Such cameras are inefficient and the drum approach was dropped from further consideration.

4.2 *Electron Beam Recording*

An electron beam machine in which a finely focussed beam writes directly on a recording medium of appropriate resolution and sensitivity is a probable approach to pattern generation. An electron beam recorder can be designed for a beam size of a few microns and a field of several centimeters.[2] Choice of a 5-cm field allows direct generation of one standard format and allows the other standard sizes to be produced in cameras using glass condenser illumination. Pattern description for this system is a simple extension of previous work for cathode ray tube systems. This technique seems to offer system compatibility; the major uncertainties which existed at the time at which a selection had to be made (November 1967) were whether the desired accuracy could be obtained, and whether the sensitivity of electron beam systems to unwanted electric and magnetic fields would limit its reproducibility. These uncertainties were sufficiently great that this approach was not chosen for our initial system, but development work was continued to provide a compatible system which might be advantageous for future large-area devices such as color and document-mode *Picturephone®* camera tubes and magnetic domain devices. This machine is described in a companion paper.[2]

4.3 *Light Deflection*

Of the three approaches, only deflection of a light beam seemed capable of meeting our anticipated requirements. Since the combination of plotting time and number of resolvable elementary areas in the pattern field requires exposure times of less than one microsecond per resolvable area, the use of a laser beam to achieve a small, very bright writing spot was indicated. Deflection of a laser beam can be accomplished by electro-optic or acousto-optic elements, but available deflector materials were not of sufficient quality to give plotting times less than one or two hours. Reflection from a spinning mirror, however, can give speeds up to and beyond those required as long as we accept a uniformly rotating mirror as the basis for our system. This led to a rotating-mirror pattern generator design where a modulated light beam would be swept across a photographic plate in one direction at a rate of about 50 scans per second, while the plate holder would move in the direction perpendicular to the scan lines. In less than ten minutes 25,000 overlapping scan lines could build up the complete pattern image. Again, in this system, we have employed continuous rotation of the higher-speed scanning member to achieve the desired plotting rate in a mechanical system. Implementing this approach requires that a lens be mounted adjacent to the rotating mirror, a diverging input beam being collimated by the lens and refocussed onto the recording medium after reflection. Because of the inverse relationship between the aperture of a lens and the diameter of the smallest spot which the lens can image and because the field angle for which a lens can be designed is sensitive to the relative aperture size, the lens and mirror sizes enlarge rapidly as the desired pattern size is diminished.[3] Specifically, the design appears impracticable at the largest standard pattern size of Table I and relatively easy at a 25-cm pattern size. Thus, the initial pattern size for this machine design is rather firmly bounded by optical-design considerations on the one hand and by considerations of plate size, governing the size of both processing equipment and reduction cameras, on the other. 8 by 10 inch photographic plates are commercially available and, in ¼ inch thickness, can be obtained with sufficient flatness. Translating to metric units gives a maximum usable area of about 13.8 cm by 23.4 cm. This puts an upper bound of 7.3 μm on the address unit size, and 7.0 μm seems a reasonable value. A review of the optical design based on this value led to reasonable sizes for the individual components and for the over-all machine.

Pattern description for the primary pattern generator (PPG) requires that the topographical data be sorted into a sequence controlled by the directions of scan, and presented to the generator at a predetermined rate. These are novel requirements relative to our experience in computer aids to mask-making.[4] While the sorting operation requires large files in the off-line data-processing system, the operation is not a costly one. A larger problem is created by the need to present data to the generator from its on-line controlled computer at a predetermined rate of about 2 million bits per second. The strategy used to meet this demand is such that most of the core memory is required for storage of coded data describing the current scan line and the changes required to go from the current line to those immediately following, and thus all characteristics of features, particularly where they include slant and curved edges, have to be computed off-line and coded for transfer by means of a magnetic tape. At this time this is a significant disadvantage in the choice of the PPG as opposed to a random-access generator such as the electron beam machine.

The characteristics of the PPG previously discussed determine the design requirements which it must meet. With reference to Table I, it is evident that for thin-film circuits optical reduction of the image plate from the PPG is required. A reduction camera that reduces the image 1.4 times is required for the bulk of the thin-film circuits that have a minimum line width of 25 μm. A second camera with a 3.5 reduction ratio is also required for 10 micrometer minimum lines on a smaller field. This camera is also used for silicon integrated circuits. A third reduction camera for 5-μm lines may be required in the future if 5-mm lines are required on small areas. Conventional glass condenser systems are not practical for these cameras, and large area diffuse sources with Fresnel lens condenser systems are used to meet our requirements.[5] The cameras have been designed with no operator adjustments for either reduction ratio or focus.

For silicon integrated circuits the image produced by the 3.5× reduction camera is used as the reticle in the step-and-repeat camera which provides an additional 10-times reduction.[6] The step-and-repeat camera, in addition, generates an array of images—each with a 5-mm maximum field size and a maximum array size of 10 cm by 10 cm.

V. SYSTEM DESIGN

In completing our account of the new mask-making system, we should recognize that not all devices are square. Many thin-film integrated circuits are rectangular. As long as a camera is to be used to image a rectangular pattern, the diagonal measure of the pattern is a dominant consideration. It is not necessary, however, to compound this penalty by fitting a square pattern field within the circular field of the cameras and then constraining a rectangular pattern to lie within the square. Thus the field of the pattern generator was enlarged from 25,000 address units square to 32,000 units (22.4 cm) and, at the same time we enlarged the width to 26,000 units (18.2 cm) since the space was available.

Fiducial marks which provide for registration of patterns in the step-and-repeat camera are plotted in the corners of the 32,000- by 26,000-unit rectangle. In addition, the pattern generator writes two strips of system data, one above and one below the rectangle. The first strip shows the identification number of the particular pattern generator used and the sequence number in octal form. The second strip contains the drawing number of the pattern in three forms. One is the normal form for the direct use of mask shop operators, but in addition the number is repeated in two binary-coded formats suitable for machine reading. One is designed to be read when the pattern generator plate is in the reduction camera and the other to be imaged by the 5-cm field-reduction camera and read when the resulting reticle is in the step-and-repeat camera.

These provisions for machine reading of the drawing number are part of a supervisory and scheduling system known as the Mask Shop Information System (MSIS).[7] Earlier experience with mask-making laboratories of more modest capacity than our 10,000 per year objective taught us that the scheduling system can be the factor determining the time to complete a job. The equipment design which has been outlined here and which will be detailed in the following papers can therefore shorten the time to complete a job only if we add a system for storage and rapid retrieval of all the data required to make and inspect the masks and keep the necessary records. Scheduling each phase of each job is included; as each step after pattern generation is due, the MSIS displays to the camera operator the drawing number of the pattern generator plate or reticle and the location of that plate in the physical storage trays provided. The system then reads the plate number and advises the operator if an error has been made. At the step-and-repeat stage, all data describing the step-and-repeat array is fed to the on-line control computer.[6]

VI. SYSTEM APPRAISAL

While we have not yet had sufficient experience with MSIS, nor with a level of demand for masks which would have fully exercised MSIS, we can make a preliminary appraisal of the remainder of the system.

The PPG has accomplished essentially everything we set out to do. For the first time in many years, artwork generation is no longer the pacing item in mask making; we have a machine which takes simple patterns or patterns of a complexity we would not previously have attempted, makes patterns in which 10 percent of the area is exposed or patterns in which 90 percent is exposed, semiconductor device patterns, thin-film patterns, test patterns—and even digitized photographs—and turns them out with inhuman regularity. While the optical-design pattern bound us into a very narrow size range, the resulting machine is the right size for the operator's convenience. This is not to say that there is no room for further improvement in the area of artwork generation. We see future device applications in which the higher resolution offered by an electron beam machine could be of major importance, sufficient to justify incorporating such a unit—compatible with the PPG system standards in format and plate size—into the mask-making laboratories.

Turning to the reduction cameras, we feel that the basic system decisions which were made—separate fixed cameras using Fresnel condenser illumination with monochromatic light—were sound. We do believe that further improvements in system performance might be obtained through achieving closer tolerances in lens fabrication; essentially the state of the art of lens design has run ahead of lens assembly techniques. This comment applies even more strongly to lenses, such as the one for the step-and-repeat camera, which are aimed at feature sizes of a few wavelengths of light. The step-and-repeat camera lens proved extremely difficult to build, and appears to have distortion of about one part in 5,000 arising from fabrication tolerances; we would argue that paper designs of lenses of higher performance—perhaps seeking comparable resolutions over a larger field —should be held suspect until actual models are built and tested.

The new step-and-repeat camera is a development of a different kind from most of the other parts of this program. No single characteristic of this unit shows an order of magnitude improvement over earlier equipment, nor does it contain conceptually new major elements. The improvements which have been made, factors of two or three in smallest feature width, in linear field dimensions, in linear array dimensions, and in speed, are cumulative in their impact and are essential to the satisfaction of our anticipated needs.

REFERENCES

1. Herriot, D. R., "Lenses for the Photolithographic System," B.S.T.J., this issue, pp. 2105–2116.
2. Samaroo, W., Raamot, J., Parry, P., and Robertson, G., "The Electron Beam Pattern Generator," B.S.T.J., this issue, pp. 2077–2094.
3. Cowan, M. J., Herriott, D. R., Johnson, A. M., and Zacharias, A., "The Primary Pattern Generator, Part I—Optical Design," B.S.T.J., this issue, pp. 2033–2041.
4. Gross, A. G., Raamot, J., and Watkins, Mrs. S. B., "Computer Systems for Pattern Generator Control," B.S.T.J., this issue, pp. 2011–2029.
5. Poulsen, M. E., and Stafford, J. W., "Reduction Cameras: Mechanical Design of the 3.5× and 1.4× Reduction Cameras," B.S.T.J., this issue, pp. 2129–2143.
6. Alles, D. S., et al., "The Step-and-Repeat Camera," B.S.T.J., this issue, pp. 2145–2178.
7. Brinsfield, Mrs. J. G., and Pardee, S., "The Mask Shop Information System," B.S.T.J., this issue, pp. 2203–2220.

Computer Systems for Pattern Generator Control

By A. G. GROSS, J. RAAMOT and MRS. S. B. WATKINS

(Manuscript received May 29, 1970)

Computer systems play a fundamental role in the operation of precision integrated-circuit pattern generators. This paper first describes the XYMASK *system which provides a language for describing the geometric shapes in a set of masks and generates graphical artwork on a number of different pattern generators. The remainder of the paper is devoted to discussions of system-design considerations and algorithms for generating input to the primary pattern generator and the electron beam machine.*

I. INTRODUCTION

Computers are indispensable today in the operation of any sizable mask-making laboratory. Nearly all precision pattern generators are either directly computer controlled or else require input of a form which can be reasonably obtained only through the use of computers. Furthermore, the complexity and sheer volume of masks currently required effectively prohibit nonautomated procedures.

The mask-making laboratory system described in this issue relies heavily on the use of computers. The first part of this paper describes a system of programs which links a circuit designer to the mask-fabrication processes; the next two sections discuss algorithms and programs for generating input to the primary pattern generator (PPG) and the electron beam machine (EBM).

1.1 *Computer-Aided Generation of IC Masks*

Masks are tools required in the fabrication of integrated circuits and other devices. The starting point in mask design is thus an electrical schematic or logic diagram of the desired device. An engineer or technician first allocates scaled geometric shapes to each of the circuit components; he then arranges and rearranges these shapes on a similarly scaled substrate area. During this placement phase, many criteria are generally involved in evaluating the suitability or desirability of one arrangement over another. Some examples are thermal interaction, packing density, and the ability to realize the required component interconnections. The latter criterion is really applied in the next phase wherein the interconnection pattern is designed in detail. For most cases, several iterations between the placement- and interconnection-design phases are required before a satisfactory layout is obtained. At this point the geometric details of all the required masks are completely known; the next step in the process is mask generation.

The draftsman or engineer is now faced with the problem of transforming the mask layouts into a form suitable for driving a pattern generator. The severity of this problem depends on two factors: the form of input required by the particular pattern generator, and the complexity of the masks. For pattern generators which are concerned solely with the outline of the geometric features, such as an automatic knife coordinatograph cutting rubylith, the solution is tedious but straightforward. Either manually or via a digitizer, the coordinates of the endpoints for each horizontal feature boundary line, followed by the coordinates of each vertical feature boundary line, can be recorded on punched paper tape for each mask level. This tape would then be processed by the coordinatograph, the rubylith master peeled, and the masks obtained after appropriate photographic processing of the rubylith master. However, for more sophisticated pattern generators which operate by filling in the interior of mask features with beams of light or electrons on photographic film, substantial use of computers is necessary to convert the mask geometry into commands acceptable by the pattern generators.

1.2 *The* XYMASK *System*

The system of programs in use at Bell Telephone Laboratories and Western Electric Company for computer-aided production of integrated-circuit masks is known as XYMASK. First operational in late 1967 and subsequently improved and modified, the current version of XYMASK evolved from two earlier generations of mask-making programs. Three of the more important system-design goals may be stated as follows.

(*i*) It should provide a standard user-input language for conveniently and efficiently describing mask-feature geometry.

(*ii*) Insofar as possible, the system should be independent of any particular graphical output device.

(*iii*) The implementation should be highly independent of the host computer to enhance portability of both the system and the mask specifications.

The first of these goals is extremely important. Its realization greatly facilitates the transmittal of device designs not only among Bell Laboratories locations but also between Bell Laboratories and Western Electric Company for production. Moreover, the user-input language is a vital factor in the interface between the mask designer and the system since its convenience and flexibility have a direct bearing on user acceptance and satisfaction.

The second goal is a necessity due to the diversity and number of graphical output devices available at Bell Laboratories locations. In an indirect manner, attainment of this goal also simplifies the addition of new output-device capability as we shall see below.

The third goal arises from the use of different large-scale computers at Bell Laboratories and Western Electric locations and the ever-present possibility of new ones being acquired. The most important user benefit is the complete independence from any particular computer of the mask descriptions encoded in machine-readable form in the input language; the same mask-description input deck will produce identical artwork on different computers. Again indirectly, attainment of this goal has simplified program implementation and maintenance. The implementation is almost exclusively in a subset of FORTRAN IV common to the IBM 360 and GE-635 computers; there is essentially one set of source-language programs which runs on the several different computers.

1.3 *The User-Input Language*

As a preliminary to discussing the system organization of XYMASK, it will be helpful to describe briefly the user-input language. A somewhat more detailed description is given by B. R. Fowler[1]. Basically, the input language provides a vehicle for describing the various geometrical shapes contained in a mask or set of masks in a computer-readable form. As such, the most primitive statements in the language are used to specify the interiors of three basic geometrical shapes: rectangles, polygons, and paths. In this context, rectangles are defined to have their edges parallel to the coordinate axes and are specified by giving the coordinates of the vertices on either diagonal. The statement

```
label   RECT   mask, 10, 20, 30, 40
```

illustrates the format of the primitive statements and defines a rectangle with the lower-left vertex at $X = 10$, $Y = 20$, and upper-right vertex at $X = 30$, $Y = 40$. The label and mask attributes are discussed below. The polygon primitive is used to define generalized polygons having either straight lines or circular arcs as edges. The shape and size are fixed by giving the coordinates of the vertices in the order in which they are encountered in either a clockwise or counterclockwise tour of the periphery. The path primitive is used to specify a path of given finite width. The size and shape are fixed by giving the width and the coordinates of the endpoints and breakpoints of the centerline as they are encountered in a tour along the path. The centerline may contain circular arcs as well as straight-line segments.

The preceding paragraph discussed only the specification of the shapes and sizes of the basic geometrical features. The positions of these features on the masks may be specified in either of two ways. If a label attribute is not specified for the feature, the coordinate values define its position as well as its shape and size. On the other hand, if a label attribute is given, separate input-language statements must be used to specify the position. In addition to position, these statements also permit the orientation of the feature to be altered by reflection about either coordinate axis together with a rotation through an arbitrary angle.

In general, a set of individual but inter-related masks is required in the fabrication sequence for an integrated-circuit device. A transistor, for example, may require geometrical features on a number of different masks for forming collector, base, and emitter regions. The XYMASK user-input language allows specification of all geometrical features occurring in all required mask levels for a device in whatever intermixed order is most convenient for the user. In order to correlate the various features with the appropriate mask levels, a mask-level identification is required as part of the specification of the rectangle, polygon and path primitives.

It is often desirable and useful to treat a group of geometrical shapes as a structural entity; for example, it is far more convenient to position a transistor at the required locations as a structural entity rather than as a set of individual primitive shapes. The user-input language allows this hierarchical nesting of structures to an arbitrary depth. In other words, it is possible to define a structure which contains structures of lower "order" as well as basic geometric shapes. The structure may be positioned on the masks, possibly with reorientation, as described above for simple geometric shapes. This hierarchical structuring in conjunction with reorientation allows the user to take advantage of repetitions and symmetries in the design in order to reduce the number of statements and effort required to encode the design in the user language.

Statements are also available in the input language to retrieve previously designed structures from XYMASK libraries and to invoke component structure-design routines. Transistor designs are typical library entries. An integrated-circuit designer generally uses transistor designs which have been thoroughly tested and characterized. These designs are stored as library entries which contain the XYMASK language specification in the form of hierarchical groupings of the appropriate primitives. Library retrieval provides a sort of shorthand for the user in that only the particular library and the entry identification need be specified in the input deck in contrast to the equivalent set of XYMASK input statements.

Computer programs have been developed to design certain components and structures used in integrated circuits. Pattern generation for thin-film meander resistors, and the generation of sheafs* of interconnection paths are examples of such programs in current use.

Versions of these programs, called design routines, have been integrated into the XYMASK system. A single statement in the input language allows the user to specify the desired routine together with whatever parameters are required. Output from the routine consists of XYMASK statements specifying the generated design. These statements are automatically incorporated into the user's input.

The final feature of the user language to be discussed deals with the specification of particular graphical devices and output options. Graphical output may be requested either in the form of outline drawings or finished artwork. The outline drawings are generally produced on line plotters and are used to verify that the mask descriptions as encoded in the input language are correct. As implied, only the outlines of the geometrical features are displayed. The finished artwork is the desired end product of the system; for plotters working on photographic film, the interiors of the geometrical features have one tonality (clear or opaque) while the area which is exterior to all figures has the opposite tonality.

A single statement is used to indicate the plotter and any pertinent parameters such as drawing type and scale factor. The user has the

* A sheaf is a family of paths each member of which can be derived from a generic member by translating each of its path segments normally through a given distance and lengthening or shortening it as required to create a nested copy of the generic path.

capability of requesting individual drawings or artwork for any or all masks. He may also request composite drawings of any two or more masks. This latter feature is widely used for error checking.

1.4 XYMASK *System Organization*

A simplified diagram of the XYMASK system is shown in Fig. 1. The major program segments are the input preprocessor, the input processor, the execute processor, and the family of device-dependent output postprocessors. Input to the system is a machine-readable description of the desired masks encoded in the XYMASK user language. This input is free format and may be generated by hand encoding and keypunching, digitizing large-scale layouts, or by other computer programs such as interconnection-routing routines.

The input first passes through the input preprocessor. All input statements other than design-routine invocations or library retrievals are transmitted to the expanded input file without change. When a design-routine invocation is found, control is passed to that design routine, and the generated XYMASK statements together with the invocation are transmitted to the expanded input file. Library retrievals

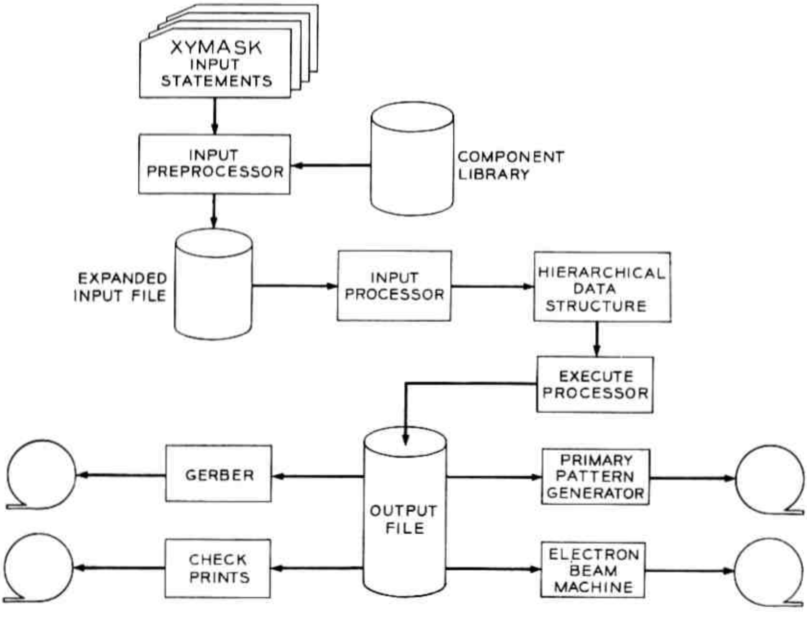


Fig. 1—The XYMASK system.

are treated similarly in that retrieval is made when the statement is encountered in the input deck; the retrieved XYMASK statements along with the retrieval statement are transmitted to the expanded input file. At the conclusion of the preprocessor phase, then, the expanded input file contains the original input statements interpolated with the results of any design-routine invocations or library retrievals.

The system design of the remainder of the XYMASK system was heavily influenced by the desired relative independence from any particular graphical output device. Accordingly, output-device dependence is relegated to a family of postprocessors each of which receives input from a common file referred to as the 'output file'.

This output file contains a representation of each of the masks requested in the XYMASK input deck in a form such that all device-independent processing has already occurred. Each mask is represented by a separate subfile, and each subfile contains only the defining coordinates of individual paths and polygons in their final positions and orientations.

The input and execute processors must then transform the expanded XYMASK input statements into the form required for the output file. The most significant aspects of this transformation are as follows: removal of all hierarchy by generating new copies of the various primitives as required while simultaneously carrying out specified reorientation and positioning; and sorting the resulting primitives into separate sets according to their individual mask-level identifications.

The above aspects of the transformation suggest that detailed descriptions of all required masks be available in memory in a convenient form prior to starting the transformation. Thus the input processor reads the input-language descriptions of the masks, makes extensive error checks, and stores the descriptions in a hierarchical data structure. Upon completion of this process, the execute processor comes into play to generate the output file from the data structure.

When output-file generation is complete, the appropriate postprocessor for the first mask is activated according to the output device specified by the user. Upon completion, processing is initiated on the second, perhaps using a different postprocessor if the user so desired. In like fashion, the remainder of the output file is processed and the job terminates.

Each postprocessor is responsible for the ultimate generation of artwork on a particular graphic-output device. In general, the postprocessor output is a magnetic tape which drives the actual device, although on-line devices, such as the STARE[2] line-drawing plotter, are easily accommodated. We can again view a postprocessor as a data transformer; it is responsible for reading each path and polygon specification from the output file and generating the proper output-device commands or codes for plotting that figure. The system design is such that all postprocessors are essentially independent programs which receive all of their input from the XYMASK output file. The system is thus open ended in the sense that new postprocessors can be easily and conveniently added.

With regard to execution times for a typical set of masks, the input and execute processors each require on the order of one-minute running time on an IBM 360/65. Postprocessor execution times are generally longer and tend to dominate other costs for the run.

The following two sections of the paper are devoted to detailed discussions of specific postprocessors for the PPG and EBM plotters described elsewhere in this issue. The two differ fundamentally in the manner in which pictures are produced. The EBM is a random-access plotter; the order in which mask features are plotted is immaterial. The PPG, on the other hand, produces pictures using a raster-scan technique. The contributions of all features intersected by each scan line must be determined and transmitted to the device in the order needed to generate the picture.

The PPG postprocessor was developed at Bell Laboratories, Murray Hill, New Jersey, by A. G. Gross. The EBM postprocessor was developed at the Western Electric Engineering Research Center, Princeton, New Jersey, by Mrs. S. B. Watkins and J. Raamot.

II. THE PPG POSTPROCESSOR

The operation and functioning of the PPG together with its control computer are discussed in this issue by A. Zacharias, et al.[3] For convenience, we will briefly review here those aspects which are of importance to the postprocessor.

For our purposes, we can consider the photographic plate plotting surface to be a rectangular lattice of 26,000 × 32,000 addressable points. A writing beam scans the lattice on a line-by-line basis, with the beam turned on at those address points interior to mask features, and off otherwise. The writing beam is controlled by a 26,000-bit display buffer in the control computer with each bit position representing one address along the scan line; the beam is turned on or off at an address according to whether the content of the corresponding bit position is one or zero. After completing a scan line, the bit configuration in the display buffer must in general be modified to correctly represent the geometric detail in the next scan line. When updating is completed, the bit configuration is again used to modulate the writing beam; this cycle continues until all 32,000 scan lines have been completed.

2.1 *Interface between Postprocessor and Control Computer*

Let us for a moment consider the subsystem comprised of the PPG postprocessor and the control computer program. The postprocessor runs on a large central computer, receiving input from the XYMASK output file discussed previously, and writing output on magnetic tape. The information is read from the magnetic tape by the control computer program and used to load and update the display buffer. The magnetic tape constitutes an interface between two computer programs: the nature of the information on the tape can thus be varied to share, in some sense, the computational load between the two computers.

At one extreme, essentially all computation can be made in the postprocessor. The magnetic tape contains 32,000 records, each representing a complete 26,000-bit display buffer configuration. In this format each mask requires transmission of something like a billion bits between the computers. At the other extreme, the control computer can process the XYMASK output file and develop the display-buffer contents. Far too much computation is relegated to the control computer since display buffer regeneration cannot in general keep up with the pattern generator plotting rate. The result is a severe degradation in plotting time.

A compromise between the above extremes can be reached by considering the basic information required to properly load the display buffer. Let us see what is involved for an extremely simple mask containing a single vertical bar. For all scan lines which do not intersect the bar, the display buffer must contain all zero bits, while the bit configuration for the remaining scan lines is invariant and need only be set once. The basic data needed to load the display buffer involves only details of the changes, if any, in the bit configuration between successive scan lines. This is true even for complex masks since a high degree of similarity generally exists between one scan line and the next. One is thus naturally led to consider a magnetic tape encoding scheme which takes advantage of these similarities by detailing only the required configuration changes from one scan line to the next.

A complete description of the various commands used in the encoding scheme appears in Ref. 3. The commands fall naturally into three groups. The first group contains commands of an incremental nature for updating the bit configuration in the display buffer. Various combinations of these commands may be used to indicate that strings of one or more bits in the buffer are to be set to zeros or ones as required for the next scan. All bits not referenced in this fashion represent recurring mask detail and are unchanged for the next scan. The second group of commands deals with complete scan-line configurations. Commands are provided for specifying that the bit configuration for the next N scan lines is invariant, contains all one bits, or contains all zero bits. Commands in the final group are used to pass various parameter values to the control computer and are not of interest here.

2.2 *Postprocessor Algorithms*

We turn now to the functioning of the postprocessor. The input data resides on the XYMASK output file in the form of various parameter values and the coordinate specifications for the individual path and polygon geometric features in the mask or masks to be generated. The output is written on magnetic tape and consists of appropriate sequences of the commands discussed above. The necessary data processing can be iteratively characterized as follows: given the set of geometric figures intersected by the previous scan line, determine the set of figures intersected by the current scan line and compare the respective display buffer configurations; the result of this comparison is expressed in the encoding scheme and written onto tape. Iteration commences with a null set of figures in scan-line zero, and terminates when scan-line 32,000 has been processed.

The practical aspects of the above characterization belie its simplicity of statement. A single mask may contain several thousand individual geometric features. Furthermore, the features occur on the XYMASK output file in random order with regard to geometric position in the mask. Finally, it is important to accelerate the scan-line comparison process by quickly detecting sequences of scan lines which have the same display buffer bit configuration. The following paragraphs give a description of the methods and algorithms which were used.

The coordinates of the mask features on the output file represent final device dimensions measured in micrometers from an arbitrary datum point. These coordinates must be scaled up by the appropriate factor to compensate for photographic reductions of the primary pattern, and converted to address units. A coordinate translation is then made to center the mask on the primary pattern plate. The postprocessor is capable, at the user's option, of generating either normal-tone masks having opaque features on a clear background, or reverse-tone masks displaying clear features on an opague background. It is an interesting and perhaps unique characteristic of the system that the two tonalities are produced with equal ease and facility. For simplicity, we will consider only normal-tone processing.

Given the set of individual mask features as input and considering the raster-scan process by which the artwork is created, it is clear that we are primarily interested in the feature boundaries. Returning to the simple mask discussed above, the writing beam is switched on at the left boundary of the bar, remains on in the interior, and is switched off at the right edge. Thus for our purpose the rectangle is totally characterized by its left and right boundary lines together with their respective tonality shifts. More generally, each polygon feature in the mask can be similarly characterized by listing all of its boundary line segments not parallel to the scan-line direction, together with the appropriate tonality transitions. Any arcs which occur are approximated by a sequence of chords and are thus reduced to sets of line segments. Since path features are described on the XYMASK output file by centerline coordinates and width, some additional computation is required. Any arcs in the centerline are first approximated by chords, and path outline then obtained by translating the centerline line segments normally through distances of plus and minus one-half the path width. The path then becomes a polygon and is treated as above.

2.3 *Postprocessor Structure*

A simplified diagram of the postprocessor is shown in Fig. 2. Each mask requires a complete pass through the system. The line segment decomposition routines read the mask-feature descriptions from the XYMASK output file, convert the coordinates into address units, decompose each feature as described above, and write the resulting line segments with their tonality shifts onto the line-segment file. The set of line segments is next sorted into an order convenient for further processing. Each line segment is described by the two coordi-

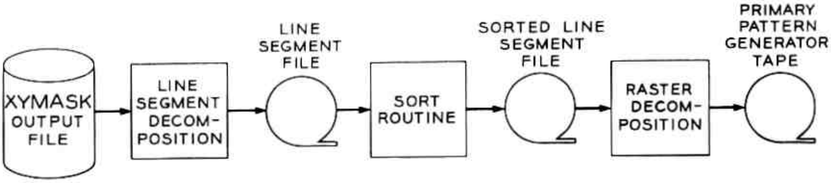


Fig. 2—PPG postprocessor system.

nate pairs of its endpoints. The endpoint which has the lower value for its Y coordinate is termed the lower endpoint. The sort is carried out using the lower endpoint Y value as the primary key, and the lower endpoint X value as secondary key. At the conclusion of the sort, the sorted-line-segment file contains the line segments in the order in which they are encountered by the raster scan. Line segments first encountered by scan N precede those first encountered by scan $N + 1$, and if several line segments are first encountered by scan N, they occur on the file in the order of increasing-scan positions.

The final section of the postprocessor reads the sorted-line-segment file, determines configuration changes between scan lines, and writes the appropriate commands on the PPG tape. This operation is carried out using a 26,000-bit image of the display buffer containing the bit configuration of the previous scan line and a linked list of all line segments contributing to the current scan line. The line-segment representation is compared to the bit-image configuration; any differences are appropriately encoded and written on the tape, and the relevant bits are changed in the bit image. When the comparison has been completed, the list of relevant line segments is updated by deleting those which do not intersect the next scan line and interpolating any new ones which do from the sorted-line-segment file. The scan routines are fairly simple but involve significant computer time. The postprocessor minimizes the number of scan comparisons by examining the line-segment list looking for scan lines which are identical to the previous one, or contain all-zeros or all-ones configurations. When such configurations are found, the scan comparison is bypassed, and the appropriate commands are written on tape. This capability allows very rapid processing for masks containing features having no slant-line boundaries.

The postprocessor execution time varies considerably with the complexity of the mask being generated. A typical interconnection mask ordinarily requires several minutes on an IBM 360/65 and writes something on the order of one-quarter-million bits on the output tape.

III. EBM POSTPROCESSING AND ALGORITHMS

This section describes a system of programs which interfaces the EBM pattern generator with XYMASK. This system consists of a postprocessor within the XYMASK system and a program for the pattern generator controller. The following short description of the EBM pattern generator will give an insight into the data transformations performed in both the XYMASK postprocessor and the control computer program.

3.1 *The EBM Pattern Generator*

The EBM is similar to a cathode ray tube; in both, a beam of electrons is focused and deflected to form a spot on a target. One difference is that in the EBM, the target is a high-resolution photographic plate, whereas in a cathode ray tube it is a phosphor screen. As the electron beam hits the target, the electrons directly expose the photographic emulsion and thereby produce a fine spot. A detailed description of the EBM pattern generator is given in this issue by W. R. Samaroo, et al.[4]

The EBM pattern generator includes a digital-control computer that drives, through appropriate interface equipment, a set of electrostatic beam deflection plates located within the EBM. The electron beam position on the target is controlled to fill mask features by drawing a sequence of adjacent line segments parallel to one coordinate axis. Fill-line data in the form of position and length are transmitted from the control computer to the interface where the digital fill-line data are converted to a sequence of analog voltages that are applied to the deflection plates.

Since a typical mask pattern may contain an estimated 10^5 fill-lines, it is impractical to read or even to store this data in the control computer. To make data processing more practical, the following strategy is used for the EBM pattern generator: While the interface controls the drawing of one fill-line segment, the control computer calculates the position and length of the adjacent line segment.

Input data to the control computer consists either of paths or of pairs of left-hand and right-hand boundaries specified by the endpoints of straight-line segments or the endpoints and centers of circular arcs. With this pattern-coding scheme, approximately 4000 words are required to represent a typical mask pattern of 10^5 fill-lines. This small volume of input data facilitates data transfer from the XYMASK postprocessor to the control computer. It is the task of the postprocessor to read the XYMASK output file, transform the data to right-hand and left-hand boundaries for the EBM, and to output this data.

3.2 *The EBM Postprocessor*

The EBM postprocessor is written in the *1 language[5] (read as star one) and in FORTRAN IV for the IBM 360/50 computer. *1 is used because of its inherent power in processing list-data structures and FORTRAN IV is used for input, output, and some of the more complex calculations.

The way the XYMASK output file describes the features of a mask does not conveniently distinguish for the EBM the areas inside and outside the periphery of each feature. Generally speaking, the more automatic the drawing device, the more work has to be done by a computer to obtain this information. Devices such as the coordinatograph and the Calcomp plotter, for example, require data in a form very similar to that of the XYMASK output file because these devices cut or draw along the periphery of each feature. Since the Calcomp plots are part of the "debug" steps and are used for alignment and correction, no further processing is required. In the case of the coordinatograph, an operator must further process the plots by deciding which sections of the rubylith are to remain as part of the mask and which are to be removed and then he manually removes the unwanted pieces. This step in mask making is computerized for the EBM.

The EBM postprocessor must interpret the XYMASK output file to determine which points are inside or outside the periphery of each feature. The EBM postprocessor converts the XYMASK output file data into sets of left-hand and right-hand boundaries whose minimum and maximum Y coordinates, when connected, are parallel to the X axis. The more nonconvex the feature, the more difficult the task becomes.

Since the EBM is a random-access plotter, the postprocessor processes one path or polygon at a time before proceeding to the next feature on the XYMASK output file. The data for a polygon are stored as a linked list in the *1 program. The program determines the lower left-hand and upper right-hand points by comparing the coordinates contained in the list. From this, two routes along the periphery are established, which eventually yield sets of left-hand and right-hand boundaries. The actual structure of the list-processing algorithm is too complex to be described here in detail.

One of the unique features of the EBM postprocessor is the interpretation of paths. As mentioned above, a path is described on the XYMASK output file as a centerline and a path width normal to the centerline. Postprocessors for drawing devices such as the coordinatograph must translate this path information into a polygon before the feature can be plotted. In other words, the postprocessor must find the periphery points for the path. The EBM postprocessor takes advantage of the form of the output file data by treating the path as the figure formed when a circular tool, having the path width as the diameter, is moved along the centerline. Rather than converting the path into polygon data and then processing the resulting polygon, the postprocessor passes the major portion of path processing onto the EBM's control computer. The description of the control computer algorithms, which follows, will explain how this data is handled.

3.3 *EBM Control Computer Algorithm*

The control computer is capable of calculating the boundary and outline points in less time than it takes for the EBM to draw fill-lines. Thereby, the interface and EBM become the limiting factors in allowing the pattern generator to maintain an average pattern drawing time of one microsecond per addressable point for a significant set of masks. The calculations of the endpoints of fill-lines along the left-hand and right-hand boundaries are based on integer arithmetic.[6] The following example of straight-line-to-arc boundaries illustrates the use of integer arithmetic in this application.

Consider a set of boundaries consisting of the straight line $Y = Y = (A/B)X$ and the circular arc $X^2 + Y^2 = R^2$. The constants A, B, and R^2 are integers calculated from the control computer input data.

In integer arithmetic, the straight line is redefined as:

$$F = BY - AX \tag{1}$$

where F represents a third dimension. Thus, the straight line can be considered as the intersection of two planes in F space, with equation (1) defining one plane, and the XY plane the other.

The introduction of the dimension F results in the following useful properties:

(*i*) F is zero on the straight line and has opposite signs for points X, Y on opposite sides of the straight line.
(*ii*) There exists a single value of F for each point in the XY plane.
(*iii*) F is an integer for all integer points X, Y.
(*iv*) There is no error in a sequence of integer solutions for F.
(*v*) The smallest integer number is 1. If this is the smallest addressable unit in the graphic field, then all points X, Y that are within 1 unit of the true solution represent the true solution in the XY plane.

These properties of F make it easy to form an algorithm for calculating integer points along a straight line. If equation (1) is evaluated at the point (0,0), then the resultant F is 0. Rather than evaluate equation (1) for F at all points, it is easier to calculate a change in F between adjacent integer points. The adjacent integer points (1,0), (0,1), and (1,1), (in the neighborhood of the straight line) have the integer F values of $-A$, $+B$ and $B - A$ respectively. According to property (*i*) the point (0,0) is on the straight line and the points (1,0) and (0,1) are on opposite sides of the line. According to properties (*ii*) and (*iv*), a step-by-step calculation of F values from the point (0,0) to (1,1) will result in the identical F value at the (1,1) point regardless of the steps taken en route. Choosing a sequence of points with the smallest F values guarantees that the points are as close to the straight lines as the address structure of the field allows.

According to property (*v*), there may exist several integer values of X and Y that represent the true solution point of the straight line. This observation is used to form a more practical algorithm where only one addition and one test for sign of F per point is required to find the next integer point along the straight line.

A circular arc is the other boundary considered in the example. The circular arc is redefined in integer arithmetic as

$$F = X^2 + Y^2 - R^2 \tag{2}$$

where again, F represents an added dimension. The circular arc is thus formed in F space by the intersection of the XY plane with a parabaloid. The properties (*i*) through (*v*) also hold true for equation (2).

A sequence of integer points along a circle is computed by taking unit increments parallel to either the X or Y axis and computing the resultant F values; for a change of 1 addressable unit in the X direction, F changes by $2X + 1$. The corresponding computation is shift left, increment, and add. A test of sign of F determines whether the next step increments X again or decrements Y. The coordinates thus generated are located along the circular arc and form the mask-feature boundary points.

It is also possible to construct an integer arithmetic algorithm to compute points along the outline of a path. According to the path definition, points on each side of the outline represent the envelope generated by a circle moving along the centerline as illustrated in Fig. 3.

The path algorithm finds points along the outline by choosing points along the circle until the normal to the circle is aligned with the normal to the path centerline. The circle is then displaced along the path centerline and the above process is repeated. A separate but identical algorithm is used for finding points on the opposite outline. Fill-lines are drawn parallel to one coordinate axis between these points. While the above algorithm appears to be complicated, surprisingly few calculations are required to find the endpoints of the fill-lines. For example, the slope of a curve in the XY plane is given by the ratio of change of F for changes in the X and Y directions, where the change of F in both directions is already available from the straight-line and circular-arc algorithms. The normal to a curve is the negative inverse of the slope, and thus the only additional computation required in the path algorithm is the comparison of a sequence of two integer ratios.

As is evident from the above discussion, only a few instructions

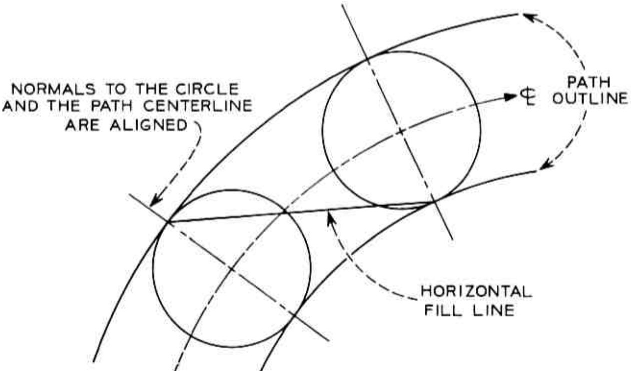


Fig. 3—Construction of a path.

are required in the control computer to calculate the endpoints of a fill line between a set of boundaries. As a result of the redefinition of the problem in integer arithmetic, the calculations in most instances are completed before fill-line generation is finished allowing the EBM pattern generator to maintain the one microsecond per addressable point drawing speed.

The postprocessor execution time varies with the complexity of the mask being generated but to a lesser degree than for the PPG postprocessor. Several minutes on an IBM 360/65 are ordinarily required for a typical interconnection mask.

IV. DISCUSSION

Several computer systems used in the generation of integrated-circuit masks have been described in the preceding sections. The first sections dealt with the XYMASK system which links the circuit designer to the mask-fabrication process. XYMASK provides a computer-independent language for describing the mask configurations, and produces either outline drawings or mask artwork on one or more of a number of different graphical output devices. The majority of all Bell Laboratories and Western Electric Company masks are produced using the XYMASK system.

The next two sections described XYMASK subsystems which generate artwork on the PPG and EBM. These two plotters fundamentally differ in that the first uses a raster-scan technique, while the second is a random-access device. Each is supported by a dedicated control computer. The subsystem descriptions indicate a degree of similarity in postprocessor functions, but different approaches toward the division of the necessary computation between the postprocessor and the control computer.

V. ACKNOWLEDGMENTS

Many persons have contributed to the development of both the XYMASK system and the PPG postprocessor. The efforts of V. A. Fasciano, B. R. Fowler and S. Pardee are singled out as being of particular importance in system design as well as implementation.

We would also like to acknowledge J. E. Gorman, A. D. Janzen, R. Sedgewick, and C. C. Wyckoff for contributions to the EBM control computer program development.

REFERENCES

1. Fowler, B. R., "XYMASK," Bell Laboratories Record, *47,* No. 6 (July 1969), pp. 204–209.
2. Christensen, C., and Pinson, E. N., "Multi-function Graphics for Large Computer System," American Federation of Information Processing Societies (AFIPS), Conference Proceedings, 1967 Fall Joint Computer Conference.
3. Dowd, P. G., Cowan, M. J., Rosenfeld, P. E., and Zacharias, A., "The Primary Pattern Generator: Part III—The Control System," B.S.T.J., this issue, pp. 2061–2067.
4. Samaroo, W., Raamot, J., Parry, P., and Robertson, G., "The Electron Beam Pattern Generator," B.S.T.J., this issue, pp. 2077–2094.
5. Newell, A., Early, J., and Haney, F., **1 Manual,* Carnegie Institute of Technology, Pittsburgh, Pennsylvania, June 26, 1967, Advanced Research Projects Agency No. SD-146.
6. Gorman, J. E., and Raamot, J., "Integer Arithmetic Technique for Digital Control Computers," Computer Design, *9,* No. 7 (July 1970), pp. 51–57.

Device Photolithography:

The Primary Pattern Generator

Introduction

By K. M. POOLE

(Manuscript received July 10, 1970)

The need for a new, high-speed pattern generator capable of producing the more complex and precise circuit patterns required in the 1970s has already been discussed.[1] This paper describes the design and operation of the Primary Pattern Generator (PPG) in some detail. For the convenience of the reader, the paper has been separated into four parts. Part I covers the optical design of the machine, including the considerations which led to the choice of an argon laser light source, a recording emulsion, and an optimum combination of spot size and brightness. The original choice of a mechanically scanned system was made on the premise that, with such an approach, the required accuracy could be built in and retained over many years of operation, and Part II discusses the principal considerations behind this premise. In that paper are discussed the dimensional stability of the structural materials and their use in an extremely stiff structure, the features provided to align the parts of the system to the required tolerances, and the design of drive systems, essentially free from both vibration and wear. The control of the machine to produce the pattern encoded on the input tape is discussed in Part III; Part IV deals with the methods used to align the assembled machine and details the pattern accuracy and reproducibility which was achieved.

The PPG, a highly automated system requiring operator action at very few points in the cycle, is part of an overall system running under computer scheduling. Operator acceptance of the system has been excellent, perhaps due to the incorporation of status displays beyond the essential minimum including a real-time display of the pattern being produced.

REFERENCE

1. Howland, F. L., and Poole, K. M., "An Overview of the New Mask-Making System," B.S.T.J., this issue, pp. 1997–2009.

Device Photolithography:

The Primary Pattern Generator
Part I–Optical Design

By M. J. COWAN, D. R. HERRIOTT, A. M. JOHNSON and A. ZACHARIAS

(Manuscript received July 10, 1970)

I. INTRODUCTION

The basic design concept of the primary pattern generator (PPG) is the production of a linearly scanning, small, constant-size light spot. The scanning system consists of a regular polygonal-prism mirror which rotates about its axis of highest symmetry. The mirror faces are used sequentially to reflect a collimated light beam into a lens (for example, the scanning lens of Fig. 1). The collimated light is focused to a spot which scans a line in the focal plane of the lens as the polygonal mirror rotates. Located in the focal plane of the lens is a flat, glass photographic plate. The glass plate is moved by the desired scan line separation during the time required to bring the succeeding mirror facet into proper position.

The collimated beam incident onto the rotating mirror is formed by the scanning lens from a diverging beam obtained from a laser. The location of the reflecting mirror facet must be close to the aperture plane of the scanning lens in order to insure that the mode is not truncated by the physical lens apertures after the light is reflected from the mirror facet. Translation of the reflecting facet will not affect the position of the focused spot; the spot position is uniquely determined by the directions of the incident collimated beam and of the reflecting mirror facet relative to the optic axis of the lens. A barrel distortion is designed into the scanning lens such that the linear velocity of the focused spot is proportional to the angular velocity of the rotating mirror.

The machine just described is basically analog along its fast-scan axis, although it is digital along the slow (substrate translation) axis. Since the required reproducibility is greater than the required accuracy,

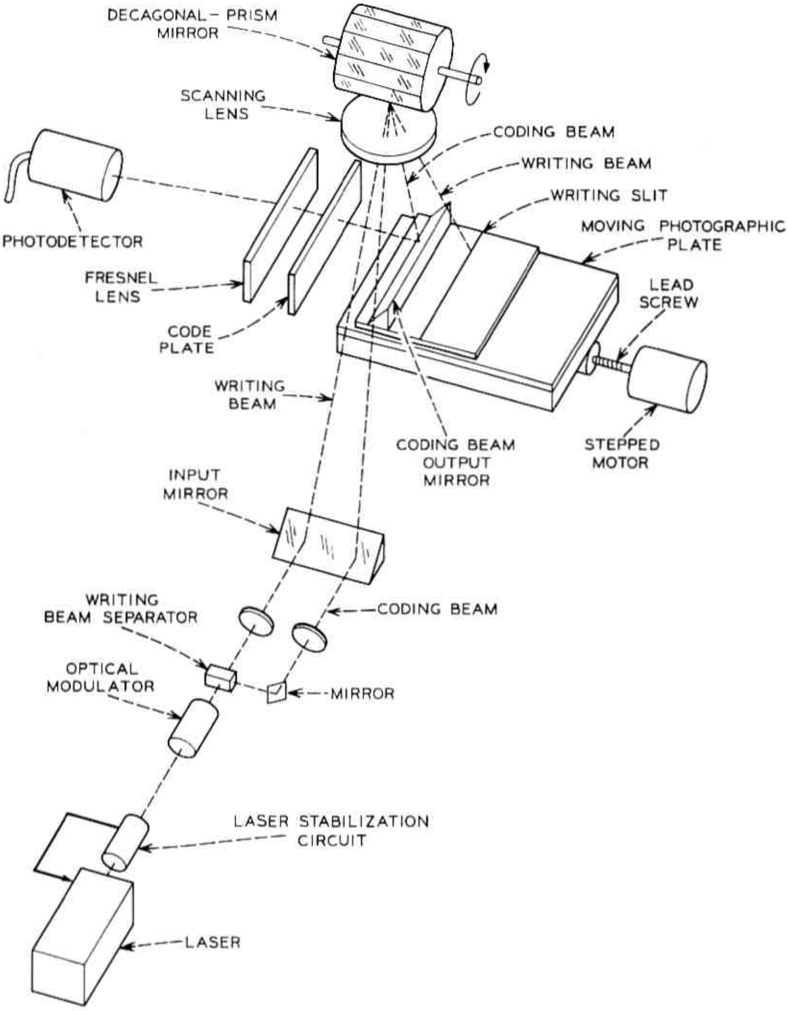


Fig. 1—Schematic of primary pattern generator.

a digitally operating machine is more desirable than an analog machine. The fast-axis can be made digital by using a separate beam to scan over a grating type of code plate. The location of this beam on the code plate tracks the position of the writing beam and generates timing pulses for a control computer. The resolution of the code plate must be as good as the reproducibility required; that is, the code plate system must be capable of resolving 26,000 positions per scan length.

The pattern size is principally established by the capabilities of the scanning lens. The minimum spot diameter is determined by the approximate diffraction limitation of equation (1),[1] obtained when a lens aperture is uniformly illuminated.

$$I(r) = \left(\frac{2J_1(x)}{x}\right)^2 I_0 \qquad x = \frac{\pi r}{\lambda f_n}. \tag{1}$$

Here, f_n is the f-number of the lens forming the image $I(r)$; r is the radial distance from the image center; I_o is a constant proportional to the intensity illuminating the aperture; and λ is the wavelength. Using this relation, we approximated the half-power diameter of a spot formed by such an illuminated lens to be

$$D \approx 0.58 f_n \qquad D \text{ in } \mu\text{m}, \lambda = 520 \text{ nm}. \tag{2}$$

We now consider that the polygonal mirror will have some wobble to its motion, and further, that all faces of the mirror will not be exactly parallel to the rotation axis. Consequently, to reduce the effect of these mirror defects on the pattern, the scanning lens should operate with as large a field angle as possible. This wide-angle requirement limits the f-number for which diffraction limited performance can be obtained in a lens. For a 48° field angle, calculations made by Tropel, Inc.,* showed that a minimum f-number of 13 could be used for good performance of the coding beam over the field. Using equation (2), a spot size of 7.5 μm half-power width is thus obtained; this will be approximately the size of the address unit. Since 26,000 address units are required for a full scan line, an address size of 7.0 μm will allow the full pattern of 26,000 by 32,000 address units to fit on a standard 8″ × 10″ photographic plate.

To produce a complete pattern in less than 10 minutes, each of the 32,000 scan lines must be traversed in less than 20 ms. Since the writing-beam diameter will be less than twice the address spacing, the beam must sweep its own diameter in less than 800 ns. To produce sufficient exposure on high-resolution emulsion[2] requires a beam brightness obtainable only from a laser. However, the writing-beam power required is only 20 μW. Orthochromatic emulsion is desirable since it will allow a safelight environment. Thus an argon laser,[3] operating at 5145 Å wavelength was chosen as the light source.

* Located at 52 West Avenue, Fairport, New York.

It operates in the lowest transverse mode,[4] thus the radial intensity distribution anywhere in the beam path is gaussian. The output of the laser is stabilized by feedback through the laser power supply to a variation of less than 1 percent, thus insuring uniform exposure of the photographic plate.

II. THE PHOTOGRAPHIC EMULSION AND THE EXPOSURE PROCESS

The sweep of the writing beam across the photographic plate results in a variation of the exposure of the emulsion in a direction normal to the scanning direction. If we use the scanning velocity as v_0 and the intensity distribution of the scanning spot as

$$I(r) = \frac{2P}{\pi w^2} \epsilon^{-2r^2/w^2} \tag{3}$$

where P is the total power in the writing beam and w is the waist radius,[5] then taking the scan to be x-directed along the line $y = y_0$, the variation of exposure in the y-direction is obtained by integration, as

$$\begin{aligned} E(y) &= \frac{2P}{\pi w^2} \epsilon^{-2(y-y_0)/w^2} \int_{-\infty}^{\infty} \epsilon^{-2(v_0 t)^2/w^2}\, dt, \\ &= \frac{P}{w v_0} \sqrt{\frac{2}{\pi}}\, \epsilon^{-2(y-y_0)^2/w^2}. \end{aligned} \tag{4}$$

The next line will scan with y_0 changed by one address spacing and the exposure produced by this scan will be added to the exposure of the first scan. The total exposure produced by N scans is thus obtained by summing N displaced gaussians given by equation (4).

A similar analysis is used to obtain the exposure resulting from modulation of the writing spot. In this case, the beam is turned off at $x = 0$ for each scan. As a first approximation, we assumed the intensity of the writing spot to decrease with a relaxation time of $\tau = d/v_0$ where d can be interpreted as a rise distance in analogy to a rise time. The exposure caused by a single trace having the beam turned off at $x = 0$ becomes

$$\begin{aligned} E(x, y) = \frac{2P}{\pi w^2} \epsilon^{-2(y-y_0)^2/w^2} \Bigg[&\int_{-\infty}^{0} \epsilon^{-2(x-v_0 t)^2/w^2}\, dt \\ &+ \int_{0}^{\infty} \epsilon^{-v_0 t/d} \epsilon^{-2(x-v_0 t)^2/w^2}\, dt \Bigg] \end{aligned} \tag{5}$$

which is evaluated in terms of the error function and its complement.[6]

$$E(x, y) = \frac{P_0}{wv_0\sqrt{2\pi}}\,\epsilon^{-2(y-y_0)^2/w^2}$$
$$\cdot\left[\operatorname{erfc}\left(\frac{x\sqrt{2}}{w}\right) + \epsilon^{-x/d}\epsilon^{w^2/8d^2}\left(\operatorname{erf}\left(\frac{x\sqrt{2}}{w} - \frac{w\sqrt{2}}{4d}\right) + 1\right)\right]. \qquad (6)$$

Application of this exposure to a high-contrast emulsion will result in the production of a density gradient at the boundaries of the exposed regions. The greatest magnitude of the gradient will occur very close to the contour of 0.5 optical transmission through the developed image. The task of determining the actual image formed by the exposure function of equation (6) is thus reduced to tracing the contour of the exposure necessary to produce 0.5 transmission and to evaluate the exposure gradient normal to this contour. A computer program was written to evaluate equation (6) over a matrix of points. Table I shows some of the results of these calculations. An exposure of 1.00 is used to produce the 0.5 transmission value.

For simplest operation, five scan lines or a five-address modulation should produce an image five address units in dimension. To obtain a best compromise between freedom from mirror facet wobble and maximum edge gradient, we chose to operate with a half-power writing beam diameter between 1.3 and 1.7 address units (9 to 12 μm). Equation (4) can now be used to calculate the beam power required to obtain proper exposure on various emulsions. For a spot velocity of approximately 16 m/s, 20 μw of beam power will produce a maximum exposure of about 120 ergs/cm². High resolution plate[2] requires over 1000 ergs/cm² for proper exposure. Eastman Kodak Company had an emulsion which reached proper exposure between 20 and 100 ergs/cm², although it was not a standard product. This emulsion, called Minicard,

TABLE I—VARIATION OF EXPOSURE PARAMETERS

Half-Power Spot Diameter	2.7	2.0	1.7	1.3	2.0	1.7	1.3
Peak Exposure of a Single Scan	1.1	1.8	2.5	4.7	0.9	1.1	1.4
Width for 5-Scan Lines	6.0	6.0	6.0	6.0	5.0	5.0	5.0
Gradient ($\partial E/\partial y$)	1.0	1.5	2.0	3.1	1.0	1.2	1.7
Peak-Exposure for Large Number of Scan Lines	2.9	3.7	4.5	6.7	2.0	2.0	2.0
Length for 5-Address Modulation	6.1	6.2	6.3	6.5	5.0	5.0	5.0
Gradient ($\partial E/\partial x$)	0.9	1.3	1.6	2.1	0.8	0.9	1.0

was available on special order; Eastman Kodak now produces 8″ × 10″ glass plates coated with Minicard emulsion.

The glass photographic plates must have a very flat emulsion surface. Fig. 2 is an illustration of the effect of plate camber. The emulsion surface will be held near the extremes of the scan line. However, plate camber will cause registration errors between plates because of the angular scan of the writing beam. The maximum angle made by the writing beam and the normal to the photographic plate is 15°. To produce less than a one-address-length error between X_1 and X_2 of Fig. 2, the plate camber must be less than ±28 μm. This specification is safely met by Kodak microflat plates, but is very far from being met by the specifications of lower grades of glass plates.

III. THE ROTATING MIRROR AND SCANNING LENS

The dimensions of the rotating polygonal mirror are determined by the scanning-lens aperture. Since the f-number, field size and field angle of the scanning lens have been determined by equations (1) and (2), the aperture size is also determined. The facet size of the polygonal mirror can be found by geometry, as well as the overall size of the polygon. Referring to Fig. 3, the radius of the polygon must be large enough to keep the vertices out of the lens aperture during the rotation producing the scan of a line.

Since a gaussian illumination of the aperture is being used, the full aperture diameter must be larger than that computed from equation (2) for a uniformly illuminated lens. A best estimate of satisfactory performance with gaussian illumination was $f/10$ and the polygonal mirror was designed not to truncate this aperture during the scan. The value of R for this condition is 9.7 cm. The location of

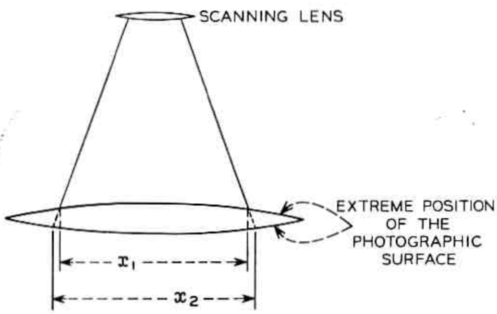


Fig. 2—The effect of photographic plate camber.

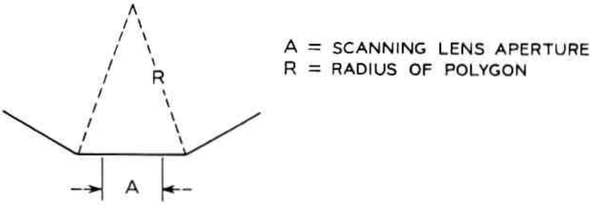


Fig. 3—Polygonal mirror-lens aperture geometry.

the aperture plane of the scanning lens must lie at the approximate location of the mirror facet. To obtain a uniform scanning velocity from constant angular velocity of the polygonal mirror, a $\theta/\tan\theta$ distortion was part of the scanning-lens design; θ is the angle between incident collimated light and the lens axis.

The number of facets on the polygonal mirror determines the ratio between the time available for writing and the unavailable time. Since the field angle of the written line is 45.4°, 22.7° of mirror rotation is spent writing a line. For the decagonal mirror used, 36° of mirror rotation is required to go from the start of one scan to the start of the next scan. Hence, 13.3° of rotation are unavailable. In order to write a complete pattern in 10 minutes, each scan line must be traversed in 18.8 ms; 11.8 ms writing and 7.0 ms waiting for the next facet to come into position. It is during this wait that the photographic plate is advanced one address spacing (7 μm).

IV. THE OPTICAL MODULATOR

The writing beam modulator used is an acoustooptic deflector.[7] The modulator operates by the interaction of the laser beam with a 50-MHz ultrasonic wave in a piece of fused silica. This device deflects approximately 2 percent of the power of the incident laser beam at an angle of 4 mrad to the incident beam when the modulator is energized. Since the modulator is located in a near field region of the laser beam, the two beams emerge from the modulator each nearly collimated but having angular separation. These beams are then passed through a 10-cm focal length lens which transforms the angular divergence into a displacement sufficient for physical separation. The separation is accomplished by a knife edged mirror which has better than a 40-dB discrimination between the beams.

The 2 percent power in the deflected beam provides more than 17-dB on-off ratio and is limited by back reflections and scattering.

However, this is sufficient for the writing-beam modulation. The undeflected beam is used as the coding beam. The modulator has a rise time of less than 200 ns, including the transistor drivers. The transducer is X-cut crystal quartz.

V. MODE-MATCHING OPTICS

A series of lenses are required to transform the output mode of the laser to modes required for the modulator and then to the modes required by the scanning lens. The output of the laser is limited to a TEM_{00} mode by use of an aperture within the laser cavity. The calculation of the positions and focal lengths of the required transforming lenses was done using the method described by H. Kogelnik.[8]

The first transformation is between the laser output and the optical modulator. The modulator requires a 300-μm waist radius in the fused silica. In turn, this mode is transformed to a 55-μm waist located at the knife-edged separation mirror. The writing beam is transformed to approximately a 9-μm waist radius at the object focal plane of the scanning lens and the proper writing spot is produced. The code beam is transformed by a pair of lenses. The first produces a mode having a waist radius of 800 μm, an essentially collimated beam for the 50-cm distance to the code plate. The second lens is a cylindrical lens which produces a 4-μm waist radius in one direction and does not change the 800-μm waist radius in the perpendicular direction. This slit-shaped spot is imaged by the scanning lens to a slit spot on the code plate.

VI. THE CODE PLATE

The code plate is a ruled grating having approximately 13,300 cycles. Each cycle consists of a 7-μm opaque region and a 7-μm clear region. The slit shaped coding beam is focussed in its narrow dimension to best resolve the grating. The long dimension of the beam is aligned to the ruling direction of the grating. In this manner, small defects in the grating, dust specks and pinholes do not significantly affect the code-plate system.

The coding beam will traverse the full-field angle over the scan of a line. In order to collect the coding beam onto a photodetector after it has passed through the code plate, a Fresnel lens is positioned beyond the code plate (see Fig. 1). This lens images the aperture of the scanning lens onto the face of a photomultiplier tube. The sensitivity of this device is required so that the coding beam can be attenuated by approximately 20 dB before it illuminates the scanning lens. If this attenuation is not used, then the scatter from the intense coding beam fogs the photographic plate and reduces the modulation capable of being obtained with the writing beam alone.

The processing and use of the code plate output is described in Part III—The Control System. The alignment of the code plate for production of an accurate scan is described in Part IV—Alignment and Conclusions.

VII. ACKNOWLEDGMENT

The acoustooptic modulator, its driver and gating amplifier were designed and constructed by R. W. Dixon and R. V. Goordman.

REFERENCES

1. Born, M., and Wolf, E., *Principles of Optics,* New York: Pergamon Press, 1959, Chapter 8.
2. *Kodak Plates and Films for Science and Industry* (P-9), Rochester, New York: Eastman Kodak Company, 1967.
3. Labuda, E. P., Gordon, E. I., and Miller, R. C., "Continuous Duty Argon Ion Lasers," IEEE J. of Quantum Elec., *QE1,* No. 6 (September 1965), p. 273.
4. Fox, A. G., and Li, Tingye, "Resonant Modes in a Maser Interferometer," B.S.T.J., *40,* No. 2 (March 1961), pp. 453–488.
5. Goubau, G., and Schwering, F., "On the Guided Propagation of Electromagnetic Wave Beams," Trans. IRE, *AP-9,* No. 3 (May 1961), pp. 248–256.
6. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions,* National Bureau of Standards, 1965, Chapter 7.
7. Gordon, E. I., "A Review of Acoustooptic Deflection and Modulation Devices," Proc. IEEE, *54,* No. 10 (October 1966), pp. 1391–1401.
8. Kogelnik, H., "Imaging of Optical Modes—Resonators with Internal Lenses," B.S.T.J., *44,* No. 3 (March 1965), pp. 455–494.

The Primary Pattern Generator Part II–Mechanical Design

By G. J. W. KOSSYK, J. P. LAICO, L. RONGVED and J. W. STAFFORD

(Manuscript received July 10, 1970)

I. INTRODUCTION

The primary pattern generator (PPG) is an electromechanical light-scanning system with an unusual combination of speed and accuracy. A 10-μm-diameter light spot can be addressed successively to any or all points of a 26,000-wide by 32,000-long rectangular point array with 7-μm vertical and horizontal spacing in about ten minutes. This corresponds to a scanning rate of one spot per 600 nanoseconds. The light spot is placed repeatedly to an accuracy of about a $\pm$7-μm total accumulated error over the whole array, and the vertical and horizontal spacing between points is maintained within ± 1 μm.

The rectangular point array is scanned one line at a time at the rate of 53 lines per second by successive sweeps of a monitored laser beam across the width of the array interposed by 7-μm steps of the photographic plate in the perpendicular direction. The essential components of the scanning system are shown in Fig. 1. The laser generates a light beam which, by various stationary mirrors, is directed to the acoustooptic modulator. When this modulator is turned on, a small portion of the laser beam is slightly deflected and is denoted the write beam. The major portion of the light beam, called here the code beam, passes through the modulator with no directional change. When the modulator is turned off, the light beam passes through unchanged. The response time of the modulator is of the order of 10 nanoseconds which is very small compared to the 600-nanosecond period it takes the scanning beam to move from one addressable point to the next.

By further fixed mirrors and lenses, the code and write beam is brought to focus at neighboring points near the edge of the photographic plate.

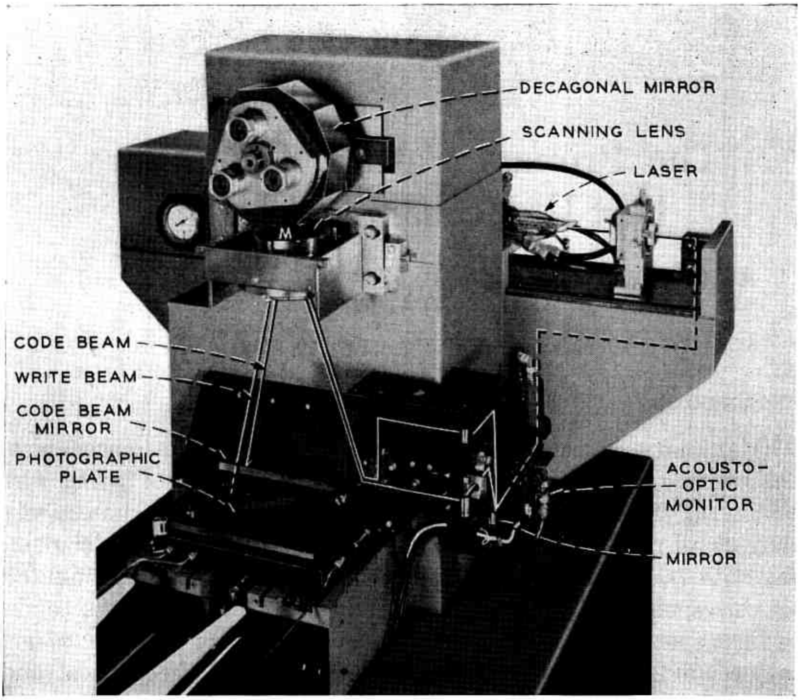


Fig. 1—Primary pattern generator.

By means of the scanning lens and the decagonal mirror, the focused spot of the write beam is imaged onto the photographic plate. The focused point of the code beam is, by the same means and one additional code beam mirror, imaged onto a code plate. The code beam is intercepted by the code plate except at 7.0-μm-wide transparent lines on 14-μm centers. The light passing through these transparent lines is collected in a photocell by means of a Fresnel lens. As the decagonal mirror turns, the two beams move together. The code beam, by pulsing the photodetector, yields positional information to the computer which, by means of the modulator, regulates the write beam on or off as required for proper exposure of the photographic plate.

The decagonal mirror spins at 300 rpm resulting in 53 write-beam sweeps per second. The 10 facets of the decagonal mirror are inclined to the mirror's radial symmetry axis at a very small angle which is identical for all facets within $\pm\frac{1}{4}$ of one second of arc. Furthermore, the mirror's radial-symmetry axis spins with a wobble less than 1/10 of one second of arc. Therefore, any sweep of the write beam when the photographic plate is fixed traces lines that are separated by no more than ±½ μm.

The 300-rpm speed of the decagonal mirror results in about 11-ms-duration sweeps across the photographic plate, with about a 7-ms-long period between the end of one and the beginning of the next sweep. The computer may write in every sweep, and the step system must be designed so that a step may be completed in the 7-ms period between sweeps. If the computer writes in every sweep, the table steps at 53 steps per second. If the computer cannot write in every sweep, one or more steps are skipped as required for the computer to catch up. This step motion is a sophisticated vibration-free one where each step is equal to the next within ±½ μm, and the total accumulative error over 32,000 steps is about ±5 μm assuming temperature control within 0.2°C.

II. MATERIALS SELECTION

The material used for the major PPG structure is Meehanite GC40. This material was chosen for its great dimensional stability with time. To insure that the material was initially stress-free, a three-step heat treatment-machining sequence was used. Briefly:

(*i*) After casting
  (*a*) Heat to 1600°F. Hold 2 hours.
  (*b*) Cool to 1250°F at 35°F per hour.
  (*c*) Hold at 1250°F for 10 hours.
  (*d*) Cool to 200°F at 20-25°F per hour.

(*ii*) After Rough Machining (allow 0.020″ for final machining)
  Thermally cycle: 210°F to 400°F to −120°F to 400°F
  to 200°F. Hold at −120°F and 400°F for 2 hours.
  Final cooling to 200°F must not exceed 25°F per hour.

(*iii*) After Dual Machining
  (*a*) Heat to 300°F. Hold for 6 hours.
  (*b*) Cool to 200°F at 20-25°F per hour.

The residual stress after heat treatment will not exceed 200 psi, resulting in a maximum relaxation strain of about 10 microinches per inch. Micro-creep tests conducted at Battelle Institute indicated that most of this relaxation occurs in the first four to six weeks which is before assembly of the pattern generator. Thus, only a few microinches per inch is expected during the life of the pattern generator.

III. TWO SPECIAL AXIAL ALIGNMENTS

Two very accurate axial alignments are made in the pattern generator. In one, the axis of an air bearing is aligned with the radial-symmetry axis of the decagonal mirror. In the other, the axis of the air bearing is aligned with the direction of motion of the step table. Both alignments use an elastic micromanipulator which was developed especially for the pattern generator. The alignments are essentially identical and only the decagonal mirror alignment is described here.

3.1 *The Elastic Micromanipulator*

The elastic micromanipulator is based upon a very elementary mechanical deamplification device. It consists of two springs that are connected in series and deflected against a support. In the static case, the total deflection of the spring, $\Delta\delta_1$, is related to the deflection of the interface of the springs, $\Delta\delta_2$, by the relationship

$$\Delta\delta_2 = \frac{k_1}{k_1 + k_2} = \Delta\delta_2$$

where k_1 and k_2 are the respective spring constants. The motion, $\Delta\delta_1$, is thus directly related to $\Delta\delta_2$ by the deamplification factor $F = k_1/(k_1 + k_2)$, which can be made as small as one pleases by choosing $k_2 \gg k_1$. In order to use such a device as a micromanipulator, one provides a fine screw to manually produce the deflection, $\Delta\delta_1$, and one attaches the body to be moved to the spring interface so that the corresponding body motion is $\Delta\delta_2$ as shown in the lower part of Fig. 2.

3.2 *Alignment of the Decagonal Mirror*

The adjustment for axial alignment of the decagonal mirror consists of three elastic micromanipulators placed 120° apart and equidistant from the symmetry axis. Between the face of the air-bearing spindle and one side of the mirror are the three stiff springs and, on the other side of the mirror directly opposite to these stiff springs, are the three soft springs which can be pressed against the mirror individually by three fine-adjusting screws.

The nature of the three stiff springs requires some explanation. The air-bearing face is machined with three raised $\frac{1}{4}'' \times \frac{1}{4}''$ areas as indicated in Fig. 3. The surface of these areas is finished machined with a stationary tool when the air bearing is spinning so that their surfaces lie in a plane normal to the air-bearing axis within about a second of arc. The decagonal mirror has an optically flat end face and this face

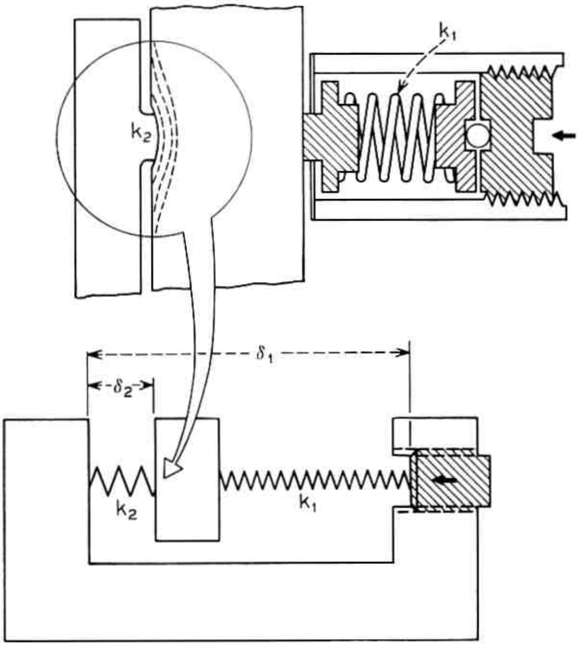


Fig. 2—Elastic micromanipulator.

is placed directly against these three pads. As the mirror is pressed against these raised areas by the soft springs on the opposite side of the mirror, the pads elastically indent the mirror as indicated on the upper part of Fig. 2. There is also some corresponding local indentation of the air-bearing face. Except for these small local regions of deformation, the mirror and the air bearing remain essentially rigid and the elastic deformation in the three small regions serves the purpose of the three stiff springs. The various mechanical elements are shown in detail in Fig. 4.

The relationship between the force exerted by the soft springs and the corresponding deflection of the stiff springs can be worked out from a classical elasticity solution due to J. Boussinesq. From this solution one can determine the effective spring constant associated with each of the three stiff springs. They are given approximately by

$$k_2 = 2 \cdot 10^6 \text{ lb/in.}$$

The soft springs on the opposite side have a spring constant given by

$$k_1 = 6 \cdot 10^2 \text{ lb/in}$$

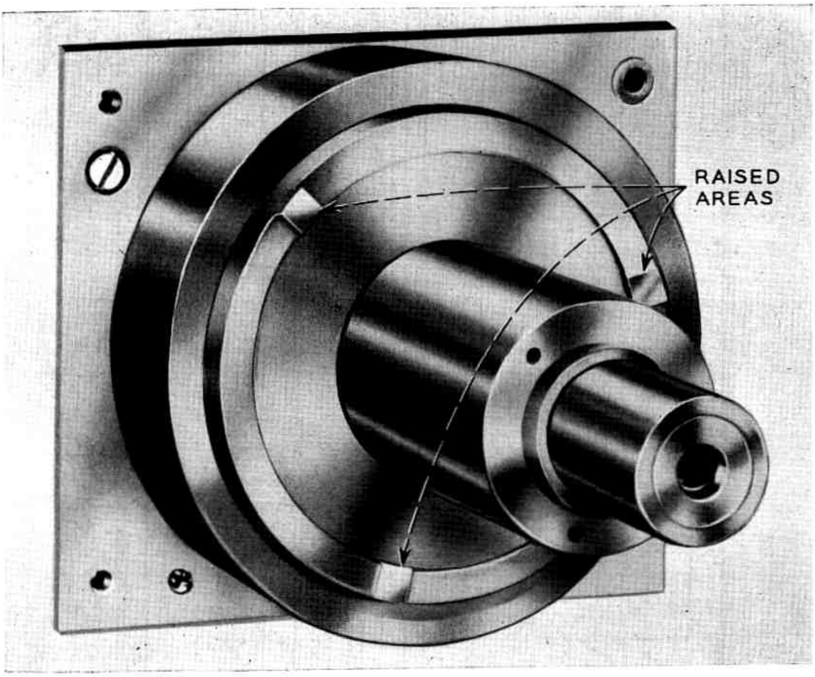


Fig. 3—Air-bearing spindle with raised areas.

and the amplification factor, F, works out to be about

$$F = 3 \cdot 10^{-4}.$$

The pitch of the adjusting screws is 40 turns per inch, and thus for one complete revolution of the adjusting screws the mirror will move about $18 \cdot 10^{-2}$ microns. When only one adjusting nut is advanced, the mirror will rotate about an axis passing through the two raised areas opposite the other two adjusting nuts. The raised areas are separated by about 7 cm, and thus the resulting rotation of the mirror equals about (0.6) second of arc per revolution of the adjusting nut. Since the adjustment is carried out together with an instrument to measure the mirror axis run out, there is no need to know this relationship exactly.

In the PPG, the mirror axis is aligned with the air-bearing axis to $\frac{1}{10}$ of a second, and we know that the adjustment remains stable to this accuracy over long periods.

When the mirror facets are measured to perform the final grinding operations, the mirror is aligned to $1/50$ of a second. This more precise adjustment has been demonstrated to be stable over several days, but it has not been evaluated on a long-term basis.

IV. THE STEPPING SYSTEM

There are two simple and fundamental concepts involved in the pattern-generator stepping system. One of these is a special electronic drive for the step motor used in the mechanical drive of the stepping system. The other is tuning of the natural frequency of the second mode of motion of the mechanical drive. Together these two concepts permit vibration-free stepping in the absence of passive damping. There are also several practical problems involved in the construction of the step table. One describes here first the two simple concepts, next the problems of construction, and last some experimental results.

4.1 *The Special Electronic Drive*

In order to describe the special electronic drive, first one describes certain characteristics of the stepping motor. The motor torque, T, as a function of the angular position of the armature, θ, is shown in Fig. 5a for a given current in the two motor windings. The amplitude of the sinusoidally varying torque is called the holding torque. The hold-

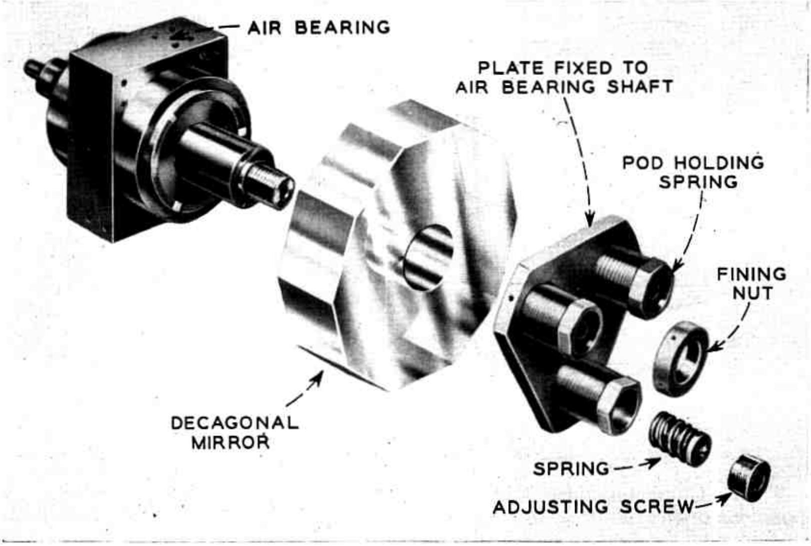


Fig. 4—Telescopic view of the decagonal mirror adjustment.

ing torque is proportional to the current in the motor windings. The magnitude of this current is usually kept constant, and only its direction is changed in the normal operation of the stepping motor. The effect of successively changing the direction of the current in each motor winding is indicated in Fig. 5b.

The mechanics of a simple operation of a stepping motor are essentially as follows: Assume the motor to be at rest in step position, n, which is one of the stable-equilibrium positions associated with the motor torque indicated by the solid curve in Fig. 6. Let the current be changed in one winding, thus bringing about the motor torque indicated by the dotted curve. The motor will now accelerate towards the step position $n + 1$ and, depending upon the damping in the motor, assumed less than critical, it will vibrate about the new position with decaying amplitude. This vibration is completely intolerable for the present application. Furthermore, if the motor is stepped continuously, vibration build-up from one step to the other occurs. To eliminate the vibration, the motor is provided with a special electronic drive. This drive provides three timed current settings for the motor per step which are applied as follows: Assume as before that the motor is at rest in position n as indicated in Fig. 6. The current is now reversed

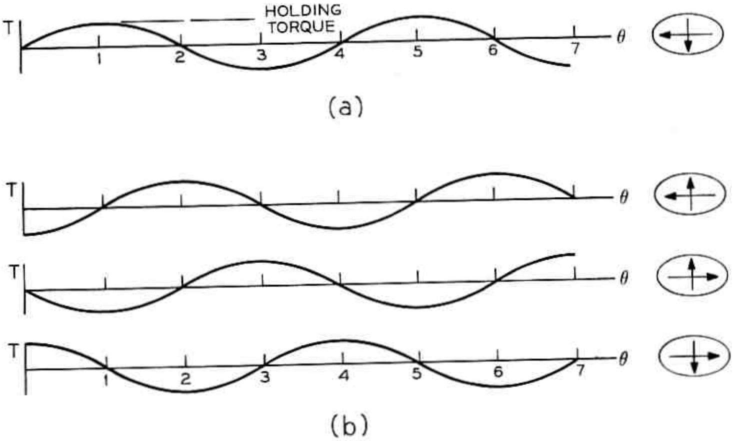


Fig. 5—Characteristics of the stepping motor. (a) 0, 1, 2, $\cdots$, N, are the step positions of the motor. 2, 6, $\cdots$, $(2 + 4I)$, where I is an integer, are equilibrium positions of the motor for a particular current direction in the two-motor windings as schematically indicated by the arrows in the ellipse on the right. (b) The motor torque as a function of theta is simply translated by one step each time the current is reversed in one winding.

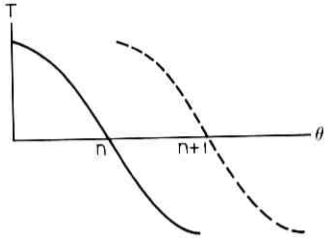


Fig. 6—Motor torque, T, as a function of angular position, θ.

in one winding for a timed period, t_1, bringing about the motor torque indicated by the dotted curve. As before, this will accelerate the motor towards its new step position, $n + 1$. However, t_1 is adjusted such that at the end of this timed period the motor is at a point about half way between n and $n + 1$, and it is of course still moving. The current is now reversed again in the same winding for another time period, t_2, bringing about the torque indicated by the solid curve in Fig. 6. This torque decelerates the step motor until it stops, and t_1 and t_2 are timed such that the point at which the motor stops coincides with the new step position, $n + 1$. The current in the same winding is now reversed a third time, producing the motor torque indicated by the dotted curve. This third current setting will hold the motor in the new equilibrium position until one wishes to make another step. This stepping technique produces vibration-free stepping without passive damping. Such an electronic device has been used previously in Bell Laboratories for a magnetic tape drive.

4.2 *A Tuned Two-Degrees-of-Freedom System*

In the previous description of the special motor drive it was tacitly assumed that the stepping motor and all that it drives behaves as a single-degree-of-freedom system, i.e., that the motion of all bodies involved can be determined from a single independent variable. This state exists if such things as backlash, elastic deformation of parts, etc., are negligible. If the time to complete a single step is made sufficiently long, say by decreasing the motor torque, our step system will behave sensibly as a single-degree-of-freedom mechanical system involving only rigid-body motion. However, if the time to complete a step is made short enough as was the case in the pattern generator, one will also excite noticeable motion involving elastic deformation in components of the system. One is then confronted with a much more complicated multidegree-of-freedom mechanical system. Specifically, there was one deformational mode of motion that could not be eliminated. The special motor drive does not then by itself yield vibration-free stepping. One describes here how we were able to control this deformational mode by tuning its natural frequency.

The stepping table is shown in Fig. 7. It consists of a stepping motor driving the shaft of a ball-lead screw, a thrust bearing preventing axial motion of the shaft relative to the rigid base, a step table on linear roller bearing ways and driven by the nut of the lead screw. There are two modes of motion that come into play in this stepping system: (*i*) The motion in which all bodies remain rigid and involving shaft rotation and linear table motion as constrained by the lead-screw pitch. One denotes this mode the ideal rigid-body mode. (*ii*) The mode of motion where the table, as in the first mode, moves as a rigid body on

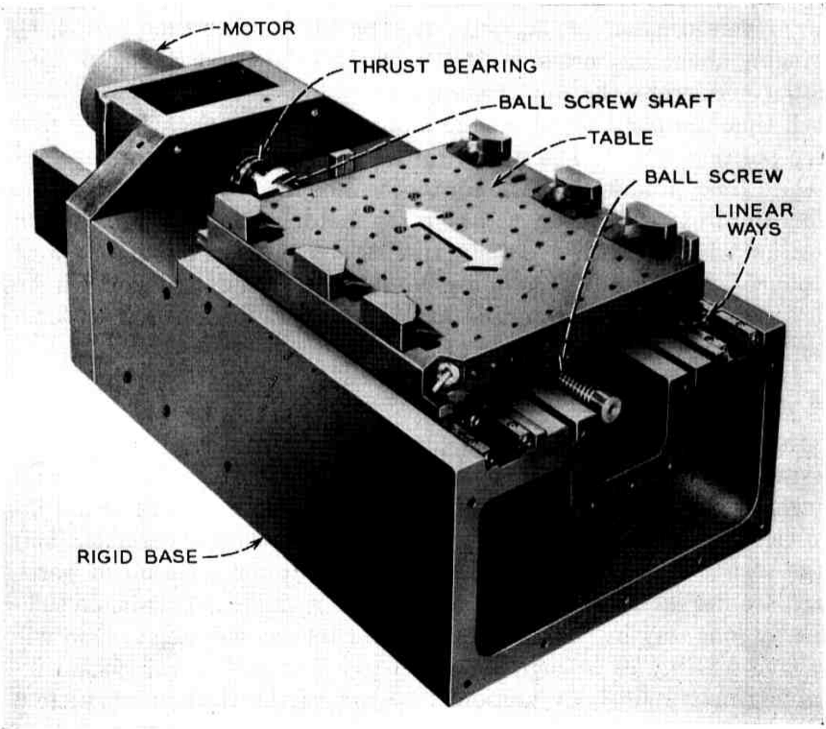


Fig. 7—Stepping system.

its ways but now as a result of elastic deformation primarily of the Hertz type that occurs in the balls and races of the lead screw.

A simple analysis of this two-degrees-of-freedom system reveals an interesting characteristic, namely, that by an adjustment of the natural frequency of the second mode of motion, the special electronic motor drive will step the table with no vibration in either mode. Subsequent experiments proved that such mechanical tuning is a practical matter. In order to describe the essential mechanics involved, some aspects of the simple analysis are given here.

Because of special mechanical characteristics of the step table the two modes of motion mentioned above, namely, the ideal rigid-body mode and the mode involving deformation in the ball screw, are very nearly the normal modes of the system. Therefore, the shaft rotation under the action of the motor torque is sensibly unaffected by the elastic deformation in the ball screw and can be calculated quite accurately, taking only the rigid-body mode into account. The second mode of motion can be equally accurately calculated, taking it to be a single-degree-of-fredom system whose support is given an inexorable motion identical to the table motion associated with the rigid-body mode. The equations for this determination of the first and second mode are

$$T = I\ddot{\theta},$$

$$x_0 = \frac{p}{2\pi}\,\theta,$$

$$\ddot{x} + \ddot{x}_0 = -\omega^2 x,$$

where T is the motor torque, I is the sum of the rotatory inertia of the motor and lead-screw shaft plus an equivalent table rotatory inertia, θ is the angular position of the motor, x_0 is the first-mode table motion, x is the second-mode table motion, p is the lead-screw pitch, ω is the circular natural frequency of the second mode, and dots indicate time derivatives. One assumes now first that the motor torque is a constant over the acceleration period t_1 and the same constant with negative sign during the deceleration period t_2. Secondly, one assumes $t_1 = t_2$ and that the constant torque is selected so that $\dot{x}_0$ is zero when x_0 is increased by one step, i.e., the special drive is adjusted to give no vibration in x_0 at the end of a step. Lastly, one assumes that x and $\dot{x}$ are zero at the beginning of a step. One obtains then for the amplitude of vibration in the second mode, A,

$$A = \bar{x}_0 \frac{\sin^2 (\pi f \bar{t}/2)}{(\pi f \bar{t}/2)^2}$$

where $\bar{x}_0$ is the length of one step, $f = \omega/2\pi$, and $\bar{t} = t_1 + t_2$. One notes now that $A = 0$ when $f\bar{t}/2$ is an integer. According to this simple analysis, there should be no vibration if $f = 286$ cps when $\bar{t} = 7 \cdot 10^{-3}$ s as required in the pattern generator. This frequency corresponds closely to the frequency determined both experimentally and from a more rigorous numerical analysis at which vibration was found to vanish. The vibration amplitude, A, is plotted in Fig. 8 as a function of f. This curve reveals another important point, namely, that where A is zero, the slope of the curve is also zero. For that reason, there is no need to adjust the frequency of the second mode accurately to effectively eliminate vibration, which would have been impractical. One notes that the above solution applies to continuous stepping only when $f\bar{t}/2$ is an integer since only then are x and $\dot{x}$ zero at the beginning of each step. If vibration in x occurs, one has to contend with vibration build-ups from one step to the next.

The rigid-body mode, $f = \infty$, is plotted together with the actual table motion in Fig. 9. The difference between these curves is essentially due to motion in the second mode. One notes that the second mode, as the first, is excited only during the times t_1 and t_2, and no subsequent motion occurs until the table is stepped again.

4.3 *Some Practical Problems of Construction*

Several problems were encountered in the construction of the step table to make it, in fact, behave as the two-degrees-of-freedom system analyzed. A major problem was to reduce the number of degrees of freedom of the system to two. This was done by increasing the natural frequency of the various other modes to a point where the step motion would not noticeably excite them. Our effort in this respect is reflected in the very massive and stiff structure of the pattern generator.

Of particular interest also is the very massive support for the thrust bearing, noticeable in Fig. 7. A particular thrust bearing was selected which enabled us to get rid of a very objectionable third mode of motion in which the ball-screw shaft would move axially by elastically deforming the thrust bearing and its support. A very difficult problem was to find lead screws with a combination of high stiffness of the nuts axial deformation relative to the shaft and low-frictional torque. We found ball-lead screws to be far superior in this respect to lead screws with acme threads.

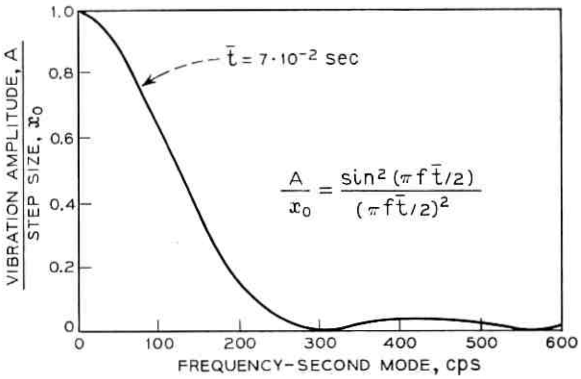


Fig. 8—Vibration amplitude, A, of the second mode as a function of its frequency for a fixed step time, $\bar{t} = 7$ ms.

4.4 *Lead Screw Life Tests*

One of the most critical mechanical requirements of the PPG is that the drive train of the system have a sufficiently long life so that many years of product can be made without changing essential items which would affect the reproducibility accuracy of the system. One sees from Fig. 8 that a drive train-table combination whose stiffness yields a frequency of about 280 cps is desirable. To insure step accu-

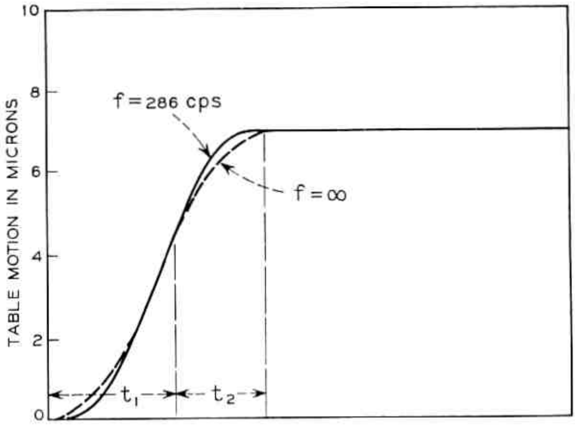


Fig. 9—Ideal rigid-body mode, $f = \infty$, superimposed on the actual table motion, $f = 286$ cps. The discrepancy between the two curves is very nearly the motion of the second mode. One notes that both modes are excited during t_1 and t_2, but no motion persists in either mode once a step is completed.

racy, it was desired that the stepping-system stiffness be great enough to yield a frequency of 280 cps and the frictional torque should be considerably lower than the stepping-motor holding torque so as to minimize step error due to friction. A preload of 25 pounds on the ball screw was found to yield the desired system stiffness and torque to break static friction.

The test setup used to establish the life test of the mechanical components of the drive train is shown in Fig. 10. The life-test setup duplicates the essential features of the PPG drive train.

The status of the life-test equipment was monitored by periodically checking the torque to break static friction and the stiffness of each system. The stiffness was measured by determining the rigid-body resonant frequency of the drive train-table combination and then calculating the stiffness. The stiffness was also checked occasionally by statically measuring the drive-train stiffness by applying a known load and measuring the table deflection relative to the thrust-bearing support.

One sees from Fig. 11, which is typical of the data taken, that there has been a pattern of decreasing torque-to-break static friction. Similarly, from Fig. 12, the stiffness measurements for the units have shown a tendency to increase with time.

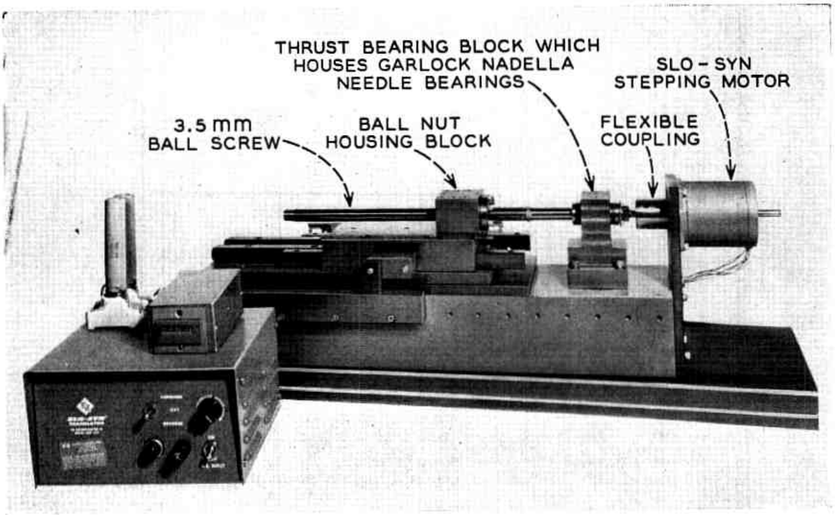


Fig. 10—Typical life-test setup.

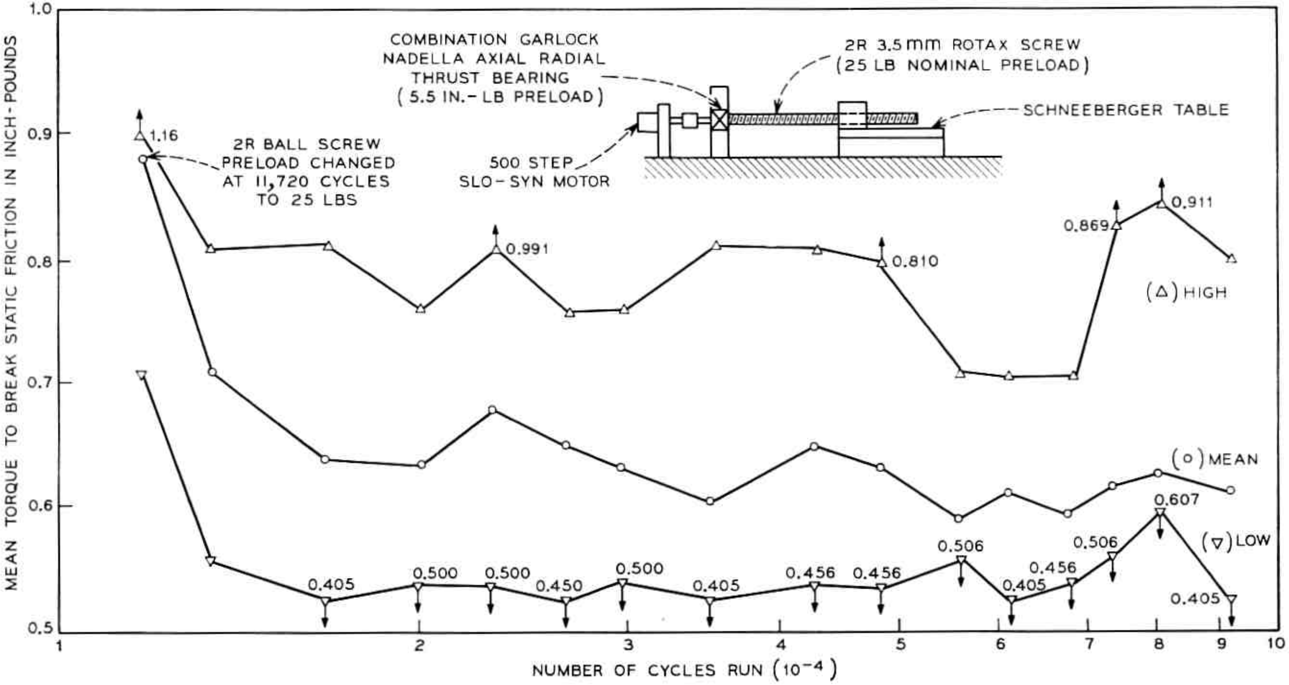


Fig. 11—Mean torque-to-break static friction versus cycles run on 2R life-test setup. Note: 1. Distance traveled per cycle = 19″; 2. Time per cycle = 190 s.

During the life test, a decrease in torque and an increase in stiffness can be attributed to the fact that the screw and bearings are being burnished (i.e., worn in) and hence, the riding surfaces are more uniform and smoother. Furthermore, as things become smoother, more balls of the ball screw and needles of the thrust bearing become fully effective.

4.5 *Stepping Test Measurements*

The accuracy of the step table as determined experimentally is briefly as follows: Steps are reproducible to ±¼ μm. This reproducibility accuracy is primarily the result of some unavoidable coulomb friction in the drive and a small amount of vibration about the equilibrium position. The absolute accuracy of steps is such that all steps are equal within ±½ μm.

Experimental determination of the table motion as a function of time is given in Fig. 13.

Straightness of table travel with minimal transverse and rotary motions is necessary to achieve reproducibility of spot positions on the photographic plate. A table mounted on preloaded roller bearings was employed to achieve the required accuracy. Measurements showed that

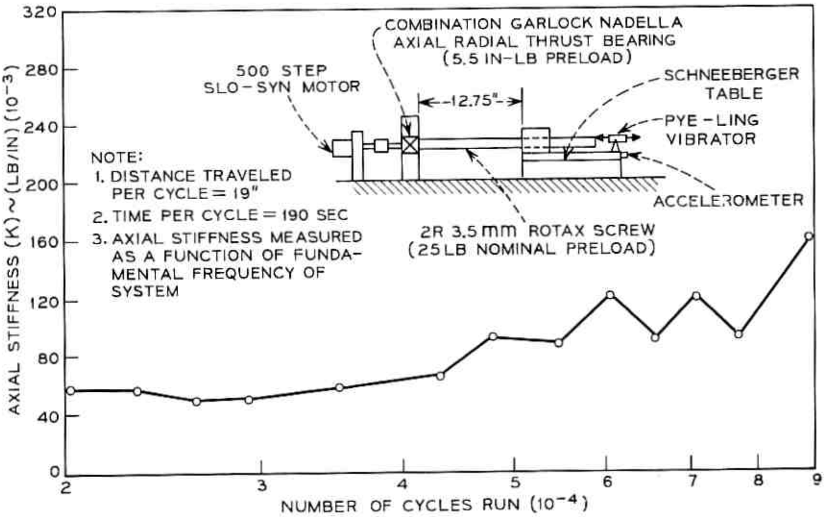


Fig. 12—Axial stiffness of 2R life-test setup versus cycles run. Note: 1. Distance traveled per cycle = 19″; 2. Time per cycle = 190 s; 3. Axial stiffness measured as a function of fundamental frequency of system; 4. Axial stiffness measured directly, using a federal gage to measure table deflection (Δ).

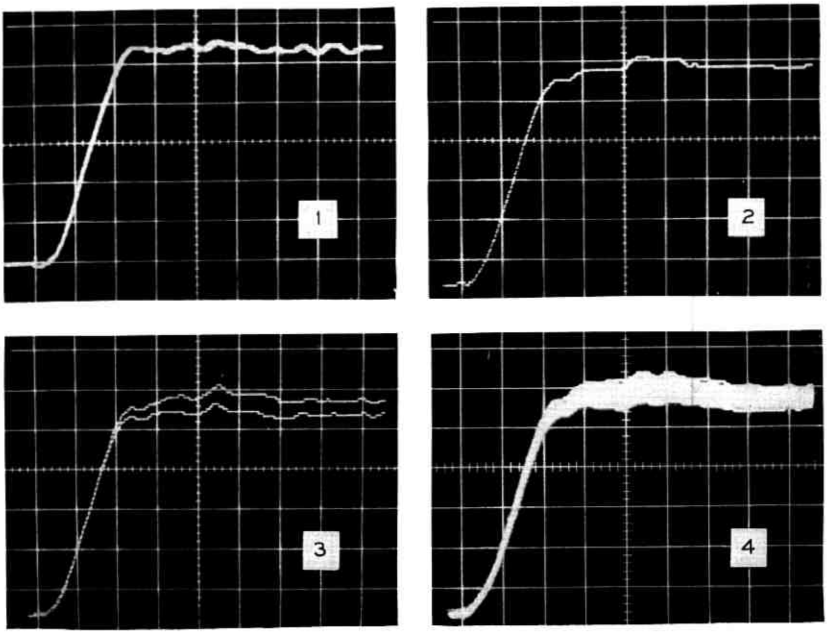


Fig. 13—Table displacement as a function of time. In the above figures, the table displacement was obtained with a laser interferometer having its digital output converted to an analogue output. The scale of the horizontal axis is 2 ms per division, and the vertical axis is 1.34 μm per division. Nominally, the table is to step 7 μm in 7 ms. The very small steps noticeable in the curves are single counts of the laser interferometer representing a displacement of about 0.079 μm. The first two curves each show a single step. The difference between them shows the effect of variations in friction and axial stiffness along the length of the ball screw. The third figure shows two successive steps. The discrepancy between them represents error introduced by the stepping motor. The fourth curve shows 50 successive steps.

the rotational motion superimposed on the translational motion was less than 10 seconds of arc and that the transverse motion was about one micron.

V. ACKNOWLEDGMENT

The authors wish to acknowledge J. W. West's contribution to the fundamental ideas involved in the stepping system and decagonal mirror adjustment.

The Primary Pattern Generator Part III–The Control System

By P. G. DOWD, M. J. COWAN, P. E. ROSENFELD and A. ZACHARIAS

(Manuscript received July 10, 1970)

I. INTRODUCTION

The primary pattern generator (PPG) writing-control system has two main functions: (*i*) interpret the commands generated by the XYMASK PPG postprocessor, and generate from these commands a bit-by-bit image of a scan line and stepping-table control; and (*ii*) check the operation of the PPG system.

The interaction between the PPG and the writing-control system must take place in synchronism with the rotating mirror on the PPG. The writing beam moves continuously across the photographic plate; once a scan has begun, a complete line must be written. One task of the control system is to assemble completely the bit image of a line in a buffer before the start of that scan line. Each line consists of 26,000 bits which must be taken from the buffer in a serial fashion in synchronism with the writing-beam position. A line is scanned in approximately 12 ms; hence, the bit rate during the writing period is 2.2 Mb/s. We thus require real time interaction with a nonstop mechanical device operating at electronic speed.

A complete pattern requires exposure of 32,000 scan lines or approximately 10^9 bits. Accurate operation requires a high degree of system reliability and thorough checking of operations. One check uses parity data generated in the XYMASK PPG postprocessor and regenerated from the signal input to the optical modulator of the PPG. Other checks are on the interface between the control system and the PPG. These checks monitor the operation of the electronics; they proved valuable during fabrication of the system.

II. THE CODE-PLATE SYNCHRONIZING SIGNAL SYSTEM

A block diagram of the computer control system is shown in Fig. 1. The code-plate synchronizing signal is obtained from the photo-

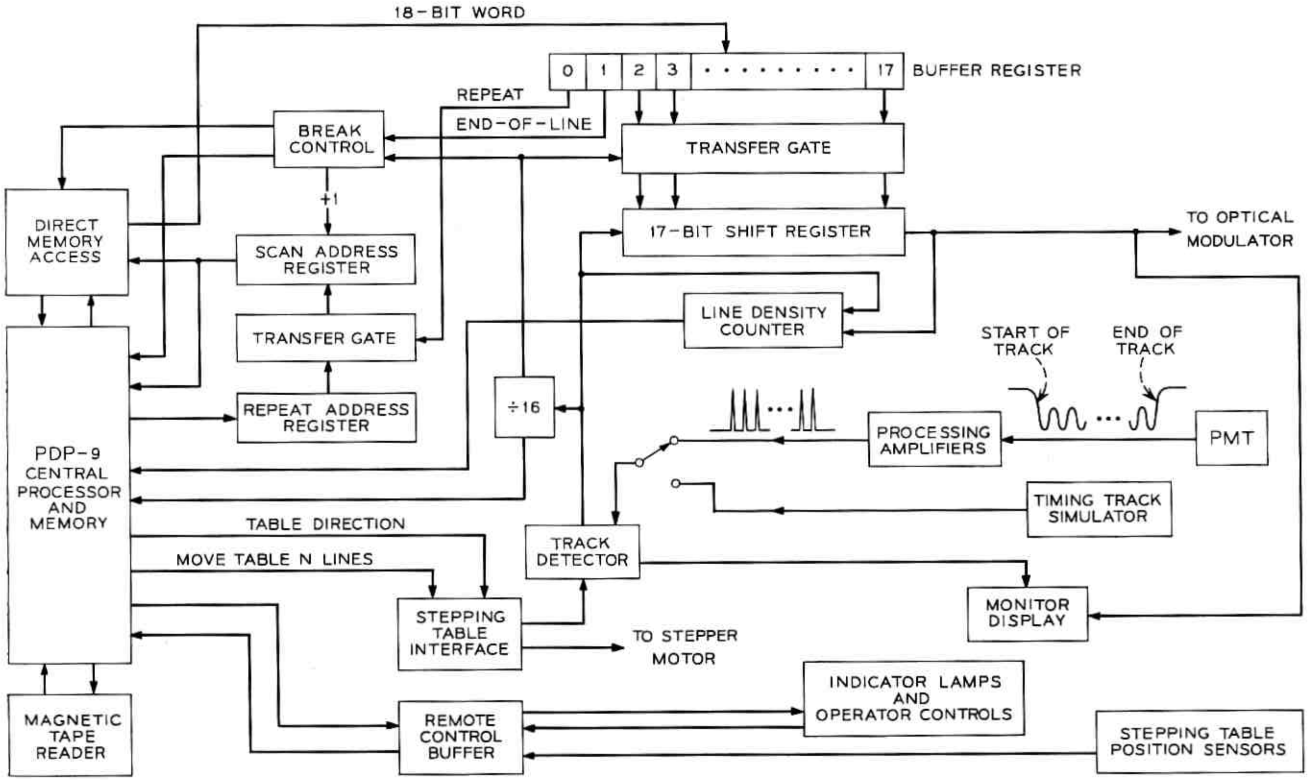


Fig. 1—Block diagram of electronics.

multiplier (PMT) used as the code-beam detector. The PMT output consists of a periodic signal superimposed on a level change. When the code beam begins its scan across the code-plate grating, the output level of the PMT changes with a relaxation time of approximately 400 ns. Similarly, when the code beam finishes its scan and leaves the code-plate grating, the PMT output level returns to zero with the same time constant. The periodic signal superimposed on this average level change represents a modulation index of approximately $\frac{1}{3}$. However, as is seen in Fig. 2 the average amplitude of the PMT output changes quite significantly over the length of the scan. Since we are solely interested in the phase of the periodic component, a limiter with controlled AM to PM conversion is employed before phase detection. The limiter will necessarily drop out during the dead period of each scan. The resulting noise into the phase detector will be unacceptable and so a silencer must be employed. This is accomplished by a gate which is opened and closed by the level changes occurring at the start and end of the scan. The laser power can be changed by a factor of five without affecting the processed output.

Referring again to Fig. 1, the processed output from the code-plate

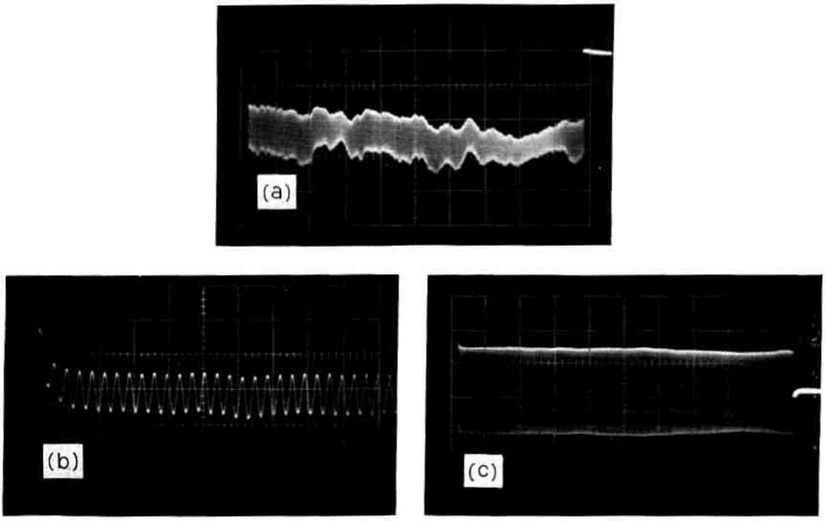


Fig. 2—Code-plate synchonizing signal. (a) Raw signal from the photomultiplier tube showing the entire scan, X sweep ≅ 1.2 ms/box. (b) An expanded view showing the start of the track, X sweep ≅ 2.2 μs/box. (c) Output of the limiter showing the entire scan, X sweep ≅ 1.2 ms/box.

synchronization system is fed to the interface between the PDP-9 control computer and the PPG. The functions of the blocks are best explained by following the sequence of steps that occur when one line is written. Before the line can be written, the bit-by-bit image of that line must be assembled in a core buffer in the PDP-9.

III. THE DATA INPUT SYSTEM AND CONTROL COMPUTER OPERATION

The data read by the PDP-9 control computer is a sequence of magnetic-tape records each of which is 1625 PDP-9 words in length. These records are of two types: one type contains a series of operation commands which define changes to be made to the current scan-line buffer in producing the succeeding scan line. Consecutive lines normally do not differ appreciably in their makeup. Therefore, only a few commands can update a scan line and thus the updates for many scan lines can be held in one record. The second type of input record is used for those instances in which a great number of update commands would be required to produce the succeding scan line. When this condition arises, a record is produced which contains all of the 26,000 bits for the new scan line rather than the update commands which would be required to produce that scan line.

Within the PDP-9 are four buffers of 1625 words each; one buffer holds the current scan-line data and another is the current-command buffer. The other two buffers allow for the overlapping of tape-reading operations with the updating and outputting of scan lines. Therefore, when a new scan-line buffer is requested or the current-command buffer is exhausted, outputting or processing of the next buffer can begin immediately. In the scan-line buffer, the rightmost 16 bits of each 18-bit PDP-9 word are used to designate whether the laser beam should be turned on or off at each of the 26,000 address locations. The magnetic-tape-handling operations were facilitated by making the update-command records the same length as the scan-line buffers.

The following is a brief description of the operation codes used to update the scan-line buffer.

(*i*) Change word N in the scan-line buffer in such a way that a specified, single transition from "beam on" to "beam off" or vice versa occurs. An index to the 32 possible single-transition words is used; this allows the word address and the index number to be packed into one PDP-9 word.

(*ii*) Change M consecutive words to all zeros.

(*iii*) Change M consecutive words to all ones.

(*iv*) Change word M to a specified bit-by-bit configuration. This covers changes entailing more than a single transition.

(*v*) Replace the current scan line with the line described in the next magnetic-tape record.

(*vi*) Write N scan lines identical to the last one.

(*vii*) Skip N scan lines. This allows the rapid coverage of blank areas of the pattern.

(*viii*) Write N consecutive lines of all ones.

In addition to the scan-updating commands, the input tape contains control commands which direct the operation of the PDP-9 control program. These control commands cover such information as: (*i*) the file number of the pattern information on the tape reel; (*ii*) the total number of scan lines in the pattern; (*iii*) addresses for locating the repeat sections of a scan line; (*iv*) end of update commands for the scan line; (*v*) the number of horizontal repeats in a scan line; and (*vi*) identification of the last line of a pattern.

IV. THE INTERFACE BETWEEN THE CODE-PLATE SYNCHRONIZING SYSTEM AND THE CONTROL COMPUTER

When the computer has finished assembling a scan line, it loads the starting address of the scan-line buffer into both the Repeat Address Register (RAR) and the Scan Address Register (SAR) (Refer to Fig. 1). It also sends signals to the break control and track detector telling them that a line may be written. The break control causes a word having address specified by the SAR to be fetched from memory and placed in the buffer register. The SAR is incremented by one so that it now points to the next word in the scan-line buffer. When the track detector finds the start of the timing track, it opens a gate and allows timing pulses to pass to the 17-bit shift register and the divide-by-sixteen counter. Each timing pulse causes the bits in the shift register to be shifted right one place. The output of the last stage of the shift register is used to control the laser writing beam, turning it on if it is a "one" and off if a "zero." The divide-by-sixteen counter produces an output pulse at each 16th timing-track pulse. This pulse causes the contents of the buffer register to be transferred to the shift register. This pulse also causes the break control to fetch another word from memory and deposit it in the buffer register. The line density counter counts the number of "ONES" that are shifted out of the shift register. This count is used in error checking.

The process of transferring words from memory continues until a word which contains a "ONE" in either bit position 0 or 1 is loaded into the buffer register. A "ONE" in bit 0 signals that the portion of the line being written is to be repeated. Therefore, the contents of the RAR are transferred to the SAR and the next word fetched by the break control will come from the location in the scan-line buffer specified by the RAR. A "ONE" in bit 1 signals that this is the last word in the scan-line buffer for this scan line. The break control logic is disabled and ignores any further pulses from the divide-by-sixteen counter. In addition, the control program is notified that the end of line has been reached; the track detector notifies the program that the end-of-track has been reached when that event occurs.

Anytime after the end-of-line is reached, the control program can command the carriage to be moved. This is done by transferring a word to the carriage control logic that specifies how many steps the carriage is to be moved. If the carriage is to be moved one line, the carriage control logic will cause the stepper motor driver to deliver the sequence of steps required to cause the carriage to move and be stopped within the 7 ms allowable time. Thus, if the next line is assembled in the core buffer of the PDP-9, the line will be written by the succeeding mirror facet. If the carriage is to be moved more than one line, then the number of lines less one to be moved must be all blank, and so carriage motion can be carried out asynchronously at high speed. After the last line is stepped, synchronism is regained by the operation of the track detector. The last line is always output by the carriage control logic as if only one line were to be moved. This effectively stops the carriage 7 ms after the last line command is issued.

The remote control buffer is used to provide communication between the operator and the computer. It consists of a flip-flop register and lamp drivers for signaling the operator, and gates to allow the computer to sense the pushbuttons the operator uses to signal it.

V. ERROR DETECTION

There are a number of safeguards in the control program which check on both hardware and software types of errors. Error detections are transmitted to the operator via teletype and light signals. Most errors are fatal and necessitate the restarting of the pattern. When this type of error is encountered, the current run is aborted and the photographic plate is unloaded from the machine. Some errors occur before the pattern is begun and, in these cases, the operator is advised but no unloading takes place.

Most hardware errors are detected in two major ways. One is a count of the number of pulses the code plate synchronizing signal system send to the track detector. If this deviates by more than ±1 pulse, then a fatal error is detected. Another track check is the occurrence of end of track before the end-of-line word has been written. The second way that hardware errors are checked is by the line density counter. If some malfunction occurred, then the number of ONES in the line written will not agree with the control command specifying the number of ONES in that line. This line-density count checks not only the functioning of the interface hardware, but also the PDP-9 assembly of the line image in the scan-line buffer. Other hardware errors are checked by comparing the SAR value at the end of a scan line with the value the SAR should have after the line is output. The carriage control is checked by comparing the reading of a shaft encoder on the stepper motor with the required reading after the pattern is completed. This shaft encoder gives an indication of its position only once in 500 lines, so continuous monitoring is not feasible. However, if the reading at the end of the pattern is not correct, then indication of a carriage error is given to the operator.

Software and magnetic-tape errors are detected by program routines in the PDP-9. Illegal update commands, magnetic-tape reading errors and other magnetic-tape controller errors are the main errors detected by these routines.

The Primary Pattern Generator Part IV–Alignment and Performance Evaluation

By A. M. JOHNSON and A. ZACHARIAS

(Manuscript received July 10, 1970)

I. REQUIREMENT FOR ALIGNMENT

The mechanical nature of the primary pattern generator (PPG) requires a precise juxtaposition of most of the machine elements in order to achieve both pattern accuracy and reliable functioning of the machine. Part II described the alignment of the rotating polygonal mirror to the air-bearing axis. The precision required in that assembly is the tightest tolerance in the PPG. This precision is required to produce a uniform scan-line spacing on the pattern. In addition, the direction of that scan line must be made as perpendicular as possible to the travel direction of the photographic plate. Therefore, the carriage of the photographic plate must move without rotation. The method for aligning the polygonal mirror axis to the carriage direction will be described, as well as other alignment needed to produce an accurate pattern. The code-plate system for controlling the fast scan was described in Parts I and III. Implicit in this description was the assumption that the code-plate grating and the photographic plate are the exact same distance from the scanning lens (see Fig. 1 in Ref 1). The positioning of the code plate to achieve accurate length of the fast scan is a critical alignment that requires a combination of optical and electronic techniques.

The accuracy goal for the PPG was 100 parts per million (ppm) deviation from an absolute coordinate system, the error reference being the overall dimension of the full PPG field. Thus the coordinate axes of the pattern must be othogonal to within 20 seconds. A second of arc is approximately 5×10^{-6} rad. The photographic-plate position is determined by a lead screw as described in Part II. The accuracy of this screw is the determining factor in the overall length error of the plate translation axis. For convenience, we will refer to this axis as the Y-axis and the fast scan axis as the X-axis.

The functional alignment includes positioning the optical modulator, obtaining separation of the coding and writing beams, positioning of the scanning lens, and positioning of various other lenses and mirrors in the optical paths of the two beams. The design and alignment of the laser cavity is described. The long-term functioning of the PPG will require replacement of the laser discharge tubes. Our design-and-alignment procedure allows tube replacement without realignment of the remainder of the optics.

II. FUNCTIONAL ALIGNMENT

The quartz laser tube is clad with a water-cooling jacket and is rigidly mounted within a solenoid which provides the axial magnetic field. By placement against pins, this assembly is located precisely on a flat plate on which the cavity mirrors are rigidly mounted. This system was devised so that a remotely located reference cavity can be used to prealign a laser-tube-solenoid assembly to the laser cavity on the PPG. The use of the reference cavity significantly reduces the down time of the PPG during laser replacement; replacement of the laser does not require realignment of the PPG.

The laser cavity is of a nearly hemispherical configuration consisting of a 0.9-m radius highly reflecting mirror and a flat, transmission mirror at the output. The separation is 0.75 m. The output is constrained to the TEM_{00} mode by using a 2-mm aperture inside the cavity near the spherical mirror. The 514.5-nm line is selected by the transmission characteristic of the output mirror.* The output mode of the laser has a $1/\epsilon$-amplitude radius[2] of 200 μm. The train of lenses and mirrors (see Parts I and II) which is used to direct the laser output to the optical modulator was aligned by autoreflection at each mirror. The lenses were inserted after the beam had been correctly positioned. Back reflections from each lens were used to center accurately that lens.

The optical modulator must be positioned to the Bragg angle.[3] The angle is set by periodically exciting the modulator and then detecting the deflected beam with a photodetector and maximizing the modulation. After the modulator is positioned, the writing-beam separation

* The reflective band of the transmission mirror is centered near 550-nm wavelength. The edge of the band is at 514.5 nm and thus the reflectivity at all the other spectral lines is insufficient for oscillation.

mirror (see Part I) is positioned. A 10-cm focal length lens placed at the modulator output produces the required spatial separation of the writing and coding beams. At the separation mirror each beam has a $1/\epsilon$-amplitude radius of 50 μm and the center-to-center beam spacing is 400 μm. At this location, the coding beam is 20 to 50 times the intensity of the writing beam. The light from the coding beam which is scattered in the writing beam direction is removed by an 0.75-mm aperture placed concentric with the writing beam. Slight tilting of lenses eliminates objectionable back reflections. After these adjustments, the on-off ratio of the writing beam is greater than 50.

III. ACCURACY ALIGNMENT

The path of the writing beam from the modulator to the scanning lens (see Fig. 1 of Ref. 1) is determined by three adjustable mirrors in addition to the writing beam separation mirror. These three mirrors are used to properly direct the writing beam into the scanning lens. However, the proper position of the scanning lens is determined partly by the positions of the rotating mirror and photographic plate. Consequently, the rotating mirror must first be aligned to the photographic plate; then the writing-beam illumination of the scanning lens can be set and finally the scanning lens is positioned.

The alignment between the rotating polygonal mirror and the translational direction of the photographic plate (Y-axis) is accomplished by use of a precision cube and an autocollimator. The cube is mounted on the photographic-plate carriage in such fashion that a cube face is normal to the Y-axis. Errors are introduced by the yaw, pitch and roll of the carriage; each contributes a few arc seconds of error. First, two faces of the cube are indicated parallel to the Y-axis by using sensors capable of detecting $1/40$ μm displacement. The cube face normal to these two faces is normal to the Y-axis. The X-axis of the pattern is the intersection of a plane normal to the axis of rotation of the polygonal mirror (this plane is also normal to all of the facets of this mirror) and the plane of the photographic plate. The plane of the photographic plate must be parallel to the Y-axis or else the X-axis as defined above will not always be in the focal plane of the scanning lens. A sufficient, but not necessary condition for the X-axis to be normal to the Y-axis is to make the carriage travel direction parallel to the rotation axis of the polygonal mirror. This is accomplished by using an autocollimator to set the reference face of the polygonal mirror (the reference face is perpendicular to all the facets of the mirror) parallel to the face of the precision cube which is normal to the Y-axis.

The actual angle between the X- and Y-axes was determined by generating a test pattern on the PPG and measuring this pattern with a coordinate-measuring machine (CMM).[4] This measuerement could be made with an error of less than 3 s. Thus, a correction to the direction of the rotating mirror was determined and used to reset the X-axis. Since this correction was less than 20 s, no other alignment was disturbed.

After the initial positioning of the rotating-mirror axis, the writing beam must be directed to the center of the entrance pupil of the scanning lens. This is set by autoreflecting the writing beam from a properly positioned polygonal mirror facet. The proper angle of the facet is calculated from the parameters of the scanning lens. The polygonal mirror facet is exactly positioned by the use of an autocollimating theodolite. The position which must be taken by the axis of the scanning lens is now fully constrained. This position is duplicated by a helium-neon laser beam which is positioned normal to a facet of the polygonal mirror. This facet is first set parallel to the X-axis. The He-Ne laser beam is also passed through the center of the scan line on the photographic plate. The scanning lens is positioned by centering its back reflections of the He-Ne laser beam thereby aligning the axis of the scanning lens with the He-Ne laser beam.

The last step in the X-axis alignment is the length-accuracy adjustment of the code-plate position. To accomplish this, a replica of the code-plate grating is produced by contact printing onto a photographic plate. This plate is then positioned in the PPG in exactly the manner a photographic plate is positioned when it is to be exposed. A long, silicon PIN photodetector is placed under the replica grating. The focused writing beam will produce a signal output from the PIN photodetector as it sweeps across the replica grating. However, the long photodetector has very little bandwidth. To circumvent this photodetector deficiency, the output of the actual code plate is used to modulate the writing beam by feeding the code plate signal into the optical modulator. Now the long photodetector under the replica grating will only have to respond to the beat frequency between the code-plate signal and the writing beam sweeping the replica. By adjusting the beat frequency to zero throughout the scan, the exact position registration between writing and coding beams is obtained. This method of alignment resulted in less than 10-ppm error in the X-axis length. Residual errors are caused by camber of the photographic plates (see Part I), inevitable temperature variations, and camber in the coding-beam output mirror (see Fig. 1 of Ref 1).

IV. PERFORMANCE EVALUATION

The design and fabrication of the necessary high-frequency mechanical components allowed the synchronization between the fast scan and the photographic-plate translation to be accomplished by a simple, computer-controlled system. Further, this step-on-command system allows flexibility in the computer control so that future work can produce a more economical division of work between the PPG control computer and the PPG postprocessor.[5] At present, very few of the patterns drawn by the PPG have required the machine to wait for the computer to finish assembly of a line.

The rotating mirror presented the most critical item in terms of tolerance. The periodic bunching and spreading of the scan lines caused by the nonideal mirror results in both a periodic variation in the optical density of exposed regions and a periodic displacement in feature edges which are parallel to the Y-axis. The optical density variation is lost when the pattern is photographed by the reduction cameras. However, the periodic displacement is still detectable after the first reduction; the peak-to-peak amplitude is less than one-third address.

The major inaccuracy in the PPG is the Y-axis length. The lead screws used are accurate to within 15 ppm at 20°C. However, the lead-screw temperature in the operating machine is 25°C and so the Y-axis length is in error by 90 to 100 ppm. However, the lead screws can be replaced and this error can be eliminated.

The measured reproducibility of the PPG cannot be separated from the reproducibility of the coordinate-measuring machine. It was found that remeasurement of a PPG plate on the CMM produced readings which showed a variance of one-third address at the extremes of the pattern field. Near the CMM reference point in the pattern, the variance of the readings was approximately one-sixth address.[4] Such behavior indicates a systematic error such as that caused by temperature differences. If the reproducibility of the CMM is accounted for, the variance in the location of a PPG-produced feature is not greater than one-third address and may be less than one-fourth address. Figure 1 shows the measured scatter of identical features drawn on 18 separate plates made over a period of two months. The (X, Y) address location of the CMM reference was (1000,1375) in the PPG field. The scale on the axes of the scatter plots are in addresses with respect to the absolute coordinate. Note the error increase in Y caused by the excess length of the Y-axis.

The PPG, as constructed, meets all of the requirements set by the mask-making system.[6]

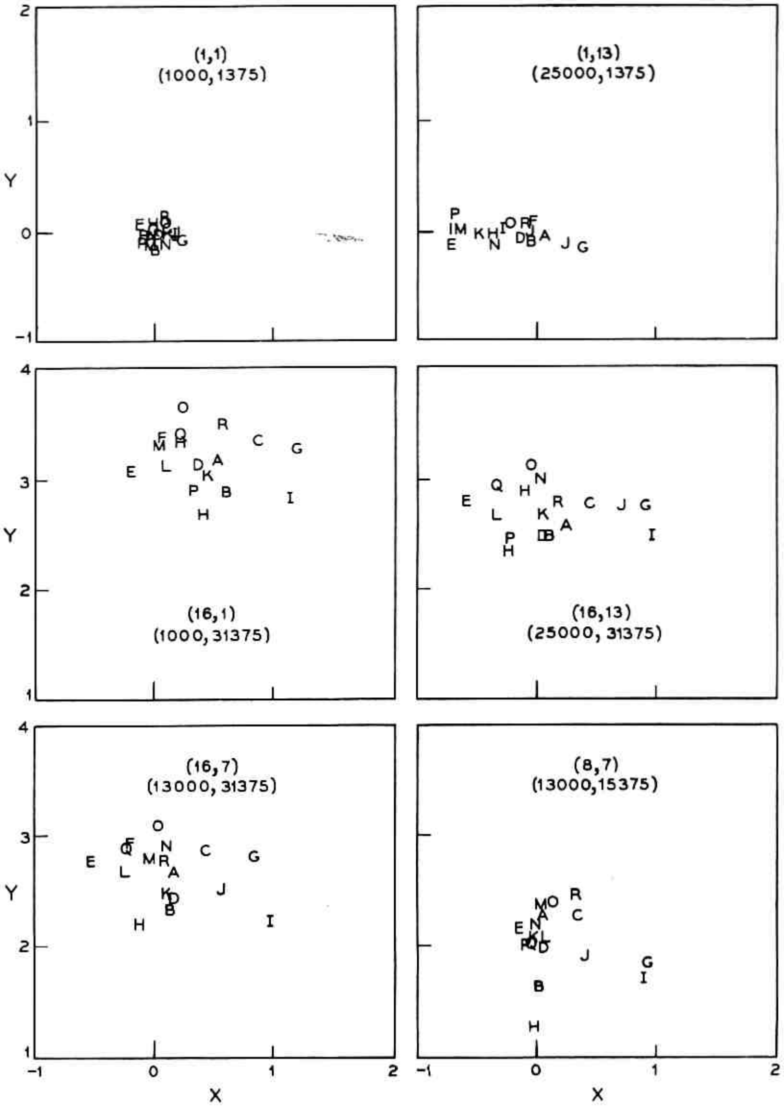


Fig. 1—Reproducibility of pattern generator.

V. ACKNOWLEDGMENT

A. D. White computed the illumination angle for the scanning lens.

REFERENCES

1. Cowan, M. J., Herriott, D. R., Johnson, A. M., and Zacharias, A., "The Primary Pattern Generator, Part I—Optical Design," B.S.T.J., this issue, pp. 2033–2041.
2. Kogelnik, H., "Imaging of Optical Modes—Resonators with Internal Lenses," B.S.T.J., *44,* No. 3 (March 1965), pp. 455–494.
3. Gordon, E. I., "A Review of Acoustooptic Deflection and Modulation Devices," Proc. IEEE, *54,* No. 10 (October 1966), pp. 1391–1401.
4. Ashley, F. R., Murphy, Miss E. B., Savard, H. J., Jr., "A Computer Controlled Coordinate Measuring Machine," B.S.T.J., this issue, pp. 2193–2202.
5. Gross, A. G., Raamot, J., Watkins, Mrs. S. B., "Computer Systems for Pattern Generator Control," B.S.T.J., this issue, pp. 2011–2029.
6. Howland, F. L., and Poole, K. M., "An Overview of the New Mask-Making System," B.S.T.J., this issue, pp. 1997–2009.

Device Photolithography:

The Electron Beam Pattern Generator

By W. SAMAROO, J. RAAMOT, P. PARRY and G. ROBERTSON

(Manuscript received June 8, 1970)

An electron beam pattern generator is being developed to write directly on photographic plates with a 4-μm diameter beam over a 5-cm by 5-cm field with an address structure of 25,000 by 25,000. Two unique features of this pattern generator are random-access computer control of the beam and a 15-bit digital-to-analog converter stable to better than ±1 part in 10^6. Capability for drawing 4-μm lines having an edge gradient less than 0.5 μm and an optical density greater than three has been demonstrated. Stability of better than ±1 μm in 24 hours over a 4-mm by 4-mm field has been achieved. Experiments still in progress have demonstrated ±1-μm stability over the entire 5-cm by 5-cm field for shorter time periods. Reticles of typical complexity are drawn routinely in less than five minutes.

I. INTRODUCTION

The demand for integrated circuits is increasing rapidly, and projections indicate that the existing mask-making facilities will be severely overloaded in the near future. The major portion of the time required to make a mask is taken in producing the reticle. The electron beam pattern generator was originally conceived to assist the mask-making shop by producing reticles rapidly.

The use of a computer-controlled electron beam also holds promise for solving other problems. As integrated circuits become more complex, it is increasingly difficult to meet the line-width and field-size requirements of the final masks. The fundamental limits set by diffraction effects are currently being approached; moreover, the depth of focus of the lens system producing these masks is so small that severe requirements are made on material tolerances. Due to the extremely short wavelength of kilovolt electrons, diffraction effects are negligible and it is possible to write with beams a few tenths of a micron in width over small fields. A. N. Broers et al.[1] have succeeded in producing interdigital surface-wave transducers of 0.3-μm width and 0.7-μm spacing using a modified form of scanning electron microscope. With electron beams it is possible to use very high *f* numbers to give a large depth of focus; this relieves the problem of extreme materials tolerances.

This paper describes an experimental machine built to prove the feasibility of drawing reticles on photographic plates. The requirements to be met are as follows. The address structure should be greater than 25,000 by 25,000 over a 5-cm by 5-cm field. The line should be 4 μm (two address units) wide and have an optical density greater than two. Stability and reproducibility should be within ± 1 μm, or ± 20 ppm (parts per million). As the machine has fast random access rather than a raster scan, the writing time is proportional to the area covered, and the machine should be able to cover 20 percent of the field within five minutes.

Although the above requirements are sufficient for all of the expected reticles for the next few years, they are also sufficient for about 90 percent of the expected masks. The specifications allow a 4-μm feature of one mask level to be registered within an 8-μm feature of another level. Because of the method of programming the computer, there is very little extra computation time involved in drawing a mask as compared with a reticle.

The electron optical column was built from commercially available parts, and does not represent the ultimate in performance. However, it has been demonstrated that the electron beam is potentially a very valuable tool in the manufacture of reticles and integrated-circuit masks.

II. DESCRIPTION OF THE SYSTEM

Figure 1 shows a block diagram of the equipment. A Digital Equipment Corporation PDP-9 computer is used to generate data for an interface to control the electron beam. Input information to the computer is obtained from design programs such as XYMASK.[2] The division of work between the computer and the interface is best explained by describing the technique used to draw patterns.

As shown in Fig. 2, patterns are drawn using line segments rather than picture points. The line segments may be up to 256 addresses long and the patterns are filled in at a rate of 1 μs per address. This is an important aspect of the system as it allows 20 percent of a 25,000 by 25,000 address structure to be covered in less than three minutes.

To draw the unfinished feature shown in the blowup of Fig. 2, the

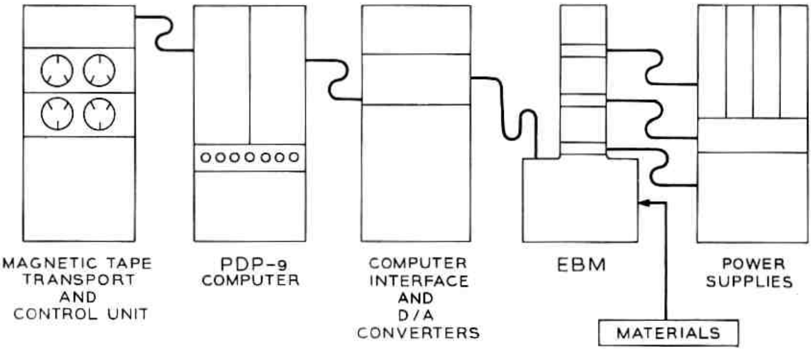

0000000
MAGNETIC TAPE TRANSPORT AND CONTROL UNIT
PDP-9 COMPUTER
COMPUTER INTERFACE AND D/A CONVERTERS
EBM
POWER SUPPLIES
MATERIALS

Fig. 1—Block diagram of the equipment.

coordinates of the four indicated points are read into the computer. The start point and length of a single line segment are fed to the interface. While the interface is controlling the beam to draw that particular line segment, the computer is calculating the position and length of the adjacent line segment. This division of work between the computer and interface minimizes the amount of information to be supplied to the system and gives the system a programming flexibility as will be discussed later.

Both familiar forms of graphics, a television like raster and point-by-point plotting, were rejected for this system. A raster-type genera-

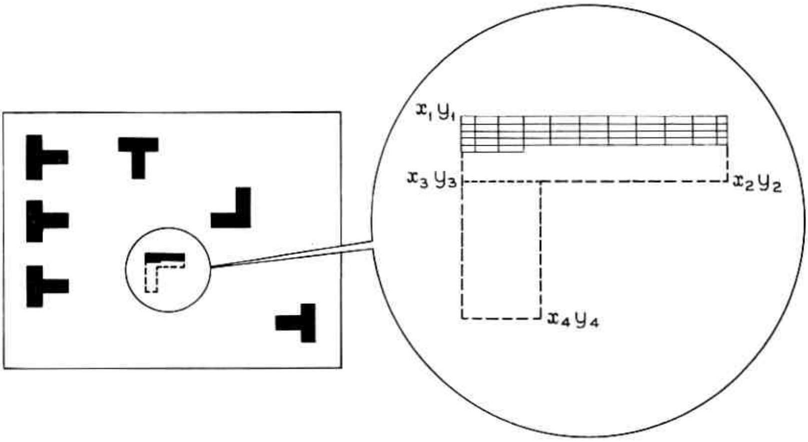


Fig. 2—Method of filling blocks using line segments.

tion is very inflexible in comparison to a random-access system and requires the transmittal of large quantities of information which is generally obtained on a large computer, thus incurring additional costs. Point-by-point plotting under the control of the computer would make generation times impractically long. Although the pattern generator uses line segments, it is a true random-access machine since a line segment may be one address long.

The Electron Beam Machine (EBM), with its associated power supplies and the photographic materials which go into the work chamber, will be discussed in detail in following sections.

2.1 *The Electron Beam Machine*

In the EBM, a beam of electrons writes directly on photographic plates thereby utilizing the good resolution inherent in electron beams. As in the case of CRTs, where the overall resolution is dependent on the phosphors, the resolution of the EBM is limited by the recording medium. For reticle and mask generation, where Kodak High Resolution Plate (HRP) emulsion is used to achieve the desired plotting speed, the resolution or edge definition of a line is limited to about 0.5 μm.

The electron optical column consists of a triode gun with a reentrant Wehnelt cylinder, two demagnifying lenses and one projection magnetic lens. The two demagnifying lenses are used to produce a 4-μm-diameter image just below the second lens, and the projection lens reproduces this image at a 35-cm working distance. The long working distance makes it possible to scan a 5-cm field with deflection angles of less than 5°. An electrostatic deflection system is used in preference to a magnetic system, in which eddy current and hystersis in the chamber walls would reduce the speed and accuracy to below acceptable limits. Because of the small deflection angles and apertures used, deflection defocusing with the electrostatic system presents no problem.

The current of the 15-kV beam is in the range of 0.1 nA to 1 nA. The large f number of the final lens (8000) enables a 4-μm beam to be achieved using commerically available lenses,* and keeps aberrations to negligible proportions. The operating pressure of 10^{-5} Torr is produced by a liquid nitrogen trapped 4″ oil diffusion pump.

2.2 *Control of the Beam*

As was indicated before, patterns are composed from line segments, which may be up to 256 addresses long. Figure 3 shows a block diagram

* Made by Canal Industrial Corp. (Canalco), Rockville, Maryland.

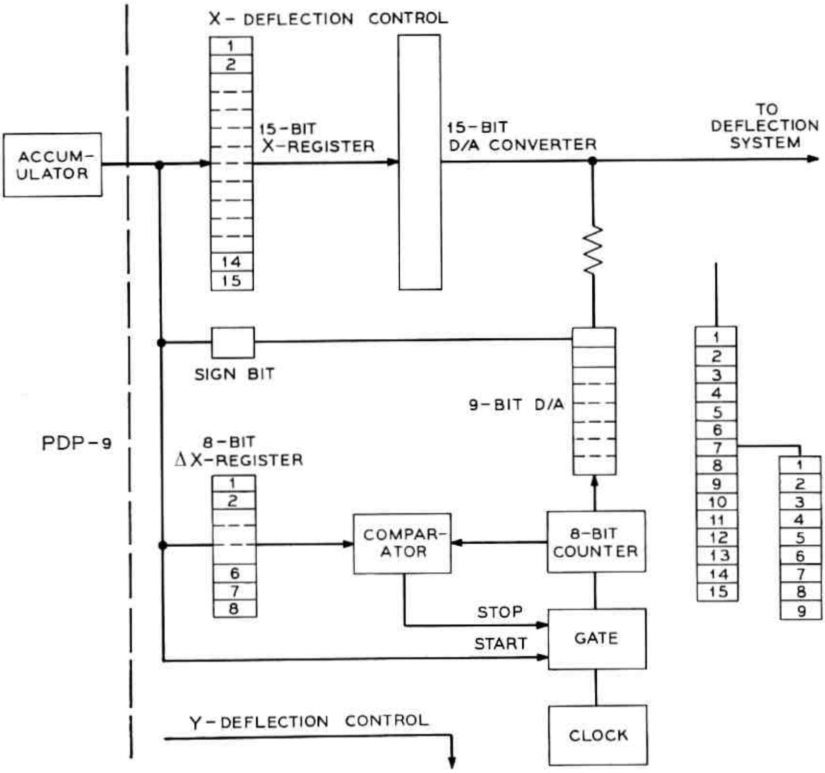


Fig. 3—Block diagram of computer interface.

of the computer interface which controls the beam during the generation of the line segment. Only the x-axis control is shown; the y-axis control is identical.

To draw a line segment $\Delta X(\Delta X \leqq 256)$ addresses long in the x direction starting at (x_1, y_1), x_1, ΔX and y_1 are loaded into the X-register, ΔX register, and Y-register, respectively. The beam is blanked at this time but the voltages corresponding to x_1 and y_1 are generated by the X and Y Digital-to-Analog Converters (DACs) and are applied to the deflection system. Therefore, when the beam is unblanked, it is at (x_1, y_1). A start signal from the computer unblanks the beam and opens the gate to allow a continuously running clock to increment the 8-bit counter, which is initially set in the zero state. The output from the 8-bit counter is converted to an analog signal by the eight less-significant bits of the 9-bit DAC. The output of the 9-bit DAC is attenuated by a factor of 64 and is added to the output of the 15-bit

DAC. In this way, a voltage ramp is generated to move the beam from x_1 to $x_1 + \Delta X$. The comparator compares the ΔX-register with the 8-bit counter and turns off the gate and blanks the beam when they are equal. The most significant bit on the 9-bit DAC allows lines to be drawn in both positive and negative x and y directions.

This method of generating line segments offers a number of advantages. It has already been mentioned that locating each address with the computer makes the generating time impractically long. With the interface described, the line segment is generated at 1 μs per point. If the X-register were incremented directly from the clock pulse (thereby eliminating the 9-bit DAC and the 8-bit counter), then, depending on the initial state of the X-register, some of the more significant bits will change states. When this happens, large transients of the output voltage will occur. The use of the ΔX converter system insures that only the less significant bits are switched while the beam is on. The sketch inserted to the right of Fig. 3 is meant to convey this idea. Experimentally it has been found that switching transients in the 9-bit DAC appearing at the deflection system are within tolerable limits.

The output of the DAC drives the deflection plates directly. At the beginning of a line segment, before the beam is unblanked, 10 μs are allowed for the output to settle. While the output is settling, the interface is loaded, which takes 11 μs. At the end of a line segment it is necessary to wait only 1 μs after the counter stops before blanking the beam.

2.3 *High-Precision Digital-to-Analog Conversion*

The best commercially available DACs have 13-bit resolution with 0.01-percent accuracy. There are DACs with 15-bit resolution and 0.01-percent accuracy, but these are not consistent with DAC because the lesser significant bits are not reproducible.

Normal practice in DAC is to switch accurately controlled voltages through precision resistors using transistor switches for their speed. The resistors form a binary series and the currents from the resistors are summed through a load. A simplified sketch of a 3-bit DAC of this type is shown in the left side of Fig. 4. It is because of the instabilities across the transistor switches that only 0.01-percent stability can be achieved.

DACs have been developed for this system whereby the voltage is regulated after switching as shown in the right side of Fig. 4. Notice that parallel current-source regulation is used instead of series-

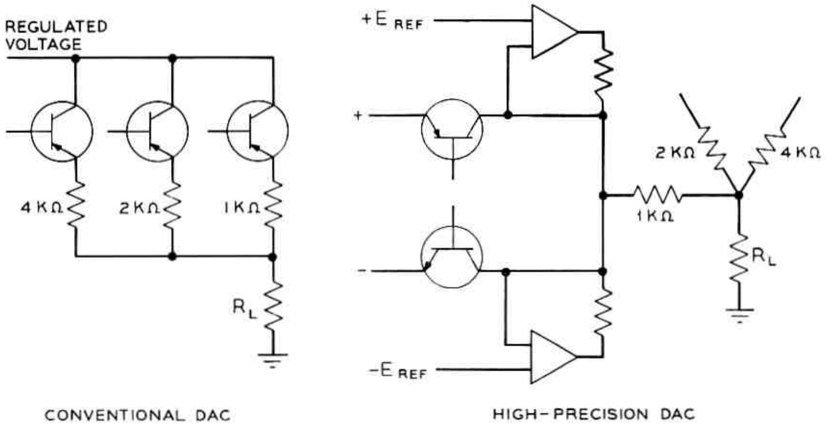


Fig. 4—Simplified diagram of conventional and high-precision DACs.

voltage regulation. Operational amplifiers which act as current sources compare the switched voltages with a reference and compensate for any variation by either supplying more or less current to the load. Details of this circuit have been described elsewhere.[3]

Using such a circuit, it has been shown experimentally that after the switch the voltage can be regulated to ±0.0001 percent or ±1 ppm. Only the more significant bits need be built in this fashion. An 18-bit system was built, in which 13 bits are conventional and only the five more significant bits are regulated in this way.

The secondary reference source used for comparison includes a primary standard and was also made using the principle of parallel-current source regulation. The resistors in the most significant bits and in the secondary voltage standard are stable to ±1 ppm/°C.*

An 18-bit DAC was built and tested successfully, but only the 15 most significant bits are used in the present application. One of the measurements made to test the DACs is illustrated in Fig. 5. The input to one DAC was held constant while the input to a second one was incremented and the output voltages were summed and recorded. The arrow indicates the step which resulted when the most significant, or first, bit was turned on and all other bits were turned off. Each step corresponds to a change of the least significant bit and equals a 4-ppm change in the total deflection voltage. It was found that the combination of the 13-bit DAC and the high-precision 5-bit DAC was cali-

* Resistors manufactured by Julie Research Laboratories, Inc., New York, New York.

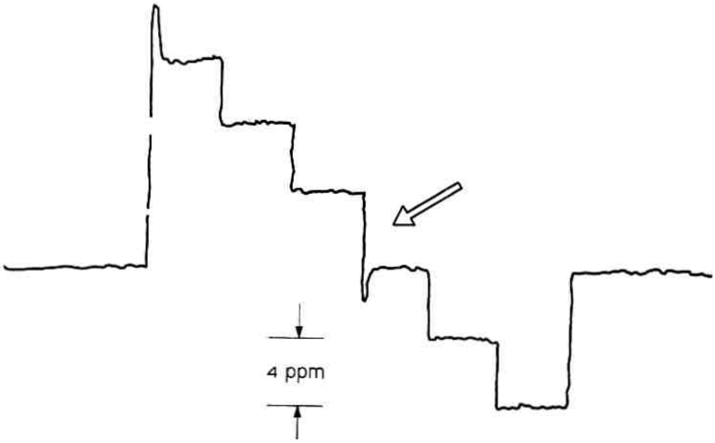


Fig. 5—Output voltage of the 18-bit DAC showing the matching of the most significant bit switch within a structure of the least-significant bits.

brated to better than $\frac{1}{2}$ ppm and that the least significant bit was clearly resolvable. Other similar measurements have shown that the 24-hour stability of the 18-bit DAC is ±1 ppm.

2.4 *Programming*

Because the system has random access and because of the particular division of work between the computer and its interface, programming of the electron beam pattern generator is simplified and results in elegant solutions. However, the biggest speed factor is the small amount of input information required by the system in comparison with other graphics systems.

XYMASK is a rather general program and the output of this design program must be "postprocessed" for the particular pattern generator. Since the computer of the electron beam pattern generator is available for large portions of the drawing time, a considerable amount of the computation normally done in postprocessing is done in real time in the PDP-9.

It has been shown how the electron beam pattern generator fills in rectangles at the rate of one address per microsecond. Circles, as well as complex geometries bounded by quadratic functions and lines of any given slope, may also be filled in at the same rate. This was made possible by "integer arithmetic," which avoids the use of multiplication and division in determining the boundaries of those geometries.[4] Algorithms based on integer arithmetic, which are run during the free time on the PDP-9, allow the programming of complex geometries in real time. The programming is described in detail elsewhere in this issue.[2]

2.5 *Electron-Sensitive Materials*

The photographic plates used in the EBM consist of 6 μm of Kodak HRP emulsion on a flat glass plate coated with a thin layer of chromium. The glass flatness is better than ±0.27 μm per linear cm and the optical density of the chromium layer is 0.04. The chromium layer has a resistivity less than 1000 Ω per square and is used to dissipate the charge of the electrons.

An interesting feature of electron beam exposure of these plates is that the image formed occurs in about the upper micrometer and a half of the emulsion. For projection exposure of the patterns produced by the electron beam system, this reduces the depth-of-field requirement on the projection system. Experiments on electron sensitization of photoresists have been performed at the Western Electric Engineering Research Center, Princeton, New Jersey, and are described in the following paper.[5]

III. STABILITY AND CALIBRATION OF THE SYSTEM

When drawing a mask, it is essential to maintain stability for at least the time required to complete the pattern, typically five minutes. In order to insure registration of mask levels made at different times, it is essential to be able to maintain calibration over a long period of time. The EBM has been designed to have a short-term stability and long-term recalibration capability of better than ±1 μm or ±20 ppm over the entire field. This section contains a discussion of systematic and random errors and a description of the calibration method.

3.1 *Systematic Errors*

The reproducibility with which the beam can be deflected a distance y is related to the stability of the accelerating voltage V_b, the deflection voltage V_d, and the distance L from the deflection plates to the sample by the equation

$$\frac{\Delta y}{y} = \frac{\Delta V_b}{2V_b} + \frac{\Delta V_d}{2V_d} + \frac{\Delta L}{2L}. \tag{1}$$

These errors are all zero at the center of the field and increase out to the edge of the field. The expression $\Delta y/y$ in equation (1) is the fractional change in y which occurs for a given set of instabilities, measured from one corner of the field. Since *every* point on the reticle must be within the specified tolerance, it is not sufficient to require, for example, that the root mean square of the three terms on the right side of equation (1) be less than ±20 ppm; rather, the sum of their absolute values must be less than 20 ppm.

The steps taken to insure good stability and low transients in the deflection voltage have already been described. High-voltage stability was obtained by using a commercially available* 0- to 20-kV supply with estimated stability under constant load of better than ±10 ppm per hour. Variations in the distance L can arise either from surface non-uniformities on the electron-sensitive material or from a lack of reproducibility in referencing successive samples to the top of the sample holder. Surface variations are held to less than ±0.5 ppm and ±5 ppm is allowed for referencing successive samples.

3.2 *Random Errors*

In addition to the sources of errors just described, there are two sources of random error which must be considered. First, any insulating material in or near the path of the beam will tend to charge to the cathode potential, and the resultant electric field will cause the beam to deflect away from the computed position on the electron-sensitive material. Experiments conducted with this system indicate that charging of the electron-optical components is not a problem when proper cold-trap techniques are used. A small charging effect was observed under worst-case conditions when photographic emulsion on glass substrates was used without any metallic underlay; however, no detectable effect was observed on hundreds of samples with metallic underlay.

Second, time-varying magnetic fields at any frequency from essentially d.c. to several kHz (in particular 60 Hz) limit reproducibility by deflecting the beam from the programmed position by an amount which is proportional to the magnitude of the field and to the square of the interaction distance. Therefore, in a system with a long working distance, it is especially important to shield against magnetic disturbances. This problem is made difficult by the fact that the shielding must extend over a wide bandwidth down to very low frequencies. However, successful shielding has been obtained in the past by using an enclosure made up of successive layers of highly conductive material and of high-permeability magnetic material. A simple multi-layer

* Power Design Model HV 1584-R, produced by Power Design, Inc., Westbury, New York.

shield which was constructed has reduced deflections due to 60-Hz magnetic fields and to slow changes in the field of the earth to less than ±1 μm. A larger shield is being designed[6,7] to enable the electron beam apparatus to operate in magnetic environments somewhat noisier than those found in most research laboratories.

3.3 *Calibration*

The calibration technique, which uses electron-beam-induced sample current, is shown schematically in Fig. 6. The alignment target consists of a gold grid on a chromium substrate. When the electron beam is swept over the target, the current through the picoammeter varies due to the different back-scattering coefficients of gold and chromium. Figure 7 shows a chart recording of the changes in the sample current when the beam passes over a gold stripe. The stripe can be detected with a S/N of better than 100 and its position can be determined to within ±0.5 μm. The calibration will be accomplished by adjusting the accelerating voltage or the deflection voltage to maintain a constant number of address units between the fiducial marks.

IV. RESULTS

4.1 *Electron Beam Writing Characteristics*

Figure 8 shows a photograph of two intersecting lines written with the electron beam on the HRP emulsion with the chromium underlay. These lines are 4 μm wide and were written by single passes of the beam. The edge fuzziness of the lines is less than 0.5 μm, and no rounding off can be observed at the corners of the intersection. The optical density of the lines is about three.

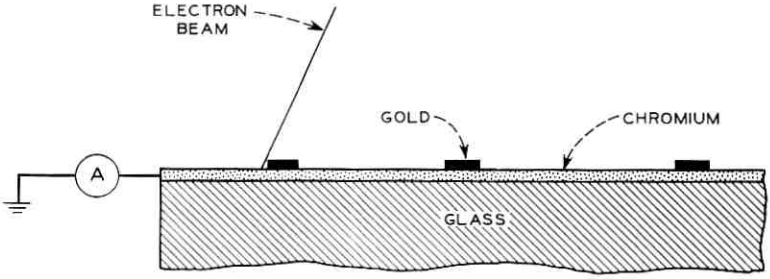


Fig. 6—Diagram of the calibration target.

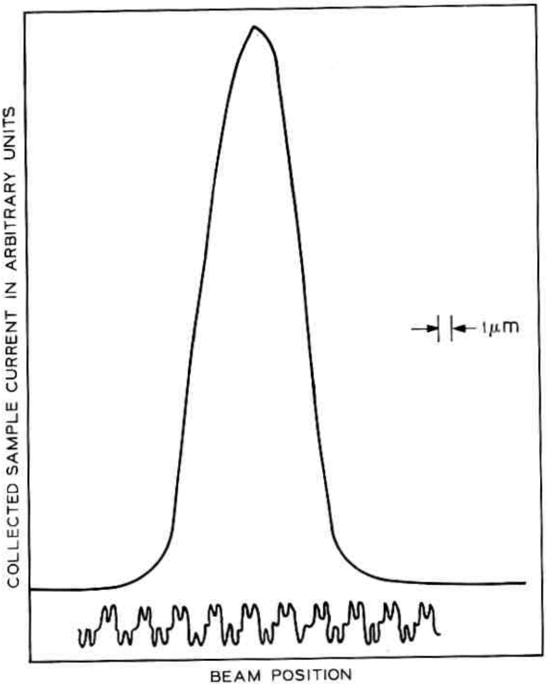


Fig. 7—Recording of calibration output. One complete cycle of the alternating signal gives four equally spaced reference marks, each of one address unit. To calibrate, number of address units between fiducial marks is held constant.

4.2 *Reticle and Mask Patterns*

Figure 9 shows a reticle produced by the electron beam pattern generator over a 5-cm by 5-cm field. This pattern is a lower-level metallization beam cross-over test reticle. The information for this pattern was obtained as an input deck to XYMASK. It was run on XYMASK on an IBM 360/50 and postprocessed on the same machine with a postprocessor written for the electron beam pattern generator. The outer edge of a corner of a path is purposely programmed with a circle. The generation time for this pattern is about three minutes. Figure 10 shows a test pattern for XYMASK, illustrating the sloped-line capability of the electron beam pattern generator. This pattern was also generated in less than three minutes.

Figure 11 shows a mask consisting of a 27 by 27 array of the patterns shown in Fig. 9. It should be emphasized that the pattern was drawn

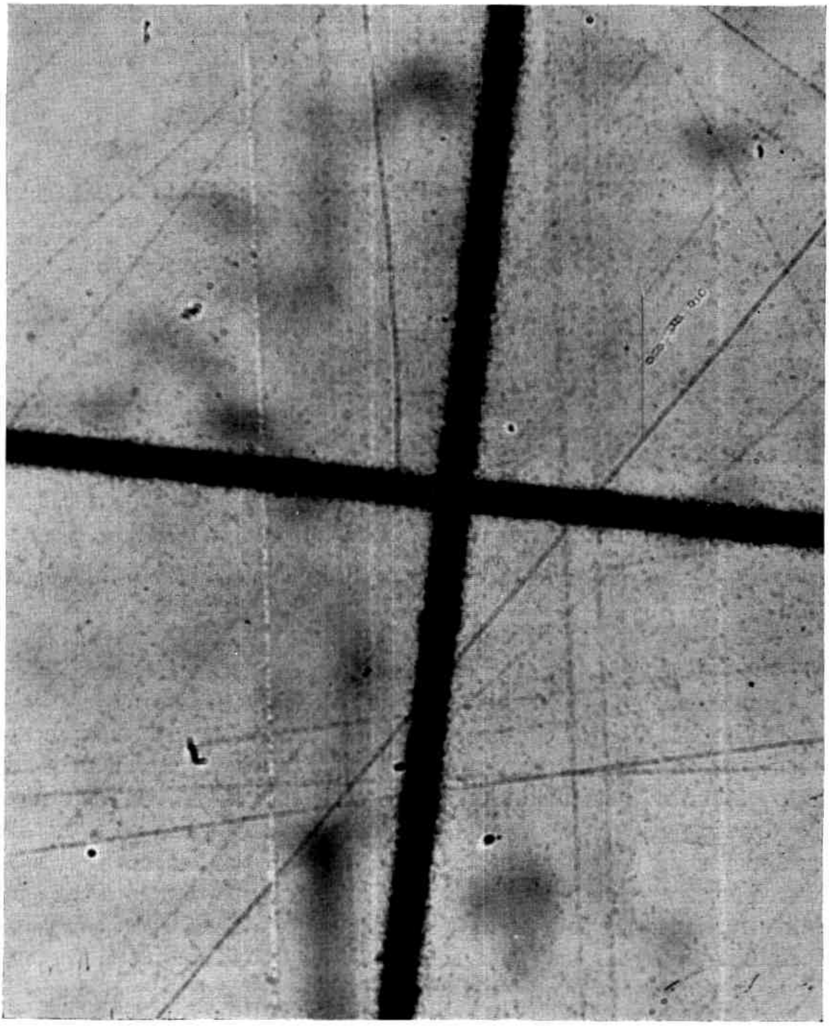


Fig. 8—Photograph of two intersecting 4-μm lines formed from single passes of the beam.

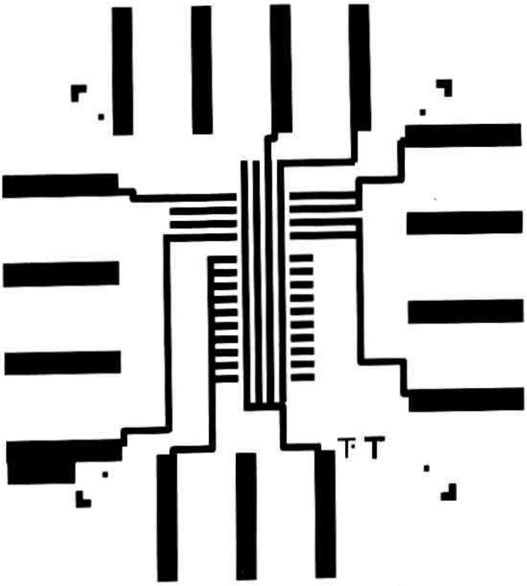


Fig. 9—Reticle produced by the electron beam.

over the 5-cm by 5-cm field without step-and-repeat in less than eight minutes.

4.3 *Stability*

From independent measurements of all the sources of error described in Section 3.2, it has been predicted that with the present equipment, the reproducibility of a pattern should be better than ±1.5 μm over the entire 5-cm by 5-cm field. Stability experiments have been performed by drawing grid patterns on the same plate at fixed-time intervals and measuring any displacements. Stability of ±1 μm has been observed for five-minute time periods. Experiments to measure the stability for longer periods of time are in progress. The largest source of instability is 60-Hz magnetic fields and the second largest is fluctuations in the accelerating voltage.

Experiments performed some time ago over a 4-mm by 4-mm field, which was the maximum field of the deflection system at that time, showed that stability better than ±1 μm could be obtained over a period of 24 hours. Measurements were made by two independent methods: by the use of a Preco comparator and by contacting two plates made 24 hours apart on a single plate using the image integration process developed by R. E. Kerwin and examining the results under a high-powered microscope.[8]

The following steps are being taken to improve long-term stability over the 5-cm by 5-cm field to ±1 μm in the near future. A large four-layer magnetically shielded enclosure is being constructed. The enclosure will have a clean-room interior and will have two sets of walls separated by a one-foot gap. Each wall will be made up of aluminum on the inside and molypermalloy* on the outside, separated by approximately two inches. This shield is designed to attenuate d.c. and 60-Hz magnetic fluctuations by factors of at least 200 and 5000, thereby reducing all beam position instabilities due to magnetic disturbances to less than ±0.1 μm. The customized high-voltage power supply is expected

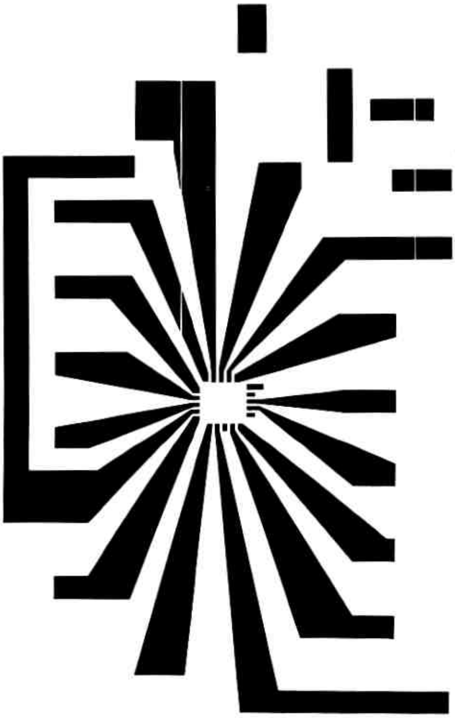


Fig. 10—Test pattern for XYMASK illustrating sloped line capability of the pattern generator.

* Supplied by Allegheny Ludlum Steel Corporation, Brackenridge, Pennsylvania.

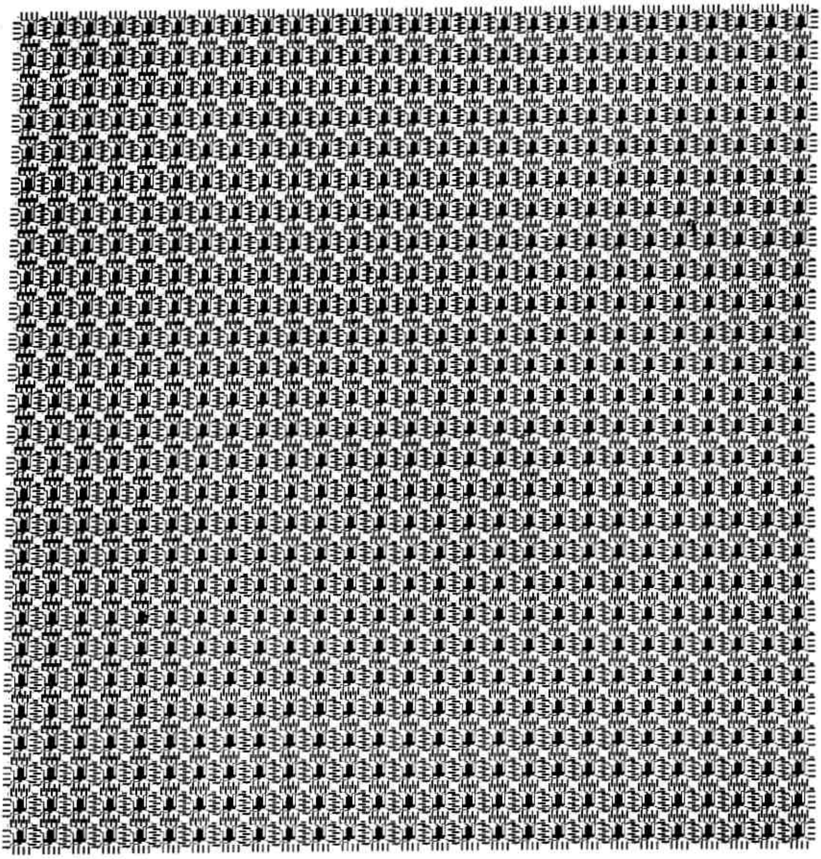


Fig. 11—Mask level containing an array of 27 × 27 of the patterns shown in Fig. 9.

to contribute less than ±0.1 μm to the beam position instability. The ±1-ppm stability of the DAC is more than adequate (±0.05 μm) and it is expected that the variations in the plate flatness and registration will contribute less than ±0.25 μm.

V. CONCLUSION

An electron beam pattern generator has been developed which has a field size of 5 cm by 5 cm. The line width is 4 μm and the lines drawn were shown to have an edge fuzziness of less than 0.5 μm. The optical density of lines produced on HRP emulsion by a single pass of the beam is about three. Experiments on a 4-mm by 4-mm field size showed a stability better than ± 1 μm over a period of 24 hours. In preliminary experiments on a 5-cm by 5-cm field an instability of ± 1 μm for five minutes has been observed. Measurements indicate that the instability on the 5-cm by 5-cm field over a five-minute time period is caused mainly by 60-Hz magnetic fields; high voltage fluctuations become important for longer time periods. It is anticipated that improved shielding and a new high voltage supply can produce long-term stability of ± 1 μm.

When completed, the electron beam pattern generator may be used to produce reticles, which will be compatible with the step-and-repeat camera described in a following paper.[9] Alternatively, the system may be used to make masks directly, with a minimum line width of 4 μm registering inside an 8-μm feature of another level. Although a 4 μm line width is not small by electron beam standards, its generation and control over a 5-cm by 5-cm field without step-and-repeat has not been reported before.

A linewidth of 4 μm over a 5-cm by 5-cm field was selected for reticle making. Smaller linewidths can be obtained with better lenses. There are many possible applications involving various combinations of linewidths and field sizes. The present system can possibly be extended to write with a 1 μm linewidth over a 5-cm by 5-cm field or with sub-micrometer linewidth over a 1-cm by 1-cm field. In addition, the long depth of focus and sub-micrometer resolution capability are important characteristics of electron beam systems for pattern generation directly on semiconductor slices.

VI. ACKNOWLEDGMENTS

A large number of persons at the Western Electric Engineering Research Center have contributed to this project. Particular credits are due to M. W. Sagal, B. Broyde, S. Badeau, A. D. Janzen, R. Sked and D. L. Torlay.

REFERENCES

1. Broers, A. N., Lean, E. G., and Hatzakis, M., "1.75 GHz Acoustic-Surface-Wave Transducer Fabricated by an Electron Beam," Appl. Phys. Letters (USA), *15*, No. 3 (August 1969), pp. 98–101.
2. Gross, A. G., Raamot, J., and Watkins, Mrs. S. B., "Computer Systems for Pattern Generator Control," B.S.T.J., this issue, pp. 2011–2029.
3. Raamot, J., "18-Bit Digital to Analog Conversion," American Federation of Information-Processing Societies (AFIPS) Conf. Proc., *36* (1970), Spring Joint Computer Conference.
4. Gorman, J. E., and Raamot, J., "Integer Arithmetic Techniques for Digital Control Computers," Computer Design, *9*, No. 7 (August 1970), pp. 51–57.

5. Broyde, B., "Electron Sensitive Materials," B.S.T.J., this issue, pp. 2095–2104.
6. Patton, B. J., and Fitch, J. L., "Design of Room Size Magnetic Shield," J. Geophys. Res., *67*, No. 3 (March 1962), pp. 1117–1121.
7. Cohen, D., "A Shielded Facility for Low-Level Magnetic Measurements," J. Appl. Phys., *38*, No. 3 (March 1967), pp. 1295–1296.
8. Kerwin, R. E., and Stanionis, C. V., "Image Integration Features of Photographic Physical Development," Electrochem. Technology, *6*, No. 11–12 (November–December 1968), pp. 463–464.
9. Alles, D. S., et al., "The Step-and-Repeat Camera," B.S.T.J., this issue, pp. 2145–2177.

Device Photolithography:

Electron-Sensitive Materials

By BARRET BROYDE

(Manuscript received May 27, 1970)

Certain additives increase the electron sensitivity of Kodak's negative photoresists by a factor of five to seven; others increase the sensitivity of AZ-1350 by a factor of two to three. With additives the contrast of the negative resists is increased, leading to sharper edges and higher resolution. Some of these additives also increase the light sensitivity of both positive and negative resists. A recording system based on a silver halide emulsion and containing a conductive underlay is also described.

I. INTRODUCTION

An electron beam pattern generator developed at the Western Electric Engineering Research Center requires novel recording systems that possess high resolution, high sensitivity at short exposure times, flat surfaces, and a conductive underlay.[1] Silver halide emulsions are best suited for the generation of reticles by this generator, while for the production of one-to-one masks or the generation of patterns directly onto silicon slices (thereby avoiding the use of masks) photoresists are the preferred recording media.

High-resolution emulsions and photoresists were chosen over other recording media since they offer the best combination of sensitivity and resolution.[2–15] (See Table I.) Systems using these two recording media are discussed in this paper.

II. SILVER HALIDE RECORDING MEDIA

Kilovolt electrons passing through silver halide grains form latent images in much the same way as photons.[16] Although electron scattering by the grains causes some loss in resolving power, the edges of lines generated by writing electron beams have been found to be as sharp as edges made by conventional processes.

The recording medium required when the writing beam is used for

TABLE I—ELECTRON BEAM RECORDING MEDIA

Recording Media	Smallest Spot Recorded (μm)	Flux of 15-keV Electrons Needed to Record (C/cm^2)	Ref.
High resolution silver halide emulsion	1–2	10^{-9}	2, 3
Photoresists			
Negative (KPR, KTFR)	0.25	$8\text{–}10 \times 10^{-6}$	4, 5
Positive (AZ-1350)	1	6×10^{-6}	6
Methacrylate resists	0.5	8×10^{-6}	7
Silicone resists	0.4	$\sim 10^{-5}$	8, 9
Polymerization of monomers absorbed from the vapor	1.5×10^{-2}	10^{-1}	10
Liquid crystals	30	$\approx 10^{-9}$	11
Ferromagnetics	>100	≈ 1	12, 13
Thermoplastics	≈ 10	$>5 \times 10^{-4}$	14
Electrostatic	>50	N.A.	15

reticle generation is made in the following way: Eastman-Kodak coats a high-resolution emulsion (649-GH) on glass manufactured by Liberty Mirror Company, Brackenridge, Pennsylvania. Seamed 6″ × 6″ glass plates, covered with a Liberty Mirror proprietary PE-81-E conductive coating, transmitting 75-80 percent of incident visible light, are used so that uniform coatings can be obtained. The majority of the transmission loss is in the glass. The surface of the glass is flat to ±27 μ inch per linear inch, which is sufficiently flat so that the total error of ±20 PPM allowed for the electron beam machine is not exceeded.[1] In order to meet this flatness specification, glass 0.235 ± 0.01 inch thick is used. The plates are cut to 3″ × 3″ before they are used in the electron beam pattern generator. A conductive coating is used to avoid charge storage by the recording medium. The proprietary Liberty Mirror coating was chosen since it is highly conductive (<1000 Ω/square), transparent (optical density <0.04), resistant to the precleaning procedure used by Kodak before coating with emulsion, and can be applied without heating the substrate.

III. ELECTRON RECORDING BY PHOTORESISTS

It is likely that the chemical reactions that the electrons cause in photoresists are the same as those induced by light. Negative resists undergo cross-linking[17,18] while positive photoresists are usually converted to carboxylic acids[17] and perhaps lactones.[19]

The direct exposure of photoresist coatings on the surface of silicon slices appears to be an attractive means of patterning semiconductor substrates. For this application the exposure time presently required for a writing beam is unacceptably long.[1] Increasing the beam current to reduce exposure times might lead to undesirable thermal effects, so work on increasing the sensitivity of photoresists was initiated. Both positive and negative resists were examined in order that the electron beam pattern generator would never be required to pattern more than half the addressable points, thereby minimizing exposure times.

3.1 *Chemical Additives to Increase the Sensitivity of Negative Photoresists*

Recently, chemical additives have been found here[20] which reduce the flux of 15 keV electrons needed to expose negative photoresists from 8–10 × 10^{-6} C/cm^2 to 1.5 × 10^{-6} C/cm^2. The absorbed energy required for full exposure corresponds to 4.2 × 10^{20} eV/cm^3.[18] Futher reductions in the required exposure are anticipated.

The additives, incorporated into the photoresist solution before it is applied, divide into two classes based on their mechanisms. The first class, alkyl and aryl compounds of heavy metals, e.g., hexaphenyldilead, reduce the required flux by acting in two ways: (*i*) they increase the capture cross section of the resist so that more energy is transferred to the resist layer, and (*ii*) since these compounds are readily dissociated into free radicals, they probably reduce the required flux by initiating more than the average number of crosslinks. The second class of additives, of which benzophenone is typical, does not increase the capture cross section but does cause more than normal crosslinking. This occurs either because the additives are readily dissociated into free radicals which initiate crosslinking, or because they are excited to low-lying triplet states after the primary process of absorption.[21,22] These triplet states decay slowly and transfer energy to more distant polymers causing them to crosslink.

Figure 1 shows the typical effects of an additive. The thickness of exposed and developed KTFR is given as a function of the exposing flux, both with and without benzophenone. Note that there is a threshold; that is, no insolubilization occurs below a certain flux of electrons. Since negative resists are a subset of crosslinking systems, they are not insolubilized until there is an average of one crosslink per molecule. Sufficient radiation to form this average number of crosslinks per molecule makes part of the radiated resist insoluble, giving rise to a threshold. Further radiation increases the thickness of the exposed photoresist by insolubilizing more and more of it until the maximum thickness is obtained. Similar results have been found

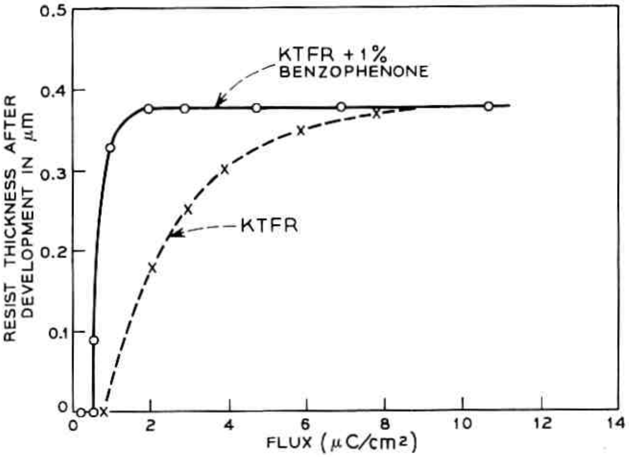


Fig. 1—Effect of benzophenone on the exposure of KTFR by 15 keV electrons. The initial thickness of the resist was 0.6 μm.

with KPR. Figure 1 also shows that the slope of the photoresist thickness versus flux curve is much higher when additives are present; that is, the resist with additives is a high contrast recording medium.

Γ has been defined here as a contrast function for photoresists. Analogous to γ used in photography to specify the contrast of a film, Γ is defined:

$$\Gamma = \frac{\text{threshold flux}}{\text{flux for maximum thickness}}$$

so that

$$0 < \Gamma \leqq 1.$$

A high Γ implies good contrast, giving sharp edges in the patterning process.

Some of the results that have been obtained on increasing both the sensitivity and the contrast of KTFR and KPR are shown in Table II. Although benzophenone and hexaphenyldilead are equally efficacious in reducing the flux required for full exposure, benzophenone is the preferred additive since it is more soluble in photoresists. Benzil and 1,4-diphenyl-1,3-butadiene show behavior quite similar to benzophenone.

3.2 *The Electron Exposure of AZ-1350—A Positive Photoresist*

The solubility of AZ-1350 films as a function of the exposing flux of 15 keV electrons is shown in Fig. 2. Typical results obtained with

TABLE II—ADDITIVES THAT INCREASE THE SENSITIVITY AND CONTRAST OF NEGATIVE PHOTORESISTS

Resist	Additive	Flux of 15-keV Electrons Needed to Expose (μC/cm^2)	Γ
KTFR	None	8	0.11
	1.0% benzophenone	1.5	0.33
	Hexaphenyldilead (sat.)	1.5	0.33
	2.0% triphenylbismuth	1.9	0.5
KPR	None	10	0.09
	1.0% benzophenone	1.5	0.33

newly purchased AZ-1350 are presented in curves A and B and are summarized below. At low exposures ($<10^{-8}$ C/cm^2) no solubilization occurs and no image can be detected in the photoresist. Fluxes between 10^{-8} and about 6×10^{-6} C/cm^2 cause an image to form after development; some of the resist is solubilized, but not all can be removed. If the irradiation is brought up to fluxes between 6×10^{-6} and 8×10^{-5} C/cm^2, then the resist can be fully solubilized. Irradiation with fluxes greater than 8×10^{-5} C/cm^2 yields an insoluble spot after development.

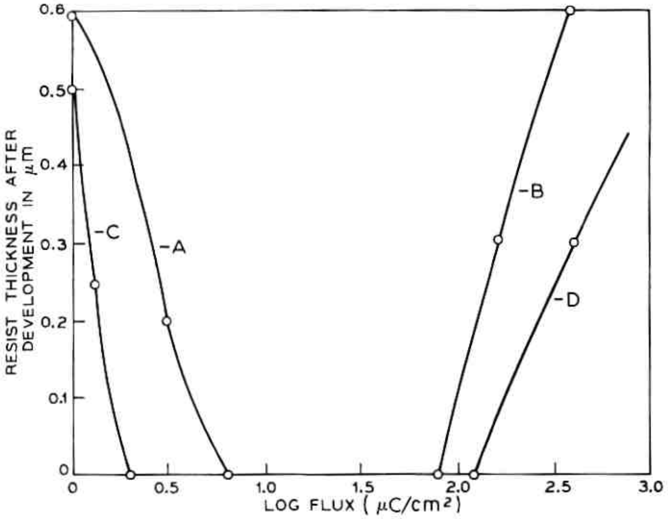


Fig. 2—Exposure of AZ-1350 by 15 keV electrons. Curves A and B newly purchased resist. Curves C and D, newly purchased AZ-1350 containing 2% benzotriazole. The initial thickness of resist was 0.6 μm.

The mechanisms of the reactions involved in the response of AZ-1350 to electron radiation have not been determined yet, but it is likely that a crosslinking reaction of its phenol-formaldehyde polymer[23] leads to the insoluble product. It is also likely that the solubilization reaction is the same one that occurs with light; quinone diazides are converted to carboxylic acids. Solubilization arising from a scission reaction of the phenol-formaldehyde polymer in which indiscriminate bond breakage occurs and soluble low molecular weight compounds result cannot be fully excluded, but it is not probably because scission would have to predominate at low exposures and crosslinking at high fluxes.

3.3 *Increasing the Sensitivity of AZ-*1350 *by the Addition of Benzotriazole*

When 2 percent solutions of benzotriazole in AZ-1350 are prepared, the flux of electrons needed to solubilize the resist is decreased (curve C, Fig. 2). Benzotriazole also inhibits crosslinking that occurs at high electron fluxes (curve D, Fig. 2). To date, the lowest fluxes required for full exposure are 2 μC/cm^2.

Benzotriazole is not the only efficacious additive. All members of the benzotriazole, imidazole and indazole families that are soluble and that have a hydrogen atom bound to a nitrogen atom also decrease the flux needed to solubilize AZ-1350.

A full explanation for the effects of benzotriazole-type additives has not been developed yet, but crosslinking inhibition (curve D versus curve B, Fig. 2) probably arises from the antioxidant properties of benzotriazole; benzotriazole reacts with the free radicals generated by the absorption of energy before they cause crosslinking.

IV. INCREASED LIGHT SENSITIVITY OF PHOTORESISTS

Patterning semiconductor slices is conventionally done by contact exposure through a mask. In this process, a mask is placed on top of and in contact with a photoresist-coated slice and the photoresist exposed by ultraviolet light through the mask. In this way, a contact print of the mask is made on the photoresist. The lifetime of a mask copy is limited by abrasion of the mask during printing to about 10 exposures for an emulsion copy and about 150 exposures for a chrome copy. More important, contact with the mask results in defects, such as pinholes in the pattern and mechanical damage to the photoresist coated slice. Defects related to contact printing have been recognized in the past, but their effect on the yield of discrete semiconductor devices is small. However, their effect on the yield of high precision, large area integrated circuits is much more severe.

An electron beam pattern generator writing on the sensitized resists discussed in the previous section offers one means of patterning silicon slices. Projection photolithography systems would benefit if more sensitive photoresists were available, since exposure times will be reduced and the problems of dust settling on the optics of the exposure system and the mechanical instabilities of the system would be minimized.

4.1 *Increasing the Sensitivity of KPR to Light*

Benzophenone and its derivatives have been reported to increase the sensitivity of polyvinyl cinnamate, the polymer in KPR, possibly because they increase the absorptivity of the system at longer wavelengths.[24] It was found here that benzophenone decreases the time required to produce the maximum thickness of KPR films (polyvinyl cinnamate and sensitizer) after development, but that the threshold flux is not decreased. (See Fig. 3.) This implies that the edge resolution of the image is increased, when benzophenone is present, and sharper etched lines should result. In neither this case nor in the

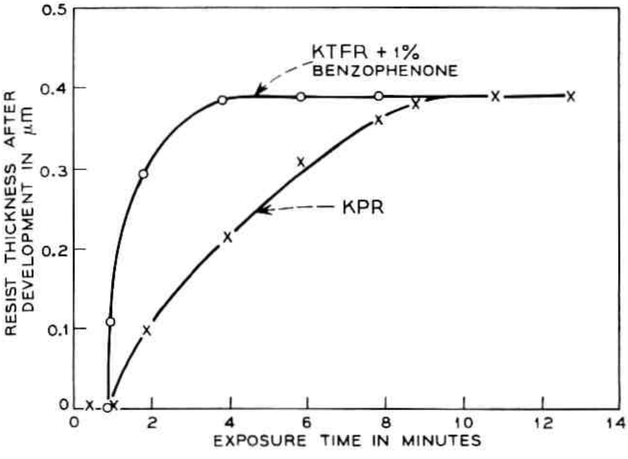


Fig. 3—Effect of benzophenone on the exposure of KPR by light. The source was a 150-W xenon lamp, 100 cm from the resist. The initial thickness of the resist was 0.6 μm.

case of the sensitization of AZ-1350 discussed below does the sensitization appear to arise from increased light absorption since the absorption spectra of the resists with and without additives are identical.

4.2 *Increased Sensitivity of AZ*-1350 *to Light*

The exposure time required to fully solubilize AZ-1350 can be reduced when low concentrations of benzotriazole or similar compounds are incorporated into the AZ-1350 film.[25] Some results are shown in Fig. 4.

V. SUMMARY

A photographic recording system suitable for use with a reticle genenator has been developed. The electron flux rquired to fully expose negative photoresists has been reduced from 8–10 $\mu C/cm^2$ to 1.5 $\mu C/cm^2$ by incorporating additives into the resists. Other additives reduce the flux required to expose AZ-1350 from 6 $\mu C/cm^2$ to 2–3 $\mu C/cm^2$. Further reductions are anticipated for both systems.

Most of the additives that sensitize the response of resists to electrons also increase their sensitivity to light. So far, the photon flux required for full exposure has been reduced by a factor of 50 to 100 percent.

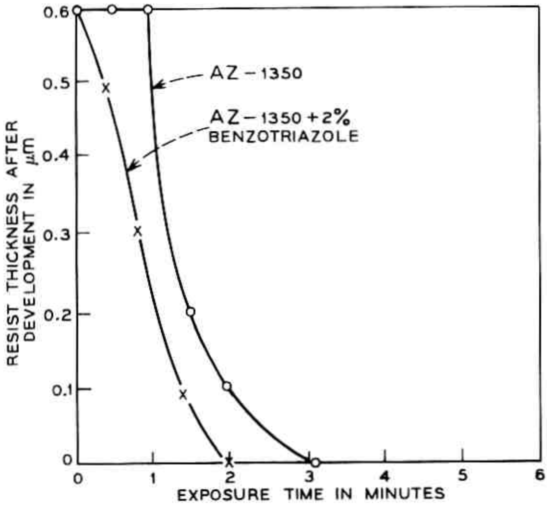


Fig. 4—Effect of benzotriazole on the exposure of AZ-1350 by a 150-W xenon lamp 100 cm from the sample. The initial thickness of the resist was 0.6 μm.

VI. ACKNOWLEDGMENTS

The author is indebted to Drs. H. H. Loar, M. W. Sagal, and W. R. Samaroo of Western Electric Company, Inc., for many stimulating discussions and to Mr. C. A. Adams for his proficient technical assistance.

REFERENCES

1. Samaroo, W. R., Raamot, J., and Parry, P. D., "An Electron Beam Pattern Generator," 1970 IEEE Convention Digest, New York, New York (March 1970).
2. Loeffler, K. H., "An Electron Beam System for Digital Recording," *Record of the Ninth Annual Symposium on Electron, Ion and Laser Beam Technology,* San Francisco: San Francisco Press, 1967, page 344.
3. Shepp, A., Whitney, R. E., and Masters, J. I., "Evaporated Silver Bromide as an Electron Beam Recording Material," Photographic Sci. and Eng., *11* (September–October 1967), p. 322.
4. Broers, A. N., "Combined Electron and Ion Beam Processes for Microelectronics," Microelectronics and Reliability, *4,* No. 1 (January 1967), p. 103.
5. Kanya, K., Yamazaki, H., and Tanaka, K., "Measurement of Spot Size and Current Density Distribution of Electron Probes by Using Electron Beam Exposure of Kodak Photoresist Films," Optik, *25* (July 1967), p. 471.
6. Matta, R. K., "High Resolution Electron–Beam Exposure of Photoresists," Electrochemical Technology, *5,* No. 4 (July–August 1967), p. 382.
7. Broers, A. N., and Hatzakis, M., "Some New Characteristics of the Methacrylate Electron Resist," 10th Symposium on Electron, Ion and Laser Beam Technology, Gaithersburg, Maryland, May 1969.
8. Roberts, E. E., "Rapid Direct Deposition of Silicaceous Diffusion Barriers by Electron Beams," 133rd Meeting Electrochem. Soc., Boston, Massachusetts, May 1968.
9. Yatsui, Y., Nakata, T., and Umehara, K., "Electron Beam Exposure of Silicones," J. Electrochem. Soc., *116,* No. 1 (1969), p. 97.
10. Pease, R. F. W., and Nixon, W. C., "Microformation of Filaments," *First Inter. Conf. on Electron and Ion Beams,* R. Bakish, Editor, New York: John Wiley, 1964, p. 250.
11. Hansen, J. R., and Schneeberger, J. R., "Liquid Crystal Media for Electron Beam Recording," Paper 15.1, IEEE Electron Devices Meeting, Washington, D. C., 1967.
12. Blanchard, J. G., and Hart, D. M., "Electron Beam Recording and Readout," IBM Technical Disclosures, *9,* No. 5 (May 1967), p. 1762.
13. Land, C. E., "Ferroelectric Ceramic Electro-Optic Storage and Display Devices," Paper 15.2, IEEE Electron Devices Meeting, Washington, D. C., 1967.
14. Glenn, W. E., Jr., "Thermoplastic Recording," J. Appl. Phys., *30,* No. 12 (December 1959), pp. 1870–1873.
15. Krittman, I. W., and Inslee, J. W., "Discussion and Applications of Electrostatic Signal Recording," RCA Rev., *24,* No. 3 (September 1963), p. 406.
16. Tarnowski, A. A., and Evans, C. H., "Photographic Data Recording by Direct Exposure With Electrons," J. Soc. Motion Picture and Television Engineers, *71,* No. 10 (October 1962), p. 765.
17. Kosar, J., *Light Sensitive Systems,* New York: John Wiley, 1965.
18. Broyde, B., "Exposure of Photoresists: Electron Exposure of Negative Photoresists," J. Electrochem. Soc., *116,* No. 9 (September 1969), p. 1241.
19. Levine, H. A., "Positive Photoresist Materials," *Polymer Preprints,* Amer. Chem. Soc., *10* (January 1969), p. 337.
20. Broyde, B., "Electron Beam Exposure of Sensitized Photoresists," 134th Meeting, Electrochem. Soc., Montreal, Quebec, October 1968.

21. Kikuchi, S., and Nakamura, K., *The Photocrosslinking of Polyvinyl Cinnamate, Postprints of Photopolymers—Principles, Processes, and Materials,* New York: Soc. Plastics Eng., 1967, p. 175.
22. Moreau, W. M., "Photosensitization of Polyvinyl Cinnamate," *Polymer Preprints,* Amer. Chem. Soc., *10* (January 1969), p. 362.
23. Harwood, M. G., and Hunter, D. N., "A Study of Some Photosensitive Resists Used in Microcircuitry," AD 846, 184.
24. Minsk, L. M., Van Deusen, W. P., and Robertson, E. M., "Photosensitization of Polymeric Cinnamic Acid Esters," U. S. Patent 2,670,287, applied for January 20, 1961, issued February 23, 1954.
25. Broyde, B., "Exposure of Photoresists: II. Electron and Light Exposure of a Positive Photoresist," 137 Meeting, Electrochem. Soc., Los Angeles, California, May 1970.

Device Photolithography:

Lenses for the Photolithographic System

By DONALD R. HERRIOTT

(Manuscript received July 10, 1970)

The edge definition, maximum complexity and accuracy of details in photolithographic masks are limited by the performance of the lenses in the system. The tolerances on exposure, sensitivity and uniformity of the photosensitive materials, and processing are dependent upon the images formed exceeding the minimum quality required. The lenses in this system have been designed and fabricated to achieve the best practical performance at this time in order to obtain the largest tolerances possible. This paper details the design parameters chosen, the constructions used and the performance obtained by each of the lenses in the system.

I. INTRODUCTION

There are two classes of photographic mask-making systems. In the first class, the pattern is generated through a lens as in a cathode-ray-tube plotter or primary pattern generator (PPG), or a lens is used to reduce the size of the pattern to that of the circuit being made. The maximum complexity of pattern in this type of system is limited by the resolution that can be obtained over the field of a lens.

A second class of systems uses a lens imaging a single small spot of light that is moved over an area and modulated to write a pattern. In this type of pattern generator the complexity of pattern is limited only by the minimum spot size and the area covered. This system must be used to draw the mask at the same scale as the final circuit or the lens in a reduction camera would limit the resolution.

Systems in the first class have been chosen for the mask laboratory in spite of the resolution limitations because of the speed and flexibility of the lens type systems for making a wide variety of masks. As a result the lenses in the system are the principal limitation on the maximum complexity of patterns that can be produced and on the quality of the images.

The performance of lenses is limited by the wavelength of light, the aperture of the lenses, and the aberration correction of the lenses. The wavelength of visible light is about half a micron, and it is theoretically possible to obtain light distributions in an image having cycles of light and dark of about one-half-micron width. Blue light can be imaged with better resolution because it has a shorter wavelength than green or red light. The wavelength that can be used in making masks is limited by the sensitivity of the photographic materials, the available light sources and the transmittance of the glasses used in the lenses and as a substrate for the photosensitive materials.

The resolution is also limited by diffraction. It would be necessary to bring light to the image from a cone subtending an angle of 180° to resolve spatial images with periods of one wavelength. A smaller angle of light to an image will limit the resolution to larger detail. The large apertures of the lenses used in this system are required for resolution of the detail in the masks rather than to collect light.

The resolution of a lens may also be limited by aberrations. A single lens element with spherical surfaces will not image the light passing through it from a point in the object to a point in the image. Aspheric surfaces could be used to do this for a point on the axis of the system but not for points off axis. These defects in the imagery can be greatly reduced by combining many elements designed to compensate for the aberrations. It is not possible to reduce these aberrations to zero but they can be made smaller than the diffraction effects by using complex combinations of lenses.

II. MODULATION TRANSFER FUNCTIONS

A convenient measure of the quality of an optical image is the modulation transfer function (MTF). This is a curve of the contrast that is obtained in the image of a sinusoidal intensity target as a function of the spatial frequency of the target. Figure 1 shows a series of MTF curves for perfect lenses of various aperture ratios. The MTF varies from 0 to 1.0 and is the ratio of the contrast in the image to that of the target. The spatial frequency scale is in cycles per mm and covers the general range of interest in mask-making systems. As you can see in Fig. 1, the smaller the *f*/number, the better the contrast and the higher in spatial frequency it extends. Thus, to get a high quality image of 25μ lines in a reduction camera may require only an *f*/8 cone angle to the image, but good 1μ lines in a step-and-repeat camera require a lens of *f*/2 or faster.

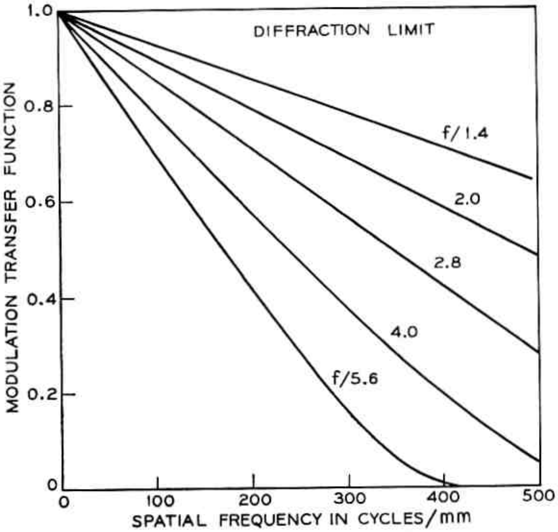


Fig. 1—MTF as a function of spatial frequency in an image formed by a cone of light of the indicated "*f*" number.

This requirement for low "*f*" numbers for high resolution may seem strange to those who are used to stopping down the lens to get a sharper image. This is because conventional camera lenses are limited in performance by their aberrations and stopping down the lens reduces these aberrations. The best resolution is probably obtained at about *f*/8; the image gets poorer when stopped down beyond that because of the diffraction limits shown in the MTF curves. Photographic lenses are often used in low-light conditions and the value of the increased speed obtained by increasing the aperture is more important than the loss in resolution caused by the aberrations.

In contrast, the large apertures of lenses for mask-making systems are almost always picked for resolution rather than speed. It is therefore necessary to reduce the aberrations to values that are small in comparison to their diffraction effects. There is still a compromise region. A lens for a 2.5μ linewidth mask should have an MTF of over 60 percent at 200 cycles/mm. This could either be obtained with a perfect *f*/4 lens or an *f*/3 lens with some aberrations. It could also be obtained with an *f*/2 lens with larger aberrations but unless the exposure speed of the lens were critical, the greater complexity of the *f*/2 lens would make it more expensive and prone to larger errors in fabrication.

A second reason to select the smaller aperture is its increased depth of focus. When projecting an image directly onto a non-flat silicon wafer, this can be of major importance. In making masks on glass it determines the flatness tolerance; in all cases it determines the accuracy to which the cameras must be focussed and the stability of this focus.

III. SYSTEM CONSIDERATIONS

The lenses used in this mask-making system have been designed for practical operation in a production system. The parameters have been selected to advance the state of the art in each area and to obtain the largest tolerance possible in each operation of the mask system.

The performance of each part of the system is limited by the lens. The 26,000 address width of the pattern generator field is near the maximum that can be obtained with the aperture limits of the scanning system. The 5000 linewidth square field of the step-and-repeat camera is even more challenging to the lens designer for the small image involved. The reduction-camera lenses are not as difficult but have been designed for higher performance and therefore greater tolerances in use.

All of the lenses have been designed without major consideration of cost as even small improvements in performance would result in operating savings in excess of any reasonable cost.

IV. LENS DESIGN

The design of specialized lenses of this type is far ahead of the ability to manufacture them with uniform quality. In recent years automatic lens design programs have been developed which efficiently find the optimum design from each starting point while placing the desired importance on each characteristic. For instance, it has been found that designs of the types used are capable of essentially zero field distortion. It would be difficult using manual design techniques to find designs completely free of distortion. With automatic design programs, a small weight on distortion will cause new designs to be selected by the programs that are free of distortion until it is necessary to compromise other characteristics. The designer can then see just what must be sacrificed in one characteristic for gain in the other.

It is either necessary for the lens designer to learn all of the other parameters of the mask system or for the system designer to understand the lens design difficulties to arrive at suitable system compromises. The development of automatic design programs has made it reasonable for the system designer to explore the design of the lens while designing the system. A variety of lens designs for the lenses of this program were explored by the systems designer although the final lenses were designed and constructed by an experienced lens design group at Tropel, Inc.* In this manner, the system parameters were selected, a suitable performance target could be determined, and a tentative choice between performance and complexity could be made prior to final lens design.

V. LENS ASSEMBLY

All of the lenses in the system have maximum wavefront aberrations of approximately $\lambda/4$. They have up to 14 air glass surfaces as well as two or more cemented surfaces. The quality of each of these surfaces must be very good so that the accumulations of the errors on the individual surfaces including the inhomogeneity of the glass does not approach the aberration tolerance. The centering and spacing of the elements must be of extraordinary quality to maintain the diffraction limited performance. Conventional techniques for measuring and controlling the centering and spacing of lens elements are not sensitive or accurate enough for lenses of this type. The lenses have been assembled by Tropel using new techniques that they have developed in recent years. We have carried out a program at the Laboratories to explore improved interferometric techniques that will make even better lens systems feasible.

VI. LENS EVALUATION

Lenses are now evaluated by photoelectrically measuring the modulation transfer function in a lens bench. This is done by scanning the image of a periodic target with a slit or the image of one slit with a second one and calculating the transfer function. For lenses of this quality, the slits must be extremely narrow and the measurement is limited by the photon noise of signals through the slits and the stability of the lens bench and air during the time of measurement. One measured curve is shown for the 3.5X lens but the measurement is not convincing as the curve goes above theoretical values at high frequency. Wavefront measuring methods are now being developed from which better MTF curves should be obtained.

* Located in Fairport, New York.

VII. PATTERN GENERATOR LENS

The pattern generator lens has very special requirements. It must both collimate the laser beam before it is reflected from the polygonal mirror and then image the reflected beam to a flat focal plane on the photographic plate. The effective aperture position for the lens is at the surface of the mirror. The gaussian light distribution in the aperture of the lens is controlled by the illuminating laser beam. Although the lens is corrected at $f/10$, the writing beam fills the aperture with an $f/22$ cone angle which gives a 10μ-diameter gaussian distribution in the image. The code beam fills a larger aperture in the scan direction so that a higher modulation is obtained when the image scans the 7-μm bars and spaces of the code beam. The lens must provide a large amount of barrel distortion so that a constant angular rate of the scanning mirror provides a uniform linear scan in the focal plane. The combination of no vignetting of the laser beam in the lens and a uniform linear velocity of the scan gives a uniform exposure over the plate. Figure 2 shows the scanning lens and Fig. 3 shows the calculated MTF of this design.

VIII. REDUCTION-CAMERA LENSES

The reduction-camera lenses image the pattern generator plate onto HRP photographic plates. The mercury 435.8-nm spectral line is used so that only the monochromatic aberrations are critical. The lenses are correct for first-order axial and lateral color at this wavelength. The field angle is a compromise between camera length and aberration correction. The entrance pupil distance is the same for both the 3.5X

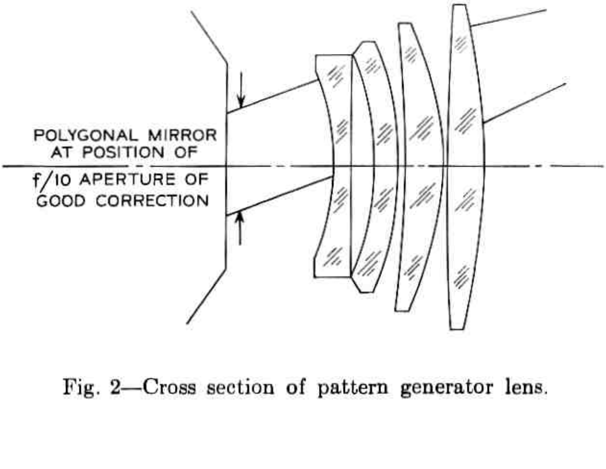


Fig. 2—Cross section of pattern generator lens.

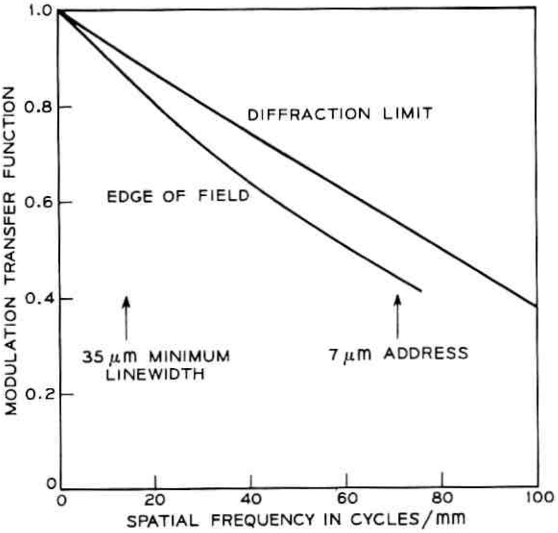


Fig. 3—MTF curves for the pattern generator lens on axis and at the edge of the field in relation to the fundamental frequency of the 7-μm address and 35-μm linewidth.

and 1.4X lenses so that the same illumination system can be used for both. Microflat glass plates are used in this camera so depth of focus is not important. The apertures have been selected to give best image quality and an iris is built into each lens so that they can be stopped down if poorer quality glass is used.

The 435.8-nm wavelength was selected as a compromise between the better resolution at the shorter wavelength than the more commonly used 546.0-nm line, and the smaller amount of scattered light in the green. The scattering in the blue is greatly reduced by using the dyed emulsion plates that are described in another article in this issue.

IX. 3.5X REDUCTION CAMERA LENS

The 3.5X reduction-camera lens shown in Fig. 4 is a seven-element double-Gauss type operating at $f/3.5$ and having a focal length of 17.7 cm. Efforts were made to use an eight-element design for better performance but the improvement was not judged sufficient to exceed the probable losses in an extra element. Figure 5 shows the MTF curves for this lens on axis and at the edge of the field along with the diffraction limit for the lens aperture used. The fundamental frequency

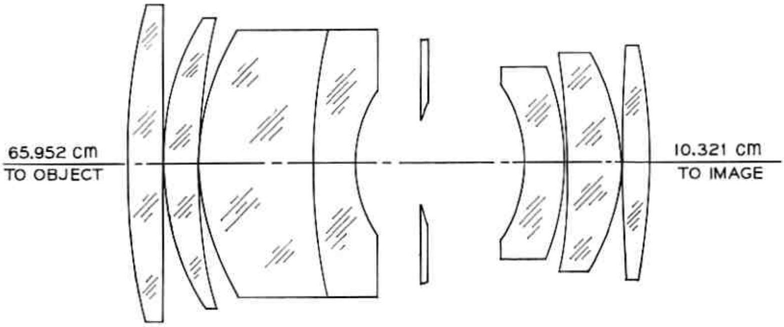


Fig. 4—Cross section of 3.5X reduction camera lens.

for a 10μ minimum linewidth used would be at 50 cycles per mm where the response is 70 percent or greater. There is significant response at a number of harmonics of this frequency to better reproduce sharp edges.

The intensity distribution for a square-wave object can be calculated from the response at the various harmonics in the source. Figure 6 shows the intensity distribution calculated for this lens from a 10μ-periodic square wave object, an isolated 10μ line at the center of

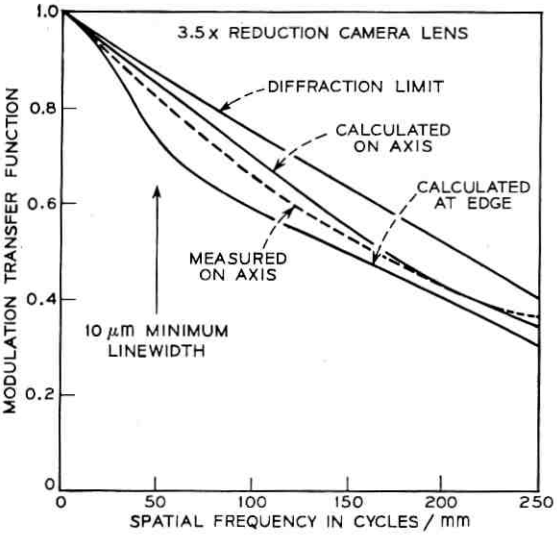


Fig. 5—Measured and calculated MTF curves for 3.5X lens.

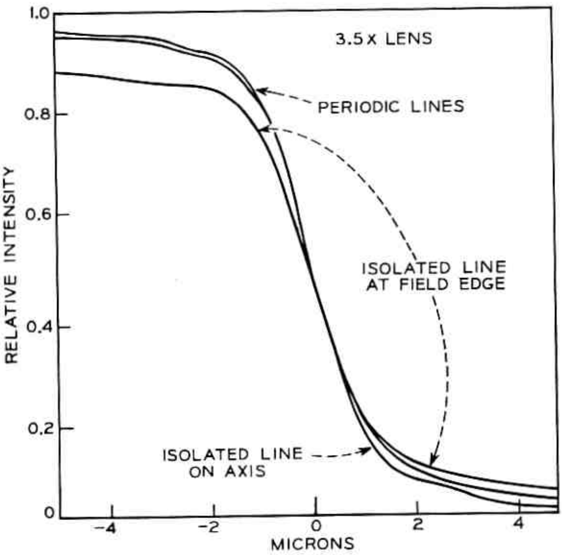


Fig. 6—Intensity distributions for 10-μm-wide periodic and isolated lines on axis and at the corner of the field of the 3.5X lens.

the field, and at the edge of the field. It is important that the slope of these curves at the edge of the line be large so that variation of exposure caused by light-source fluctuation, photographic-material sensitivity variation, and developing chemistry, time or temperature will not have a large effect on the linewidth developed from the image. As can be seen here, the isolated line and periodic lines would require a

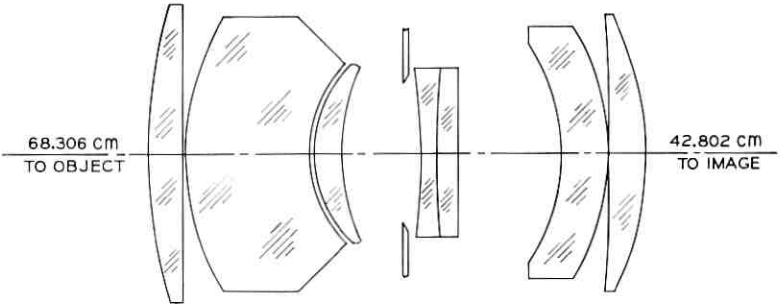


Fig. 7—Cross section of 1.4X reduction-camera lens.

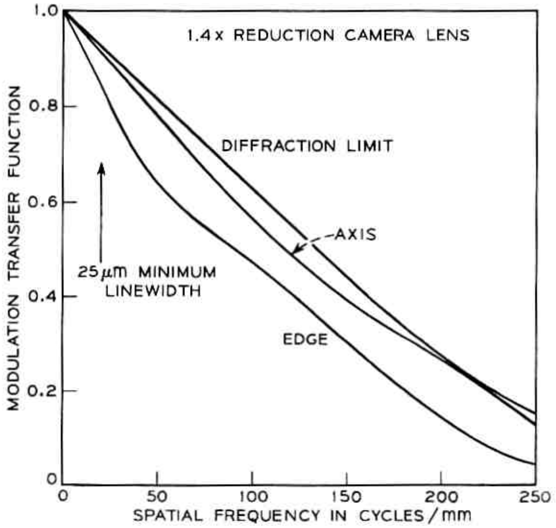


Fig. 8—MTF curves for the 1.4X reduction-camera lens.

slightly different exposure to both have correct linewidth. While this different exposure can be used to obtain accurate linewidth on masks having predominantly isolated or periodic lines, only a lens with a good MTF will give consistently accurate dimensions on all types of features.

X. 1.4X REDUCTION-CAMERA LENS

The outline of the 1.4X lens is shown in Fig. 7. While a double-Gauss type could have been used for this lens, this rather unusual configuration gave better performance for the specific requirement and the size is much smaller than the double-Gauss type.

The focal length is 32.4 cm and the overall length is 128.4 cm. The $f/4.15$ aperture provides a smaller cone to the image than the 3.5X lens but accepts a larger cone of light from the object providing better resolution compared to the finest line.

Figure 8 shows the MTF curves for the 1.4X reduction-camera lens and Fig. 9 shows the corresponding intensity distribution for periodic and isolated 25-μm lines. The 80 percent MTF at the fundamental frequency of the line results in a sharper line edge in the intensity profile and a resulting larger tolerance in exposure.

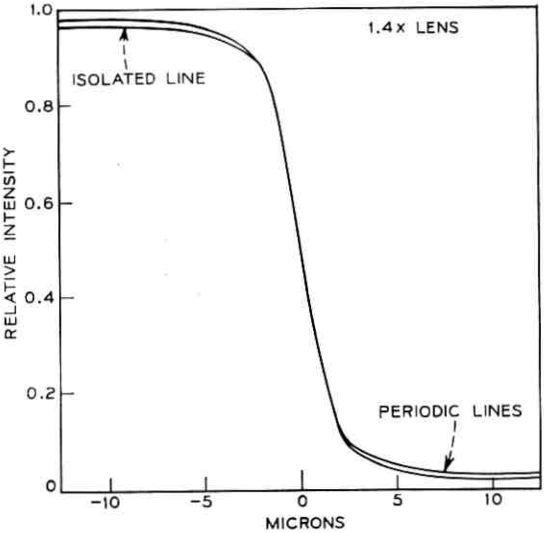


Fig. 9—Intensity distributions for isolated and periodic lines imaged by the 1.4X lens.

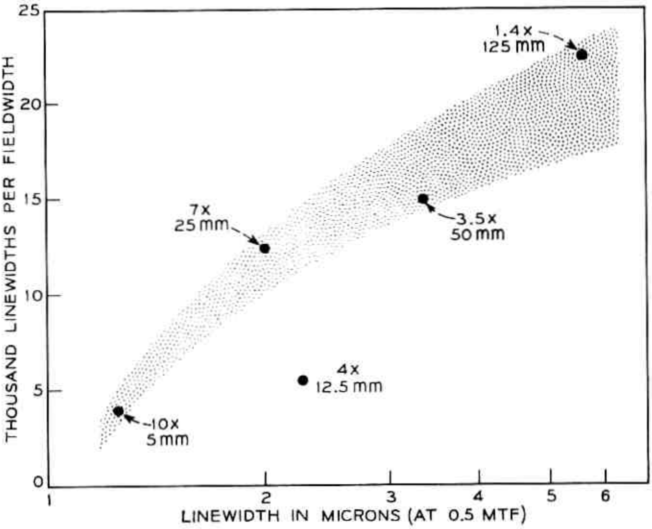


Fig. 10—Performance of a group of photolithographic lenses plotted as the number of linewidths per field width as a function of the linewidth at which 0.5 MTF is obtained.

XI. CAPABILITY OF GENERAL PHOTOLITHOGRAPHIC LENSES

The designs of the lenses in this system, including a 7X reduction-camera lens that has not been used, show the general range of performance that can be obtained. Figure 10 shows the number of thousands of linewidths per field as a function of the linewidth at 0.5 MTF. The shaded region indicates the area of reasonable design. There is not a smooth curve through these points as different lens types are used. A smoother curve could be drawn for each lens type. The 4X projection lens below the shaded area is limited in aperture and therefore resolution because of the required depth of focus. The 10X step-and-repeat lens is a very reliable point as many designers have designed lenses having these parameters. The step-and-repeat lens is described in detail in another paper in this issue.

Reduction Cameras: Optical Design and Adjustment

By ERIC G. RAWSON

(Manuscript received June 2, 1970)

This paper describes the optical design of the photolithographic reduction cameras and discusses in detail several aspects of the illumination system including the light source spectrum, the method of attaining even illumination, and the use of a Fresnel condenser lens. The camera design provides for first-order correction of focus and magnification shifts due to changes in ambient temperature. To adjust the cameras for best focus and proper magnification, a new technique using a special test reticle and digital computers was developed. It automates much of the procedure and processes much more data than would otherwise be possible. The reticle allows simultaneous measurement of focus and magnification errors throughout the image field, and a time-shared computer calculates the required corrective shifts on the object- and image-spacer bars.

I. INTRODUCTION

This paper and the paper immediately following describe the two reduction cameras which have been developed to serve as part of the photolithographic mask-making facility described in this issue.

The primary pattern generator[1] generates artwork masks which are nominally 17.5 cm square. This size was determined by various optical and mechanical considerations. The two reduction cameras reproduce these masks at the two specific, reduced sizes required for use as masters for tantalum thin film circuits and interconnection substrates; the reduced masks from one of these cameras (the 3.5X camera, shown in Fig. 1) can also serve as the reticle in the step-and-repeat camera.[2] The two reduction ratios, together with the corresponding mask size and minimum linewidths, are summarized in Table I. In each size the minimum linewidth is 1/5000 of the width of the mask.

This reduction in mask size carried out by the reduction cameras must be accomplished without significant loss of resolution in the

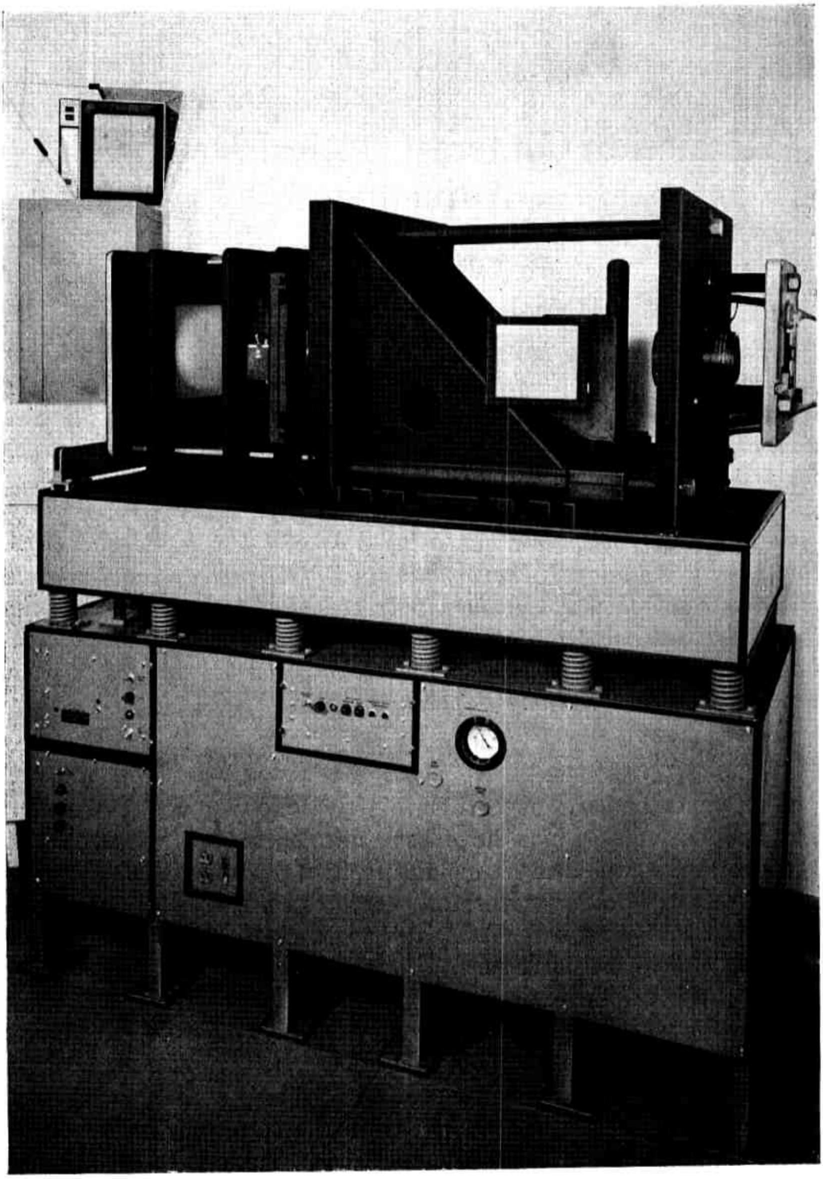


Fig. 1—The 3.5× reduction camera.

TABLE I—COMPARISON OF THE ARTWORK PLATE WITH THE OUTPUT PLATES OF THE 1.4X AND THE 3.5X CAMERAS

	Artwork	1.4X Reduction	3.5X Reduction
Size (max. nominal)	17.5 cm sq	12.5 cm sq	5 cm sq
Minimum Linewidth	35 μm	25 μm	10 μm
Use	Input to Reduction Cameras	Tantalum Thin Film Masks	Tantalum Thin Film Masks, Step-and-Repeat Reticle

minimum-width details and without introducing distortions greater than about half of the minimum linewidth. The degree to which these two requirements can be met is primarily determined by the resolution and distortion characteristics of the reduction lens; the design considerations of such lenses are a major topic in themselves, presented elsewhere in this issue.[3] Given a lens of suitable quality, however, the camera's performance is still critically dependent on three factors: (*i*) The mechanical design of the camera must be such as to maintain its performance in the presence of environmental perturbations such as vibration, changes in ambient temperature, and variations in the operators' handling techniques. (*ii*) Camera performance is dependent upon the proper design and adjustment of the illumination system. (*iii*) The performance of the camera can be no better than one's ability to adjust the completed camera for best focus and proper magnification over the whole image field, not a trivial task for cameras in this performance class. The mechanical design of the reduction cameras is discussed in the following paper. In this paper, we consider the problems of the optical design and the final adjustment.

II. OPTICAL DESIGN

Figure 2 shows the optical layout of the reduction cameras. The light source is a 100-watt mercury arc lamp operating at a pressure of about 10 atmospheres. This light source, suitably filtered, diffused, and modulated as described below, is imaged by a two-element Fresnel condenser lens onto the entrance pupil of the reduction lens. The convergent beam illuminates the artwork plate, which the reduction lens images onto the output image plate at the right. The image plate used is a Kodak Microflat High Resolution Plate with ground edges.

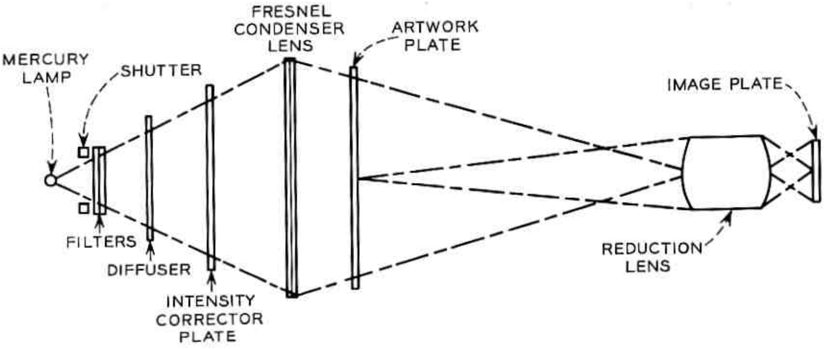


Fig. 2—Optical layout of the reduction cameras.

One of the design requirements for the two reduction lenses was that they have equal entrance pupil distances. This allowed us to use a single illumination system design for both the 1.4X and the 3.5X cameras. It was found that such a restriction could be imposed on the reduction lenses without significantly compromising their design performance.

A mercury arc light source was chosen for reliability and spectral narrowness. The 4358 Å blue line was selected rather than the 5461 Å green line because of the extra resolution afforded by the shorter wavelength, and because it left open the possibility of direct exposure onto photoresist surfaces which are sensitive to the blue but not to the green light. Although the scattering of light within the emulsion (which varies as the fourth power of the light frequency) is greater for the blue line, it is not a serious problem in this case, where the emulsion is 6 μm thick and the finest structure to be written on it is 10 μm wide.

The lamp chosen, a General Electric H100 A4/T, represents a compromise between brightness and spectral narrowness. High-pressure lamps, though brighter, exhibit sufficient pressure (Lorentz) broadening (see for example Ref. 4) of the 4358 Å line to complicate the color correction of the reduction lens, which would necessarily compromise its overall performance. The lamp brightness results in exposure times of 3–4 seconds for the 3.5X camera and 20–25 seconds for the 1.4X camera.

Two glass filters (Corning Filters No. 3389 and No. 5543) are used to isolate the Hg 4358 Å spectral line.

The image of the arc source projected onto the entrance pupil by the

Fresnel condenser, although magnified slightly by the Fresnel lens, is still too small to fill the entrance pupil. Therefore, a ground-glass diffusing screen is placed in front of the mercury lamp to increase the apparent size of the light source. The amount of this increase can be controlled by adjusting the axial position of the ground-glass diffuser. This position is adjusted until the diffused image of the source just overfills the entrance pupil.

The accumulation of reflection losses at air-glass interfaces throughout the camera results in considerable transmission loss. While this in itself is not serious, the difference in the losses experienced by paraxial rays and by rays near the edge of the field, which arises because of differences in the angles of incidence, results in an illumination intensity which falls off seriously with field angle. This intensity fall-off is compensated by the intensity corrector plate (see Fig. 2) which has a thin, vacuum-deposited absorbing layer of Inconel. The amount of deposited Inconel decreases radially from the plate center so as to compensate for the field-angle-dependent losses of the rest of the camera. Figure 3 shows

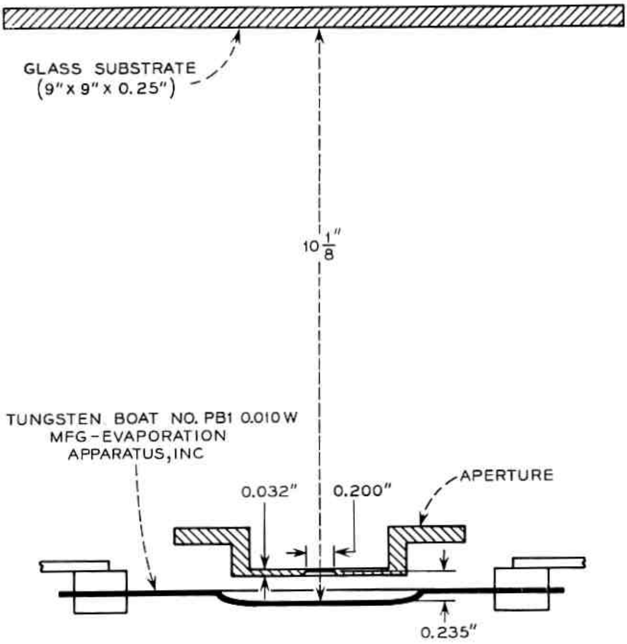


Fig. 3—Arrangement used in the vacuum evaporation of the Inconel coatings onto the intensity corrector plate.

the arrangement used in the vacuum evaporation process. The aperture defines a finite area source for the Inconel vapor. The aperture diameter and height, and the height of the glass substrate, were adjusted empirically to yield the flattest intensity distribution in the image plane. This was determined by scanning the image plane with a pinhole, integrating sphere and photomultiplier tube assembly. The result is shown in Fig. 4 which is a plot of the measured illumination intensity as a function of position within the image plane. Each horizontal scan extends beyond the active image area; the vertical tic marks on each scan delimit the actual image field. Figure 4 also shows the plane of constant intensity which best fits the measured data. It can be seen that the measured intensity distribution deviates from this plane of best fit everywhere by less than ±7 percent.

The Fresnel condenser lens* was designed specifically for these cameras and provides for zero spherical aberration at the particular object and image distances of our illumination system. The material is Rohm and Haas VM plexiglass. The lens is made up of two elements, each approximately 0.060″ thick, cemented around the rim face-to-face, as illustrated in Fig. 5. Opposing facets on the two halves have dissimilar facet angles; these angles were chosen to equalize the optical power of opposing facets, thereby minimizing reflection losses. The ability to specify the angle of each facet is equivalent to allowing general aspheric surfaces on a conventional lens. The result is that axial spherical aberration can be eliminated completely from the lens design, minimizing the problem of illumination fall-off with field angle. In addition, the ability to specify the angle of the cutback facet assures that the scattering of light by this facet will be minimal. The Fresnel lens is laterally located in the camera with three corner pins riding in radially oriented slots, so as to allow free thermal expansion without buckling or decentering.

A Fresnel condenser was chosen rather than a conventional glass condenser largely because of the difficulty in obtaining large glass lenses sufficiently free of bubbles. Such bubbles, if larger than a millimeter or so, modulate the illumination light cone at sufficiently low spatial frequencies to seriously perturb the intensity distribution in the image plane. On the other hand, the ring pattern of the Fresnel lens facets is at a sufficiently high spatial frequency that its effect is not detectable in the image plane. The moiré effects and erratic illumina-

* Designed and fabricated by the Alliance Tool and Die Co., Rochester, New York.

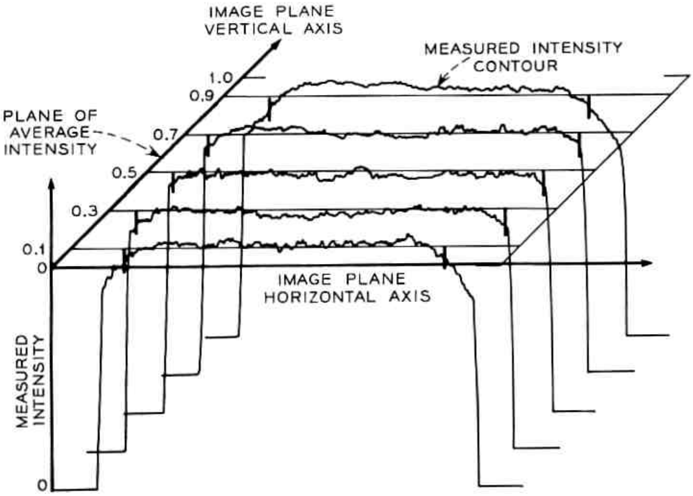


Fig. 4—Measured intensity distribution in the image plane. The maximum deviations of the measured intensity contours from the plane of average intensity are ± 7 percent.

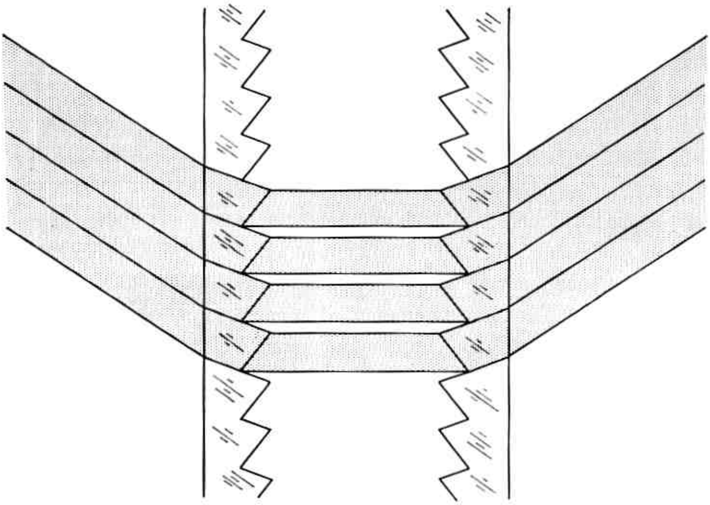


Fig. 5—Optical design of the special two-element Fresnel lens.

tion uniformity usually seen in Fresnel lens combinations are eliminated by maintaining a tight tolerance on the alignment of the two Fresnel elements during fabrication.

The design and performance of the 3.5X and the 1.4X reduction lenses are discussed in another paper in this issue.[3]

III. TEMPERATURE COMPENSATION

The reduction cameras are designed to operate in an environment in which the temperature is regulated to ±0.25°F. Nonetheless, as an additional precaution, it was decided to provide first-order compensation for changes in focus and magnification due to changes in ambient temperature. For this purpose a computer program was written* which determines the effects of temperature changes on the focus and magnification of a lens, taking into account the thermal coefficients of volume expansion, refractive index, and dispersion of each glass element, and calculating approximate changes in air gaps from the thermal expansion coefficient of the lens barrel material. It is then possible to calculate, using either the ACCOS† lens optimization program or an equivalent optimization program, how the object and image distances must change with temperature in order to maintain best focus and proper magnification. First-order temperature compensation is then attained by selecting the spacer materials to provide the appropriate thermal expansion coefficients.

IV. FOCUS AND MAGNIFICATION ADJUSTMENT

As the resolution and magnification accuracy demanded of photolithographic lenses increase, it has become apparent that traditional methods of focus and magnification adjustment are impractically slow. In order to adjust the reduction cameras properly it was necessary to develop an adjustment system which automates much of the procedure and processes much more data than would otherwise be possible. The system is broadly divisible into three parts: (*i*) a special test reticle which allows simultaneous measurement of focus and magnification errors at nine points distributed over the image plane; (*ii*) a computer-controlled, interferometric, coordinate measuring machine[5] to locate fiducial marks on the test image plates and punch their co-

* The Fortran IV TEACOPS program (*T*emperature *E*ffects *A*nalysis of *C*omplex *O*ptical *S*ystems) is available on request from the author.

† ACCOS (*A*utomatic *C*orrection of *C*omplex *O*ptical *S*ystems) is copyrighted by Scientific Calculations, Inc., Rochester, N. Y.

ordinates on paper tape; and (*iii*) a set of computer programs to analyze the paper tapes, establish the current camera errors, and calculate the necessary corrective shifts on the spacer rods. The rod adjustment mechanism is based on the elastic compression of Belleville spring washers and is described in detail in the following paper.

The test reticle is shown in Fig. 6. An 8″ × 10″ photographic plate has a test pattern consisting of horizontal and vertical bar patterns of various spatial frequencies, arranged in continuous vertical stripes. Nine glass prisms are cemented to the face of the plate in a 3 × 3 array covering the desired object field. Nine identical prisms are cemented to the back of the plate to compensate for the refraction of the illumination beam. The effect of the first nine prisms is to tilt the apparent plane of the object test pattern seen in the prism. Nine flat spacer pads displace the test reticle to the rear when mounted in the camera, so that the tilted object test patterns straddle the true

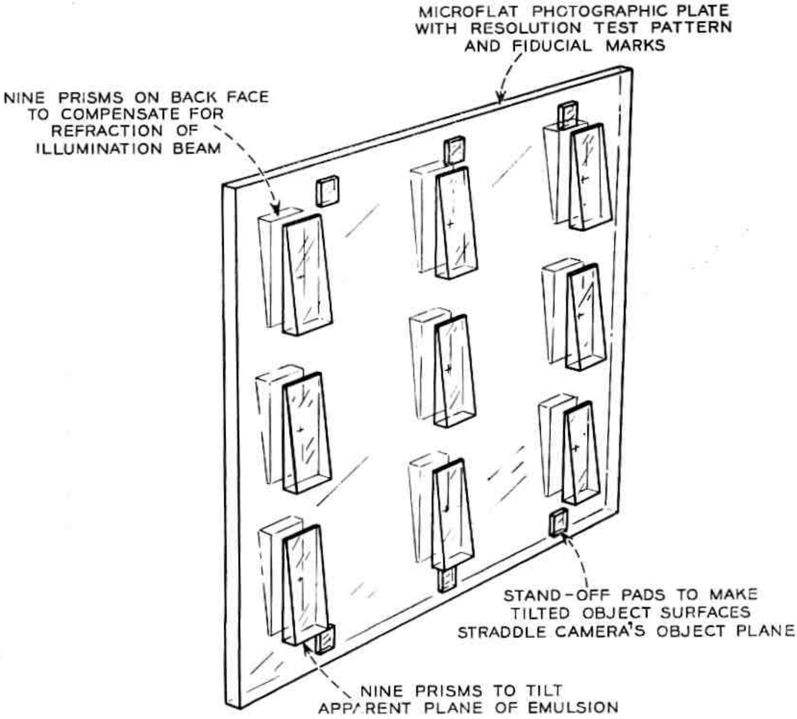

MICROFLAT PHOTOGRAPHIC PLATE WITH RESOLUTION TEST PATTERN AND FIDUCIAL MARKS
NINE PRISMS ON BACK FACE TO COMPENSATE FOR REFRACTION OF ILLUMINATION BEAM
STAND-OFF PADS TO MAKE TILTED OBJECT SURFACES STRADDLE CAMERA'S OBJECT PLANE
NINE PRISMS TO TILT APPARENT PLANE OF EMULSION

Fig 6—The special test reticle used for simultaneous adjustment of focus and magnification.

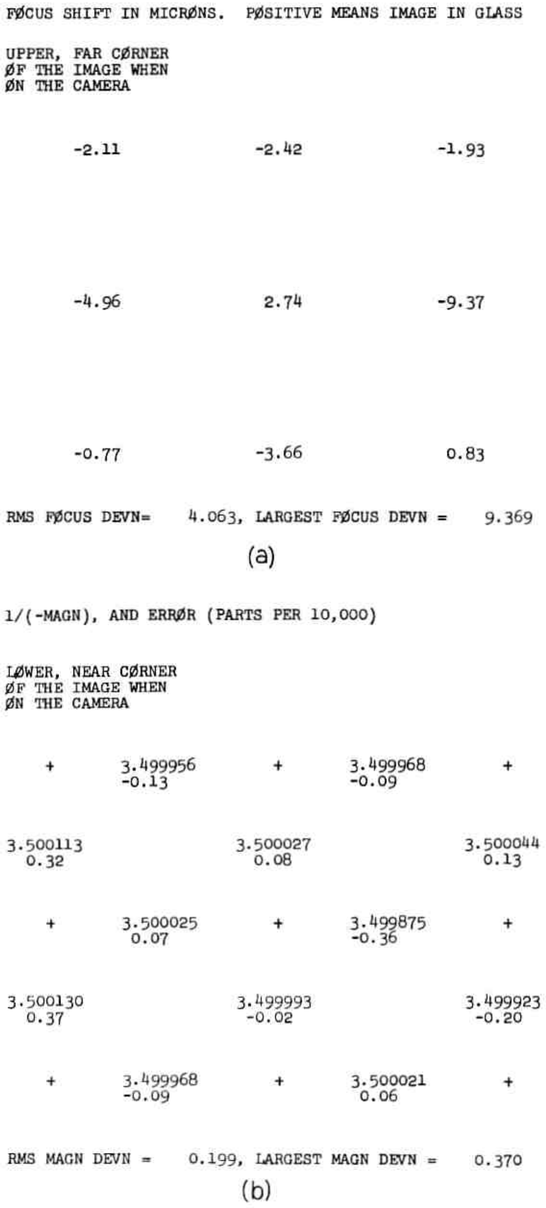
object plane. In addition, fiducial marks are placed in the center of each prism field, to be used to check magnification.

The test reticle is placed in the camera and a test image plate is exposed and developed. Measurements are made, in each of the nine prism areas on the image plate, of the vertical position where the pattern appears sharpest. Additionally, the coordinates of the nine fiducial marks are measured. This information, when compared with reference measurements made on the test reticle itself, yields the focus errors in each of the nine prism regions and twelve magnification errors, corresponding to twelve distances between the nine fiducial marks. Figure 7 shows focus and magnification error maps printed by our time-sharing computer program. The data shown is that for a well-adjusted 3.5X camera. Note that the magnification errors, Fig. 7(b), are at worst 0.37:10,000 and average 0.20:10,000. This is to be compared to the maximum allowable error of 1:10,000. Similarly, the focus errors, Fig. 7(a), are at worst 9.4 μm and average 4.1 μm. These numbers are comparable to the diffraction limited depth of focus.

Other parts of the time-shared computer program calculate (using paraxial optical equations) the object- and image-distance shifts necessary to bring each of the nine prism regions into best focus and magnification. These shifts are then fitted (using the method of least squares) onto tilted and axially displaced object and image planes. Finally, the program calculates the length changes required on each of the six spacer rods to bring the existing object and image planes into conjunction with the desired planes.

In general, approximately 6–8 iterations of this correction cycle are required to bring the camera into adjustment such as is shown in Fig. 7. During the last few iterations, a modified procedure is followed in which a test reticle without prisms is used in addition to the prism test reticle: the former provides the magnification error data, and the latter provides the focus error data, as before. This procedure eliminates the small magnification error introduced by the prisms. Such errors amount to about 1:10,000 and can be neglected during the first several iterations.

Figure 8 compares a resolution test pattern and the corresponding image taken with a well-adjusted 3.5X camera. The narrow lines in each of the five "L" patterns are (on the reduction camera plate) 4, 6, 8, 10, and 12 μm. (The finest detail required in normal operation is 10 μm.) It can be seen that the 4 μm detail is adequately resolved



```
FØCUS SHIFT IN MICRØNS.  PØSITIVE MEANS IMAGE IN GLASS

UPPER, FAR CØRNER
ØF THE IMAGE WHEN
ØN THE CAMERA


       -2.11               -2.42               -1.93




       -4.96                2.74               -9.37




       -0.77               -3.66                0.83

RMS FØCUS DEVN=     4.063, LARGEST FØCUS DEVN =     9.369
                            (a)

1/(-MAGN), AND ERRØR (PARTS PER 10,000)


LØWER, NEAR CØRNER
ØF THE IMAGE WHEN
ØN THE CAMERA

    +        3.499956        +        3.499968        +
             -0.13                    -0.09

3.500113                 3.500027                 3.500044
  0.32                     0.08                     0.13

    +        3.500025        +        3.499875        +
              0.07                    -0.36

3.500130                 3.499993                 3.499923
  0.37                    -0.02                    -0.20

    +        3.499968        +        3.500021        +
             -0.09                     0.06

RMS MAGN DEVN =     0.199, LARGEST MAGN DEVN =     0.370
                           (b)
```

Fig. 7—Computer-generated maps of focus errors (a) and magnification errors (b) for a well-adjusted 3.5× reduction camera.

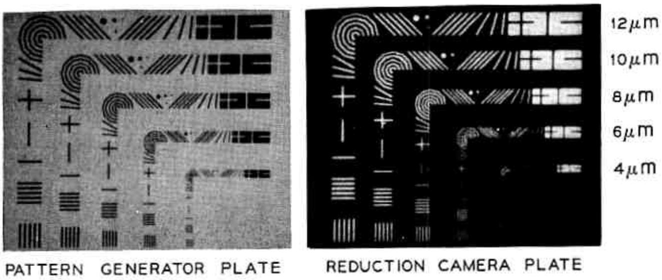


Fig. 8—Photomicrographs of an artwork resolution test pattern (left) and the corresponding image plate (right) taken with the 3.5× camera. The image-plane linewidths are indicated at the right. The finest image linewidth required in normal use is 10μ.

and that the 10 μm (fundamental) detail is well resolved with sharp edges.

V. ACKNOWLEDGMENTS

I would like to acknowledge the assistance of R. G. Murray who carried out the focus and magnification procedure and E. T. Doherty for his work on the intensity corrector plate and the illumination system alignment.

REFERENCES

1. Poole, K. M., et al. "The Primary Pattern Generator," B.S.T.J., this issue, pp. 2031–2076.
2. Alles, D. S., et al., "The Step-and-Repeat Camera," B.S.T.J., this issue, pp. 2145–2177.
3. Herriot, D. R., "Lenses for the Photolithographic System," B.S.T.J., this issue, pp. 2105–2116.
4. Mitchell, A., and Zemansky, M., *Resonance Radiation and Excited Atoms,* Cambridge, Mass.: Cambridge University Press, 1961. p. 98.
5. Ashley, F. R., Murphy, Miss E. B., and Savard H. J., Jr., "A Computer Controlled Coordinate Measuring Machine," B.S.T.J., this issue, pp. 2193–2202.

Reduction Cameras: Mechanical Design of the 3.5X and 1.4X Reduction Cameras

By M. E. POULSEN and J. W. STAFFORD

(Manuscript received July 14, 1970)

I. INTRODUCTION

The 3.5X and 1.4X reduction cameras basically employ the same structural features differing only in the lenses and focal distances required to achieve the desired reductions. Both cameras have been designed as fixed-focus cameras in that no adjustment is made on individual components to optimize the focus and magnification.

The camera incorporates the following design features: (*i*) isolation of the camera from building vibrations; (*ii*) temperature compensation in the long and short conjugates to compensate for changes in the lens due to changes in the ambient temperature; (*iii*) sufficient structural mass of individual components and material conductivity to avoid local distortions due to rapid changes in ambient temperatures; (*iv*) artwork and image plates automatically positioned to within an accuracy of about one micron; and (*v*) exposure control which can be varied, with a high degree of reproducibility.

II. PHYSICAL DESCRIPTION OF CAMERAS

A rigid camera bed supports the elements as shown in Fig. 1. The camera bed is mounted on springs to provide vibrational isolation. The welded frame which supports the camera bed-spring assembly contains the pneumatic controls, lamp power supply, shutter-control electronics, and the Mask Shop Information System (MSIS) lamp and read-out supplies.

The camera bed is made of two GA50 Meehanite cast iron I-beams which are connected laterally. The faces of the I-beams have been ground flat and parallel in pairs to provide an accurate support for

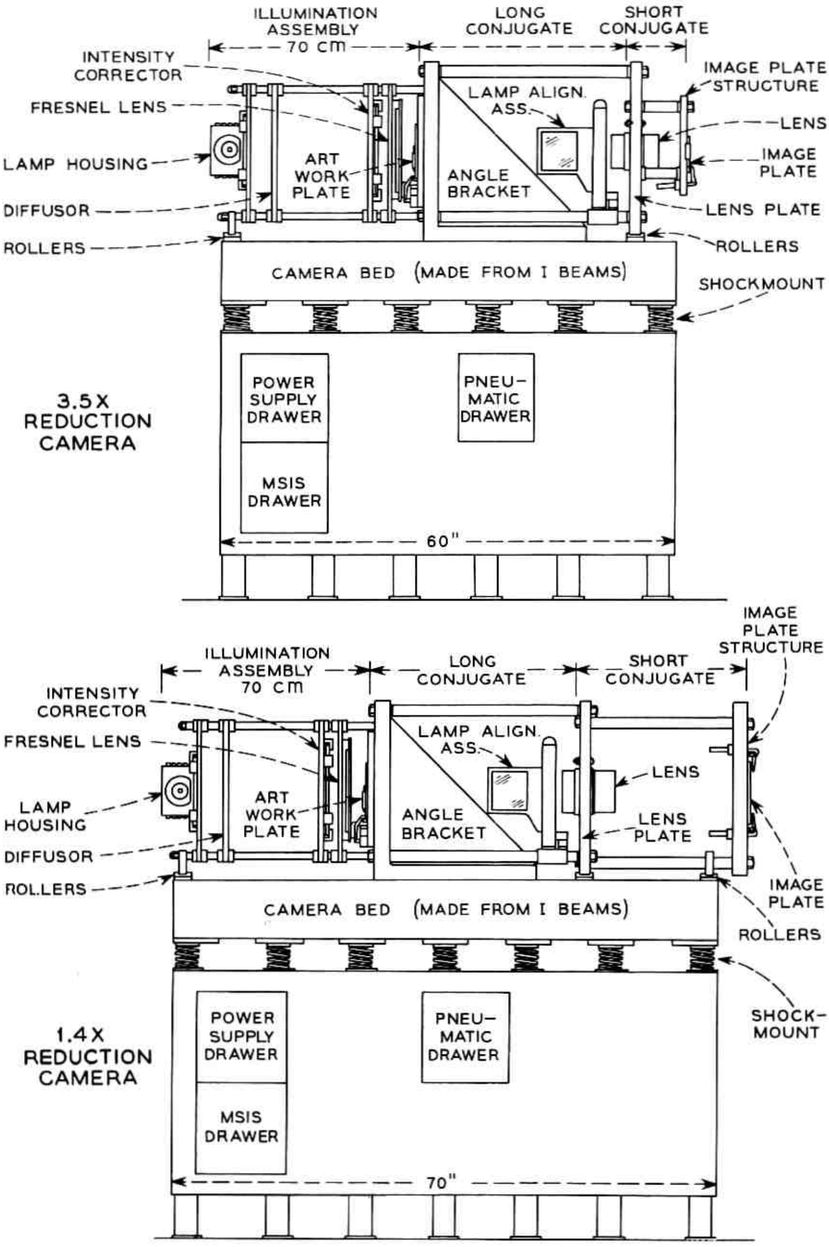


Fig. 1—Schematic of 3.5X and 1.4X reduction cameras.

the camera elements. As shown in Fig. 1, a large gusseted Meehanite angle bracket is bolted and pinned to the camera. This bracket provides the fixed support for the illumination assembly and the lens-and-image plate assembly. As will be discussed later, the rods of the conjugates are machined along with the assembled lens-and-image plates to obtain the correct theoretical lengths so as to yield the desired focus and magnification. Fine trimming of the lengths is performed utilizing a computer program until the optimum lengths are achieved (see Ref. 1).

The lens plate is mounted on rollers and is free to move along the camera bed with changes in ambient temperature. Similarly, the end of the rod supporting the illumination assembly is mounted on rollers. The illumination assembly consists of the fresnel lens, intensity corrector, diffusion screen, and lamp-housing shutter assembly.

As shown in Fig. 1, a roller support is provided for the image-plate structure of the 1.4X camera. On the 3.5X camera, this additional support is not needed because of the relatively small short-conjugate distance. Figures 2 and 3 are photographs of the 3.5X and 1.4X reduction cameras.

III. VIBRATION ISOLATION

Providing a vibration-free environment is essential if high-quality reductions are to be made. If excited, vibration of the camera bed would result in bending of the bed in many modes and thus could destroy the focus and magnification of the camera along with the alignment of its image relative to the artwork. To eliminate this, the 3.5X and 1.4X camera beds were designed to have a free-free natural frequency of 100 Hz and the bed shock mounted on springs to yield a rigid body natural frequency of 3 Hz. Reference 2 shows that if this is the case the natural frequency of the bed coupled to the springs is the 3 Hz rigid body mode with the next resonant frequency occurring at 100 Hz and other frequencies occurring from 100 Hz on up. From Fig. 4 taken from Ref. 3 one can see that if the exciting frequency ω is three times greater than the rigid-body natural frequency ω_n, the rigid body is essentially isolated from the perturbing force. Normally, for most building floor slabs one can expect floor slabs to have a resonance of from 12 to 15 Hz and the foundation (i.e., base slab or cellar) to have a resonance of around 30 Hz or higher. Hence, mounting the reduction camera bed at 3 Hz isolates it from all the disturbing building frequencies above 10 Hz. Furthermore, because of the rigidity of the camera bed, it will behave

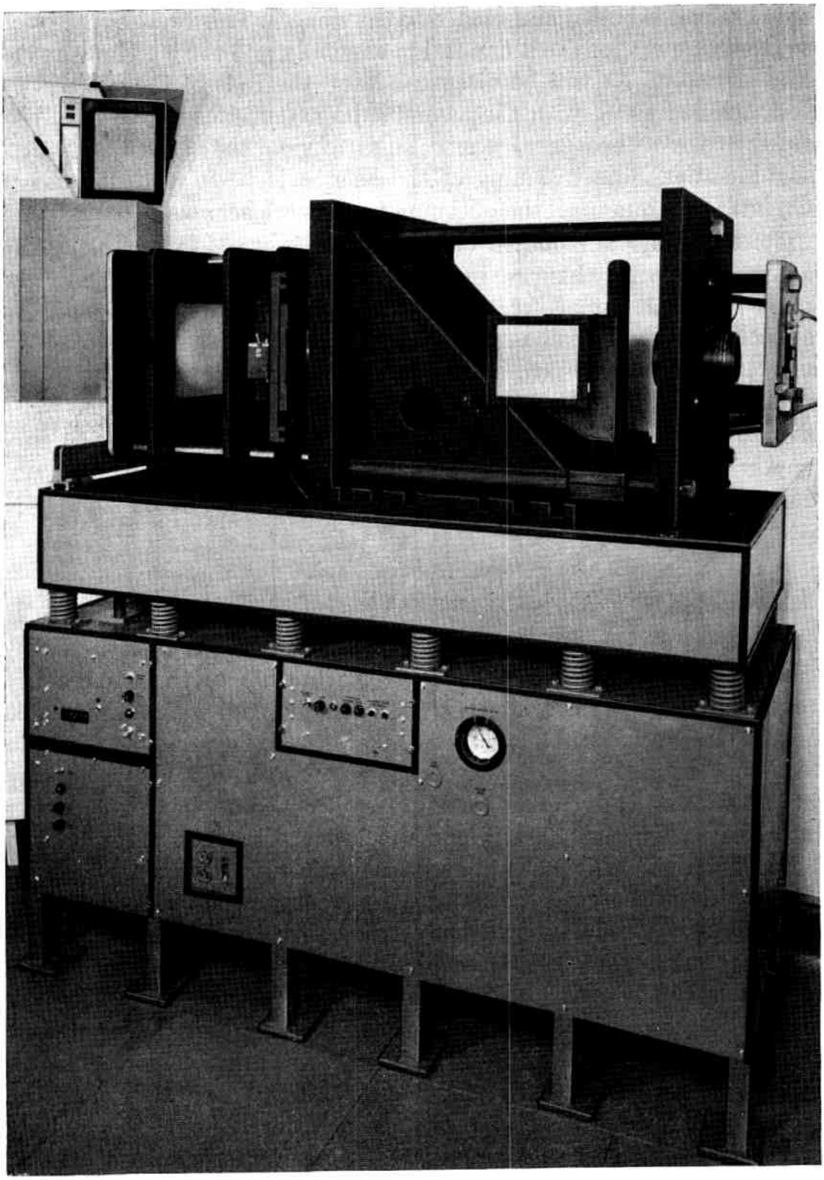


Fig. 2—3.5X reduction camera.

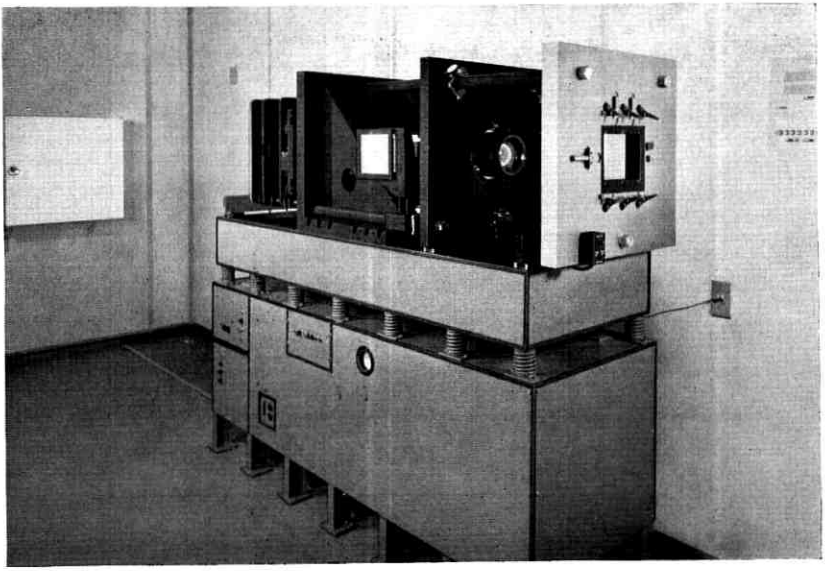


Fig. 3—1.4X reduction camera.

as a rigid body if subjected to excitation below 10 Hz. The 3-Hz shock mounts are provided with inorganic (metal mesh) damping and the camera bed I-beams have an elastomeric damping compound bounded on their webs to provide damping should the camera bed be inadvertently excited.

The frame to which this shock-mounted camera is attached was designed so that its resonant frequency is 50 Hz, hence eliminating any possibility of the support structure being the source of a rigid body excitation near 3 Hz.

The three support rods of the illumination system have a natural frequency of 40 Hz in the lowest mode which is lateral bending. The three rods of the long and short conjugates have a natural frequency of 200 Hz.

IV. THERMAL DESIGN CONSIDERATIONS

The reduction cameras are operated in a clean room which is temperature-controlled to within ±0.15°C to maintain artwork and image sizes as well as their relative positions. To further assure reproducability even under more adverse ambient condtions, additional design features were incorporated. The rod material of the long and short

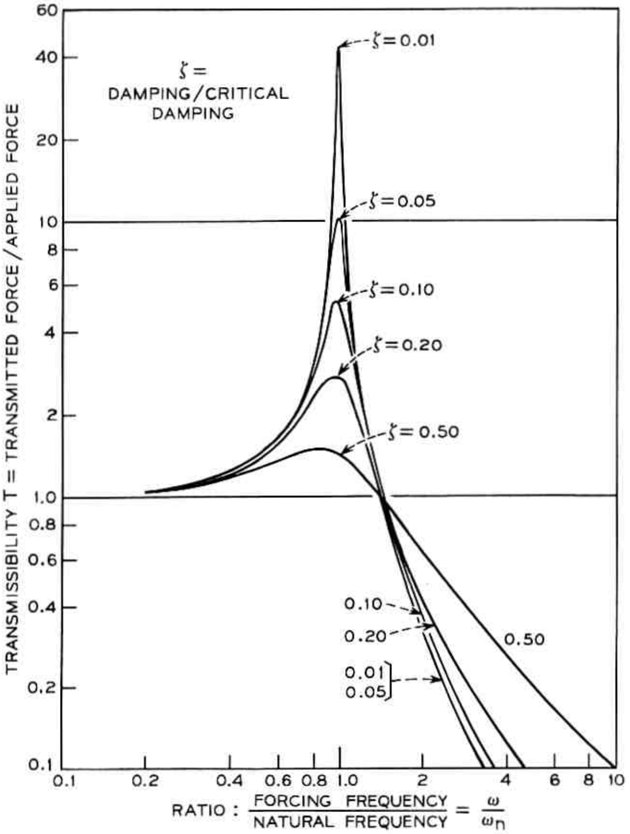


Fig. 4—Transmissibility versus frequency for a single-degree-of-freedom rigid-body system.

conjugates on both cameras was selected to compensate for both focus and magnification errors due to the effect of temperature fluctuations of the lens itself over ±5°C. The length of the long and short conjugates vary with temperature in a prescribed manner to accomplish this compensation. The variation is linear with temperature and is obtained by selecting the rod material with the appropriate coefficient of thermal expansion.

The good conductivity of the GA50 Meehanite and the large mass of the bed insures that only negligible thermal gradients through the bed structure will be encountered and, hence, bending distortion of the bed is effectively eliminated. The bed temperature will change uniformly should a change in room temperature occur, thus, preventing degradation of focus, magnification, and artwork-image alignment.

V. MATERIAL SELECTION

GA50 Meehanite was selected for the I-beams and angle bracket of the camera bed because of its good dimensional stability with time and its good conductivity. Both the I-beams and angle bracket were furnace annealed prior to rough machining and given a vibration stress relief after rough machining and prior to final machining. This was done to insure the stability of the parts.

The illumination element holders were made from ground tool plate which was annealed to avoid warpage during final machining.

The lens-holder plate and image-plate structure were made from AZ-31 magnesuim plate which has exceptionally good dimensional stability. This material provided a rigid yet lightweight structure.

For the 3.5X reduction camera, the rods of the long conjugate were made from Hastelloy X and those of the short conjugate from a 49 percent nickel iron alloy. These materials were selected because they had coefficients of thermal expansion which provided the required temperature compensation for the lens.

For the 1.4X reduction camera, the long conjugate rods were made from a composite two-material rod of Invar 36 and 49 percent nickel iron alloy, and the short conjugate made from a composite two-material rod of stainless steel and 49 percent nickel iron alloy to obtain the appropriate coefficient of thermal expansion.

VI. ADJUSTMENT OF THE LONG AND SHORT CONJUGATES

A relatively gross adjustment in the mils range (i.e., 10^{-3}″ range) has been provided on both the long and short conjugates of the reduction cameras. In addition, an adjustment in the microinch range (i.e., 10^{-6}″ range) has also been provided utilizing the technique developed for the mirror of the PPG (see Ref. 4). The gross adjustment is provided by compressing Belleville springs as shown in Fig. 5, and the fine adjustment uses elastic compression of rectangular pads into the metal surface to provide the microinch adjustment, the soft spring being the bolt itself as shown in Fig. 5.

The long conjugate is bolted to the angle bracket reference surface. This end contains the pad washers which provide the microinch adjustment. The other end of the long conjugate is bolted to the lens plate and contains the Belleville springs used for gross adjustment.

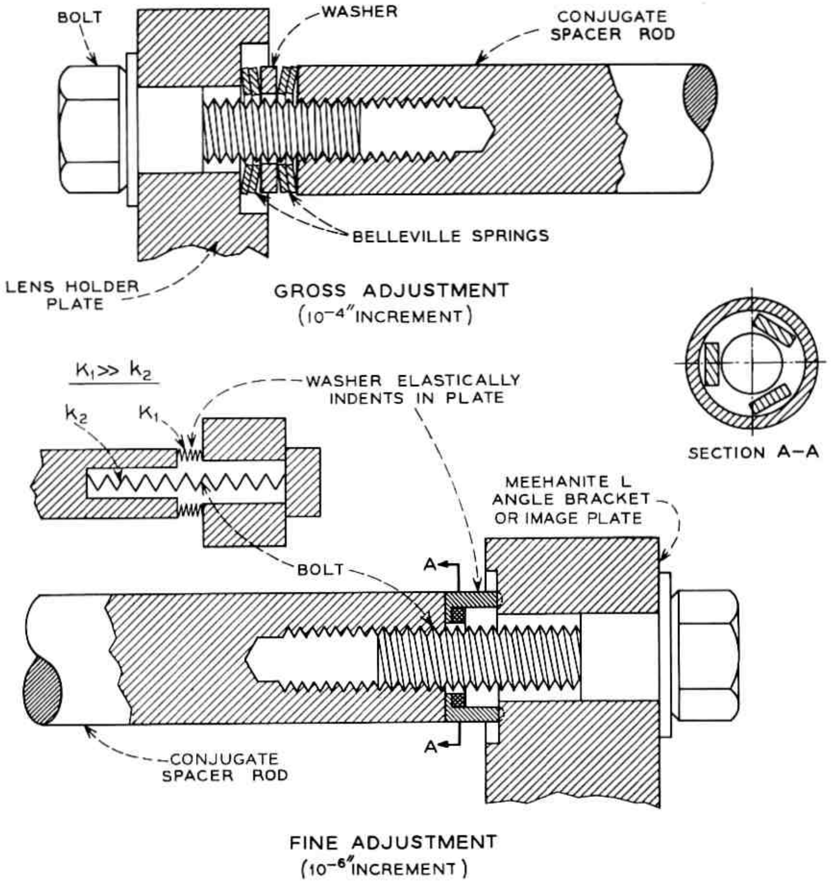


Fig. 5—Adjustment mechanism for conjugates of reduction camera.

The short conjugate is bolted to the lens plate and to the image-plate structure. The end bolted to the lens plate contains, the Belleville washers for gross adjustment and the end bolted to the image-plate structure contains the pad washer for microinch adjustment. The gross adjustment provided for each conjugate rod at the lens plate is monitored with a dial indicator (later removed) capable of being read accurately to within 0.0001″ and having a range of ±0.01″. This permits the adjustment of the long and short conjugate rods according to the computer program discussed in Ref. 1. The pad-washer end of each conjugate provides the fine adjustments in the microinch range within a range of ±0.00025″. For the current reduction cameras it was not necessary to use the microinch adjustment to bring the camera into focus and magnification.

VII. ARTWORK AND IMAGE PLATE POSITIONING

Both the artwork and image plates must be positioned reproducibly against the locating pins. The artwork plate, locating pins on the camera are positioned and constructed exactly as they are on the pattern generator. Similarly, the image plate is positioned against pins which are constructed exactly as they are on the artwork side of the step and repeat camera.

To position the artwork and image plates in the reduction camera successfully, the applied forces holding plates against their location pin must be greater than the frictional forces. A static analysis, knowing the coefficient of friction, allows one to adjust the relative forces in the horizontal, the vertical, and the axial directions, such that the plate will always seat.

The artwork is placed into a vertical holder and pneumatically held. This holder, supported on bearings, is then pushed into the camera. Upon contacting a microswitch, the plate is released from the holder and clamped against vertical and horizontal pins. The holder is ejected and the plate then located onto the axial pins (i.e., pins parallel to the optical axis). To accomplish this, a system of miniature pneumatic cylinders utilizing dry nitrogen are used (see Fig. 6). The image plate is also located pneumatically. The operations are controlled by pneumatic and electrical components in a drawer located in the bench assembly (see Fig. 7).

Artwork and image plates have been loaded repeatedly into the cameras. Statistical analysis of the data shows that the plates index reproducibly. For the nominal eight-inch by ten-inch, one-quarter-inch-thick artwork plate it was found that the plates had a mean seating error of from eight to thirteen microinches depending on the axis measured, with a standard deviation of from five to ten microinches. For the nominal four-inch by five-inch, one-quarter-inch-thick image plate, it was found that the plates had a mean seating error of from three to four microinches depending on the axis measured, with a standard deviation of from two to eight microinches.

VIII. SHUTTER

For both cameras, it was deemed desirable to be able to vary the exposure time from 30 ms to 100 s. To provide uniform exposure, it

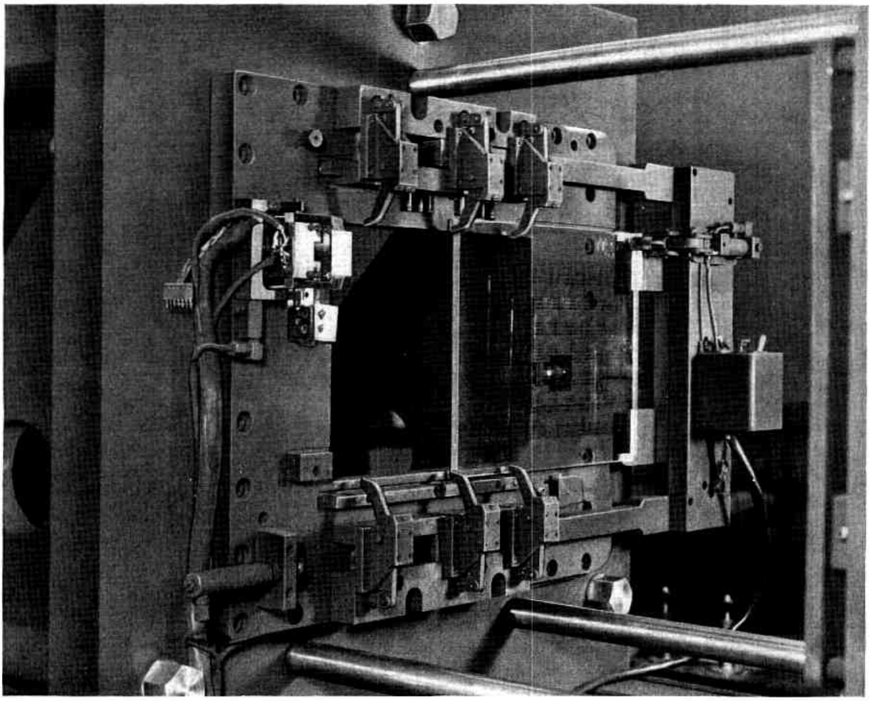


Fig. 6—Insertion mechanism for artwork.

is necessary to have opening and closing speeds which are small compared to the exposure. It was not possible to purchase a shutter having the required 6.3-cm aperture and the necessary range and precision of exposure with an opening and closing time of 10 ms.

A commercial spring-activated shutter was modified to meet this requirement. The case A and plate B were retained as shown in Fig. 8. The leaves were reinforced in the high impacted area, and the leaf-activating mechanism was designed to ride on ball bearings to reduce the frictional forces. The driving mechanism consists of an opening and closing solenoid with their armatures joined at the driving arm of the shutter. The solenoids with the shutter are mounted on the lamp housing, aligned, and pinned. At the ends of the armatures are damping cushions to reduce bounce, and at the ends of the solenoids are adjusting screws to insure that the impact force is not absorbed by the shutter leaf rotating slot. The shutter driving arm is coupled to the solenoid armatures through a slot. The slot is longer than the

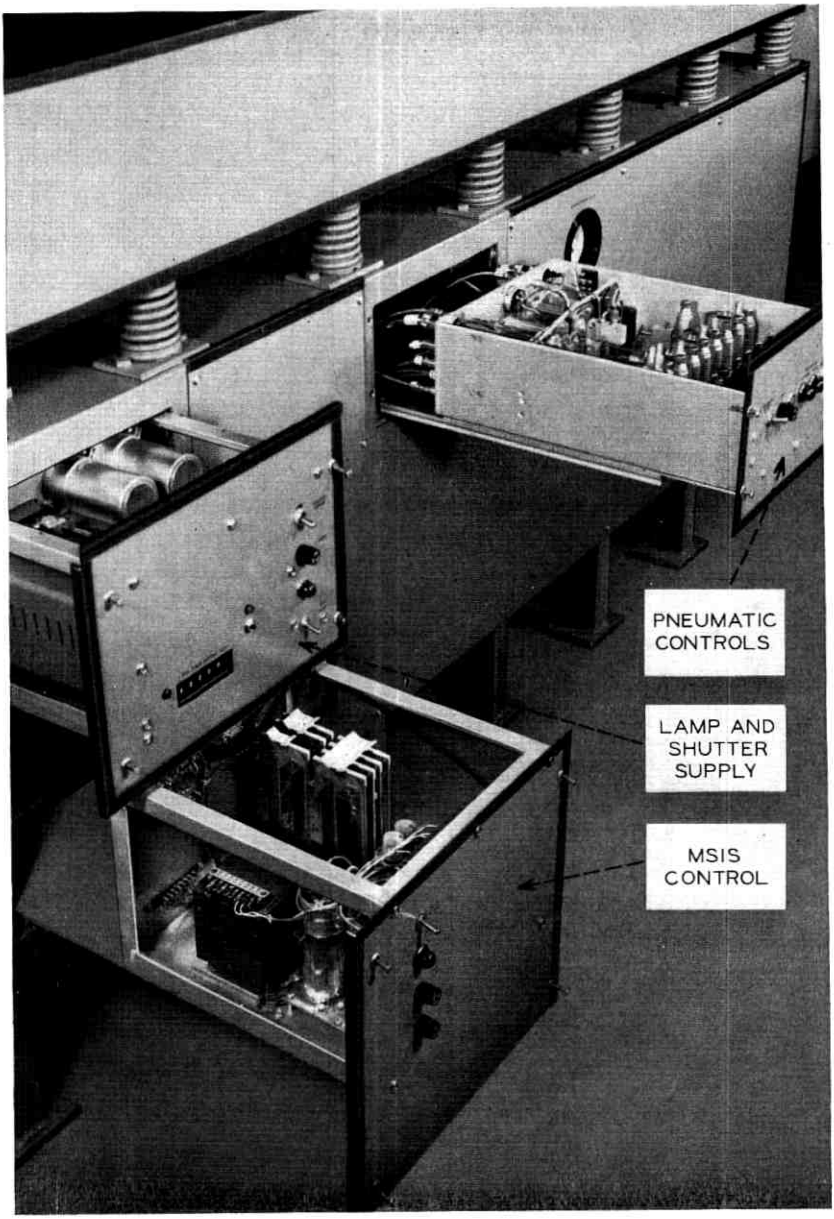


Fig. 7—Pneumatic and electrical controls.

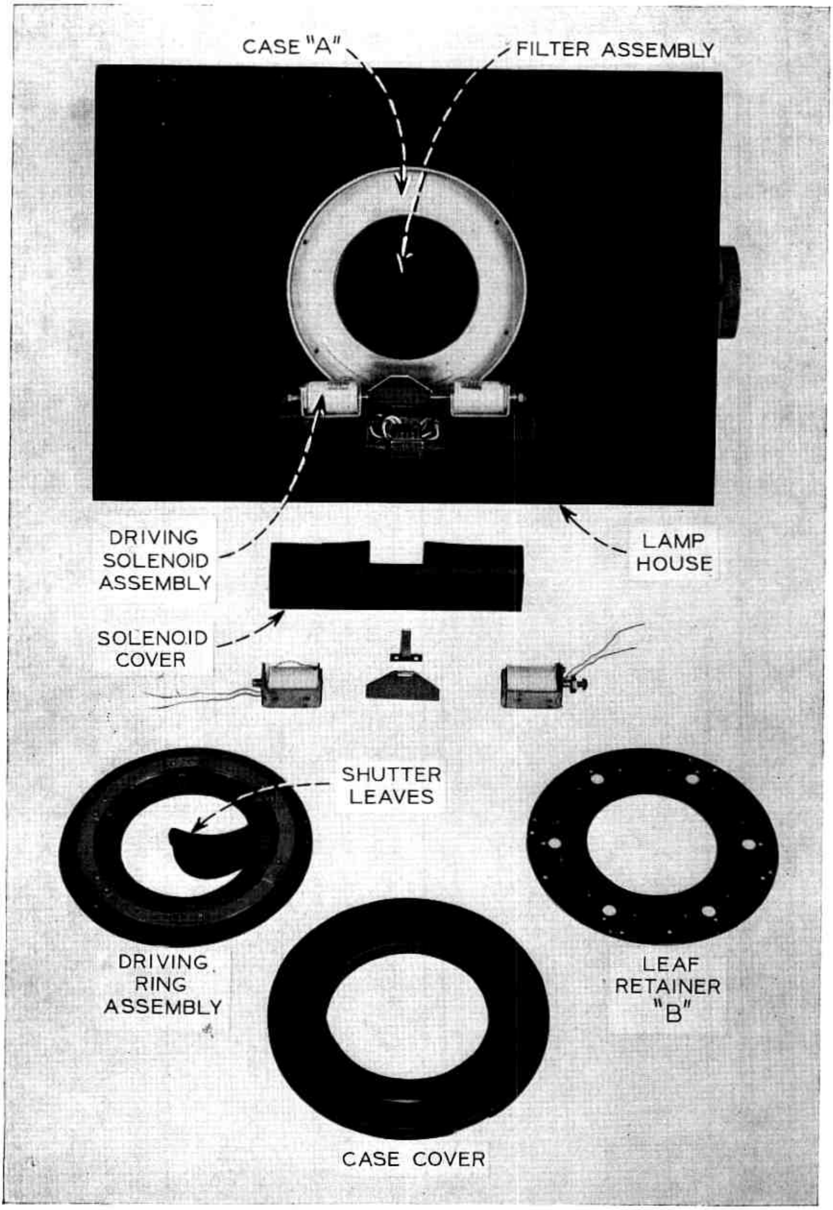


Fig. 8—Shutter assembly and components.

actuating arm is wide, providing for a 1/16″ space at one end towards the activated solenoid. The armature travels 1/16″ before contacting the activating arm, thereby accelerating without the load of the shutter mechanism. The total armature travel necessary to achieve the shutter fully open or closed is 0.188″. The opening and closing time takes 10 to 12 ms, depending on the friction in the assembly. This is accomplished by the solenoid operating at five times its rated voltage. Since the duty cycle is very low, this causes no damage. The solenoid voltage is applied for 30 ms. The additional 20 ms is necessary to keep the shutter from bouncing and to provide damping-down time. The shutter solenoids are controlled by a digital timer, consisting of a 1-kHz crystal clock oscillator, a five-decade-selector switch, and associated integrated circuitry.

All shutters are acceptance tested to 3000 cycles; life-tested shutters have run over 100,000 cycles. The life-tested shutter showed signs of wear but no signs of imminent failure.

IX. MASK SHOP INFORMATION SYSTEM

The primary pattern generator (PPG) records identification codes on the plate. This information is used both for visual inspection and for automatic identification in the reduction cameras. This consists of human-readable and machine-readable information. The machine-readable information is encoded as a series of clear or opaque rectangles located outside the primary pattern area. This binary information is read in the camera by a linear array of phototransistors and sent to the MSIS computer which verifies that the proper plate has been loaded.

The detector array is made up of 8 silicon chips, each with six phototransistors positioned in a row on a gold interconnection pattern on a sapphire substrate. This is mounted on a Bakelite assembly and attached to the artwork support structure (see Fig. 9). The diodes are located 80 mils from the emulsion side of the artwork plates. On the opposite side of the artwork plate is located a special illuminator housing consisting of lamps and condenser lenses. To prevent the artwork plate from striking the MSIS illuminator housing, it was necessary to swivel the illuminator housing out of the way during artwork insertion. This was accomplished by means of pneumatic cylinders actuated in conjunction with the plate clamping pneumatics. Since space did not permit an in-line illuminator source, the housing was set off to the side and the light beams were brought into line

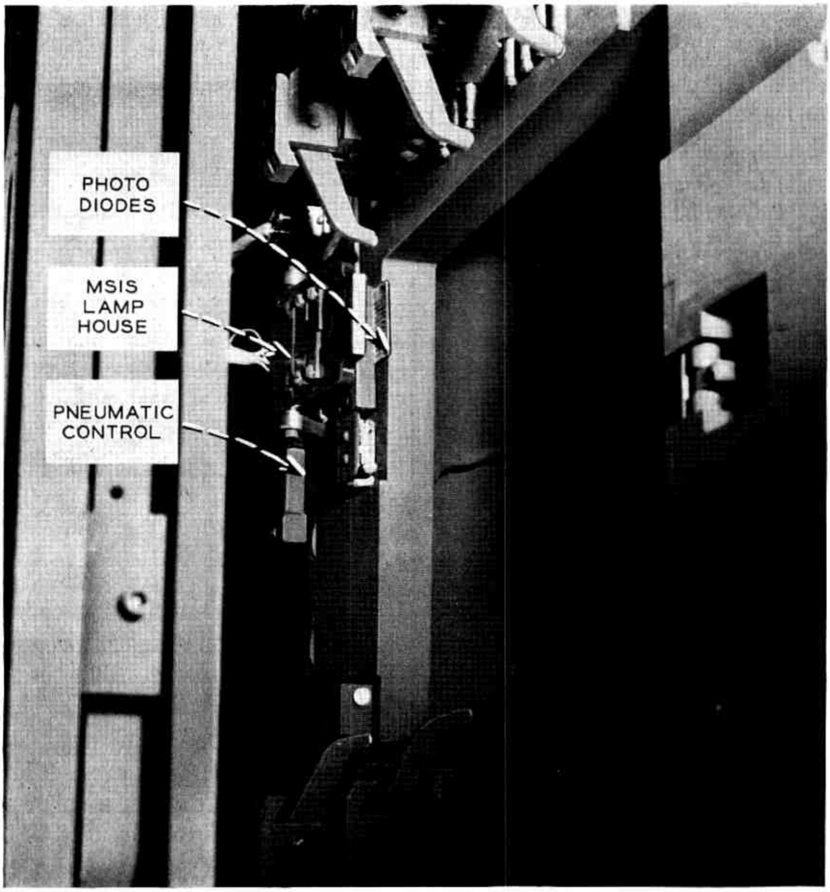


Fig. 9—MSIS assembly.

by means of a prism. Each pair of silicon chips (12 phototransistors) is illuminated by one of four lamps independently switched on by the computer. Each lamp is separated by septums to prevent interference, and each lamp has a lens to image the filament at infinity. The collimated beam is directed into the prism which reflects the light through a linear array of twelve fly-eye lenses which in turn illuminate the twelve phototransistors through the information strip on the artwork plate. The four lamps are turned on in sequence, and the information is sent to the MSIS computer.

X. SUMMARY

High-quality reduction cameras have been designed which are unperturbed by normal building vibrations and which, due to their mass, good material conductivity and temperature compensation of the conjugates are unaffected by reasonable changes in the ambient temperature. Insertion of the artwork and image is reproducible. A high-speed, wide-aperture shutter, capable of being opened or closed in 10 to 12 ms, has also been designed with a life of over 100,000 cycles.

XI. ACKNOWLEDGMENT

The authors acknowledge the work and assistance performed by D. H. Berry on the MSIS circuitry.

REFERENCES

1. Rawson, E. G., "Reduction Cameras: Optical Design and Adjustment," B.S.T.J., this issue, pp. 2117–2128.
2. Stafford, J. W., "Natural Frequencies of Beams and Plates on An Elastic Foundation with a Constant Modulus," J. Franklin Institute, *209*, No. 4 (October 1967), pp. 262–264.
3. *Shock and Vibration Handbook,* edited by C. M. Harris and C. E. Crede, New York: McGraw-Hill, 1961, Vol. 1, pp. 2–12.
4. Kossyk, G. J. W., Laico, J. P., Rongved, L., and Stafford, J. W., "The Primary Pattern Generator: Mechanical Design," B.S.T.J., this issue, pp. 2043–2059.

The Step-and-Repeat Camera

By D. S. ALLES, J. W. ELEK, F. L. HOWLAND, B. NEVIS, R. J. NIELSEN, W. A. SCHLEGEL, J. G. SKINNER and C. E. STOUT, Jr.

(Manuscript received July 10, 1970)

We discuss in this paper the design of a new high-precision step-and-repeat camera with respect to its optics, mechanical design, control system, and control computer program. One micrometer images from a 5-mm-square lens field can be placed within 0.12 μm over a 10-cm × 10-cm area on photographic glass plate. Features such as, image plane control, interferometric metering, and automatic reticle pattern alignment, are used to accomplish these objectives. The control computer with CRT message displays for the operator result in an efficient operator-machine interaction.

I. INTRODUCTION

In previous papers, the equipment for converting the designer's topography into a primary pattern and the subsequent reduction in size have been described. For thin film integrated circuits, the output of the reduction camera is the master mask from which working copies can be produced for use in fabricating the device. For semiconductor devices, however, the output of the reduction camera is ten times larger than the required final image size. Thus, a further reduction in size is required. In addition, a mask for a semiconductor device consists of an array of images that are precisely placed on the master mask. Thus, the step-and-repeat camera is both a reduction camera and, through the use of a moving X-Y stage, permits the placement of images in an array covering the desired field of the mask.

1.1 *Requirements*

If a final mask consisted of a single image and if only one mask level were required to produce a functioning semiconductor device, the step-and-repeat camera would be a simple tool to design and build. In actuality, a mask is complex. As shown in Fig. 1, a mask consists of an array of images, the majority of which are the primary images required for fabricating the specific device. In addition to the primary images, a wide range of test, secondary continuity and alignment images used during the fabrication process are included. Thus, before a mask can be produced on the step-and-repeat camera, all of the reticles containing the required images must be available.

In fabricating a semiconductor device, multiple masks are needed, each corresponding to a specific processing step. Typically, nine to twelve distinct levels are required for integrated circuits. For the mask set to be useful, the images in the various levels must be in registration from one level to the next. As device features have become smaller, the requirement for registration of the mask images from level to level

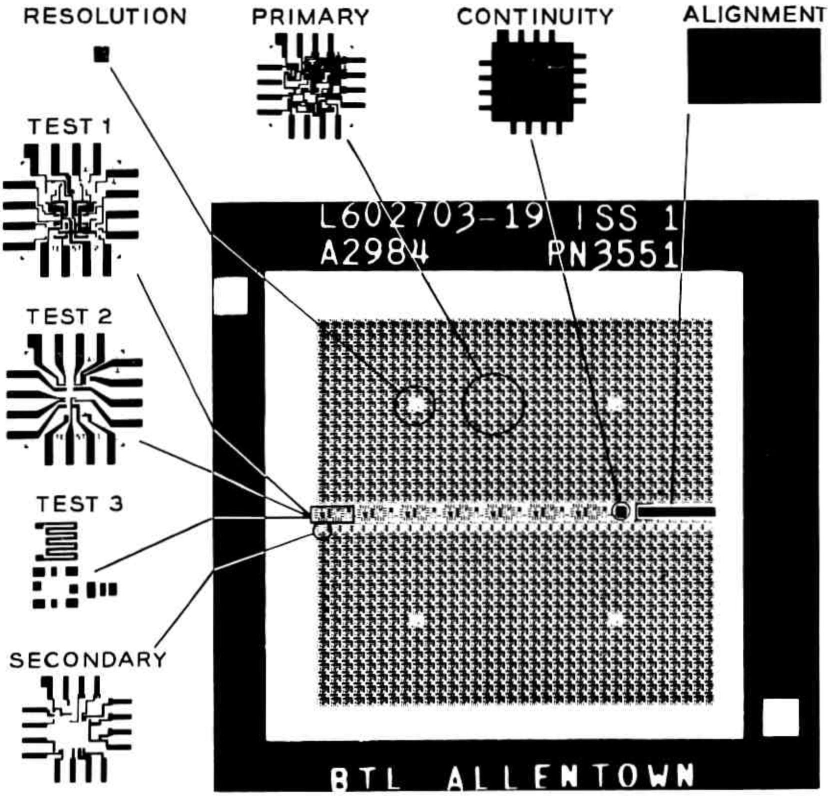


Fig. 1—Typical integrated-circuit photo mask and its various patterns.

has become more stringent. In addition, certain classes of devices such as the silicon target for the *Picturephone*® camera tube, though basically simple, must be produced by field-butting techniques that require the step-and-repeat camera to provide precise image placement.

The design objectives for the step-and-repeat camera are summarized in Table I. The first three items, resolution, image field size and distortion, are established by the optics of the system. The image placement accuracy and array size, items 4 and 5, are determined by the mechanical stage and the position sensing and control system. The last item, the operating time, is of interest because it relates to the balanced operational capability of the entire mask-making system.

Past experience with step-and-repeat camera operations has revealed that a camera capable of the performance listed in Table I does not guarantee error-free operation, i.e., high yield. The major cause of low camera output is operator error. Thus considerable attention has been given to eliminating, where possible, operator tasks that have been shown to result in errors.

1.2 *General Description*

The step-and-repeat camera described in the following sections, is a single-head, ten-times-reduction camera mounted over a moving X-Y stage supported and guided by air bearings. Table-position is determined using double-pass interferometers for both X and Y axes. The physical arrangement of the completely assembled camera is shown in Fig. 2. The physical size is approximately 1.2 m in width and depth and 1.5 m high. The camera and stage systems are on the operator's left; operator displays and controls are on his right.

The glass photographic plate on which the mask will be made is positioned in a fixture on the camera's stage. The reticle whose pattern will be projected onto the photographic plate is located below the hinged cover which supports the flash lamp and condenser housing. The camera's status and operator instructions are given on the lighted message board and all normal operator controls are provided on the operator keyboard. A more extensive set of controls is available for camera maintenance.

All of the camera functions are controlled and monitored by a computer which is located outside the camera's temperature-controlled clean room. The use of a small computer rather than a hard-wired controller has allowed the camera's operation to be flexible, and provides nearly automatic operation with a minimum of operator intervention.

The mask-making sequence is initiated by the operator's request

TABLE I—DESIGN GOALS FOR THE STEP-AND-REPEAT CAMERA

1. Lens Resolution	1 μm
2. Maximum Image Size	5 mm square
3. Image Distortion	0.1 μm
4. Image-Placement Accuracy	0.12 μm
5. Maximum Size of the Array	10 cm
6. Typical Step-and-Repeat Time	1200 s

(at the operator keyboard) for a new job. The computer in turn requests a job from the Mask Shop Information System (MSIS). The list of required reticles and other pertinent information is displayed for the operator on a CRT. The operator loads both the photographic plate and the correct reticle and then commands the camera to continue. The array information is transmitted from the MSIS computer and the pattern is step-and-repeated. When the mask is completed or a new reticle is needed, the operator is alerted by a message on the display board

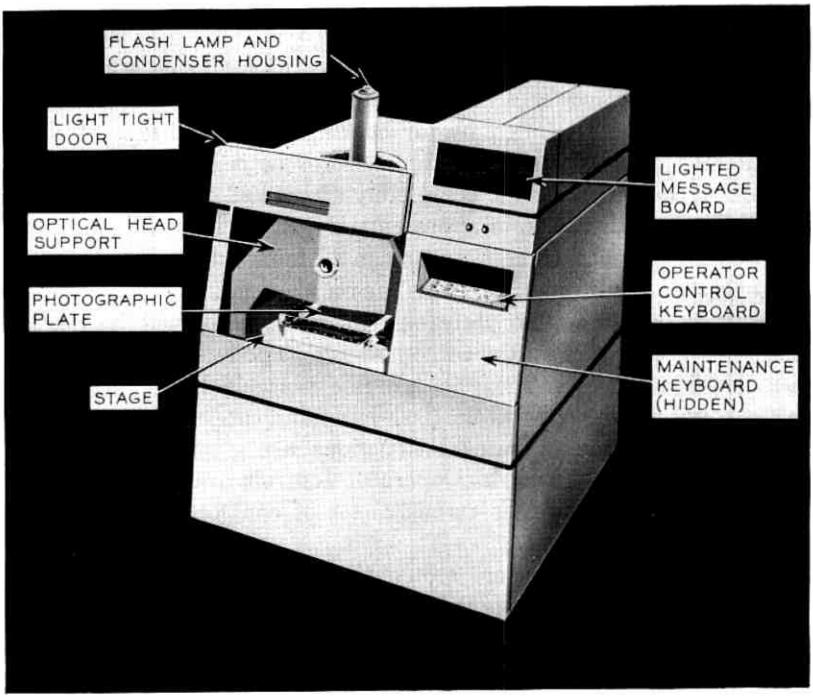

FLASH LAMP AND CONDENSER HOUSING
LIGHT TIGHT DOOR
OPTICAL HEAD SUPPORT
PHOTOGRAPHIC PLATE
STAGE
LIGHTED MESSAGE BOARD
OPERATOR CONTROL KEYBOARD
MAINTENANCE KEYBOARD (HIDDEN)

Fig. 2—The step-and-repeat camera.

and by an auditory alarm. Operator errors or equipment malfunctions are also indicated on the message board.

The following discussion of the camera is divided into four major sections: Section II, Optical Head Assembly; Section III, Stage Assembly; Section IV, Control System; and Section V, Program. Although these are discussed separately, the design of all systems, both hardware and software, were developed simultaneously with close collaboration between personnel to assure a smoothly functioning camera design.

II. OPTICAL HEAD ASSEMBLY

2.1 *General Features*

The optical head assembly is a complete unit containing all the optics and electronics necessary to project a pattern on to the mask. The salient features of the optical head are:

(*i*) It projects a 5-mm-square image with a line-width capability of 1 μm on photographic emulsion.

(*ii*) It can deliver four times the energy needed to expose dyed KHRP plates.

(*iii*) It has the power capacity to project six formats per second.

(*iv*) Exposures are made while the table is in motion with negligible line-width errors.

(*v*) It has a built-in auxillary projection system to facilitate centering the flash-lamp.

(*vi*) The reticle is held in place by six pneumatic plungers which operate in a programmed sequence to ensure that it is correctly positioned against its support pads. It is then optically automatically aligned to an accuracy of ±0.25 μm, which is equivalent to a positioning accuracy of ±0.025 μm in the image plane.

(*vii*) The reticle number is electronically identified after it is aligned.

(*viii*) A 5 × 7 lamp array can be projected through the main projection lens for writing identifying alpha-numeric information on the mask.

(*ix*) The main structure is thermally compensated to maintain focus and magnification over a temperature range of ±3°C.

(*x*) The assembly is "fixed-focus" with no external adjustments.

(*xi*) The lens is automatically protected from above by a shutter whenever the reticle is removed.

2.2 *Reticle Format*

The reticle for the step-and-repeat camera is the 4″ × 5″ output plate from the 3.5X reduction camera. The reticle format, shown in Fig. 3, has a pattern area of 5.2 cm × 6.4 cm with the corner bounded by a radius of 3.536 cm. At one end of the format is a secondary information strip containing 42 binary bits each 0.5 mm square. The data recorded in this strip is the drawing number of the reticle pattern.

At the other end of the format are two fiducial marks that are used for automatic-positioning of the reticle when it is mounted in the camera.

2.3 *Projection Lens*

The lens, which was designed and manufactured by Tropel, Inc.,* to Bell Telephone Laboratories specifications, is a nine-element double-Gauss design (See Fig. 4). The object to image distance is 48.37 cm at a 10 : 1 image reduction, the effective focal length is 4.12 cm, the f-number is 1.4 at infinity and the spectral range is 436 nm ± 7.5 nm. One of the design criteria for the lens was to maintain a large working distance between the lens and the image plane to provide for the lens air bearing which is described in Section 3.1.

The computed image distortion does not exceed 0.1 μm at any point in the projected field. The calculated modulation transfer function (MTF) in the center, and at the edge of the field of view is shown in Fig. 5. Allowing for a slight decrease in these values due to manufacturing errors, the lens can produce 1-micron lines having a MTF of 0.4, which is adequate for exposing KHRP plates.

2.4 *Illuminating & Condenser System*

The purpose of the condenser system is to provide adequate illumination in the correct spectral range uniformly over the projected field. In the step-and-repeat camera, the exposure is made while the table is moving so as to reduce the time required to make a mask. Thus it is necessary to use a short exposure time to minimize the image blur.

In this camera the illumination is supplied by an EG&G, FX-76 flash lamp with a special glass envelope having a ripple-free front surface to minimize illumination non-uniformities. The lamp has a maximum rated power input of 15 W which is sufficient to expose six patterns per second on dyed KHRP emulsion. The flash duration is about 15 μs

* Located in Fairport, New York.

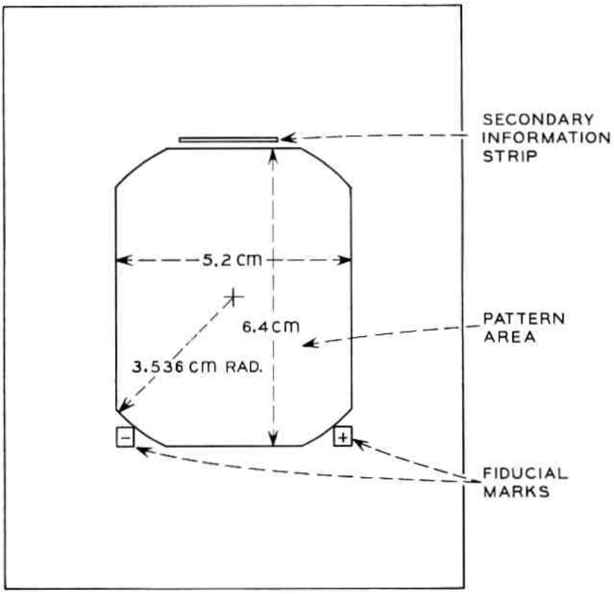


Fig. 3—Reticle format.

which allows a maximum table velocity of 0.5 cm per second with a line edge blur of less than 0.1 μm. Another point of concern is that the jitter-time between the computer command and the onset of the flash should be small and constant; experiments of over a quarter of a million flashes show that the positioning error due to the jitter-time is less than 0.005 μm. The maximum rated energy input to the flash lamp is 10 J which is four times that required to expose dyed KHRP emulsion.

A six-element condenser assembly with a large collection angle images the flash-lamp discharge onto the entrance pupil of the projection lens. The assembly contains a filter having a transmission bandwidth of 15 nm centered at 436 nm and having a uniform transmission to within 5 percent over its working area.

2.5 *Mechanical Construction*

The optical head, including its pneumatic control system and reticle-positioning electronics, is mounted on a Meehanite casting which bridges the interferometrically controlled stage. The head contains the projection optics, flash lamp, reticle positioning system, alpha-numeric

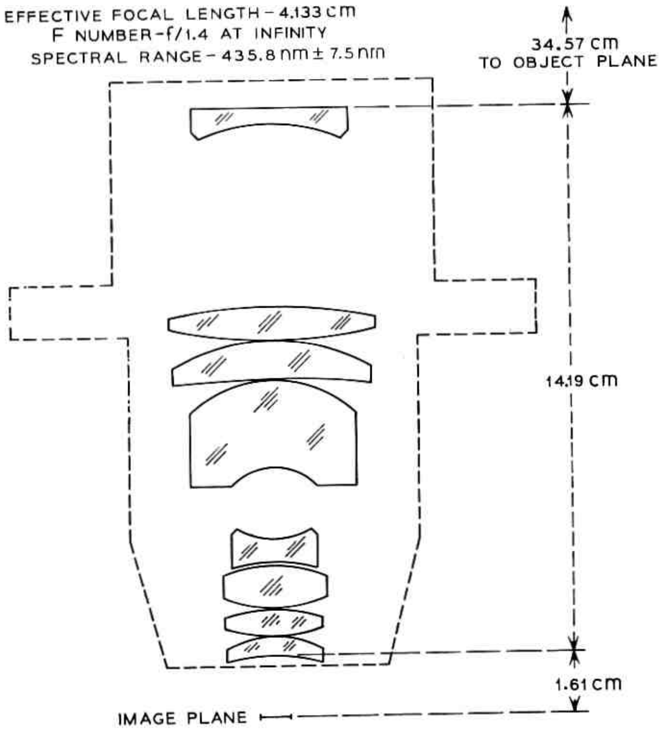

EFFECTIVE FOCAL LENGTH – 4.133 cm
F NUMBER – f/1.4 AT INFINITY
SPECTRAL RANGE – 435.8 nm ± 7.5 nm
34.57 cm
TO OBJECT PLANE
14.19 cm
1.61 cm
IMAGE PLANE

Fig. 4—Projection lens.

writing system and the machine language reading detectors (See Fig. 6).

The structure consists of three parallel plates separated by rods, the center or main plate being fastened to the top of the bridge. The upper, or reticle plate is supported on slender rods which permit lateral displacement of the plate for positioning the reticle. The plate displacement is controlled by stepping motors driving a low angle cam through a large gear reduction. This drive mechanism is mounted on the cylindrical housing that encloses the space from the main plate to above the reticle plate. A pneumatically operated cover, incorporating the flash lamp and condenser system, is hinged at the rear of this housing.

The lower, or lens, plate is suspended from the main plate by three rods and attached to the underside of the bridge by a diaphragm that prevents vibrations normal to and around the optical axis but perimts displacement along the axis. A housing, which provides the mounting for the projection lens, the lens air bearing and the retro-reflectors for the reference legs of the interferometer (See Section 3.4), is attached to the lower surface of the lens plate.

The plate and rod structure is designed to reduce the thickness tolerances on the plates. This has been done by grinding flat one reference surface on each plate. The rods then terminate on only these surfaces (See Fig. 6). Where required, the rods are inserted through holes in the plate and fastened to steel disks screwed to the reference surface. A similar construction is employed in mounting the projection lens, i.e., the lens housing is fastened to the reference surface of the lens plate and the lens flange is screwed to the interface surface of the lens housing.

The changes in the lens conjugates necessary to maintain the focus and magnification over a temperature range of ±3°C was calculated from the lens and lens holder design. These values were then used in selecting the materials for the rods and the lens housing so that the focus and magnification are compensated over the specified temperature range.

The head is aligned and focussed before it is installed in the bridge. The reference surfaces of the plates are set parallel to each other within

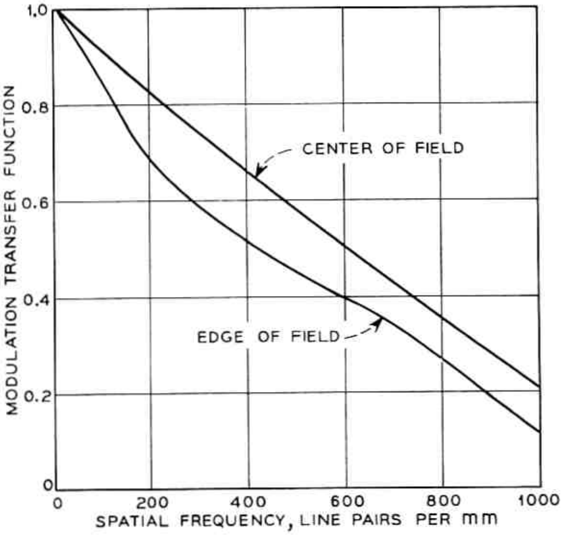


Fig. 5—Calculated modulation transfer function curves for the projection lens.

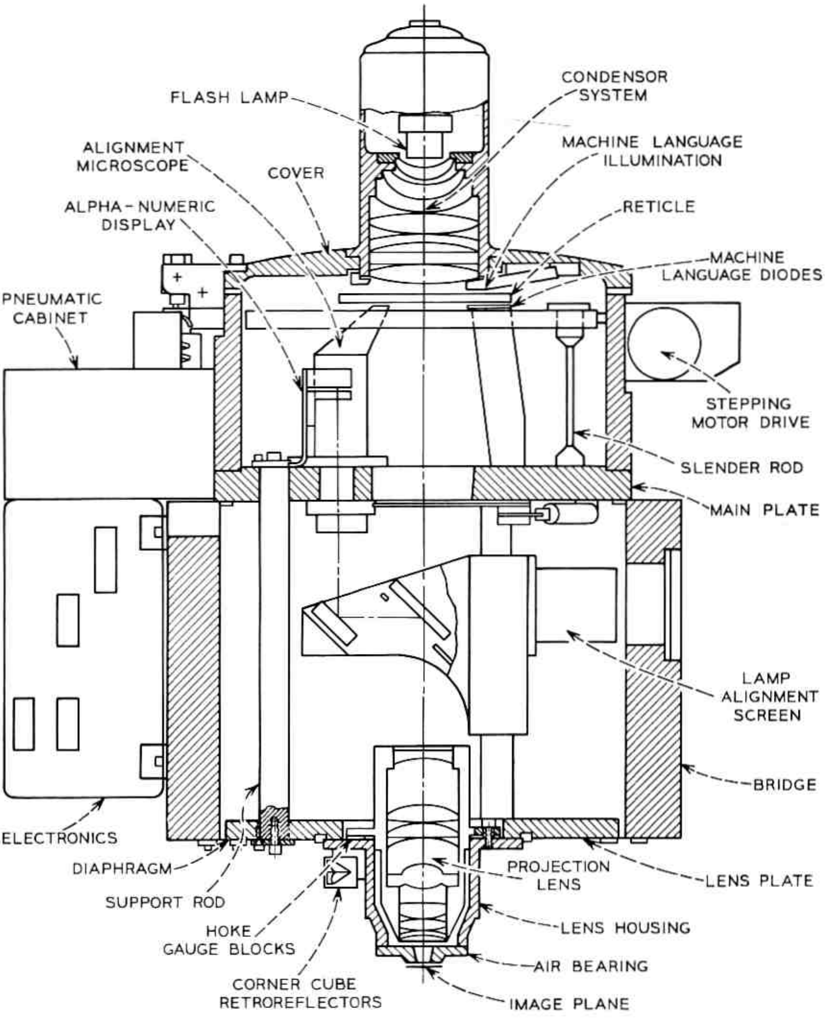


Fig. 6—Optical head assembly.

25 μrad by adjusting the length of the rods. Centering of all parts, except the reticle plate, is controlled by the initial machining. The reticle plate is offset toward the stepping motor drives so that when the plate is centered the slender rods are flexed. The flexure of these rods causes the plate to maintain contact with the positioning drives.

The lens conjugates were measured on an optical bench and the rods machined to these dimensions. This procedure can give only approximate positioning and adjustments are therefore provided in the design. The lens-flange-to-air-bearing distance is adjusted by interposing Hoke gauge blocks under the lens flange. This permits controlled adjustment in 2.5 μm step and angular adjustments of 50 μrad. Final adjustments are made by varying the air space between the lens air bearing and the mask. The long conjugate is varied by changing three pads that support the reticle. A 0.5-mm range in 50-μm steps is provided. Since these adjustments are critical, they are not available for operator manipulation, i.e., the camera is a fixed focus and magnification instrument.

The air bearing holds the mask to within 0.25 μm of its theoretical position (See Section 3.1) which provides excellent control for image quality and introduces an image distortion of only 6 parts per million (ppm).

2.6 *Reticle Alignment*

The reticle is secured to its mount by six pneumatic plungers fed through throttling valves to give them a programmed sequence to insure that the reticle is correctly positioned against the locating pads. The locating pads are in the same location as those used in the reduction camera to help minimize the centering errors. The two fiducial marks on the reticle are imaged with a 5.5X magnification on to two EG&G, SGD-444 photodiodes. One fiducial mark is a cross pattern and is imaged on to a quadrant detector, for X-Y positioning, and the second mark is a straight bar pattern that is imaged on to a bi-cell detector, for θ orientation. The arms of the X-Y fiducial mark are 30 μm wide by 710 μm long, and θ fiducial mark is 40 μm wide by 1500 μm long. The X-Y mark is imaged on to the quadrant photo-diode as shown in Fig. 7.

The microscopes that image the fiducial marks on to the photodiodes incorporate the following features:

(*i*) A bent optical path, provided by two dove prisms, prevents the structure of the microscope from infringing upon the main pattern projection area.

(*ii*) Diode illumination, from an external light source via fiber optics, for visual alignment of the diodes.

(*iii*) External rotation adjustment of diodes for initial alignment.

(*iv*) A refractor block, which can be adjusted through the reticle plate, for aligning the pattern with the X-Y axes of the table to within an accuracy of 5 μrad.

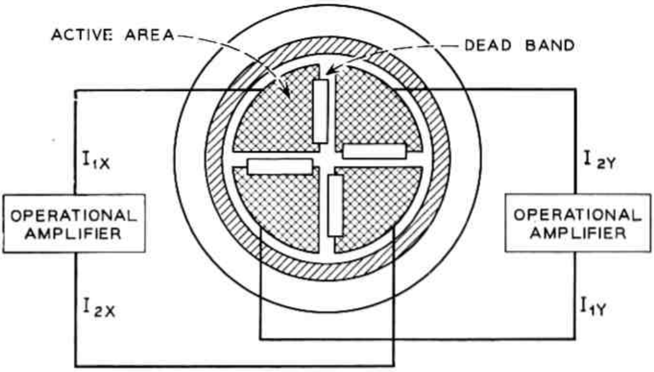


Fig. 7—Photo diode.

The output current from each quadrant in the photo-diodes is proportional to the amount of light focussed on to it, and the difference in output from conjugate quadrants is a measure of the reticle displacement. A schematic of the electronics is shown in Fig. 8. The diode outputs are amplified and converted to a voltage signal to give a reticle displacement sensitivity, near the balance point, of 1.6 V/μm. This signal is monitored by a voltage comparator having a dead-space window of ±0.2 V which is equivalent to a dead-band of 0.25 μm. The signal-to-noise ratio at the balance position is about 20 to 1 thus ensuring adequate signal strength. The outputs from the comparators

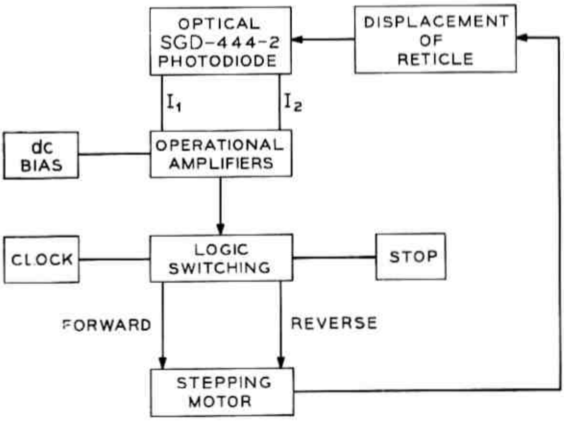


Fig. 8—Block diagram of reticle alignment electronics.

control switching logic which in turn drives three stepping motors. These move the reticle via the three low-angle cams to the balance position. One step of the motor displaces the reticle 0.02 μm. Experimentally, the system positioning accuracy was measured to be ±0.25 μm. The total range of motion is ±125 μm; however, in order to avoid driving the cam through its transition region, the pattern on the reticle is required to be within ±63 μm of its correct position. The total balancing time is about 15 s, which includes a 5-s pause at the end of the positioning period to ensure that the system is completely stable. The switching logic is then shut down, the stepping motors are put in the "hold" position, but the power to the amplifier and the photodiodes remains on at all times. Measurements to date indicate an overall drift of about 0.2 μm per day, and the drift during the period required to expose a complete mask is negligible. The complete electronic package includes a sensitivity calibration system that displaces the reticle one micrometer and measures the corresponding output voltage.

2.7 *Machine Language Read-Out*

The secondary information strip is monitored by the computer when the reticle has been positioned. The 42 bits are read as four separate groups for the convenience of transferring the information, each group having its own lamp and condenser system. The lamp is imaged on to each bit-area by means of a flys-eye lens array mounted just above the reticle, and under each bit-area is a photo-transistor for monitoring whether the bit area is clear or opaque.

2.8 *Alpha-Numeric Display*

A 5 by 7 lamp array is available for producing characters 2 mm high in the center of the field of view of the projection lens. A flys-eye lens array with 35 lenses each 2.5 mm square is placed just in front of the lamp array. The effect of the lens array is to increase the amount of light collected from each lamp, as opposed to using no lens array at all, and to image each lamp as a square on the mask. The required exposure time is 1 s which requires that the mask be stationary during the exposure.

2.9 *Operation*

The reticle is loaded in the camera by the operator and is clamped into position by activating an air switch. The camera cover is closed by an air-piston, activated by the operator, and the system is then transferred over to computer-control. The computer initiates the reticle-positioning electronics which then proceeds to center the reticle. If for any reason the reticle has not been positioned after a given length of time, the computer turns off the servo-system and informs the operator. When the computer has been informed that the reticle is correctly positioned, it will then read the secondary information strip to ensure that the correct plate is in the camera. The optical head is now ready for operation.

III. STAGE ASSEMBLY

3.1 *Focus Control*

For a lens capable of 1.0-μm line resolution and precise control of magnification, the depth of focus is less than 1.0 μm. Since photographic plates having a submicrometer flatness are unavailable, an automatic focusing system is required to assure well-focused images over the entire step-and-repeat array. The flattest commercially available photographic plates for use in this camera, are Mirco Flat KHRP.* These have an over-all flatness of 6.5 and 16 μm respectively for the 4″ × 5″ and 8″ × 10″ plates. If the surface at the plate's perimeter is positioned at the image plane, the remaining portions of the plate will be above or below the image plane by as much as the flatness specification and the resulting images will be unacceptable. In addition to the plate's lack of flatness, its emulsion is 6 μm thick, thus only a thin layer of the emulsion can be in focus, the rest being out of focus. This difficulty was overcome by R. E. Kerwin[1] who dyed the emulsion with a material which absorbs strongly at 436 nm. The dyed plate is then only exposed at its top surface and it is only this surface which must be maintained in the lens image plane.

To maintain the plate in focus, it is mounted in a softly suspended fixture attached to the stage. The plate's emulsion surface is allowed to glide under a stiff air bearing which is attached to the lens housing. Since the lens air bearing is in effect a very stiff spring and the plate support system is relatively soft, the distance between the photographic plate and the lens bearing will be nearly constant for large deflections of the plate holding system. From Fig. 9 it is possible to predict the performance of this focus control scheme. If K_f and K_b are the stiffnesses of the plate support and the lens bearing and ΔY is the deflection of the plate supporting fixture due to photo plate distortion, the resulting change in plate-to-lens bearing force is Δf. This results in a plate-

* Manufactured by Eastman Kodak, Inc., Rochester, New York.

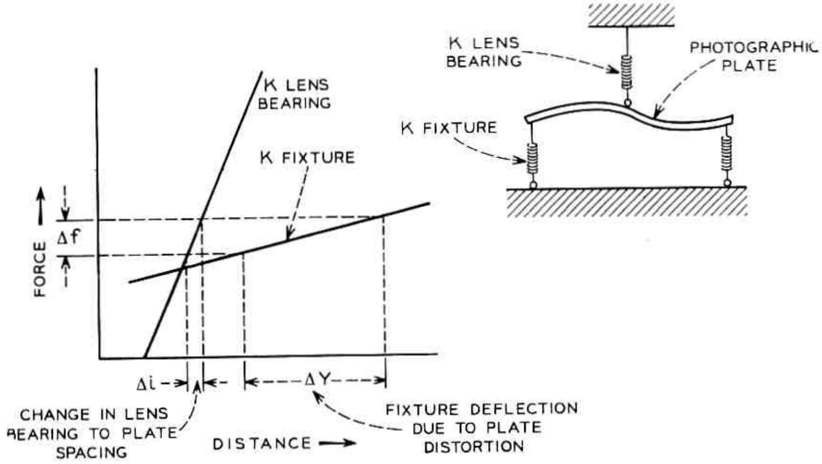


Fig. 9—Theoretical prediction of focus control system performance.

to-lens spacing change of Δi. For a given plate distortion the error in the emulsion surface location is proportional to K_f/K_b, which in practice is about 40. This results in focus errors of 0.15 and 0.3 μm for 4″ × 5″ and 8″ × 10″ plates respectively.

The mechanical realization of the focus control may be seen in Fig. 10. The photographic plate is pneumatically clamped in a fixture. The emulsion side, which is up, rests against 14 co-planar pins which locate the perimeter of the top surface parallel to the lens' image plane. The plate clamping fixture is in turn attached to the movable step-and-repeat camera stage by four pairs of stressed parallel springs. This suspension system is stiff to all rotations and translations excepting motion in the vertical direction. In order to minimize the vertical stiffness (k_f) the springs must be horizontal; however, in this position their load-bearing capacity is zero. Therefore, to support the weight of the plate and its clamping fixture an annular groove covered with a flexible diaphragm is provided under the plate-supporting fixture. This is inflated until the diaphragm, acting against the main stage, makes the springs horizontal and brings the edges of the clamped plate to the elevation of the image plane. By inflating the chamber each time a new plate is loaded, the correct starting elevation is achieved regardless of differences in plate weight or barometric pressure. The pneumatic counterbalance would add no stiffness to the plate support system if it were operated at constant pressure. However, it is more

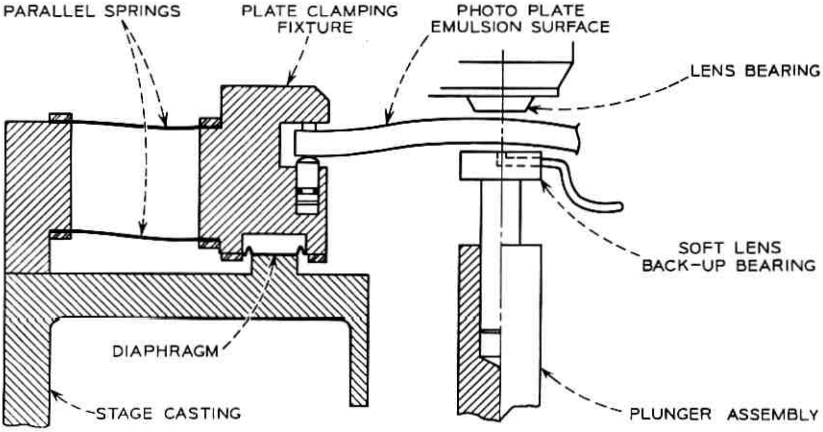


Fig. 10—Focus control system.

practical to inflate the system to the proper height and close the valve than to establish and maintain the correct pressure; therefore, it is operated at a constant volume. The combined stiffness of the parallel springs and pneumatic counter balance system is between 5.0×10^4 and 8.5×10^4 N/m.

The lens bearing is 2.54 cm in diameter and has a central 0.94-cm hole through which the image is projected. The placement of this bearing between the lens and the surface is only possible because of the lens' large working distance. In operation the lens bearing-to-plate spacing is 12.5 μm, its stiffness is 4.4×16^6 N/m, and the normal force which it exerts against the plate is 44 N. If this load is transmitted to the fixture and diaphragm through the photographic plate, it will result in a large deflection of the plate and increase the difficulty of maintaining good focus. Therefore, an equal but opposing force is applied to the lower surface of the plate by a soft air bearing placed directly below the lens bearing. This bearing is mounted on a low friction pneumatic plunger which is raised into place after the plate is loaded. This bearing provides nearly constant upward force on the plate regardless of plate bow or taper and does not contribute to the system stiffness. A further discussion of the air-bearing design is given in Section 3.3.

The focus control system was evaluated by using a laser interferometer in place of the projection lens to measure the relative motion between a mirrored plate clamped in the fixture and the lens housing. Using typical plates the focus control is able to maintain the elevation of the plate's surface on the optical center line within ±25 μm of the image plane.

3.2 *Stage Design*

All major parts of the camera are supported on a one-meter-square block of granite. The top surface of the granite is flat to 2.5 μm and three 15-cm-square areas under the stage support bearings are parallel to within 5 μrad. The block is supported on three special Barry Control Corporation* Serva-Level® mounts which provide vibration isolation in the vertical and horizontal directions. The vertical and horizontal natural frequencies are 0.82 Hz and 0.87 Hz respectively. A 1200-kg lead ballast is attached to the underside of the granite to lower the camera's center of gravity and assure the stability of the Serva-Level® system. The extra ballast also increases the working pressure in the air mounts which minimizes the effects of sudden changes in the ambient pressure due to opening and closing doors in the air conditioned facility.

The stage is supported on three two-inch-diameter air bearings. Each bearing is attached to the stage through a spherical bearing assembly and a spacer made up of Hoke gauge blocks (See Fig. 11). The spherical bearing assembly allows some initial angular adjustment of the bearing during assembly and alignment; however, once the weight of the stage is resting on the bearing the static friction in the spherical bearing prevents further movement. The gauge block stack between the bearing assembly and the stage allows the elevation and tilt of the stage to be adjusted more accurately than is possible by machining. Since the stage rests directly on the granite surface rather than on an intermediate stage as is customary in machine-tool construction, its attitude depends only on the granite surface and is constant within 2.5 μrad.

The stage is guided by an intermediate structure which is called the cross. The cross is constrained to move only in the X (right-left) direction by two pairs of 2.5 cm-diameter air bearings and two colinear quartz guide blocks which are secured to the granite base. The sides of the guide blocks are straight and parallel to 0.25 μm and the blocks are optically aligned on the granite surface to achieve a cross yaw of less than 1.25 μrad over 10 cm. The cross is supported on four 2.5-cm-diameter air bearings, two resting directly on the granite surface and two on the top surface of the quartz guides. All of the cross support and

* Located in Watertown, Massachusetts.

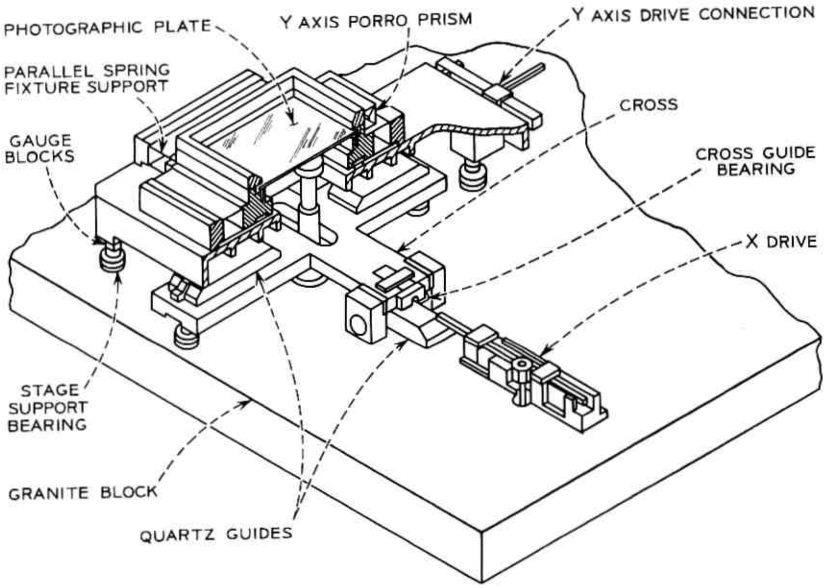


Fig. 11—Cross section of stage showing cross and drives.

guide bearings are also assembled using sperical bearings and gauge blocks. The stage is also guided by four air bearings and a pair of quartz guide blocks mounted on the top surface of the cross parallel to the camera's *Y* axis. The straightness of travel along this axis is similar to the *X* axis giving a total stage yaw of less than 2.5 μrad over the entire 10- × 10-cm travel of the stage. Therefore, this is the maximum rotational error between any two images on the final mask due to the guiding system.

The parallel springs of the plate holding fixture are attached at the outer edges of the stage. This fixture and its diaphragm support rest on the top surface of the stages. The elevator bearing for the focus control system is attached to the granite surface on the optical axis and passes up through a slot in the cross.

The *X*-axis drive is attached to the cross through a slender flexible column which allows both vertical and horizontal misalignment between the cross and drive. Under worst-case conditions, the maximum cross rotation due to the force required to deflect this member is 0.1 μrad. The *Y*-axis drive is coupled directly to the stage through two air bearings and a guide bar attached to the stage.

All of the camera's major components are cast from Meehanite GC-40 because of its good stability and good damping qualities. The castings were X-rayed to assure their soundness and were heat treated prior to initial machining, prior to the final grinding operation, and again after all machining operations. This heat treatment provides phase stability (ferrite and graphite) and assures low creep rates under the low stress condtions of its use. All nonmating surfaces were then painted with an air dry vinyl paint.

3.3 *Air Bearings*

The stringent requirements of position and focus control necessitated that the camera be supported and guided by stiff low-friction bearings. Investigation of various bearing types revealed that gas hydrostatic thrust bearings possessed the necessary characteristics to meet these requirements.[2] They operate in a nearly frictionless manner (frictional resistance is about 1/4000 of that of a light oil bearing). They are very simple in construction permitting relative ease in meeting mechanical tolerance requirements. Their load-stiffness characteristics are such that camera requirements can be met with low gas-supply pressures ($<3.4 \times 10^5$ N/m^2) and total gas consumption ($<2.3 \times 10^{-3}$ standard m^3/s).

The performance and space requirements for the various air bearings employed on the camera are indicated in Table II. The plate bearing is a central-jet type and the remaining bearings are ring-jet varieties (See Fig. 12).

The design and development of the gas bearings involved both analytical and experimental programs. The analytical program consisted of computer predictions of steady[2] and transient[3] behavior of the aforementioned bearing types. The experimental program was used to verify analytical predictions as well as prove-in the focus-control system. Typical results of these programs are shown in Table III and Fig. 13.

TABLE II—AIR BEARING REQUIREMENTS

Bearing	Load	Stiffness	Space
Main Stage	160–220 N	1.2–1.8 $\times 10^7$ N/m	5.08 cm OD
Cross Support	36–54 N	0.35–0.53 $\times 10^7$ N/m	2.54 cm OD
Cross Guides	22–45 N	0.18–0.35 $\times 10^7$ N/m	2.54 cm OD
Lens	36–54 N	0.35–0.53 $\times 10^7$ N/m	2.54 cm OD
Plate	Force equal to Lens *B*	Zero or Finite but small	3.81 cm OD

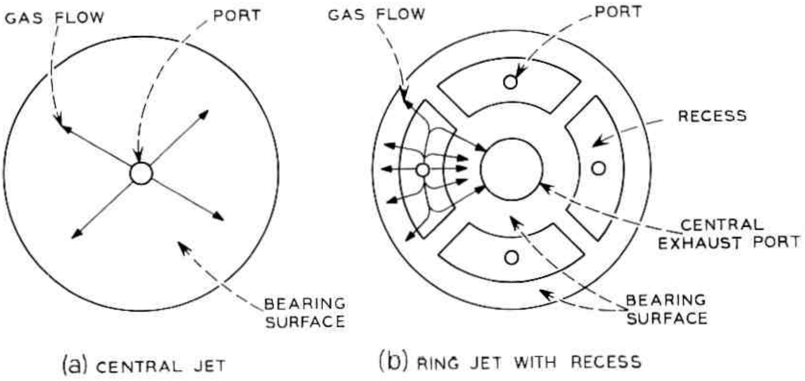


Fig. 12—Air bearing configurations.

3.4 *Interferometer Design*

The correct placement of every image on the photographic mask is a primary requirement of a step-and-repeat camera. Therefore, special attention was given to the design of the camera's interferometers. Both the X and Y interferometers are identical and they share the output of a frequency stabilized HeNe laser. Their outputs indicate both the direction of stage motion and the distance traveled. Each output pulse represents 0.04-μm stage travel or 1/16 wave length (λ). The interferometers are arranged in a double pass configuration so that each fringe represents a $\lambda/4$ displacement of the stage (See Fig. 14). Two photocells monitor the output light beams whose phases differ by 90°. This phase difference is achieved by the use of circularly polarized light and polarizers before each photocell, and it is used to indicate the stage's direction of travel and to further divide the output fringe

Table III—Actual and Theoretical Air Bearing Characteristics

Bearing	Load Range N	Gas Supply Pressure N/m²	Average Stiffness N/m	
			Theoretical	Experimental
Lens & Cross Support	38–53	1.4×10^5 2.8×10^5	4.3×10^6 5.5×10^6	2.6×10^6 4.4×10^6
Main Stage Support	154–220	2.1×10^5 2.8×10^5	14×10^6 18×10^6	11×10^6 19×10^6
Guide	25–43	2.1×10^5 2.8×10^5	5.6×10^6 5.8×10^6	2.9×10^6 4.2×10^6

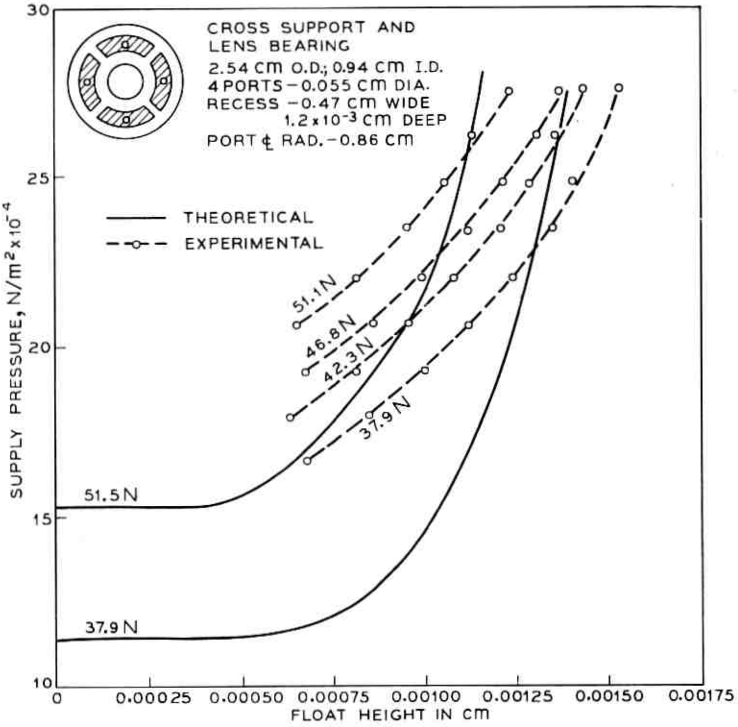


Fig. 13—Theoretical and experimental air-bearing characteristics.

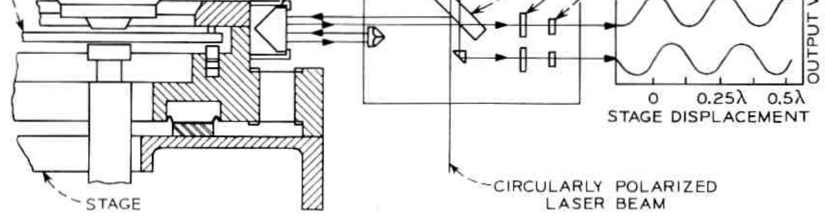


Fig. 14—Arrangement of stage position interferometer.

interval by four through the use of the two signals' zero crossings. Judicious placement of the interferometers is required to assure the maximum accuracy of the camera. For instance the X and Y measuring legs intersect the camera's optical axis thus making the measured location of the image independent of small stage rotations about the vertical axis. For machines in which the measuring legs do not intersect the optical axis, legitimate translations and errors due to rotation are indistinguishable. (This is always the case with some heads of a multiple-head camera.) Similarly, the measuring beams are at the same elevation as the photographic plate making the image position independent of small amounts of stage pitch and roll.

The measuring legs are terminated at the stage by 12-cm-long porro prisms. The photographic plate is rigidly attached to these prisms through the plate-clamping fixture, thus assuring that motions of the porro prisms are identical to those of the photo plate. Since the stage-drive system continually moves the stage to maintain a particular fringe count, its location along the two axes is dependent only on the straightness of these prisms and not on the straightness of the stage guides. Similarly the orthogonality of the two axes is only dependent on the relative mounting of these prisms. On the camera these prisms are set at right angles to within 1.25 μrad.

The reference leg retro-reflector is attached to the optical head as close to the image plane as is practical. In this way the interferometer output represents the relative locations of the stage and the optical head rather than the location of the stage with reference to a leg fixed in the interferometer body as is the usual practice. This allows compensation for deflections of the optical head which would otherwise go unnoticed.

The optical parts were fabricated from fused silica because of its excellent stability. They use total internal reflection and are not antireflection coated in order to eliminate the mechanical distortions which frequently accompany the deposition of these coatings. They have a wave-front accuracy of one-tenth wave over their 12-cm length and they are mounted in a nearly stress-free state so their accuracy will not be reduced by mechanical strain.

The laser wave length changes with variations in atmospheric conditions. Since the room temperature is maintained constant to ±0.13° C, corrections for temperature are not necessary. However, barometric corrections are made prior to each exposure because pressure variation of 3.44×10^3 N/m^2 results in 1.0-μm errors. When wave length corrections are calculated, the actual difference in the measuring and reference leg lengths must be known. Therefore, the granite table has been fitted with sensors which allow the establishment of one absolute stage position from which all corrections for varying ambient conditions are made.

3.5 *Stage Drive*

The stage and cross are driven by identical low-backlash drive systems. Motion in the X and Y directions is imparted to the cross and stage through 1.3-cm-square bars which are guided on all sides by air bearings, and driven longitudinally by a capstan which is an extension of the motor shaft (See Fig. 15). Sufficient driving force is achieved by pinching the drive bar between the driven capstan and a spring-loaded idler. The capstan and idler shafts are mounted so that no net transverse force is transmitted to the drive bar and they are prevented from moving in the direction of the drive bar by flexures attached to the body of the drive unit. The motors are mounted below the granite table and their shafts pass vertically through holes in the granite in an

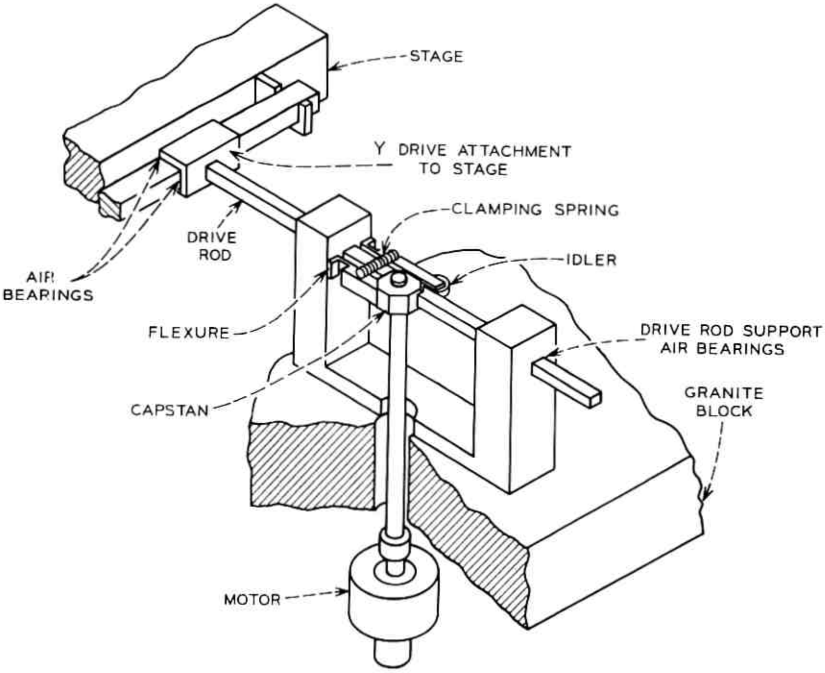


Fig. 15—Y axis drive.

attempt to minimize any heat transfer from the motors to the camera structure. The lack of gearing and the use of flexures minimizes the drive train backlash and enhances the system stiffness.

The drives can move the stage at any uniform velocity up to 0.5 cm/s and they are capable of stopping the stage from this velocity in less than 2^{11} interferometer counts (80 μm). They can also maintain the stage to within plus or minus two interferometer counts (0.08 μm) of its desired position. This is accomplished through the servo system which is shown in block diagram form in Fig. 16.

IV. CONTROL SYSTEM

4.1 *Servo System*

Since the X- and Y-axis servo systems are identical only one will be discussed. They operate in three modes: constant speed slewing, decelerating from a contant speed and holding at a fixed location. The stage velocity is determined by the value in the twelve-bit speed register which is converted to an analog signal in the digital-to-analog (D/A) converter. Thus when the number in the speed register remains constant, the D/A output is fixed and the stage runs at a constant speed. The stage velocity is stabilized through a digital tachometer feedback loop which uses the rate of interferometer pulses to determine the stage velocity. The drive system stability was further enhanced by the addition of a small viscous damper on the motor shaft.

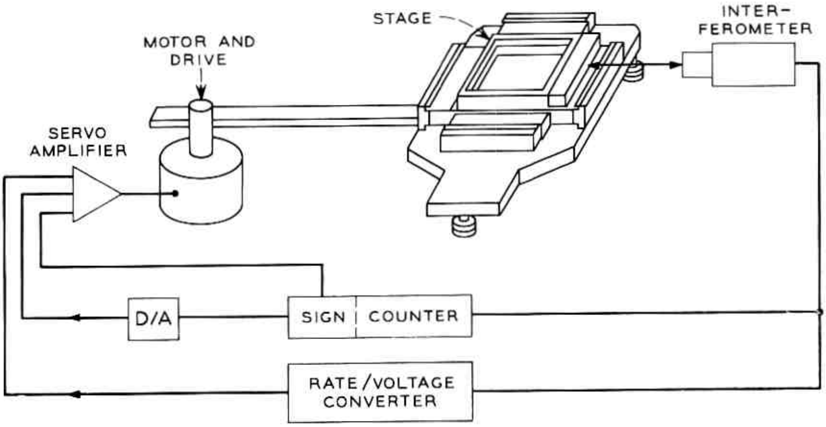


Fig. 16—Block diagram of stage-positioning servo system.

When the stage must stop at a given location, the value in the speed register is made equal to the distance from the stopping location in interferometer counts (0.04 μm per count) and the D/A output voltage decreases toward zero as the stopping location is approached. When the stage is at the desired location, it is imperative that the value in the counters become zero and remain or recross zero in a small limit cycle so that each image on the step-and-repeat mask will be in the correct location. Ideally when the position error is zero, the stage should stop and if it moves slightly, say one count, the motor should again drive it to zero. Unfortunately small amounts of drift in the D/A converter or the servo amplifier will cause the stage to stop and remain at some point with other than zero in the counter and speed register. Even without this electronic drift, the system's static friction, although small, will require unreasonably large gains to assure that errors of one count in the speed register (0.04-μm stage position errror) will be corrected.

Both problems have been overcome by demanding that the stage execute a small limit cycle which includes the zero location. This is accomplished by adding to the D/A output a two-valued function which is positive when the speed register is positive and negative otherwise (See Fig. 17). The magnitude of this step voltage is just sufficient to cause the motor to drive the stage toward zero in spite of electronic drift and mechanical stiction. This assures a continual recrossing of the zero-minus-one transition point. In addition it reduces the positioning error by removing the dead band of 0.04 μm which occurs if the stage location corrections are made only when the value in the speed register becomes plus or minus one.

4.2 *Digital Computer and Interface*

The control of the camera is coordinated through a Digital Equipment Corporation PDP-8/L computer. This has a 12-bit word length, a cycle time of 1.6 μs and 4096 words of core memory. Its function is to provide communication between the operator, the camera and the MSIS and to make the necessary conversions and calculations for the camera's operation. The connection to the MSIS PDP-9 computer is through an interface and a high-speed data link. The interface provides buffering to allow data transfer between the computers to occur asynchronously. The information transferred across this link includes step-and-repeat array data from the information system, operating status, and requests for information from the camera. The computer

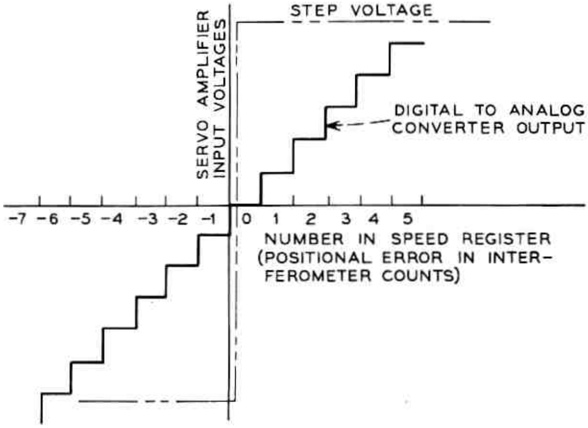


Fig. 17—Servo amplifier input voltages versus stage-positional error.

and interface racks are located outside the step-and-repeat camera room to minimize any heat transfer to the camera structure.

The information transferred between the camera and the PDP-8/L is stored, acted on, and relayed by three other interface sections, each of which consists of about 180 DTL integrated circuits on a board with wire wrapped interconnections (See Fig. 18). The three sections are the identical X- and Y-axis control interfaces and the accessory interface. The latter interface monitors camera functions and provides special tests, which are not related to the stage positioning system, such as reticle identification, interlock testing, and atmospheric pressure monitoring.

Operator/computer communication is also affected through the accessory interface which controls both an illuminated message board and an auditory alarm as well as storing input from the operator keyboard. An additional interface allows the computer to output supplemental instructions of a variable nature on a CRT display.

The X- and Y-axis interfaces interconnect their respective interferometers and drives. These interfaces, once loaded from the PDP-8/L, are capable of positioning the camera stage at any location within the limits of its travel and exposing an image on the plate at that point without further intervention from the computer, thus leaving the computer free for other work.

Each axis interface has two storage registers which may be loaded from the computer. A 24-bit register contains the address of the stage's

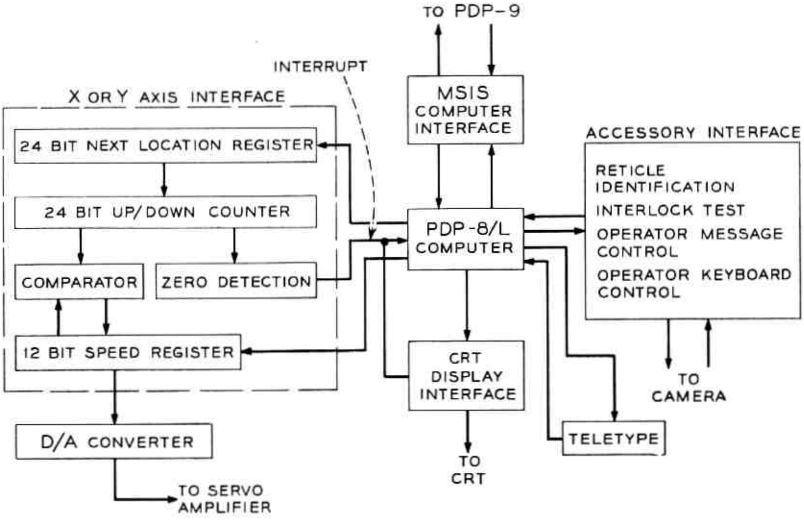


Fig. 18—Block diagram of computer/camera interface.

"next location" relative to the current stage destination, and a 12-bit register (the previously mentioned speed register) which contains the binary equivalent of the speed at which the stage is to move. The heart of the interface is a 24-bit binary up/down counter for accumulating interferometer output pulses. It consists of four six-bit parallel-carry counters connected in series to allow counting in either direction at rates of 4 MHz. This counter is initially loaded with a number from the "next location" register and the stage is moved until the counter becomes zero. At this time the optical head flash lamp may be triggered and the counter may be automatically reloaded from the "next location" register.

The speed register and the counter are connected to a comparator which, upon a command to stop at the next location, will compare their contents. When they become identical it will continually transfer the value of the counter into the speed register keeping these two equal. This makes the value in the speed register proportional to the distance from the stopping location thus providing automatic stage deceleration. The interface also initiates a computer program-interrupt when its counter has become zero. This immediately alerts the computer to the fact that its previous requests have been carried out and that the status of the interface may be updated.

The size and complexity of the axis interfaces justify the inclusion of several special functions. These allow the computer, using a special program, to perform maintenance tests on these interfaces to identify malfunctions and enumerate possible corrective actions.

V. COMPUTER PROGRAM

The program stored in the step-and-repeat camera control computer, the PDP 8/L, couples the camera's systems into an automatic production tool. Logic sequences of this program could have been provided by hardware logic components; however, implementation in this manner would not allow the flexibility of a computer program and would require many more electronic components. The balance between program logic and hardware logic has been established by providing hardware functions that greatly reduce either program complexity or computer time and by utilizing program logic where decision making or complex hardware logic would be required. The division of logic functions between hardware and program has been greatly influenced by experience with previous Bell Telephone Laboratories computer controlled photolithographic equipment.

Features utilized in the program to accomplish the control objectives are:

(*i*) Live interaction with the operator at all times using non-interrupt programming;

(*ii*) Input data conditionally accepted at any of three input terminals with two of the terminals serving a dual use for operator control;

(*iii*) Message board and CRT display used to communicate with the operator;

(*iv*) Multiple use of the axis-control routine and all other routines when posible;

(*v*) Overwriting loader areas to provide maximum utilization of computer core; and

(*vi*) Self starting of the program on loading with a checking routine to verify correct loading.

5.1 *Functions Performed by the Program*

The control program's objective is to provide automatic control of pattern placement on a photographic plate. In this operation the only operator tasks are: initializing the computer program, installation and removal of the photographic plate, and installation of reticles as required. However, in cases of mask shop operational decisions and equipment malfunctions, operator intervention is also needed. These requirements have resulted in a set of necessary control program functions:

(*i*) Initialize and terminate a table maneuver.
(*ii*) Transfer data as needed to the interface.
(*iii*) Reproduce a series of text characters at specified locations on the photographic plate using the 5 × 7 light array.
(*iv*) Automatically zero the table (initialize the interferometer counters).
(*v*) Read a decimal input format and convert it to binary values suitable for interface use.
(*vi*) Summon the operator when human intervention is needed.
(*vii*) Communicate with the operator through a message board and a CRT display.
(*viii*) Check installed reticle for correct identification number and control its alignment procedure.
(*ix*) Receive input from either the operator, a paper tape in the teletype or the MSIS computer as specified by the operator.
(*x*) Provide a maintenance founction which transfers control of the table to the maintenance keyboard.

These ten functions are either self descriptive or have been described in prior sections.

The format used to transfer data from outside sources to the step-and-repeat control computer include nine code characters:

Y—Indicates *Y*-axis coordinate value;
X—Indicates *X*-axis coordinate value;
D—Spacing between images;
N—Number of images on (*D*) spacing;
R—Repeat the last line of data;
A—Repeat the data preceding the last line;
E—End the run;
*—Following characters represent a reticle number; and
″—Enclosed characters are to be written as text on the mask.

The first four code characters require that a minimum of one digit follow them to specify their magnitude with a maximum of seven digits. For the convenience of paper-tape input, a decimal point may be used with the digits to the left of the decimal point indicating the distance in millimeters. If one or two digits are specified and no decimal point is used, the decimal point is assumed to be after the last digit. For more than two digits, the decimal point is assumed to be after the third digit.

Whether rows or columns will be run is determined by the sequence of the first two code characters following a reticle number or a text request, or upon starting a new mask array. If the sequence is *Y*---*X*--- the program assumes the data following is to be placed in rows with all subsequent *Y*s being the *Y* coordinates of their respective rows. Alternately, the sequence *X*---*Y*--- signifies column data with the columns located at the specified *X* coordinates. Image locations along a row (column) are specified by subsequent *X*---s or by a *D* followed by an *N*, each followed by its appropriate digit value. Numbers following *N*s are evaluated as integers and not by the aforementioned decimal format. The only restriction on this format is that all values along a row or column must be in increasing magnitude.

5.2 *Program Philosophy*

A block flow diagram of the control program is shown in Fig. 19. In this figure bold lines indicate the main path of control through the program, light lines indicate paths that are taken when the block's function has been requested and broken lines are paths taken in going to the control routines when a waiting point has been reached. Most of these waiting points are labeled as "gates". These gates are points in the program that a control function dare not pass until some task has been completed. For example, the run and the main gates prevent simultaneous operation in either the run area, which is controlling the table motion, or the loading area, which is bringing in data and deciphering coded characters. The keyboard monitor routine maintains contact with the operator, allowing intervention in the camera's operation.

The program for running the camera has been developed using a foreground-background philosophy. Since the primary purpose of the program is to control the camera, the foreground program is the set of routines which directly control the table motions. The main routine of the foreground program is a general running routine which will control an axis through its most general maneuver, that of running along a row or column and exposing images at specified locations. To accomplish this, the routine requires the table of data for image placement to be available in the computer core. Because of the urgency in transferring information to the interface hardware when a task is completed, the foreground program is interrupt addressed.

Maneuvers of the stage other than the most general are accomplished by defeating inappropriate functions in the general running routine to

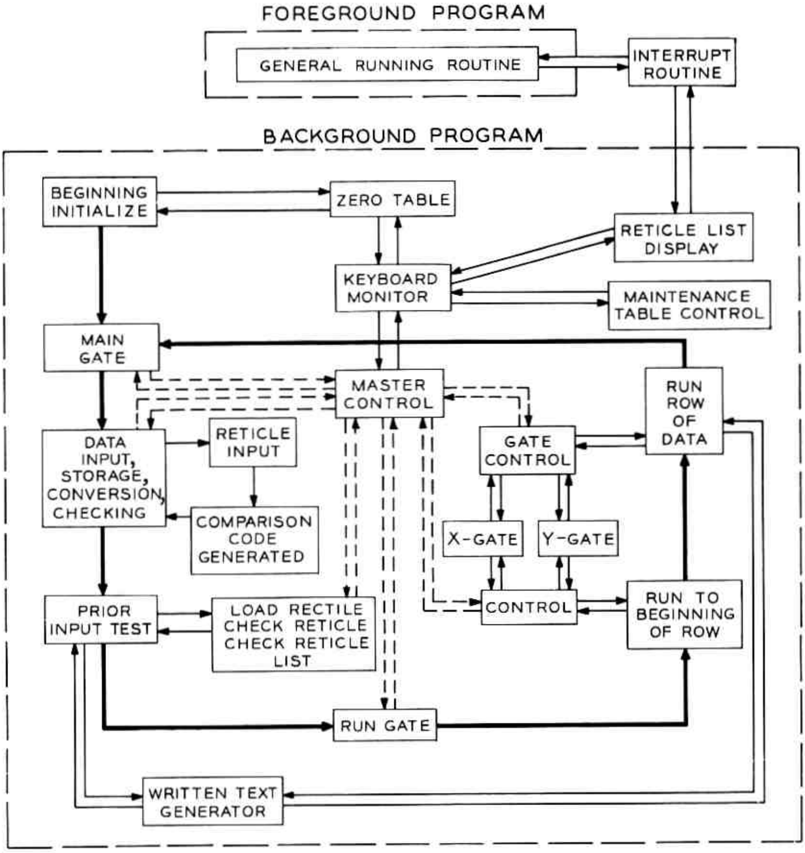


Fig. 19—Program block flow diagram.

make it perform as required. These changes are made prior to the desired maneuver and the general running status is restored after the maneuver is completed.

The background program provides input points from the communication equipment and communicates with the operator. This program operates on a noninterrupt philosophy. Noninterrupt programming is used because D.E.C. teletypewriter philosophy will cause the computer to overflow its limited storage capacity when operating with paper-tape input. The teletypewriter's hardware interrupt has been disabled to allow this noninterrupt philosophy.

To eliminate waiting time for slow input terminals like the teletypewriter, the machine waiting points are programmed to return control of the computer to a master-control routine. The master-control routine then cycles through the other waiting points searching for work to be done. This technique allows the general running routine to be initialized and, while waiting for the addressed axis to complete its maneuver, the second axis can be initialized and data received from the input terminals.

Control information from the operator is entered through the operator keyboard or the teletypewriter. However, if the teletypewriter is being used to input data, operator control through the teletypewriter is not allowed.

Output to the CRT display is controlled through the computer interrupt facility after being initialized by the background program. This implementation was made because interrupt techniques minimize the asynchronous nature of this transmission.

5.3 *Implementation of the Program*

To implement the required functions with the control program, all locations in the 4096-word core of the PDP-8/L have been assigned an operational use. In doing this all program loaders are written over the background routine. The program logic occupies all locations from 0 through 5777_8. Locations 6000_8 through 6777_8 are used to store incoming data, the first half from 6000_8 through 6377_8 being table 1 and the second half, 6400_8 through 6777_8 being table 2. Each table contains the information for one row or column of images. The technique of using two data storage tables allows reading data into one table while the second table is being used to run a row or column of images. It also allows a mask using only two types of spacings to be produced by only transferring the Y- (or X-) coordinate values for all rows (or columns) following the initial table information transfer.

Locations 7000_8 through 7777_8 are used to store the list of reticles used to make a mask. This area is also divided into two equal table areas. One area will store the list of reticles for the current mask. When the data is transferred from the MSIS computer, the table is loaded immediately following the initiation of a job. However, when the job is being entered through the teletypewriter, this table area will store each reticle number at the time the operator is requested to load the reticle into the camera.

The second table area contains the list of reticles used to generate the preceding mask. Either of these two lists may be displayed on the CRT display at the operator's request.

Initial loading of the program has been reduced to a "push button" operation through a hardware deposited elementary loader. When this loading is complete and the computer is started, any tape which is in the high-speed, perforated tape reader will be read. In the case of the program tape for the step-and-repeat camera computer, the D.E.C. binary loader has been incorporated at the beginning with a sufficient number of statements added to cause the computer to switch from the elementary loader to the binary loader. The binary loader then loads the program's binary tape without the computer coming to a stop. At the end of the program, binary tape statements have been included which overwrite the binary loader and switch control of the computer to a test area to check for correct loading of the program. If the program has been properly loaded, the computer starts the camera control program. If the computer must be stopped, a restart location has been provided for the operator.

VI. CONCLUSION

We have discussed the design of a step-and-repeat camera capable of meeting the most exacting integrated circuit mask requirements. The requirements for precision image placement, 1-μm line-width resolution, and minimum operator intervention have influenced every aspect of the camera's design. A system capable of maintaining the photographic surface in focus to within ±0.25 μm was developed in order to assure maximum image resolution and correct image magnification. The stage guide, drive and measuring systems utilize air bearings and multiple-pass interferometers to achieve precise image placement. The computer program provides, in addition to camera control, operator checking and communication to simplify the operator's job and to minimize errors.

VII. ACKNOWLEDGMENTS

The authors wish to pay tribute to P. Fritz, E. S. Jarzena, R. C. Joseph, D. C. Noack, and E. A. Toepel whose work in assembly, alignment, and testing have made the realization of this camera possible.

REFERENCES

1. Kerwin, R. E., "Thin Photosensitive Materials," B.S.T.J., this issue, pp. 2179–2192.
2. Gross, W. A., *Gas Film Lubrication,* New York: John Wiley & Sons, Inc., 1962, pp. 255–306.
3. Roudebush, W., "An Analysis of the Effect of Several Parameters on the Stability of an Air-Lubricated Hydrostatic Thrust Bearing," National Advisory Committee for Aeronautics Technical Note (NACATN) 4095, 1957.

Device Photolithography:

Thin Photosensitive Materials

By R. E. KERWIN

(Manuscript received June 3, 1970)

New camera systems, utilizing lenses of high numerical aperture and concomitant shallow depth of focus, require thin recording media. A number of materials potentially fulfilling this requirement are discussed. These include photoresist-coated metal or semitransparent masks, some unconventional photographic processes, and dyed photographic emulsions. The use of dyed photographic emulsions is recommended on the basis of sensitivity and improved resolution and modulation of the recorded image.

I. INTRODUCTION

In this paper we discuss some recent developments in thin recording media in light of their suitability for exposure in the new step-and-repeat camera system. The requirements of integrated-circuit pattern generation have led to the development of a wide-field objective lens for this camera having high numerical aperture (N. A.) corrected for diffraction limited performance using monochromatic light. Specifically the lens has a 7.1-mm field diameter, $f/1.5$ at 10:1 conjugate ratio, and is corrected for $\lambda = 436$ nm.

This lens has a depth of focus shallower than the thickness of the photosensitive emulsion on the thinnest high-resolution photographic plates. Kodak High Resolution Plates (KHRP) consist of a 6-μm-thick Lippman-type emulsion of small (< 0.1 μm) silver halide grains in gelatin on a flat glass substrate. For any projection lens of f-number less than $f/1.7$ the depth of focus is less than 6 μm. This is illustrated in Fig. 1 which is an idealized ray diagram, drawn to scale, showing in cross section a 6-μm-thick photographic emulsion of refractive index 1.56 into which a linear array of diffraction-limited spots on 1-μm centers have been projected through an $f/1.5$ lens using 436-nm light. Even a perfect lens images a point source as a diffraction patch, the Airy Disc, having a radius r to the first dark ring given by:

$$r = \frac{0.61\lambda}{\text{N.A.}} = \frac{0.61\lambda}{n \sin \theta}. \tag{1}$$

In accordance with this equation, the cone angles (θ) of illumination in Fig. 1 are functions of the lens aperture, the refractive index (n) of the medium, and the wavelength. The width of each rectangular shaded area is equal to the radius of the Airy Disc, and the depth of focus is approximated as the region of overlap of this with the cone of illumination. Light scattering due to the difference in refractive indices of the silver halide and gelatin is not indicated in the figure although it is recognized as a major source of image spread.

From Fig. 1 it is evident that some out-of-focus illumination is capable of acting on the photosensitive emulsion. A fraction of this out-of-focus illumination will be recorded as a function of the sensi-

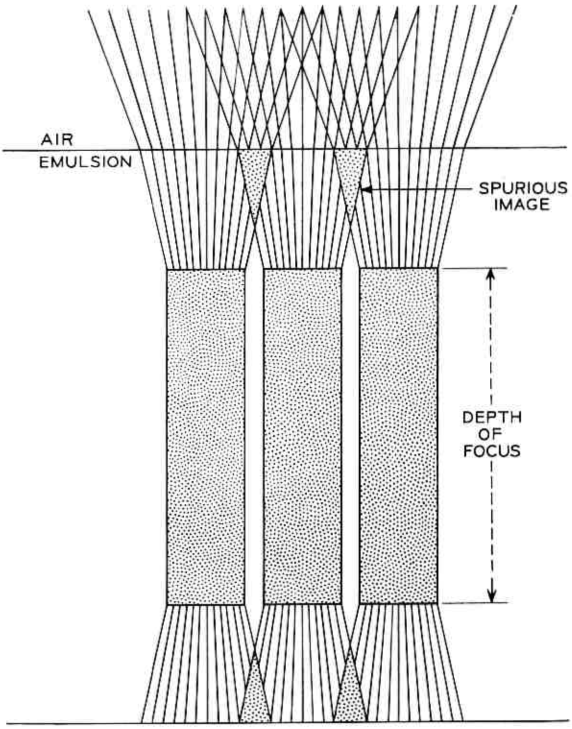


Fig. 1—Ray diagram indicating the depth of focus of a linear array of diffraction-limited spots 1 μm apart projected into a 6-μm emulsion of refractive index 1.56 through an ideal $f/1.5$ lens using 436-nm light.

tivity of the emulsion and the efficiency of the developing process, thus degrading the overall image quality. In the case of large images or uniform-sized small images this could be circumvented by darkroom "clipping" so that exposures below a defined threshold would not be developed. This is not possible in microelectronic photomask production since the intensities of fine lines near the diffraction limit vary as a function of line width.[1] Figure 2 shows the intensity profiles obtained with our 10X lens for isolated lines of widths 1, 2, 4, and 10 μm, all normalized to the same width W. These have been calculated by convoluting the modulation transfer function of the lens at the edge of the field with the light-distribution function of the object, indicated by the dashed rectangle.[2] Since the intensity profiles are symmetrical, only one half-cycle is shown. A focal-plane image of a 1-μm line would have at its center only 69 percent of the light intensity at the center of a neighboring 10-μm line. Thus, "clipping" would result in a loss of fine-line image detail.

Another problem of a depth of focus shallower than the recording-medium thickness is the formation of spurious images. As indicated in Fig. 1, the regions of overlap of adjacent cones of illumination may provide sufficient intensity for exposure giving rise to spurious images between the real images.[3]

Thus, it is apparent that even high-resolution photographic plates must be regarded as three-dimensional systems and for optimum use of the new lenses thinner recording media must be obtained. A number of approaches to the solution of this problem have been tried and are discussed below.

II. THIN PHOTORESIST FILMS

It is immediately attractive to those familiar with microelectronic photolithography to use photoresists, which have demonstrable high-resolution capabilities in thin films, as the required thin recording medium. Figure 3 illustrates this approach using the step-and-repeat camera to project images into a photoresist coating on metal or semi-transparent films on glass substrates. In this case, the thin photoresist film (0.3 μm) records the high-resolution image without depth-of-focus limitations and, after development, controls the transfer of this image by etching into the 0.1-μm film of chromium or iron oxide to provide the optical density and hardness required of a photomask.

However, there are serious problems related to this approach and each of these must be solved before this approach can become prac-

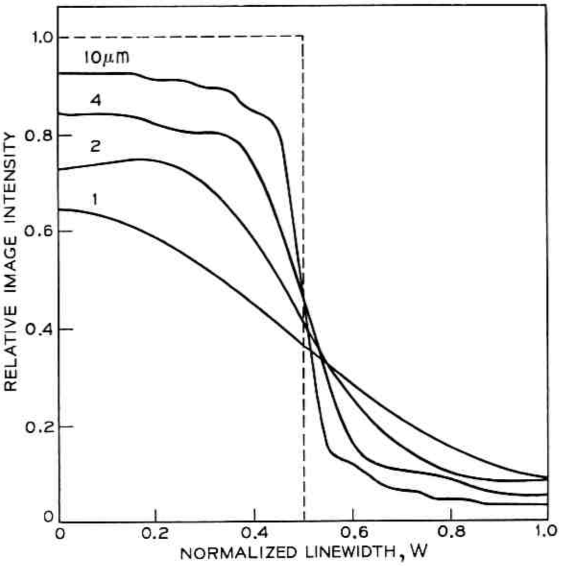


Fig. 2—Calculated intensity distributions as a function of linewidth for isolated slits as imaged by the 10X camera lens.

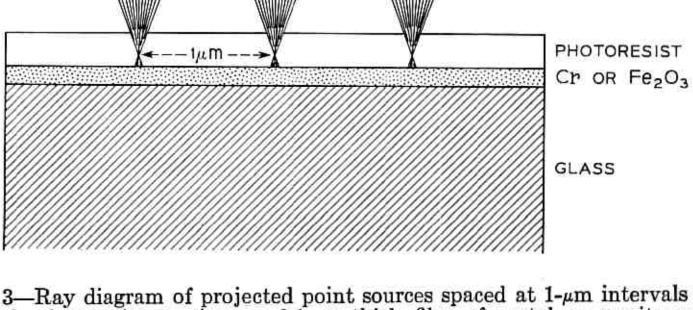


Fig. 3—Ray diagram of projected point sources spaced at 1-μm intervals in 0.3-μm-thick photoresist coating on 0.1-μm-thick film of metal or semitransparent mask material (*f*/1.5 lens, 436-nm light).

ticable. These difficulties are the standing-wave effect due to the reflective substrate, the apparent high-modulation requirements of photoresist, and the low photographic speed of photoresist.

It is known that exposure of photoresist films on reflective substrates leads to the formation of standing waves due to the interference of the incident and reflected light waves, and these in turn produce nonuniformly exposed strata in the photoresist.[4,5] This effect is already seriously detrimental in contact printing with polychromatic light, $340 \leqq \lambda \leqq 440$ nm, and will be enhanced in our projection printing with monochromatic light, $\lambda = 436 \pm 8$ nm. It has been demonstrated that the first node or minimum in intensity lies 0.07 μm above a chromium mask surface so that normal exposure of a negative photoresist in these circumstances would result in a 0.07-μm developed film which is too thin to withstand etching solutions.[6] Recently, semitransparent masks consisting of 0.1- to 0.2-μm films of Fe_2O_3 formed on glass by the vapor-phase decomposition of iron pentacarbonyl have been developed to facilitate alignment procedures during contact printing onto photoresist-coated silicon wafers.[7] The reflectivity of this material is a function of its film thickness but is approximately only 50 percent that of chromium at 436 nm. The substitution of this mask for chromium masks will alleviate somewhat the standing-wave problem. Further improvement will be obtained through the use of darker resists, in which the reflected light will be a small fraction of the incident light.

Recent measurements of the characteristic curves of photoresists (developed film thickness versus exposure, the slope of which is referred to as the gamma of the system) indicate that sharp image-formation requires relatively high intensity modulation in the projected image.[8] Specifically an 80 percent modulation is required for a normally developed 0.4-μm film of Kodak Thin Film Resist (KTFR). This is a stringent demand on the optics of the system since it implies (Fig. 4) a usable resolution in the resist of only 0.18 the limiting frequency of an aberration-free system.[9] Although the use of gentler development conditions and dilute developers leads to some lowering of the constrast requirement,[5] the resolution of this difficulty awaits the development of higher gamma photoresists.

The most obvious limitation of present photoresist systems with respect to their use in a flash source step-and-repeat camera is their low sensitivity. In Fig. 5 the measured values of the spectral sensitivities of four types of photoresist and KHRP are presented. The KHRP sensitivity refers to the reciprocal of the energy necessary in

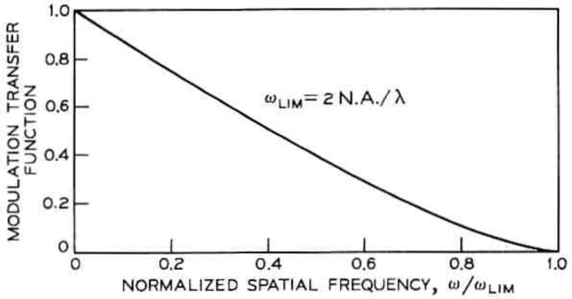


Fig. 4—The modulation transfer function of an aberration-free system as a function of the normalized spatial frequency. ω is the line frequency in cycles per millimeter, and ω_{lim} is the high frequency limit imposed by diffraction effects, a function only of the numerical aperture (N.A.) of the lens and the wavelength of light (λ).

exposure to reach an optical density of 1.5 on development in Kodak type D-19 developer. The measurements on photoresist were carried out using 0.2-μm films of negative resist and 0.5-μm films of positive resist. The box straddling the 436-nm line represents the measured energy output of the camera using type FX-76 xenon flash lamps, manufactured by the EG & G Company of Boston, Massachusetts, and a 15-nm-wide bandpass filter. Obviously, all the resists fall short of the camera requirements and again the need for the development of a new class of resists is implied.

III. UNCONVENTIONAL PHOTOGRAPHIC PROCESSES

There are a number of recently developed photographic processes which would appear, at first glance, to be candidates for the required thin recording medium. It is not within the scope of this paper to present each of these in detail or to analyze their current uses; however, we merely seek to correlate their sensitivity and resolution limits with the requirements of our system. This information is presented in Fig. 6 in the form of a correlation diagram of the measured or reported maximum photosensitivities and resolution limits. The box outline in the center of the diagram serves as the goal with an ordinate range corresponding to the output per flash of the step-and-repeat camera, 50 to 250 μJ/cm^2, and an abscissa range of 250 to 1000 cycles/mm or equivalently 2 to 0.5 μm lines.

Point *A*, for Kodak Plus-X film, is shown merely to relate the scales used to a familiar system having an ASA rating of 125. *B* represents KHRP with the vertical spread representing the speed difference between plates processed in Kodak D-19 and Kodak HRP developers. *C* represents a dyed version of the same high resolution plates which will be discussed in Section IV of this paper. Similarly *H* represents the sensitivity limits of KOR and AZ1350 photoresists at $\lambda = 436$ nm as shown in Fig. 5 together with resolution limits found in contact printing these systems. These serve to summarize the other sections of this paper showing that photoresists fall short of the goal while the dyed plates, with their enhanced modulation, fall within the goal.

Line *D* represents an interesting extrapolation of the common photo-

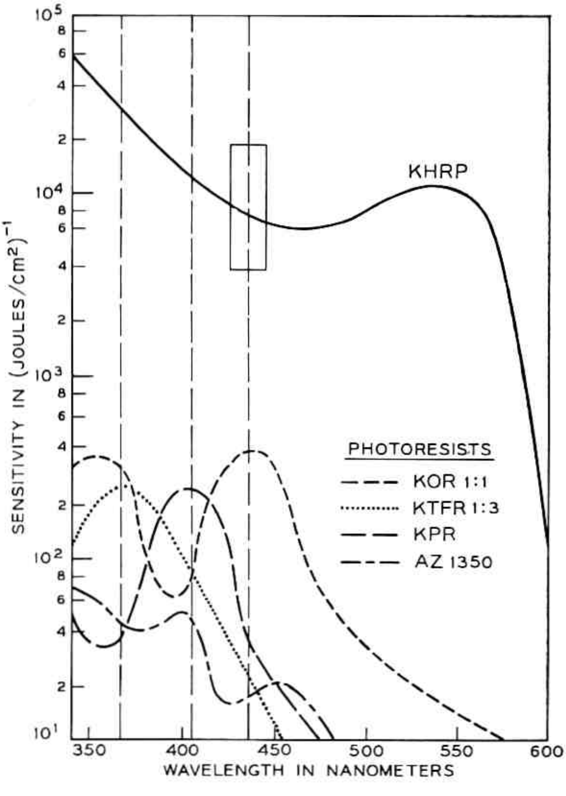


Fig. 5—Spectral sensitivity curves of Kodak High Resolution Plates and some common photoresists. The box about the 436-nm line indicates the reciprocal of the available flash energy range in the step-and-repeat camera. KOR, KTFR, and KPR are photoresist formulations manufactured and sold by Eastman Kodak Company, Rochester, New York; and AZ1350 is a photoresist formulation sold by the Shipley Company, Newton, Massachusetts.

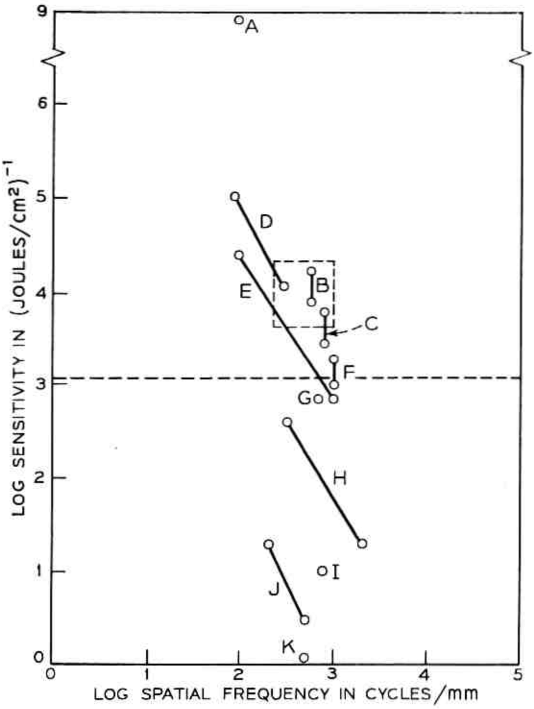


Fig. 6—Sensitivity-resolution correlation diagram showing the target region of the step-and-repeat camera design and the demonstrated performance of a number of unconventional photographic processes identified in the text.

graphic system to its thinnest version, one without the gelatin matrix. This consists of a 0.3-μm film of AgBr evaporated on a glass substrate which, after exposure, may be developed in common photographic developing solutions.[10] While this process does involve amplification and falls within the goal in Fig. 6, it fails as a photomask system because of its very low gamma ($0.3 \leqq \gamma \leqq 1.5$) and its tendency towards infectious fogging on development, i.e., several AgBr grains develop for each exposed AgBr grain.

The line *E* and point *G* represent systems which use photographic physical development as their amplification step. The Philips PD* process is represented by *E* and the Itek RS† process by *G*. [11, 12] The highest-resolution PD-MD1 version of the Philips process does not have the speed necessary for our camera. Neither of these systems is

* N. V. Philips Gloeilampenfabrieken, Eindhoven, The Netherlands.

† Itek Corporation, Lexington, Massachusetts.

at present commercially available in a form suitable for photomask applications.

Another candidate for the appropriate speed range is photopolymerization in which free radical chain propagation steps should provide the necessary amplification. As indicated by *F* one such system has been developed.[13] This is based on the photopolymerization of barium diacrylate to form either an opaque light-scattering image or a clear phase-only image with 0.5-μm resolution. Its speed lies within a factor of five of our requirement; but this data is for a polymerization in aqueous solution approximately 178 μm thick.

In Fig. 6, *I* and *J* refer to organic color-forming photographic systems. The Dupont* Dylux® system *I* develops an intense blue image from colorless precursors on exposure to ultraviolet light; photodeactivation, or fixing, is carried out by exposure to visible light.[14] The "free-radical photography" *J* developed by Horizons Inc.† involves the photochemical reaction of arylamines and carbon tetrabromide leading to a variety of colored images which may be fixed by heating.[15] The resolution capability of each is inherently high because of the molecular nature of the imaging species but their sensitivity is low because they lack amplification steps.* One may calculate the minimum energy necessary to expose at 436 nm a unit quantum yield process to achieve an optical density of 1.0 assuming an extinction coefficient of 10^6 cm^{-1} (the highest known value) for an organic molecule of density 1.0 and molecular weight 300 and result in 0.9 mJ/cm^2. This upper limit to nonamplified photochemical processes is indicated in Fig. 6 by the dashed horizontal line.

Finally, point *K* represents lead-iodide photography.[16] Thin evaporated layers of PbI_2 become transparent when exposed to blue or ultraviolet light at temperatures in excess of 160°C. Similar behavior has been observed in other halides, such as BI_3 and CdI_2, and chalcogenides, such as PbS and Cds. In all cases the sensitivity is very low requiring approximately 1 J/cm^3 for an optical density change of 0.6.

A wide variety of classes of unconventional photographic processes is represented by the above selection, *D* through *K* in Fig. 6, none of which fulfill the requirements of the step-and-repeat camera. This survey does serve to focus our attention on amplified versus nonamplified photographic processes.

* E. I. duPont de Nemours & Company, Inc., Wilmington, Delaware.

† Horizons, Inc., Cleveland, Ohio.

* Note added in proof. One version of Horizon's system is capable of amplification by photo development, but this produces 0.5-μm grain size,

IV. DYED PHOTOGRAPHIC EMULSIONS

Another approach to the solution of the problem is to utilize thinner coatings of the high-resolution photographic emulsion. However, the suppliers claim that thinner coatings cannot be produced with the same degree of uniformity and quality control. We have suggested that as an alternative approach we need only make the usual emulsion effectively thinner.[17] It is known that exposure of photographic emulsions to ultraviolet light in the region of strong absorption by the silver halide results in images confined to the top layers of the emulsion.[18] This behavior may be duplicated in other spectral regions by dyeing the emulsion such that only the top few microns can be effectively exposed. The focal plane of the projection system may also be confined to this same region by the use of distance pieces or pneumatic gauging which ride on the top surface of the emulsion.

What is required is a nonfluorescent water-soluble dye, strongly absorbing of the exposure wavelength, which may be readily and uniformly imbibed by the gelatin and yet may be subsequently removed in normal processing so as not to lower the overall image contrast. Specifically for $\lambda = 436$ nm, I have characterized three suitable dyes, that is, metanil yellow, tartrazine, and naphthol yellow S. In aqueous solution these have somewhat broad absorption peaks with maxima at 435 nm for metanil yellow, 425 nm for tartrazine, and at 390 and 425 nm for naphthol yellow S. The specular optical density (D) of each dyed plate at 436 nm is a linear function of the weight percent (C) dye in the dyeing solution in the low-concentration region of interest. The molar extinction coefficient and the D versus C relationship for each of the dyes at 436 nm are presented in Table I.

The recommended procedure is to dye the plate by five-minute immersion in a gently rocking solution of C weight percent dye plus 0.02 percent nonionic wetting agent. To maintain the initial plate quality, all solutions are filtered to remove particles larger than 0.1 μm, and the dyeing is carried out in clean hoods equipped with type-1A safelights.

TABLE I—DYE ABSORPTION PARAMETERS AT 436 nm

Dye	E (liter/cm–mole)	Dyed Plate Density (D)
Metanil yellow	2.12×10^4	$11.5C + 0.3$
Tartrazine	1.77×10^4	$1.70C + 0.3$
Naphthol yellow S	1.40×10^4	$3.75C + 0.3$

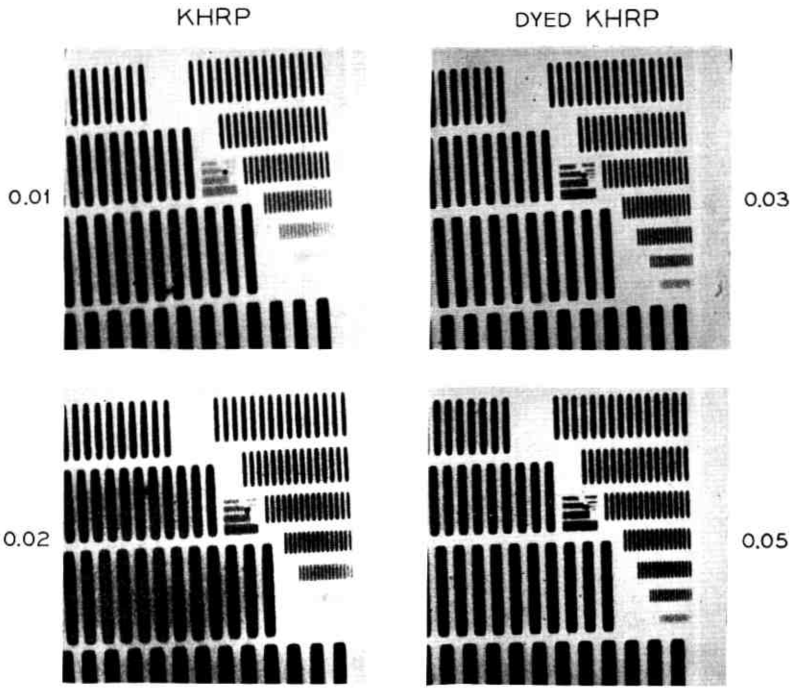


Fig. 7—Photomicrographs of high resolution test target images projected through an $f/1.5$ lens recorded in dyed and nondyed photographic emulsions at the indicated exposure times using 436-nm light.

To characterize the influence of the dye on the photographic response of the plate, monochromatic exposures of a series of plates of varying dye concentration were made through calibrated step tablets. The specular optical density of each step of the developed plates was then measured. The resulting family of characteristic curves for tartrazine-dyed plates indicated that both the speed and gamma, the slope of the characterize curve, of the system decrease with increasing dye concentration. For all further evaluation plates dyed in a 0.2 percent tartrazine solution were selected since their fourfold decrease in speed lies within the exposure capabilities of the step-and-repeat camera. These plates have a specular optical density of 0.64 at 436 nm which is sufficient to eliminate the necessity for an antihalation backing.

A series of exposures of nondyed KHRP and 0.2 percent tartrazine-dyed KHRP were made in a test camera employing the $f/1.5$ lens and 436-nm illumination. Photomicrographs of the results of highest

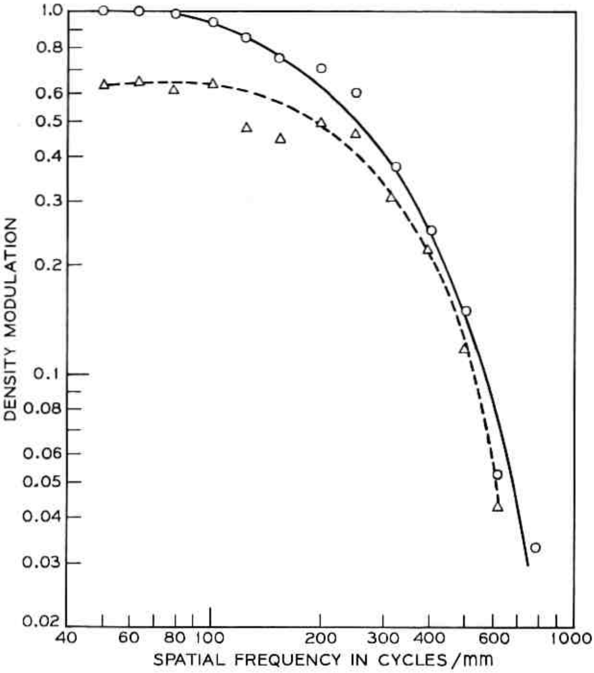


Fig. 8—Density modulation versus spatial frequency of the test images recorded in the dyed plate (circles, 0.05-second exposure of Fig. 7) and the nondyed plate (triangles, 0.02-second exposure of Fig. 7).

resolution are presented in Fig. 7. The bar widths in microns for the eleven 15-bar patterns in the second group on the Ealing* #22-863 test target at 10X reduction are (in the order in which they appear in the photomicrographs, clockwise from 6 o'clock): 5.0, 4.0, 3.15, 2.5, 1.99, 1.58, 1.26, 1.0, 0.79, 0.63 and 0.50 μm. On qualitative comparison the dyed plates appear to resolve finer lines while at the same time providing better modulation of the low frequency bar patterns.

Quantitative measurements of the apparent modulation improvement were carried out on the Ansco Model 4 recording microdensitometer. The 0.02-second exposure of the undyed plate and the 0.05-second exposure of the dyed plate, selected as having the highest resolution on microscopic evaluation, were measured using a 20X, 0.4 N.A. objective with a 5-μm illuminating slit and a 1-μm scanning slit in the microdensitometer. The results are presented in Fig. 8

* The Ealing Corporation, Cambridge, Massachusetts.

in which the density modulation (M) is defined:

$$M = \frac{\bar{D}_{\max} - \bar{D}_{\min}}{\bar{D}_{\max} + \bar{D}_{\min}}. \tag{2}$$

The averages were taken over the seven bars and six spaces at the center of each 15-bar pattern. The data demonstrate the improved modulation of the dyed plate at all frequencies plus the slightly higher resolution capability.

The images thus formed in a dyed emulsion were used to control the exposure of a 0.2-μm-thick film of KTFR photoresist by routine contact printing procedures resulting in usable 1-μm lines, demonstrating that the photographic image had sufficient developed optical density. Thus, all the speed and resolution requirements of the camera are fulfilled by these dyed photographic emulsions with the added benefits of increased modulation of low-frequency images and the elimination of the antihalation backing. At the time of this writing, commercial versions of this dyed high-resolution plate are coming on the market.

REFERENCES

1. Altman, J., "Photography of Fine Slits Near the Diffraction Limit," Photographic Sci. and Eng., *10,* No. 1 (January 1966), pp. 140–143.
2. Skinner, J. G., unpublished work.
3. Stevens, G. W. W., *Microphotography,* New York: John Wiley and Sons, 1968, p. 216.
4. Kerwin, R. E., "Reflections on Reflections in Photoresist," presented at the Kodak Seminar on Microminiaturization, Cherry Hill, New Jersey, June 1, 1967.
5. Middelhoek, S., "Projection Masking, Thin Photoresist Layers and Interference Effects," IBM J. Res. Develop., (March 1970), pp. 117–124.
6. Altman, J., and Schmitt, H. C., Jr., "On the Optics of Thin Films of Resist Over Chrome," Kodak Photoresist Seminar Proceedings, *2* (1968), pp. 12–19.
7. MacChesney, J. B., O'Connor, P. B., and Sullivan, M. V., "Preparation of Iron Oxide Thin Films for Selectively Semitransparent Photomasks," presented at the Electrochemical Society Meeting, Los Angeles, California, May 11, 1970.
8. Goldrick, M. R., and Curran, R. K., unpublished work.
9. Smith, W. J., *Modern Optical Engineering,* New York: McGraw-Hill, 1966, p. 319.
10. Shepp, A., Goldberg, G., Masters, J., and Lindstrom, R., "Evaporated Silver Bromide as a Photographic Recording Medium," Photographic Sci. and Eng., *11,* No. 5 (September–October 1967), pp. 316–321.
11. Jonker, H., Dippel, C. J., Houtman, H. J., Janssen, C. J. G. F., and van Beek, L. K. H., "Physical Development Recording Systems. I. General Survey and Photochemical Principles," Photographic Sci. and Eng., *13,* No. 1 (January–February 1969), pp. 1–8.
12. Berman, E., "Reduction Reactions with Irradiated Photoconductors," Photographic Sci. and Eng., *13,* No. 2 (March–April 1969), pp. 50–53.
13. Brault, R. G., Jenney, J. A., Margerum, J. D., Miller, L. J., and Rust, J. B., "Rapid Access Photopolymerization Imaging," *Applications of Photopoly-*

mers, Washington, D. C.: Society of Photographic Scientists and Engineers, 1970, pp. 113–131.

14. Dessauer, R., "Dylux Instant Access Photosensitive Materials," presented at the Annual Conference of Photographic Sci. and Eng., Los Angeles, California, May 12–16, 1969.
15. Sprague, R. H., Fichter, H. L., and Wainer, E., "New Photographic Processes. I. The Arylamine-Carbon Tetrabromide System Giving Print-Out Dye Images," Photographic Sci. and Eng., *5*, No. 2 (March–April 1961), pp. 98–103.
16. Tubbs, M. R., "High Resolution Image Recording on Photosensitive Halide Layers," J. Photographic Sci., *17*, No. 5 (1969), pp. 162–169.
17. Kerwin, R. E., "Dyed Photographic Emulsions for Improved Recording of Projected Images," Applied Optics, *8*, No. 9 (September 1969), pp. 1891–1895.
18. Mees, C. E. K., and James, T. H., *The Theory of the Photographic Process,* New York: MacMillan Co., 1966, p. 234.

Device Photolithography:

A Computer Controlled Coordinate Measuring Machine

By F. R. ASHLEY, Miss E. B. MURPHY and H. J. SAVARD, Jr.

(Manuscript received June 29, 1970)

In development and operation of the mask-making laboratory, a precise positional measurement system is needed. This paper describes a system based on a Do-all coordinate measurement machine, controlled by a PDP-8 computer. The computer handles all sequential operations as well as computation necessary for coordinate transformation and feature location. The result is a system which can measure an array of 208 points to an accuracy of ±1 μm in less than two hours. Without computer control, measurement of such an array is not feasible.

I. INTRODUCTION

In the design of the mask-making laboratory, the need for a precise positional-measurement system was recognized. This system is needed for alignment and adjustment of the primary pattern generator (PPG), the reduction cameras and the step-and-repeat camera. It is also needed for mask inspection. The measurement system should be at least ten times more precise than the tolerance on the masks being measured, and it should be capable of measuring a large number of points in a reasonable time. For example, the test pattern to align the PPG is an array of 208 points, and this should require no more than two hours to measure. Table I summarizes the requirements on the measurement system.

1.1 *System Description*

A Do-all Coordinate Measurement Machine (CMM) controlled by a PDP-8 computer forms the basis for the measurement system to meet these needs. This is shown schematically in Fig. 1. The Do-all machine consists of two air-bearing slides at 90° on black granite ways. The plate to be measured is mounted on one slide (x-axis) and

Table I—Performance Objectives for CMM System

Field X	18 CM
Y	22 CM
Optical Power	250
Projected Field	400 μm
Feature Location	±200 μm
System Precision	±0.08 μm
Slew Rate (Both Axes)	0.5 CM/SEC
Plate Measurement Time (208 Points)	<2 hrs.

a microscope with projection screen is mounted above the plate on the other slide (y-axis). The range of travel on the x and y axes is 18 cm and 22 cm respectively; by appropriate adjustment of the x- and y-axes slides, the microscope can be positioned above any point on a plate within an 18-cm by 22-cm range. Stepping motors move the x and y slides through taut wire capstan drives. Fine manual positional adjustment is provided for by two torque transmitter and receiver pairs on a separate control panel. Fringe counting interferometers using a HeNe laser source provide precise positional information on the x and y axes. The counters display the total counts (1

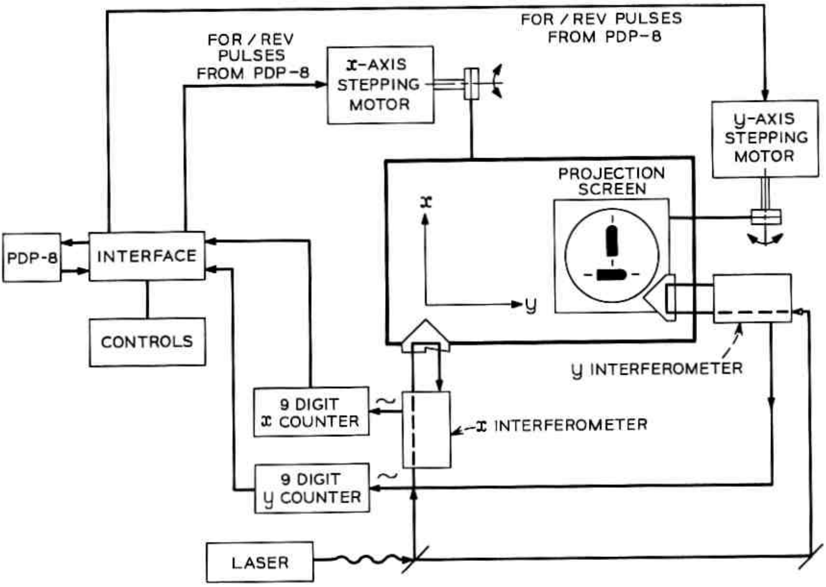


Fig. 1—Block diagram of measurement system.

count = 0.0791 μm) that the x and y slides have moved from some predetermined origin. Plate features to be measured are optically projected onto a screen to allow the operator, operating as a feedback element, to trim the location of the plate feature with respect to a reticle on the projection screen. The interface provides an interaction channel between the CMM, the PDP-8 and the operator.

II. ADVANTAGES OF COMPUTER CONTROL

The use of a general purpose computer as a control element for the measurement system has a number of advantages over a "hard wired" controller. First, there is the flexibility that the stored program allows. Modifications of the control functions and correction of errors are done by changing the program, not wiring. Very often this involves changing just a few instructions in memory and can be done right at the computer console in a matter of minutes. Second, the computer offers much greater input-output capacity; the system can be expanded to use disk or magnetic tape if required for future needs. The third advantage is that the computer can transform the coordinate system of a plate to the CMM coordinate system. This transformation can take into account (*i*) plate rotation with respect to the CMM, (*ii*) the deviation of the angle between the x and y axes of the CMM from 90° (skew angle), and (*iii*) conversion of units of counts to metric units or address units of the plate. A fourth advantage is that the computer allows feature location on a plate. Since the computer is interfaced to read the x and y counters and to pulse the x- and y-stepping motors, a computer program can be written to position the CMM microscope over any desired point on a plate.

III. MEASUREMENT SYSTEM DESIGN

The two basic areas of design in the CMM system—design of the interface and program design—are discussed in the following paragraphs.

3.1 *Interface*

The interface was designed with simplicity as the objective, at the possible expense of more programming. This is feasible because of the high speed of the PDP-8 relative to the mechanical speed of the CMM. The block diagram for the interface is shown in Fig. 2. The interface allows the computer to read information (counter readings, control switches and data inputs) into its accumulator, and allows the com-

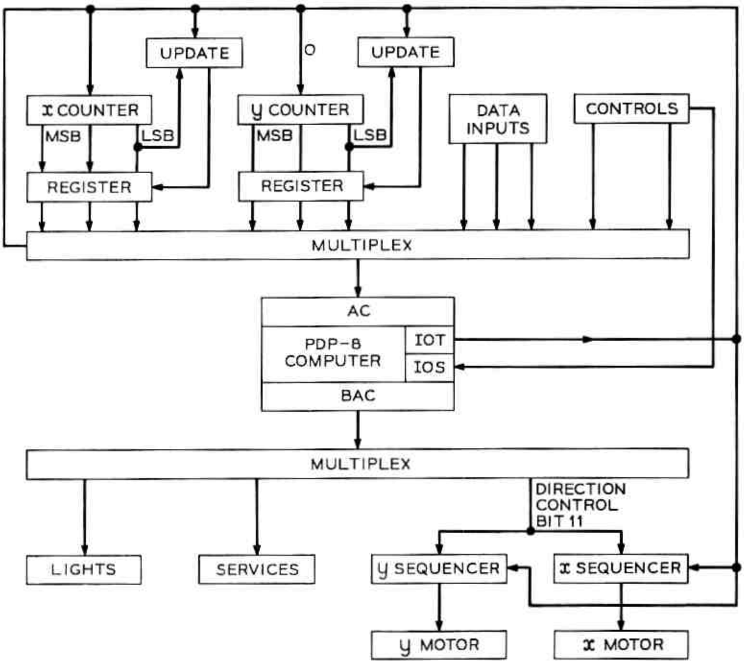


Fig. 2—Interface block diagram.

puter to transfer data from its accumulator to external devices. It also allows input-output transfer (IOT) pulses to be output by the computer.

The x- and y-axes stepping motors are driven by IOT pulses which occur in response to IOT instructions in the computer program. The direction of the step is controlled by the accumulator. One step of the motor causes a displacement of about 250 counts (20 μm) on either the x- or y-axis. The maximum rate of the motors is 200 steps per second giving a slew rate of 4 mm per second.

The x and y counters are nine-digit counters with binary-coded decimal (BCD) output. Since one computer word is only 12 bits, it is necessary to provide a 36-bit storage register for each counter. This register is read into the computer in three 12-bit bytes. Care must be exercised in transferring the outputs of a counter to its storage register, in that the transfer must not occur when the counter is in a transition. This is taken care of by an update circuit which transfers the counter outputs to the register only at a fixed time delay after a change in the least-significant bit (LSB). This time delay is chosen to be less than the time between input transitions to the counter. The routine to read one counter and store its BCD output in memory requires 33 μs.

There are a number of manual controls available for the operator. These are operable only when the program is in the manual control mode (Fig. 3). These controls consist of switches whose state is sensed by the computer through the accumulator bus. A toggle switch MANL allows the operator to enter the manual mode. Push-button switches X+, X−, Y+, Y−, allow the operator to manually position the microscope by pulsing the x- or y-stepping motors at a 200 pps rate. Push button switch CLEAR causes a present counter reading to be stored in core and then both counters to bet set to zero. Push button switch OUTPUT causes the present position in address units to

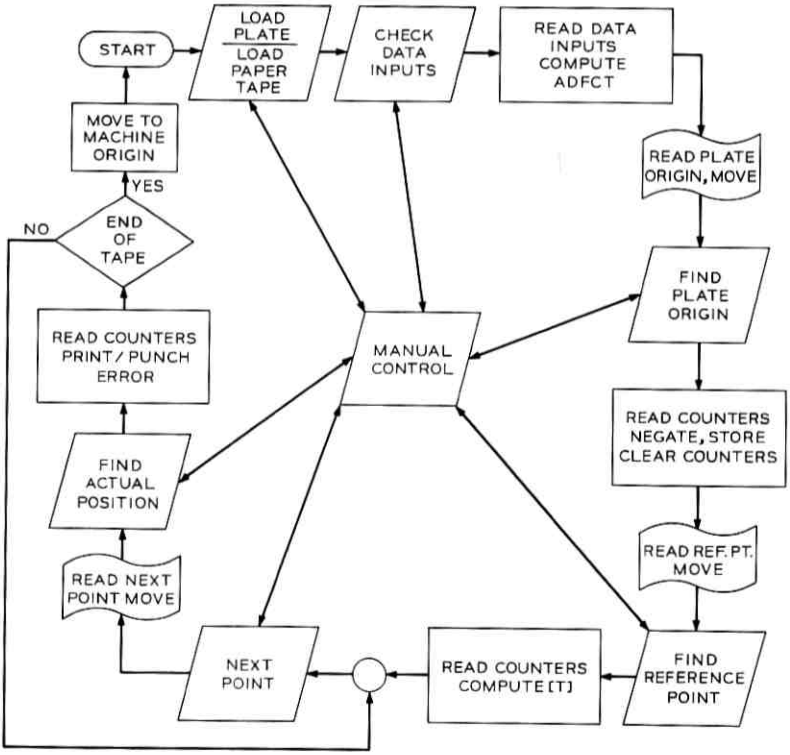


Fig. 3—Flowchart of control program.

be output on the teletype. Toggle switch TTY allows the operator to input x-y coordinates in plate-address units; the MOVE routine, described later, is then entered and causes the desired coordinates to be located.

3.2 *The Computer Program*

A simplified flow chart for the control program is shown in Fig. 3. The function of the control program is to provide a proper sequence of steps that will result in measurement of a plate. The slanted side boxes represent message lights that are illuminated when that point of the program is reached. The program progresses from one slanted side box to the next as the operator presses a continue push button on the control panel. The computer cycles in a loop while waiting for this operation. The manual mode can be accessed from any of the waiting loops, and is represented by the slanted side box in the center of Fig. 3. The boxes with curved tops and bottoms represent paper-tape input from the high-speed reader. In reading coordinates from the paper tape, a program XYINPT is called to transform plate coordinates to CMM coordinates. This is accomplished by the following matrix equation:

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix} = \frac{ADFCT}{\cos \varphi} \begin{bmatrix} \cos(\theta - \varphi) & -\sin(\theta - \varphi) \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_p \\ y_p \end{bmatrix}. \tag{1}$$

Figure 4 shows the plate and CMM coordinates systems. The rotation of the plate with respect to the CMM is denoted by θ, while φ denotes the skew angle. The quantity $ADFCT$ is a constant to convert address units of a plate to counts of the CMM.

$$ADFCT = 8\ N/\lambda \text{ counts/address.} \tag{2}$$

N is the address size in μm, and λ is the wavelength ($\approx .6328\ \mu$m) of the HeNe laser.

To compute positional errors, it is necessary to convert the CMM coordinates to plate coordinates. This is done by inverting the transformation (1)

$$\begin{bmatrix} x_p \\ y_p \end{bmatrix} = \frac{1}{ADFCT} \begin{bmatrix} \cos \theta & \sin(\theta - \varphi) \\ -\sin \theta & \cos(\theta - \varphi) \end{bmatrix} \begin{bmatrix} x_m \\ y_m \end{bmatrix}. \tag{3}$$

3.2.1 *Feature Location*

When the coordinates of a feature on a plate are known, a routine called MOVE can be used to cause the microscope to be positioned over

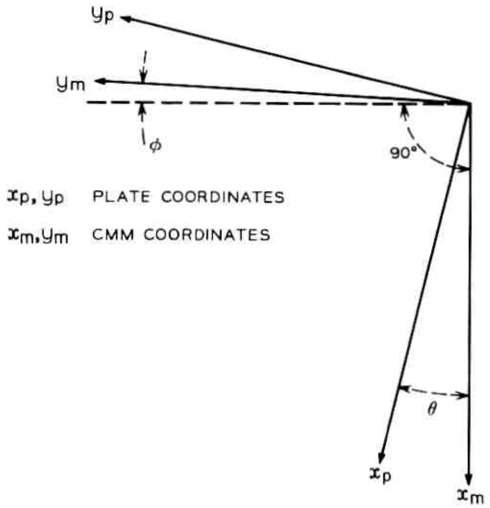


Fig. 4—CMM and plate coordinate systems.

the feature. This is done through the stepping motors. Since one step is 20 μm, the positioning accuracy is ± 20 μm on each axis, which is well within the field of the microscope. Thus, to locate a feature, the following procedure takes place for each axis:

(*i*) The desired counter reading is computed using equation (1) and is stored in the PDP-8 memory.

(*ii*) The quantity Δ, which is the difference between the desired counter reading and the present counter reading, is computed.

(*iii*) If $|\Delta| < 500$ counts, the feature is located and the procedure terminates; otherwise go to (*iv*).

(*iv*) A pulse is applied to the stepping motor; the sign of Δ determines the direction of the step. After a 5-ms delay (to allow the motor to complete its step) go to (*ii*).

Thus, while slewing to a position over a feature, the x- and y-stepping motors are pulsed continuously, and between pulses, the counters are read to see if the desired readings are obtained. The sequencing of MOVE insures that the first time the desired condition of being within 500 counts (40 μm) of the feature is obtained, then the motion of that axis is complete. No more pulses are output to that stepping motor. This eliminates hunting that would occur due to the time lag between application of a pulse to a stepping motor and motion of the slide.

IV. MEASUREMENT OF A PLATE

The process of measurement of a plate is illustrated by the flow chart of Fig. 3. Initially, the microscope is positioned over a fixed point that is the CMM origin. Prior to measurement of a plate, a paper tape is made containing the distance from the CMM origin to to the origin of the plate coordinate system. Following on the tape are the coordinates of a reference point on the x-axis, followed by the coordinates of all points to be measured; these points are expressed in the plate coordinate system. This paper tape is placed in the high-speed reader of the PDP-8, and is read under program control. Also prior to measurement, three data-input thumbwheel switches are set. These contain the CMM skew angle in seconds of arc, the factor N which is the address size in microns and the three least-significant digits of the wavelength of the HeNe laser. The wavelength of the HeNe laser is represented to an accuracy of seven digits in the PDP-8 -0.6328XXX μm, where XXX is read in on thumbwheel switches.

During the preliminary steps of the measurement, the plate origin is located and both counters are cleared. Prior to being cleared, both counters are read and the negative of their readings are stored in the PDP-8 memory. This enables the program to return the CMM microscope to the CMM origin at the end of the measurement. The reference point on the x-axis is located next. This enables the PDP-8 to compute the angle θ, $\theta = \tan^{-1} (y_{\text{ref}}/x_{\text{ref}})$, where $(x_{\text{ref}}, y_{\text{ref}})$ are the coordinates of the reference point on the x-axis as read from the CMM counters. The program now has enough information to compute the coordinate transformations (1) and (3). The program then proceeds to measure points as they are read from the tape, printing the errors on the ASR 33. An end of tape character signals the last point to be measured and causes the programs to terminate with the microscope positioned over the CMM origin.

V. CONCLUSIONS

The measurement system as described has been successfully used to measure plates from the PPG, the 3.5X reduction camera and the step-and-repeat camera. The main limitation on accuracy is the ability of the operator to align the desired feature with the reticle of the microscope. This in turn depends very much on the line-edge definition. For example, features generated by the PPG have been measured with ±1-μm accuracy; the PPG generates line edges that are typically defined over a 5-μm distance.

The objectives on measurement time also have been met. A test array of 208 points for alignment of the PPG can be measured in less than two hours. Without computer control, the measurement time would be so long that drift problems and operator fatigue would make the measurement unfeasible.

VI. ACKNOWLEDGMENTS

The authors are grateful to A. Zacharias and K. M. Poole whose guidance and support provided the basis for this work. J. W. Stafford and L. Rongved provided consultation on mechanical problems. Thanks also are due to Mrs. E. E. Yamin for the use of her assembly program written on the GE 635 to assemble code for the PDP-8.

APPENDIX A

Skew Angle Measurements

The skew angle Φ as defined in Fig. 4 is the deviation from 90° of the x and y CMM axes. To measure Φ, a glass plate having three marks is used. An origin mark and marks on lines approximately 90° apart define the xp and yp axes as shown in Fig. 5. First the plate is placed on the CMM as shown in Fig. 5a with the axis of the plate roughly aligned with the CMM axis. Angles A and B may now be accurately measured by use of the CMM counters resulting in the following equation:

$$A + 90° + \Phi = B + 90° + \Phi p. \tag{4}$$

The measurement is then repeated with the plate rotated approxi-

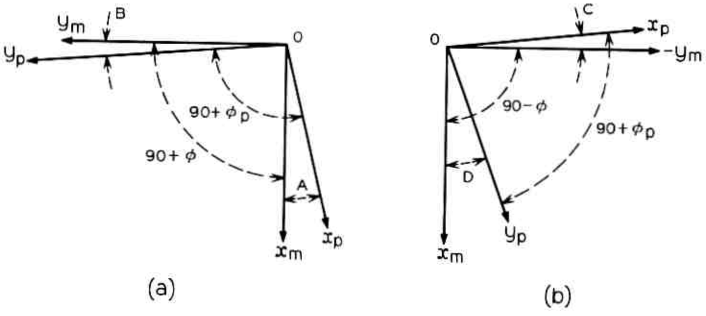


Fig. 5—Skew angle measurements. (a) Plate aligned with CMM axes. (b) Plate rotated 90° from CMM axes.

mately 90° with respect to the CMM axis as shown in Fig. 5b. Angles C and D are now measured resulting in the equation:

$$C + 90^\circ - \Phi = D + 90^\circ + \Phi p. \tag{5}$$

Equations (4) and (5) may be solved simultaneously to give the equation:

$$\Phi = \frac{(B + C) - (A + D)}{2}. \tag{6}$$

Measurements of this type indicate that Φ is about 11.5 seconds of arc. Φ must be measured after any disassembly of the y-axis support.

The Mask Shop Information System

By MRS. J. G. BRINSFIELD and S. PARDEE

(Manuscript received June 17, 1970)

The Mask Shop Information System (MSIS) is a set of computer tasks which exist in a specially designed multi-programming environment within a PDP-9 computer, and which control the flow of jobs through the new mask-making facility. The main functions of MSIS are to accept job descriptions and to assign tasks and pass data to the various shop facilities so that these jobs can be efficiently processed. In addition, MSIS keeps statistics on the progress and problems of the shop and issues reports both periodically and upon demand.

I. INTRODUCTION

A computer-based information and control system, referred to as the Mask Shop Information System (MSIS), assists in running the Bell Telephone Laboratories mask-making facilities at Murray Hill, New Jersey, and Allentown, Pennsylvania. In the planning stage for the new mask-making facility, it was realized that the scheduling and processing information required for efficient control of the flow of jobs would be too complicated to handle with paper work. Furthermore, keeping track of the large number of glass plates passing through the facility would be a problem. Thus, it was decided to develop MSIS. Briefly, MSIS controls the entire mask-making facility and serves as a repository of information on the status of each job, the location of each plate, and the performance of the overall facility. The equipment that makes up the new facilities has been discussed elsewhere.[1–3]

The first part of this paper will deal with the functions performed by MSIS and how the system appears to the user; the second part will describe the organization of computer programs and data required to implement MSIS.

II. MSIS FUNCTIONS

The significance of MSIS can best be grasped by reviewing the various functions that are performed.

2.1 *Scheduling*

MSIS schedules each process step required to complete a mask. This scheduling is done on-line and allows for the inclusion of jobs of varying priority. As each task is completed by either a human operator or a machine, such as the primary pattern generator (PPG), MSIS determines the highest-priority task waiting and assigns it to the operator or machine that is idle.

2.2 *Control Information*

MSIS transmits control information over wide-band data links to the two control computers attached to the PPG and the step-and-repeat camera. For the PPG, this information indicates the magnetic tape reel number and the file number within that reel that contains the information describing the artwork to be generated next. For the step-and-repeat camera, the identification of the specific reticles needed to make a particular mask as well as the step-and-repeat array information is transmitted from MSIS to the control computer. Other control information is transmitted directly to human operators via special displays and teletypewriters.

2.3 *Information Storage*

MSIS maintains an extensive disc file containing information such as

(*i*) the status of every job in process,

(*ii*) performance statistics covering each process step as well as the overall mask-making facility,

(*iii*) inspection information required to define special mask features that should be inspected in detail, and

(*iv*) the step-and-repeat array information necessary to define a complete mask for silicon circuits.

2.4 *Glass-Plate Handling*

To avoid the confusion of human operators sorting through a mountain of glass plates to find a specific reticle, piece of artwork, or mask, MSIS assigns each piece of glass to a numbered slot within a numbered carrier. The location of each piece of glass is remembered so that when it is needed as the input to another process step, its exact location can be supplied to the operator.

2.5 *Inquiries and Reports*

MSIS will allow certain on-line inquiries to be made from a teletypewriter terminal. Some on-line inquiries might be:

(*i*) What is the status of my job?
(*ii*) What is the backlog of work for the reduction cameras?
(*iii*) How many pieces of artwork have been generated this shift?

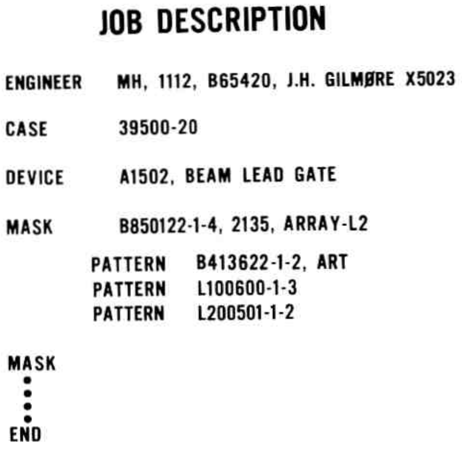
In addition to these short on-line inquiries, more detailed management reports will be generated by MSIS on a daily, weekly, monthly, or quarterly basis. Certain of these management reports will also be available on demand.

III. USER/COMPUTER INTERFACE

There are two sets of users that must interface with MSIS. The first user is the engineer or designer who wishes to request that a particular set of masks be manufactured. The other is the mask-shop operator who must exchange information with the computer system while completing his job.

The engineer or designer will communicate his needs to MSIS by a set of instructions on punched cards that can be included in his XYMASK input deck or can be submitted separately with the post-processed XYMASK tape that is required. Figure 1 shows an example of these instructions for a typical set of masks for a silicon circuit. Both tantalum- and silicon-circuit masks can be handled with equal ease; but a silicon circuit is used in this example because in general it requires that more information be supplied.

The first two cards contain standard identifying information; an engineer might have a number of these cards duplicated to have when



```
                JOB DESCRIPTION

ENGINEER   MH, 1112, B65420, J.H. GILMØRE X5023

CASE       39500-20

DEVICE     A1502, BEAM LEAD GATE

MASK       B850122-1-4, 2135, ARRAY-L2
               PATTERN   B413622-1-2, ART
               PATTERN   L100600-1-3
               PATTERN   L200501-1-2

MASK
 •
 •
 •
END
```

Fig. 1—Job-description information.

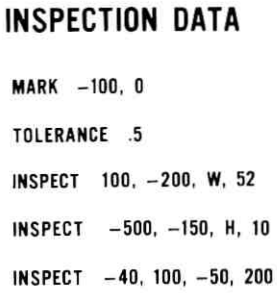
needed. The third card identifies the circuit and is used primarily on reports for ease of identifying the jobs. A MASK card is included for each mask of the complete job. This MASK card contains the drawing-level-issue number of the particular mask and an optional process number. The process number will be imaged by the step-and-repeat camera onto the final mask for use during circuit fabrication. The final field on the MASK card indicates, in this case, that a prestored step-and-repeat array (L2) is to be used in making the mask. It is hoped that most masks can be specified using one of a number of prestored array definitions. For those masks that require special arrays, a means is provided for defining the array along with the job description. In the example of Fig. 1, assume that three patterns are required to complete the desired mask, that is, the primary pattern for which artwork must be generated and two standard test patterns (L100600-1-3 and L200501-1-2) for which reticles already exist. Similarly, for each mask that makes up the job a MASK and three PATTERN cards would be required.

3.1 *Inspection Data*

If special features on a mask are to receive specific inspection, a series of cards, as shown in Fig. 2, can be included. These cards indicate

(*i*) the coordinates of a fiducial mark;
(*ii*) the tolerance, in microns, to be maintained;
(*iii*) the coordinates of a feature and its desired width;
(*iv*) the coordinates of a feature and its desired height; or
(*v*) the coordinates of two vertices of a feature whose edges do not parallel the X and Y axes.

At inspection time, MSIS scales this information appropriately, depending on the inspection being carried out, and presents the scaled information to the inspector.



INSPECTION DATA

```
MARK  -100, 0

TOLERANCE  .5

INSPECT  100, -200, W, 52

INSPECT  -500, -150, H, 10

INSPECT  -40, 100, -50, 200
```

Fig. 2—Mask-inspection information.

3.2 *Operator Interface*

Operator-to-computer communications is carried out by two different means. Operators involved with the reduction cameras, contact printing, chrome etching, and the step-and-repeat camera will communicate with the system via a combination of cathode-ray-tube displays and keyboards. Administrative information and inspection data is communicated via standard KSR35 Teletypewriters.

3.3 *Secondary Information Strip*

Another medium for conveying information required by both the users and the computer is the secondary information strip. Figure 3 shows the relationship of this strip to the primary artwork as it comes from the PPG (not drawn to scale). Two items of information are placed in the strip in both human- and machine-readable form. These are, the drawing number (for example, B123456-4-3) and the magnification that was used in drawing the artwork (for example, 35). The human-readable portion is intended to allow the operators to verify visually that they have the proper piece of glass or are using the proper reduction camera. The machine-readable representation is repeated twice as a series of coded clear and opaque spots. One set of coded information can be read by an array of photo-diodes mounted in the reduction camera. The other set is imaged onto the reticle produced by the reduction process and can be read by photo-diodes in the step-and-repeat camera. In both cases, MSIS uses the machine-readable information to insure that the proper artwork, or reticle, is mounted in the proper device before allowing a job step to proceed.

Across the top of the artwork is another piece of encoded information. This represents the particular PPG on which the artwork was manufactured, and a sequential serial number. The MSIS does not make use of this latter information.

IV. EQUIPMENT CONFIGURATION

Figure 4 shows the overall equipment configuration for the mask shop. In the center of Fig. 4 is the MSIS main computer, a Digital Equipment Corporation PDP-9. The characteristics of this machine and its associated hardware are shown in Table I. The MSIS computer is interfaced via high-speed data links directly to the control computers associated with the PPG (PDP-9) and the step-and-repeat camera (PDP-8). Two model 35 KSR Teletypewriters are connected to the system. One is for administrative purposes and the other for

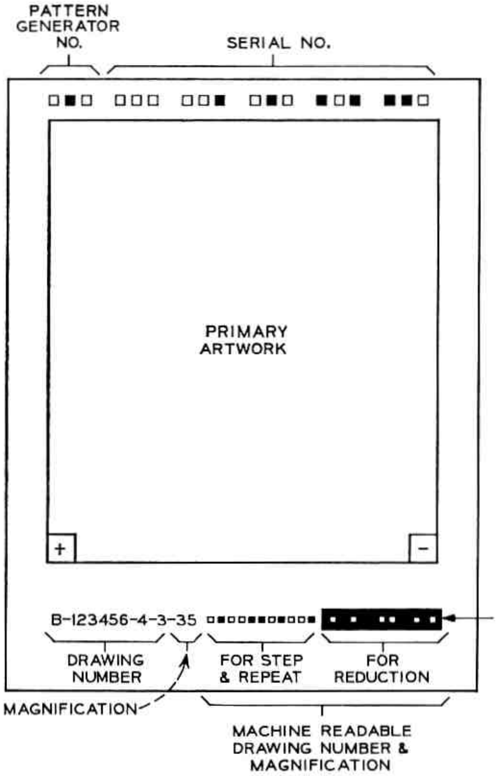


Fig. 3—Secondary information strip.

use by the inspectors. Three keyboard display positions are also connected to the system. Each position consists of a Tectronix 611 Storage Display and a 16-position keyboard. These are used to communicate with operators in the reduction, contact-printing, and chrome-etching areas. Additional keyboards and displays are connected to the two control computers.

V. A TYPICAL JOB

To help understand the functioning of the MSIS, it would be instructive to trace a typical mask as it flows through the system. A silicon-circuit mask will be used as the example (although the system is designed to handle both silicon- and thin-film-circuit masks)

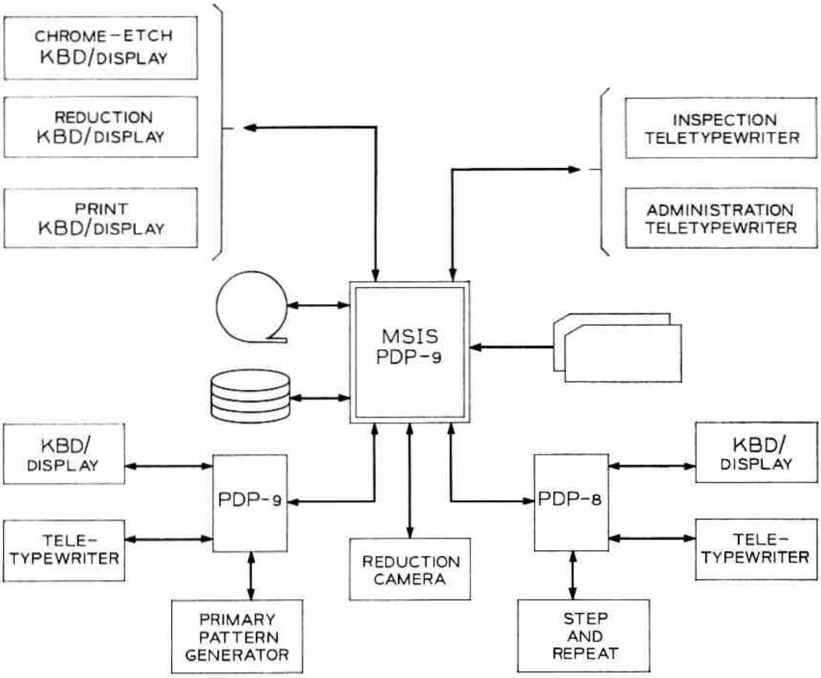


Fig. 4—System configuration.

since it uses more facets of the system. Figure 5 shows the flow of the mask through the system.

When an engineer feels he has adequately debugged his circuit masks using XYMASK, he will add the job description cards described earlier to his XYMASK deck and make a final computer run to generate a computer tape for use by the PPG. This tape will also

TABLE I—MSIS/PDP-9 HARDWARE CHARACTERISTICS

Devices	Characteristics
Core Memory	$8K$ words, 18 bits, 1-μs cycle time
Disk Memory	1 million words, 17-ms average access time
Magnetic Tape	9 track, IBM compatible, $30K$-character/second
Paper Tape	8 level, 300-characters/second reader, 60-characters/second punch

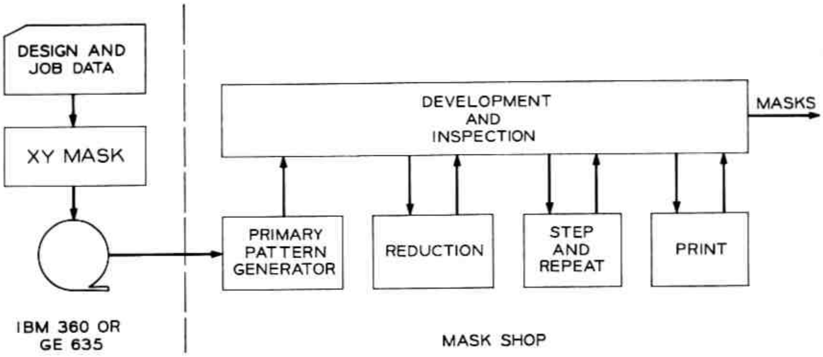


Fig. 5—Trace of a typical mask as it flows through MSIS.

contain the job-description information. The reel of magnetic tape is submitted to the mask shop and is mounted on the magnetic-tape unit attached to the MSIS. The job-description information is read by the MSIS and an instruction is given to save the reel in a particular numbered bin. The designations of the desired masks are placed in a queue awaiting the PPG. Their position in the queue is based on the priority assigned to the job.

When the mask in question reaches the top of the PPG queue, a message is sent from the MSIS to the PPG control computer indicating the reel number, bin, and file within the reel where the XYMASK data can be found for the desired mask. When the artwork has been completed, the PPG control computer transmits an appropriate message to the MSIS. The mask designation will be removed from the PPG queue and assigned to a table of those masks undergoing photographic development.

No attempt is made to schedule or control the work as it passes through the photographic development area. After completing photographic development, the artwork is passed to inspection and it is "logged in" at a teletypewriter. The MSIS instructs the operator to place the artwork in a particular numbered slot of a numbered carrier. At the same time the mask designation is placed on the inspection queue. When it reaches the top of the queue, the inspector is told by the MSIS where to locate the artwork and what unique features should be inspected.

Assuming the artwork passes inspection, the operator signals MSIS via the teletypewriter and the mask designation is placed on the reduction queue. Again the artwork is assigned to a unique slot. When the mask again reaches the top of the queue, a message is displayed to the reduction camera operator to mount the artwork in a particular reduction camera. The MSIS checks to insure that the proper artwork is mounted in the proper camera by scanning the secondary information strip. If it is properly mounted, the MSIS initiates exposure. The exposed reticle is then passed to photographic development.

The development-and-inspection cycle is repeated again, and the reticle is passed to the step-and-repeat camera. When the mask designation reaches the top of the step-and-repeat queue, the MSIS transmits all the step-and-repeat array information, including all reticles required, to the step-and-repeat control computer. Again at each stage of the step-and-repeat process the secondary information strip is checked to insure that the proper reticle has been mounted in the camera.

Upon completion of the step-and-repeat process, another development-and-inspection cycle occurs with the mask being passed to the print area. At the proper time, the number and type of prints required are displayed to the operator. After the prints have been made, a final development-and-inspection cycle occurs and the finished masks are available.

VI. MSIS SYSTEM PROGRAM STRUCTURE

During the early design stage of MSIS, it became obvious that:

(*i*) To keep and continually update the data required to process jobs in the shop, such a large number of disk-memory accesses would be required that the system performance would be limited by the ability to read from disk memory.

(*ii*) MSIS would be continually receiving requests for service either directly or indirectly from about 12-15 shop operators and would have to answer within reasonable human-response times. Furthermore, due to the difference in characteristics of the shop facilities, some of this communication could be handled via speedy interfaces such as data links while other input/output would have to be handled at relatively slow teletypewriter speeds.

(*iii*) To allow demand and periodic reports on shop progress and periodic checks of data to prevent potential problems, there would have to be some programs with long processing times included in the system.

It was decided that the best computer system for solving these problems would be a multi-programming system with a task-priority scheme. Since such an operating system did not exist for the PDP-9, the programming of a monitor had to be included in the MSIS project.

With the present MSIS multi-programming monitor, execution of one program can go on simultaneously with block transfers of data to and from disk for another program. Furthermore, by keeping those programs that have long processing times or that use slow input/output devices in separate execution areas from faster running programs, it is possible to provide quick operator responses and still run lengthy programs. Using a task-priority scheme, those tasks* which must provide quick response can be given high priority and thus processed much more quickly than the slower report and data-checking programs.

With only 8K words of core memory, the luxury of a complex monitor that allows dynamic allocation of execution areas and relocatable programs cannot be afforded. Thus the core memory is divided into fixed execution areas. Each task is assigned to an execution area according to its characteristics. Another restriction used to simplify the monitor is that swapping of tasks in the midst of execution is not permitted. That is, once a task is in execution in an execution area, no other task that requires the same area can be executed until the present task is completed.

The layout of the PDP-9 core memory is illustrated in Fig. 6. The first 3300 words are taken up by the monitor. Approximately the same amount of space is divided into 6 execution areas for task processing. The remainder of core memory is used as a "common" area to provide communication of data between the tasks.

6.1 *Monitor*

The monitor comprises three main modules: the task sequencer, interrupt handler and input/output control:

The heart of the monitor is the task sequencer; its basic function is to determine the sequence in which the tasks should be executed. The relative priority of the various tasks that make up the MSIS is determined by their relative order as they appear in the task list.

Whenever it is necessary to determine which task is to be executed next, the task sequencer scans the task list from the beginning. It searches for the first task which is activated, which belongs to an

* Throughout this paper "task" is used to indicate a collection of computer subroutines that perform a particular function. These tasks are usually stored on disk memory and brought into core memory only when needed.

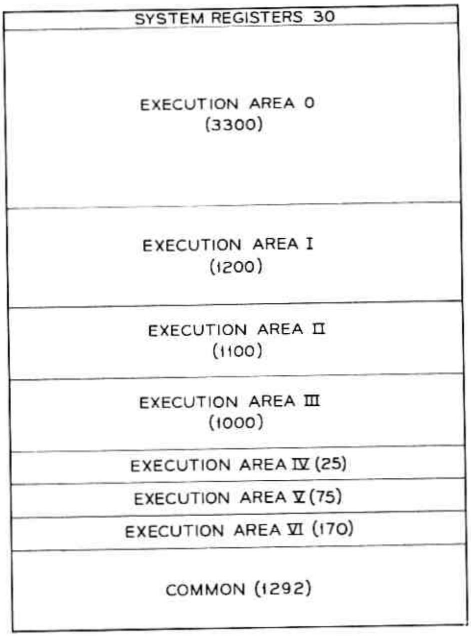


Fig. 6—MSIS PDP-9 core memory layout.

execution area that is not already in use by another task, and which is not still waiting for the completion of an I/O transfer. If the chosen task has already started execution, the task sequencer restores the registers, and returns to the place where it was interrupted. If the selected task is ready to start from the beginning and in core, the task sequencer transfers to its starting location. If the task has yet to be read in from disk, the sequencer calls I/O control to perform the disk transfer and continues its search.

The interrupt handler executes whenever an interrupt occurs as the result of the completion of an I/O transfer or an overflow of the real-time clock. The interrupt handler immediately saves the registers for the program in execution, and then decides what caused the interrupt. If the interrupt was caused by a special keyboard, a reduction-camera signal, a data link, or a request for attention (carriage-return) from one of the teletypewriters, the interrupt handler sets up some common data words and activates the task required to handle the input.

If the interrupt was one of a series of interrupts that occur in the process of completing an I/O transfer (such as the transfer of one character of a teletypewriter message), the interrupt handler stores the data and/or sets up transfer of the next data, and updates some common words to keep track of the status of the overall message. If the interrupt marks the end of a transfer (such as the last column of a card), the interrupt handler sets up the appropriate common words and activates the I/O control program. When all interrupts have been handled, control is given to the task sequencer.

The purpose of the I/O control program is to make I/O transfers via the card reader, magnetic tape, disk, paper tape, and both teletypewriters appear to the application tasks as fully buffered operations that can be handled immediately through subroutine calls. Actually, fully buffered I/O occurs only with the magnetic tape and disk. And when I/O control is called by an application task, it simulates an interrupt and locks the task from execution so that the task sequencer will allow other tasks to execute while the data transfer is taking place. Input/output requests for busy devices are queued and initiated as soon as the device is free. Five retries are made when disk-and-tape parity errors occur. For all other errors, an error is returned to the application task.

6.2 *Execution Areas*

There are six execution areas for the MSIS application tasks. Their size and arrangement are illustrated in Fig. 6.

Execution area I contains the highest-priority disk tasks. These are tasks which lower-priority tasks activate to accomplish activities requiring an update of shared data blocks that are stored on disk. Since all tasks that are allowed to update these shared data blocks at crucial times are included in execution area I, they can never be in execution simultaneously and thus no updates can be lost due to "race" conditions between two tasks. The scheduler task, which handles all queue manipulations, plate carrier assignments and job-status updates, is located in execution area I. The allocate task, which dynamically allocates and restores disk space, is also located in execution area I.

All low-priority disk tasks use execution area II. These are tasks which can afford to wait for their execution area to be free without appreciably slowing up the processing of tasks. There are 25 tasks presently assigned to this area whose functions include the following: entering and deleting jobs from the system, initializing tables at the beginning of a shift, asking for plate carriers to be moved between facilities, listing the contents of a plate carrier upon request, asking for shop output to be delivered to the engineer, and reporting on shop progress and shop problems.

Disk tasks which have medium priority use execution area III. A medium-priority task is one which, in general, doesn't have another task waiting for its completion, but which does have an operator waiting for a response. The tasks which use this execution area include those which control the task assignments for the PPG, reduction cameras, step-and-repeat camera, and inspectors.

Execution areas V, VI, and VII contain small, high-priority in-core tasks. One of these tasks provides a check against the failure of an I/O device to respond, which would cause a tie-up in the system. The other two tasks accept requests from the two teletypewriters and decode the messages to decide what disk task should be activated to handle the request.

6.3 *Common*

As has already been pointed out, most of the common data area is used to pass data between tasks. Another use for this common data area is to allow a task to save crucial data from one execution time to another without requiring the data to be stored on disk. For instance, the scheduling task saves the top of each of its facility queues in core so that it can perform most queue manipulations without taking the time to access disk. Also, the facility control tasks save the description of the mask presently in process in core because the control task is usually activated several times before passing one mask through the facility.

VII. MSIS DATA STRUCTURE

The bulk of the MSIS data is kept on disk in data sets called description blocks. While a particular description block is part of MSIS, its location on disk remains constant so that a "pointer" to its location on disk is a unique and unchanging number. These disk pointers are used to set up a structure of rings and linked lists that unite the data for one job even though the data is not in one contiguous area. In this way, disk segments (64 words) can be allocated in a random fashion, avoiding the problem of collecting a contiguous data area large enough for a particular job. The disk pointers are used in tables that correlate the data with names that have meaning to the shop operators so that teletypewriter requests for specific data can be made.

7.1 *Description Blocks*

There are five kinds of description blocks: job-description block (JDB), mask description block (MDB), pattern description block (PDB), inspection description blocks and step-and-repeat-array description blocks. The job, mask, and pattern description blocks have a fixed length of one segment. The inspection and step-and-repeat-array description blocks can be any number of linked segments.

When a job is entered in the MSIS, its processing data is split into description blocks. As the job is processed, additional data is added to these blocks and this data can be retrieved at any time.

The data common to all masks of a job, such as the engineer's name and the circuit code, are kept in a JDB. The data common to all patterns of a mask, such as the mask-identification number and the present status of the mask, are kept in an MDB. The data particular to one pattern of a mask, such as the pattern-identification number and the current location of the glass plate, are kept in a PDB.

As explained in Section 3.1 of this paper, the engineer may specify particular features of a mask that he wants inspected. All the inspection information for one mask or pattern is kept in an inspection description block.

A description of the array of patterns to be used in making a mask is kept in a step-and-repeat-array description block. The placements of each pattern in terms of X and Y coordinates are given in micron dimensions. The pattern names may be given in general form according to the order of the PDB. In this way, one step-and-repeat-array description block can be used by many masks as mentioned in Section III of this paper.

7.2 *Job-Data Structure*

All the data for one job are linked together by a ring and linked list structure. The ring structure is used to unite the job, mask, and pattern description blocks as illustrated in Fig. 7. An inspection description block consists of a linked list of segments; the pointer to the first of these segments is placed in the description block of the mask or pattern described by the inspection data. The step-and-repeat-array description block is also a linked list of segments; the pointer to the first of these segments is placed in all MDBs using this array description.

With this data structure, data for one job may be scattered in random fashion over the disk and yet one pointer to any one of these segments can lead to all the data for the job. Using this arrangement, data can be easily added to or deleted from a job description. Furthermore, no data structure rules cause limitations to be placed on the number of masks in a job, the number of patterns in a mask, the

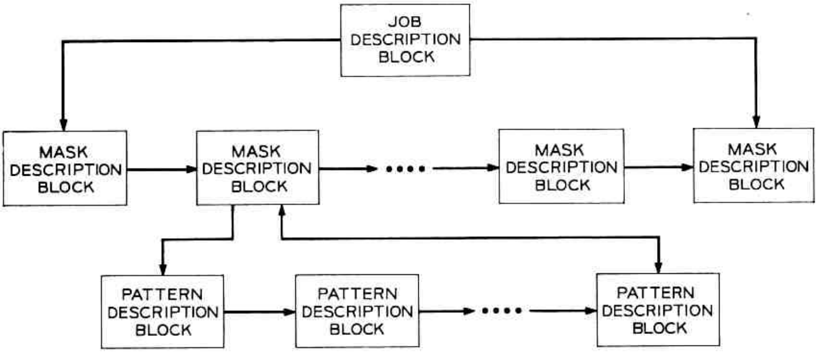


Fig. 7—Description block ring structure.

number of critical features to be inspected, or the complexity of an array description.

7.3 *System-Data Structure*

MSIS uses a table structure to connect the external world with its data. MSIS can retrieve all information about a particular job (if the request for the data includes the job number) by referencing a table which correlates the job number with the JDB pointer. Furthermore, the status and location of any glass plate in the shop, whether it be a piece of artwork, a reticle, a master mask, or a working copy can be obtained through a table which correlates plate identification numbers with the appropriate description block.

Also, the facility queues need contain only a one-word description block pointer for each entry on the queue because all the data needed when an operator or a facility requests a new assignment can be obtained with the use of that pointer.

Shop statistics are kept in status tables on permanent disk segments. The figures in these tables are continually updated by the application tasks. Thus when system statistics, such as the number of jobs in process, or the average time to get a job through the shop, are requested by a demand report, the answer is immediately available.

VIII. SIMULATION

Concurrent with the design of MSIS, a program which simulates the flow of jobs through the new mask-making facility was written using the IBM General Purpose System Simulator. The predictions made with this simulator were used to set upper limits on table lengths, to design the scheduling algorithm, and to design the MSIS—inspector interface.

With a fairly good knowledge of the processing times at each facility and the number and type of jobs that would pass through the shop on an average day, it was possible to set up a reasonably accurate simulation of the shop. However, two items were particularly hard to describe: the length of time required to inspect a plate and the rejection rate for inspected plates. Based on an earlier generation mask shop which existed at Allentown, some figures were obtained. However, two of the main purposes of the new shop were to provide more reliable making of plates and a better inspection facility. Thus these figures were considered worst-case values. An educated guess was used to establish a more likely set of numbers. Using both sets of inputs and varying the load on the shop, a large number of simulation runs were made. The results of the simulation for a load of 25 masks/shift are summarized in Table II.

8.1 *Table Lengths*

With a million-word disk, it was possible to be safe and allow extra space for table lengths. Nevertheless, some numbers were needed to decide just what "safe" meant. Here the simulation was invaluable. By having numbers for the maximum number of jobs in process and the total number of plate carriers used, it was possible to define an upper limit for the plate and carrier tables.

8.2 *Scheduling Algorithm*

The scheduling algorithm maintains a queue of masks that are awaiting processing by each of the facilities that make up the mask shop (e.g., PPG, reduction camera, etc.). One problem in designing the scheduling algorithm was the setting of a limit on the size of a facility queue. If a facility queue was allowed to be indefinitely long, the scheduling program would be cumbersome. Even if the queue lengths were set at a large number, the queues would have to be located on disk and the number of disk accesses involved in the continual queue manipulations would be too time-consuming. However, the possibility of a facility breakdown and a queue build-up prevented the placing of a tight restriction on queue length.

Thus it was decided to have a set of in-core facility queues that would be large enough for use during normal shop operation and queue

TABLE II—INSPECTION RESULTS

	Long Inspection Times—High Rejection Rates	Short Inspection Times—Low Rejection Rates
Shop through-put	22 masks/shift	24 masks/shift
Turn-around time (normal)	3 days	1½ days
Turn-around time (priority)	1½ days	1 day
Average length of a queue	8	5
Maximum number of jobs in process	56	40
Total number of carriers used	48	32
Average percent of time 8 inspectors were busy	85%	35%

extensions on disk to be used during abnormal operation. Using the results of the simulation, the number of in-core queue entries was established as ten; with this number, most queue manipulations can take place without disk accesses.

Other decisions that had to be made in designing the scheduling algorithm were whether a first-come, first-served system would be adequate, whether priority jobs should be allowed in shop operation, and, if so, what the number of priority levels should be. The simulation was used to test the possibilities. It was discovered that if the shop is keeping up with the input load (this occurs at a load of 25 masks/shift), most jobs can be completed in a couple days. Also, if two levels of priority are used for jobs entering the shop and the number of high-priority jobs is limited to 5 percent, a high-priority job can be completed in one day. Thus a simple two-level priority scheme is considered adequate, at least for a first version of MSIS.

8.3 *Inspection-Station Design*

For most of the shop facilities, the hardware defines the number of jobs that can be in process at one facility at one time and thus no facility limitation problems were encountered in the design of MSIS. However, the inspection facility is limited only by the number of inspectors and the capacity of the communication device. Considering the length of time required to complete one inspection process, it was realized that one teletypewriter could readily service five to ten inspectors. And, if necessary, another teletypewriter could easily be added to the MSIS hardware.

The number of inspectors allowed was critical to the design of the inspection control task and the allocation of core area for description blocks for masks in process at the inspection facility. Thus the simulation was invaluable again. The results showed that the number of inspectors could be limited to seven without impairing shop efficiency.

IX. MSIS FUTURE

At present, MSIS is controlling the making of masks at Murray Hill in a shop that is using one PPG and two reduction cameras to handle a small load of work. When the contact printing facility becomes available at Murray Hill, MSIS will also assign tasks in this area. The information system will then be installed in the Allentown shop; at this time, the communication with the step-and-repeat camera will be included. Before the end of the year, MSIS will be controlling the operation of both the Murray Hill and Allentown shops.

It is too early to know what problems will be encountered during a long period of shop operation with MSIS. Realistically, it must be assumed that even with the simulation, some unexpected demands on the system will turn up and some additions and changes will be required. However, the monitor is sufficiently general that it is doubtful that it will undergo any major change. And the method of communication between tasks, the description block data arrangement, the handling of queues, the plate-carrier assignments, and the allocation of disk areas are basic enough to remain permanent. Since the coding for these functions has been kept in separate tasks from those dealing directly with the outside world, these tasks can be kept intact even when major shop changes take place. Thus a change in shop operation will probably lead to the rewriting of a task or two, and it will be possible to add the new tasks without causing havoc to the rest of the system.

This modular arrangement of system functions will also work out well when MSIS expands. A study is now underway to see how the adding of a data link between the coordinate measuring machine in the inspection area and the MSIS computer will improve shop operation. It is believed that with a new inspection task added to the MSIS task list and a small addition to the interrupt handler, this improved inspection capability can readily be added to MSIS.

REFERENCES

1. Poole, K. M., et. al., "The Primary Pattern Generator," B.S.T.J., this issue, pp. 2031–2076.
2. Rawson, E. G., Poulsen, M. E., and Stafford, J. W., "Reduction Cameras," B.S.T.J., this issue, pp. 2117–2143.
3. Alles, D. S., et al., "The Step-and-Repeat Camera," B.S.T.J., this issue, pp. 2145–2177.

Response of Periodically Varying Systems to Shot Noise—Application to Switched RC Circuits

By S. O. RICE

(Manuscript received June 16, 1970)

This paper is concerned with the statistical properties of the output $y(t)$ of a periodically varying linear system when the input is random shot noise.

Usually $y(t)$ can be divided into a noise part, $y_N(t)$, and a periodic part, $y_{\rm per}(t)$. Expressions are obtained for the Fourier components of $y_{\rm per}(t)$ and the power spectrum of $y_N(t)$. Various averages associated with $y(t)$ are studied. Some of the results for shot noise input can be converted into corresponding results for white noise input.

Some of the theoretical results are illustrated by applying them to two examples. In both examples the system consists of an arrangement of a resistance, a condenser, and a switch which opens and closes periodically. The output is the voltage across the condenser.

I. INTRODUCTION

Consider a circuit, shown in Fig. 1a, consisting of a resistance R shunted by a switch and condenser C. The circuit is driven by a Poisson shot noise current. The elementary charges q arrive at random at an average rate of ν per second. The switch operates in a cycle with period T. It is closed during the intervals $nT < t < nT + \alpha T$ and open during the intervals $nT + \alpha T < t < (n + 1)T$ where n is an integer and $0 < \alpha \leq 1$. We are interested in the statistical properties of the voltage $V(t)$ across the condenser. In particular, we want an expression for the two-sided power spectrum $W_V(f)$ of $V(t)$.

This problem was encountered by D. D. Sell[1] during the development of a new type of spectrophotometer. The determination of an exact expression for $W_V(f)$ turned out to be unexpectedly difficult, and led to the present investigation of the more general case in which

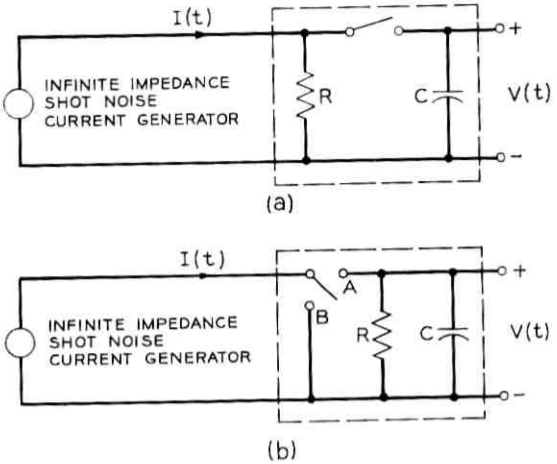


Fig. 1—RC circuits with periodically operating switch.

the switched RC circuit is replaced by a general linear network which varies periodically with time.

The systems shown in Fig. 1 are "cyclo-stationary" (this term was introduced by W. R. Bennett). Cyclo-stationary systems have been studied by a number of writers. A detailed treatment and many references are given by H. L. Hurd[2] in his thesis on periodically correlated stochastic processes. However, I have been unable to find any references dealing specifically with periodically varying systems having shot noise input. The nearest approach is contained in seven pages of anonymous handwritten notes[3] obtained by Sell. These notes give approximate results for the case of Fig. 1a with white-noise input instead of shot noise input.

In Section II, we make some general remarks about the notation and type of analysis used in this paper. Section III contains a statement of results for the general system shown in Fig. 2. In Sections IV and V, the general results are applied to the RC circuits shown in Figs. 1a and 1b. Representative curves giving $W_V(f)$ for various values of the circuit parameters in Fig. 1a are plotted in Fig. 3. Sections VI, VII, and VIII contain the derivation of the expressions stated in Section II for the various ensemble averages and the output power spectrum. The results for shot noise input can be carried over into corresponding results for white gaussian noise input. This correspondence is developed in Section IX. Appendix A gives an outline of the analysis required in applying the general theory to get the power spectrum $W_V(f)$ of the output $V(t)$ in the RC circuit of Fig. 1a.

Roughly speaking, the shot effect formulas for a periodically varying system differ from the shot-effect formulas for a time invariant system[4] by containing an additional integration. This extra integral represents an average taken over the period.

II. REMARKS CONCERNING NOTATION AND ANALYSIS

In this paper ensemble averages are denoted by the angle bracket $\langle\ \rangle$ and time averages by over-bars. For example, consider $V(t)$ in Fig. 1. We can write $V(t) \equiv V(t, \varphi)$ where φ represents the family of random arrival times of the charges q comprising the shot noise current. When t is held fixed, $V(t)$ can be regarded as a random variable and $\langle V^l(t)\rangle$ as the average value of the lth power of $V(t)$ at time t. On the other hand, for a fixed set φ of arrival times, i.e., for a particular member of the ensemble, the time average of $V^l(t)$ is denoted by

$$\overline{V^l(t)} = \operatorname*{limit}_{T_1 \to \infty} \frac{1}{T_1} \int_0^{T_1} V^l(t)\, dt. \tag{1}$$

Let $z(t)$ be an output function (e.g., $V^l(t)$) of our periodic system such that its ensemble average $\langle z(t)\rangle$ is periodic with period T, the period of the system. We assume that the time average $\overline{z(t)}$ has the same value for almost all members of the ensemble. From this assumption and the periodicity of $\langle z(t)\rangle$ it follows, upon averaging both sides of the equation

$$\overline{z(t)} = \operatorname*{limit}_{T_1 \to \infty} \frac{1}{T_1} \int_0^{T_1} z(t)\, dt$$

over the ensemble, that

$$\overline{z(t)} = \frac{1}{T} \int_0^{T} \langle z(t)\rangle\, dt. \tag{2}$$

In addition to ensemble and time averages, we shall use E to denote expected values of time invariant random variables associated with the amplitudes of the shot noise impulses.

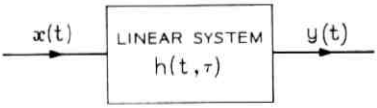


Fig. 2—Time-varying linear system specified by $h(t, \tau)$.

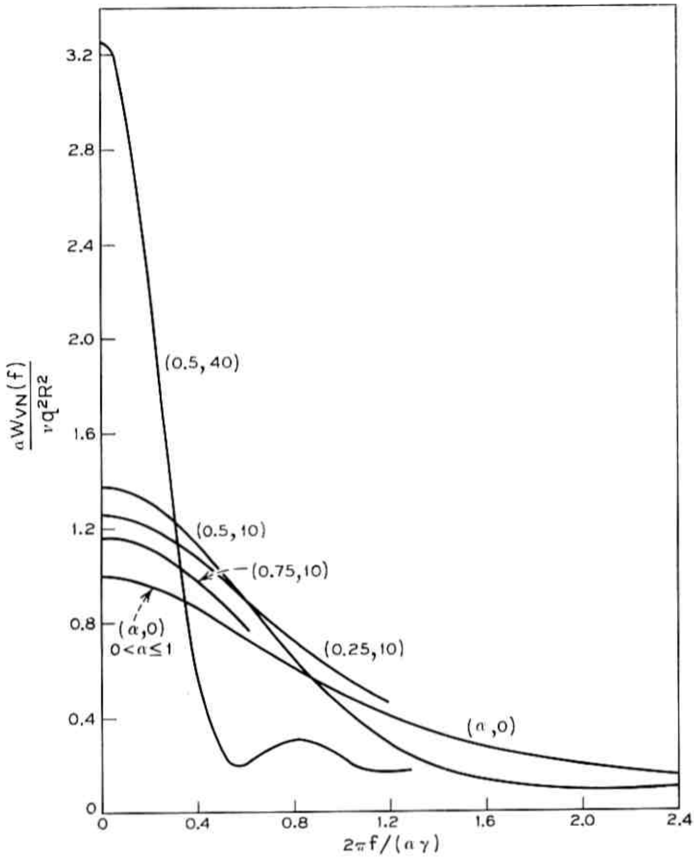


Fig. 3—Power spectrum of $V(t)$ in Fig. 1a.

$W_{V_N}(f)$ = 2-sided power spectrum of $V(t)$ minus DC spike due to $V_{dc} = \nu qR$.
$I(t) = \sum_k q\delta(t - t_k)$, ν = Arrival Rate, $\gamma = 1/(RC)$;
α = Fraction of time switch is closed;
T = Length of switch cycle;
$(\alpha, \gamma T)$ = Curve parameters.

We use the term "periodic" to mean "singly periodic." The more difficult case of "multiply periodic" variation is not considered. An example of the latter is given by the circuit of Fig. 1a in which the switch is operated by the function $f(t) = P \cos pt + Q \cos qt$, p and q being incommensurable. The switch is closed when $f(t) > 0$, and is open when $f(t) < 0$. Possibly such cases could be handled by the method used by Bennett[5] to obtain the output of a rectifier when $P \cos pt + Q \cos qt$ is applied.

The (two-sided) power spectrum $W_V(f)$ [where $W_V(-f) = W_V(f)$] can be interpreted physically as follows. Let $V(t)$ be applied to an ideal filter which passes only the narrow band $f_1 < |f| < f_1 + \Delta f$, and let the filter be terminated in a resistance of one ohm. Then

$$[W_V(-f_1) + W_V(f_1)]\,\Delta f = 2W_V(f_1)\,\Delta f$$

is the time average of the power which would be dissipated in the one ohm resistance. The average must be taken over an interval long in comparison with $1/\Delta f$.

The analysis used here makes no attempt at mathematical rigor. Orders of summation and integration are interchanged freely, and assumptions are made which are physically plausible but which may be difficult to express in precise mathematical terms.

III. STATEMENT OF RESULTS FOR GENERAL SYSTEM

The results given in this section pertain to the general system shown in Fig. 2. The system is linear and is specified by the response $y(t) = h(t, \tau)$ to a unit impulse $x(t) = \delta(t - \tau)$ applied at time τ. The system varies periodically with period T so that

$$h(t + nT, \tau + nT) = h(t, \tau) \qquad n = \text{integer}. \tag{3}$$

In most of our work, the input $x(t)$ is the shot noise

$$x(t) = \sum_{k=-\infty}^{\infty} a_k \delta(t - t_k) \tag{4}$$

where the random "arrival times" t_k occur at an average rate of ν/second and constitute a Poisson process. The impulse amplitudes a_k are independent random variables with

$$E(a_k) = E(a), \qquad E(a_k^2) = E(a^2). \tag{5}$$

Since the system is linear, the output corresponding to equation (4) is

$$y(t) = \sum_{k=-\infty}^{\infty} a_k h(t, t_k). \tag{6}$$

The function $h(t, \tau)$ is assumed to be such that the steps in the analysis are legitimate. In particular, it is assumed that when $0 \leq \tau \leq T$ and $|t| \to \infty$, $|h(t, \tau)|$ tends to 0 with sufficient rapidity to (*i*) make the various integrals converge, and (*ii*) ensure that the times at which a long interval of operation begins and stops have no appreciable influence on the output during the major portion of the interval.

In Section VIII it is shown that $y(t)$ is the sum of a noise component $y_N(t)$ and a periodic (including dc) component $y_{\text{per}}(t)$:

$$y(t) = y_N(t) + y_{\text{per}}(t). \tag{7}$$

The power spectrum of $y_N(t)$ is

$$W_{\nu N}(f) = \frac{\nu E(a^2)}{T} \int_0^T |\, s(f, \tau) \,|^2 \, d\tau \tag{8}$$

where

$$s(f, \tau) = \int_{-\infty}^{\infty} e^{-i\omega t} h(t, \tau) \, dt, \qquad \omega = 2\pi f. \tag{9}$$

The periodic component of $y(t)$ is

$$\begin{aligned} y_{\text{per}}(t) &= \nu E(a) \sum_{m=-\infty}^{\infty} s_0(m/T) e^{i2\pi mt/T}, \\ &= \nu E(a) s_0(0) + 2\nu E(a) \text{ Real} \sum_{m=1}^{\infty} s_0(m/T) e^{i2\pi mt/T}, \end{aligned} \tag{10}$$

where

$$s_0(f) = \frac{1}{T} \int_0^T s(f, \tau) \, d\tau. \tag{11}$$

The dc part of $y(t)$ is given by the constant term in equation (10):

$$y_{\text{dc}} = \nu E(a) s_0(0). \tag{12}$$

Note that $y_{\text{per}}\ (t)$ is zero when $E(a)$ is zero.

The ensemble average $\langle y^l(t) \rangle$, which gives the lth moment of the distribution of the ensemble of $y(t)$'s at time t, is a periodic function of t of period T. For $l = 1$ and $l = 2$

$$\langle y(t) \rangle = \nu E(a) \sum_{n=-\infty}^{\infty} \int_0^T h(t + nT, \tau) \, d\tau, \tag{13}$$

$$\langle y^2(t) \rangle - \langle y(t) \rangle^2 = \nu E(a^2) \sum_{n=-\infty}^{\infty} \int_0^T h^2(t + nT, \tau) \, d\tau. \tag{14}$$

These equations give the first and second cumulants of the distribution of the $y(t)$'s. The lth cumulant at time t is

$$\kappa_l(t) = \nu E(a^l) \sum_{n=-\infty}^{\infty} \int_0^T h^l(t + nT, \tau) \, d\tau. \tag{15}$$

The periodic and noise components of $y(t)$ are related to the ensemble averages by

$$y_{\text{per}}(t) = \langle y(t) \rangle = \kappa_1(t), \tag{16}$$

$$\langle y_N^2(t) \rangle = \langle y^2(t) \rangle - \langle y(t) \rangle^2 = \kappa_2(t). \tag{17}$$

The mean square value of $y_N^2(t)$, averaged over time, may be expressed in several ways:

$$\begin{aligned} \overline{y_N^2(t)} &= \frac{1}{T} \int_0^T \langle y_N^2(t) \rangle \, dt = \frac{1}{T} \int_0^T \kappa_2(t) \, dt, \\ \overline{y_N^2(t)} &= \int_{-\infty}^{\infty} W_{VN}(f) \, df, \\ &= \frac{\nu E(a^2)}{T} \int_0^T d\tau \int_{-\infty}^{\infty} df \mid s(f, \tau) \mid^2, \\ &= \frac{\nu E(a^2)}{T} \int_0^T d\tau \int_{-\infty}^{\infty} dt \, h^2(t, \tau). \end{aligned} \tag{18}$$

All of the foregoing results pertain to the case in which the input $x(t)$ is the shot noise (4). Now let the input be zero-mean white gaussian noise with the power spectrum

$$W_x(f) = \begin{cases} N_0, & \mid f \mid < F; \\ 0, & \mid f \mid > F; \end{cases} \tag{19}$$

where $F \rightarrow \infty$. It is shown in Section IX that results for this input can be obtained from the preceding shot noise formulas by taking $a_k = \pm (N_0/\nu)^{\frac{1}{2}}$ with equal probability and letting $\nu \rightarrow \infty$. Then

$$\nu E(a) \rightarrow 0, \quad \nu E(a^2) \rightarrow N_0, \quad \text{and} \quad \nu E(a^l) \rightarrow 0 \quad \text{for} \quad l > 2. \tag{20}$$

Therefore $y_{\text{per}}(t) = \langle y(t) \rangle = 0$, and consequently $y(t)$ consists entirely of the noise component $y_N(t)$. Expressions for the output power spectrum $W_y(f)$ and the mean square values $\langle y^2(t) \rangle$, $\overline{y^2(t)}$ are obtained by replacing $\nu E(a^2)$ by N_0 in equations (8), (14), and (18):

$$\begin{aligned} W_y(f) &= \frac{N_0}{T} \int_0^T \mid s(f, \tau) \mid^2 d\tau, \\ \langle y^2(t) \rangle &= N_0 \sum_{n=-\infty}^{\infty} \int_0^T h^2(t + nT, \tau) \, d\tau, \\ \overline{y^2(t)} &= \frac{N_0}{T} \int_0^T d\tau \int_{-\infty}^{\infty} df \mid s(f, \tau) \mid^2, \\ &= \frac{N_0}{T} \int_0^T d\tau \int_{-\infty}^{\infty} dt \, h^2(t, \tau). \end{aligned} \tag{21}$$

In these expressions, $s(f, \tau)$ is still the Fourier transform (9) of $h(t, \tau)$. Equations (15) and (20) show that all of the cumulants except the second are zero. Therefore the ensemble of $y(t)$'s at time t is normally distributed about 0 with variance $\langle y^2(t)\rangle$ given by equation (21). The probability that $y(t)$ will lie between Y and $Y + dY$ at a time t picked at random is given by expression (113) in Section IX.

IV. RC CIRCUIT OF FIGURE 1a

In this section the results stated in Section III for general systems will be applied to the RC circuit shown in Fig. 1a. In this case the input $x(t)$ is the input $I(t)$ from the shot-noise current generator,

$$I(t) = \sum_{k=-\infty}^{\infty} q\delta(t - t_k) \tag{22}$$

where the individual charges (of q coulombs) arrive at an average rate of ν per second.

Comparison with the series (4) for $x(t)$ shows that $a_k = q$ and

$$E(a) = q, \qquad E(a^2) = q^2. \tag{23}$$

The output $V(t)$ is constant for intervals of length $(1 - \alpha)T$ while the switch is open. When the switch is closed $V(t)$ drifts either up or down, depending upon whether the input current is temporarily greater than or less than the leakage through R. The average value of $V(t)$ is $V_{dc} = \nu qR$ where νq is the average current flowing through R. It turns out that the mean square value of $V(t) - V_{dc}$ is $qV_{dc}/(2C)$. Furthermore, the circuit of Fig. 1a is unusual in that the distribution of the ensemble of $V(t)$'s at time t does not vary periodically with t.

Some insight into the behavior of the system can be obtained by considering the case when $T/\mathrm{RC} \ll 1$. If the switch were closed all of the time ($\alpha = 1$), the usual shot effect formulas would hold and the two-sided power spectrum of $V(t)$ would be

$$\begin{aligned} W_V(f) &= \nu \left| \int_{-\infty}^{\infty} e^{-i\omega t} F(t)\, dt \right|^2 + V_{dc}^2\, \delta(f), \\ &= \frac{\nu q^2 R^2}{1 + (\omega\, \mathrm{RC})^2} + V_{dc}^2\, \delta(f), \qquad \omega = 2\pi f, \end{aligned}$$

where $F(t)$ is the $V(t)$ due to a charge q arriving at time 0; $F(t) = (q/C) \exp(-t/\mathrm{RC})$ for $t > 0$, and $F(t) = 0$ for $t < 0$. The first term in $W_V(f)$ is $W_{V_N}(f)$, the power spectrum of the noise component $V_N(t) =$

$V(t) - V_{dc}$, and the second term is the spike due to V_{dc}. Now, instead of $\alpha = 1$ let α be anywhere in $0 < \alpha < 1$, but take $T/\mathrm{RC} \ll 1$. The cycles are so brief that $V(t)$ does not change much during one cycle; and the situation is much like that for $\alpha = 1$ except that, in effect, ν is reduced to $\nu\alpha$, and $F(t)$ becomes $(q/C)\exp(-t\alpha/\mathrm{RC})$ because the condenser current flows only the fraction α of the time. Replacing ν by $\nu\alpha$ and $F(t)$ by its new expression leads to

$$W_{V_N}(f) \approx \frac{\nu q^2 R^2/\alpha}{1 + (\omega\ \mathrm{RC}/\alpha)^2}.$$

When T/RC is not small, the expressions for the power spectrum become much more complicated. We now turn to the general case in which T/RC and α are unrestricted except for $0 < \alpha < 1$.

The first step is to determine the response (the condenser voltage) $h(t, \tau)$ at time t to a *unit* impulse of current arriving at time τ where $0 < \tau < T$. When $\alpha T < \tau < T$, the impulse arrives when the switch is open, no charge reaches the condenser, no voltage appears across the condenser, and hence

$$h(t, \tau) \equiv 0 \quad \text{for all} \quad t \quad \text{when} \quad \alpha T < \tau < T. \tag{24}$$

When $0 < \tau < \alpha T$ the switch is closed, and the unit impulse of current arriving at time τ deposits a unit charge on the condenser. This charges the condenser to the voltage $1/C$. The voltage decreases exponentially as the charge leaks off through R until the switch opens at time αT. The voltage remains constant throughout the interval $\alpha T < t < T$ during which the switch is open. It resumes its exponential decay during $T < t < T + \alpha T$, remains constant during $T + \alpha T < t < 2T$, and so on. Hence when $0 < \tau < \alpha T$ the values of $h(t, \tau)$ are

$$\begin{aligned}
&0, && -\infty < \mathrm{t} < \tau;\\
&C^{-1}\exp[-\gamma(t - \tau)], && \tau < t < \alpha T;\\
&C^{-1}\exp[-\gamma(n\alpha T - \tau)], && (n-1)T + \alpha T < t < nT;\\
&C^{-1}\exp[-\gamma(n\alpha T - \tau) - \gamma(t - nT)], && nT < t < nT + \alpha T;
\end{aligned} \tag{25}$$

where $n = 1, 2, 3, \cdots$ and

$$\gamma = 1/(\mathrm{RC}). \tag{26}$$

Equation (6) for the output $y(t)$ becomes

$$V(t) = \sum_{k=-\infty}^{\infty} qh(t, t_k) \tag{27}$$

where $h(t, t_k)$ can be obtained from the relation $h(t + nT, \tau + nT) = h(t, \tau)$ and the values (24) and (25). From equation (9), the Fourier transform $s(f, \tau)$ of $h(t, \tau)$ is 0 when $\alpha T < \tau < T$ because, from (24), $h(t, \tau)$ is 0 in the same interval:

$$s(f, \tau) = 0, \qquad \alpha T < \tau < T. \tag{28}$$

For $0 < \tau < \alpha T$ we have, from equations (9) and (25),

$$\begin{aligned} s(f, \tau) &= \int_{-\infty}^{\infty} e^{-i\omega t} h(t, \tau)\, dt, \qquad \omega = 2\pi f; \\ &= \int_{\tau}^{\alpha T} C^{-1} \exp\left[-\gamma(t - \tau) - i\omega t\right] dt \\ &\quad + \sum_{n=1}^{\infty} C^{-1} \exp\left[-\gamma(n\alpha T - \tau)\right] \\ &\quad \times \left(\int_{(n-1)T+\alpha T}^{nT} e^{-i\omega t}\, dt + \int_{nT}^{nT+\alpha T} e^{-\gamma(t-nT)-i\omega t}\, dt \right). \end{aligned} \tag{29}$$

When the integrations are performed, the series summed, and the notation

$$z = e^{-i\omega T}, \qquad b = e^{-\gamma\alpha T} \tag{30}$$

introduced, some algebra carries equation (29) into

$$s(f, \tau) = \frac{C^{-1}e^{-i\omega\tau}}{\gamma + i\omega} + \frac{C^{-1}be^{\gamma\tau}}{1 - bz}(z - z^{\alpha})\left(\frac{1}{\gamma + i\omega} - \frac{1}{i\omega}\right) \tag{31}$$

for $0 < \tau < \alpha T$.

The integral (11) for $s_0(f)$ becomes

$$\begin{aligned} s_0(f) &= \frac{1}{T}\int_0^T s(f, \tau)\, d\tau, \qquad \omega = 2\pi f; \\ &= \frac{1}{T}\int_0^{\alpha T} s(f, \tau)\, d\tau. \end{aligned} \tag{32}$$

The function $s_0(f)$ is used solely to compute the periodic portion of the output, and therefore only the values of $s_0(m/T)$, where m is an integer, are of interest. For $f = m/T$ the value of $\omega = 2\pi f$ is $\omega = 2\pi m/T$, and $\omega T = 2\pi m$. Evaluation of the integral (32) for $s_0(f)$ leads to

$$\begin{aligned} s_0(0) &= 1/(C\gamma) = R, \\ s_0(m/T) &= 0, \qquad m \neq 0. \end{aligned} \tag{33}$$

As in equation (7), the output $V(t)$ can be expressed as the sum of a noise component and a periodic component,

$$V(t) = V_N(t) + V_{\text{per}}(t). \tag{34}$$

Since $s_0(m/T)$ is zero for $m \neq 0$, equation (10) shows that for Fig. 1a, the periodic component consists only of the dc component:

$$V_{\text{per}}(t) = V_{\text{dc}} = \nu E(a) s_0(0) = \nu q R. \tag{35}$$

The quantity νq is the average shot noise current (in amperes if q is measured in coulombs) flowing through R; and V_{dc} is the average IR drop across the resistance.

The value $\nu q R$ for $V_{\text{per}}(t)$ can also be obtained from equations (16) and (13),

$$\begin{aligned} V_{\text{per}}(t) &= \langle V(t) \rangle = \kappa_1(t), \\ &= \nu E(a) \int_0^T d\tau \sum_{n=-\infty}^{\infty} h(t + nT, \tau), \\ &= \nu E(a) C^{-1}/\gamma = \nu q C^{-1}/\gamma = \nu q R, \end{aligned} \tag{36}$$

where the expressions (24) and (25) for $h(t, \tau)$ are used in summing the series and evaluating the integral.

The values of the higher order cumulants $\kappa_l(t)$ follow almost immediately from equation (36). First observe that the expression (15) for $\kappa_l(t)$ can be obtained from the expression (13) for $\langle y(t) \rangle$ $(= \kappa_1(t))$ by replacing $E(a)$, $h(t + nT, \tau)$ by $E(a^l)$, $h^l(t + nT, \tau)$, respectively. Furthermore, $h^l(t + nT, \tau)$ can be obtained from $h(t + nT, \tau)$ by replacing C^{-1} and γ by C^{-l} and $l\gamma$, respectively. Therefore from equation (36),

$$\kappa_l(t) = \nu E(a^l) C^{-l}/(l\gamma) = \nu q^l C^{-l}/(l\gamma). \tag{37}$$

In particular, the variance of the distribution of the ensemble of $V(t)$'s at time t is

$$\langle V^2(t) \rangle - \langle V(t) \rangle^2 = \kappa_2(t) = \nu q^2 C^{-2}/(2\gamma) = \frac{q}{2C} V_{\text{dc}}. \tag{38}$$

The fact that this does not depend on t shows that the mean square value of the fluctuation about V_{dc} is also $qV_{\text{dc}}/(2C)$:

$$\overline{(V(t) - V_{\text{dc}})^2} = \overline{V_N^2(t)} = \frac{q}{2C} V_{dc} = \frac{\nu q^2 R}{2C}. \tag{39}$$

Equation (31) leads to the equation

$$\begin{aligned}\ln \varphi(z) &= \sum_{l=1}^{\infty} \kappa_l(t)(iz)^l/l!,\\ &= \frac{\nu}{\gamma}\sum_{l=1}^{\infty}\frac{(izq/C)^l}{l!\,l}, \\ &= \frac{\nu}{\gamma}\int_0^{zq/C}(e^{i\theta}-1)\,d\theta/\theta,\end{aligned} \tag{40}$$

for the characteristic function $\varphi(z)$ of the distribution of the ensemble of $V(t)$'s at time t. The probability density of the distribution is given by

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\varphi(z)e^{-iVz}\,dz \tag{41}$$

where $\ln \varphi(z)$ can be expressed in terms of sine and cosine integrals. The integral (41) also gives the probability density of the value of a particular member of the ensemble at a time selected at random.

The power spectrum $W_{V_N}(f)$ of $V_N(t) = V(t) - V_{\text{dc}}$ is obtained by substituting the value (31) of $s(f, \tau)$ in (8).

$$\begin{aligned}W_{V_N}(f) &= \frac{\nu q^2}{T}\int_0^{\alpha T}|s(f,\tau)|^2\,d\tau,\\ &= \frac{\nu q^2R^2}{1+(\omega/\gamma)^2}\,\text{Real}\left[\alpha + \frac{(1-bz^{\alpha})(1-z^{1-\alpha})\gamma(\gamma-i\omega)}{T(1-bz)\omega^2(\gamma+i\omega)}\right],\end{aligned} \tag{42}$$

where $\omega = 2\pi f$, $\gamma = 1/\text{RC}$, and z and b are given by (30):

$$z = e^{-i\omega T}, \qquad b = e^{-\gamma\alpha T}.$$

An outline of the evaluation of the integral is given in Appendix A. The curves plotted in Fig. 3 were computed from equation (42). It can be shown that

$$\begin{aligned}&W_{V_N}(0) = \nu q^2R^2[2-\alpha+\tfrac{1}{2}\gamma T(1-\alpha)^2(1+b)(1-b)^{-1}];\\ &W_{V_N}(f) \to \nu q^2R^2\gamma^2\alpha/\omega^2, \quad \text{as} \quad f \to \infty;\\ &W_{V_N}(f) \to \frac{\nu q^2R^2/\alpha}{1+[\omega/(\gamma\alpha)]^2}, \quad \text{as} \quad T \to 0.\end{aligned} \tag{43}$$

In Fig. 3, the quantity $\alpha W_{V_N}(f)/(\nu q^2R^2)$ is plotted as a function of $\omega/(\gamma\alpha) = \omega\ \text{RC}/\alpha = 2\pi f\ \text{RC}/\alpha$. The parameters are α and $\gamma T = T/(\text{RC})$. These coordinates were chosen because the exact computations made from equation (42) give nearly the same values as does the last approximation ($T \to 0$) in (43) for values of γT less than, say, 5.0. From

$$\int_0^\infty W_{V_N}(f)\,df = \tfrac{1}{2}\overline{V^2(t)} = \frac{1}{2}\frac{\nu q^2 R}{2C}$$

it can be shown that the area under any curve in Fig. 3 is $\pi/2$. As $\gamma T \to \infty$, the ordinate at $f = 0$ ultimately increases as

$$\alpha(2-\alpha) + \frac{\alpha\gamma T}{2}(1-\alpha)^2$$

which, for γT fixed, has a maximum at

$$\alpha = \frac{1}{3} + \frac{4}{3\gamma T}.$$

The oscillations in the curves for the large values of γT can be correlated with the oscillations in a $(\sin f/f)^2$ type of spectrum associated with the flat portions of length $(1-\alpha)T$ in $V(t)$.

When the shot noise current generator in Fig. 1a is replaced by a zero-mean white noise current generator with a flat, two-sided power spectrum $W_I(f) = N_{I0}$, the dc component of $V(t)$ becomes 0 and $V(t)$ is distributed normally about zero with variance

$$\langle V^2(t)\rangle = \overline{V^2(t)} = N_{I0}/(2\gamma C^2). \tag{44}$$

This $V(t)$ is an example of a stochastic process in which the distribution of the ensemble at time t is normal and does not change with t, but the process is still not a stationary gaussian process because $dV(t)/dt$ is zero during the intervals that the switch is open.

The power spectrum $W_V(f)$ is given by equations (42) and (43) with the multiplier νq^2 replaced by N_{I0}. In the particular case in which the period T is small compared to the time constant RC, the last approximation given in equation (43) goes into

$$W_V(f) \to \frac{N_{I0}R^2/\alpha}{1 + (\omega\,\mathrm{RC}/\alpha)^2}. \tag{45}$$

The Princeton Applied Research notes[3] obtained by Sell give results associated with this approximation.

By Thevenin's theorem, the portion of Fig. 1a consisting of the infinite impedance shot noise current generator plus the resistance R shunting the generator can be replaced by a zero impedance shot noise voltage generator in series with R. The currents and voltages in the remaining portion of the circuit are unchanged by this replacement. The voltage of the new generator is $V_g(t) = I(t)R$; and its two-sided power spectrum $W_{V_g}(f)$ is flat and equal to $N_{V0} =$

$N_{I0}R^2$. The statistical results for the voltage $V(t)$ across C can be expressed in terms of N_{V0} by replacing N_{I0} by N_{V0}/R^2 (that is, νq^2 by N_{V0}/R^2). For example, equation (44) becomes

$$\overline{V^2}(t) = N_{V0}/(R^2 2\gamma C^2) = N_{V0}/(2RC). \tag{46}$$

V. RC CIRCUIT OF FIGURE 1*b*

The input shot noise current $I(t)$ in Fig. 1b is the same as in Fig. 1a, and is given by the sum (22) of impulses of weight q. The switch is in position a during the first part of the cycle, $nT < t < nT + \alpha T$; and in position b during the second part, $nT + \alpha T < t < (N + 1)T$.

The condenser voltage $V(t)$ increases more often than not during the first part of the cycle. It always decreases during the second part. Unlike the circuit Fig. 1a, $V(t)$ has a periodic portion $V_{\text{per}}(t)$ which includes variable terms in addition to V_{dc}.

Just as in Fig. 1a, we have $a_k = q$ and $E(a) = q$, $E(a^2) = q^2$ .The response $h(t, \tau)$ at time t to a unit impulse of current arriving at τ, where $0 < \tau < T$, is

$$\begin{gathered} 0 \quad \text{for} \quad -\infty < t < \tau, \\ C^{-1} \exp\left[-\gamma(t - \tau)\right] \quad \text{for} \quad 0 < \tau < \alpha T \quad \text{and} \quad \tau < t, \\ 0 \quad \text{for} \quad \alpha T < \tau < T \quad \text{and all} \quad t. \end{gathered} \tag{47}$$

As before, $\gamma = 1/(\text{RC})$ and

$$V(t) = \sum_{k=-\infty}^{\infty} qh(t, t_k). \tag{48}$$

The Fourier transform of $h(t, \tau)$ is

$$\begin{aligned} s(f, \tau) &= \int_\tau^\infty e^{-i\omega t} C^{-1} e^{-\gamma(t-\tau)}\, dt, \\ &= \frac{C^{-1} e^{-i\omega\tau}}{\gamma + i\omega}, \qquad \omega = 2\pi f; \end{aligned} \tag{49}$$

for $0 < \tau < \alpha T$, and $s(f, \tau) = 0$ for $\alpha T < \tau < T$. The integral $s_0(f)$ used in computing $V_{\text{per}}(t)$ in $V(t) = V_N(t) + V_{\text{per}}(t)$ is

$$\begin{aligned} s_0(f) &= \frac{1}{T} \int_0^T s(f, \tau)\, d\tau, \\ &= (1 - e^{-i\omega\alpha T})/[i\omega TC(\gamma + i\omega)], \\ s_0(0) &= \alpha/C\gamma = \alpha R. \end{aligned} \tag{50}$$

The dc portion of $V_{\text{per}}(t)$ is

$$V_{\text{dc}} = \nu E(a)s_0(0) = \nu q\alpha R \tag{51}$$

and, from the general expression (10) for $V_{\text{per}}(t)$,

$$V_{\text{per}}(t) = V_{\text{dc}} + 2 \text{ Real} \sum_{m=1}^{\infty} \left[\frac{V_{\text{dc}}}{1 + i(\omega/\gamma)} \left(\frac{1 - e^{-i\omega\alpha T}}{i\omega\alpha T}\right) e^{i\omega t}\right]_{\omega=2\pi m/T}. \tag{52}$$

By working with

$$V_{\text{per}}(t) = \langle V(t) \rangle = \nu q \sum_{n=-\infty}^{\infty} \int_0^T h(t + nT, \tau)\, d\tau \tag{53}$$

it can be shown that $V_{\text{per}}(t)$ increases from A exp $(-\gamma T)$ at $t = 0$ to A exp $(-\gamma\alpha T)$ at $t = \alpha T$, and then decreases to A exp $(-\gamma T)$ at $t = T$ and so on, where

$$A = \frac{V_{\text{dc}}}{\alpha} \frac{e^{\gamma\alpha T} - 1}{1 - e^{-\gamma T}}. \tag{54}$$

The power spectrum $W_{VN}(f)$ of the noise portion $V_N(t)$ of $V(t)$ is given by equation (8) and the expression (49) for $s(f, \tau)$.

$$\begin{aligned} W_{VN}(f) &= \frac{\nu E(a^2)}{T} \int_0^T |s(f, \tau)|^2\, d\tau, \\ &= \frac{\nu q^2 C^{-2}}{T} \int_0^{\alpha T} \frac{d\tau}{\gamma^2 + \omega^2}, \qquad \omega = 2\pi f; \\ &= \frac{\nu q^2 C^{-2} \alpha}{\gamma^2 + \omega^2} = RqV_{\text{dc}}/[1 + (\omega \text{ RC})^2]. \end{aligned} \tag{55}$$

Integrating $W_{VN}(f)$ from $f = -\infty$ to $f = +\infty$ shows that the time average of $V_N^2(t)$ is

$$\overline{V_N^2(t)} = \frac{q}{2C} V_{\text{dc}} \tag{56}$$

just as in the case of Fig. 1a [see equation (39)]. However, in Fig. 1a, $V_{\text{dc}} = \nu qR$; whereas in Fig. 1b, $V_{\text{dc}} = \nu q\alpha R$.

When the shot noise current generator in Fig. 1b is replaced by a zero-mean white noise current generator with flat power spectrum $W_I(f) = N_{I0}$, the periodic component $V_{\text{per}}(t)$ vanishes and the power spectrum of $V(t)$ is obtained by replacing νq^2 in equation (55) by N_{I0}:

$$W_V(f) = W_{VN}(f) = N_{I0} \frac{C^{-2}\alpha}{\gamma^2 + \omega^2}, \qquad \omega = 2\pi f. \tag{57}$$

The time average of $V^2(t)$ obtained by integrating equation (57) is

$$\overline{V^2(t)} = N_{I0} \frac{\alpha R}{2C}. \tag{58}$$

Although the periodic component $V_{\text{per}}(t) = \langle V(t) \rangle$ is zero, the ensemble variance $\langle V^2(t) \rangle$ at time t is a periodic function of t. It may be calculated from the second of equations (21) in which $h^2(t, \tau)$ is obtained by squaring the expressions (47) for $h(t, \tau)$.

VI. THE ENSEMBLE AVERAGE $\langle y(t) \rangle$

In this section and the two following ones, the arguments used to deal with shot noise will be used to determine the power spectrum and the moments (more precisely, the cumulants) of the distribution of the output $y(t)$ of the periodically varying system shown in Fig. 2. The input $x(t)$ is taken to be shot noise consisting of a train of randomly arriving impulses.

Let the system of Fig. 2 start operating at time $t = 0$ and run to $t = T_1$ where $T_1 = NT$ with $N \gg 1$. Let the number of impulses arriving in $0 < t < T_1$ be the random variable K, and let the input be

$$\begin{aligned} x(t) &= \sum_{k=1}^{K} a_k \delta(t - t_k), \qquad && K \geqq 1; \\ x(t) &= 0, && K = 0; \end{aligned} \tag{59}$$

where, as in equation (4), the impulse amplitudes a_k are independent random variables with probability density $q(a)$ and expected value

$$E(a_k) = E(a), \qquad E(a_k^2) = E(a^2). \tag{60}$$

The arrival times $t_1, t_2, \cdots t_k$ are independent random variables with

$$\text{Prob}\,[t < t_k < t + dt] = dt/T_1\,. \tag{61}$$

The number of arrivals K has the Poisson distribution

$$\begin{aligned} \text{Prob}\,[K = L] &= (\nu T_1)^L e^{-\nu T_1}/L!, \\ E(K) &= \nu T_1\,, \\ E(K^2 - K) &= (\nu T_1)^2, \end{aligned} \tag{62}$$

where ν is the expected number of arrivals per second.

The output produced by the input (59) is

$$\begin{aligned} y(t) &= \sum_{k=1}^{K} a_k h(t, t_k), \qquad && K \geqq 1; \\ y(t) &= 0, && K = 0. \end{aligned} \tag{63}$$

When t is fixed, $y(t)$ may be regarded as a random variable since it depends on the random variables K, a_k, t_k. The lth moment of the distribution of $y(t)$ is the ensemble average $\langle y^l(t)\rangle$. Usually $\langle y^l(t)\rangle$ will depend on t and be periodic with period T. We shall be concerned with the first moment $\langle y(t)\rangle$ in the remainder of this section.

When the right side of the first part of equation (63) is averaged over the ensemble of a_k's, it becomes

$$E(a) \sum_{k=1}^{K} h(t, t_k), \qquad K \geqq 1. \tag{64}$$

Averaging this over the ensemble of t_k's gives

$$E(a) \sum_{k=1}^{K} \frac{1}{T_1} \int_0^{T_1} dt_k\, h(t, t_k) = KE(a) \frac{1}{T_1} \int_0^{T_1} dt_k\, h(t, t_k) \tag{65}$$

where use has been made of the fact that all of the terms in the series on the left are equal. Finally, averaging over the ensemble of K's with the help of $E(K) = \nu T_1$ gives

$$\langle y(t)\rangle = \nu E(a) \int_0^{T_1} dt_k\, h(t, t_k). \tag{66}$$

Dividing the interval $(0, T_1)$ into N equal intervals of length T, setting $t_k = nT + \tau$, and using the periodic property $h(t + nT, \tau + nT) = h(t, \tau)$ leads to

$$\begin{aligned}\langle y(t)\rangle &= \nu E(a) \sum_{n=0}^{N-1} \int_{nT}^{(n+1)T} dt_k\, h(t, t_k),\\ &= \nu E(a) \sum_{n=0}^{N-1} \int_0^{T} d\tau\, h(t - nT + nT, nT + \tau),\\ &= \nu E(a) \sum_{n=0}^{N-1} \int_0^{T} d\tau\, h(t - nT, \tau).\end{aligned} \tag{67}$$

Equation (67) holds when the system starts operating at $t = 0$ and stops at $t = T_1$. The following heuristic argument suggests that when (i) the system runs from $t = -\infty$ to $+\infty$, and (ii) $h(t, \tau)$ is such that only recent arrivals are of importance in determining the present state of the system, the analogue of equation (67) is

$$\langle y(t)\rangle = \nu E(a) \sum_{n=-\infty}^{\infty} \int_0^{T} d\tau\, h(t - nT, \tau). \tag{68}$$

We assume that, for $0 < \tau < T$, $h(u, \tau)$ becomes negligible when $|u| \geq mT$ where m is a small integer. We define t to be in the "interior" of $(0, T_1)$ when

$$mT < t < T_1 - mT.$$

If t is in the interior of $(0, T_1)$, the summation in equation (67) can be written as

$$\sum_{n=0}^{N-1} = \sum_{|t-nT|<mT} = \sum_{n=-\infty}^{\infty}$$

because $h(t - nT, \tau)$ is negligible except when $|t - nT| < mT$. Hence when t is in the interior of $(0, T_1)$ and the system runs from 0 to T_1 $< y(t) >$ is given by both (67) and (68).

In the interior of $(0, T_1)$ the starting and stopping transients near 0 and T_1 have died out, and $y(t)$ is the same irrespective of whether the system runs from 0 to T_1 or from $-\infty$ to $+\infty$. Hence when t is in the interior of $(0, T_1)$ and the system runs from $-\infty$ to $+\infty$, $\langle y(t) \rangle$ is again given by both (67) and (68).

The right side of equation (68) is a periodic function of t of period T. Physical considerations suggest that when the system runs from $-\infty$ to $+\infty$ $\langle y(t) \rangle$ is also a periodic function of period T. Since $\langle y(t) \rangle$ and the right side of equation (68) are equal when t lies in the interior of $(0, T_1)$ (which extends over more than one period), it is plausible to say that the equality holds for all values of t. This is what we wished to show.

Equation (68) appears as equation (13) in Section III. The sign of the index of summation n has been changed to make it easier to apply the formula.

VII. THE CUMULANTS FOR $y(t)$

The lth moment $\langle y^l(t) \rangle$ may be expressed in terms of the first l cumulants $\kappa_1(t), \cdots \kappa_l(t)$ of the distribution and conversely. For $l = 1$ and $l = 2$.

$$\begin{aligned} \kappa_1(t) &= \langle y(t) \rangle, \\ \kappa_2(t) &= \langle y^2(t) \rangle - \langle y(t) \rangle^2. \end{aligned} \tag{69}$$

The cumulants are defined by

$$\ln \varphi(z) = \sum_{l=1}^{\infty} \kappa_l(t)(iz)^l/l! \tag{70}$$

where $\varphi(z)$ is the characteristic function

$$\varphi(z) = \langle \exp [izy(t)] \rangle. \tag{71}$$

The method of averaging over the ensemble used in the preceding section to obtain $\langle y(t) \rangle$ will now be applied to calculate $\langle \exp [izy(t)] \rangle$. We have, because of the independence of the a_k's and t_k's,

$$\begin{aligned}\langle \exp [izy(t)] \rangle &= \left\langle \exp \left[iz \sum_{k=1}^{K} a_k h(t, t_k) \right] \right\rangle, \\ &= \sum_{K=0}^{\infty} \frac{(\nu T_1)^K}{K!} e^{-\nu T_1} \langle \exp [iza_k h(t, t_k)] \rangle^K, \\ &= \exp [-\nu T_1 + \nu T_1 \langle \exp [iza_k h(t, t_k)] \rangle].\end{aligned} \tag{72}$$

Therefore, upon using the definition (71) of $\varphi(z)$ and the probability densities of a_k and t_k,

$$\ln \varphi(z) = -\nu T_1 + \nu T_1 \int_{-\infty}^{\infty} da_k \, q(a_k) \int_0^{T_1} \frac{dt_k}{T_1} \exp [iza_k h(t, t_k)]. \tag{73}$$

Expanding both sides in powers of z and equating coefficients of $(iz)^l/l!$,

$$\begin{aligned}\kappa_l(t) &= \nu \int_{-\infty}^{\infty} da_k \, q(a_k) a_k^l \int_0^{T_1} dt_k \, h^l(t, t_k), \\ &= \nu E(a^l) \int_0^{T_1} dt_k \, h^l(t, t_k).\end{aligned} \tag{74}$$

When $l = 1$, equation (74) reduces to equation (66) for $\langle y(t) \rangle$. The steps that lead from equation (66) to the final expression for $\langle y(t) \rangle$ carry (74) into

$$\kappa_l(t) = \nu E(a^l) \sum_{n=-\infty}^{\infty} \int_0^T d\tau \, h^l(t - nT, \tau). \tag{75}$$

This appears as equation (15) in Section III with n replaced by $-n$.

VIII. THE POWER SPECTRUM OF $y(t)$

When $h(t, \tau)$ is such that $y(t)$ has the two-sided power spectrum $W_y(f)$, it is given by[4]

$$W_y(f) = \operatorname*{limit}_{T_1 \to \infty} \langle | \, S(f, T_1) \, |^2 \rangle / T_1 \tag{76}$$

where

$$\begin{aligned} S(f, T_1) &\equiv S(f, T_1\,; K; a_1\,, \cdots, a_K\,; t_1\,, \cdots, t_K) \\ &= \int_0^{T_1} dt\, e^{-i\omega t} y(t), \qquad \omega = 2\pi f; \\ &= \sum_{k=1}^{K} a_k \int_0^{T_1} dt\, e^{-i\omega t} h(t, t_k); \\ &= \sum_{k=1}^{K} a_k \int_{-t_k}^{T_1 - t_k} du\, e^{-i\omega (t_k + u)} h(t_k + u, t_k). \end{aligned} \tag{77}$$

In the derivation of equation (68) for $\langle y(t) \rangle$, the limits of summation $n = 0$, $n = N-1$ were replaced by $n = -\infty$, $n = \infty$. In much the same way, we assume that in (77) the limits of integration $-t_k$, $T_1 - t_k$ can be replaced by $-\infty$, $+\infty$ in all but a negligible fraction of the terms (those with t_k near 0 or T_1). This presupposes a sufficiently rapid decrease in the value of $|h(t_k + u, t_k)|$ as $|u| \to \infty$. Heuristically, we picture $h(t_k + u, t_k)$ as being negligible except when u is small. When T_1 is very large, most of the t_k's and $(T_1 - t_k)$'s will be large. Consequently, for most of the t_k's, $h(t_k + u, t_k)$ will be negligible when u is less than $-t_k$ or greater than $T - t_k$.

This assumption allows us to replace equations (76) and (77) by

$$W_y(f) = \operatorname*{limit}_{T_1 \to \infty} \langle | S_a(f, T_1) |^2 \rangle / T_1 \tag{78}$$

and

$$\begin{aligned} S_a(f, T_1) &= \sum_{k=1}^{K} a_k \int_{-\infty}^{\infty} du\, e^{-i\omega (t_k + u)} h(t_k + u, t_k), \\ &= \sum_{k=1}^{K} a_k s(f, t_k), \end{aligned} \tag{79}$$

where

$$s(f, \tau) = \int_{-\infty}^{\infty} dt\, e^{-i\omega t} h(t, \tau). \tag{80}$$

From equation (79)

$$\begin{aligned} | S_a(f, T_1) |^2 &= S_a(f, T_1) S_a^*(f, T_1), \\ &= \sum_{k=1}^{K} \sum_{l=1}^{K} a_k a_l s(f, t_k) s^*(f, t_l), \end{aligned} \tag{81}$$

where the star denotes conjugate complex. The terms in equation (81) can be divided into two types. For Type *I*, $l = k$, and for Type II, $l \neq k$. It is convenient to take their ensemble averages separately.

The typical Type I term is

$$a_k^2 \mid s(f, t_k) \mid^2. \tag{82}$$

There are K terms of Type I in the double sum (81), and all of them are of the form (82). Therefore, when use is made of $E(K) = \nu T_1$, the contribution of the Type I terms to $\langle |S_a(f, T_1)|^2 \rangle$ is found to be

$$\nu E(a^2) \int_0^{T_1} d\tau \mid s(f, \tau) \mid^2. \tag{83}$$

The typical Type II term in (86) is

$$a_k a_l s(f, t_k) s^*(f, t_l), \qquad l \neq k.$$

When averaged with respect to a_k, a_l, t_k, t_l it becomes

$$\left| E(a) \int_0^{T_1} \frac{dt_k}{T_1} s(f, t_k) \right|^2. \tag{84}$$

There are $K^2 - K$ terms of Type II in the double sum (81) and all of them have the average value (84). Therefore, when use is made of $E(K^2 - K) = \nu^2 T_1^2$, the contribution of Type II terms to $\langle \mid S_a(f, T_1) \mid^2 \rangle$ is found to be

$$\left| \nu E(a) \int_0^{T_1} d\tau \, s(f, \tau) \right|^2. \tag{85}$$

Adding the contributions of Type I and Type II, and inserting the resulting expression for $\langle |S_a(f, T_1)|^2 \rangle$ in equation (78) for the power spectrum gives, with $\omega = 2\pi f$ and $s(f, \tau)$ given by (80),

$$W_y(f) = \lim_{T_1 \to \infty} \frac{1}{T_1} \left[\nu E(a^2) \int_0^{T_1} d\tau \mid s(f, \tau) \mid^2 + \left| \nu E(a) \int_0^{T_1} d\tau \, s(f, \tau) \right|^2 \right] \tag{86}$$

provided $s(f, \tau)$ [i.e., $h(t, \tau)$] is such that the limit exists. If, for certain frequencies, the function of T_1 following the limit sign ultimately increases linearly with T_1, $W_y(f)$ has an infinite spike at these frequencies. This means that $y(t)$ has sinusoidal components at these frequencies.

So far in this section, the time variation of the system has not been assumed to be periodic. Now we apply (86) to the case in which the system varies periodically with period T and, in accordance with equations (3) and (80),

$$\begin{aligned} h(t - nT + nT, \tau + nT) &= h(t - nT, \tau), \\ s(f, \tau + nT) &= e^{-i\omega nT} s(f, \tau). \end{aligned} \tag{87}$$

In (86) set $T_1 = NT$ and let $N \to \infty$. Then

$$\begin{aligned} \frac{1}{T_1}\int_0^{T_1} d\tau \mid s(f,\tau) \mid^2 &= \frac{1}{T}\int_0^{T} d\tau \mid s(f,\tau) \mid^2, \\ \int_0^{T_1} d\tau\, s(f,\tau) &= T\sum_{n=0}^{N-1} e^{-i\omega nT} s_0(f), \\ &= Ts_0(f)\,\frac{1-e^{-i\omega NT}}{1-e^{-i\omega T}}, \end{aligned} \tag{88}$$

where

$$s_0(f) = \frac{1}{T}\int_0^T d\tau\, s(f,\tau), \qquad \omega = 2\pi f. \tag{89}$$

The contribution of the second term in (86) contains the factor

$$\begin{aligned} \operatorname*{limit}_{N\to\infty} \frac{T}{N}\left| \frac{1-e^{-i\omega NT}}{1-e^{-i\omega T}} \right|^2 &= \operatorname*{limit}_{N\to\infty} \frac{T}{N}\left| \frac{\sin\dfrac{\omega NT}{2}}{\sin\dfrac{\omega T}{2}} \right|^2 \\ &= \sum_{m=-\infty}^{\infty} \delta\left(f-\frac{m}{T}\right), \end{aligned} \tag{90}$$

where the last step follows from the relations used in the proof of Fejér's theorem in the theory of Fourier series. When these results are used in equation (86), it goes into

$$\begin{aligned} W_y(f) = \nu E(a^2)\,\frac{1}{T}\int_0^T d\tau \mid s(f,\tau) \mid^2 \\ + \nu^2E^2(a) \sum_{m=-\infty}^{\infty} \delta\left(f-\frac{m}{T}\right)\left| s_0\left(\frac{m}{T}\right)\right|^2. \end{aligned} \tag{91}$$

Equation (91) shows spikes in $W_y(f)$ at $f = m/T$ where m is an integer. The spike at $f = 0$ corresponds to the dc component y_{dc} of $y(t)$, and the spikes at $\pm m/T$ to the sinusoidal component

$$A_m \cos\,[2\pi m(t/T) + \theta_m] \tag{92}$$

in $y(t)$. The expression (91) for $W_y(f)$ shows that the (time) average powers in these components are

$$\begin{aligned} y_{\text{dc}}^2 &= [\nu E(a)s_0(0)]^2, \\ \tfrac{1}{2}A_m^2 &= [\nu E(a)]^2[\mid s_0(-m/T) \mid^2 + \mid s_0(m/T) \mid^2], \\ &= 2[\nu E(a) \mid s_0(m/T) \mid]^2. \end{aligned} \tag{93}$$

Equations (93) tell us nothing about the sign of y_{dc} or about the phase angle θ_m. One of the several ways to get this information is to imagine $y(t)$ expanded in a Fourier series of long period T_1,

$$\begin{aligned} y(t) &= \sum_{-\infty}^{\infty} c_n \exp(i2\pi nt/T_1), \\ c_n &= \frac{1}{T_1}\int_0^{T_1} dt\, y(t) \exp(-i2\pi nt/T_1), \\ &= \frac{1}{T_1} S\left(\frac{n}{T_1}, T_1\right), \\ &\approx \frac{1}{T_1} S_a\left(\frac{n}{T_1}, T_1\right), \\ &= \frac{1}{T_1}\sum_{k=1}^{K} a_k s\left(\frac{n}{T_1}, t_k\right), \end{aligned} \tag{94}$$

where we have used equations (77) and (79) for $S(f, T_1)$ and its approximation $S_a(f, T_1)$. The expression (91) for $W_y(f)$ shows that the c_n's may be divided into two classes; those corresponding to the frequencies $n/T_1 = m/T$, i.e., $n = mN$ (discrete sinusoidal components) and those corresponding to $n \neq mN$ (noise). For the first class, c_n is $O(1)$ and nearly the same for most $y(t)$'s of the ensemble. For the second class, c_n is $O(T_1^{-\frac{1}{2}})$ and varies greatly from member to member.

To obtain the discrete sinusoidal component in $y(t)$ of frequency m/T, we set $n = mN$ in equation (94) and apply the procedure used in Section VI (to obtain $\langle y(t)\rangle$) to average c_n over the ensemble.

$$\begin{aligned} \langle c_n\rangle_{n=mN} &= \frac{1}{T_1} E(K)E(a)\int_0^{T_1}\frac{dt_k}{T_1} s\left(\frac{m}{T}, t_k\right), \\ &= \nu E(a)T\sum_{n=0}^{N-1}\frac{1}{T_1}\exp\left[\frac{-i2\pi m(nT)}{T}\right] s_0\left(\frac{m}{T}\right), \\ &= \nu E(a)s_0(m/T), \end{aligned} \tag{95}$$

where we have used equations (88) and (89) with $\omega = 2\pi m/T$. We therefore write $y(t)$ as the sum of a noise component $y_N(t)$, consisting of the sum of terms of the second class, and a periodic component $y_{\mathrm{per}}(t)$, consisting of the sum of terms of the first class:

$$y(t) = y_N(t) + y_{\mathrm{per}}(t). \tag{96}$$

The power spectrum of $y_N(t)$ is the first term in the expression (91) for $W_y(t)$:

$$W_{yN}(f) = \nu E(a^2) \frac{1}{T} \int_0^T d\tau \mid s(f, \tau) \mid^2. \tag{97}$$

The periodic component is, from (95),

$$y_{\text{per}}(t) = \nu E(a) \sum_{m=-\infty}^{\infty} s_0(m/T) \exp [i2\pi mt/T]. \tag{98}$$

The parts $y_N(t)$ and $y_{\text{per}}(t)$ of $y(t)$ are related to the ensemble averages by

$$y_{\text{per}}(t) = \langle y(t) \rangle = \kappa_1(t), \tag{99}$$

$$\langle y_N^2(t) \rangle = \langle y^2(t) \rangle - \langle y(t) \rangle^2 = \kappa_2(t). \tag{100}$$

Equation (99) can be proved by showing that the mth Fourier coefficients of $y_{\text{per}}(t)$ and $\langle y(t) \rangle$ are equal for all integers m, i.e., by showing that

$$\nu E(a) s_0(m/T) = \frac{1}{T} \int_0^T \langle y(t) \rangle \exp(-i2\pi mt/T)\, dt. \tag{101}$$

When the series (68) for $\langle y(t) \rangle$ is substituted on the right, the summation and the integration with respect to t from 0 to T combine to give an integral in t with limits $\pm\infty$. This integral can be evaluated with the help of the integral (80) for $s(f, \tau)$ and leads to the verification of (99). Equation (100) follows from the ensemble average of the square of

$$y_N(t) = y(t) - y_{\text{per}}(t) = y(t) - \langle y(t) \rangle.$$

Setting $l = 2$ in the expression (75) for $\kappa_l(t)$ and using (100) gives an expression for the ensemble average of $y_N^2(t)$ at time t,

$$\langle y_N^2(t) \rangle = \kappa_2(t) = \nu E(a^2) \sum_{n=-\infty}^{\infty} \int_0^T h^2(t - nT, \tau)\, d\tau. \tag{102}$$

It follows from (102) that when the variance $\langle y_N^2(t) \rangle$ varies with t, it varies periodically with period T. When equation (102) is averaged over a period and use is made of the ergodic relation (2), we get the time average

$$\overline{y_N^2(t)} = \overline{\langle y_N^2(t) \rangle} = \frac{\nu E(a^2)}{T} \int_0^T d\tau \int_{-\infty}^{\infty} dt\, h^2(t, \tau). \tag{103}$$

From the expression (97) for $W_{yN}(f)$, we get a second expression for $\overline{y_N^2(t)}$

$$y_N^2\overline{(t)} = \int_{-\infty}^{\infty} W_{yN}(f)\, df = \frac{\nu E(a^2)}{T} \int_0^T d\tau \int_{-\infty}^{\infty} df \mid s(f, \tau) \mid^2. \tag{104}$$

The equality of (103) and (104) can also be proved directly by using the Fourier integral (80) relating $s(f, \tau)$ and $h(t, \tau)$.

IX. WHITE GAUSSIAN NOISE INPUT

Let the input $x(t)$ of the periodically varying system shown in Fig. 2 be white gaussian noise with zero mean. Here we show that the output $y(t)$ has no dc or sinusoidal components, and that the power spectrum of $y(t)$is

$$W_y(f) = \frac{N_0}{T}\int_0^T |\, s(f, \tau)\,|^2\, d\tau \tag{105}$$

where the power spectrum of $x(t)$ is $W_x(f) = N_0$ for $|f| < F$ and $W_x(f) = 0$ for $|f| > F$ with $F \to \infty$.

Consider Fig. 4 in which an ideal low pass filter which passes only the frequencies $|f| < F$ has been inserted between the input and the periodically varying network specified by $h(t, \tau)$.

When $x(t)$ is a unit impulse applied at time τ, $x(t) = \delta(t - \tau)$, the filter output at time $t = t_1$ is

$$z(t_1) = \frac{\sin 2\pi F(t_1 - \tau)}{\pi(t_1 - \tau)}\,, \tag{106}$$

and the system output at time t is

$$y(t) = \int_{-\infty}^{\infty} h(t, t_1)\,\frac{\sin 2\pi F(t_1 - \tau)}{\pi(t_1 - \tau)}\, dt_1\ . \tag{107}$$

Thus, when $h(t, t_1)$ satisfies conditions associated with the Fourier integral theorem, $y(t)$ tends to $h(t, \tau)$ as $F \to \infty$; a result which follows immediately from physical considerations.

Take $x(t)$ to be the shot noise given by (4) in which, for given values of N_0 and ν, $a_k = \pm(N_0/\nu)^{\frac{1}{2}}$ with equal probability. Then

$$\begin{aligned}\nu E(a) &= \nu E(a_k) = 0,\\ \nu E(a^2) &= \nu E(a_k^2) = N_0\,,\end{aligned} \tag{108}$$

and the filter output is the zero-mean shot noise

$$z(t) = \sum_{-\infty}^{\infty} a_k\,\frac{\sin 2\pi F(t - t_k)}{\pi(t - t_k)} \tag{109}$$

with the power spectrum

$$\begin{aligned} W_z(f) &= \nu E(a^2)\left|\int_{-\infty}^{\infty}\frac{\sin 2\pi F t}{\pi t}\, e^{-i\omega t}\, dt\right|^2, \qquad \omega = 2\pi f;\\ &= \begin{cases} N_0, & |f| < F;\\ 0, & |f| > F.\end{cases}\end{aligned} \tag{110}$$

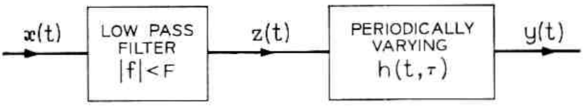


Fig. 4—Conversion of shot noise $x(t)$ to white noise $z(t)$.

Now hold F fixed and let $\nu \rightarrow \infty$. The individual pulses comprising $z(t)$ become smaller and smaller, and overlap more and more. In the limit $z(t)$ becomes zero-mean gaussian noise with the power spectrum (110).

Finally, let $F \rightarrow \infty$. Then $z(t)$ becomes white gaussian noise with the flat power spectrum $W_z(f) = N_0$. According to equation (107), the response of the Fig. 4 system at time t to a unit impulse applied at time τ tends to $h(t, \tau)$ as $F \rightarrow \infty$. Therefore, the results obtained in Sections VI, VII, and VIII for shot noise input in Fig. 2 are carried into corresponding results for white noise input (i.e., $x(t)$ in Fig. 2 is white gaussian noise) by the substitutions (108), namely $\nu E(a) = 0$ and $\nu E(a^2) = N_0$.

Setting $\nu E(a) = 0$ in equation (98) for $y_{\text{per}}(t)$ shows that $y_{\text{per}}(t)$ is zero for zero-mean white noise input. Consequently, $y(t)$ contains no dc or sinusoidal components.

Setting $\nu E(a) = 0$, and $\nu E(a^2) = N_0$ in equation (91) for $W_y(f)$ shows that the power spectrum of $y(t)$ is given by equation (105) when the input is white gaussian noise. Furthermore, $y(t)$ is composed entirely of $y_N(t)$; and equations (102), (103), and (104) become

$$\langle y^2(t)\rangle = N_0 \sum_{n=-\infty}^{\infty} \int_0^T h^2(t - nT, \tau)\, d\tau, \tag{111}$$

$$\begin{aligned}\overline{y^2(t)} &= \frac{N_0}{T} \int_0^T d\tau \int_{-\infty}^{\infty} dt\, h^2(t, \tau), \\ &= \frac{N_0}{T} \int_0^T d\tau \int_{-\infty}^{\infty} df\, |s(f, \tau)|^2.\end{aligned} \tag{112}$$

The fraction of time any particular member of the ensemble of outputs spends in the infinitesimal interval $Y < y(t) < Y + dY$ is

$$\frac{dY}{T} \int_0^T dt[2\pi\langle y^2(t)\rangle]^{-\frac{1}{2}} \exp\,[-Y^2/(2\langle y^2(t)\rangle)] \tag{113}$$

where $\langle y^2(t)\rangle$ is the function of t defined by equation (111).

X. ACKNOWLEDGMENTS

I wish to express my appreciation for the help I have received from many discussions with D. D. Sell, and for his computation of the power spectrum curves given in Fig. 3. I also wish to thank H. A. Bodner and R. E. Langseth for their helpful comments.

APPENDIX A

The Power Spectrum for Figure 1a

Here we give some of the steps leading from the first line to the second line of equation (42) for $W_{V_N}(f)$. The first line is

$$W_{V_N}(f) = \frac{\nu q^2}{T} \int_0^{\alpha T} |s(f, \tau)|^2 \, d\tau \tag{114}$$

where $s(f, \tau)$ is given by equation (31). Multiplying (31) by its complex conjugate gives

$$\begin{aligned} |s(f, \tau)|^2 = {} & \frac{C^{-2}}{\gamma^2 + \omega^2} + \frac{C^{-2} b^2 e^{2\alpha\tau} \gamma^2}{|(\gamma + i\omega)\omega|^2} \left| \frac{z - z^\alpha}{1 - bz} \right|^2 \\ & + 2 \text{ Real} \left[\frac{C^{-2} b e^{\gamma\tau + i\omega\tau} (z - z^\alpha)(-\gamma)}{(\gamma - i\omega)(1 - bz)(\gamma + i\omega)(i\omega)} \right]. \end{aligned} \tag{115}$$

Then

$$\begin{aligned} \int_0^{\alpha T} |s(f, \tau)|^2 \, d\tau = {} & \frac{C^{-2}}{\gamma^2 + \omega^2} \left[\alpha T + \frac{b^2 (e^{2\gamma\alpha T} - 1)\gamma}{2\omega^2} \left| \frac{z - z^\alpha}{1 - bz} \right|^2 \right. \\ & \left. + 2 \text{ Real} \frac{b(e^{\gamma\alpha T + i\omega\alpha T} - 1)(z - z^\alpha)(-\gamma)}{(1 - bz)(\gamma + i\omega)(i\omega)} \right]. \end{aligned} \tag{116}$$

Upon introducing the values $b = \exp(-\gamma\alpha T)$, $z = \exp(-i\omega T)$, and using the identity

$$\tfrac{1}{2}(1 - b^2) \left| \frac{z - z^\alpha}{1 - bz} \right|^2 = -\text{Real} \frac{(z^{-\alpha} - b)(z - z^\alpha)}{1 - bz} \tag{117}$$

the quantity within the square brackets in equation (116) becomes

$$\alpha T + \text{Real} \frac{(1 - bz^\alpha)(1 - z^{1-\alpha})\gamma(\gamma - i\omega)}{(1 - bz)\omega^2(\gamma + i\omega)} \tag{118}$$

and thus leads to the expression of $W_{V_N}(f)$ given by the second line of equation (42).

REFERENCES

1. Sell, D. D., "A Sensitive Spectrophotometer for Optical Reflectance and Transmittance Measurements," J. Applied Optics, *9,* No. 8 (August 1970), pp. 1926–1930.
2. Hurd, H. L., "An Investigation Periodically Correlated Stochastic Processes," Ph.D. Thesis, Duke University, 1969.
3. Unpublished application notes obtained from Princeton Applied Research Corp., Princeton, N.J.
4. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23,* No. 3 (July 1944), pp. 282–332.
5. Bennett, W. R., "New Results in the Calculation of Modulation Products," B.S.T.J., *12,* No. 2 (April 1933), pp. 228–243.

A New Approach for Evaluating the Error Probability in the Presence of Intersymbol Interference and Additive Gaussian Noise

By E. Y. HO and Y. S. YEH

(Manuscript received June 25, 1970)

The determination of the error probability of a data transmission system in the presence of intersymbol interference and additive gaussian noise is a major goal in the analysis of such systems. The exhaustive method for finding the error probability calculates all the possible states of the received signal using an N-sample approximation of the true channel impulse response. This method is too time-consuming because the computation involved grows exponentially with N. The worst-case sequence bound avoids the lengthy computation problem but is generally too loose.

In this paper, we have developed a new method which yields the error probability in terms of the first 2k moments of the intersymbol interference. A recurrence relation for the moments is derived. Therefore, a good approximation to the error probability of the true channel can be obtained by choosing N large enough, and the amount of computation involved increases only linearly with N. The series expansion is shown to be absolutely convergent, and an upper bound on the series truncation error is given. In order to show the improvement provided in this new method, it is compared with the Chernoff bound technique in three representative cases. An order of magnitude improvement in accuracy is obtained.*

I. INTRODUCTION

An important problem in the analysis of binary digital data systems is the determination of the system performance in the presence of intersymbol interference and additive gaussian noise. Since it is usually the most meaningful criterion in designing a digital data

* In April 1970, the authors were advised by R. W. Pulleyblank that a similar method was discovered independently by M. Celebiler and O. Shimbo to be presented in a paper which will be published in Conference Record, ICC, 1970.

system, the error probability is chosen as the measure of the system performance.

Two alternatives are available at present. The first alternative[1,2] considers a truncated N-pulse-train approximation of the true channel. The error probability is calculated by evaluating the conditional error probability of each of 2^N possible data sequences and averaging over all 2^N sequences. Since each calculation of the conditional error probability takes a great deal of computer time, the number of sequences must be held to several thousand.[3] This limitation leads to a poor approximation of the true channel, and the error probability so obtained is not very useful. The second alternative evaluates an upper bound of the error probability by either the worst-case sequence[3] or the Chernoff inequality.[4,5] In many cases, the bound is too loose.

In this study we have developed a new way to evaluate the error probability in terms of the first $2k$ moments of the intersymbol interference. It provides a significant improvement in accuracy over the worst-case sequence bound or the Chernoff bound. The computations increase only linearly with N. Thus a good approximation of the true channel may be obtained. The convergence of this alternative is proved. Throughout, additive gaussian noise and independence of information digits are assumed. The generalization to a multilevel system is straightforward; hence, only binary systems will be considered in this study.

II. BRIEF DESCRIPTION OF THE SYSTEM

A simplified block diagram of a binary amplitude modulation (AM) data system is shown in Fig. 1. We assume that a single $s(t)$ having amplitude a_ℓ is transmitted through the channel every T seconds. The system transfer function is

$$R(\omega) = S(\omega)T(\omega)E(\omega) \tag{1}$$

where $s(t)$ and $r(t)$ are the Fourier transform pair of $S(\omega)$ and $R(\omega)$, respectively. In the absence of channel noise, a sequence of input channel signals

$$\sum_{\ell=-\infty}^{\infty} a_\ell s(t - \ell T), \tag{2}$$

will generate a corresponding output sequence

$$\sum_{\ell=-\infty}^{\infty} a_\ell r(t - \ell T), \tag{3}$$

where $\{a_\ell\}$ is a sequence of independent binary random variables,

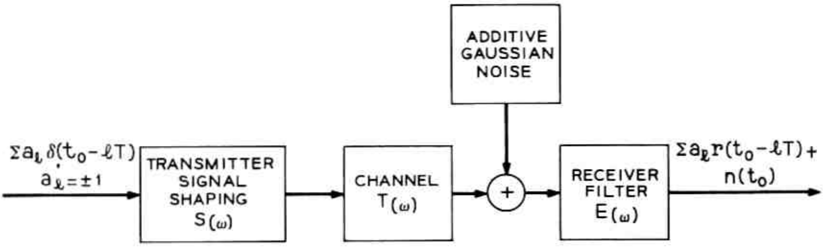


Fig. 1—Simplified block diagram of a binary AM data system.

$a_\ell = \pm 1$, and satisfies

$$P_r(a_\ell = 1) = P_r(a_\ell = -1) = \tfrac{1}{2}$$
$$\ell = -\infty, \cdots, -1, 0, 1, \cdots \infty. \qquad (4)$$

We also assume that additive gaussian noise is present in the system. Thus the corrupted received sequence at the input to the receiver detector is

$$y(t) = \sum_{\ell=-\infty}^{\infty} a_\ell r(t - \ell T) + n(t), \qquad (5)$$

where $n(t)$ is additive gaussian noise with a one-sided power spectral density of σ^2 watts/cps.

At the detector, $y(t)$ is sampled every T seconds to determine the transmitted signal. At sampling instant t_0, the sampled signal is

$$y(t_0) = a_0 r(t_0) + \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} a_\ell r(t_0 - \ell T) + n(t_0). \qquad (6)$$

The first term is the desired signal while the second and the third terms represent the intersymbol interference and gaussian noise respectively.

It is well known that the optimum (minimum error probability) decision level is zero. Thus the error probability is given by

$$P_e = P_r\left\{\left[\sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} a_\ell r(t_0 - \ell T) + n(t_0)\right] \geqq r(t_0)\right\}. \qquad (7)$$

For the real system we are interested in, we may assume that the $a_\ell r(t_0 - \ell T)$'s are uniformly bounded and $\sum_{\ell \neq 0} a_\ell r(t_0 - \ell T)$ converges absolutely.* For example, in a system having an open binary eye,

[*] Finite truncated pulse-train approximation will be used for those pulses with absolutely divergent intersymbol interference.

$\sum_{\ell \neq 0} | r(t_0 - \ell T) |$ is less than $r(t_0)$. Thus by Kolmogorov's Three-Series criterion,[6] it can be easily shown that $\sum_{\ell \neq 0} a_\ell r(t_0 - \ell T)$ converges absolutely to a random variable.

Equation (7) can be calculated by evaluating the expected value of the conditional expectation of the error probability for a given random variable $\sum_{\ell \neq 0} a_\ell r(t_0 - \ell T)$; therefore,

$$P_e = \int_{\text{all } X} \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{0} \exp\left[-\{y - r(t_0) - X\}^2/2\sigma^2\right] dy\, dF(x), \qquad (8)$$

where $F(X)$ is the distribution function of the random variable X, and $X = \sum_{\ell \neq 0} a_\ell r(t_0 - \ell T)$.

III. SERIES EXPANSION OF P_e

With the exception of a few special cases, equation (8) is generally difficult to solve. The existing solutions are either too time-consuming[1,2] or inaccurate.[3,4,5]

We have found that equation (8) can be evaluated in terms of an absolutely convergent series involving moments of the intersymbol interference. Furthermore, the moments can be obtained readily through recurrence relations. Therefore, the computation time is significantly reduced in comparison with the exhaustive method.[1,2] The absolute convergence and the recurrence relations for the moments are given in Appendix A and B respectively.

Expanding equation (8), we obtain the following expression for the error probability,

$$\begin{aligned} P_e &= \tfrac{1}{2} \operatorname{erfc}\left(-\frac{r(t_0)}{\sqrt{2}\,\sigma}\right) + \sum_k \frac{1}{(2k)!} \cdot \left(\frac{1}{2\sigma^2}\right)^k \cdot M_{2k} \cdot \frac{1}{\sqrt{\pi}} \\ &\qquad \cdot \left[\exp - \left(\frac{r^2(t_0)}{2\sigma^2}\right)\right] \cdot H_{2k-1}\left(\frac{r(t_0)}{\sqrt{2}\,\sigma}\right) \qquad k = 1, 2, 3, \cdots, \\ &= P_{e_0} + \sum_{k=1}^{\infty} P_{e_{2k}}, \end{aligned} \qquad (9)$$

where $H_{2k-1}(x)$ is a Hermite polynominal, M_{2k} is the $2k$th moment of the random variable X, and

$$\operatorname{erfc}(-x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{-x} \exp(-z^2)\, dz. \qquad (10)$$

The first term in equation (9) represents the nominal system error probability due to additive gaussian noise alone while the summation represents the degradation of the system performance due to intersymbol interference in the additive gaussian noise environment.

3.1 *Convergence Property*

In Appendix A we have shown that equation (9) is an absolutely convergent series. Therefore, the error probability can be evaluated by taking a finite number of terms,

$$P_e = \sum_{k=0}^{K-1} P_{e_{2k}} + R_{2K} \, , \tag{11}$$

where R_{2K} represents the truncation error and is upper bounded by

$$R_{2K} = \sum_{k=K}^{\infty} P_{e_{2k}} \leqq \frac{(2K-3)!!}{(2K)!} \sqrt{4K-2} \cdot \frac{1}{2\sigma^{2k}} \cdot \frac{1}{\sqrt{\pi}} \cdot \frac{\left[\exp - \left(\frac{r^2(t_0)}{4\sigma^2}\right)\right] \cdot \left[\sum_{\ell \neq 0} \mid r(t_0 - \ell T) \mid\right]^{2K}}{\left[1 - \frac{1}{2K}\left(\frac{\sum_{\ell \neq 0} \mid r(t_0 - \ell T) \mid}{\sigma}\right)^2\right]},$$

$$= U_{2K} \, . \tag{12*}$$

Thus for a given truncation error bound, ϵ, we may always find a positive integer, K, such that

$$U_{2K} \leqq \epsilon. \tag{13}$$

For a real system, the truncation error is generally much smaller than ϵ. Therefore, fewer terms are needed in evaluating the error probability.

3.2 *Evaluation of Moments*

The series expansion of equation (9) can be readily evaluated if we can determine the moment, M_{2k}. The M_{2k}'s are given by

$$M_{2k} = \int_{\text{all } X} X^{2k} \, dF(X). \tag{14}$$

To evaluate M_{2k} according to equation (14) requires the knowledge of $dF(X)$; this is just as difficult to obtain as the evaluation of the error probability given by equation (8). However, we have found it possible to obtain a recurrence relation for M_{2k} by examining the first deriva-

* $(2K - 3)!! = (2K - 3) \cdot (2K - 5) \cdots 3 \cdot 1.$

tive of the characteristic function. The recurrence formula makes the series expansion approach feasible, and is derived in Appendix B:

$$M_{2k} = -\left\{ \sum_{i=1}^{k} \binom{2k-1}{2i-1} (-1)^i M_{2(k-i)} f^{(2i-1)}(0) \right\}, \tag{15}$$

where

$$M_0 = 1 \tag{16}$$

$$f^{(2i-1)}(0) = \frac{2^{2i}(2^{2i}-1)}{2i} \mid B_{2i} \mid \sum_{\ell \neq 0} [r(t_0 - \ell T)]^{2i} \tag{17}$$

and B_{2i}'s are Bernoulli numbers.

3.3 *Truncated Pulse-Train Approximation*

For any real binary system, the message must be time-limited to a finite number of symbol durations, or we may even assume that $r(t)$ is time-limited to, say, N symbol durations. Thus the error probability may be calculated by evaluating the conditional error probability for each of 2^N possible data sequences and then averaging over all 2^N sequences. Since the number of possible data sequences grows exponentially with N, it would be impractical to evaluate the error probability by this straightforward method even with a digital computer. Hence, N must be confined to a small number; the error probability so obtained could at best be a poor approximation of the true error probability. However, in equation (9), the amount of computation involved grows only linearly with N. Therefore, the pulse train can be truncated at any desired point to assure a good approximation of the true channel.

IV. APPLICATIONS

The error probabilities for certain cases are calculated by equation (9) to determine the accuracy and the convergence of this new method.

4.1 *Case 1: Data Set* 203[7]

A 2400-baud DDD option of the Data Set 203 operating over a channel having symmetrical parabolic delay distortion, as shown in Fig. 2, is considered in this case. The group delays at the carrier and the lower 3-dB frequencies are 0.6 ms relative to the center of the signal spectrum. The channel we considered is worse than a worst-case-C2 line. A 5-tap mean-square equalizer is used by the receiver to equalize the channel. A truncated 34-pulse-train approximation (19 samples after and 15 samples before the sampling instant t_0) for the

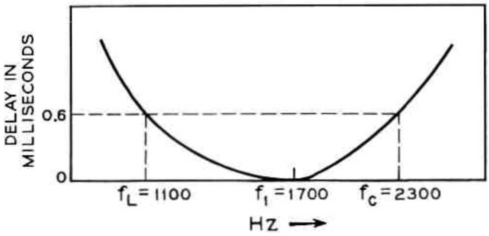


Fig. 2—Channel group-delay-frequency response.

equalized output impulse response was used. The equalized binary eye is about 70 percent open in this case. The input signal-to-noise ratio is 14 dB. The error probabilities at the equalizer output evaluated by equation (9) and the Chernoff inequality are shown in Fig. 3. Curve (a) is the Chernoff bound. Curve (b) is the error probability evaluated by taking a finite number of terms in equation (9). Curve (c) is the truncation error bound given by equation (12). It can be seen that taking the first nine terms in equation (9) assures less than one percent truncation error in evaluating the error probability. In this case, however, the actual series converges after only four terms. An improvement in accuracy by a factor of 15 is realized by this series expansion method compared to that obtained by Chernoff inequality.

4.2 *Case 2: Ideal Channel and Ideal Band-Limited Pulse*

The received pulse is assumed to have the form,

$$r(t)_k = \frac{\sin \pi t/T}{\pi t/T}. \tag{18}$$

The signal-to-noise ratio at the nominal sampling instant is taken to be 16 dB. In the absence of intersymbol interference, the system error probability is 10^{-10}. For a truncated 11-pulse-train approximation, the exact error probabilities and the error probabilities evaluated by taking a finite number of terms in equation (9) for different values of sampling instant and number of terms and equation (12) are shown in Figs. 4–5. It can be seen from these figures that the series converges more rapidly for smaller values of the quantity

$$q(t_0) = \left(\sum_{\substack{\ell=-5 \\ \ell\neq 0}}^{5} | r(t_0 - \ell T) | / \sigma \right)^2$$

[e.g., in this case, $q(0.05T) = 1.96$].

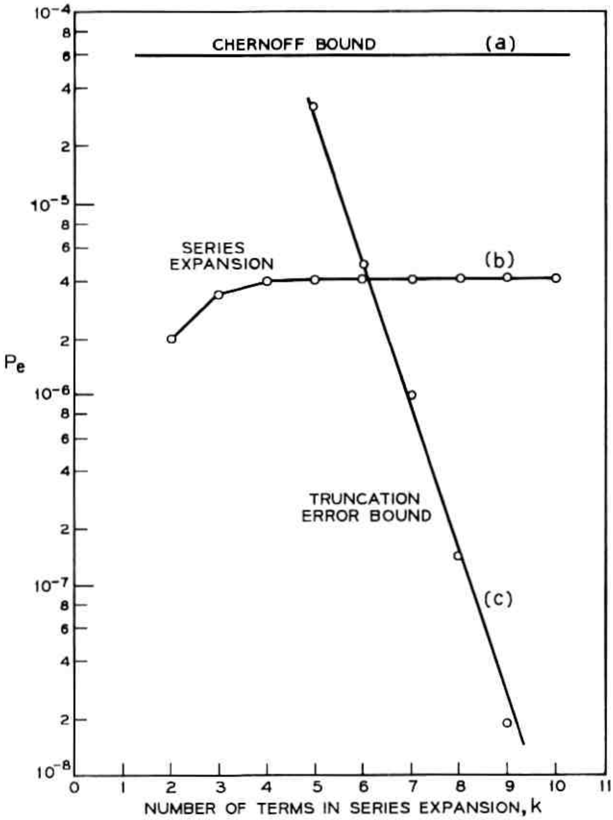


Fig. 3—Comparison of error probabilities obtained by Chernoff bound and series expansion method. $(S/N)_{\text{input}} = 14$ dB; data set 203 (2400-Baud Option); 5-tap mean square equalizer; parabolic delay distortion channel (see Fig. 2).

The series starts to oscillate when $q(t_0)$ is not small [e.g., $q(0.2T) = 30.8$]. At $t_0 = 0.2T$, the series did not converge well for the first eight terms in equation (9). However, it will converge to the exact value eventually. The error probabilities obtained by Chernoff bound,[5] exact calculation, and equation (9) are shown in Fig. 6. It is clear that this new alternative provides a significant improvement over the Chernoff bound.

4.3 *Case 3: Ideal Channel and Fourth-Order Chebyshev Pulse*[5]

In this case, a fourth-order Chebyshev filter is used. The received pulse is

$$r(t) = A_1 \cos(\omega_1 \mid t \mid /T - \Phi_1) \cdot \exp[-\alpha_1 \mid t \mid /T] + A_2 \cos(\omega_2 \mid t \mid /T - \Phi_2) \cdot \exp[-\alpha_2 \mid t \mid /T], \qquad (19)$$

with

$$A_1 = 0.4023, \qquad A_2 = 0.7163,$$
$$\omega_1 = 2.839, \qquad \omega_2 = 1.176,$$
$$\Phi_1 = 0.7553, \qquad \Phi_2 = 0.1602,$$
$$\alpha_1 = 0.4587, \qquad \alpha_2 = 1.107.$$

The signal-to-noise ratio at the nominal sampling instant is taken to be 16 dB. For a truncated 11-pulse-train approximation, the exact error probabilities and the error probabilities obtained by taking a finite number of terms in equation (9) for various sampling instants and numbers of terms are shown in Figs. 7–8. The error probabilities obtained by the Chernoff[5] bound, the exact calculation, and equation (9) are shown in Fig. 9. The same results as in case 2 are observed.

V. SUMMARY AND CONCLUSIONS

In this study we have developed a new method of evaluating the error probability for synchronous data systems in the presence of intersymbol interference and additive gaussian noise under the fol-

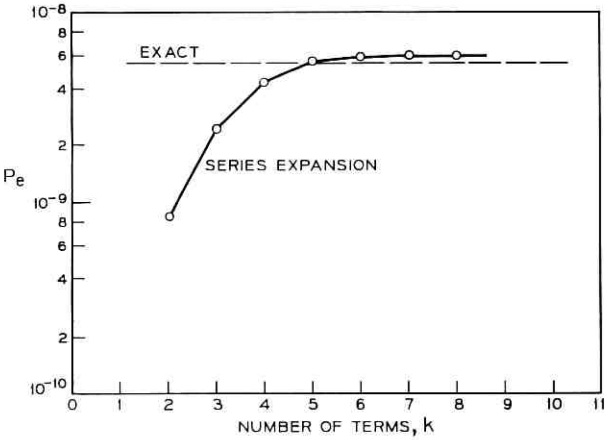


Fig. 4—Error probabilities versus number of terms in equation (9). Ideal band-limited signal. 11-pulse truncation approximation; sampling instant, $t = 0.05\ T$; (S/N) = 16 dB.

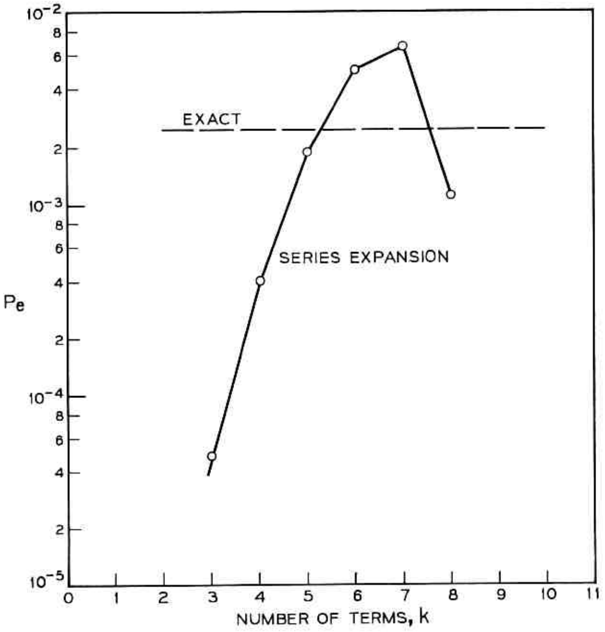


Fig. 5—Error probabilities versus number of terms in equation (9). Ideal band-limited signal. 11-pulse truncation approximation; sampling instant, $t = 0.2\ T$; (S/N) = 16 dB.

lowing assumptions. First, the information digits are identically and independently distributed. Second, the intersymbol interference converges absolutely. (For those pulses with absolutely divergent intersymbol interference, only finite truncated approximation of the real pulse will be used.) Three cases, which are representative of practical situations, are considered. The results show that this new method has a significant improvement in accuracy over Chernoff bound. For example, we consider the 2400-baud DDD option of the Data Set 203 operating over a channel having symmetrical delay distortion in excess of that of a worst-case C-2 line. A 5-tap mean-square equalizer is used by the receiver to equalize the channel. With a 14-dB input signal-to-noise ratio, the series expansion method provides a factor of 15 improvement over the Chernoff bound in estimating the error probability at the equalizer output.

The absolute convergence of the series expansion method is proved in Appendix A. An estimate of the terms required to reach the neighborhood of the true error probability is provided by equation (12). In

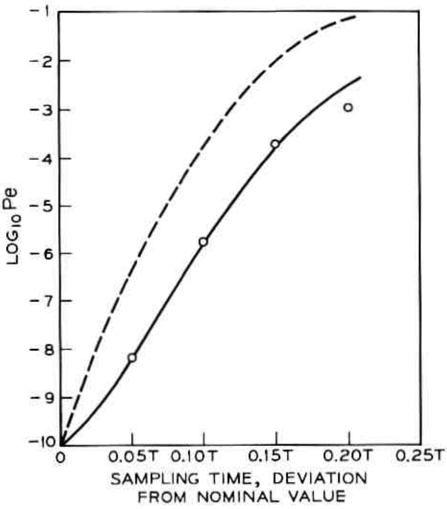


Fig. 6—Comparison of error probabilities obtained by Chernoff bound, exhaustive method, and series expansion method. Ideal band-limited signal (S/N) = 16 dB. [- - -Chernoff bound, ———— exhaustive method (11-pulse truncation), ooo series expansion (8-terms).]

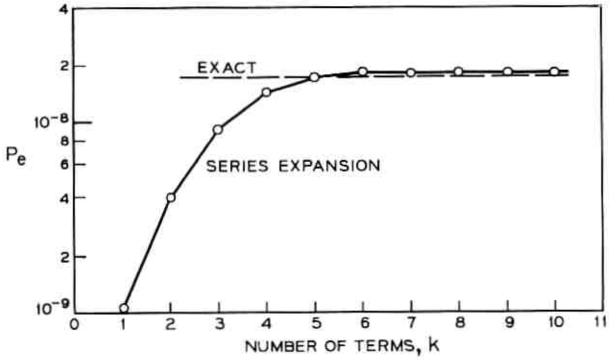


Fig. 7—Error probability versus number of terms in equation (9). Fourth-order Chebyshev pulse, 11-pulse truncation approximation; sampling instant, $t = 0.05$ T;(S/N) = 16 dB.

$$
\begin{aligned}
a(t) = {} & A_1 \text{ COS } (\omega_1 |\text{t}|/T - \phi_1) \text{ EXP } (-\alpha_1 |\text{t}|/T) \\
& + A_2 \text{ COS } (\omega_2 |\text{t}|/T - \phi_2) \text{ EXP } (-\alpha_2 |\text{t}|/T).
\end{aligned}
$$

$$
\begin{array}{ll}
A_1 = 0.4023, & A_2 = 0.7163, \\
\omega_1 = 2.839, & \omega_2 = 1.176, \\
\phi_1 = 0.7553, & \phi_2 = 0.1602, \\
\alpha_1 = 0.4587, & \alpha_2 = 1.107.
\end{array}
$$

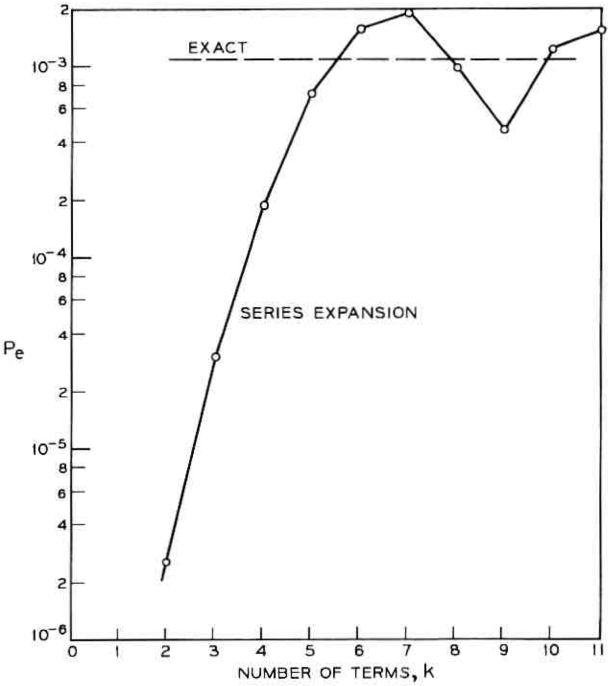


Fig. 8—Error probabilities versus number of terms in equation (9). Fourth-order Chebyshev pulse; 11-pulse truncation approximation; sampling instant, $t = 0.2\ T$; (S/N) = 16 dB.

actual systems, however, the true value is usually reached with only a small number of expansion terms. For example, in Fig. 3, the truncation error is less than 2×10^{-8} after taking into account the 9th term of the series expansion (which involves the 18th moment of the intersymbol interference); practically speaking, however, only three or four terms would be required for the series to converge in this example. In all the examples we considered, it is observed that a small error is assured by taking into account the first ten terms of the series.

The convergence is somewhat slower if the ratio of intersymbol interference to noise power $[q(t_0)]$ is large (see Section IV, case 2.), as indicated in Figs. 5 and 8. Under this condition, either the intersymbol interference is so bad that the system is not of practical interest, or the input signal-to-noise ratio is so high that the Chernoff bound already assures that the system performance is acceptable. For both cases, there is no need to evaluate the error probability.

For computation purposes every system must be approximated by a finite-memory-system. Since the computations involved in this new method increase only linearly with the length of the memory, a good approximation of the true channel may be obtained without excessive computation.

APPENDIX A

Convergence of the Series Expansion Method

In this Appendix, we shall prove that equation (9) is an absolutely convergent series. We know that

$$\begin{aligned} M_{2k} &= \int_{\text{all } X} X^{2k}\, dF(x) \\ &\leq \int_{\text{all } X} (\sup X)^{2k}\, dF(x), \\ &= \{\sum_{\ell \neq 0} | r(t_0 - \ell T) |\}^{2k} \qquad (20) \end{aligned}$$

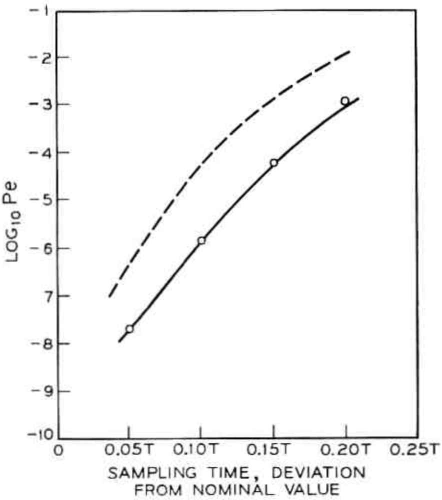


Fig. 9—Comparison of error probabilities obtained by Chernoff bound, exhaustive method, and series expansion method. Fourth-order Chebyshev pulse, (S/N) = 16 dB. [---Chernoff bound, ———— exhaustive method (11-pulse truncation), ooo series expansion (8-terms).]

and

$$H_{2K+1}(x) = (-1)^K 2^{K+\frac{1}{2}}[(2K-1)!!]\sqrt{2K+1} \cdot \exp\,(x^2/2) \cdot \left[\sin\,(\sqrt{4k+3}\,x) + O\left(\frac{1}{4\sqrt{K}}\right)\right]\cdot \qquad (21)^*$$

Hence,

$$| P_{e_{2K}} | \leqq \frac{(2K-3)!!}{(2K)!}\sqrt{2K-1}\,\frac{1}{\sqrt{2}}\cdot\left(\frac{1}{\sigma^2}\right)^K \frac{1}{\sqrt{\pi}} \cdot \exp\left[-\left(\frac{r^2(t_0)}{4\sigma^2}\right)\right]\cdot\{\sum_{\ell\neq 0} | r(t_0 - \ell T) | \},^{2K}$$

$$= S_{2K}\,. \qquad (22)$$

The ratio of S_{2K+2} to S_{2K} is given by

$$\frac{S_{2K+2}}{S_{2K}} = \frac{\sqrt{2K-1}}{(2K+2)\sqrt{2K+1}}\left[\frac{\sum_{\ell\neq 0} | r(t_0-\ell T) |}{\sigma}\right]^2. \qquad (23)$$

For K sufficiently large, equation (23) is always less than unity. Therefore equation (9) is an absolutely convergent series.

APPENDIX B

Derivation of the Recurrence Relations for the Moment of Intersymbol Interference

It has been shown that the intersymbol interference converges absolutely to a random variable[6] X. The characteristic function of the random variable X is given by,

$$\Phi(\omega) = \int_{\text{all } X} e^{j\omega X}\, dF(X),$$

$$= 1 + j\omega M_1 + \frac{(j\omega)^2}{2!}M_2 + \cdots + \frac{(j\omega)^k}{k!}M_k + \cdots\,. \qquad (24)$$

Therefore, we obtain

* See Ref. 8.

$$\left.\begin{aligned} \Phi(0) &= 1 \\ \frac{d^2\Phi(\omega)}{d\omega^2}\bigg|_{\omega=0} &= \Phi^2(0) = -M_2 \,, \\ &\vdots \\ \frac{d^{2k}\Phi(\omega)}{d\omega^{2k}}\bigg|_{\omega=0} &= \Phi^{2k}(0) = (-1)^k M_{2k} \,, \\ &\vdots \end{aligned}\right\}. \tag{25}$$

Since a_ℓ's are identically and independently distributed random variables and with zero mean,

$$M_1 = M_3 = \cdots = M_{2k+1} = \cdots = 0 \quad \text{for} \quad k = 0, 1, 2, \cdots , \tag{26}$$

and

$$\Phi(\omega) = \prod_{\ell=1}^{N} \cos \omega r(t_0 - \ell T), \tag{27}$$

where a truncated N-pulse-train approximation of the channel impulse response is assumed.

The even-order moments could be obtained by differentiating equation (27) $2k$ times, but the right hand side expressions could become untractable. However, if we differentiate equation (27) once and regroup the terms, we obtain the following,

$$\begin{aligned} \Phi'(\omega) &= -\left[\sum_{\ell=1}^{N} r(t_0 - \ell T) \tan \omega r(t_0 - \ell T) \right] \cdot \Phi(\omega), \\ &= -f(\omega) \cdot \Phi(\omega). \end{aligned} \tag{28}$$

By successive differentiation of equation (28), a recurrence relation can now be obtained. Differentiating equation (28) $2k - 1$ times, we obtain

$$\Phi^{2k}(0) = -\left\{ \sum_{i=1}^{k} \binom{2k-1}{2i-1} \Phi(0)^{2(k-i)} f^{2i-1}(0) \right\}, \tag{29}$$

where

$$f^{2i-1}(0) = \frac{d^{2i-1}}{d\omega^{2i-1}} f(\omega) \bigg|_{\omega=0} . \tag{30}$$

The power series expansion of $\tan \omega r(t_0 - \ell T)$ around origin is

$$\tan \omega r(t_0 - \ell T) = \omega r(t_0 - \ell T) + \frac{(\omega r(t_0 - \ell T))^3}{3!} + \cdots + \frac{2^{2k}(2^{2k} - 1)}{(2k)!} \mid B_{2k} \mid (\omega r(t_0 - \ell T))^{2k-1} + \cdots, \tag{31}$$

where B_{2k} is the Bernoulli number. It can be seen that

$$\frac{d^k}{d\omega^k} \tan \omega r(t_0 - \ell T) \Big|_{\omega=0} = [r(t_0 - \ell T)]^k \frac{2^{k+1}(2^{k+1} - 1)}{(k+1)} \mid B_{k+1} \mid, \quad \text{for } k = \text{odd positive integers,} \tag{32a}$$

$$= 0, \quad \text{for } k = \text{even positive integers.} \tag{32b}$$

Thus,

$$f^k(0) = \frac{d^k}{d\omega^k} f(\omega) \Big|_{\omega=0} = \frac{2^{k+1}(2^{k+1} - 1)}{(k+1)} \mid B_{k+1} \mid \cdot \lambda_{k+1}, \quad \text{for } k = \text{odd positive integers,} \tag{33a}$$

$$= 0, \quad \text{for } k = \text{even positive integers.} \tag{33b}$$

where

$$\lambda_{k+1} = \sum_{\ell=1}^{N} [r(t_0 - \ell T)]^{k+1}. \tag{33c}$$

Since

$$M_{2k} = (-1)^k \Phi^{2k}(0). \tag{34}$$

Combining equations (34) and (29), we obtain the recurrence relation for M_{2k},

$$M_{2k} = -\left\{ \sum_{i=1}^{k} \binom{2k-1}{2i-1} (-1)^i M_{2(k-i)} f^{2i-1}(0) \right\} \tag{35}$$

where $f^{2i-1}(0)$'s are given by equation (33a).

Knowing that $M_0 = 1$, all the higher order moments can be obtained via equation (35) without the knowledge of $dF(x)$.

REFERENCES

1. Aein, J. M., and Hancock, J. C., "Reducing the Effects of Intersymbol Interference With Correlation Receivers," IEEE Trans. Information Theory, *IT-9*, No. 3 (July 1963), pp. 167–175.

2. Aaron, M. R., and Tufts, D. W., "Intersymbol Interference and Error Probability," IEEE Trans. Information Theory, *IT-12*, No. 1 (January 1966), pp. 26–34.
3. Lucky, R. W., Salz, J., and Welson, E. J., Jr., *Principle of Data Communication,* New York: McGraw-Hill Book Company, 1968, p. 65.
4. Saltzberg, B. R., "Intersymbol Interference Error Bounds With Application to Ideal Bandlimited Signaling," IEEE Trans. Information Theory, *IT-14*, No. 4 (July 1968), pp. 563–568.
5. Lugannani, R., "Intersymbol Interference and Probability of Error in Digital System," IEEE Trans. Information Theory, *IT-15*, No. 6 (November, 1969), pp. 682–688.
6. Loève, M., *Probability Theory,* 2nd ed., Princeton, New Jersey: Van Nostrand, 1960.
7. Holzman, L. W., and Lawless, W. J., "Data Set 203, A New High-speed Voiceband Modem," Conference Record, 1970 ICC, pp. 12-7–12-12.
8. Gradshteyn, I. S., and Ryzhik, I. M., *Table of Integrals and Series and Products,* New York and London: Academic Press, 1965.

Upper Bound on the Efficiency of dc-Constrained Codes

By TA-MU CHIEN

(Manuscript received June 2, 1970)

We derive the limiting efficiencies of dc-constrained codes. Given bounds on the running digital sum (RDS), the best possible coding efficiency η, for a K-ary transmission alphabet, is $\eta = \log_2 \lambda_{\max}/\log_2 K$, where $\lambda_{\max}$ is the largest eigenvalue of a matrix which represents the transitions of the allowable states of RDS. Numerical results are presented for the three special cases of binary, ternary and quaternary alphabets.

I. INTRODUCTION

In digital transmission systems, the transmission channel often does not pass dc. This causes the well-known problem of baseline wander. One way to overcome this difficulty is to restrict the dc content in the signal stream using suitably devised codes.[1–3] As a result many codes having a dc-constrained property have been studied.[4–9] The coding requirement is represented by the constraint put upon the running digital sum (RDS) of the coded signal stream. We expect that the efficiency of a dc-constrained code is related to the limits of RDS in some definite way. This is the subject to which we address ourselves in this paper. More specifically, we intend to answer the question: What is the best possible efficiency of any dc-constrained code satisfying a given limit on RDS?

Let $\{a_1, a_2, \cdots\}$ be the sequence of the transmitted symbols, the RDS of the signal stream at instant k is defined to be the sum $\sum_{i=1}^{k} a_i$. Taking the RDS at any instant as the state of the signal stream at that point, the limits on RDS define a set of allowable states, and each additional signal symbol may be considered as a transition from one state to another. This transition can be represented by a matrix-called naturally the transition matrix. For a K-ary signal alphabet, the best possible efficiency η of dc-constrained codes is found to be

$$\eta = \frac{\log_2 \lambda_{\max}}{\log_2 K} \tag{1}$$

where $\lambda_{\max}$ is the largest eigenvalue of the transition matrix.

The efficiency of a code is defined to be the ratio of the average bits per symbol of the coded signal stream to that of the random (uncoded) signal stream.

McCullough[4] has derived the same result (1) for the special cases of $K = 2$ and 3. His approach is quite different from what will be presented in the sequel.

We first describe in detail the construction of a mathematical model for the case of a binary alphabet. Then we generalize the result of the binary case to include any alphabet set. Methods of effecting numerical calculation are discussed as well as approximation formulas. The numerical results for three important cases are presented and known codes are compared with the theoretical limits.

II. LIMITING EFFICIENCY OF THE BINARY CODES

In this section we confine our discussion to binary signals and direct our attenion to the intuitive reasoning which leads to the construction of a simple mathematical model and its interpretation.

Let M, a positive integer, be the desired bound on the RDS of the coded binary signal stream. This defines a subset $S_M(\infty)$ of the set $S(\infty)$ of all infinite binary sequences in the following way: An infinite sequence is in the subset $S_M(\infty)$ if the RDS of the sequence is nowhere larger than M or less than $-M$, i.e., $|\sum_{i=1}^{k} a_i| \leqq M$ for $k = 1, 2, \cdots$. A sequence in $S_M(\infty)$ is called an allowable sequence. Denoting by $N_M(\infty)$ and $N(\infty)$ the number of infinite sequences in $S_M(\infty)$ and $S(\infty)$ respectively, the average information per symbol for the sequences in $S_M(\infty)$ is given by

$$\eta = \frac{\log_2 N_M(\infty)}{\log_2 N(\infty)}, \tag{2}$$

assuming the ratio exists. If we interpret the set $S(\infty)$ as source data and $S_M(\infty)$ as the transmitted signal, then η defined in equation (2) is the efficiency of a dc-constrained code which maps one-to-one from $S(\infty)$ onto $S_M(\infty)$.*

Clearly for any code which satisfies the requirement that RDS be bounded by M, the coded signal stream must be a member of $S_M(\infty)$. Therefore, the set of allowable infinite sequences defined by any code satisfying the desired constraint on RDS must be a subset of $S_M(\infty)$.

* The puzzle of mapping a large set to a small set can be cleared mathematically by observing that the cardinality of both $S(\infty)$ and $S_M(\infty)$ are that of a continuum, and physically by demanding that the transmitter has a higher baud than the source.

Thus we conclude that the formal expression in (2) indeed gives the best possible efficiency for a given bound M. Our next step is to find a way to count the number of allowable sequences in $S_M(\infty)$.

Let us start by counting the allowable sequences of finite length L. Define an occupancy vector of the allowable states $\mathbf{u}_L$, $'$ denoting the transpose of a vector (or matrix),

$$\mathbf{u}_L = [u_{-M} \cdots u_0 \cdots u_{+M}]', \tag{3}$$

where u_k, $k = -M, \cdots, M$, is the number of allowable sequences of length L with their RDS at end equal to k, i.e., $\sum_{i=1}^{L} a_i = k$. The total number of allowable sequences of length L, $N_M(L)$ is simply

$$N_M(L) = \sum_{k=-M}^{M} u_k . \tag{4}$$

As $L \to \infty$, $N_M(L) \to N_M(\infty)$ and the total number of sequences of length L, $N(L) = 2^L \to N(\infty)$. Hence we can rewrite (2) as

$$\eta = \lim_{L\to\infty} \frac{\log_2 N_M(L)}{L}. \tag{5}$$

Now our job is to find a formula for the number of allowable sequences of finite length.

Suppose we know the occupancy vector $\mathbf{u}_L$ and we want to calculate the occupancy vector $\mathbf{u}_{L+1}$. Clearly for any allowable sequence of length $L + 1$, its first L elements must be one of the allowable sequences of length L. We generate, therefore, the allowable sequences of length $L + 1$ from that of length L by adding one more binary symbol (+1 or −1). Therefore, the sequences of length $L + 1$ in the $-M$th state are generated by adding −1 to the sequences of length L in the $-M +$ 1st state; the sequences of length $L + 1$ in the $-M +$ 1st state are generated by adding +1 to the sequences of length L in the $-M$th state and by adding −1 to the sequences of length L in the $-M +$ 2nd state; etc. It is not difficult to see that the new state occupancy vector is

$$\mathbf{u}_{L+1} = \begin{bmatrix} u_{-M+1} \\ u_{-M} + u_{-M+2} \\ u_{-M+1} + u_{-M+3} \\ \vdots \qquad \vdots \\ u_{M-2} + u_M \\ u_{M-1} \end{bmatrix}. \tag{6}$$

Equivalently, $\mathbf{u}_{L+1}$ can be written as

$$\mathbf{u}_{L+1} = A_{2M+1}\mathbf{u}_L \tag{7}$$

where

$$A_{2M+1} = \begin{bmatrix} 0 & 1 & & & & 0 \\ 1 & 0 & 1 & & & \\ & 1 & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & 1 \\ & 0 & & & 1 & 0 \end{bmatrix} \tag{8}$$

is a square matrix of size $2M + 1$ with ones in the superdiagonal and the subdiagonal and zeros elsewhere. A_{2M+1} is the transition matrix of the allowable states. By the same reasoning, we have,

$$\begin{aligned} \mathbf{u}_L &= A_{2M+1}\mathbf{u}_{L-1} \\ \mathbf{u}_{L-1} &= A_{2M+1}\mathbf{u}_{L-2} \\ &\vdots \end{aligned} \tag{9}$$

and

$$\mathbf{u}_1 = A_{2M+1}\mathbf{u}_0$$

where $\mathbf{u}_0$ is the occupancy vector of the sequence of zero length. It is defined naturally with one at the zeroth state and zeros elsewhere,

$$\mathbf{u}_0 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{10}$$

From equation (9), we obtain, by successive substitution,

$$\mathbf{u}_L = A_{2M+1}^L\mathbf{u}_0 \ . \tag{11}$$

The total number of allowable sequences, from equation (4), is

$$N_M(L) = \mathbf{1}' A_{2M+1}^L u_0 \tag{12}$$

where

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{13}$$

The class of matrices A_n defined in equation (8) has many interesting properties. Their investigation is relegated to the Appendix.

Using the result derived in the Appendix, and adapting the following ordering or eigenvalues of A_{2M+1}:

$$\lambda_{-M} < \lambda_{-M+1} < \cdots < \lambda_{M-1} < \lambda_M ,$$

we can rewrite equation (12),

$$N_M(L) = \mathbf{1}' P D_A^L P' \mathbf{u}_0 \tag{14}$$

where

$$D_A = \begin{bmatrix} \lambda_{-M} & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{bmatrix}, \tag{15}$$

$$P = \begin{bmatrix} \phi_0(\lambda_{-M})/\phi(\lambda_{-M}) & \cdots & \phi_0(\lambda_M)/\phi(\lambda_M) \\ \phi_1(\lambda_{-M})/\phi(\lambda_{-M}) & \cdots & \phi_1(\lambda_M)/\phi(\lambda_M) \\ \vdots & & \vdots \\ \phi_{2M}(\lambda_{-M})/\phi(\lambda_{-M}) & & \phi_{2M}(\lambda_M)/\phi(\lambda_M) \end{bmatrix} \tag{16}$$

from Lemma 7 of the Appendix and $\phi(\lambda)$ and $\phi_i(\lambda)$ are defined in equation (54) and (70). By straightforward multiplication, we can write, from equation (14),

$$N_M(L) = \sum_{i=-M}^{M} \lambda_i^L \phi_{M+1}(\lambda_i) \sum_{j=0}^{2M} \phi_j(\lambda_i), \tag{17}$$

where the normalization constants $\phi(\lambda_i)$ are omitted for simplicity. Denote by $\lambda_{\max}$ the largest absolute value of the eigenvalues of A_n, and from Lemmas 3 and 6 of the Appendix we know that

$$\lambda_{\max} = \lambda_M = -\lambda_{-M} . \tag{18}$$

Then from equation (17) and Lemma 2,

$$
\begin{aligned}
N_M(L) &= \lambda_M^L\left\{\phi_{M+1}(\lambda_M)\sum_{j=0}^{2M}\phi_j(\lambda_M) + (-1)^L\phi_{M+1}(\lambda_{-M})\sum_{j=0}^{2M}\phi_j(\lambda_{-M})\right\} \\
&\qquad + \sum_{i=-M+1}^{M-1}\lambda_i^L\phi_{M+1}(\lambda_i)\sum_{j=0}^{2M}\phi_j(\lambda_i) \\
&= \lambda_M^L\phi_{M+1}(\lambda_M)\sum_{j=0}^{2M}\left[1+(-1)^{L+M+j+1}\right]\phi_j(\lambda_M) \\
&\qquad + \sum_{i=-M+1}^{M-1}\lambda_i^L\phi_{M+1}(\lambda_i)\sum_{j=0}^{2M}\phi_j(\lambda_i). \qquad (19)
\end{aligned}
$$

Since $\phi_j(\lambda_M) > 0$ for all j (see the proof of Lemma 4), the coefficient of the λ_M^L term in equation (19),

$$
\phi_{M+1}(\lambda_M)\sum_{j=0}^{2M}\left[1+(-1)^{L+M+j+1}\right]\phi_j(\lambda_M) > 0 \qquad (20)
$$

independent of L.

Substituting equation (19) in (5) and using (18), we have

$$
\begin{aligned}
\eta &= \lim_{L\to\infty}\frac{1}{L}\left\{\log_2\lambda_{\max}^L\left[z_{\max}+\sum_{i=-M+1}^{M-1}\left(\frac{\lambda_i}{\lambda_{\max}}\right)^L z_i\right]\right\} \\
&= \log_2\lambda_{\max} + \lim_{L\to\infty}\frac{1}{L}\log_2\left[z_{\max}+\sum_{i=-M+1}^{M-1}\left(\frac{\lambda_i}{\lambda_{\max}}\right)^L z_i\right] \qquad (21)
\end{aligned}
$$

where $z_{\max}$ and z_i are the coefficients of λ_M and λ_i, $i \neq \pm M$ in equation (19). The second term in equation (21) approaches zero as a limit since $z_{\max} > 0$. Thus we have the desired result for the binary case

$$
\eta = \log_2\lambda_{\max}. \qquad (22)
$$

Actually, we have proved a result more general than (22). Observe that, in passing to the limit, the crucial point is that $z_{\max}$ in (21) be nonzero. From equation (20) and the fact that $\phi_i(\lambda_M) \neq 0$ for $i \leqq 2M$, we conclude that the particular $\mathbf{u}_0$ we use, though natural, is immaterial, and any vector with non-negative coordinates will serve the purpose. Observe also the actual values of the allowable RDS state nowhere enter into our discussion, hence, it is immaterial whether the bound on RDS be symmetric or not. We can consolidate our discussion by stating the following theorem.

Theorem 1: For a binary alphabet, if the RDS of a coded signal stream is required to be within some bound M^+ and M^-, where M^+ and

M^- are integers, then the best possible coding efficiency is given by

$$\eta = \log_2 \lambda_{\max}$$

where $\lambda_{\max}$ is the largest positive eigenvalue of the transition matrix A_n of size $n = M^+ - M^- + 1$ as defined in equation (8).

III. GENERALIZATION TO K-ary CODES

We now wish to extend the result derived in the previous section to an arbitrary K-ary alphabet set, $\{\alpha_1, \cdots, \alpha_K\}$. We shall restrict ourselves to symmetric alphabets. Namely, if K is even, α_i takes on the values $-(K-1), -(K-3), \cdots, -1, +1, \cdots, (K-1)$; if K is odd, α_i takes on the values $-(K-1)/2, -(K-2)/2, \cdots, -1, 0, 1, \cdots, (K-1)/2$. The transition matrix of allowable states is then given by

$$A_n = \sum_{\alpha_i \geqq 0} H_n^{\alpha_i} + \sum_{\alpha_i < 0} F_n^{-\alpha_i} \tag{23}$$

where the size of the matrices is

$$n = M^+ - M^- + 1, \tag{24}$$

and M^+ and M^- are the desired upper and lower bound on RDS. If $\alpha_i = 0$ is a member of the alphabet, we follow the usual convention that $H_n^0 = 1_n$. The matrices

$$H_n = \begin{bmatrix} 0 & 1 & & & 0 \\ & 0 & 1 & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ 0 & & & \cdot & 1 \\ & & & & 0 \end{bmatrix} \tag{25}$$

and

$$F_n = \begin{bmatrix} 0 & & & & 0 \\ 1 & 0 & & & \\ & 1 & \cdot & & \\ & & \cdot & \cdot & \\ 0 & & & \cdot & \cdot \\ & & & 1 & 0 \end{bmatrix} \tag{26}$$

are known as superdiagonal and subdiagonal matrices respectively. To see that A_n given in equation (23) is indeed the transition matrix, we observe that each symbol α_i will generate a sequence in any allowable state to a state α_i unit away. Each term in (23) represents the transition of states due to a particular alphabet. As an example, taking the quaternary alphabet set $\{-3, -1, +1, +3\}$, the transition matrix is

$$A_n = \begin{bmatrix} 0 & 1 & 0 & 1 & & & 0 \\ 1 & 0 & 1 & 0 & 1 & & \\ 0 & 1 & 0 & 1 & \cdot & \cdot & \\ 1 & 0 & 1 & \cdot & \cdot & \cdot & 1 \\ & 1 & \cdot & \cdot & \cdot & \cdot & 0 \\ & & \cdot & \cdot & \cdot & \cdot & 1 \\ 0 & & & 1 & 0 & 1 & 0 \end{bmatrix}.$$

With these preliminaries out of the way, we now state a general result on the limiting efficiency of dc-constrained codes:

Theorem 2: If the RDS of the coded K-ary signal stream is required to be within some bound M^+ and M^-, then the best possible coding efficiency is given by

$$\eta = \frac{\log_2 \lambda_{\max}}{\log_2 K}, \tag{27}$$

where $\lambda_{\max}$ is the largest eigenvalue of the transition matrix A_n defined by equations (23) and (24).

Before we embark on the proof of Theorem 2, we need to establish an important auxiliary result.

Let $N_M(L)$ denote again the number of allowable K-ary sequences, then the limiting efficiency η, corresponding to equation (5), is

$$\eta = \lim_{L \to \infty} \frac{\log_2 N_M(L)}{L \log_2 K}. \tag{28}$$

In the set of allowable sequences $S_M(L)$, we can define a subset $S_{M|\alpha_i}(L)$ by restricting the first symbol to be α_i. Similarly we define a subset $S_{M|\alpha_i}, - \alpha_i(L)$ by restricting the first two symbols to be α_i and $-\alpha_i$. Clearly

$$S(L) \supset S_M(L) \supset S_{M|\mathrm{L}_i}(L) \supset S_{M|\alpha_i,-\alpha_i}(L), \tag{29}$$

and it follows that

$$\eta \geqq \eta_{\alpha_i} \geqq \eta_{\alpha_i,-\alpha_i} \tag{30}$$

where η_{α_i} and $\eta_{\alpha_i,-\alpha_i}$ are the limiting efficiencies given by equation (28) with the additional restriction on the leading elements.

Considering now all the sequences in $S_{M|\alpha_i,-\alpha_i}(L+2)$, it is not difficult to see that the number of sequences in $S_{M|\alpha_i,-\alpha_i}(L+2)$ is equal to that in $S_M(L)$. Hence the efficiency,

$$\begin{aligned}\eta_{\alpha_i,-\alpha_i} &= \lim_{L\to\infty} \frac{\log_2 N_M(L)}{\log_2 K^{L+2}}, \\ &= \lim_{L\to\infty} \frac{\log_2 N_M(L)}{(L+2)\log_2 K}, \\ &= \eta.\end{aligned} \tag{31}$$

Coupled with equation (30), we have shown that

$$\eta = \eta_{\alpha_i} = \eta_{\alpha_i,-\alpha_i}\,. \tag{32}$$

A little reflection should convince us that any finite pattern at the beginning of the sequences does not affect the limiting efficiency η. In other words, the limiting efficiency is independent of starting point —a fact we observed in the previous section after the detailed study of the transition matrix. This fact enables us to prove Theorem 2 without going through a tedious mathematical analysis.

Proof of Theorem 2: The matrix A_n defined in equation (23) is real and symmetric. It can be diagonalized by an orthogonal transformation, i.e.,

$$A_n = PD_AP',\ PP' = 1 \tag{33}$$

where D_A is a diagonal matrix of real elements $\lambda_1, \cdots, \lambda_n$.

Using any $\mathbf{u}_0$, a constant vector with non-negative elements, we can generate a sequence of vectors $\mathbf{u}_L$,

$$\mathbf{u}_L = A_n^L\mathbf{u}_0\,,\ \text{for } L = 1, 2, \cdots. \tag{34}$$

Since A_n is a matrix with non-negative elements, it is easy to see that all $\mathbf{u}_L$'s are vectors with non-negative elements. Write

$$P = [\mathbf{p}_1\mathbf{p}_2 \cdots \mathbf{p}_n] \tag{35}$$

where $\mathbf{p}_i$ is a column vector. From equation (34) we have, using $\mathbf{u}_0$ with only a 1 in jth position,

$$\begin{aligned}\mathbf{u}_L &= PD_A^LP'\mathbf{u}_0 \\ &= \sum_{i=1}^{n} \lambda_i^L p_{ji}\mathbf{p}_i\end{aligned} \tag{36}$$

where p_{ji} is the jth element in $\mathbf{p}_i$.

Let $\lambda_{\max}$ denote the absolute value of the largest eigenvalue of A_n and assume that, in general,*

$$\lambda_1 = \lambda_2 = \cdots = \lambda_r = \lambda_{\max}, \tag{37}$$

$$\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_{r+s} = -\lambda_{\max}.$$

We can rewrite (36)

$$\mathbf{u}_L = \lambda_{\max}^L \left\{ \sum_{i=1}^{r} p_{ji}\mathbf{p}_i + (-1)^L \sum_{i=r+1}^{r+s} p_{ji}\mathbf{p}_i + \sum_{i=r+s+1}^{n} \left(\frac{\lambda_i}{\lambda_{\max}}\right)^L p_{ji}\mathbf{p}_i \right\}. \tag{38}$$

Denote by $\mathbf{z}$ the first two sums in equation (38),

$$\mathbf{z} = \sum_{i=1}^{r} p_{ji}\mathbf{p}_i + (-1)^L \sum_{i=r+1}^{r+s} p_{ji}\mathbf{p}_i . \tag{39}$$

$\mathbf{z}$ must be non-negative for any j and L. If not, then for some large enough L, $\mathbf{u}_L$ will have negative elements, which is a contradiction. Since $\mathbf{z}$ is a linear combination of $\mathbf{p}_1, \cdots, \mathbf{p}_{r+s}$, a set of linearly independent vectors, $\mathbf{z} = \mathbf{0}$ only if $p_{ji} = 0$, $i = 1, \cdots, r + s$. Furthermore, if $p_{ji} = 0$ for all $j = 1, \cdots, n$, then the transformation matrix P has a row of zeros, which is again a contradiction. Thus we conclude that, for some choice of $\mathbf{u}_0$, i.e., for some j,

$$\mathbf{u}_L = \lambda_{\max}^L \left\{ \mathbf{z} + \sum_{i=r+s+1}^{n} \left(\frac{\lambda_i}{\lambda_{\max}}\right)^L p_{ji}\mathbf{p}_i \right\} \tag{40}$$

with $\mathbf{z}$ non-negative independent of L. The total number of allowable sequences is

$$N_M(L) = \lambda_{\max}^L \left\{ \mathbf{1}'\mathbf{z} + \sum_{i=r+s+1}^{n} \left(\frac{\lambda_i}{\lambda_{\max}}\right)^L p_{ji}\mathbf{1}'\mathbf{p}_i \right\}. \tag{41}$$

Substituting $N_M(\mathrm{L})$ in equation (28), and passing to limit, we get equation (27). The proof is now complete.

IV. NUMERICAL RESULTS AND DISCUSSIONS

4.1 *Numerical Calculation*

Using the digital computer, the calculation of $\lambda_{\max}$ of any transition matrix A_n is not difficult except maybe for large n. In the following, we discuss several alternative approaches to evaluating $\lambda_{\max}$, and we present results for three important cases.

* It can be shown that $r = 1$ and $s = 0$ or 1. But the proof that follows does not require this fact.

(*i*) Find $\lambda_{\max}$ by direct diagonalization of the matrix A_n. There are computer programs developed for this purpose. This is done for the binary case and the quaternary case, and the limiting efficiency η is plotted (solid curve) as a function of allowable states n in Figs. 1 and 2 respectively.

(*ii*) In the binary case, the characteristic polynomial $\phi_n(\lambda)$ of A_n satisfies a simple recursive relation (56). Treating (56) as a difference equation of $\phi_n(\lambda)$'s, one can express $\phi_n(\lambda)$ in an alternate form:[4]

$$\phi_n(\lambda) = \frac{\sin[(m+1)\cos^{-1}(\lambda/2)]}{\sin[\cos^{-1}(\lambda/2)]}. \tag{42}$$

The roots of $\phi_n(\lambda)$ are, as easily seen in equation (42),

$$\lambda_k = 2\cos\frac{k\pi}{n+1}, \qquad k = 1, \cdots, n. \tag{43}$$

(*iii*) The ternary case can be reduced to the binary case by replac-

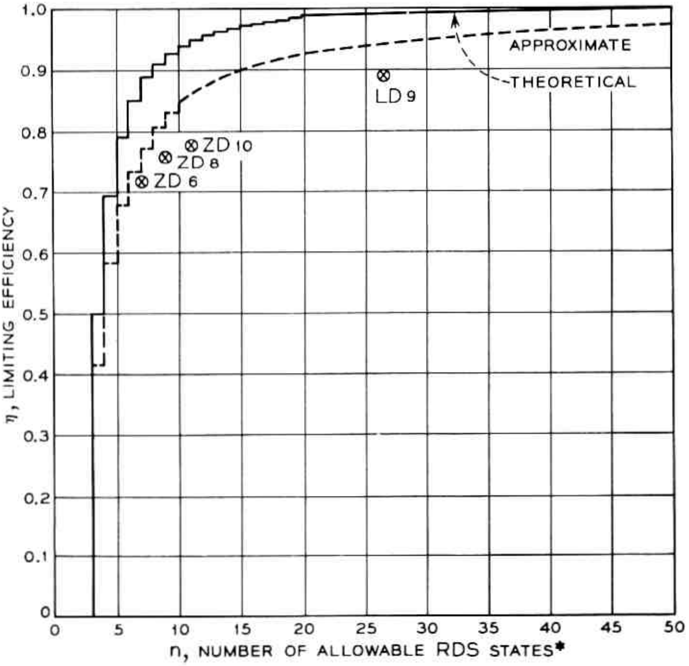


Fig. 1—Limiting efficiency vs allowable states binary alphabet (+1, −1). *$n = M^+ - M^- + 1$, where M^+ and M^- are the upper and lower bound of the RDS.

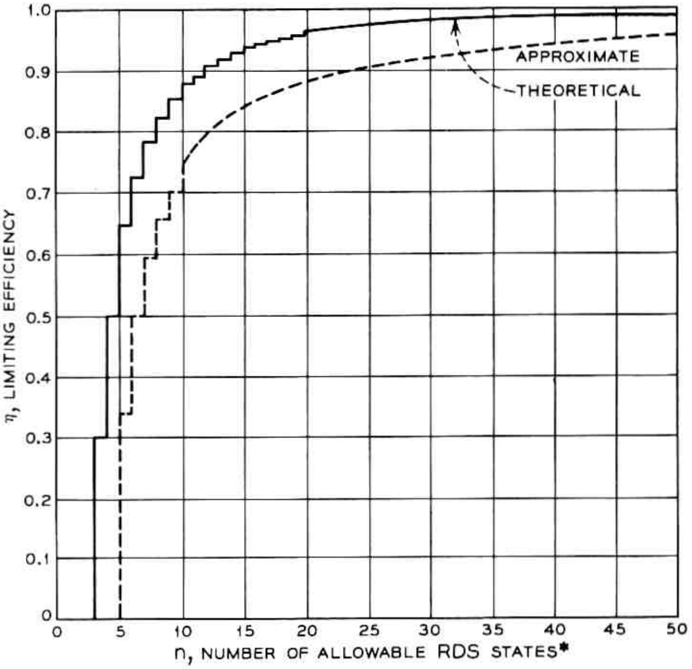


Fig. 2—Limiting efficiency vs allowable states quarternary alphabet (+3, +1, −1, −3). $*n = M^+ - M^- + 1$, where M^+ and M^- are the upper and lower bound of RDS.

ing λ in $\phi_n(\lambda)$ by $\lambda + 1$. Therefore, one gets $\lambda_{\max}$ of the ternary case by adding 1 to the corresponding (same n) $\lambda_{\max}$ of the binary case. The top curve in Fig. 3 is plotted in this way.

(*iv*) From the well-known formula[10]

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} \mathbf{x}' A_n \mathbf{x} \tag{44}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \tag{45}$$

and the norm of a vector $||x||$, is

$$||x|| = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \tag{46}$$

we have

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} \left\{ \sum_{\alpha_i \geq 0} \sum_{i=1}^{n-\alpha_i} x_i x_{i+\alpha_i} + \sum_{\alpha_i < 0} \sum_{i=1}^{n+\alpha_i} x_i x_{i-\alpha_i} \right\} \tag{47}$$

where the α_i's are members in the alphabet set. For example,

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} 2 \sum_{i=1}^{n-1} x_i x_{i+1}, \tag{48}$$

in the binary case;

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} \left\{ \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i x_{i+1} \right\} \tag{49}$$

in the ternary case; and

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} 2 \left\{ \sum_{i=1}^{n-1} x_i x_{i+1} + \sum_{i=1}^{n-3} x_i x_{i+3} \right\} \tag{50}$$

in the quaternary case.

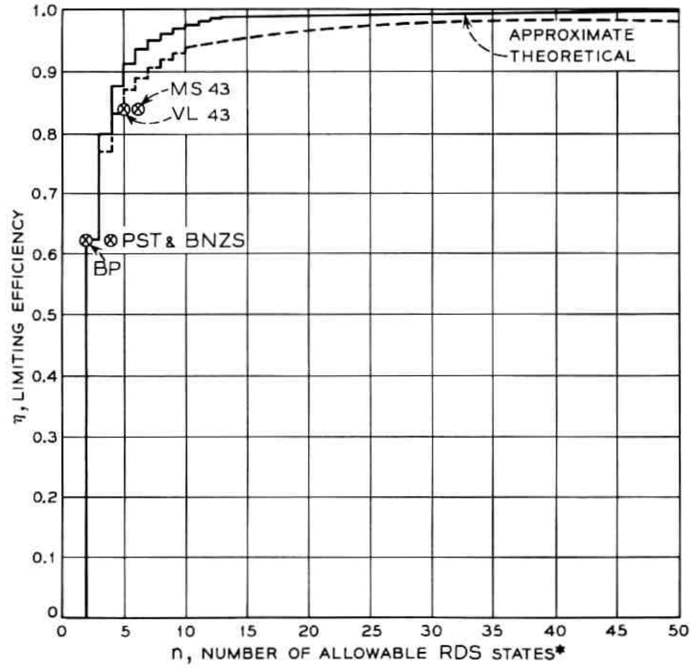


Fig. 3—Limiting efficiency vs allowable states ternary alphabet (+1, 0, −1). *$n = M^+ - M^- + 1$, where M^+ and M^- are the upper and lower bound of the RDS.

In this formulation, $\lambda_{\max}$ becomes the extreme value of a quadratic form with an equality constraint. There are a number of ways to effect a numerical solution.

4.2 *Approximation Formula*

To search for $\lambda_{\max}$ using equation (47) is not an easy alternative, but it leads to an estimation of $\lambda_{\max}$. If we let $x_1 = x_2 = \cdots = x_n$, then, from equation (47), we have

$$\lambda_{\max} \geqq \sum_{\alpha_i \geqq 0} \frac{n - \alpha_i}{n} + \sum_{\alpha_i < 0} \frac{n + \alpha_i}{n}. \tag{51}$$

Any other choice of the x_i's will lead to a different estimate of $\lambda_{\max}$ which may be better or worse than that of equation (51). We justify the present choice by noting the simplicity of equation (51). Using equation (51), we obtain an approximation formula for the limiting efficiency η,

$$\eta = \frac{\log_2 \left(\sum_{\alpha_i \geqq 0} \frac{n - \alpha_i}{n} + \sum_{\alpha_i < 0} \frac{n + \alpha_i}{n} \right)}{\log_2 K} \tag{52}$$

where α_i's are members of the K-ary alphabet set. The approximate η are also plotted in Figs. 1 to 3 (dashed curve).

As expected, the approximation is reasonably good for large n and it's for large n that we may have to rely on the approximation formula.

4.3 *Discussion*

(i) It is of some interest to see how the efficiencies of various codes with the dc-constrained property compare with the limiting curves. We have located the following known codes:

ZDN (Zero-Disparity Binary Code of Block Length N)[5] and
LDN (Low-Disparity Binary Code of Block Length N)[6] in Fig. 1;
and
BP (Bipolar Code)[1], $n = 2$, $\eta = 0.63^{\dagger}$,
PST (Paired Selected Ternary Code)[7], $n = 4$, $\eta = 0.63$,
BNZS (Bipolar with N Zero Substitution Codes)[8],
$n = 4$, $\eta = 0.63$,
VL43 (Variable Length Ternary Code)[9], $n = 5$, $\eta = .84$, and
MS43 (Fixed Length Ternary Code)[9], $n = 6$, $\eta = 0.84$, in Fig. 3.

* $N = 6$. The allowable states for BP are -1 and 0 or 0 and $+1$ depending upon whether the first pulse transmitted be -1 or $+1$. A similar situation exists for PST, BNZS.

The BP and BNZ* codes are used in T-1 and T-2 systems[1,2] respectively and PST is used in the experimental T-4 system.[3] In comparison with their limiting efficiencies, one gets the impression that, barring the fact that these codes have other properties in addition to the dc-constraint, there is some room for improving the coding efficiency. VL43, and MS43 are examples along this direction.

It should be pointed out* that the real engineering problem is to control baseline wander. A true comparison of different coding schemes should therefore be done on this basis. The relation between RDS and baseline wander is an elusive one. It depends upon the detailed structure of the code in question and the channel to which the signal is applied. In terms of RDS, it depends not only on the bounds and the distribution of the allowable RDS but also on its dynamics. By dynamics we mean the "speed" of moving from one state to another, the dynamic behavior is of importance because the channel has "failing memory", so to speak. For example, it can be shown[11] under certain conditions that a quick jump from one extreme state to the other results in a larger amount of baseline wander than would occur from staying in an extreme state for a long time.

(*ii*) As a general observation, the limiting curves saturate rapidly. This implies that, to make a high efficiency code possible, a physical system should be designed to operate beyond the fast rising portion of the curves. It would be reasonable then to expect that a simple ternary block code could be found with 90 percent efficiency or better for, say, $n = 15$.

(*iii*) An interesting question which arises naturally in connection with the limiting efficiency is its realizability. If one accepts infinite delay, the answer is affirmative. If one thinks in terms of block codes of finite length, then the limiting efficiency cannot be realized.

V. CONCLUSION

We have shown that, for a dc-constrained code, the limiting efficiency is related to the number of allowable RDS states in a very simple way. The result is effective in the sense that it lends itself easily to numerical evaluation.

The underlying mathematical fact in our proof is the property of non-negative matrices and vectors. Using the theorem of Frobenius on non-negative matrices,[10] our result can be proved in a few steps.

* The discussion here is heuristic in nature. A thorough treatment of the subject is beyond the scope of this paper.

We retain our approach for the reasons that it only requires elementary knowledge of matrix theory and it gives more insight to the problem.

The technique developed in this paper can be used to investigate bounds for some other classes of codes, say timing codes. This will be done elsewhere.

VI. ACKNOWLEDGMENTS

The author would like to express his thanks to V. I. Johannes and P. E. Rubin for their challenging comments and useful suggestions and to R. W. Kolor for his effort and interest in calculating the numerical results.

APPENDIX

The class of matrices which we want to investigate has the following general form*

$$A_n = \begin{bmatrix} 0 & 1 & & & 0 \\ 1 & 0 & 1 & & \\ & 1 & \cdot & \cdot & \\ & & \cdot & \cdot & 1 \\ 0 & & & 1 & 0 \end{bmatrix}. \tag{53}$$

Each matrix A_n is a square matrix of size n and it has ones in the super- and sub-diagonal and zeros elsewhere.

Let $\phi_n(\lambda)$ denote the characteristic polynomial of A_n. By definition,

$$\begin{aligned} \phi_n(\lambda) &\triangleq \det\,[\lambda 1_n - A_n] \\ &= \det \begin{bmatrix} \lambda & -1 & & & 0 \\ -1 & \lambda & -1 & & \\ & -1 & \cdot & \cdot & \\ & & \cdot & \cdot & -1 \\ 0 & & & -1 & \lambda \end{bmatrix}. \end{aligned} \tag{54}$$

The first few polynomials $\phi_n(\lambda)$ are

* A special class of Jacobi matrix.[12]

$$\begin{aligned}\phi_0(\lambda) &= 1,\\ \phi_1(\lambda) &= \lambda,\\ \phi_2(\lambda) &= \lambda^2 - 1,\\ \phi_3(\lambda) &= \lambda^3 - 2\lambda,\\ \phi_4(\lambda) &= \lambda^4 - 3\lambda^2 + 1,\end{aligned} \tag{55}$$

where $\phi_0(\lambda)$ is defined to be 1. The polynomials $\phi_n(x)$ have some interesting properties. We state them as lemmas.

Lemma 1: $\phi_n(x)$ satisfies the following recursive relations:
For $n \geqq 2$

$$\phi_n(\lambda) = \lambda\phi_{n-1}(\lambda) - \phi_{n-2}(\lambda), \tag{56}$$

and for $n > m$,

$$\phi_n(\lambda) = \phi_m(\lambda)\phi_{n-m}(\lambda) - \phi_{n-1}(\lambda)\phi_{n-m-1}(\lambda). \tag{57}$$

Proof: To prove equation (56), we expand the determinant in (54) with respect to the first column. To prove (57), we expand the determinant with respect to the first m columns and observe that the only two nonzero products of minors are of size m and $n - m$.

Lemma 2: ϕ_n (λ) is an even (odd) polynomial if n is even (odd), i.e.,

$$\phi_n(\lambda) = (-1)^n\phi_n(-\lambda). \tag{58}$$

Proof: Assume that (58) is true for $\phi_{n-1}(\lambda)$ and $\phi_{n-2}(\lambda)$, then by (56),

$$\begin{aligned}\phi_n(-\lambda) &= -\lambda\phi_{n-1}(-\lambda) - \phi_{n-2}(-\lambda),\\ &= (-1)^n\lambda\phi_{n-1}(\lambda) - (-1)^{n-2}\phi_{n-2}(\lambda),\\ &= (-1)^n[\lambda\phi_{n-1}(\lambda) - \phi_{n-2}(\lambda)],\\ &= (-1)\phi_n(\lambda).\end{aligned}$$

Since (58) is true for $\phi_0(\lambda)$ and $\phi_1(\lambda)$ by inspection of (55), (58) is true for any n by induction.

Lemma 3: If λ_0 is a root of $\phi_n(\lambda)$ and $\lambda_0 \neq 0$, then $-\lambda_0$ is also a root of $\phi_n(\lambda)$.

Proof: Follows directly from the previous lemma.

By definition, the roots of $\phi_n(\lambda)$ are the eigenvalues of matrix A_n. Since A_n is real symmetric, it follows that all the roots of $\phi_n(\lambda)$ are real. Let $\lambda_{\max}^{(n)}$ denote the root of $\phi_n(\lambda)$ with largest absolute value. In view of the result in Lemma 3, $\lambda_{\max}^{(n)}$ can always be taken to be positive.

Lemma 4: For any finite n, $\lambda_{\max}^{(n)}$ of $\phi_n(\lambda)$ has the following ordering:

$$\lambda_{\max}^{(1)} < \lambda_{\max}^{(2)} < \cdots < \lambda_{\max}^{(n)} < 2. \tag{59}$$

Proof: We will prove this lemma again by induction. Clearly, $\phi_n(\lambda) \to \infty$ as $\lambda \to \infty$ for any $n = 1, 2, \cdots$. It follows that $\phi_n(\lambda) > 0$ for $\lambda > \lambda_{\max}^{(n)}$ the largest positive root of $\phi_n(\lambda)$. Now, assume that (59) holds for $\lambda_{\max}^{(n-1)}$ and $\lambda_{\max}^{(n-2)}$, then, from (56),

$$\begin{aligned} \phi_n(\lambda_{\max}^{(n-1)}) &= \lambda_{\max}^{(n-1)}\phi_{n-1}(\lambda_{\max}^{(n-1)}) - \phi_{n-2}(\lambda_{\max}^{(n-1)}) \\ &= -\phi_{n-2}(\lambda_{\max}^{(n-1)}) < 0. \end{aligned} \tag{60}$$

Hence $\phi_n(\lambda)$ changes sign at least once between $\lambda_{\max}^{(n-1)}$ and ∞. This implies that

$$\lambda_{\max}^{(n)} > \lambda_{\max}^{(n-1)}. \tag{61}$$

Since (59) holds for $n = 1$ and 2, it holds for any n.

To show that $\lambda_{\max}^{(n)} < 2$ for any finite n, we make the observation that for $\lambda > 2$, the matrix $\lambda 1_n - A_n$ is a dominant matrix,[13] and therefore nonsingular. For $\lambda = 2$,

$$\begin{aligned} \phi_n(2) &= n\phi_1(2) - (n - 1)\phi_0(2), \\ &= n + 1 \neq 0, \end{aligned} \tag{62}$$

by repeated use of the recursive relation (56).

Lemma 5: For some number λ_0, if $\phi_n(\lambda_0) = 0$, then $\phi_{n-1}(\lambda_0) \neq 0$.

Proof: Assume the contrary, i.e.,

$$\phi_n(\lambda_0) = \phi_{n-1}(\lambda_0) = 0. \tag{63}$$

Then from the recursive formula (56),

$$\phi_{n-2}(\lambda_0) = 0. \tag{64}$$

Repeating the same argument, we conclude

$$\phi_n(\lambda_0) = \cdots = \phi_1(\lambda_0) = \phi_0(\lambda_0) = 0 \tag{65}$$

which is impossible.

Lemma 6: $\phi_n(\lambda)$ has only simple roots.

Proof: Let λ_0 be a root of $\phi_n(\lambda)$, then the matrix $[\lambda_0 1_n - A_n]$ is singular. On the other hand, from the previous lemma, $[\lambda_0 1_{n-1} - A_{n-1}]$ is nonsingular. Hence the null space of the matrix $[\lambda_0 1_n - A_n]$

is one-dimensional. From the fact that A_n is diagonalizable, we conclude that λ_0 must be a simple root of $\phi_n(\lambda)$.

Lemma 7: Write

$$A_n = PD_AP' \tag{66}$$

where

$$D_A = \operatorname{diag}\,[\lambda_1\,,\,\cdots\,,\,\lambda_n] \tag{67}$$

and

$$PP' = 1; \tag{68}$$

then in general, P can be expressed in terms of $\phi(\lambda)$'s

$$P = \begin{bmatrix} \dfrac{\phi_0(\lambda_1)}{\phi(\lambda_1)} & \dfrac{\phi_0(\lambda_2)}{\phi(\lambda_2)} & \cdots & \dfrac{\phi_0(\lambda_n)}{\phi(\lambda_n)} \\ \vdots & \vdots & & \vdots \\ \dfrac{\phi_{n-1}(\lambda_1)}{\phi(\lambda_1)} & \dfrac{\phi_{n-1}(\lambda_2)}{\phi(\lambda_2)} & & \dfrac{\phi_{n-1}(\lambda_n)}{\phi(\lambda_n)} \end{bmatrix} \tag{69}$$

where

$$\phi(\lambda) = \left[\sum_{i=0}^{n-1} \phi_i^2(\lambda)\right]^{\frac{1}{2}} \tag{70}$$

is a normalization constant.

Proof: Let $\mathbf{x}$ be an eigenvector corresponding to an eigenvalue λ of A_n,

$$A_n\mathbf{x} = \lambda\mathbf{x}. \tag{71}$$

Write

$$\mathbf{x} = [x_1\,,\,\cdots\,,\,x_n]'$$

and expand (71), we have

$$\begin{aligned} x_2 &= \lambda x_1\,, \\ x_3 &= \lambda x_2 - x_1\,, \\ x_4 &= \lambda x_3 - x_2 \\ &\vdots \\ x_n &= \lambda x_{n-1} - x_{n-2}\,, \end{aligned} \tag{72}$$

and

$$x_{n-1} = \lambda x_n .$$

Delete the last equation in (72) and then compare with equations (55) and (56). We can make the following identification:

$$\begin{aligned} x_1 &= \phi_0(\lambda), \\ x_2 &= \phi_1(\lambda) \\ &\vdots \\ x_n &= \phi_{n-1}(\lambda). \end{aligned} \tag{73}$$

To normalize **x**, we divide (73) by the inner product of **x**. Denoting the inner product by $\phi(\lambda)$,

$$\phi(\lambda) = \left[\sum_{i=0}^{n-1} \phi_i^2(\lambda) \right]^{\frac{1}{2}}, \tag{74}$$

the normalized eigenvector corresponding to the eigenvalue λ is given by

$$\left[\frac{\phi_0(\lambda)}{\phi(\lambda)} \cdots \frac{\phi_{n-1}(\lambda)}{\phi(\lambda)} \right]'. \tag{75}$$

It is well known that eigenvectors corresponding to different eigenvalues are orthogonal. Therefore, for P as defined in (69), $P'P = 1_n$. Since each column of P is an eigenvector of A_n, it follows that

$$A_n P = P D_A$$

and equation (66) is immediate.

REFERENCES

1. Aaron, M. R., "PCM Transmission in the Exchange Plant," B.S.T.J., *41,* No. 1 (January 1962), pp. 99–141.
2. Davis, J. H., "T2: A 6.3 Mb/s Digital Repeatered Line," IEEE International Conference on Communications, Boulder, Colorado, June 1969.
3. Dorros, I., Sipress, J. M., and Waldhauer, F. D., "An Experimental 224 Mb/s Digital Repeatered Line," B.S.T.J., *45,* No. 7 (September 1966), pp. 993–1043.
4. McCullough, R. H., "Ternary Codes for Regenerative Digital Transmission Systems," Ph.D. Dissertation, Polytechnic Institute of Brooklyn, June 1967.
5. Cattermole, K. W., "Low-Disparity Codes and Coding for PCM," Proc. IEEE Conf. Transmission Aspects of Commun. Networks, London, February 1964, pp. 179–182.
6. Carter, R. O., "Low-Disparity Binary Coding System," Elec. Letters, *1,* No. 3 (May 1965), pp. 67–68.
7. Sipress, J. M., "A New Class of Selected Ternary Pulse Transmission Plans for Digital Transmission Lines," IEEE Trans. on Commun. Technology, *13,* No. 3 (September 1965), pp. 366–372.

8. Johannes, V. I., Kaim, A. G., and Walzman, T., "Bipolar Pulse Transmission with Zero Extraction," Trans. IEEE Commun. Technology, *17*, No. 2 (April 1969), pp. 303–310.
9. Franaszek, P. A., "Sequence-State Coding for Digital Transmission," B.S.T.J., *47,* No. 1 (January 1968), pp. 143–157.
10. Gantmacher, F. R., *The Theory of Matrices,* Vol. II, New York: Chelsea Publishing Co., 1959, pp. 50–60.
11. Johannes, V. I., unpublished work.
12. Marcus, M., and Ming, H., *A Survey of Matrix Theory and Matrix Inequalities,* Boston: Allyn and Bacon, 1964, pp. 166–167.
13. Bellman, R., *Introduction to Matrix Analysis,* New York: McGraw-Hill, 1960, p. 295.

Pull-In Range of a Phase-Locked Loop With a Binary Phase Comparator

By JAMES F. OBERST

(Manuscript received May 25, 1970)

We develop a method for calculating the pull-in range of a phase-locked loop with a binary phase comparator and an arbitrary loop filter. Complete numerical results are presented for loop filters of the phase-lag and low-pass types. The problem of stability is also considered, and it is proved that with these loop filters no steady-state phase jitter can exist after frequency acquistion has been achieved.

I. INTRODUCTION

The phase-locked loop (PLL) is an important element of many modern communication and control systems. A PLL block diagram is shown in Fig. 1. The input $v_1(\omega_0 t + \theta_1)$ is a narrow-band signal with carrier frequency ω_0 and phase $\theta_1(t)$. This phase is compared with the phase $\theta_2(t)$ of the voltage-controlled oscillator (VCO) in the phase comparator (PC). The PC output $f(\phi)$, where $\phi = \theta_1 - \theta_2$, is filtered by the loop filter $H(p)$ and applied to the VCO control terminal.

Depending on the values of the PLL parameters, the phase error ϕ can be kept small even with input phase modulation. Thus with $\theta_1(t) = \Omega t + \theta_{10}$, which represents a constant input frequency offset, the system can produce a synchronized signal $v_2(t)$ with frequency $\omega_0 + \Omega$. This synchronization capability leads to PLL applications in carrier extraction,[1] frequency synthesis,[2] narrow-band filtering,[3] FM demodulation,[3] timing extraction in PCM and data transmission systems,[4] etc.

In this paper, we examine the acquisition, or pull-in, range of a PLL with a binary PC. We present numerical results for the special case of a second-order PLL with either low-pass or phase-lag loop filter.

The PC characteristic considered here is the binary curve shown in Fig. 2. It is of interest in at least three situations. First, since many synchronization systems are designed to operate with very small

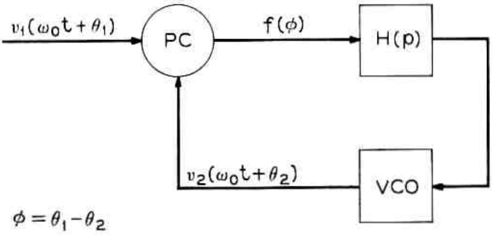

$v_1(\omega_0 t + \theta_1)$
PC
$f(\phi)$
H(p)
$v_2(\omega_0 t + \theta_2)$
VCO
$\phi = \theta_1 - \theta_2$

Fig. 1—PLL block diagram.

phase errors, dynamic range limitations in the PC circuitry often produce severe saturation effects. The characteristic of Fig. 2 corresponds to the extreme case of vanishing linear range near zero phase error. However, it is a useful approximation for systems with small but nonzero dynamic range for the purpose of studying pull-in performance. Second, a binary PC can be easily implemented with logic circuits. The resulting characteristic differs from the ideal of Fig. 2 by exhibiting small hysteresis about the zeros at $\phi = n\pi$, but this hysteresis has no effect on the pull-in range achieved. Finally, J. J. Stiffler[5] has shown that for a first-order PLL with additive white gaussian noise and no frequency offset, the cross-correlation type PC which minimizes Pr $\{ |\phi| > \phi_0 \}$ for all ϕ_0 has the characteristic of Fig. 2. Although such a square-wave correlation function is unrealizable, this result suggests that PLLs employing other types of PC having this characteristic are worthy of consideration. In addition, similar "bang-bang" control characteristics are known to be optimum for PLL acquisition.[6]

II. CALCULATION OF PULL-IN RANGE

The phase model corresponding to the PLL block diagram in Fig. 1 is shown in Fig. 3. We assume that the gain of the loop filter $H(p)$ is unity at DC. The input-signal frequency differs from the VCO center frequency by Ω rad/s:

$$\theta_1(t) = \Omega t + \theta_{10}. \tag{1}$$

It is convenient to normalize the detuning Ω to the dc loop gain* α:

$$\gamma = \Omega/\alpha. \tag{2}$$

* Since no gain can be defined for the binary PC being considered, the symbol α does not represent the usual small-signal loop gain.

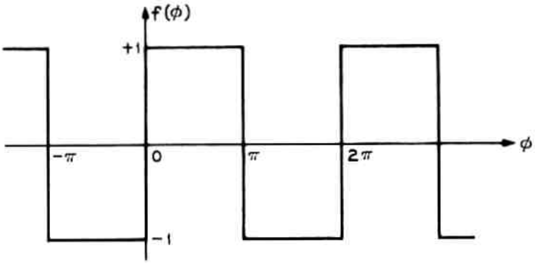


Fig. 2—Phase comparator characteristic.

When the normalized detuning γ is not too great, the VCO frequency changes toward the input frequency, and eventually the PLL synchronizes to the input signal with zero frequency error and finite phase error. The normalized lock range γ_L is the maximum detuning $|\gamma|$ for which the PLL can remain locked after synchronization has been established. Inspection of dc conditions in the phase model of Fig. 3 shows that $\gamma_L = 1$. The normalized pull-in range γ_P is the maximum $|\gamma|$ for which eventual synchronization is assured from any initial conditions in the loop filter and VCO. In general $\gamma_P < 1$ for PLLs of order higher than first. The order of a PLL is defined as one plus the number of poles in $H(p)$. Calculation of γ_P is the subject of this paper.

The method employed here is similar to that used by A. J. Goldstein[7] to calculate γ_P for a PLL with a sawtooth PC. Due to the binary nature of $f(\phi)$, the PC output waveform $f[\phi(t)]$ is piecewise constant, assuming only the values ± 1. Assume that the PLL is not synchronized to the input signal. Then $\phi(t)$ increases with time (for $\Omega > 0$), and the waveforms $\phi(t)$ and $f[\phi(t)]$ appear as shown in Fig. 4. The time origin has been selected so that $\phi(0) = 0$. The transition instants are

$$0 = t_{02} < t_{11} < t_{12} < \cdots < t_{j1} < t_{j2} < \cdots \tag{3}$$

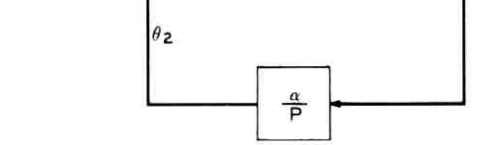


Fig. 3—Phase model of PLL.

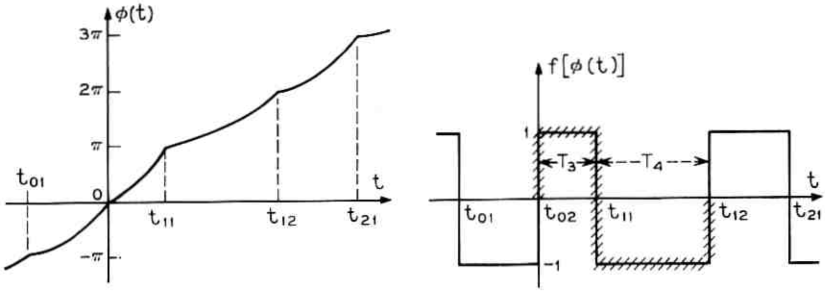


Fig. 4—Phase and PC output waveforms.

where

$$\phi(t_{j1}) = (2j - 1)\pi \text{ (negative transition)},$$
$$\phi(t_{j2}) = 2j\pi \text{(positive transition)}. \tag{4}$$

An expression can be written for $f[\phi(t)]$ by summing all of the segments corresponding to 2π increments in $\phi(t)$:

$$f[\phi(t)] = \sum_{j=-\infty}^{\infty} [u(t - t_{j-1,2}) - 2u(t - t_{j1}) + u(t - t_{j2})]. \tag{5}$$

The $j = 1$ segment is shown crosshatched in Fig. 4. Since the PLL is not synchronized to the input signal, the steady-state PC waveform $f_{ss}[\phi(t)]$ is periodic, and the transition instants can be written:

$$\begin{aligned} t_{j2} &= j(T_3 + T_4), \\ t_{j1} &= j(T_3 + T_4) - T_4 . \end{aligned} \tag{6}$$

T_3 and T_4 are the times between transitions as indicated in Fig. 4. Using equation (6), equation (5) becomes:

$$\begin{aligned} f_{ss}[\phi(t)] = \sum_{j=-\infty}^{\infty} & [u(t - [j - 1][T_3 + T_4]) \\ & - 2u(t - j[T_3 + T_4] + T_4) + u(t - j[T_3 + T_4])]. \end{aligned} \tag{7}$$

From equation (1) and Fig. 3, the phase error in steady state is:

$$\phi_{ss}(t) = \Omega t + \phi_0 - f_{ss}[\phi(t)] * \mathcal{L}^{-1}\left\{\frac{\alpha H(p)}{p}\right\}, \tag{8}$$

where ϕ_0 is some constant. Since $f_{ss}[\phi(t)]$ is composed only of step functions, the last term in equation (8) can be written as a sum of delayed functions $g(t)$, where

$$\mathcal{L}\{g(t)\} = G(p) = \frac{H(p)}{p^2}. \tag{9}$$

The general expression for $\phi_{ss}(t)$ is then

$$\phi_{ss}(t) = \Omega t + \phi_0 - \alpha \sum_{j=-\infty}^{\infty} [g(t - [j - 1][T_3 + T_4]) \\ - 2g(t - j[T_3 + T_4] + T_4) + g(t - j[T_3 + T_4])]. \tag{10}$$

From equation (4) and Fig. 4, we have the following conditions on $\phi_{ss}(t)$:

$$\begin{aligned} \phi_{ss}(0) &= 0, \\ \phi_{ss}(T_3) &= \pi, \\ \phi_{ss}(T_3 + T_4) &= 2\pi. \end{aligned} \tag{11}$$

These conditions are sufficient to determine the unknown constants ϕ_0, T_3, and T_4 in equation (10). Thus equations (10) and (11) together define the relationship between the normalized detuning γ and the loop-filter parameters [through $g(t)$] which must be satisfied for the PLL to be out-of-lock in the steady state. Then clearly γ_P is the minimum value of γ for which these equations possess a solution.

III. RESULTS FOR SECOND-ORDER PLL

In this section, the method derived above is applied to the second-order PLL with loop filter

$$H(p) = \frac{1 + T_2 p}{1 + T_1 p}. \tag{12}$$

This is a phase-lag filter for $0 < T_2 < T_1$. It becomes a simple low-pass filter for $T_2 = 0$, and setting $T_2 = T_1$ reduces the PLL to first order. The corresponding $g(t)$, from equation (9), is

$$g(t) = [t - (T_1 - T_2)(1 - \exp(-t/T_1))]u(t). \tag{13}$$

In Appendix A, equation (10) is rewritten using equation (13) and is evaluated at $t = 0$, T_3, and $T_3 + T_4$. Equation (11) is then applied, along with the normalization

$$\tau_i = \alpha T_i, \qquad i = 1, 2, 3, 4. \tag{14}$$

The equations which result are

$$\gamma = \frac{2\pi + \tau_3 - \tau_4}{\tau_3 + \tau_4}, \tag{15}$$

$$4(\tau_1 - \tau_2)(\tau_3 + \tau_4) = [\tau_4(\tau_3 + \pi) + \tau_3(\tau_4 - \pi)]\left[\coth \frac{\tau_3}{2\tau_1} + \coth \frac{\tau_4}{2\tau_1}\right]. \tag{16}$$

According to the discussion following equation (11), the pull-in range is

$$\gamma_p = \min_{\tau_3, \tau_4 > 0} \frac{2\pi + \tau_3 - \tau_4}{\tau_3 + \tau_4}, \tag{17}$$

subject to the constraint equation (16). It is important to note that equation (15) is simply the dc balance equation for the PLL model of Fig. 3, and therefore holds for any $H(p)$ with unity dc gain. Since equation (15) gives γ explicitly in terms of τ_3 and τ_4, it can always be used to eliminate γ from a constraint equation corresponding to equation (16). Therefore γ_p can be calculated from equation (17) for any $H(p)$ subject to the appropriate constraint equation which relates the loop-filter parameters to the transition-time parameters τ_3 and τ_4. Hence only $\phi_{ss}(0)$ and $\phi_{ss}(T_3)$ actually had to be evaluated in Appendix A.

The method employed to evaluate γ_p for various values of τ_1 and τ_2 was to choose $\tau_3 > 0$, use equation (16) to obtain the corresponding τ_4, and calculate γ from equation (15). The τ_3, τ_4 relationship was found to be single-valued for all loop filters investigated. Examples of the behavior of γ with τ_3 are given in Fig. 5. The filter parameter r is defined as

$$r = T_2/T_1. \tag{18}$$

In all cases the curves $\gamma(\tau_3)$ are smooth and exhibit a single local minimum. This minimum is found by computing a sequence $\gamma(\tau_{3n})$, where $\tau_{3n} > \tau_{3,n-1}$. When this sequence begins to increase, the minimum γ_p has just been passed, and can be estimated accurately from the last three computed values of γ.

Curves of γ_p versus τ_1 with r as a parameter are presented in Fig. 6. Several characteristics are notable. First, $\gamma_p = 1$ for $r \geqq 0.5$ independent of τ_1. Second, as τ_1 increases with r constant, γ_p approaches an asymptotic value γ_{pa} which is a function only of r. Later we shall derive an explicit formula for γ_{pa}. The same results are presented in a different way in Fig. 7, which shows curves of constant γ_p on the τ_1, τ_2

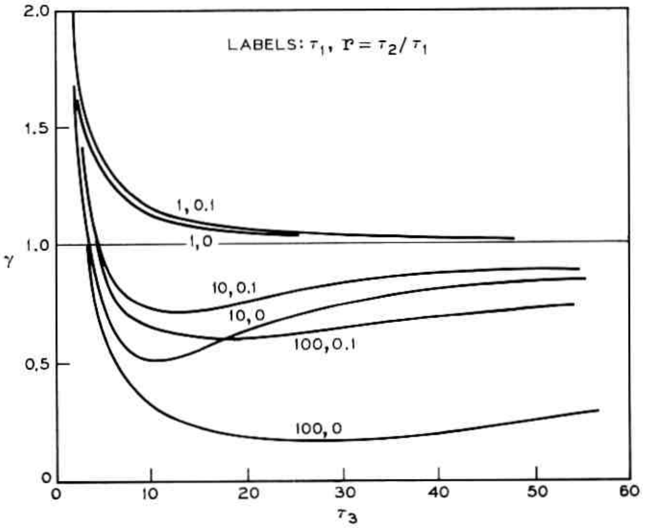


Fig. 5—γ versus τ_3.

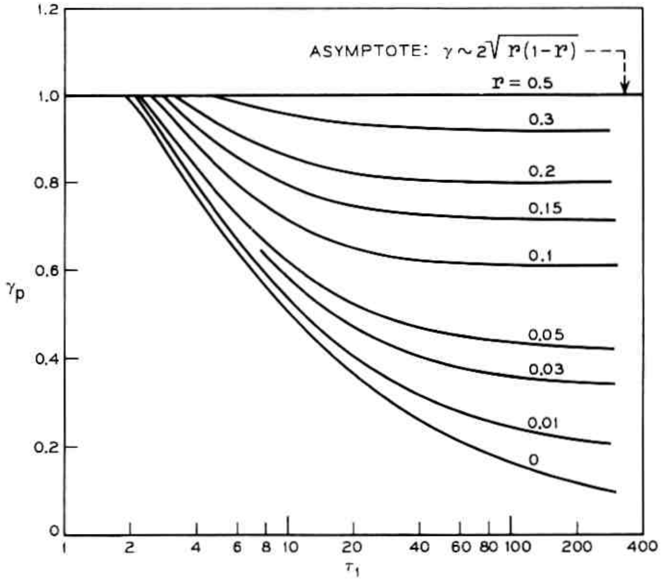


Fig. 6—γ_p versus τ_1 with parameter r.

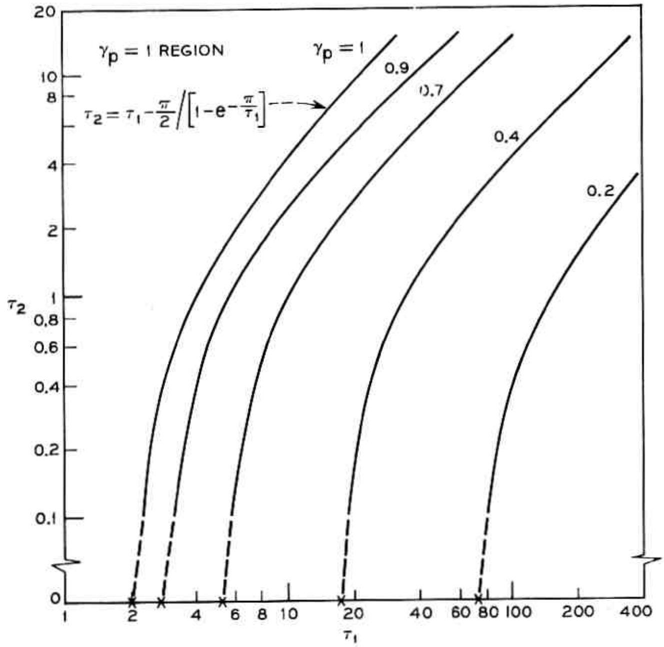


Fig. 7—τ_2 versus τ_1 with parameter γ_p.

parameter plane. Below we derive the equation for the curve in this figure corresponding to $\gamma_p = 1$.

IV. FURTHER RESULTS

Let us consider the important case of very large τ_1, which corresponds to strong filtering in the PLL. Noting that

$$\lim_{x \to 0} \coth x = \frac{1}{x} \tag{19}$$

and using equation (18), equation (16) becomes for large τ_1:

$$4\tau_1(1 - r)(\tau_3 + \tau_4) = [\tau_4(\tau_3 + \pi) + \tau_3(\tau_4 - \pi)]\left[\frac{2\tau_1}{\tau_3} + \frac{2\tau_1}{\tau_4}\right]. \tag{20}$$

This simplifies to

$$\tau_4 = \frac{\pi\tau_3}{2r\tau_3 + \pi}. \tag{21}$$

Substituting equation (21) into equation (15) gives the result for γ

with τ_1 large:

$$\gamma = \frac{r\tau_3^2 + 2\pi r\tau_3 + \pi^2}{r\tau_3^2 + \pi\tau_3}. \tag{22}$$

All minima of γ must satisfy

$$\frac{d\gamma}{d\tau_3} = 0. \tag{23}$$

Applying equation (23) to equation (22) yields a single positive value of τ_3:

$$\tau_3 = \frac{\pi\left[1 + \left(\frac{1}{r} - 1\right)^{\frac{1}{2}}\right]}{1 - 2r}, \qquad r < 0.5. \tag{24}$$

Substituting equation (24) into equation (22) gives the result

$$\gamma_{pa} = \begin{cases} 2[r(1 - r)]^{\frac{1}{2}}, & 0 \leqq r < 0.5, \\ 1, & r \geqq 0.5. \end{cases} \tag{25}$$

This agrees with the numerical results in Fig. 6 and with a result obtained by M. V. Kapranov by a different method in an untranslated Russian paper.[8] The existence of only a single minimum of $\gamma(\tau_3)$ for large τ_1 supports our hypothesis of a single minimum for all τ_1 which is based on the curves in Fig. 5.

The region $\gamma_p = 1$ in Fig. 7 corresponds to τ_1, τ_2 such that $\gamma \geqq 1$ for all τ_3, τ_4. From equation (15), this implies that

$$\tau_4 \leqq \pi. \tag{26}$$

Thus from equation (16), τ_1, τ_2, and τ_3 on the $\gamma_p = 1$ boundary must satisfy

$$4(\tau_1 - \tau_2)(\tau_3 + \pi) = \pi(\tau_3 + \pi)\left[\coth\frac{\tau_3}{2\tau_1} + \coth\frac{\pi}{2\tau_1}\right]. \tag{27}$$

Examination of Fig. 5 shows that the critical γ curves approach $\gamma = 1$ from above for very large τ_3. Using

$$\operatorname*{Lim}_{x\to\infty} \coth x = 1, \tag{28}$$

equation (27) becomes

$$4(\tau_1 - \tau_2)\tau_3 = \pi\tau_3\left[1 + \coth\frac{\pi}{2\tau_1}\right]. \tag{29}$$

This reduces to the simple expression for the $\gamma_p = 1$ curve:

$$\tau_2 = \tau_1 - \frac{\pi}{2} \Big/ [1 - \exp(-\pi/\tau_1)]. \tag{30}$$

Finally, let us consider the following problem. The periodic behavior of $\phi_{ss}(t)$ assumed throughout this paper which led to equations (15) and (16) is known as a limit cycle of the second kind in the phase plane.[9] The nonexistence of such limit cycles for $|\gamma| < \gamma_p$ proves that frequency lock is eventually attained for these values of γ. Physically speaking, $\phi_{ss}(t)$ cannot increase monotonically with time as assumed in Fig. 4. However, proper synchronization of the PLL requires that phase lock also be achieved. This means that after a long enough time, the system comes to rest:

$$\lim_{t\to\infty} \phi(t) = 2n\pi, \qquad |\gamma| < \gamma_p . \tag{31}$$

Because of the gross nonlinearity $f(\phi)$ in the system considered here, it is not obvious that equation (31) will be satisfied. Specifically, it is conceivable that a series of self-sustaining overshoots in $\phi(t)$ could become established after pull-in which would produce a periodic phase jitter. This behavior corresponds to a limit cycle of the first kind in the phase plane. Although this problem is not solved in general here, a test which is valid for any $H(p)$ is applied to the phase-lag filter case in Appendix B. It is found that in this case, phase lock described by equation (31) is always achieved.

V. CONCLUSION

A method has been presented for calculating the pull-in range γ_p of a PLL with a binary phase comparator and an arbitrary loop filter. The result is obtained as the minimum value of a function of two variables, subject to a constraint equation which relates these variables to the parameters of the loop filter. Complete numerical results for γ_p were obtained for loop filters of the phase-lag and low-pass types. Explicit formulas were given in this case for the asymptotic value of γ_p with strong loop filtering, and for the set of filter parameters which result in unity pull-in range. Finally, it was proved that no steady-state phase jitter can exist after pull-in with these loop filters.

VI. ACKNOWLEDGMENT

The author thanks M. R. Aaron for his suggestions and encouragement throughout the course of this work.

APPENDIX A

Evaluation of $\phi_{ss}(t)$

Consider one period of $\phi_{ss}(t)$, and define the functions $\phi_i(t)$ as:

$$\phi_{ss}(t) = \begin{cases} \phi_1(t), & 0 \leqq t \leqq T_3\,, \\ \phi_2(t), & T_3 \leqq t \leqq T_3 + T_4\,. \end{cases} \tag{32}$$

From Fig. 4, $\phi_1(t)$ includes all terms in equation (10) which correspond to transitions prior to $t_{11} = T_3$. Using equation (13) in (10), $\phi_1(t)$ becomes:

$$\begin{aligned} \phi_1(t) = \Omega t + \phi_0 - \alpha \sum_{j=-\infty}^{1} & [t - (j-1)(T_3 + T_4) - (T_1 - T_2) \\ & \cdot (1 - \exp[-(t - [j-1][T_3 + T_4])/T_1])] \\ & + 2\alpha \sum_{j=-\infty}^{0} [t - j(T_3 + T_4) + T_4 - (T_1 - T_2) \\ & \cdot (1 - \exp[-(t - j[T_3 + T_4] + T_4)/T_1])] \\ & - \alpha \sum_{j=-\infty}^{0} [t - j(T_3 + T_4) - (T_1 - T_2) \\ & \cdot (1 - \exp[-(t - j[T_3 + T_4])/T_1])]. \end{aligned} \tag{33}$$

Absorbing all constant terms into ϕ_0, equation (33) becomes:

$$\begin{aligned} \phi_1(t) &= \Omega t + \phi_0' - \alpha t \\ & \quad - \alpha(T_1 - T_2) \exp(-t/T_1) \sum_{j=-\infty}^{0} [\exp[j(T_3 + T_4)/T_1] \\ & \quad - 2 \exp[(j[T_3 + T_4] - T_4)/T_1] + \exp[j(T_3 + T_4)/T_1]], \\ &= \Omega t + \phi_0' - \alpha t - 2\alpha(T_1 - T_2) \\ & \quad \cdot \exp(-t/T_1) \frac{1 - \exp(-T_4/T_1)}{1 - \exp[-(T_3 + T_4)/T_1]}. \end{aligned} \tag{34}$$

ϕ_0' is eliminated from equation (34) using

$$\phi_1(0) = 0, \tag{11a}$$

which yields

$$\begin{aligned} \phi_1(t) &= (\Omega - \alpha)t + 2\alpha(T_1 - T_2) \\ & \quad \cdot \frac{1 - \exp(-T_4/T_1)}{1 - \exp[-(T_3 + T_4)/T_1]} (1 - \exp(-t/T_1)). \end{aligned} \tag{35}$$

Now requiring that

$$\phi_1(T_3) = \pi \tag{11b}$$

results in

$$\pi = (\Omega - \alpha)T_3 + 2\alpha(T_1 - T_2) \cdot \frac{(1 - \exp(-T_3/T_1))(1 - \exp(-T_4/T_1))}{1 - \exp[-(T_3 + T_4)/T_1]} \tag{36}$$

which after some manipulation becomes

$$\pi = (\Omega - \alpha)T_3 + 4\alpha(T_1 - T_2) \Big/ \left[\coth\frac{T_3}{2T_1} + \coth\frac{T_4}{2T_1}\right]. \tag{37}$$

Next, $\phi_2(t)$ is obtained from equation (35) by adding the term in equation (10) which corresponds to the transition time $t_{11} = T_3$:

$$\begin{aligned}\phi_2(t) &= (\Omega - \alpha)t + 2\alpha(T_1 - T_2) \\ &\quad \cdot \frac{1 - \exp(-T_4/T_1)}{1 - \exp[-(T_3 + T_4)/T_1]}(1 - \exp(-t/T_1)) \\ &\quad + 2\alpha(t - T_3) - 2\alpha(T_1 - T_2)(1 - \exp[-(t - T_3)/T_1]).\end{aligned} \tag{38}$$

Requiring that

$$\phi_2(T_3 + T_4) = 2\pi \tag{11c}$$

and performing some simple manipulation leads to the result

$$2\pi = \Omega(T_3 + T_4) - \alpha(T_3 - T_4). \tag{39}$$

Equations (37) and (39) can be normalized by letting

$$\gamma = \Omega/\alpha, \tag{2}$$

$$\tau_i = \alpha T_i, \qquad i = 1, 2, 3, 4. \tag{14}$$

Equation (39) becomes

$$\gamma = \frac{2\pi + \tau_3 - \tau_4}{\tau_3 + \tau_4} \tag{15}$$

which is equation (15) in Section III. Equation (37) becomes

$$\gamma = 1 + \frac{1}{\tau_3}\left\{\pi - 4(\tau_1 - \tau_2) \Big/ \left[\coth\frac{\tau_3}{2\tau_1} + \coth\frac{\tau_4}{2\tau_1}\right]\right\}. \tag{40}$$

Eliminating γ between equations (15) and (40) gives the constraint equation (16).

APPENDIX B

The Phase Jitter Problem

In this appendix it is proved that whenever the second-order PLL studied here achieves frequency lock, it also achieves phase lock:

$$\lim_{t\to\infty} \phi(t) = 2n\pi, \qquad |\gamma| < \gamma_p . \tag{31}$$

It can be shown that for this system any steady-state phase jitter $\phi(t)$ is confined within the $\pm\pi$ neighborhood of some lock point $\phi = 2n\pi$. This is a result of the periodicity of the phase-plane geometry in the ϕ direction. However, since this proof requires a rather lengthy description of the properties of the phase-plane trajectories, it will be omitted. Below we prove that equation (31) holds when the phase error remains within such a $\pm\pi$ neighborhood of a lock point.

The technique used previously to calculate γ_p can also be employed here. Assume that in the steady state, a periodic phase jitter $\phi(t)$ exists. Then the waveform $f[\phi(t)]$ is binary and periodic, and the phase error is again given by equation (8). Since we are now considering a phase jitter within $\pm\pi$ of $\phi = 2n\pi$, the requirements on $\phi_{ss}(t)$ are:

$$\phi_{ss}(0) = \phi_{ss}(T_3) = \phi_{ss}(T_3 + T_4) = 2n\pi. \tag{41}$$

Proceeding as in Appendix A, we obtain the equations

$$\gamma = \frac{\tau_3 - \tau_4}{\tau_3 + \tau_4}, \tag{42}$$

$$2(\tau_1 - \tau_2)(\tau_3 + \tau_4) = \tau_3\tau_4\left[\coth\frac{\tau_3}{2\tau_1} + \coth\frac{\tau_4}{2\tau_1}\right]. \tag{43}$$

These equations may be written directly from equations (15) and (16) by replacing π with 0. Below it is demonstrated that $\tau_3 = \tau_4 = 0$ is the only solution of these equations when $|\gamma| < 1$, which proves equation (31).

Using $|\gamma| < 1$ in equation (42) yields

$$\tau_3 , \tau_4 > 0, \tag{44}$$

so that only equation (43) must be considered. Using r of equation (18) and dividing by $\tau_3\tau_4$ gives

$$2\tau_1(1 - r)\left(\frac{1}{\tau_3} + \frac{1}{\tau_4}\right) = \coth\frac{\tau_3}{2\tau_1} + \coth\frac{\tau_4}{2\tau_1}. \tag{45}$$

Defining

$$x_j = \frac{\tau_j}{2\tau_1}, \qquad j = 3, 4, \tag{46}$$

equation (45) becomes

$$(1 - r)\left(\frac{1}{x_3} + \frac{1}{x_4}\right) = \coth x_3 + \coth x_4 . \tag{47}$$

Recalling that $r \geqq 0$, we have for $x > 0$:

$$\coth x > \frac{1}{x} \geqq \frac{1 - r}{x}. \tag{48}$$

Thus the only possible nonnegative solution of equation (47) is

$$x_3 = x_4 = 0 \tag{49}$$

which is the desired result.

REFERENCES

1. Viterbi, A. J., *Principles of Coherent Communications,* New York: McGraw-Hill, 1960.
2. Noordanus, J., "Frequency Synthesizers—A Survey of Techniques," IEEE Trans. Commun. Tech., *COM-17,* No. 2 (April 1969), pp. 257–271.
3. Gilchriest, C. E., "Application of the Phase-Locked Loop to Telemetry as a Discriminator or Tracking Filter," IRE Trans. Telemetry Rem. Cont., *TRC-4,* No. 1 (June 1958), pp. 20–35.
4. Saltzberg, B. R., "Timing Recovery for Synchronous Binary Data Transmission," B.S.T.J., *46,* No. 3 (March 1967), pp. 593–622.
5. Stiffler, J. J., "On the Selection of Signals for Phase Locked Loops," IEEE Trans. Commun. Tech., *COM-16,* No. 2 (April 1968), pp. 239–244.
6. Shaft, P. D., and Dorf, R. C., "Minimization of Communication–Signal Acquisition Time in Tracking Loops," IEEE Trans. Commun. Tech., *COM-16,* No. 3 (June 1968), pp. 495–499.
7. Goldstein, A. J., "Analysis of the Phase-Controlled Loop with a Sawtooth Phase Comparator," B.S.T.J., *41,* No. 3 (March 1962), pp. 603–633.
8. Kapranov, M. V., "The Asymptotic Value of the Locking Band in Phase Automatic Frequency Control" [in Russian], Radiophysics, *11,* No. 7 (1968), pp. 1028–1040.
9. Minorsky, N., *Nonlinear Oscillations,* Princeton, New Jersey: Van Nostrand, 1962.

A Fast Method of Generating Digital Random Numbers

By C. M. RADER,* L. R. RABINER and R. W. SCHAFER

(Manuscript received June 12, 1970)

In this article we propose a fast, efficient technique for generating a pseudorandom stream of uniformly-distributed numbers. The arithmetic operations required are an L bit exclusive-or, a rotation, and a shift to update the state of the number generator. With moderately large values of L we have been able to generate sequences of numbers whose periods are quite long (on the order of 2×10^7 *long). Its simplicity of construction, as well as its ability to generate long streams of independent pseudorandom uniformly-distributed integers make this noise generator a worthy candidate for use in high-speed digital systems.*

I. INTRODUCTION

Almost all methods of generating digital random numbers use as input a set of previously generated random numbers which were produced by an iterative arithmetic process (modulo a large integer)—e.g.

$$X_n \equiv F(X_{n-1}, X_{n-2}, \cdots, X_{n-J}) \bmod N \qquad n = 0, 1, \cdots$$

with initial conditions

$$\begin{aligned} X_{-1} &= C_1 , \\ X_{-2} &= C_2 , \\ &\cdot \\ &\cdot \\ &\cdot \\ X_{-J} &= C_J . \end{aligned}$$

For each new random number, X_n, this arithmetic process is repeated. The integer N and the initial values C_1, C_2, $\cdots$, C_J are chosen to

* Mr. Rader is with Lincoln Laboratories, Massachusetts Institute of Technology, Lexington, Massachusetts. Lincoln Laboratories is operated with support from the U. S. Air Force.

guarantee a large period of repetition. Methods of the type described above involve a considerable propagation delay, representing at the least one addition or one multiplication time, between the time the nth random number is put into its storage register, and the time at which the $(n + 1)$st random number is available. This delay is not generally a problem in most applications, because computational delays in other parts of most digital systems are far greater than those encountered in generating random numbers. However, it is conceivable that in the future a need will arise for which random numbers must be computed far more rapidly than is now necessary. With such a time in mind we propose the following algorithm, for which the time necessary to produce a new random number is equal to the sum of a flip-flop settling time, and the propagation delay of an exclusive-or gate.

II. THEORY

The algorithm for generating the L-bit random number X_n from the two previous L-bit numbers X_{n-1} and X_{n-2} may be stated as

$$X_n = T_P(X_{n-1} \oplus X_{n-2})$$

where $T_P(\cdot)$ denotes a cyclic rotation of P places to the right and $\oplus$ denotes exclusive-or. The algorithm requires L flip-flops to store X_{n-1} and L flip-flops to store X_{n-2} . Each bit of the new random number X_n is derived by an exclusive-or operation on a pair of corresponding bits in the previous random numbers. These bits are not stored in the bit positions from which they were produced, but instead each new bit is rotated cyclically to the right by P bit positions, with overflow on the right being fed into the left. The process is illustrated in Fig. 1 for $P = 1$. At each cycle, the new random number generated (X_n) is clocked into the lower flip-flops (X_{n-1}); at the same time the number stored in the lower flip-flops (X_{n-1}) is clocked into the upper flip-flops (X_{n-2}). For maximum period, P should be chosen mutually prime to L. Otherwise, the bit rotation may be shown to be composed of several interleaved rotations of shorter words. As seen from Fig. 1, the bit rotation does not constitute a separate hardware operation. In physical terms, the exclusive-or of two flip-flops in corresponding bit positions is clocked into a third flip-flop while one of the pair of flip-flops is clocked into the other. An example of the process is given in Table I for $L = 3$, $P = 2$. (The starting values of $X_{-1} = 000$, $X_{-2} = 001$ are used here.) The period of this generator is 15. It is easily shown that

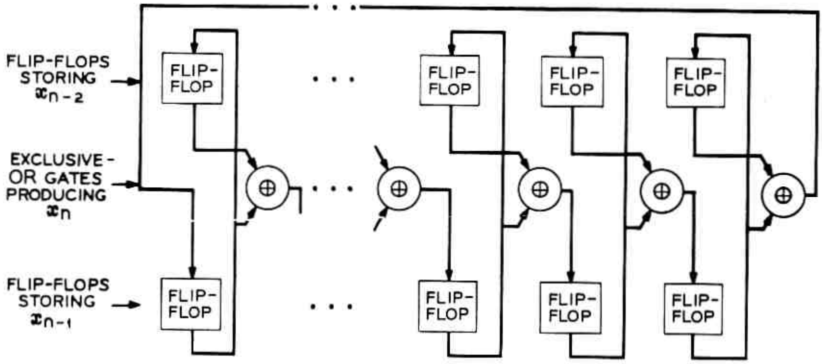

FLIP-FLOPS STORING x_{n-2}
EXCLUSIVE-OR GATES PRODUCING x_n
FLIP-FLOPS STORING x_{n-1}
FLIP-FLOP
⊕

Fig. 1—Schematic diagram showing how the random numbers are generated and stored.

with the output depending on the state of six ($2L$) flip-flops, the maximum theoretical period of any random number generator is 64, or in general (2^{2L}), and the maximum period of any random number generator using only flip-flops and exclusive-or elements is 63, or in general $(2^{2L} - 1)$, since the state when all the flip-flops are zero is succeeded only by itself.

Even though the period of the generator of Table I, and the periods

Table I—Typical Output Sequence for Noise Generator with $L = 3$, $P = 2$

Clock Number	X_{n-1} Most Recent	X_{n-2} Next Most Recent	X_n New Random Number
0	000	001	010
1	010	000	100
2	100	010	101
3	101	100	010
4	010	101	111
5	111	010	011
6	011	111	001
7	001	011	100
8	100	001	011
9	011	100	111
10	111	011	001
11	001	111	101
12	101	001	001
13	001	101	001
14	001	001	000
(15)	(repeats 000)	(repeats 001)	(repeats 010)

for other values of L, are small fractions of the theoretical maximum, it is possible to choose L to obtain a very long period. It is also possible, as we shall see, to combine the results of several generators to get still longer periods. We have theoretically predicted all the word lengths ($L \leqq 25$) for which very short periods result*, and we have measured the periods associated with the remaining values of L. The longest periods result for word lengths of $L = 11, 13, 17, 19, 22, 23$ and 25. For 25 bits, the longest period results, and is 17,825,775. However, since 2^{25} is about 3.3×10^7, not all the possible 25 bit numbers appear at the output of the generator. The computation of the period assumes that the word is rotated a number of bits mutually prime to the word length (e.g., $P = 1$) and that the starting states are reasonable (e.g., $X_{-1} = 0$, $X_{-2} = 1$). There exist unreasonable starting states, such as all zeros in one word and all ones in the other word, which have much shorter periods than those prescribed, but these can be avoided in all cases by restricting the starting values to 0 and 1.

Table II lists the periods of the generators for values of L from 1 to 25, as well as the factorization of these periods. The factorization is useful in predicting the periods associated with generators made up of two or more of these simple generators with interleaved bits. For example, it is possible to generate a 48 bit random number by interleaving the bits of a 25 bit word with the bits of a 23 bit word. The periods of the two generators may be found to have only the prime factor 3 in common, as seen from Table II. Thus the joint period (the least common multiple) is $\frac{1}{3}$ of the product of the periods of the individual generators, resulting in a period of about 2×10^{13}. Similarly a 24 bit word could be made up of an 11 bit word and a 13 bit word, with a joint period of about 2×10^9. The disparity of prime factors among several interesting cases seems surprisingly fortuitous.

III. TESTS FOR RANDOMNESS

It is clear that a long period does not by itself indicate a good random number generator (e.g. the iteration $r_n = r_{n-1} + 1$ would have a

* It is possible to equate any bit (as a function of time) to the mod 2 sum of the same bit delayed by various amounts. For example, with $L = 3$ the equation for any bit of the 3 bit word is:

$$b_n = b_{n-3} \oplus b_{n-4} \oplus b_{n-5} \oplus b_{n-6}.$$

This equation describes a particular 6 bit shift register with feedback. The analysis of such shift registers is described in Ref. 1. Specifically it is possible to obtain the maximum period associated with a given shift register, and therefore with a given random number generator, by obtaining the factors of certain characteristic polynomials over the Galois Field mod 2.

TABLE II—THE PERIOD AND ITS DECOMPOSITION INTO ITS PRIME FACTORS FOR NOISE GENERATORS WITH VALUES OF L FROM 1 TO 25

L	Period	Factors
1	3	3
2	6	(2) (3)
3	15	(3) (5)
4	12	$(2)^2$ (3)
5	255	(3) (5) (17)
6	30	(2) (3) (5)
7	63	$(3)^2$ (7)
8	24	$(2)^3$ (3)
9	315	$(3)^2$ (5) (7)
10	510	(2) (3) (5) (17)
11	33825	(3) $(5)^2$ (11) (41)
12	60	$(2)^2$ (3) (5)
13	159783	(3) (13) (17) (241)
14	126	(2) $(3)^2$ (7)
15	255	(3) (5) (17)
16	48	$(2)^4$ (3)
17	65535	(3) (5) (17) (257)
18	630	(2) $(3)^2$ (5) (17)
19	14942265	(3) (5) (13) (19) (37) (109)
20	1020	$(2)^2$ (3) (5) (17)
21	4095	$(3)^2$ (5) (7) (13)
22	67650	(2) (3) $(5)^2$ (11) (41)
23	4194303	(3) (23) (89) (683)
24	120	$(2)^3$ (3) (5)
25	17825775	(3) $(5)^2$ (11) (17) (31) (41)

very long period, but would be unacceptable to most users). A better indication of the acceptability of a random number generator is the autocorrelation function of the output of the generator, $R_x(n)$. It is difficult to obtain $R_x(n)$ theoretically for the generators described here; so instead we have estimated $R_x(n)$ by standard techniques for several cases of interest. In Fig. 2 we show the estimated autocorrelation function for the generator with $L = 13$. Approximately $N = 15000$ samples were used in the estimate, and Fig. 2 shows the results for up to 512 delays. It seems clear that there are no irregularities present in the autocorrelation function. The peak values of the autocorrelation function of Fig. 2, for $n \neq 0$, are ± 0.026. It is easily shown that estimates of the autocorrelation function (for $n \neq 0$) tend to be normally distributed random variables with zero mean, and variance of $1/N$. For the data of Fig. 2, the standard deviation, σ, was calculated to be 0.008. Cramer[2] shows that the expected value of the upper extreme of 512 values from a normal population with zero mean, and standard deviation of σ, is approximately 3.25 σ, or 0.026 in this example. The 50 percent

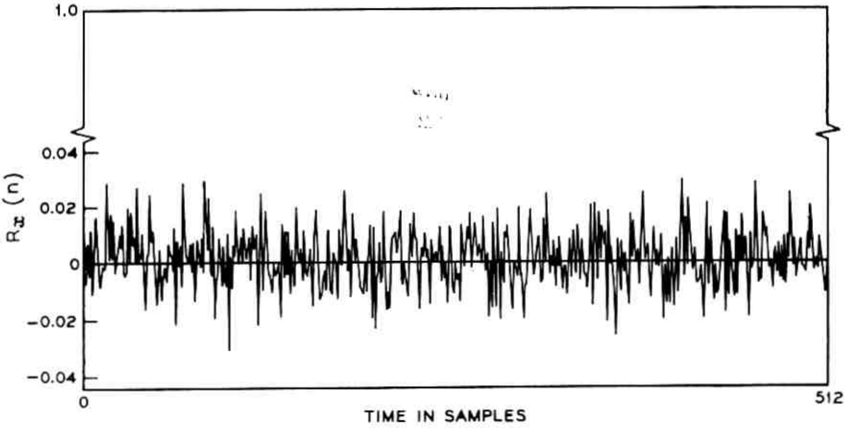


Fig. 2—Autocorrelation function of noise generator with $L = 13$.

confidence interval for the upper extreme is about 0.4 σ wide, thus peak values of the autocorrelation estimates from 0.024 to 0.028 would be quite common. Therefore the peak values of ±0.026 in Fig. 2 are not inconsistent with the above theoretical results based on a true normal population.

To empirically test the uniformity of the output of the generators for $L = 11$ and $L = 13$, we measured the number of occurrences of each of the 2^L output states during a single period. For both cases all states were present at the output of the generator. In Fig. 3, we show plots for $L = 11$ and 13 of the number of occurrences of each of the 128 cells specified by the 7 most significant bits of the output as a function of cell number. The upper plot shows the measured result for $L = 11$, and the lower plot shows the measured result for $L = 13$. The solid lines across the plots indicate the expected number of occurrences of each state based on uniformity assumptions. The plots of Fig. 3 tend to validate the assertion that the amplitude distribution of the output of the noise generator is uniform for certain values of L.

We have also measured the mean value for the entire sequence for $L = 11$ and $L = 13$ and it is near to 2^{L-1} if the L bits are interpreted as a positive integer. The nearest means obtained were 1024.3169 for $L = 11$ and 4095.8326 for $L = 13$. (Starting values for the two sequences were $X_{-1} = 341$, $X_{-2} = 0$ for $L = 11$, and $X_{-1} = 151$, and $X_{-2} = 0$ for $L = 13$.) Also, it was found that a scatter diagram from the generator with $L = 17$ showed no tendencies to order, such as are common

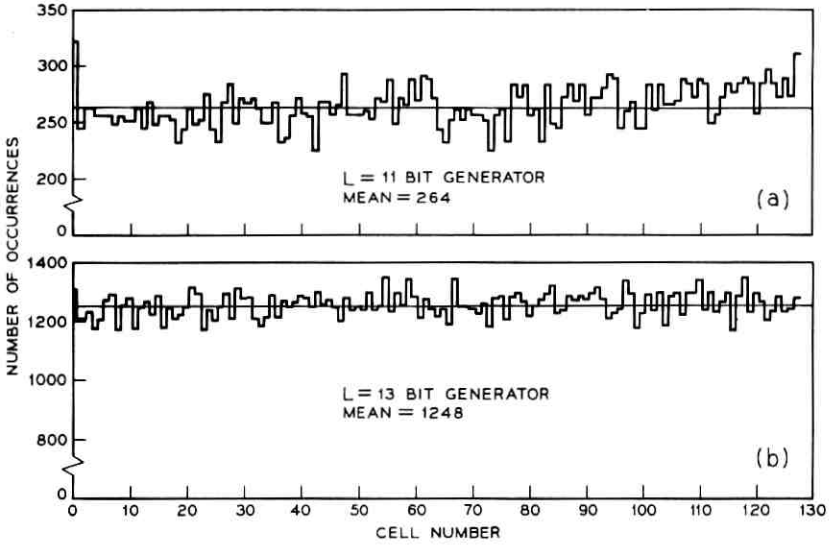


Fig. 3—Measured distribution functions for noise generators with $L = 11$ and $L = 13$.

in the simple multiplicative congruence generators. It is expected that this result would be true for any of the generators with reasonably long periods.

It was stated earlier that it is possible to generate longer random numbers than 25 bits by interleaving bits of shorter generators. Besides the prerequisite that the periods of the individual generators have few or no prime factors in common, it is important that the outputs of the individual noise generators be uncorrelated. To check whether

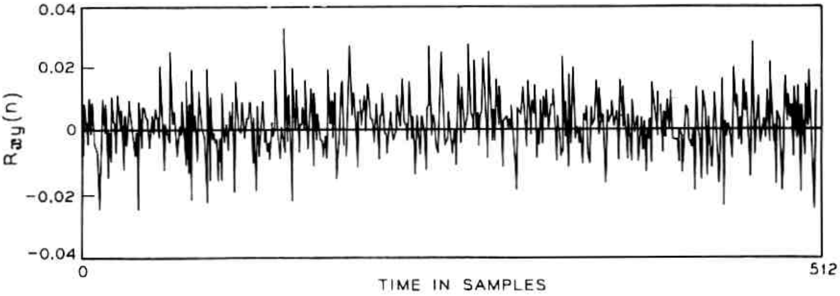


Fig. 4—Cross-correlation function of noise generators with $L = 11$ and $L = 13$.

or not the individual outputs of the noise generators, for the desirable values of L, were uncorrelated, we again used standard statistical techniques to measure the cross correlation function. Figure 4 shows the cross-correlation function, $R_{xy}(n)$, for the case where one input was the output of the generator with $L = 11$, and the other input was the output of the generator with $L = 13$. The cross-correlation function is plotted for delays up to 512 samples, and is again based on approximately $N = 15{,}000$ samples. There are no apparent irregularities seen in this figure, and the peak values of the cross-correlation function, ± 0.027, are quite close to similar peaks observed for the autocorrelation function of Fig. 2, and again consistent with the assumption that these 512 estimates are from a normal population with $\sigma = 0.008$.

The reader should be cautioned that a poor choice of the rotation P can cause an ordering in the pseudorandom outputs which may be harmful for some applications. For example, consider the set of all triples (x_n, x_{n+1}, x_{n+2}) when $P = 1$. If the sign bits (using two's complement integer notation) of x_n and x_{n+1} are the same, the most significant bit of x_{n+2} will always be zero. Thus, if (x_n, x_{n+1}) lies in quadrants 1 or 3, x_{n+2} is constrained to either

$$0 \leqq x_{n+2} < 2^{L-2}$$

or

$$-2^{L-1} \leqq x_{n+2} < -2^{L-2}.$$

A similar constraint results when (x_n, x_{n+1}) lies in quadrants 2 or 4. This effect is eliminated by choosing $P \approx L/2$, but mutually prime to L. We have not considered what ordering might exist in quadruples, quintuples, etc., for various choices of P.

IV. CONCLUSION

In conclusion, we have presented a fast and efficient technique for generating digital random numbers. The simple statistical tests to which we have subjected several of these noise generators indicate they are more than adequate for use in simulation programs for communications systems.

REFERENCES

1. Golomb, S., *Shift Register Sequences,* San Francisco: Holden-Day Inc., 1967.
2. Cramér, H., *Mathematical Methods of Statistics,* Princeton, N. J.: Princeton Univ. Press, 1946, pp. 363–378.

Nonorthogonal Optical Waveguides and Resonators

By J. A. ARNAUD

(Manuscript received May 4, 1970)

The modes of propagation in optical systems which do not possess meridional planes of symmetry (nonorthogonal systems) are investigated in the case where the effect of apertures and losses can be neglected. The fundamental mode of propagation is obtained with the help of a complex ray pencil concept. An integral transformation of the field, based on a quasi-geometrical optics approximation and a first-order expansion of the point characteristic of the optical system, is given; it shows that the complex (three-dimensional) wavefront of the fundamental mode is transformed according to a generalized "ABCD law." A simple expression is also obtained for the phase-shift experienced by the beam. The higher order modes of propagation are obtained from a power series expansion of the fundamental mode. These higher order modes are expressed, in oblique coordinates, as the product of the fundamental solution and finite series of Hermite polynomials with real arguments. In the special case of systems with rotational symmetry, these series reduce to the well-known generalized Laguerre polynomials. The theory is applicable to media such as helical gas lenses and optical waveguides suffering from slowly varying deformations in three dimensions. Nonorthogonal resonant systems are also investigated. An expression for the resonant frequencies, applicable to any three-dimensional resonator, is derived. Numerical results are given for the resonant frequencies and the resonant field of a twisted path cavity which exhibits interesting properties: the usual polarization degeneracy is lifted and the intensity pattern of all of the modes possesses a rotational symmetry.

I. INTRODUCTION

An optical system, or a resonator, is called "nonorthogonal" when it is not possible to define two mutually orthogonal meridional planes of symmetry (Ref. 1, p. 240). The helical gas lens[2,3] is an example of a nonorthogonal lenslike medium. A conventional ring type cavity generally ceases to be orthogonal when its path is twisted, i.e., becomes nonplanar.[4]

Let us briefly review the major approaches in the theory of optical resonators. The field in a resonator can be expressed exactly in terms of known functions only for a few simple boundary surfaces. No exact solution is available for nonorthogonal systems. However, we are interested only in the high frequency operation of large resonators. In that limit, the waves have a tendency to follow closed curves in the resonator, either clinging to the concave parts of the boundary (whispering gallery modes[5]) or connecting opposite points of the boundary (bouncing ball modes). One defines the axial mode number as the number of wavelengths existing along such closed curves. The nodes of the field in the transverse planes define the transverse mode numbers. More insight concerning the mode structure and the resonant frequencies can be gained by using a geometrical optics approximation, or a paraxial form of the Huygens diffraction principle. The geometrical optics approach was developed by Keller and Rubinow.[6] It consists of setting up in the resonator a manifold of rays tangent to a caustic. The location of the caustic and the resonant frequencies are obtained from the condition that the variations of the eikonal along three independent closed curves are equal to an integral number of wavelengths (or an integer plus one-half or one-quarter). This theory, which is analogous to the Born approximation of quantum mechanics, gives the exact resonant frequencies of paraxial modes. The geometrical optics field, when extended in the shadow of the caustic by analytic continuation, provides an acceptable approximation to the exact field for large transverse mode numbers but, for the fundamental mode, it differs vastly from the exact field. The caustic line however, does coincide, in two dimensions, with the mode profile.[7,8] This geometrical optics method has been extended to nonorthogonal resonators incorporating homogeneous media by Popov,[9] who gave an expression for the resonant frequencies. Within the paraxial approximation, exact solutions for the field can be obtained from the Huygens principle; for that reason, the geometrical optics method, in spite of its general interest, will not be discussed further in this paper.

For the case of resonators incorporating inhomogeneous media, the Huygens principle must be supplemented by a quasi-geometrical optics approximation. This approximation consists of assuming that a point source at the input plane of the system creates at the output plane a field which can be adequately represented by the geometrical optics field. This approximation is generally applicable to optical waveguides and resonators if one disregards the effect of apertures and assumes that no diffraction gratings or other wavelength-dependent scatterers are present. This quasi-geometrical optics method provides an integral transformation for the field which is equivalent to a partial differential equation of the parabolic type (see Section II). The similarity between this parabolic equation and the Schroedinger equation has often been pointed out.[10–13] The matched modes of propagation in uniform lens-like media with hyperbolic secant refractive index laws, for instance, can be found in Landau and Lifshits' *Quantum Mechanics*[14] [whereas the ray trajectories are given in Ref. (1), p. 179]. The more general problem of unmatched beams in nonuniform lens-like media corresponds to the time-dependent Schroedinger equation with time-varying potentials. The adiabatic approximation usually applied to this problem, is based on conditions[12] which are too stringent for most optical systems. Generalized modes, where allowance is made for a wavefront curvature, were introduced by Goubau and Schwering[15] and Pierce[16] for the free-space case, in agreement with the theory of confocal resonators proposed by Boyd and Gordon.[17] These results were extended to orthogonal square law media.[18,19,20] The transformation of the complex curvature of beams through arbitrary optical systems with rotational symmetry and the resonant frequency of linear cavities was obtained by Kogelnik.[21,22] Vlasov and Talanov[23] have observed that, in two dimensions, the phase shift experienced by a matched beam in an optical system is equal to the phase of one of the two ray-matrix eigenvalues. This result is easily demonstrated and generalized to astigmatic orthogonal systems by using a complex ray pencil concept.[4,24]

The generalization to nonorthogonal systems is substantially more intricate. Arnaud and Kogelnik[24] have obtained a generalized gaussian mode of propagation in free space by giving complex values to the three parameters which define an astigmatic ray pencil, i.e., the position of the focal lines and the angular orientation of one of them. This solution can be used to obtain the beam transformation in a sequence of thin astigmatic lenses arbitrarily oriented, by matching the complex wavefronts at each lens. This method does not give, however, a general expression for the phase shift experienced by the beam, knowledge of which is essential in studying resonators. For that reason, a somewhat different approach is used here, where the ray pencil is defined by two of its rays. The field of the fundamental mode of propagation is obtained (Section III) by allowing these two rays to assume complex positions while remaining solutions of the ray equations.

The higher order modes of propagation are studied in Section IV. They are obtained by application of differential operators related to those used in quantum mechanics. An oblique coordinate system is introduced which diagonalizes the complex wavefront of the fundamental mode. In this oblique coordinate system, the higher order modes can be expressed as the product of the fundamental solution and finite series of Hermite polynomials with real arguments. An alternative procedure is also given which leads to Hermite polynomials in two complex variables. The simple formula for the resonant frequencies of linear resonators given by Popov[9,26] is shown to be applicable to ring type resonators incorporating inhomogeneous media (Section V). Finally these general results are applied to a new type of optical resonator called "cavity with image rotation" which presents interesting resonance and polarization properties (Section VI). Numerical results are presented.

The present theory is limited to paraxial first-order solutions in loss-less isotropic media. As indicated before, it is assumed that no apertures or diffraction gratings are present in the system, and the problem of mode selection is not discussed. The electromagnetic field is treated as a scalar quantity and the polarization effects are introduced only at a later stage; this is permissible within the paraxial approximation. Fresnel reflection at surfaces of discontinuity is also neglected.

II. PARABOLIC WAVE EQUATION AND INTEGRAL TRANSFORMATION OF THE FIELD

In this section an approximate form of the scalar Helmholtz equation is derived which is applicable to paraxial beams, i.e., to beams propagating at small angles with respect to the system axis. It is subsequently compared to an integral transformation derived from Huygens principle.

The scalar Helmholtz equation can be written in a x_1, x_2, z rectangular coordinate system

$$\frac{\partial^2 E}{\partial x_1^2} + \frac{\partial^2 E}{\partial x_2^2} + \frac{\partial^2 E}{\partial z^2} + k^2 n^2(x_1, x_2, z)E = 0, \tag{1}$$

where E is a component of the field and $n(x_1, x_2, z)$ the refractive index of the medium. Let us introduce a reduced field

$$\psi(x_1, x_2, z) = E(x_1, x_2, z) \exp\left[jk \int_0^z n(0, 0, z)\, dz\right], \tag{2}$$

and neglect the second derivative of ψ with respect to z. This approximation physically means that only waves propagating in a direction close to the z axis are considered. Denoting $n(0, 0, z)$ by n_0, for brevity, one obtains

$$\frac{\partial^2 \psi}{\partial x_1^2} + \frac{\partial^2 \psi}{\partial x_2^2} - 2jkn_0 \frac{\partial \psi}{\partial z} - jk\psi \frac{dn_0}{dz} + k^2(n^2 - n_0^2)\psi = 0. \tag{3}$$

This equation can be simplified if one introduces the following changes of function and variables[24]

$$\Psi = n_0^{\frac{1}{2}}\psi, \tag{4}$$

$$\zeta = \int_0^z dz/n_0 . \tag{5}$$

One obtains

$$\frac{\partial^2 \Psi}{\partial x_1^2} + \frac{\partial^2 \Psi}{\partial x_2^2} - 2jk \frac{\partial \Psi}{\partial \zeta} + k^2(n^2 - n_0^2)\Psi = 0. \tag{6}$$

Let us further assume that $n^2 - n_0^2$ is a quadratic form in x_1, x_2

$$n^2 = n_0^2 + n_{11}x_1^2 + 2n_{12}x_1x_2 + n_{22}x_2^2 . \tag{7a}$$

n_{11}, n_{12} and n_{22} are *real* functions of z since the losses in the medium are neglected. The quadratic form given in equation (7a) describes a nonorthogonal optical system when the directions of its axes change as z varies. In that case, the diagonal term $2n_{12}x_1x_2$ cannot be eliminated by rotating the coordinate system about z. We discuss this general case.

Let us rewrite equation (7a), for brevity, in matricial form

$$n^2 = n_0^2 + \tilde{r}\eta r \tag{7b}$$

where r denotes a column matrix with elements x_1, x_2 and η denotes a 2 × 2 real symmetrical matrix. The sign $\sim$ indicates a transposition. Inserting equation (7b) in equation (6), the wave equation assumes the form

$$\mathfrak{L}\Psi \equiv \left(\nabla^2 - 2jk \frac{\partial}{\partial \zeta} + k^2\tilde{r}\eta r\right)\Psi = 0 \tag{8}$$

where ∇^2 denotes the laplacian operator in the transverse x_1, x_2 plane.

It is henceforth assumed that $n^2 - n_0^2$ is small compared with unity. Within this (first-order) approximation, the refractive index law, equation (7b), becomes

$$n \simeq n_0 + \tilde{r}\eta r/(2n_0). \tag{9}$$

Let us now consider the ray trajectories. A ray $\mathcal{R}$ is defined at any transverse plane z by its position $q(z)$ and by the projection $p(z)$ on that plane of a vector directed along the ray, of length equal to the refractive index n. $q(z)$ and $p(z)$ are called respectively the position vector and the direction vector of the ray. It is convenient to represent these vectors by column matrices whose elements are the vector components on x_1 , x_2 . As long as only fixed coordinate systems are used, such matrices can be denoted without ambiguity $q(z)$ and $p(z)$, or simply q and p. The exact ray equations are (see, for instance, Ref. 1, p. 90)

$$\dot{p} = -n_0 \nabla H(r, p), \tag{10a}$$

$$\dot{q} = n_0 \nabla_p H(r, p), \tag{10b}$$

at $r = q$. In equation (10) the upper dots denote differentiations with respect to ζ, and $H(r, p)$ denotes the Hamiltonian of the system defined by

$$H(r, p) = -(n^2 - \tilde{p}p)^{\frac{1}{2}}; \tag{10c}$$

∇ denotes the gradient operator in the transverse x_1 , x_2 plane, and ∇_p denotes a gradient operator relative to the p variables. Within the first order approximation [equation (9)], equations (10c), (10a) and (10b) reduce respectively to

$$H(r, p) = -n_0 - (\tilde{r}\eta r - \tilde{p}p)/(2n_0), \tag{11c}$$

$$\dot{p} = \eta q, \tag{11a}$$

and

$$\dot{q} = p. \tag{11b}$$

Equations (11a) and (11b) are called the paraxial ray equations.

Let us now consider two arbitrary rays, $\mathcal{R}$ and $\hat{\mathcal{R}}$, defined by their position and direction vectors q, p and $\hat{q}$, $\hat{p}$, respectively, and let the "product" of these two rays be defined by the scalar expression

$$(\mathcal{R}; \hat{\mathcal{R}}) \equiv \tilde{q}\hat{p} - \tilde{\hat{q}}p. \tag{12}$$

$(\mathcal{R}; \hat{\mathcal{R}})$ is sometimes called the *Lagrange invariant* (see Ref. 1, p. 251).

It is easy to show that this quantity is independent of ζ (or z). Indeed, applying equations (11a) and (11b) to both $\mathcal{R}$ and $\hat{\mathcal{R}}$, and remembering that η is a symmetric matrix, one obtains†

$$\frac{d}{d\zeta}(\mathcal{R};\hat{\mathcal{R}}) \equiv \frac{d}{d\zeta}(\tilde{q}\hat{p} - \tilde{\hat{q}}p) = \dot{\tilde{q}}\hat{p} + \tilde{q}\dot{\hat{p}} - \dot{\tilde{\hat{q}}}p - \tilde{\hat{q}}\dot{p} = 0. \qquad (13)$$

The Lagrange invariant $(\mathcal{R};\hat{\mathcal{R}})$ plays an important role in the present theory. Notice that n_0 does not appear explicitly in equations (8), (11a) and (11b). It can therefore be assumed, without loss of generality, that $n_0 \equiv 1$.

The properties of propagating beams are sometimes more easily understood by considering the transformation of the field between the input plane and the output plane of an optical system described by its point characteristic. Let us now choose as optical axis, for generality, an arbitrary ray $\mathcal{A}$ which need not be a straight line nor even a plane curve. Let us further define, at a distance z' from an origin 0, a rectangular coordinate system x_1', x_2', whose axes are oriented respectively along the principal normal and the binormal to $\mathcal{A}$ (see Fig. 1). At any given transverse plane, a ray is defined by its position vector q and its direction vector p. Let us assume that there is one ray, and only one ray which goes from a point r at $z = 0$ (input plane) to a point r' at $z = z'$ (output plane). This assumption implies, in particular, that the planes $z = 0$ and $z = z'$ are not conjugate. The optical length $\mathcal{V}(r, r')$ of such a ray is called the point characteristic of the optical system. As is well known, the direction vectors of a ray can be obtained from $\mathcal{V}$ by differentiation (Ref. 1, p. 97)

$$p = -\nabla\mathcal{V}(r, r'), \qquad (14a)$$

$$p' = \nabla'\mathcal{V}(r, r'), \qquad (14b)$$

at $r = q$, $r' = q'$. The primes always denote quantities at the output plane $z = z'$.

The law of transformation of the field can be obtained from the Huygens principle supplemented by a quasi-geometrical optics approximation.[28] The Huygens principle states that each point of an incident wavefront can be considered as the source of a secondary wave. The quasi-geometrical optics approximation consists of assuming that the field created at the output plane of the system by a point source at the input plane is adequately represented by the geometrical optics

† Recall also that, for any conformable matrices a and b, $(ab)^{\sim} = \tilde{b}\tilde{a}$ and that, for any scalar (one element matrix) c, we have $\tilde{c} \equiv c$.

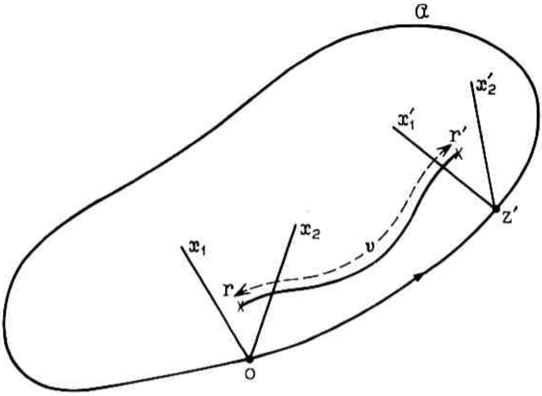


Fig. 1—Optical axis of a ring type resonator. $\mathcal{V}$ denotes the point characteristic of the system included between two transverse planes, $z = 0$ and $z = z'$.

field. These two assumptions allow us to express the field $E'(r')$ at the output plane as a function of the field $E(r)$ at the input plane. Within the paraxial approximation, we have

$$E'(r') = \pm\lambda^{-1} \iint_{-\infty}^{+\infty} E(r)K(r, r')\, d^2r, \tag{15a}$$

where

$$K(r, r') \equiv |\, \partial^2\mathcal{V}/\partial x_i\, \partial x'_j \,|^{\frac{1}{2}} \exp\left[-jk\mathcal{V}(r, r')\right]. \tag{15b}$$

The term $|\, \partial^2\mathcal{V}/\partial x_i\, \partial x'_j \,|^{\frac{1}{2}}$, where the bars denote a determinant, is obtained by recognizing that the power flowing through a small area at the output plane is equal to the power flowing in the corresponding cone of rays leaving the point source at the input plane, and using equation (14a).

To first order, the quantity $\mathcal{S} \equiv \mathcal{V} - z'$ is a quadratic form in x_1, x_2, x'_1, x'_2 which can be written, in matricial notation

$$\mathcal{S} = \tfrac{1}{2}(\tilde{r}Ur + \tilde{r}Vr' + \tilde{r}'\tilde{V}r + \tilde{r}'Wr'), \tag{16}$$

where U and W are 2×2 symmetric real matrices and V is a 2×2 real matrix. Equation (16) can be rewritten, more concisely

$$\mathcal{S} = \tfrac{1}{2}[\tilde{r}\; \tilde{r}'] \begin{bmatrix} U & V \\ \tilde{V} & W \end{bmatrix} \begin{bmatrix} r \\ r' \end{bmatrix} \equiv \tfrac{1}{2}[\tilde{r}\; \tilde{r}'][\mathcal{S}] \begin{bmatrix} r \\ r' \end{bmatrix}. \tag{17}$$

Introducing equation (16) in equations (14a) and (14b), one obtains linear relations between p, p' and q, q' in the form

$$\begin{bmatrix} -p \\ p' \end{bmatrix} = \begin{bmatrix} U & V \\ \tilde{V} & W \end{bmatrix} \begin{bmatrix} q \\ q' \end{bmatrix} \equiv [\mathcal{S}] \begin{bmatrix} q \\ q' \end{bmatrix}. \tag{18}$$

It is sometimes convenient to introduce a *ray matrix* which relates q', p' to q, p. Simple relations exist between $[\mathcal{S}]$ and the ray matrix; they are given in Appendix A.

Let us now go back to the integral transformation and observe that, if $\mathcal{S}$ is a quadratic form [equation (16)], the determinant

$$| \partial^2 \mathcal{V}/\partial x_i \, \partial x_j' | = | \partial^2 \mathcal{S}/\partial x_i \, \partial x_j' | = | V | \tag{19}$$

is independent of r and r'. This term can consequently be taken out of the integral in equation (15). The integral transformation of the reduced field ψ [$\psi \equiv E \exp(jkz)$] becomes

$$\psi'(r') = \pm \lambda^{-1} | V |^{\frac{1}{2}} \iint_{-\infty}^{+\infty} \psi(r) \exp(-jk\mathcal{S}) \, d^2 r, \tag{20}$$

whose kernel is essentially

$$K_0 \equiv | V |^{\frac{1}{2}} \exp(-jk\mathcal{S}). \tag{21}$$

Let us show that, in a rectangular coordinate system, K_0 represents the Green function of the parabolic wave equation, equation (8), i.e., that

$$\mathcal{L}'K_0 \equiv \left(\nabla'^2 - 2jk \frac{\partial}{\partial z'} + k^2 \tilde{r}' \eta' r' \right) [| V |^{\frac{1}{2}} \exp(-jk\mathcal{S})] = 0. \tag{22}$$

The first term in equation (22) can be written, using equation (16)

$$\begin{aligned} \nabla'^2 [| V |^{\frac{1}{2}} & \exp(-jk\mathcal{S})] \\ &= | V |^{\frac{1}{2}} (-jk\nabla'^2 \mathcal{S} - k^2 \nabla' \mathcal{S} \cdot \nabla' \mathcal{S}) \exp(-jk\mathcal{S}) \\ &= | V |^{\frac{1}{2}} \exp(-jk\mathcal{S})(-jk \,\mathrm{Spur}\, W - k^2 \nabla' \mathcal{S} \cdot \nabla' \mathcal{S}). \end{aligned} \tag{23}$$

To evaluate the second term in equation (22) one needs to know the derivative of $\mathcal{S}$ with respect to z'. We have (Ref. 1, p. 97)

$$\frac{\partial \mathcal{V}}{\partial z'} = -H(r', p') \cong 1 + (\tilde{r}' \eta' r' - \tilde{p}' p')/2, \tag{24}$$

where the paraxial approximation of H, equation (11c), has been used. Therefore, introducing the expression for p', equation (14b) in equation (24), one obtains

$$\frac{\partial \mathcal{S}}{\partial z'} = \frac{\partial \mathcal{V}}{\partial z'} - 1 = (\tilde{r}' \eta' r' - \nabla' \mathcal{S} \cdot \nabla' \mathcal{S})/2. \tag{25}$$

One also needs to know the derivative of V with respect to z'. It is obtained by introducing the quadratic form, equation (16), in both sides of equation (25)

$$\tilde{r}\,\frac{dU}{dz'}\,r + 2\tilde{r}\,\frac{dV}{dz'}\,r' + \tilde{r}'\,\frac{dW}{dz'}\,r' = \tilde{r}'\eta' r' - (\tilde{r}V + \tilde{r}'W)(\tilde{V}r + Wr'). \qquad (26)$$

Equation (26) shows, upon identification, that

$$\frac{dV}{dz'} = -VW. \qquad (27)$$

Therefore (see Ref. 29)

$$\frac{d}{dz'}\,|\,V\,|^{\frac{1}{2}} = \tfrac{1}{2}\,|\,V\,|^{-\frac{1}{2}}\,\frac{d}{dz'}\,|\,V\,| = \tfrac{1}{2}\,|\,V\,|^{\frac{1}{2}}\,\mathrm{Spur}\left(V^{-1}\,\frac{dV}{dz'}\right)$$
$$= -\tfrac{1}{2}\,|\,V\,|^{\frac{1}{2}}\,\mathrm{Spur}\,W. \qquad (28)$$

Upon substitution of equations (23), (25) and (28), one finds that equation (22) is satisfied.

Consequently, within the first-order approximation, one may use indifferently the parabolic wave equation, equation (8), or the integral transformation, equation (20). Most of the demonstrations given in the following sections are based on both formulations.

III. FUNDAMENTAL MODE OF PROPAGATION

We know that in the high frequency limit, propagating beams closely resemble ray pencils. Let us therefore consider first the field of such ray pencils, and subsequently see how this solution can be generalized to take into account diffraction effects.

A ray pencil is, in general, astigmatic; it can be defined, in free space, as the manifold of rays which intersect two mutually perpendicular focal lines. At any point, a surface exists, called the wavefront, which is perpendicular to all of these rays. The field of ray pencils propagating in inhomogeneous media can be written in a x_1, x_2, z rectangular coordinate system

$$E(x_1\,,\,x_2\,,\,z) = Ae^{-ikS}, \qquad (29)$$

where A and S are real functions of x_1, x_2 and z. A is an amplitude factor and S is called the eikonal of the geometrical optics field. The surfaces S = constant are the equations of the wavefronts associated with the manifold of rays. Let us assume that one of the rays coincide with the z-axis and that the refractive index of the medium is unity on that axis. Within the first-order approximation, $\Phi \equiv S - z$ is a quadratic form in the transverse variables x_1 and x_2, whose coefficients are slowly varying functions of z, and A is independent of x_1, x_2. Φ can be written, in matrix notation

$$\Phi(r, z) \equiv \tfrac{1}{2}\tilde{r}\mu(z)r, \tag{30}$$

where $\mu(z)$ is a 2×2 symmetrical matrix which generally depends on z. The law of conservation of power dictates that A and μ cannot be independent; a wavefront with a positive curvature, for instance, corresponds to a contraction of the ray pencil as z increases, which necessarily results in an increased intensity. To express this relation between A and μ (transport equation), let us choose any two rays of the ray pencil such as $\mathcal{R}$ and $\hat{\mathcal{R}}$. Since $\mathcal{R}$ and $\hat{\mathcal{R}}$ are both perpendicular to the wavefront, one has, from equation (30)

$$p = \nabla\Phi(q) = \mu q, \tag{31a}$$

$$\hat{p} = \nabla\Phi(\hat{q}) = \mu\hat{q}, \tag{31b}$$

where ∇ denotes as before the gradient operator in the x_1, x_2 plane. Equations (31a and b) can be written more concisely:

$$P = \mu Q, \tag{32a}$$

where we have defined

$$Q = [q\ \hat{q}], \tag{32b}$$

$$P = [p\ \hat{p}]. \tag{32c}$$

Equations (31) and (32) show that the product of $\mathcal{R}$ and $\hat{\mathcal{R}}$, defined in equation (12), is equal to zero at any plane

$$(\mathcal{R}; \hat{\mathcal{R}}) \equiv \tilde{q}\hat{p} - \tilde{\hat{q}}p = 0. \tag{33}$$

Any ray defined by a linear combination of $\mathcal{R}$ and $\hat{\mathcal{R}}$ also belongs to the ray pencil since its product with either $\mathcal{R}$ or $\hat{\mathcal{R}}$ is equal to zero. Therefore, the one-parameter manifold of rays $\epsilon\mathcal{R}$, $\epsilon\mathcal{R} + \hat{\mathcal{R}}$, $\epsilon\hat{\mathcal{R}}$, $\mathcal{R} + \epsilon\hat{\mathcal{R}}$, with $0 < \epsilon < 1$, defines a tube of rays in the ray pencil whose cross section is a parallelogram with sides ϵq, $\epsilon q + \hat{q}$, $\epsilon\hat{q}$ and $q + \epsilon\hat{q}$ (see Fig. 2). The area of this parallelogram is given by the length of the vector product of q and $\hat{q}$

$$h = q_1\hat{q}_2 - q_2\hat{q}_1 = |\,Q\,|. \tag{34}$$

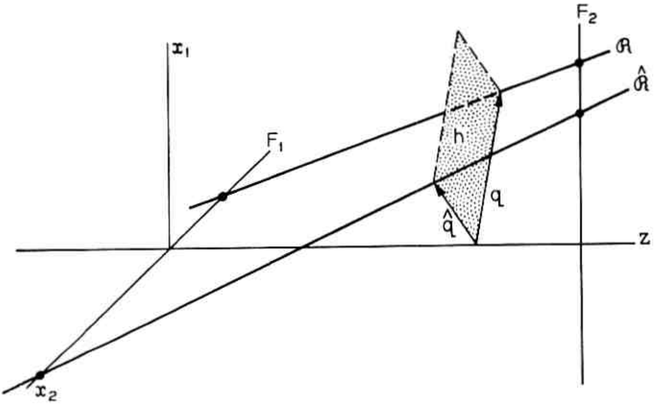


Fig. 2—An astigmatic ray pencil is defined in free space by the manifold of rays which intersect two mutually perpendicular focal lines such as F_1 and F_2. At any transverse plane the intensity of the field is inversely proportional to the square root of the area defined by q and $\hat{q}$, the position vectors of any two rays of the ray pencil ($\mathcal{R}$ and $\hat{\mathcal{R}}$).

Conservation of power requires that $A^2(z)h(z)$ be a constant. $A(z)$ can therefore be obtained from equation (34). Notice that, at a focal line, the sign of $h(z)$ changes from positive to negative. Therefore $A(z) \propto [h(z)]^{-\frac{1}{2}}$ becomes imaginary. If one insists on keeping $A(z)$ real, a $\pi/2$ phase shift must be subtracted from S at such points (anomalous phase shift).

The elements of the wavefront matrix μ can also be obtained from the components of two rays satisfying equation (33). One obtains, solving for μ equation (32a)

$$\mu_{11} = (\hat{q}_2 p_1 - q_2 \hat{p}_1)h^{-1}, \tag{35a}$$

$$\mu_{22} = (q_1 \hat{p}_2 - \hat{q}_1 p_2)h^{-1}, \tag{35b}$$

$$\begin{aligned} \mu_{12} = \mu_{21} &= (q_1 \hat{p}_1 - \hat{q}_1 p_1)h^{-1}, \\ &= (\hat{q}_2 p_2 - q_2 \hat{p}_2)h^{-1}. \end{aligned} \tag{35c}$$

The reduced field of the ray pencil is therefore

$$\psi(r, z; \mathcal{R}, \hat{\mathcal{R}}) = \pm h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right), \tag{36}$$

where h and μ are given by equations (34) and (35a, b and c) respectively. The sign ambiguity in the expression of ψ can be resolved only by counting the number of focal lines along the ray pencil, from some reference plane.

Let us now show that the field of a ray pencil, as given by equation (36), is a solution of the parabolic wave equation, equation (8), i.e., that

$$\mathcal{L}\psi(r, z; \mathcal{R}, \hat{\mathcal{R}}) \equiv \left(\nabla^2 - 2jk\frac{\partial}{\partial z} + k^2\tilde{r}\eta r\right) \cdot \left[h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right)\right] = 0. \tag{37}$$

The first term on the right side of equation (37) is

$$\nabla^2\left[h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right)\right] = h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right) \times (-jk \operatorname{Spur} \mu - k^2\tilde{r}\mu^2 r). \tag{38}$$

The second term is

$$-2jk\frac{\partial}{\partial z}\left[h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right)\right] = h^{-\frac{1}{2}} \exp\left(-j\frac{k}{2}\tilde{r}\mu r\right) \times (jk\dot{h}/h - k^2\tilde{r}\dot{\mu}r). \tag{39}$$

Using now equations (34), (32a) and (11b), one notices that

$$\dot{h}/h = \operatorname{Spur} \mu. \tag{40}$$

Differentiating both sides of equations (31a and b) with respect to z and using the paraxial ray equations [equations (11a) and (11b)] one obtains

$$(\dot{\mu} + \mu^2 - \eta)q = 0, \tag{41a}$$

$$(\dot{\mu} + \mu^2 - \eta)\hat{q} = 0. \tag{41b}$$

Since q and $\hat{q}$ are generally linearly independent, it results from equation (41) that

$$\dot{\mu} + \mu^2 = \eta. \tag{42}$$

Upon substitution of equations (38), (39), (40) and (42) in equation (37), one finds that the field of a paraxial ray pencil is, as expected, a solution of the parabolic wave equation.

It is important to remark that it has nowhere been specified that q, p, $\hat{q}$ and $\hat{p}$ are real quantities. The right side of equation (36) therefore remains a solution of the wave equation if $\mathcal{R}$ and $\hat{\mathcal{R}}$ are allowed to be complex valued while remaining solutions of the paraxial ray equations.

In that case, $\mu(z)$, whose elements are given by equations (35a, b and c), becomes a complex matrix and the exponential term in equation (36) describes the intensity pattern of the beam as well as its wavefront. As observed before[25] the axes of the constant intensity ellipse do not coincide, in general, with the axes of the wavefront surface. It is possible, however, to define at any plane an oblique coordinate system in which both the real part of μ, corresponding to the beam wavefront, and the imaginary part of μ, corresponding to the beam intensity, are diagonal. This coordinate transformation is given at the end of this section and used in Section IV to express in a convenient form the higher order modes of propagation.

$h(z)$, given by equation (34), and therefore the amplitude term $A(z)$, become complex quantities too. The $\pm j$ ambiguity pointed out for the case of ray pencils does not exist any more since the phase of $A(z)$ changes in a continuous manner along the z axis.

Let us now consider an optical system described by its point characteristic matrix $[\mathbb{S}]$ and calculate the transformation experienced by an incident gaussian beam whose reduced field has the form given in equation (36). Introducing this expression in equation (20), one obtains a reduced field at the output plane

$$\psi'(r') = \pm\lambda^{-1} \mid V \mid^{\frac{1}{2}} h^{-\frac{1}{2}}$$
$$\cdot \iint_{-\infty}^{+\infty} \exp\left\{-j\,\frac{k}{2}\,[\tilde{r}(U + \mu)r + 2\tilde{r}Vr' + \tilde{r}'Wr']\right\} d^2r. \tag{43}$$

The integral in equation (43) is easily integrated if one notices that, for any nonsingular square matrix m and any comformable column matrices r and s one has

$$\tilde{r}mr + 2\tilde{r}s = (\tilde{r} + \tilde{s}m^{-1})m(r + m^{-1}s) - \tilde{s}m^{-1}s. \tag{44}$$

Using equation (44) one finds that, if m is a symmetric matrix

$$\iint_{-\infty}^{+\infty} \exp\left[-(\tilde{r}mr + 2\tilde{r}s)\right] d^2r = \pi \mid m \mid^{-\frac{1}{2}} \exp\,(\tilde{s}m^{-1}s), \tag{45}$$

provided the integral is defined, i.e., provided: $\tilde{r}$ (real part of m) r is a positive definite form. Substituting

$$m = j\,\frac{k}{2}\,(U + \mu) \tag{46a}$$

and

$$s = j\,\frac{k}{2}\,Vr' \tag{46b}$$

in equation (45), one obtains a reduced field at the output plane

$$\psi'(r') = \pm(-h \mid U + \mu \mid \mid V \mid^{-1})^{-\frac{1}{2}} \cdot \exp\left\{-j\frac{k}{2}\tilde{r}'[W - \tilde{V}(U + \mu)^{-1}V]r'\right\}. \tag{47}$$

This field has the same general form as the input field and describes a gaussian beam with a wavefront matrix

$$\mu' = W - \tilde{V}(U + \mu)^{-1}V, \tag{48a}$$

or, in terms of the ray matrix (see Appendix A)

$$\mu' = (C + D\mu)(A + B\mu)^{-1}. \tag{48b}$$

This interesting relation† generalizes the "ABCD law" which describes the transformation of the complex wavefront in two dimensions.[21,22] In some applications, it is also of interest to know the phase shift experienced by the beam through the optical system. It is given, from equation (47), to within Π, by the simple expression

$$\Theta = kz' - \tfrac{1}{2} \text{ Phase of } (\mid U + \mu \mid \mid V \mid^{-1}), \tag{49}$$

where kz' is the geometrical optics phase shift. Equation (49) reduces to the expression given in Ref. 24 in the case of systems with rotational symmetry. One also verifies, after a few rearrangements, that the amplitude of the beam at the output plane assumes the form given in equation (34), i.e., that

$$h' \equiv -h \mid U + \mu \mid \mid V \mid^{-1} = q_1'\hat{q}_2' - q_2'\hat{q}_1' , \tag{50}$$

where q' and $\hat{q}'$ denote the output (complex) ray position vectors. Equations (48), (49) and (50) completely define the transformation of fundamental gaussian beams propagating along the axis of nonorthogonal optical systems.

These solutions are easily generalized to the case where the axis of the incident beam is a ray $\bar{\mathcal{R}}(\bar{q}, \bar{p})$, distinct from the system axis. Let $\psi(r, z)$ denote the field of an arbitrary beam and $\bar{\mathcal{R}}$ denote an arbitrary ray; one can show[24] that

$$\psi(r, z; \bar{\mathcal{R}}) \equiv \psi(r - \bar{q}, z) \exp[-jk(\tilde{r}\bar{p} - \tfrac{1}{2}\tilde{\bar{q}}\bar{p})] \tag{51}$$

is a solution of the parabolic wave equation, equation (8). An equivalent

† The transformation of the complex curvature of gaussian beams through nonorthogonal systems has been given before (in a very complicated form) by Y. Suematsu and H. Fukinuki.[30] Equation (48b) can alternatively be obtained without integration by writing down the laws of transformation of (real) astigmatic ray pencils, as suggested in Ref. 25.

result can alternatively be obtained from the integral transformation, equation (20), by introducing a change of variables. According to equation (51), a general form for the propagation of gaussian beams is obtained by introducing the expression for the field obtained before [equation (36)] into equation (51)

$$\psi(r, z; \mathcal{R}, \hat{\mathcal{R}}) = h^{-\frac{1}{2}} \exp\left\{-j\frac{k}{2}[(\tilde{r} - \tilde{\bar{q}})\mu(r - \bar{q}) + 2\tilde{r}\bar{p} - \tilde{\bar{q}}\bar{p}]\right\}\cdot \qquad (52)$$

Notice that $\bar{q}$ and $\bar{p}$ need not be real for the right side of equation (52) to satisfy the parabolic wave equation. It is merely required that they satisfy the paraxial ray equations. When $\bar{\mathcal{R}}$ assumes complex values, however, it cannot be interpreted any longer as a beam axis. Such solutions, with $\bar{\mathcal{R}}$ complex, are of interest to generate higher order modes of propagation, as shown in the next section.

Let us now show that the fundamental mode of propagation can be written in a form resembling the form obtained in the case of orthogonal systems. This can be done by introducing, at each transverse plane, a coordinate system in which μ is diagonal.

The reduced eikonal Φ can be written

$$\Phi \equiv \tfrac{1}{2}\tilde{r}\mu r \equiv \tfrac{1}{2}(\tilde{r}\mu^{r} r + j\tilde{r}\mu^{i} r), \qquad (53)$$

where μ^{r} and μ^{i} denote the real and imaginary parts of μ, respectively. Two quadratic forms such as $\tilde{r}\mu^{r}r$ and $\tilde{r}\mu^{i}r$ can be simultaneously diagonalized if a proper (generally oblique) coordinate system is introduced.[31] The explicit expression for this transformation is not necessary here, because we are interested only in the general form of the field; it is given in Appendix C. To deal with oblique coordinates, it is convenient to introduce a tensorial notation. The expression for the scalar product of two real vectors[†] q and p in oblique coordinates assumes the form

$$q \cdot p = q^{i}p_{i}\ , \qquad (54)$$

where q^{1} and q^{2} denote the (contravariant) components of q, obtained by drawing lines parallel to the axes from the tip of the vector q as shown in Fig. (3), and where p_1 , p_2 denote the (covariant) components of p, obtained by drawing lines perpendicular to the axes. For brevity the summation sign over repeated indices is omitted.

[†] The following relations are also applicable to complex vectors since such vectors can be defined as linear combinations $V_r + jV_i$ of two real vectors V_r and V_i.

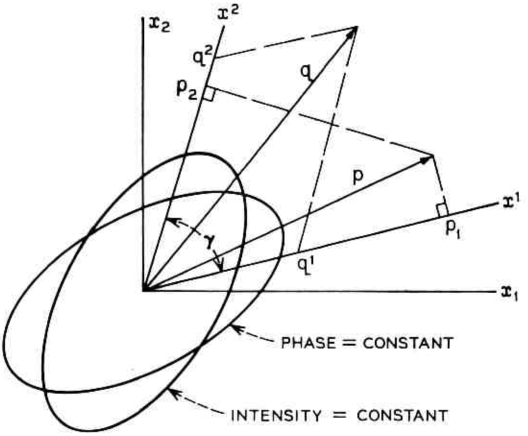


Fig. 3—This figure represents the oblique coordinate system, defined in the x_1x_2 transverse plane, which diagonalizes both the real and the imaginary parts of the wavefront (represented schematically by ellipses). The contravariant components of the position vector q, and the covariant components of the direction cosine vector p are also represented. It is assumed that the unit vectors of the coordinate system have a unit length. The index in the rectangular coordinate system is placed at a lower position only to distinguish it from the oblique coordinate system.

The reduced eikonal Φ is now written

$$\Phi \equiv \tfrac{1}{2}\mu_{ij}x^i x^j \tag{55}$$

where x^i, $i = 1, 2$ (or x^j, $j = 1, 2$) denote the contravariant components of a position vector r, and μ denotes a twice covariant tensor. With this notation, equations (31) and (32) are valid in any coordinate system. Therefore, in the coordinate system in which μ is diagonal ($\mu_{12} = 0$), we have

$$p_1 = \mu_{11}q^1, \tag{56a}$$

$$p_2 = \mu_{22}q^2, \tag{56b}$$

$$\hat{p}_1 = \mu_{11}\hat{q}^1, \tag{56c}$$

$$\hat{p}_2 = \mu_{22}\hat{q}^2. \tag{56d}$$

Let us set

$$\mu_{11} = \frac{p_1}{q^1} = \frac{\hat{p}_1}{\hat{q}^1} \equiv C_1 - 2jk^{-1}w_1^{-2}, \tag{57a}$$

$$\mu_{22} = \frac{p_2}{q^2} = \frac{\hat{p}_2}{\hat{q}^2} \equiv C_2 - 2jk^{-1}w_2^{-2}. \tag{57b}$$

In equation (57), C_i and $-2k^{-1}w_i^{-2}$ denote the real and imaginary parts of μ_{ii}, respectively. The reason for the notation $2k^{-1}w_i^{-2}$ is that w_i represent the beam radii along the coordinate axes, as shown in equation (59).

In the new coordinate system (with base vectors of unit lengths), the area of the parallelogram constructed on the vectors q and $\hat{q}$ is

$$\sin\gamma\ (q^1\hat{q}^2 - q^2\hat{q}^1) \tag{58}$$

where γ is the angle between the two coordinate axes. The field of the fundamental mode, equation (36), can consequently be rewritten

$$\begin{aligned}\psi_{00}(x^1, x^2, z; \mathcal{R}, \hat{\mathcal{R}}) = (\sin\gamma)^{-\frac{1}{2}}&(q^1\hat{q}^2 - q^2\hat{q}^1)^{-\frac{1}{2}} \\ &\cdot\exp\{-[(x^1/w_1)^2 + (x^2/w_2)^2]\} \\ &\cdot\exp\left\{-j\frac{k}{2}[C_1(x^1)^2 + C_2(x^2)^2]\right\}\cdot\end{aligned} \tag{59}$$

The first exponential term in equation (59) describes the beam intensity pattern and the second one describes the wavefront of the beam.

In the special case where the lens-like medium is orthogonal, one may choose $\mathcal{R}$ and $\hat{\mathcal{R}}$ in two mutually orthogonal planes. Assuming that these planes coincide with the x_1z and x_2z planes, respectively, we have

$$q_2 = p_2 = \hat{q}_1 = \hat{p}_1 = 0, \tag{60}$$

and equation (59) reduces to the known form (see, for example, Ref. 24)

$$\begin{aligned}\psi_{00}(x_1, x_2, z; \mathcal{R}, \hat{\mathcal{R}}) = q_1^{-\frac{1}{2}}&\exp\left[-(x_1/w_1)^2 - j\frac{k}{2}C_1x_1^2\right] \\ &\cdot\hat{q}_2^{-\frac{1}{2}}\exp\left[-(x_2/w_2)^2 - j\frac{k}{2}C_2x_2^2\right]\cdot\end{aligned} \tag{61}$$

IV. HIGHER ORDER MODES OF PROPAGATION

Two procedures are given in this section to obtain the higher order modes of propagation. One is based on the power series expansion of the field of off-set gaussian beams and the other is based on the application of differential operators on the fundamental mode. These two methods can be shown to be equivalent. They lead however to two different representations of the field, one in terms of Hermite polynomials in two complex variables and the other in terms of finite series of ordinary Hermite polynomials. Both representations are of interest.

It has been shown in the previous section that the field of a gaussian beam propagating along the axis of an optical system is fully described by two complex rays, denoted $\mathcal{R}$ and $\hat{\mathcal{R}}$, satisfying equation (33). This solution of the wave equation can be generalized to include the case where the beam axis is a ray $\bar{\mathcal{R}}$. It was pointed out also that $\bar{\mathcal{R}}$ may assume complex values provided its position and direction vectors $(\bar{q}, \bar{p})$ remain solutions of the ray equations. Let us define $\bar{\mathcal{R}}$ as a linear combination of the rays $\mathcal{R}^*$ and $\hat{\mathcal{R}}^*$, conjugate of $\mathcal{R}$ and $\hat{\mathcal{R}}$, respectively. We have

$$\bar{q} = \alpha_1 q^* + \alpha_2 \hat{q}^* \equiv Q^* \alpha, \tag{62}$$

$$\bar{p} = \alpha_1 p^* + \alpha_2 \hat{p}^* \equiv P^* \alpha, \tag{63}$$

where α_1 and α_2 are two arbitrary parameters. Introducing these expressions in equation (52) one obtains

$$\psi(r, z; \mathcal{R}, \hat{\mathcal{R}}; \alpha_1, \alpha_2) = \psi_{00}(r, z; \mathcal{R}, \hat{\mathcal{R}}) \times \exp(\tilde{\alpha} y - \tfrac{1}{2} \tilde{\alpha} \nu \alpha), \tag{64}$$

where $\psi_{00}(r, z; \mathcal{R}, \hat{\mathcal{R}})$ denotes the fundamental mode field and where we have defined

$$\alpha \equiv \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \tag{65a}$$

$$\nu \equiv \begin{bmatrix} \nu_{11} & \nu_{12} \\ \nu_{12} & \nu_{22} \end{bmatrix} \equiv -2k\tilde{Q}^* \mu^i Q^*, \tag{65b}$$

$$y \equiv \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \equiv -2k\tilde{Q}^* \mu^i r. \tag{65c}$$

Notice that ν is a symmetric matrix as a result of equation (33).

One now observes that the exponential term in equation (64) is the generating function for Hermite polynomials in two variables[32]

$$\exp(\tilde{\alpha} y - \tfrac{1}{2} \tilde{\alpha} \nu \alpha) = \sum_{m,n=0}^{\infty} \frac{\alpha_1^m}{m!} \frac{\alpha_2^n}{n!} H_{mn}(\nu^{-1} y; \nu), \tag{66}$$

where the polynomials H_{mn} have the form

$$H_{mn}(y_1, y_2; \nu) \equiv y_1^m y_2^n - \left[\frac{1}{2} \frac{m(m-1)}{1} \nu_{11} y_1^{m-2} y_2^n + \frac{mn}{1} \nu_{12} y_1^{m-1} y_2^{n-1} + \frac{1}{2} \frac{n(n-1)}{1} \nu_{22} y_1^m y_2^{n-2} \right] + \cdots \tag{67}$$

and, for $m + n \leqq 3$

$$\begin{aligned}
H_{00} &= 1, \\
H_{10} &= y_1 , \\
H_{01} &= y_2 , \\
H_{20} &= y_1^2 - \nu_{11} , \\
H_{11} &= y_1 y_2 - \nu_{12} , \\
H_{02} &= y_2^2 - \nu_{22} , \\
H_{30} &= y_1^3 - 3\nu_{11} y_1 , \\
H_{21} &= y_1^2 y_2 - 2\nu_{12} y_1 - \nu_{11} y_2 , \\
H_{12} &= y_1 y_2^2 - 2\nu_{12} y_2 - \nu_{22} y_1 , \\
H_{03} &= y_2^3 - 3\nu_{22} y_2 .
\end{aligned} \tag{68}$$

Each coefficient in the expansion of $\psi(r, z; \mathcal{R}, \hat{\mathcal{R}}; \alpha_1 , \alpha_2)$ in power series of α_1 , α_2 is necessarily a solution of the wave equation since α_1 and α_2 are arbitrary numbers. New solutions of the wave equation are therefore obtained in the form

$$\psi_{mn}(r, z; \mathcal{R}, \hat{\mathcal{R}}) = \psi_{00}(r, z; \mathcal{R}, \hat{\mathcal{R}}) H_{mn}(Q^{*^{-1}} r; \nu). \tag{69}$$

It is demonstrated in Appendix D that this set of solutions forms an orthogonal system, provided the condition $(\mathcal{R}; \hat{\mathcal{R}}^*) = 0$ is satisfied [in addition to equation (33)]. The fact that y_1 and y_2 are complex does not raise any particular difficulty in calculating $H_{mn}(y_1 , y_2 ; \nu)$ from equations (67) and (68). This prevents us, however, from identifying y_1 and y_2 with real coordinates. It is important to notice that multiplication of $\mathcal{R}$ by a factor λ (i.e., $q \rightarrow \lambda q$, $p \rightarrow \lambda p$) and $\hat{\mathcal{R}}$ by a factor $\hat{\lambda}$ ($\hat{q} \rightarrow \hat{\lambda}\hat{q}$, $\hat{p} \rightarrow \hat{\lambda}\hat{p}$) leaves essentially unchanged the field given in equation (69); it is merely multiplied by a constant. This property results from equation (65) and the general form of H_{mn} given in equation (67). Consequently $\mathcal{R}$ and $\hat{\mathcal{R}}$ need to be defined only to within constant factors.

In the special case where the optical system is orthogonal, one may choose $\mathcal{R}$ and $\hat{\mathcal{R}}$ in two mutually perpendicular meridional planes coincident with the $x_1 z$ and $x_2 z$ planes respectively ($q_2 = p_2 = 0$, $\hat{q}_1 = \hat{p}_1 = 0$). The matrix ν becomes diagonal and the Hermite polynomials in two variables reduce to a product of two Hermite polynomials in one variable. To within a constant one has, in that special case[32]

$$H_{mn}(y_1\ ,\ y_2\ ;\nu) = \nu_{11}^{m/2}\nu_{22}^{n/2}H_m(2^{-\frac{1}{2}}y_1\nu_{11}^{-\frac{1}{2}})H_n(2^{-\frac{1}{2}}y_2\nu_{22}^{-\frac{1}{2}}) \tag{70}$$

where $H_k(x)$ denotes a Hermite polynomial in one variable of order k (as defined in Ref. 33). Using equation (57), the right side of equation (70) can be written, to within a constant

$$(q_1^*/q_1)^{m/2}(\hat{q}_2^*/\hat{q}_2)^{n/2}H_m(2^{\frac{1}{2}}x_1/w_1)H_n(2^{\frac{1}{2}}x_2/w_2), \tag{71}$$

in agreement with previous results.[11]

The procedure just described for obtaining new solutions of the wave equation can be applied to an arbitrary field $\psi(r, z)$. The coefficients of the power series expansion are obtained in that case by repeated differentiation. If one calculates the coefficients for the few first orders, one finds that they assume the form

$$\psi_{mn}(r, z; \mathcal{R}, \hat{\mathcal{R}}) = \Lambda^m(\mathcal{R}^*)\,\Lambda^n(\hat{\mathcal{R}}^*)\psi(r, z), \tag{72}$$

where $\Lambda(\mathcal{R})$ and $\Lambda(\hat{\mathcal{R}})$ are differential operators defined by

$$\Lambda(\mathcal{R}) \equiv \tilde{p}r - jk^{-1}\tilde{q}\nabla, \tag{73a}$$

$$\Lambda(\hat{\mathcal{R}}) \equiv \tilde{\hat{p}}r - jk^{-1}\tilde{\hat{q}}\nabla. \tag{73b}$$

It is not difficult to show, using equation (33), that these two operators commute with one another. For generality, let us demonstrate equation (72) on the basis of the integral transformation, equation (20).†

Let $\psi(r)$ and $\psi'(r')$ denote fields at the input and output planes, respectively, of an optical system described by its reduced point characteristic $\mathcal{S}$. Let us prove that a field $\Lambda(\mathcal{R})\psi(r)$ is transformed into $\Lambda'(\mathcal{R}')\psi'(r')$ at the output plane, i.e., that

$$(\tilde{p}'r' - jk^{-1}\tilde{q}'\nabla')\left\{\iint_{-\infty}^{+\infty} \psi(r)\exp(-jk\mathcal{S})\,d^2r\right\}$$
$$= \iint_{-\infty}^{+\infty} [(\tilde{p}r - jk^{-1}\tilde{q}\nabla)\psi(r)]\exp(-jk\mathcal{S})\,d^2r. \tag{74}$$

Notice that the constant term $\pm\lambda^{-1}\,|\,V\,|^{\frac{1}{2}}$ in equation (20) can be dropped. The primes in equation (74) refer as before to quantities taken at the output plane.

Using equation (16), one finds that

$$\nabla'\exp(-jk\mathcal{S}) = -jk(\tilde{V}r + Wr')\exp(-jk\mathcal{S}). \tag{75}$$

† Alternatively one can show that the operator $\Lambda(\mathcal{R})$ [or $\Lambda(\hat{\mathcal{R}})$] commutes with the wave equation operator, equation (8). This result has been obtained before by Popov[26] for a special form of $\Lambda(\mathcal{R})$.

Therefore the left side of equation (74) can be written

$$\iint_{-\infty}^{+\infty} [\tilde{p}'r' - \tilde{q}'(\tilde{V}r + Wr')]\psi(r) \exp(-jk\mathcal{S})\, d^2r. \tag{76}$$

To evaluate the right side of equation (74), notice that, for any function $F(x_1, x_2)$ which tends exponentially to zero as $x_1, x_2 \to \pm \infty$, one has

$$\iint_{-\infty}^{+\infty} \nabla F(x_1, x_2)\, dx_1\, dx_2 = 0. \tag{77}$$

Therefore, setting

$$F(x_1, x_2) \equiv \psi(r) \times \exp(-jk\mathcal{S}) \tag{78}$$

in equation (77), one obtains

$$\iint_{-\infty}^{+\infty} \nabla\psi(r) \times \exp(-jk\mathcal{S})\, d^2r$$

$$= -\iint_{-\infty}^{+\infty} (-jk\nabla\mathcal{S})\psi(r) \times \exp(-jk\mathcal{S})\, d^2r. \tag{79}$$

Using again equation (16) to evaluate $\nabla\mathcal{S}$, the right side in equation (74) becomes

$$\iint_{-\infty}^{+\infty} [\tilde{p}r + \tilde{q}(Ur + Vr')]\psi(r) \exp(-jk\mathcal{S})\, d^2r. \tag{80}$$

The identity of the two terms in brackets in equations (76) and (80) results from the ray equations, equation (18).

The property established for $\Lambda(\mathcal{R})$ clearly holds true also for the operator $\Lambda^m(\mathcal{R})$ corresponding to m applications of $\Lambda(\mathcal{R})$, and for the operator $\Lambda^n(\hat{\mathcal{R}})$ associated with another ray $\hat{\mathcal{R}}$.

When applied to a gaussian beam (defined by $\mathcal{R}$, $\hat{\mathcal{R}}$), the operators $\Lambda(\mathcal{R})$ and $\Lambda(\hat{\mathcal{R}})$ give a result identically equal to zero. Higher order modes are obtained, however, if one considers the operators associated with the *conjugate* rays $\mathcal{R}^*$, $\hat{\mathcal{R}}^*$. One therefore calculates

$$\psi_{mn}(r, z; \mathcal{R}, \hat{\mathcal{R}}) = \Lambda^m(\mathcal{R}^*)\Lambda^n(\hat{\mathcal{R}}^*)\psi_{00}(r, z; \mathcal{R}, \hat{\mathcal{R}}). \tag{81}$$

To give a convenient form to the right side of equation (81), let us write down explicitely in tensorial notation (see Section III) the operator $\Lambda(\mathcal{R}^*)$, defined by equation (73a)

$$\Lambda(\mathcal{R}^*) = p^*_i x^i - jk^{-1}q^{i*}\nabla_i$$

$$\equiv \left(p^*_1 x^1 - jk^{-1}q^{1*}\frac{\partial}{\partial x^1}\right) + \left(p^*_2 x^2 - jk^{-1}q^{2*}\frac{\partial}{\partial x^2}\right). \tag{82}$$

Using equations (82) and (59) and the relation[33]

$$\frac{d}{dx} H_k(x) = 2xH_k(x) - H_{k+1}(x), \tag{83}$$

where $H_k(x)$ denotes a Hermite polynomial of order k, one finds that

$$\begin{aligned}\Lambda(\mathfrak{R}^*)\{H_m(2^{\frac{1}{2}}x^1/w_1)H_n(2^{\frac{1}{2}}x^2/w_2)\psi_{00}(x^1, x^2, z; \mathfrak{R}, \hat{\mathfrak{R}})\} \\ = \psi_{00}(x^1, x^2, z; \mathfrak{R}, \hat{\mathfrak{R}})[q^{1*}/w_1H_{m+1}(2^{\frac{1}{2}}x^1/w_1)H_n(2^{\frac{1}{2}}x^2/w_2) \\ + q^{2*}/w_2H_m(2^{\frac{1}{2}}x^1/w_1)H_{n+1}(2^{\frac{1}{2}}x^2/w_2)].\end{aligned} \tag{84}$$

A similar relation holds for $\Lambda(\hat{\mathfrak{R}}^*)$. These two relations show, by recurrence, that the field of the mode m, n can be written

$$\begin{aligned}\psi_{mn}(x^1, x^2, z; \mathfrak{R}, \hat{\mathfrak{R}}) &= \Lambda^m(\mathfrak{R}^*)\Lambda^n(\hat{\mathfrak{R}}^*)\psi_{00}(x^1, x^2, z; \mathfrak{R}, \hat{\mathfrak{R}}) \\ &= \psi_{00}(x^1, x^2, z; \mathfrak{R}, \hat{\mathfrak{R}}) \times [q^{1*}/w_1H(2^{\frac{1}{2}}x^1/w_1) \\ &\quad + q^{2*}/w_2H(2^{\frac{1}{2}}x^2/w_2)]^m \times [\hat{q}^{1*}/w_1H(2^{\frac{1}{2}}x^1/w_1) \\ &\quad + \hat{q}^{2*}/w_2H(2^{\frac{1}{2}}x^2/w_2)]^n,\end{aligned} \tag{85}$$

where the convention is made that, *after* multiplication of the two binomials, $H^k(x)$ actually represents a Hermite polynomial of order k: $H_k(x)$. This form of the field shows that the higher order modes of propagation can be obtained by multiplying the fundamental solution by a finite series of Hermite polynomials in one real variable.† Since q^2/q^1 and $\hat{q}^2/\hat{q}^1$ are generally complex, the wavefronts are different for each mode. It is shown in the next section that q^2/q^1 and $\hat{q}^2/\hat{q}^1$ happen to be real, however, at the end mirrors of linear resonators. From this observation, it results that the wavefronts of all of the resonating modes generally coincide with the end mirror surfaces.

Another special case of interest is the case where q^2/q^1 and $\hat{q}^2/\hat{q}^1$ are both equal to j. This happens in the case of systems with rotational symmetry, such as the "cavities with image rotation" which are investigated in Section VI.

V. NONORTHOGONAL RESONATORS

We are concerned in this section with the resonant fields and the resonant frequencies of nonorthogonal resonators. Ring-type resonators

† In the case where $m = 0$ (or $n = 0$) it is not difficult to show that the two expressions given for the mode mn in equations (69) and (85), respectively, coincide. This result can be obtained by writing equation (69) in the coordinate system in which μ (but not necessarily ν) is diagonal and using an expansion formula [equation (22), p. 371 of Ref. 32] for H_{mn} and a condensation formula [equation (31), p. 345 of Ref. 32] for the right side of equation (85). In the general case a direct comparison of the two expressions appears to be difficult.

being conceptually simpler than linear resonators, their properties are considered first. A ring-type resonator is essentially a section of waveguide closed on itself. An optical beam is a mode of the resonator, if, after a round trip, its field reproduces itself exactly.

The general form of the solutions obtained in the previous sections (Sections III and IV) is preserved as the beams propagate through an optical system. In general, however, the field distribution at the output plane of a section of waveguide does not coincide with the field distribution at the input plane [see, for instance, the transformation law, equation (48b) for the fundamental mode]. By a proper choice of the mode parameters it is possible, however, to achieve coincidence between the fields at the two planes (except, perhaps, for a constant phase factor). In that case, the beam is said to be *matched* to the section of waveguide considered. Clearly, such a beam would also be matched to a sequence of identical sections, forming a periodic waveguide. For the fundamental mode, the matching condition can be obtained by specifying that $\mu' = \mu$ in equation (48b) and solving for μ. However it is more convenient to look first for rays which reproduce themselves after a round trip in the system (except for a constant factor) and calculate the wavefront matrix μ associated with these rays. Such rays are called eigenrays; they are always complex in the case of stable resonators.

To obtain the eigenrays, let us replace q' and p' by λq and λp, respectively, in equation (18). One obtains the relations

$$-p = (U + \lambda V)q, \tag{86a}$$

$$p = (\lambda^{-1}\tilde{V} + W)q, \tag{86b}$$

and, by addition and subtraction

$$0 = (U + W + \lambda V + \lambda^{-1}\tilde{V})q, \tag{87a}$$

$$p = \tfrac{1}{2}(W - U + \lambda^{-1}\tilde{V} - \lambda V)q. \tag{87b}$$

Equation (87a) actually represents a system of two homogeneous linear equations which admit a solution only if

$$| U + W + \lambda V + \lambda^{-1}\tilde{V} | = 0, \tag{88}$$

where the bars denote a determinant. Equation (88) can be rewritten as a second-degree equation in $(\lambda + \lambda^{-1})$ as shown previously[27] for a special case. One obtains

$$| V | (\lambda + \lambda^{-1})^2 + [V_{11}K_{22} + K_{11}V_{22} - K_{12}(V_{12} + V_{21})](\lambda + \lambda^{-1}) \\ + | K | - (V_{12} - V_{21})^2 = 0, \tag{89}$$

where we have defined

$$K \equiv U + W. \tag{90}$$

The resonator is stable when the solutions of equation (89) for

$$(\lambda + \lambda^{-1})/2 \equiv \cos \theta \tag{91}$$

are real and are in the range -1 to $+1$; this is assumed henceforth. In that case, two real characteristic angles, denoted θ and $\hat{\theta}$, are obtained, the two other characteristic angles being clearly $-\theta$ and $-\hat{\theta}$.

If one introduces one of the four eigenvalues $\lambda = \exp(j\theta)$, $\hat{\lambda} = \exp(j\hat{\theta})$, $\lambda^* = \exp(-j\theta)$ or $\hat{\lambda}^* = \exp(-j\hat{\theta})$ in equations (87a) and (87b), one obtains (to within arbitrary constants) the components of the four eigenrays denoted respectively $\mathcal{R}$, $\hat{\mathcal{R}}$, $\mathcal{R}^*$ and $\hat{\mathcal{R}}^*$. Let us show that the product of $\mathcal{R}$ and $\hat{\mathcal{R}}$ [defined in equation (12)] is equal to zero.

Since $(\mathcal{R}; \hat{\mathcal{R}})$ is invariant, one may choose a reference plane along the path where the matrix V is symmetric. At such a plane, equations (87a) and (87b) assume the form

$$0 = [U + W + (\lambda + \lambda^{-1})V]q, \tag{92a}$$

$$p = \tfrac{1}{2}[W - U + (\lambda^{-1} - \lambda)V]q. \tag{92b}$$

Since both $U + W$ and V are symmetric, one has[31]

$$\tilde{q}V\hat{q} = \tilde{\hat{q}}Vq = 0, \tag{93}$$

provided the absolute values of θ and $\hat{\theta}$ are distinct. Therefore

$$\begin{aligned}(\mathcal{R}; \hat{\mathcal{R}}) &\equiv \tilde{q}\hat{p} - \tilde{\hat{q}}p \\ &= \tfrac{1}{2}\tilde{q}[W - U + (\hat{\lambda}^{-1} - \hat{\lambda})V]\hat{q} - \tfrac{1}{2}\tilde{\hat{q}}[W - U + (\lambda^{-1} - \lambda)V]q \\ &= \tfrac{1}{2}(\hat{\lambda}^{-1} - \hat{\lambda})\tilde{q}V\hat{q} - \tfrac{1}{2}(\lambda^{-1} - \lambda)\tilde{\hat{q}}Vq = 0. \end{aligned} \tag{94a}$$

One also has, replacing $\hat{\lambda}$ by $\hat{\lambda}^{-1}$ and/or λ by λ^{-1}

$$(\mathcal{R}^*; \hat{\mathcal{R}}^*) = 0; \qquad (\mathcal{R}; \hat{\mathcal{R}}^*) = 0; \qquad (\mathcal{R}^*; \hat{\mathcal{R}}) = 0. \tag{94b}$$

Therefore, according to the results of Section III, each pair of eigenrays in equation (94) defines a gaussian beam. The choice between the four pairs of eigenrays can be made by giving either a positive or a negative sign to θ and $\hat{\theta}$. It is made in such a way that the imaginary part of μ is a negative definite form. This ensures that the power carried by the beam in finite. After traversing a period of the optical system, the position q and the direction p of $\mathcal{R}$ become $q \exp(j\theta)$ and $p \exp(j\theta)$ respectively. Similarly, $\hat{q}$ and $\hat{p}$ become $\hat{q} \exp(j\hat{\theta})$ and $\hat{p} \exp(j\hat{\theta})$. Equations (35a through c) and (34) show that μ assumes its original value after a period (round trip) in the optical system; h, however, is multiplied by $\exp [j(\theta + \hat{\theta})]$. The field of the fundamental gaussian beam defined by $\mathcal{R}$ and $\hat{\mathcal{R}}$ consequently reproduces itself after a period except for an additional phase shift equal to $kL - (\theta + \hat{\theta})/2$, where L denotes the period (round trip) path length.

To clarify the above discussion let us observe that the modal matrix

$$\begin{bmatrix} jkQ^* & Q \\ jkP^* & P \end{bmatrix}, \tag{94c}$$

where Q and P were defined before in equations (32b and c), *is itself a ray matrix*, i.e., satisfies equation (112). As shown in Appendix D, the imaginary part of μ can be written $-(k^{-1}/2)(Q\tilde{Q}^*)^{-1}$; this is clearly a negative definite form, as required. It can also be shown, using equations (111), (16) and (21), that the mode generating function $\psi(r; \alpha)$ given in equation (64) is precisely the output field created by a point source located at the input plane of a (lossy) optical system whose ray matrix is the modal matrix, equation (94c).

Considering now the form, equation (85), obtained for the higher order modes, it appears that the operators Λ^m and Λ^n are responsible for an increase of the phase of ψ_{mn} equal to $-m\theta - n\hat{\theta}$. Therefore, the general expression for the resonant frequencies is

$$k_{\ell mn}L = (m + \tfrac{1}{2})\theta + (n + \tfrac{1}{2})\hat{\theta} + 2\ell\pi, \tag{95}$$

where ℓ is an integer defining the number of wavelengths along the system axis. This result was obtained by Popov[9,26] for the special case of linear nonorthogonal resonators incorporating homogeneous internal media. It is shown here to be applicable to the general case.

Let us now investigate the case of linear cavities (cavities with folded optical axis). It is convenient to replace the two curved end mirrors of such resonators by plane mirrors and thin lenses, and take the reference plane at one of the end mirrors. In a round trip along the folded optical axis two optical systems are encountered which are mirror images of one another. It is shown in Appendix B that the point characteristic matrix assumes in that case the simple form

$$[\mathcal{S}] = \begin{bmatrix} U & V \\ V & U \end{bmatrix}, \tag{96}$$

where both U and V are real and symmetric. The characteristic equations (92a) and (92b) become simply

$$(U + \cos\theta V)q = 0, \tag{97}$$

$$p = -j \sin \theta V q, \tag{98}$$

where θ is the characteristic angle.

Let q, p and $\hat{q}$, $\hat{p}$ denote two solutions of equations (97) and (98) (eigenrays). Equation (97) shows that the ratio of the two components of q and $\hat{q}$, q_2/q_1 and $\hat{q}_2/\hat{q}_1$, respectively, are real (in any coordinate system) since, for a stable resonator, the solutions θ and $\hat{\theta}$ of the characteristic equation

$$| U + \cos \theta V | = 0 \tag{99}$$

are real. One also observes that the wavefront matrix μ is imaginary. This result shows that, at the end mirrors of linear resonators, the wavefront of all of the modes coincide with the mirror surfaces, except perhaps in some cases of degeneracy. Since U and V are symmetrical, one further notices that[31]

$$\tilde{q} V \hat{q} = 0, \tag{100}$$

provided the absolute values of θ and $\hat{\theta}$ are distinct. Therefore, from equation (98),

$$\tilde{q}\hat{p} = \tilde{\hat{p}} q^* = 0. \tag{101}$$

This relation is useful in checking numerical calculations.

VI. CAVITIES WITH IMAGE ROTATION

As an example of application of the general theory discussed in the previous sections, let us calculate the resonant frequencies and the resonant field of a new type of optical cavity that one may call "cavity with image rotation."

Consider a nonplanar closed path (see Fig. 4) and let Ω be the rotation experienced after a round trip by rays parallel to the optical axis. (The value of Ω for a given orientation of the mirrors can be found in Ref. 4.) The case where the optical system has a rotational symmetry is of particular interest. Let $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ be the 2 × 2 ray matrix of the optical system with rotational symmetry introduced along the path. The round trip point characteristic matrix of the resonator is, in rectangular coordinates

$$[S] = b^{-1} \left[\begin{array}{cc|cc} a & 0 & -\cos \Omega & -\sin \Omega \\ 0 & a & \sin \Omega & -\cos \Omega \\ \hline -\cos \Omega & \sin \Omega & d & 0 \\ -\sin \Omega & -\cos \Omega & 0 & d \end{array}\right]. \tag{102}$$

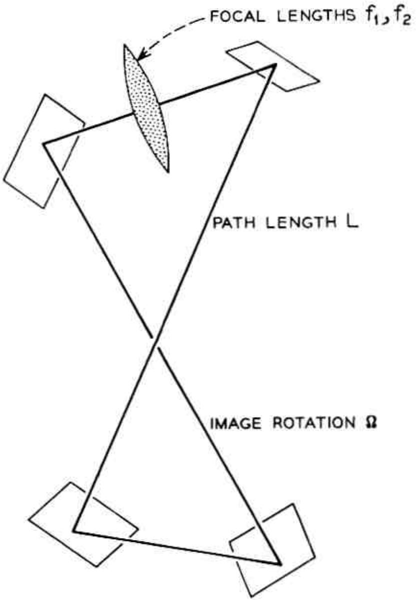


Fig. 4—A cavity with image rotation is represented. It incorporates a lens and four plane mirrors which define a nonplanar path. As a result of the twist of the path, this resonator is nonorthogonal. When the lens is astigmatic, the resonating modes do not exhibit the same patterns as in the case of more conventional cavities.

Equations (89) and (102) show that the characteristic angles are simply

$$\theta = \theta_0 + \Omega, \tag{103a}$$

$$\hat{\theta} = \theta_0 - \Omega, \tag{103b}$$

where we have defined

$$\cos \theta_0 \equiv (a + d)/2. \tag{104}$$

The resonant frequencies are therefore given, from the general relation equation (95), by

$$k_{\ell mn}L = (m + n + 1)\theta_0 + (m - n \pm 1)\Omega + 2\ell\pi. \tag{105}$$

The additional term ±Ω in equation (105) is to be introduced when polarization effects are taken into account. It has been assumed that the mirrors are perfect conductors, even in number, and that the medium is isotropic. In that case the polarization vector experiences the same transformation as an image,[4] i.e. a rotation Ω. The + and − signs in equation (105) refer to the clockwise and counterclockwise polarization states, whose degeneracy is therefore lifted.

The eigenvectors q, p and $\hat{q}$, $\hat{p}$ have respectively clockwise and counterclockwise circular polarizations too, as one expects from the rotational symmetry of the system; they are independent of the image rotation Ω. The components of $\mathcal{R}(q,\ p)$ and $\hat{\mathcal{R}}(\hat{q},\ \hat{p})$ are respectively, to within arbitrary constants

$$\mathcal{R}\begin{cases} q(jb,\ b) \\ p(-\sin\ \theta_0\ ,\ j\sin\ \theta_0) \end{cases} \qquad \hat{\mathcal{R}}\begin{cases} \hat{q}(jb,\ -b) \\ \hat{p}(-\sin\ \theta_0\ ,\ -j\sin\ \theta_0). \end{cases} \tag{106}$$

Setting for brevity, $2^{\frac{1}{2}}x^1/w = x_1$ and $2^{\frac{1}{2}}x^2/w = x_2$, where w is the beam radius, the mode ψ_{mn} assumes in rectangular coordinates, from equation (85), the form

$$\psi_{mn} = [H(x_1) + jH(x_2)]^m[H(x_1) - jH(x_2)]^n\psi_{00}\ . \tag{107a}$$

This expression, being independent of Ω, should coincide with known forms (see Ref. 15 or 11) which can be written

$$(-1)^n n!\, 2^{m+n}Z^{m-n}L_n^{m-n}(ZZ^*)\psi_{00} \qquad \text{if} \quad m \geqq n, \tag{107b}$$

$$(-1)^m m!\, 2^{m+n}Z^{*n-m}L_m^{n-m}(ZZ^*)\psi_{00} \quad \text{if} \quad m \leqq n, \tag{107c}$$

where L_p^l denotes a generalized Laguerre polynomial, and

$$Z \equiv x_1 + jx_2\ . \tag{108}$$

A relation between Hermite polynomials and generalized Laguerre polynomials was given before, in a different form, by J. R. Pierce and S. P. Morgan (private communication). The identity of the right side of equation (107a) and equations (107b) and (107c) is easily demonstrated for the special cases n (or m) $= 0$, and $m = n$, using well-known formulas,* and verified for the first values of m, n. The field consequently assumes the same form as in ordinary cavities. A rotation Ω about the z axis of the beam pattern can be expressed by a multiplication of Z by exp $(j\Omega)$ and consequently, from equation (107), by a *phase shift* $(m - n)\Omega$, in agreement with equation (105). The distinctive feature of cavities with image rotation compared with ordinary cavities, in addition to the polarization properties mentioned before, is that the intensity pattern of the resonant field has necessarily a circular symmetry.

When the optical system introduced along the nonplanar path is

* See Ref. 33. Notice that a factor 2^{2n} is missing in equation (32), p. 195, of this book.

astigmatic, one must use the general expressions given in the previous sections. This case has been studied numerically, using equation (85), for the case of a resonator incorporating a single spherical mirror of radius $R = 6m$, operating at an incidence angle of 30° and an odd number of plane mirrors. The spherical mirror is equivalent to an astigmatic lens of focal lengths $f_1 = 2.6m$ and $f_2 = 3.47m$. Assuming a round trip path length $L = 1m$ and an image rotation $\Omega = 20°$, one obtains for the point characteristic matrices, from equation (115), with $d = 1, \nu = 0$

$$U = \begin{bmatrix} 0.615 & 0 \\ 0 & 0.712 \end{bmatrix},$$

$$V = \begin{bmatrix} -0.94 & -0.34 \\ 0.34 & -0.94 \end{bmatrix},$$

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The characteristic angles are

$$\theta = -13°3,$$

$$\hat{\theta} = -54°,$$

from which the resonant frequencies can be obtained. The components of the eigenrays $\mathcal{R}$ and $\hat{\mathcal{R}}$ are respectively, in a rectangular coordinate system

$$\mathcal{R}\begin{cases} q(1, -j1.35) \\ p(0.19 - j0.66, -0.62 - j0.19), \end{cases}$$

$$\hat{\mathcal{R}}\begin{cases} \hat{q}(1, j0.91) \\ \hat{p}(0.19 - j0.57, 0.49 + j0.13). \end{cases}$$

These two eigenrays fulfill, as expected, the condition $\tilde{q}\hat{p} = \tilde{\hat{q}}p$; they define a wavefront matrix

$$\mu = \begin{bmatrix} 0.096 - j0.305 & 0.0206 \\ 0.0206 & 0.072 - j0.246 \end{bmatrix}$$

whose imaginary part is a negative definite form, as required. The intensity pattern for the mode ψ_{20} is shown in Fig. 5. It is intermediate between the circularly symmetric patterns observed when

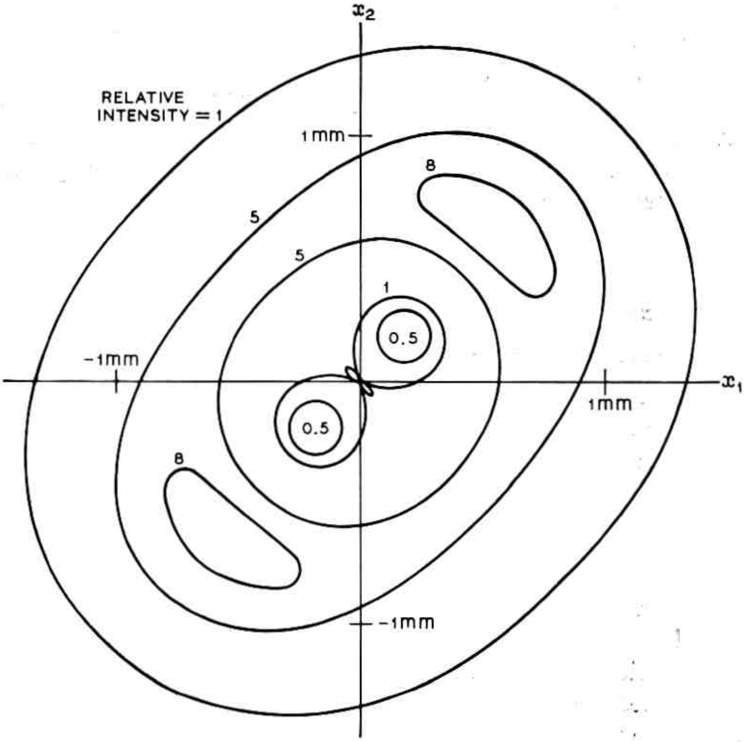


Fig. 5—This figure represents the constant intensity curves of the TEM_{20} mode in a nonorthogonal cavity incorporating a 6m radius mirror with an incidence angle of 30° and an image rotation of 20°. The optical axis path length is 1m, and the wavelength is 1μm.

$f_1 = f_2$, and the usual orthogonal patterns observed for $\Omega = 0$ (see, for instance, the TEM_{20} mode in Fig. 7 of Ref. 11).

VII. CONCLUSION

It has been shown that, within the first order approximation, the solutions of the scalar wave equation can be expressed in terms of the solutions of the (simpler) ray equations. The fundamental mode of propagation in nonorthogonal media was obtained by generalizing the expression for the field of astigmatic ray-pencils. An oblique coordinate system has been introduced which reduces this solution to the form assumed by ordinary gaussian beams. The higher order modes of propagation were also obtained; they can be expressed as the product of the fundamental solution and Hermite polynomials in one real variable.

The results of Popov[9,26] for the resonant frequencies of nonorthogonal resonators were extended to resonators incorporating arbitrary lens-like media and were applied to a new type of cavity which exhibits interesting resonance and polarization properties. This theory may also be useful for special optical waveguides such as the helical gas lenses, and for analysis of optical systems which are nominally orthogonal, but which suffer from small distortions in three dimensions.

APPENDIX A

Relations Between the Point Characteristic Matrix and the Ray Matrix

It has been shown in the main text [equation (18)] that the direction vectors p, p' of a ray at the input and output plane of an optical system are related to the position vectors q, q' by the following matricial relation

$$\begin{bmatrix} -p \\ p' \end{bmatrix} = \begin{bmatrix} U & V \\ \tilde{V} & W \end{bmatrix} \begin{bmatrix} q \\ q' \end{bmatrix} \equiv [\mathcal{S}] \begin{bmatrix} q \\ q' \end{bmatrix}, \tag{109}$$

where U and W are 2×2 real symmetric matrices, V is a 2×2 real matrix. $[\mathcal{S}]$ is a 4×4 symmetric matrix which has been called the point characteristic matrix. One also sometimes defines a *ray matrix*, $[\mathcal{M}]$ which relates the position and direction vectors of a ray at the output plane to the values assumed at the input plane

$$\begin{bmatrix} q' \\ p' \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} q \\ p \end{bmatrix} \equiv [\mathcal{M}] \begin{bmatrix} q \\ p \end{bmatrix}, \tag{110}$$

where A, B, C, D are 2×2 real matrices. Since, from equation (109), only 10 numbers suffice to define the optical system, the elements of the 4×4 ray matrix $[\mathcal{M}]$ must be related by $16 - 10 = 6$ relations. To obtain these relations, let us compare equations (109) and (110). One obtains readily

$$U = B^{-1}A, \tag{111a}$$

$$V = -B^{-1}, \tag{111b}$$

$$\tilde{V} = C - DB^{-1}A, \tag{111c}$$

$$W = DB^{-1}. \tag{111d}$$

Since U and W are symmetrical one has

$$A\tilde{B} - B\tilde{A} = 0, \tag{112a}$$

$$\tilde{B}D - \tilde{D}B = 0, \tag{112b}$$

and, by comparing the expressions obtained for V and $\tilde{V}$, equations (111b) and (111c), and using equation (112b), one finds that

$$\tilde{D}A - \tilde{B}C = 1. \tag{112c}$$

Equations (112a, b and c) are equivalent to those given by Luneburg.[1] They effectively correspond to six independent relations. The relations inverse of equations (111a through d) are

$$A = -V^{-1}U, \tag{113a}$$

$$B = -V^{-1}, \tag{113b}$$

$$C = \tilde{V} - WV^{-1}U, \tag{113c}$$

$$D = -WV^{-1}. \tag{113d}$$

APPENDIX B

Point Characteristic Matrix of a Sequence of Thin Lenses and Mirrors-Symmetrical Systems

The point characteristic matrix $[\mathbb{S}]$ of a sequence of thin astigmatic lenses and plane mirrors, arbitrarily oriented in space, can be obtained in closed form.

Let us first consider a thin astigmatic lens oriented at an angle ν with respect to the x_1 axis of a x_1x_2z rectangular coordinate system, with focal lengths f_1, f_2. This lens is followed by a section of free space of length d. For generality, one further assumes that the output coordinate system is rotated by an angle Ω about the z axis. This rotation has to be introduced in the case of non planar paths.[4,34] Using the expression for the optical thickness of a lens, and the paraxial approximation of the length of tilted rays in free space, one obtains

$$[\mathbb{S}] \equiv \begin{bmatrix} U & V \\ \tilde{V} & W \end{bmatrix}, \tag{114}$$

with

$$U = \begin{bmatrix} \dfrac{1}{d} - \left(\dfrac{\cos^2\nu}{f_1} + \dfrac{\sin^2\nu}{f_2}\right) & \cos\nu\sin\nu\left(\dfrac{1}{f_1} - \dfrac{1}{f_2}\right) \\ \cos\nu\sin\nu\left(\dfrac{1}{f_1} - \dfrac{1}{f_2}\right) & \dfrac{1}{d} - \left(\dfrac{\cos^2\nu}{f_2} + \dfrac{\sin^2\nu}{f_1}\right) \end{bmatrix},$$

$$V = -d^{-1}\begin{bmatrix} \cos\Omega & \sin\Omega \\ -\sin\Omega & \cos\Omega \end{bmatrix},$$

$$W = d^{-1}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{115}$$

These expressions are applicable to curved mirrors under oblique incidence with little modification since a curved mirror is equivalent to a plane mirror and a lens, in the most general case.[34] It remains to calculate the point characteristic matrix of a sequence of optical systems such as the one described by equations (114) and (115).

The point characteristic matrix $[\mathcal{S}_t]$ of a sequence of two optical systems whose point characteristic matrices are respectively $[\mathcal{S}_1]$ and $[\mathcal{S}_2]$ is obtained by using equation (18) of the main text, and specifying that the rays are continuous at the junction between the two systems. One obtains

$$[\mathcal{S}_t] = \begin{bmatrix} U_1 - V_1(W_1 + U_2)^{-1}\tilde{V}_1 & -V_1(W_1 + U_2)^{-1}V_2 \\ -\tilde{V}_2(W_1 + U_2)^{-1}\tilde{V}_1 & W_2 - \tilde{V}_2(W_1 + U_2)^{-1}V_2 \end{bmatrix}. \tag{116}$$

In the special case where the second optical system is the mirror image of the first system with respect to their common plane, $[\mathcal{S}_t]$ reduces to

$$[\mathcal{S}_t] \equiv \begin{bmatrix} U_t & V_t \\ \tilde{V}_t & W_t \end{bmatrix} = \begin{bmatrix} U - \frac{1}{2}VW^{-1}\tilde{V} & -\frac{1}{2}VW^{-1}\tilde{V} \\ -\frac{1}{2}VW^{-1}\tilde{V} & U - \frac{1}{2}VW^{-1}\tilde{V} \end{bmatrix} \tag{117}$$

where the index 1 has been omitted. Equation (117) shows that, in a symmetric system, U_t is equal to W_t and V_t is a symmetric matrix.

Repeated applications of equation (116) and equations (114) and (115) give the point characteristic matrix of an arbitrary sequence of lenses or mirrors.

APPENDIX C

Diagonalization of a Complex Wavefront

The need for introducing an oblique coordinate system at each transverse plane has been outlined in the main text. Detailed transformation formulas are given in this appendix.

Let e_1, e_2 be the base vectors, of unit length, of the original rectangular coordinate system, and $\mathbf{e}_1$, $\mathbf{e}_2$ the base vectors, also of unit length, of a new coordinate system.[†] The $\mathbf{e}_i$, $i = 1, 2$ are linearly related to the e_j , $j = 1, 2$ by

$$\mathbf{e}_i = \delta_i^j e_j \, , \tag{118}$$

where δ_i^j is the mixed tensor which expresses the coordinate transformation. The reduced eikonal Φ is a complex quadratic form which was written in the original coordinate system [equation (53)]

$$\Phi \equiv \tfrac{1}{2}\tilde{r}\mu r \equiv \tfrac{1}{2}\tilde{r}(\mu^r + j\mu^i)r, \tag{119}$$

where μ^r and μ^i are real symmetric matrices.

By stipulating that, in the new coordinate system, the off-diagonal terms of μ^r and μ^i are both equal to zero, one obtains the transformation $[\delta]$ which diagonalizes Φ

$$[\delta] \equiv \begin{bmatrix} \delta_1^1 & \delta_1^2 \\ \delta_2^1 & \delta_2^2 \end{bmatrix} = \begin{bmatrix} (1 + v^2)^{-\frac{1}{2}} & v(1 + v^2)^{-\frac{1}{2}} \\ u(1 + u^2)^{-\frac{1}{2}} & (1 + u^2)^{-\frac{1}{2}} \end{bmatrix}, \tag{120}$$

where

$$u = (c/a)v = [-b + (b^2 - 4ac)^{\frac{1}{2}}]/2a, \tag{121}$$

$$a \equiv \mu_{11}^r \mu_{12}^i - \mu_{11}^i \mu_{12}^r \, , \tag{122a}$$

$$b \equiv \mu_{11}^r \mu_{22}^i - \mu_{11}^i \mu_{22}^r \, , \tag{122b}$$

$$c \equiv \mu_{22}^i \mu_{12}^r - \mu_{22}^r \mu_{12}^i \, . \tag{122c}$$

The law of transformation of the contravariant components of a vector q, denoted respectively q^i in the old system and $\mathbf{q}^i$ in the new system is (omitting the summation sign)

$$q^i = \delta_j^i \mathbf{q}^j. \tag{123a}$$

This relation is also applicable to the coordinate x^i

$$x^i = \delta_j^i \mathbf{x}^j. \tag{123b}$$

The covariant components of a vector p, denoted p_i in the old system and $\mathbf{p}_i$ in the new system, transform according to the inverse relation

$$\mathbf{p}_i = \delta_i^j p_j \, . \tag{124}$$

Expressions for the new components of μ are derived in the main text.

[†] Quantities relative to the new system are denoted by bold face letters in this Appendix. Ordinary letters are used in the main text, where there is no risk of confusion.

APPENDIX D

Orthogonality of the Modes

Let q, p and $\hat{q}$, $\hat{p}$ be any two solutions of the paraxial ray equations, equations (11a and b), and assume that the matrix PQ^{-1}, where

$$Q \equiv [q \; \hat{q}], \tag{125a}$$

$$P \equiv [p \; \hat{p}], \tag{125b}$$

is symmetric.

An infinite set of solutions of the parabolic wave equation has been obtained in the main text in the form [equation (69)]

$$\psi_{mn}(r, z; Q, P) = | Q |^{-\frac{1}{2}} \exp \left(-j \frac{k}{2} \tilde{r} P Q^{-1} r \right) H_{mn}(\chi ; \nu), \tag{126}$$

where H_{mn} denote the Hermite polynomial in two variables χ_1 , χ_2

$$\chi \equiv Q^{*^{-1}} r, \tag{127}$$

associated with the quadratic form $\tilde{\chi} \nu \chi$, where

$$\begin{aligned} \nu &\equiv jk\tilde{Q}^*(PQ^{-1} - P^*Q^{*^{-1}})Q^* \\ &= JQ^{-1}Q^*, \end{aligned} \tag{128}$$

$$J \equiv -jk \begin{bmatrix} \tilde{q}p^* - \tilde{p}q^* & \tilde{\hat{q}}p^* - \tilde{\hat{p}}q^* \\ \tilde{q}\hat{p}^* - \tilde{p}\hat{q}^* & \tilde{\hat{q}}\hat{p}^* - \tilde{\hat{p}}\hat{q}^* \end{bmatrix}. \tag{129}$$

Let us now impose on the rays the additional condition

$$(\mathfrak{R}; \hat{\mathfrak{R}}^*) \equiv \tilde{q}\hat{p}^* - \tilde{p}\hat{q}^* = 0 \tag{130}$$

and assume that the diagonal terms of J are positive. Since, as pointed out in the main text, the two rays need be defined only to within constants, they can be normalized in such a way that J is the unit matrix. In that case we have, from equations (127) and (128)

$$\nu^{-1} = \nu^*, \tag{131}$$

$$\nu\chi = \chi^*. \tag{132}$$

Consequently

$$H^*_{m'n'}(\chi; \nu) = H_{m'n'}(\chi^*; \nu^*) = G_{m'n'}(\chi; \nu) \tag{133}$$

where we have introduced the adjoint polynomials G_{mn} defined by

$$G_{mn}(\chi; \nu) \equiv H_{mn}(\nu\chi; \nu^{-1}). \tag{134}$$

The orthogonality condition for two solutions ψ_{mn} and $\psi_{m'n'}$ [equation (126)] can now be written in the form

$$\begin{aligned}
&\iint_{-\infty}^{+\infty} \psi_{mn}(r)\psi^*_{m'n'}(r)\, d^2r \\
&\quad = |\, Q^{-1}Q^* \,|^{\frac{1}{2}} \iint_{-\infty}^{+\infty} \exp\left(-\tfrac{1}{2}\tilde{\chi}\nu\chi\right) H_{mn}(\chi;\nu) G_{m'n'}(\chi;\nu)\, d^2\chi, \\
&\quad = 2\pi m!\, n! \quad \text{if} \quad m' = m \quad \text{and} \quad n' = n, \\
&\quad = 0 \qquad\qquad \text{if} \quad m' \neq m \quad \text{or} \quad n' \neq n.
\end{aligned} \tag{135}$$

The biorthogonality property[32] of the polynomials H_{mn} and G_{mn} has been used in equation (135).

REFERENCES

1. Luneburg, R. K., *Mathematical Theory of Optics,* Los Angeles: University of California Press, 1964.
2. Tien, P. K., Gordon, J. P., and Whinnery, J. R., "Focusing of a Light Beam of Gaussian Field Distribution in Continuous and Periodic Lens-Like Media," Proc. IEEE, *53* (February 1965), pp. 129–136.
3. Marié, P., "Guidage de la lumière cohérente par un guide hélicoïdal. Extension à la focalisation continue des particules de haute énergie," Annales de Télécommunication, *24* (May/June 1969), pp. 177–189.
4. Arnaud, J. A., "Degenerate Optical Cavities," Applied Optics, *8,* No. 1 (January 1969), pp. 189–195.
5. Lord Rayleigh, *Scientific Papers,* Vol. V, p. 617, and Vol. VI, p. 212, New York: Dover Public. Inc., 1964.
6. Keller, J. R., and Rubinow, S. I., "Asymptotic Solution of Eigenvalue Problems," Annals of Physics, *9* (January 1960), pp. 24–75.
7. Bykov, V. P., and Vainshtein, L. A., "Geometric Optics of Open Resonators," Soviet Physic: J.E.T.P., *20,* No. 2 (February 1965), pp. 338–344.
8. Kahn, W. K., "Geometrical Optics Derivation of Formula for the Variation of the Spot-Size in a Spherical Mirror Resonator," Appl. Optics, *5* (June 1966), pp. 1023–1029.
9. Popov, M. M., "Resonators for Lasers with Unfolded Directions of Principal Curvatures," Opt. Spectrosc., *25,* No. 3 (September 1968), pp. 213–217.
10. Gordon, J. P., "Optics of General Guiding Media," B.S.T.J., *45,* No. 2 (February 1966), pp. 321–332.
11. Kogelnik, H., and Li, T., "Laser Beams and Resonators," Applied Optics, *5,* No. 10 (October 1966), pp. 1550–1567.
12. Milder, D. M., "Ray and Wave Invariants for SOFAR Channel Propagation," J. of the Acoustical Society of Amer., *46,* No. 5, Part 2 (November 1969), pp. 1259–1263.
13. Gloge, D., and Marcuse, D., "Formal Quantum Theory of Light Rays," J. Opt. Soc. Amer., *59,* No. 12 (December 1969), pp. 1629–1631.
14. Landau, L. D., and Lifshits, E. M., *Quantum Mechanics,* Reading, Massachusetts: Pergamon Press Ltd., (1958), p. 69.
15. Goubau, G., and Schwering, F., "On the Guided Propagation of Electromagnetic Wave Beams," I.R.E. Trans. on Antenna and Propagation, *AP. 9* (May 1961), pp. 248–256.
16. Pierce, J. R., "Modes in Sequences of Lenses," Proc. Nat'l. Acad. Sci., *47* (November 1961), pp. 1808–1813.
17. Boyd, G. D., and Gordon, J. P., "Confocal Multimode Resonator for Milli-

meter Through Optical Wavelength Masers," B.S.T.J., *40,* No. 2 (March 1961), pp. 489–508.

18. Marcatili, E. A. J., "Modes in a Sequence of Thick Astigmatic Lens-Like Focusers," B.S.T.J., *43,* No. 6 (November 1964), pp. 2887–2904.
19. Marcatili, E. A. J., "Effect of Redirectors, Refocusers, and Mode Filters on Light Transmission Through Aberrated and Misaligned Lenses," B.S.T.J., *46,* No. 6 (October 1967), pp. 1733–1752.
20. Collins, S. A., "Analysis of Optical Resonators Involving Focusing Elements," Applied Optics, *3* (November 1964), pp. 1263–1275.
21. Kogelnik, H., "On the Propagation of Gaussian Beams of Light Through Lenslike Media Including those with a Loss or Gain Variation," Applied Optics, *4* (December 1965), pp. 1562–1569.
22. Kogelnik, H., "Imaging of Optical Modes-Resonators with Internal Lenses," B.S.T.J., *44,* No. 3 (March 1965), pp. 455–494.
23. Vlasov, S. N., and Talanov, V. I., "On The Relation Between the Ray and Wave Descriptions of Electromagnetic Beams in Quasi-Optical Systems," Soviet Radiophysics, *8,* No. 1 (January/February 1965), pp. 145–147.
24. Arnaud, J. A., "Degenerate Optical Cavities. II: Effect of Misalignments," Applied Optics, *8,* No. 9 (September 1969), pp. 1909–1917.
25. Arnaud, J. A., and Kogelnik, H., "Gaussian Light Beams with General Astigmatism," Applied Optics, *8,* No. 8 (August 1969), pp. 1687–1693.
26. Popov, M. M., "Resonators for Lasers with Rotated Directions of Principal Curvatures," Opt., and Spectrosc., *25,* No. 2 (August 1968), pp. 170–171.
27. Kahn, W. K., and Nemit, J., "Ray Theory of Astigmatic Resonators and Beam Waveguides," *Proceedings Symposium on Modern Optics,* J. Fox, Editor, Brooklyn, N. Y.: Polytechnic Press, 1967.
28. Kiselev, V. A., "Mode of Open Resonators with Optically Inhomogeneous Regions between Mirrors," J. Prikladnoi Spektroskopii, *4,* No. 1 (January 1966), pp. 37–45 (in Russian).
29. Bodewig, E., *Matrix Calculus,* Amsterdam: North-Holland Publishing Co., 1959, p. 41.
30. Suematsu, Y., and Fukinuki, H., "Matrix Theory of Light Beam Waveguides," Bull. Tokyo Inst. Technology, *88* (March 1968), pp. 33–47.
31. Hildebrand, F. B., *Methods of Applied Mathematics,* New York: Prentice Hall, Inc. (1965), pp. 48 and 70.
32. Appel, P., and Kampé de Fériet, J., *Fonctions Hypergéométriques et Hypersphériques/Polynomes d'Hermite,* Gauthier-Villars et C^{ie}, Paris, 1926.
33. Erdelyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G., *Higher Transcendental Functions,* New York: McGraw-Hill Inc., (1953), Vol. 2.
34. Arnaud, J. A., and Ruscio, J. T., "Focusing and Deflection of Optical Beams by Cylindrical Mirrors," Applied Optics, *9,* No. 10 (October 1970), pp. 2377–2380.

Optical Resonators With Variable Reflectivity Mirrors

By H. ZUCKER

(Manuscript received May 27, 1970)

In this paper we investigate circular optical resonators with gaussian profiles of the mirror reflectivities. Closed form solution to the integral equations for such resonators are obtained. The dominant $TEM_{0,0}$ mode characteristics of a resonator consisting of one variable reflectivity mirror (VRM) and one uniform reflectivity mirror (URM) are considered in detail for a variety of parameters. This resonator is particularly suitable for high-gain lasers. Its advantages in comparison to the conventional type are: (i) there is larger mode volume utilization, and (ii) the power transmitted at the variable reflectivity mirror can in principle be utilized as the power output. We discuss dependence of the spot sizes on laser gain and mirror-curvature tolerances and present a specific design of a Fabry-Perot resonator for fundamental mode operation and the expected performance.

I. INTRODUCTION

The dominance of the fundamental mode in optical resonators with uniform reflectivity mirrors is due to the lowest diffraction loss of this mode. The power output of this mode is commonly obtained by using a partially transparent mirror. These two features could be combined in a resonator consisting of one uniform reflectivity mirror (URM) and one variable reflectivity mirror (VRM).

Resonators with VRM were investigated previously. S. N. Vlasov and V. I. Talanov[1] considered symmetrical two-dimensional resonators with two types of variations of the mirror reflectivities including the gaussian and obtained solutions for the eigenvalues of the resonator integral equations. N. G. Vakhimov[2] investigated the natural resonant frequencies and field distributions of symmetrical resonators with gaussian VRM by using an asymptotic method of solution to the wave equation subject to impedance boundary conditions. N. Kumagai and others[3] investigated Fabry Perot resonators with VRM of finite dimensions by solving numerically the resonator integral equation for different mirror reflectivities. Y. Suematsu and others[4] studied beam waveguides with gaussian transmission filters for the improvement of the stability of beam transmission.

In this work we investigate nonsymmetrical circular resonators consisting of one URM and one VRM. The radii of curvature of the mirrors are arbitrary. The reflection coefficients of the VRM are assumed to have gaussian profiles in the radial direction. For such resonators with infinite mirrors, solutions to the resonator integral equations are obtained in terms of Laguerre functions with complex arguments. The modal fields decrease off-axis very rapidly and consequently these solutions are also applicable to resonators with finite mirrors.

Resonators of the type considered seem to be particularly suitable for high-gain lasers as for example CO_2 lasers. It is shown subsequently that for the fundamental $TE_{0,0}$ mode, the spot sizes obtainable are considerably larger than those obtained with URM resonators of the same length and the same fundamental mode threshold gain ratio. This should result in a larger mode volume utilization. Furthermore, the power loss due to the transparency of VRM could be utilized as the power output.

In the following sections the solutions for the resonator modes and eigenvalues of nonsymmetrical resonators are obtained. It is shown that the solutions for symmetrical resonators consisting of two identical VRM are readily obtainable as special cases. Mode-stability criteria are established as functions of the resonator geometries and VRM parameters. The spot sizes of the fundamental $TE_{0,0}$ mode are computed as a function of the threshold-gain ratio for a variety of parameters. A comparison is made between the obtainable spot sizes with Fabry Perot resonators with VRM and URM. We show that much larger spot sizes can be achieved with the VRM resonator. We show also that the spot-size diameters of VRM mirrors depends basically on the threshold-gain ratio and on the mirror-curvature tolerances. A specific design of a Fabry Perot resonator with a VRM is examined and the expected performance is presented.

II. NONSYMMETRICAL RESONATORS

2.1 *Solutions to the Integral Equation (IE)*

The geometry of the nonsymmetrical resonator is shown in Fig. 1. It consists of one mirror with variable reflectivity (M1) and one mirror

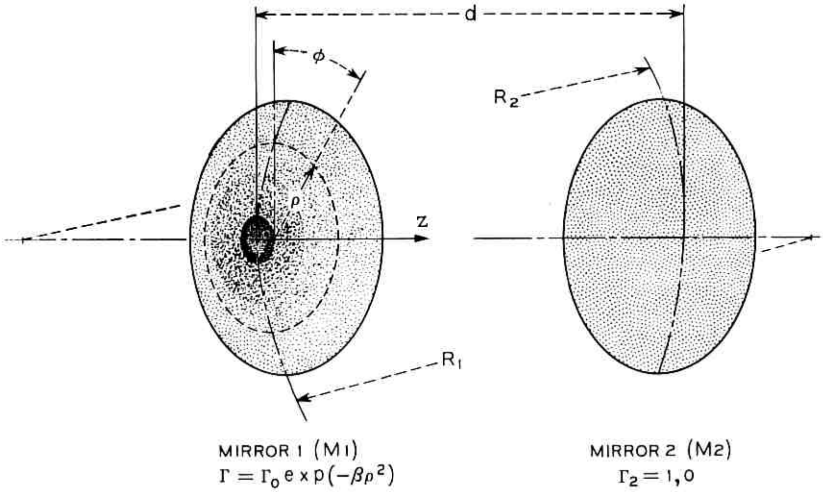


Fig. 1—Nonsymmetrical resonator.

with uniform reflectivity (M2). The separation between the mirrors is d, and the radii of curvature are designated by R_1 and R_2 respectively. The reflection coefficients of the VRM, Γ, is assumed to vary in the radial direction ρ as follows:

$$\Gamma = \Gamma_0 \exp(-\beta\rho^2) \tag{1}$$

where Γ_0 and β are constants with $| \Gamma_0 | \leqq 1$.

The reflection coefficient of the URM is assumed to be unity. (A reflection coefficient different than unity can readily be included in the solution.)

The integral equations for this resonator are obtained in a manner analogous to a URM resonator, by imposing the condition that the field should reproduce itself after a round trip. With the azimuthal dependence for the electric field $E(\rho, \phi) = \exp(-j_\ell\phi)F_\ell(\rho)$, the two simultaneous integral equations are:[5]

$$K_\ell^{(2)}F_\ell^{(2)}(\rho_2) = j^{\ell+1}\exp(-jkd)M\int_0^\infty F_\ell^{(1)}(\rho_1)\exp[-jM(g_1\rho_1^2 + g_2\rho_2^2)/2] \cdot J_\ell(M\rho_1\rho_2)\rho_1\,d\rho_1, \tag{2}$$

$$K_\ell^{(1)}F_\ell^{(1)}(\rho_1) = j^{\ell+1}\exp(-jkd)M\Gamma_0\exp(-\beta\rho_1^2)\int_0^\infty F_\ell^2(\rho_2) \cdot \exp[-jM(g_1\rho_1^2 + g_2\rho_2^2)/2]\cdot J_\ell(M\rho_1\rho_2)\rho_2\,d\rho_2, \tag{3}$$

where $F_\ell^{(1)}(\rho_1)$, $F_\ell^{(2)}(\rho_2)$ are the radial field distributions at (M1) and (M2) respectively, $K_\ell^{(1)}$ and $K_\ell^{(2)}$ are the associated eigenvalues, J_ℓ is a Bessel function of order ℓ, $M = 2\pi/\lambda d$, λ is the free-space wavelength and

$$g_1 = \left(1 - \frac{d}{R_1}\right), \tag{4}$$

$$g_2 = \left(1 - \frac{d}{R_2}\right). \tag{5}$$

The integral equations (2) and (3) are solved by using the self-reciprocal properties of the Laguerre functions on the Hankel transform.[6,7] These properties are

$$\int_0^\infty x^{\nu+1} \exp(-\beta x^2) L_n^\nu(\alpha x^2) J_\nu(xy)(xy)^{\frac{1}{2}}\, dx$$
$$= 2^{-\nu-1} y^{\nu+\frac{1}{2}} (\beta - \alpha)^n \beta^{-n-\nu-1} \exp(-y^2/4\beta) L_n^\nu\left(\frac{\alpha y^2}{4\beta(\alpha - \beta)}\right) \tag{6}$$

where L_n^ℓ is a Laguerre function of order ℓ, n.

Based on equation (6), the modal solutions to the integral equations are:

$$F_\ell^{(1)}(\rho_1) = \exp(-\gamma_1 \rho_1^2/2) L_n^\ell(\alpha_1 \rho_1^2)(\sqrt{\alpha_1}\, \rho_1)^\ell, \tag{7}$$

$$F_\ell^{(2)}(\rho_2) = \exp(-\gamma_2 \rho_2^2/2) L_n^\ell(\alpha_2 \rho_2^2)(\sqrt{\alpha_2}\, \rho_2)^\ell. \tag{8}$$

After some algebraic manipulations, the following relations are obtained for the parameters.

$$\gamma_1 = \alpha_1 + \beta, \tag{9}$$

$$\alpha_1^2 = \frac{M^2}{4g_2^2} + \left[\beta + jM\left(g_1 - \frac{1}{2g_2}\right)\right]^2. \tag{10}$$

It is convenient to express α_1 in terms of a complex trigonometric function as follows

$$\alpha_1 = \frac{M}{2g_2} \cosh \delta \tag{11}$$

with $\delta = \Lambda + j\Delta$. In equations (9) and (10), Λ and Δ are related to the resonator geometry and reflectivity parameters by

$$\sinh \Lambda \cos \Delta = \frac{2g_2\beta}{M}, \tag{12}$$

$$\cosh \Lambda \sin \Delta = (2g_1g_2 - 1). \tag{13}$$

Furthermore

$$\gamma_2 = \alpha_2, \tag{14}$$

$$\alpha_2 = Mg_2 \frac{(\cos \Delta + j \sinh \Lambda)}{\sin \Delta + \cosh \Lambda}. \tag{15}$$

The associated eigenvalues $K_\ell^{(1)}$ and $K_\ell^{(2)}$ are:

$$K_\ell^{(1)} = \Gamma_0 \exp(-jkd)_j n \left[\frac{1 + j \exp(-\delta)}{2g_2}\right]^{\ell+1} \exp(-n\delta), \tag{16}$$

$$K_\ell^{(2)} = \exp(-jkd) j^n \left[\frac{j2g_2}{\exp(\delta) + j}\right]^{\ell+1} \exp(-n\delta). \tag{17}$$

The eigenvalue K_ℓ which gives the decrease of the reflected field after a double pass is the product of the above two eigenvalues given by equations (16) and (17).

$$K_\ell = K_\ell^{(1)} K_\ell^{(2)} = (-1)^n j^{\ell+1} \exp(-j2kd)\Gamma_0 \exp[-(2n + \ell + 1)\delta]. \tag{18}$$

The eigenvalues K_ℓ are exponentially decreasing with ℓ, n. The largest eigenvalue is obtained for $n = \ell = 0$, corresponding to the fundamental $\text{TEM}_{0,0}$ mode. The next eigenvalue corresponds to the $\text{TEM}_{1,0}$ mode ($\ell = 1$, $n = 0$). Since the eigenvalues are related to the power loss, the fundamental mode selectivity will depend primarily on the eigenvalues for the $\text{TE}_{0,0}$ and $\text{TE}_{1,0}$ modes.

It is of interest to examine the special case $g_2 = 1$, which corresponds to a resonator with a perfectly reflecting planer mirror M2. This resonator is completely equivalent to a symmetrical resonator consisting of two identical VRM separated by $2d$. Both the eigenvalues and the fields are the same with the fields beyond d being equal to the reflected fields of the nonsymmetrical resonator.

The modes of the nonsymmetrical resonator are orthogonal at the uniform mirror M2, since $\gamma_2 = \alpha_2$. Howerer, at the VRM neither the incident modes nor the reflected modes are orthogonal. This is shown in Appendix A. In addition it is shown that for any particular mode, the ratio of the reflected to the incident power at the VRM is precisely equal to the absolute value square of the eigen value K_ℓ. Physically this condition corresponds to conservation of power.

2.2 *Stability Criteria*

For a resonator made to be stable, it is necessary that the exponential factors γ_1 and γ_2 in equations (7) and (8) be finite and have a positive real part. Both these factors are dependent on cos Δ and g_2. The limits of the stability regions are thus: (a) $g_2 = 0$ and (b) $\cos \Delta = 0$. The second condition can be expressed in terms of the resonator parameters by using equation (13).

$$\cos \Delta = \left[1 - \left(\frac{2g_1g_2 - 1}{\cosh \Lambda}\right)^2\right]^{\frac{1}{2}} \tag{19}$$

and the second limit of the stability region is

$$g_2 = \frac{1 \pm \cosh \Lambda}{2g_1}. \tag{20}$$

Equation (20) contains the special case of the uniform reflectivity resonator ($\cosh \Lambda = 1$). For this special case equation (20) reduces to the stability criterion derived by G. D. Boyd and H. Kogelnik.[8]

In Fig. 2 illustrative stability diagrams are shown as a function of g_1 and g_2 with exp (2Λ) as a parameter. (The choice of this parameter is discussed subsequently.)

A few special cases are considered

(*i*) $g_1 = 0$.
This resonator is stable for all values of g_2, except $g_2 = 0$.

(*ii*) $g_2 = 0$.
This resonator is unstable independent of the curvature of M1.

(*iii*) $g_1 = g_2 = 0$.
This is the very special case of the confocal resonator and is in general unstable, except for a URM resonator ($\beta = 0$).

(*iv*) $g_1 = g_2 = 1$.
This is the Fabry-Perot resonator and is stable with a variable reflectivity mirror.

2.3 *The Threshold-Gain Ratio*

To sustain oscillations in a laser resonator a necessary condition is that the active medium should have enough gain such that after a double transit the field has the same amplitude. This condition can be written in terms of the eigenvalues of the resonator modes as[9]

$$G\,|K_\ell|^2 = 1 \tag{21}$$

where G is the power gain per double transit.

In particular for the $\text{TEM}_{0,0}$ and $\text{TM}_{1,0}$ modes, equation (21) can be written using equation (18) as:

$$G_{0,0}\Gamma_0^2 \exp(-2\Lambda) = 1, \tag{22}$$

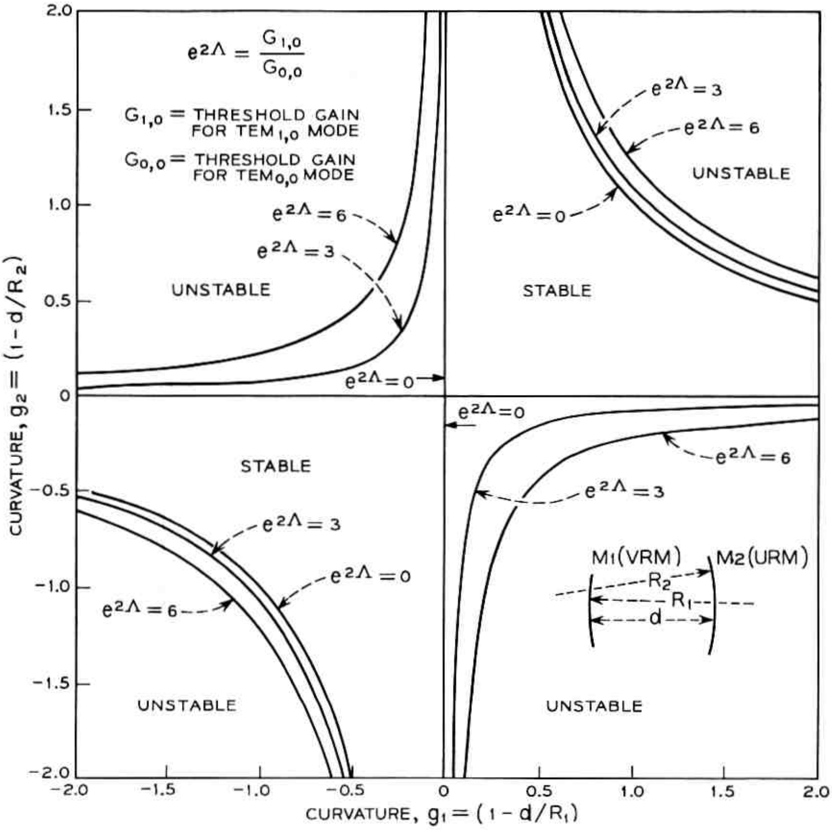


Fig. 2—Stability diagrams.

$$G_{1,0}\Gamma_0^2 \exp(-4\Lambda) = 1, \tag{23}$$

where $G_{0,0}$ and $G_{1,0}$ is the threshold-power gain required for oscillation in the respective modes. A quantity of interest is the threshold-gain ratio, t defined by:

$$t = \frac{G_{1,0}}{G_{0,0}} = \exp(2\Lambda). \tag{24}$$

This ratio is a measure of the gain tolerance required for oscillation in the dominant $TEM_{0,0}$ mode, and is independent of Γ_0.

The threshold-gain ratio may also be expressed in terms of the loss per round trip $L_{0,0}$ for the $TEM_{0,0}$ mode. Since

$$L_{0,0} = 1 - \Gamma_0^2 \exp(-2\Lambda). \tag{25}$$

The threshold-gain ratio can be written,

$$t = \frac{\Gamma_0^2}{1 - L_{0,0}}. \tag{26}$$

Equation (26) is shown in Fig. 3 as a function of $L_{0,0}$. It is evident that the threshold-gain ratio increases with the loss and hence better mode discrimination is obtained as the loss increases.[1] Furthermore, the power output is related to the power loss per transit. Therefore different values of Γ_0^2 can be used to shape the spatial distribution of the power output.

A comparison is made (similar to that in Ref. 1) between the threshold-gain ratio of a Fabry-Perot resonator with URM and a resonator with VRM as a function of the loss per transit.

Based on the Vainshtein resonator theory[10] for the URM resonator the threshold gain ratio

$$t = \left(\frac{1}{1 - L_d}\right)^{[(v_{1,0}/v_{0,0})^2 - 1]} \tag{27}$$

where L_d is the diffraction loss for the $\mathrm{TE}_{0,0}$ mode and $v_{n,0}$ is the first nonzero root of $J_n(v) = 0$.

Equation (27) is also plotted in Fig. 3. It is evident from this figure that the threshold-gain ratio as a function of the diffraction loss is higher for the Fabry-Perot resonator with uniform mirrors than the corresponding ratio for the variable reflectivity resonator. This con-

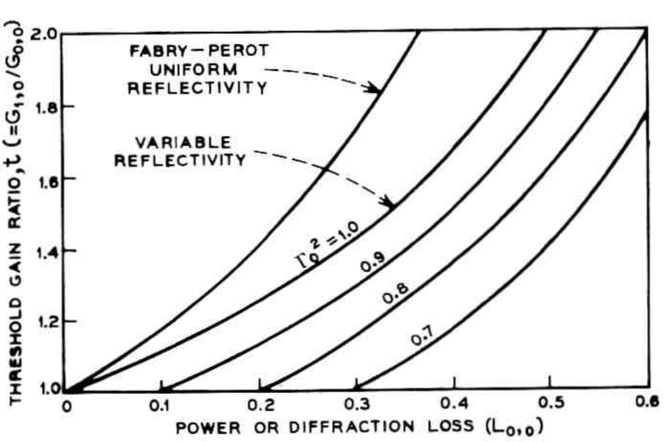


Fig. 3—Comparison of uniform and variable reflectivity mirror resonators.

clusion was previously reached by Vlasov and Talanov,[1] who also showed that the highest threshold-gain ratio is obtained with confocal resonators with uniform reflectivity mirrors. However, the high threshold-gain ratio is only of primary importance for low-gain lasers, where the output power obtained by partial transmittivity of the mirrors is only a fraction of the power lost by diffraction. For high-gain lasers, however, the mode utilization volume is of prime importance and the threshold-gain ratio can be kept at a specified level by the proper choice of the loss per pass. For such lasers the resonators with variable reflectivity mirrors have the advantage that the power loss which is necessary for mode discrimination can also be utilized as the power output. The mode volume utilization aspect is discussed later.

III. COMPUTED TEM$_{0,0}$ MODE CHARACTERISTICS

3.1 *Spot Sizes*

The TEM$_{0,0}$ mode is of particular interest since it is the fundamental mode having the highest eigenvalue and hence the lowest loss. For this mode the field distributions are gaussian with quadratic phase variations. Specifically the field distributions for the TE$_{0,0}$ mode from equations (7) and (8) are

$$F_{i1} = \exp\left\{-\frac{M}{4g_2}\left[\exp(-\Lambda)\cos\Delta + j\sinh\Lambda\sin\Delta\right]\rho_1^2\right\}, \tag{28}$$

$$F_{r1} = \exp\left\{-\frac{M}{4g_2}\left[\exp(\Lambda)\cos\Delta + j\sinh\Lambda\sin\Delta\right]\rho_1^2\right\}, \tag{29}$$

$$F_{r2} = \exp\left\{-\frac{Mg_2}{2}\left(\frac{\cos\Delta + j\sinh\Lambda}{\sin\Delta + \cosh\Lambda}\right)\rho_2^2\right\}, \tag{30}$$

where F_{i1} and F_{r1} are the field distributions of the incident and reflected fields at M1, and F_{r2} is the reflected field at M2.

The reflection coefficient can be expressed in terms of Λ and Δ is by using equation (12) as

$$\Gamma = \Gamma_0 \exp(-M/2g_2 \sinh\Lambda \cos\Delta\ \rho_1^2). \tag{31}$$

The eigenvalue for the TEM$_{0,0}$ mode is

$$K_0 = \Gamma_0 \exp(-j2kd) \exp[-(\Lambda + j\psi_0)] \tag{32}$$

with

$$\psi_0 = \frac{\pi}{2} - \cos\Delta. \tag{33}$$

The amplitudes of the field distributions at the mirrors are completely characterized by the spot sizes defined by that radius when the above quantities assume the value of $1/e$. Since the exponents in equations (28) through (31) are proportional to M, it is convenient to introduce the Fresnel numbers of the spot sizes. For example, for F_{i1} the spot size is defined by

$$\frac{M}{4g_2} \exp(-\Lambda) \cos \Delta \rho_{i1}^2 = 1 \tag{34}$$

or

$$N_{i1} = \frac{2}{\pi} \frac{\exp(\Lambda)}{\cos \Delta} g_2 \tag{35}$$

where $N_{i1} = \rho_{i1}^2/\lambda d$ is the Fresnel number. The corresponding Fresnel numbers of equations (29) through (31), N_{r1}, N_{r2}, N_m are defined in an analogous manner.

The above Fresnel numbers have been computed as a function of the threshold-gain ratio $t = \exp(2\Lambda)$, with the radii of curvature as parameters. Two types of resonators were considered: (*i*) a resonator with a uniformly reflecting plane mirror and a curved mirror with variable reflectivity with radius of curvature as a parameter, and (*ii*) a resonator with plane mirror with a variable reflectivity and a uniformly reflecting mirror with radius of curvature as a parameter.

For resonator (*i*), Figs. 4 through 6 show the Fresnel numbers of:

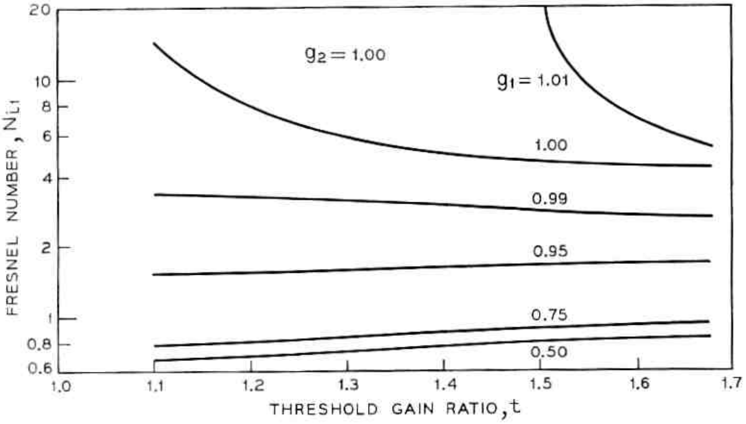


Fig. 4—Spot size of incident beam, N_{i1}.

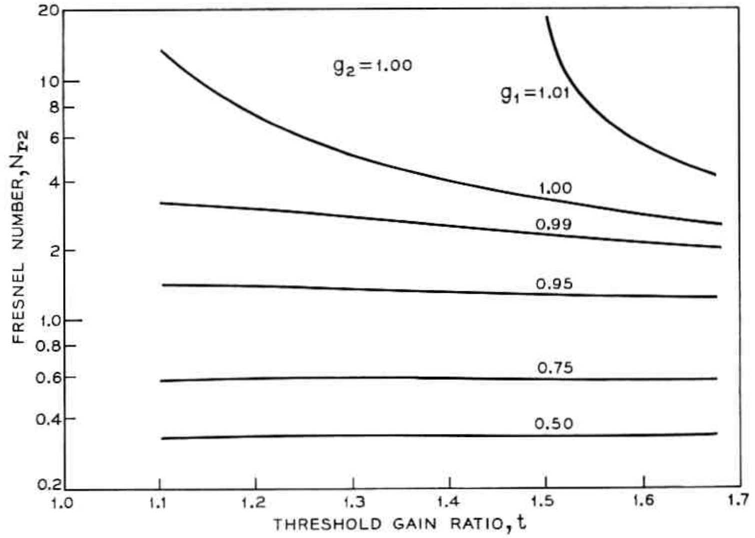


Fig. 5—Spot size of reflected beam, N_{r2}.

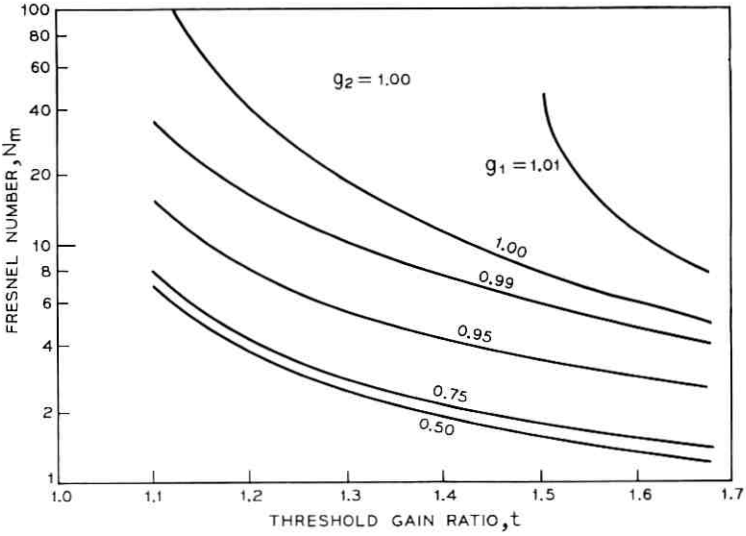


Fig. 6—Spot size of variable reflectivity mirror, N_m.

the incident beam at M1, N_{i1}, the reflected beam at M2, N_{r2} and the Fresnel number of the variable reflectivity mirror N_m. Figure 7 shows the phase of the eigenvalue ψ_0, equation (33). Figures 8 through 10 show the corresponding quantities for the type (*ii*) resonator. The phase ψ_0 is the same as in Fig. 7 but with g_1 and g_2 interchanged.

A comparison of the characteristics for the two types of resonators shows that the most pronounced differences are when either of the mirrors have curvatures g_1 or $g_2 = 0.5$. Larger spot sizes are obtainable with the type (*i*) resonator. The Fabry-Perot resonator for which g_1 and $g_2 = 1.0$ is a special case for both types. It also may be noted that a large increase in the spot size occurs when one of the mirrors is slightly convex, e.g., g_1 or $g_2 = 1.01$. This increase is caused by the curvature of the resonator mirror which approaches the unstable region, Fig. 2.

For a finite resonator, the resonator diameter will be limited by the minimum obtainable reflectivity at the mirror edges. At the spot-size diameter the reflection coefficient of the mirror has the value $1/e$. The spot-size diameter may be considered as a measure for the diameter of the VRM. The Fresnel number of the spot size of the incident beam, which for a variety of geometries is the maximum spot size of the beam along the resonator is related to the Fresnel number of the

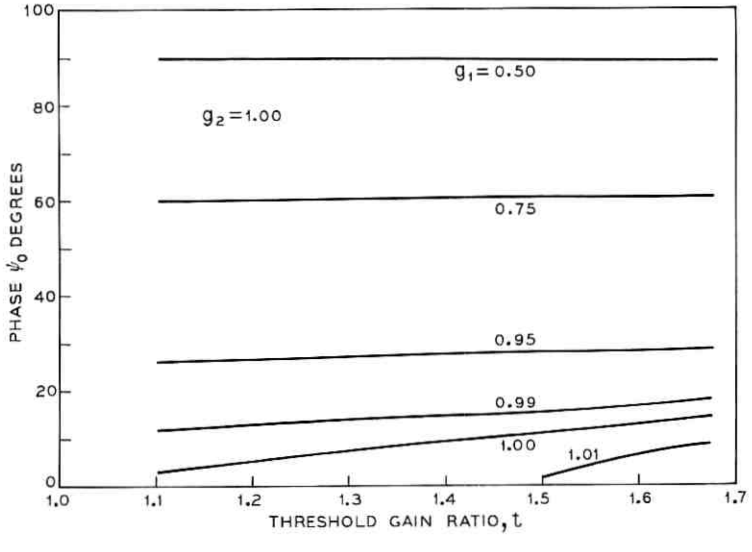


Fig. 7—Phase of the eigenvalue, $K_{0,0}$.

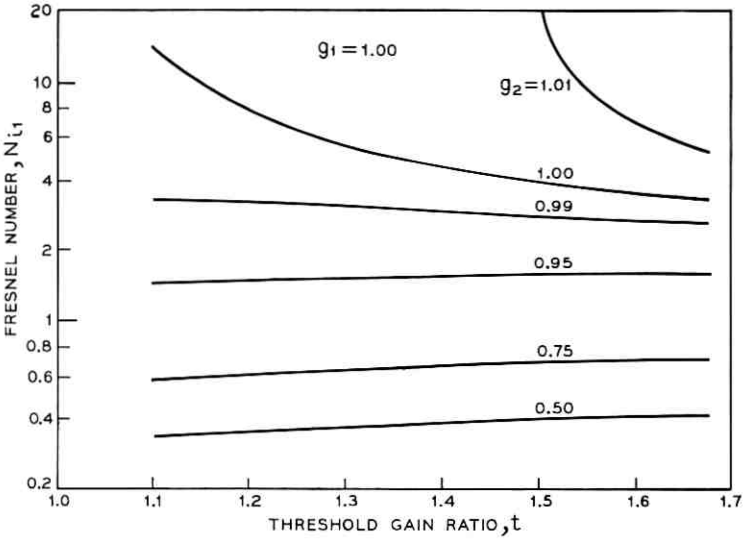


Fig. 8—Spot size of incident beam, N_{i1}.

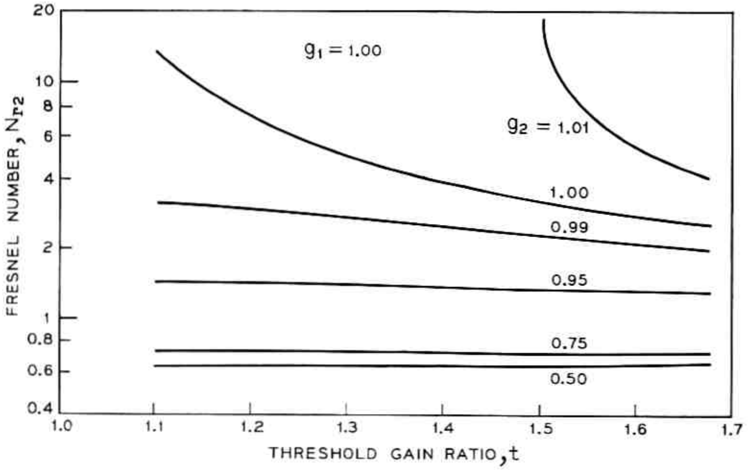


Fig. 9—Spot size of reflected beam, N_{r2}.

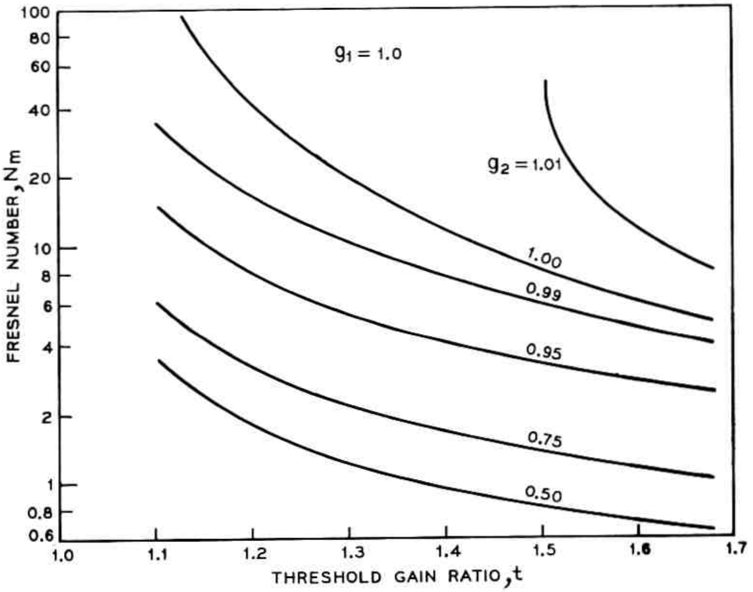


Fig. 10—Spot size of variable reflectivity mirror, N_m.

mirror spot size by:

$$\frac{N_{i1}}{N_m} = [\exp(2\Lambda) - 1] = t - 1. \tag{36}$$

For a resonator with finite dimensions to be a good approximation to the infinite resonators, it is necessary that beam power outside the mirrors be small. To obtain an estimate of this power, a resonator is assumed with a diameter equal to the mirror spot size diameter.

The ratio p, of the incident power outside the mirror spot-size diameter to the total incident power is from equations (28) and (1) given by

$$p = \frac{\displaystyle\int_{1/\sqrt{\beta}}^{\infty} \exp[-M/2g_2 \exp(-\Lambda) \cos \Delta\rho_1^2]\rho_1 \, d\rho_1}{\displaystyle\int_0^{\infty} \exp[-M/2g_2 \exp(-\Lambda) \cos \Delta\rho_1^2]\rho_1 \, d\rho_1}. \tag{37}$$

After performing the integration and substituting equation (12), this ratio can be written as

$$p = \exp\{-2/[\exp(2\Lambda) - 1]\}. \tag{38}$$

It readily follows from equation (38) that for a threshold-gain ratio $t = \exp(2\Lambda)$ smaller than 1.43, p is less than one percent.

A comparison is made between the characteristics of Fabry-Perot resonators with one large (such that the diffraction loss is negligible) uniformly reflecting mirror and with the other mirror being either of uniform or variable reflectivity. For the resonator with VRM, the Fresnel number for a given diffraction loss has twice the value than that if both mirrors are of the same size. Figure 11 shows the Fresnel number of the uniform mirror as a function of threshold-gain ratio. The curve is based on the Vainshtein resonator theory.[10] In the same figure is also shown the Fresnel number of the spot size of the incident field N_{i1}. It is evident from this figure that the spot-size diameter at the VRM is considerably larger than the diameter of the uniform reflectivity mirror for the same values of t. As an example the special case of a resonator with a uniform mirror with a Fresnel number of two is considered. For this resonator, the field at the mirror has been computed by T. Li.[5] Though the field distribution for the $\text{TEM}_{0,0}$ is not gaussian, for comparison purposes the Fresnel number of the spot size (based on the $1/e$ value for the field) is estimated to be about 1.6. For the same value of t the Fresnel number of spot size for the

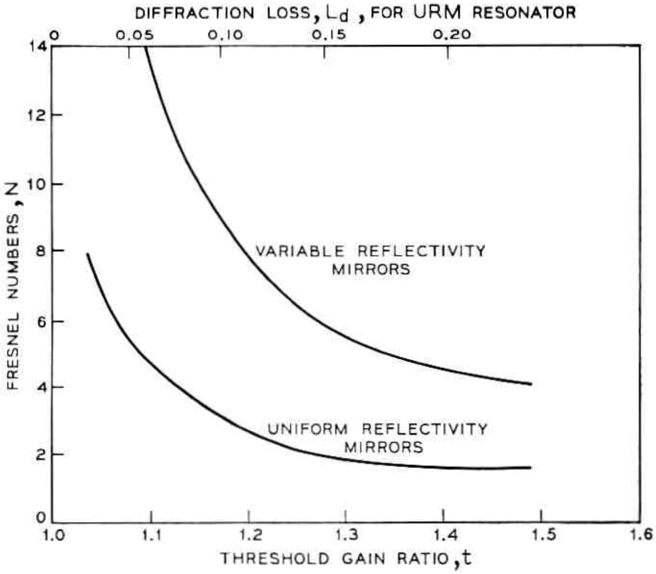


Fig. 11—Comparison of the beam sizes of Fabry-Perot resonators.

VRM resonator is 5.2. For lower values of t the difference in the spot sizes is even more pronounced. One of the advantages of VRM resonator is therefore the larger spot sizes and hence the large potential for mode volume utilization.

3.2 *Field Distributions in the Resonator*

For efficient mode volume utilization, the field along the resonator should be reasonably uniform. The uniformity of the fields is strongly dependent on the mirror curvatures. Referring to Fig. 1, let $B1$ be the reflected beam from M1 and $B2$ the reflected beam from M2. The functional dependence of the two beams on the longitudinal z coordinate has been obtained from the fields at the mirrors, and is given by the following equations

$$B1 = \exp\,[-\gamma_1(z)]\,\frac{\rho_1^2}{2}\,, \tag{39}$$

$$B2 = \exp\,[-\gamma_2(z)]\,\frac{\rho_2^2}{2}\,, \tag{40}$$

with

$$\gamma_1(z) = \frac{M\left\{2g_2\left(\dfrac{d}{z}\right)^2 \exp(\Lambda)\cos\Delta + j\dfrac{d}{z}\left[(\exp(\Lambda) + a\sin\Delta)^2 + a^2\cos^2\Delta - 2g_2\dfrac{d}{z}(\exp(\Lambda)\sin\Delta + a)\right]\right\}}{(\exp(\Lambda) + a\sin\Delta)^2 + a^2\cos^2\Delta}\,, \tag{41}$$

$$a = 1 + 2g_2\left(\frac{d}{z} - 1\right), \tag{42}$$

and

$$\gamma_2(z) = \frac{M\left[2g_2\left(\dfrac{d}{d-z}\right)^2 \exp(\Lambda)\cos\Delta + j\left(\dfrac{d}{d-z}\right)\left\{(\exp(\Lambda)b - \sin\Delta)^2 + \cos^2\Delta + \dfrac{d}{d-z}\left[\exp(2\Lambda)b + 2\exp(\Lambda)\sin\Delta(g_2 - 1) - 1\right]\right\}\right]}{\left[\exp(\Lambda)\left(b + \dfrac{d}{d-z}\right) + \left(\dfrac{d}{d-z} - 1\right)\sin\Delta\right]^2 + \left(\dfrac{d}{d-z} - 1\right)^2\cos^2\Delta} \tag{43}$$

with

$$b = (2g_2 - 1). \tag{44}$$

The real part of equations (41) and (43) has been evaluated as a function of z, in terms of the spot-size Fresnel number $N_1(z)$ and $N_2(z)$ defined in accordance with equation (34) as $N_{1,2}(z) = 2/\gamma_{1,2}(z)\lambda d$. For a resonator with $g_2 = 1.0$, Fig. 12 shows the spot-size Fresnel number as a function z/d for a number of parameters.

It is characteristic of resonators with VRM, that minimum-beam spot size even for symmetrical resonators does not occur at half the mirror spacings in contrast to resonators with uniform reflectivity mirrors. In Fig. 12 this characteristic is particularly evident for the equivalent confocal resonator $g_1 = 0.5$.

The uniformity of the beams along the resonator increases with the

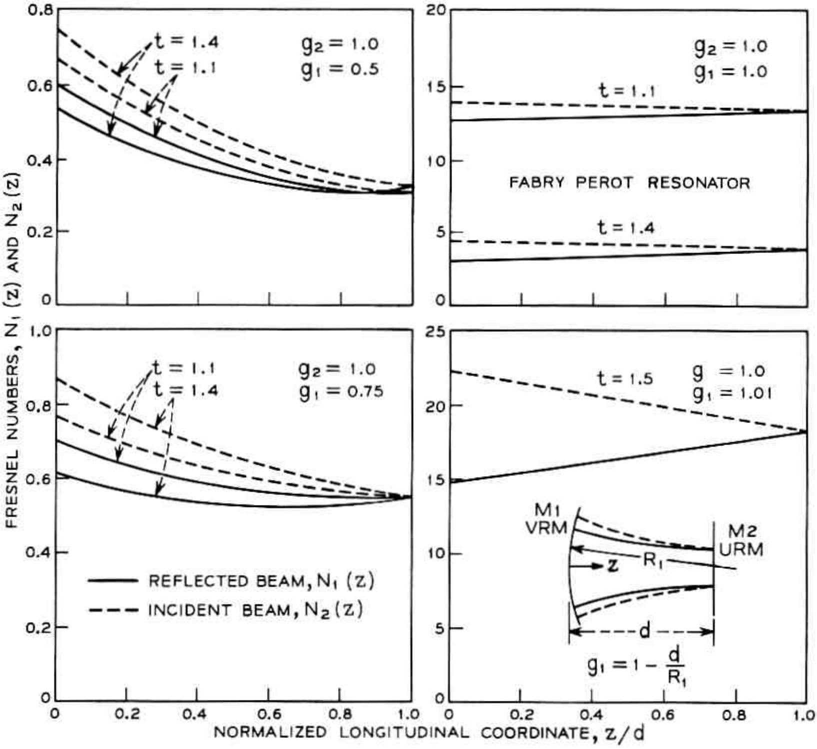


Fig. 12—Spot size of the beams along the resonator.

increasing value of g_1 up to $g_1 = 1.0$ which corresponds to the Fabry-Perot resonators. For this resonator the beams are most uniform. For the higher value of $g_1 (g_1 = 1.01)$ the uniformity of the beams within the resonator decreases.

3.3 *Minimum Spot Sizes*

The uniformity of the beams within the resonator depends on the locations of the minimum spot-size diameters. The beam diameters are more uniform if the virtual minimum spot-size diameters occur at distances, away from the mirrors, which are large in comparison to the separation of the mirrors. The positions of the minimum spot sizes are obtained either by determining the maxima or by setting the imaginary parts of equations (41) and (43) equal to zero. Either condition gives the same result (i.e., at the minimum spot-size positions the beams have constant phase).

The minimum positions for the two beams $(z_1/d)_{\min}$ and $(z_2/d)_{\min}$ are:

$$\left(\frac{z_1}{d}\right)_{\min} = 2g_2 \frac{[2g_2 - 1 - \exp(\Lambda) \sin \Delta]}{[\exp(2\Lambda) + 2(1 - 2g_2) \exp(\Lambda) \sin \Delta + (1 - 2g_2)^2]} \tag{45}$$

and

$$\left(\frac{z_2}{d}\right)_{\min} = 2g_2 \exp(\Lambda) \frac{[\exp(\Lambda)(2g_2 - 1) - \sin \Delta]}{[\exp(2\Lambda)(1 - 2g_2)^2 + 2(1 - 2g_2) \exp(\Lambda) \sin \Delta + 1]}. \tag{46}$$

The Fresnel numbers of the minimum spot sizes $[N_1(z)]_{\min}$ and $[N_2(z)]_{\min}$ are:

$$[N_1(z)]_{\min} = \frac{2g_2 \cos \Delta \exp(\Lambda)}{\pi[\exp(2\Lambda) + 2(1 - 2g_2) \exp(\Lambda) \sin \Delta + (1 - 2g_2)^2]} \tag{47}$$

and

$$[N_2(z)]_{\min} = \frac{2g_2 \cos \Delta \exp(\Lambda)}{\pi[\exp(2\Lambda)(1 - 2g_2)^2 + 2(1 - 2g_2) \exp(\Lambda) \sin \Delta + 1]}. \tag{48}$$

The minimum positions and minimum spot sizes have been computed for a resonator with a VRM and variable radius of curvature and a uniform plane mirror $g_2 = 1.0$. Since this resonator is equivalent to a symmetrical resonator with two VRM, the two minimum positions are mirror images with respect to the uniform mirror, and the minimum spot sizes are the same. For this resonator

$$[N_1(z)]_{\min} = [N_2(z)]_{\min} = \frac{1}{\pi} \frac{\cos \Delta}{[\cosh \Lambda - \sin \Delta]}. \tag{49}$$

For specified Λ equation (49) has a maximum for $\cosh \Lambda \sin \Delta = 1$, which from equation (13) corresponds to the Fabry-Perot resonator ($g_1 = g_2 = 1.0$). The corresponding Fresnel number N_M is:

$$N_M = \frac{1}{\pi \sinh \Lambda}. \tag{50}$$

Figure 13 shows $(z_2/d)_{\min}$ as a function of g_1 with $t = \exp(2\Lambda)$ as a parameter. As g_1 increases, so does $(z_2/d)_{\min}$ assuming relatively large values in the vicinity of $g_1 = 1.0$. The large values of $(z_2/d)_{\min}$ explain the uniformity of the beams in the Fabry-Perot resonator. Figure 14 shows the dependence of the minimum spot on the mirror curvature g_1 with t as a parameter.

3.4 *Dependence of the Spot Sizes on the Curvatures of Spherical Mirrors*

The previous calculations show that large spot sizes are obtainable in the vicinity of the instability region. How critically the spot sizes depend on mirror curvatures is of importance.

The Fresnel number of the spot size for the incident beam N_{i1} is directly related to the Fresnel number of the VRM by equation (36).

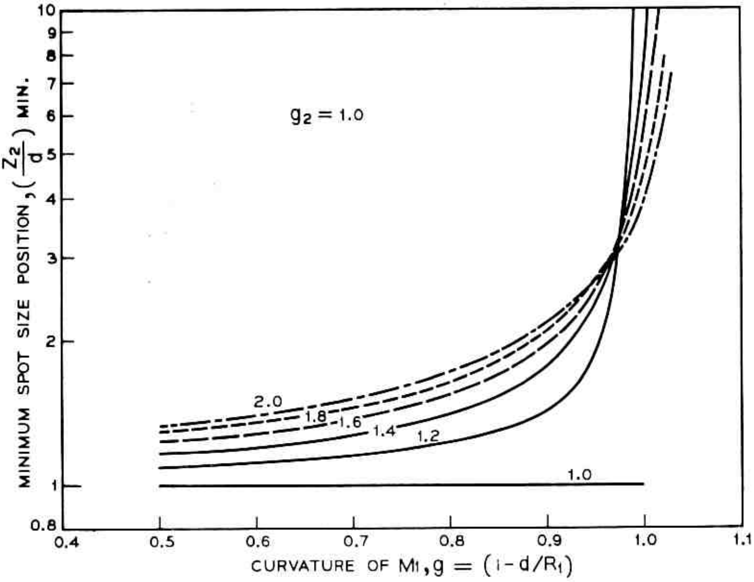


Fig. 13—Location of the minimum spot size.

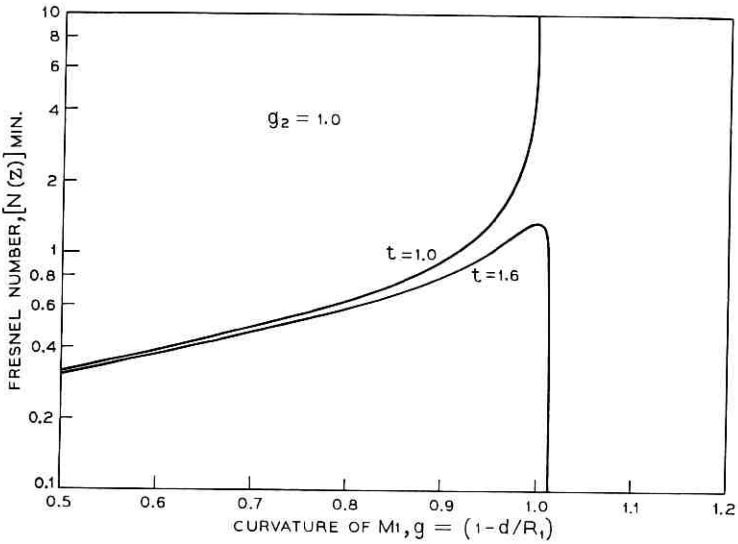


Fig. 14—Fresnel number of minimum spot size as a function of curvature of M1.

Equations (12) and (13) give the relation between the Fresnel number of VRM, the mirror curvature and Λ. Solving these equations for g_1 gives:

$$g_1 = \frac{\text{ctnh}\,\Lambda[(\pi N_m \sinh \Lambda)^2 - g_2^2]^{\frac{1}{2}} + \pi N_m}{2\pi g_2 N_m} \tag{51}$$

where N_m is related to β in equation (1) by $N_m = (1/\beta\lambda d)$.

Equation (51) has been computed as a function of g_2 with t as a parameter. Figures 15 through 18 show the computed characteristics for $N_m = 10{,}20{,}40{,}100$.

The critical dependence of the beam spot size N_{i1} on the mirror curvatures is evident from these figures, particularly as N_m increases. A small change in g_1 or g_2 results in a large change in t and there is a very large change in the beam spot size N_{i1} for a specified N_m.

The conclusion based on these computations is that though large Fresnel numbers for the beam spot sizes are in principle possible, the critical tolerance requirements for the mirror curvatures may set a practical limit on spot sizes relative to those obtainable with a Fabry-Perot resonator with one VRM.

3.5 *Resonator Design*

As an application, a resonator design is considered for a CO_2 laser. In view of the realizable high gain per pass, the power loss per pass and the related power output should be large. A Fabry-Perot resonator with a VRM mirror seems to be most suitable for this application. The remaining parameter to be specified is the threshold-gain ratio, t. This ratio should be as low as possible in order to obtain large beam diameters (see Fig. 11). For the fundamental $TEM_{0,0}$ mode operation, the limitation on t is based on the accuracy with which the gain can be controlled. A value of t of 1,2 is assumed.

Using the above parameters in equations (28) and (31) (Figs. 4 and 6) the Fresnel number of the spot size of the VRM is 38.35 and that of the incident beam at the VRM is 7.65. For a resonator with length d = 100 cm and for a wavelength $\lambda = 10^{-3}$ cm (10 micron), the radius of the spot size of the VRM, a_m = 1.95 cm.

Some of the characteristics of this resonator are shown in Fig. 19. Illustrated is the dependence of the power-reflection coefficient Γ^2 as a function of the normalized radius ρ/a_m. Also shown is the normalized-incident power density at the VRM, and power loss density at the VRM for different values of the reflectivity at the center, Γ_0^2, as functions of

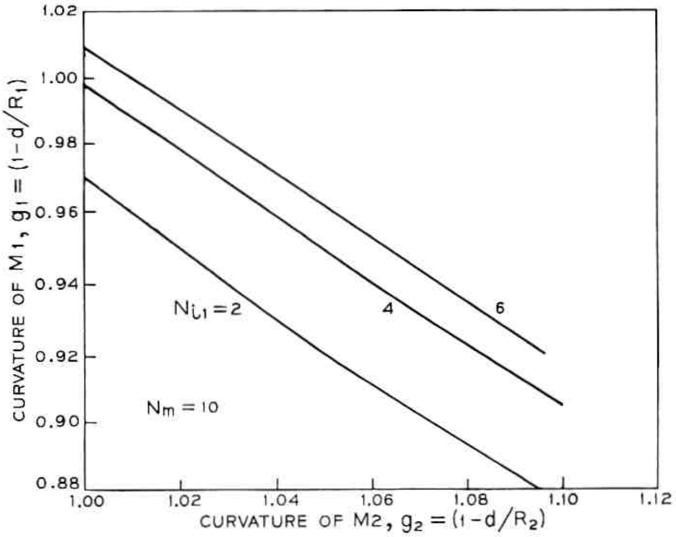


Fig. 15—Relation between mirror curvature parameters g_1 and g_2 and Fresnel number N_{i1} of spot size at M1.

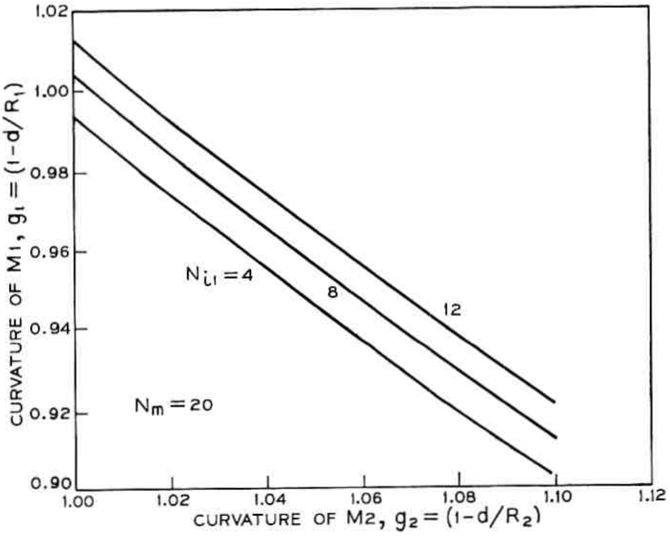


Fig. 16—Relation between mirror curvature parameters g_1 and g_2 and Fresnel number N_{i1} of spot size at M1.

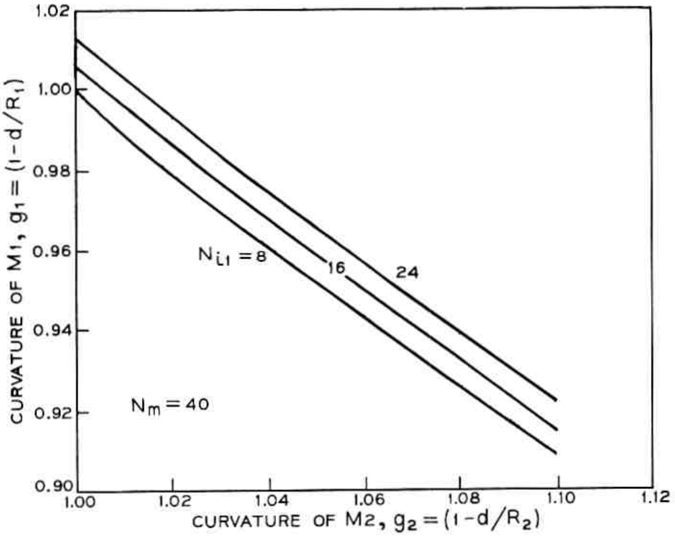


Fig. 17—Relation between mirror curvature parameters g_1 and g_2 and Fresnel number N_{i1} of spot size at M1.

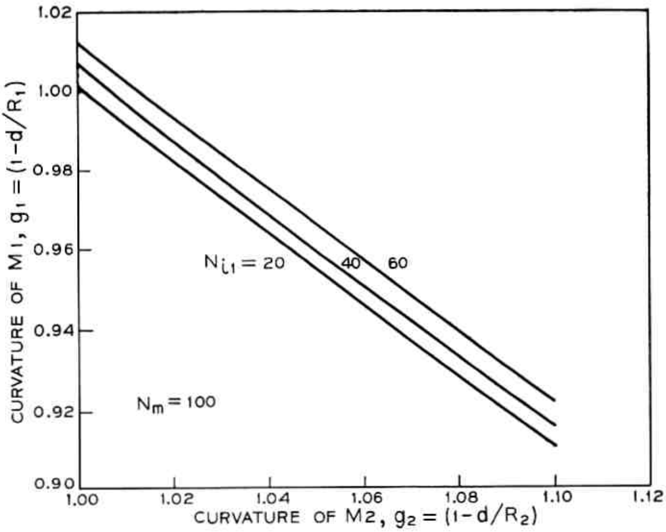


Fig. 18—Relation between mirror curvature parameters g_1 and g_2 and Fresnel number N_{i1} of spot size at M1.

ρ/a_m. The power-loss density is the laser-power output when the absorptivity of the VRM is zero. Figure 20 shows the ratio of the power loss to the incident power as a function ρ/a_m with Γ_0^2 as parameters.

The actual diameter of the VRM can presently be determined only by assuming that a finite resonator will behave similarly to a resonator with infinite mirrors when the beam power outside a certain diameter is small. For the resonator considered, 0.5 percent of the incident beam power is contained outside the mirror radius of 0.73 a_m and one percent outside the radius 0.68 a_m which corresponds for the above value of a_m to 1.41 cm and 1.31 cm radii. For a resonator with a VRM of radius a_m the perturbation of the fields should therefore be very small.

IV. CONCLUSIONS

The characteristics of optical resonators with gaussian radial variations of the mirror reflectivities have been investigated. These variable reflectivity mirror (VRM) resonators seem to be particularly suitable for high-gain and high-power laser application such as the 10.6 micron CO_2 laser. For the fundamental $TEM_{0,0}$ mode generation, these resonators have the advantage in comparison to conventional resonators that larger beam spot sizes are obtainable (with better mode-volume

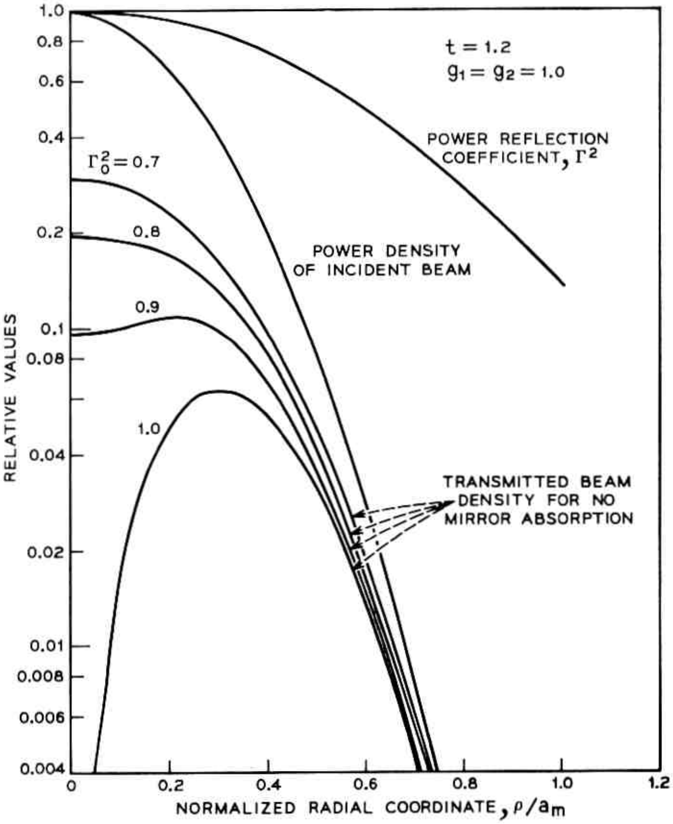


Fig. 19—Power distribution at the variable reflectivity mirror.

utilization) and the power loss necessary for mode discrimination can be utilized as the power output.

The factors limiting the spot are the threshold-gain ratio and the mirror-curvature tolerances.

The Fabry-Perot resonator with a VRM is stable and furthermore the field distribution along the resonator is more uniform in diameter relative to other resonator geometries. A specific design of such a resonator with a gain threshold ratio $(G_{1,0}/G_{0,0})$ of 1.2 shows that a spot size Fresnel number of 7.65 with a power loss (or power output depending on the absorptivity of the mirrors) as high as 40 percent of the incident power are obtainable.

In this investigation it was assumed that the mirrors are infinite.

The results presented should be a good approximation to a finite resonator when the beam power outside a certain circular region has a negligible value (say less than one percent).

V. ACKNOWLEDGMENTS

The author gratefully acknowledges the valuable comments and suggestions of N. Amitay and H. G. Cooper.

APPENDIX A

Integrals of Laguerre Functions

In order to determine the orthogonality and power relations for the modes in resonators with variable reflectivity mirrors, the following integral $I_{m,n}^{\ell}$ of product of Laguerre functions is evaluated.

$$I_{m,n}^{\ell} = 2 \int_0^{\infty} \exp(-s\rho^2) L_m^{\ell}(\alpha\rho^2) L_n^{\ell}(\beta\rho^2) \rho^{2\ell+1} \, d\rho,$$

$$= \int_0^{\infty} \exp(-st) L_m^{\ell}(\alpha t) L_n^{\ell}(\beta t) t^{\ell} \, dt. \tag{52}$$

For the special case $m = n$ the integral is known.[11,12] The integral

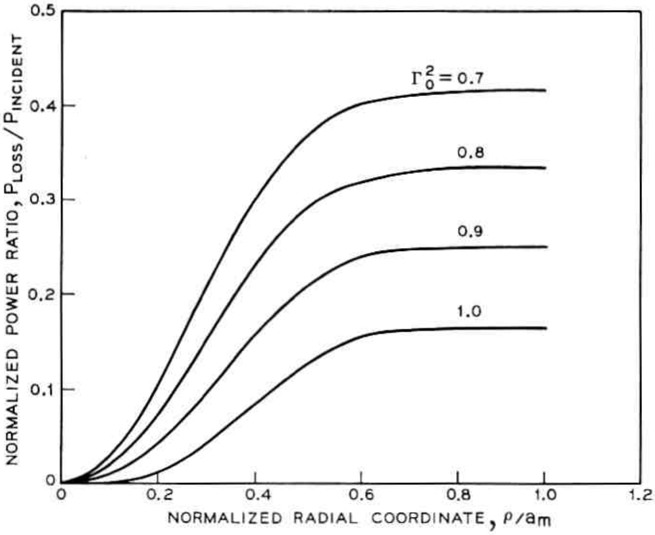


Fig. 20—Ratio of the loss to the incident beam power as a function of ρ/a_m.

(52) is evaluated by considering this integral as a Laplace transform of two functions $f_1(t)$ and $f_2(t)$ and using the Faltung relation

$$\int_0^\infty \exp(-st) f_1(t) f_2(t)\, dt = \frac{1}{2\pi j} \int_{\gamma - j\infty}^{\gamma + j\infty} F_1(z) F_2(s - z)\, dz \tag{53}$$

where $F_1(z)$ and $F_2(z)$ are the Laplace transforms of $f_1(t)$ and $f_2(t)$, γ is a constant with $\mathrm{Re}(s) > \mathrm{Re}(\gamma) > 0$.

Let

$$f_1(t) = L_m^\ell(\alpha t) = \sum_{k=0}^{m} (-1)^k \binom{m+\ell}{m-k} \frac{(\alpha t)^k}{k!} \tag{54}$$

and

$$f_2(t) = L_m^\ell(\beta t) t^\ell. \tag{55}$$

The Laplace transform of equations (54) and (55) are readily obtained. Furthermore with the transformation $\zeta = 1/z$ together with equation (53), equation (52) reduces to:

$$I_{m,n}^\ell = \frac{(n+\ell)!}{n!\, s^{m+1}} \left(-\frac{1}{\alpha}\right)^\ell \frac{1}{2\pi j} \oint \frac{(1 - \beta\zeta)^n}{\left(\zeta - \frac{1}{s}\right)^{m+1}} [(s - \alpha)\zeta - 1]^{m+\ell}\, d\zeta. \tag{56}$$

In equation (56) the contour of integration encloses the point $\zeta = 1/s$. Equation (56) is therefore evaluated by determining the residue at $z = 1/s$, which yields

$$I_{m,n}^\ell = \frac{(n+\ell)!}{n!\, s^{m+1}} \left(-\frac{1}{\alpha}\right)^\ell \frac{d^m}{d\zeta^m} \{(1 - \beta\zeta)^n [(s - \alpha)\zeta - 1]^{m+\ell}\}_{\zeta = 1/s}. \tag{57}$$

Equation (56) can be expressed in terms of Jacobi polynomials[12] by rotating and translating the coordinate system with the result that

$$I_{m,n}^\ell = \frac{(n+\ell)!}{n!\, s^{n+\ell+1}} (s - \alpha - \beta)^m (s - \beta)^{n-m} P_m^{\ell, n-m}(\eta) \tag{58}$$

with

$$\eta = \frac{s^2 + 2\alpha\beta - s(\alpha + \beta)}{s(s - \alpha - \beta)} \tag{59}$$

and $P_m^{\ell, n-m}$ is a Jacobi ploynomial defined by [12]

$$P_m^{a,b}(x) - \frac{(-1)^m}{2^m m!} (1 - x)^{-a} (1 + x)^{-b} \frac{d^m}{dx^m} [(1 - x)^{a+m} (1 + x)^{b+m}]. \tag{60}$$

As an application let $F_m^\ell(\rho)$ and $F_n^{\ell *}(\rho)$ designate the reflected fields at the VRM given by equation (7) and the* indicate the complex conjugate. The integrals which enter in the evaluation of the total reflected power due to several modes of the same index ℓ, can be written as

$$\int_0^\infty F_m^\ell(\rho) F_n^{\ell *}(\rho) \rho \, d\rho = \frac{(\alpha_1 \alpha_1^*)^{\ell/2}}{2} (I_{m,n}^\ell)_r \tag{61}$$

where α_1 is given by equation (11). Using equations (7), (9), (11), (12) and (58), it follows that

$$\frac{(\alpha_1 \alpha_1^*)^{\ell/2}}{2} (I_{m,n}^\ell)_r = \frac{(-1)^n (n + \ell)! \, g_2}{n! \, M \cos \Delta} \exp \left[-\Lambda(n + n + 1 + \ell) \right]$$
$$\cdot \left[1 + \frac{\sin h^2 \Lambda}{\cos^2 \Delta} \right]^{\ell/2} \left[\frac{\sin h\Lambda}{\cos \Delta} \exp(-j\Delta) \right]^{m-n} P_m^{\ell, n-m}(\eta) \tag{62}$$

with

$$\eta = -\left[1 + 2 \frac{\sin h^2 \Lambda}{\cos^2 \Delta} \right]. \tag{63}$$

For stable resonators with VRM equation (62) is not equal to zero for $m \neq n$. Hence, the total reflected power is in general not equal to the sum of the powers of the individual modes.

The reflected field $F_m^{(\ell)}(\rho)$ for any particular mode is related to the incident field $F_m^{(l)}(\rho)_i$ by reflection coefficient (1). The evaluation of the corresponding integral (61) for the incident fields yields the same value for η. Furthermore setting $m = n$ the following relation if obtained

$$\frac{(I_{n,n}^\ell)_r}{(I_{n,n}^\ell)_i} = \Gamma_0^2 \exp \left[-2\Lambda_0(2n + \ell + 1) \right] = | K_\ell |^2. \tag{64}$$

The meaning of equation (64) is that the ratio of the reflected to incident power for a particular mode is precisely equal to absolute value of the eigenvalue squared.

REFERENCES

1. Vlasov, S. N., and Talanov, V. I., "Selection of Axial Modes in Open Resonators," Radio Eng. and Elec. Phys., *10,* No. 3 (March 1965), pp. 469–470.
2. Vakhimov, N. G., "Open Resonators With Mirrors Having Variable Reflection Coefficients," Radio Eng. and Elec. Phys., *10,* No. 9 (September 1965), pp. 1439–1446.
3. Kumagai, N., et. al., "Resonant Modes in a Fabry-Perot Resonator Consisting of Nonuniform Reflectors," Elec. and Comm. Japan, *49,* No. 7 (July 1966), pp. 1–8.

4. Suematsu, Y., et al., "A Light Beam Waveguide Using Gaussian Mode Filters," Elec. and Comm. in Japan, *51-B,* No. 4 (April 1968), pp. 67–74.
5. Li, Tingye, "Diffraction Loss and Selection of Modes in Maser Resonators With Circular Mirrors," B.S.T.J., *44,* No. 5 (May–June 1965), pp. 917–932.
6. Howell, W. T., "On a Class of Function Which are Self Reciprocal in the Hankel Transform," Phil. Mag, *25,* Series 7 (April 1938), pp. 622–628.
7. Erdelyi, A., et. al., *Tables of Integral Transform,* Vol. 2, New York: McGraw-Hill, 1954, p. 43.
8. Boyd, G. D., and Kogelnik, H., "Generalized Confocal Resonator Theory," B.S.T.J., *41,* No. 4 (July 1962), pp. 1347–1369.
9. La Tourette, et. al., "Improved Laser Angular Brightness Through Diffraction Coupling," Appl. Opt., *3,* No. 8 (August 1964), pp. 981–982.
10. Vainshtein, L. A., "Open Resonators for Lasers," Sov. Phys., JETP, *17* (September 1963), pp. 709–719.
11. Buchholz, H., *Die Konfluente Hypergeometrische Function,* Berlin: Springer-Verlag, 1953, p. 144.
12. Gradshteyn, I. S., and Ryzhik, I. M., *Tables of Integrals, Series and Products,* New York: Academic Press, 1965, pp. 845 and 1035.

Projecting Filters for Recursive Prediction of Discrete-Time Processes

By ALLEN GERSHO and DAVID J. GOODMAN

(Manuscript received June 19, 1970)

We consider the design of time-invariant recursive filters of constrained order for one-step prediction of discrete-time stationary processes. For this purpose, we introduce the projecting-filter concept. An nth-order projecting filter for a given process has the characterizing property that with the process as input, the output at each instant is the optimal linear combination of the n previous output and n latest input samples. This definition implies that (i) the filter is stable, (ii) any $n + 1$ consecutive samples of the prediction error sequence are mutually uncorrelated, (iii) the mean-square prediction error is at least as low as that of the best nth order nonrecursive predictor, and (iv) if the spectral density of the process is rational of order $2n$ or less, then the nth-order projecting filter coincides with the optimal (unconstrained) linear predictor.

A design algorithm for nth-order projecting filters iteratively generates successive sets of coefficients of a time-varying nth-order recursive filter which asymptotically approaches the desired time-invariant filter. The only input data needed for the algorithm are the autocovariance coefficients of the process to be predicted. When the order of the filter is matched to the order of the process, the time-varying filter is the same as the Kalman predictor. The algorithm has yielded effective projecting filters for several specific processes. Our results indicate that near optimal prediction may often be obtained with filters of order lower than that of the optimal unconstrained predictor.

I. INTRODUCTION

Although the optimal linear predictor of a random process must make use of the entire past of the process, any practical predictor can store only a finite number of data. One way to design a finite storage predictor is to determine the best linear combination of the n latest sample values of the process. However, for many processes, a large value of n is required to achieve a performance quality approaching that of the unconstrained optimal linear predictor. An alternate approach is to find the best recursive predictor constrained to operate only on the n latest data samples and the n latest predictions. This approach has the advantage of using condensed information from the entire past of the process with the consequence that optimal or near optimal prediction can often be achieved with a relatively small amount of storage.

The purpose of this paper is to introduce the projecting-filter approach to recursive prediction and to present an algorithm for the design of projecting filters that has yielded effective low-order predictors not otherwise attainable. So far, a complete theory of projecting filters has not been established. We do not yet know how broad is the class of processes which possess projecting filters of a given order; nor have we determined the class of processes for which our design algorithm is effective. However, we can report very favorable experience in the design of projecting filters for a variety of specific processes. We have also established some important theoretical properties of projecting filters.

1.1 *Optimal and Finite Memory Predictors*

In certain special cases the optimal (least mean-square error) unconstrained predictor is realizable with a finite-storage filter.[1] In particular, for an nth-order autoregressive, or wide-sense Markov, process the optimal unconstrained predictor is a finite-memory nonrecursive filter operating only on the n latest data samples. More generally, the optimal unconstrained predictor of any stationary process whose spectral density is rational of order $2n$ may be implemented as an nth-order recursive filter. The characteristics of the optimal filter may be determined by applying the discrete-time form of Wiener's spectral factorization technique. Even more generally, consider any nonstationary process which can be modeled as the response of an nth-order linear time-varying recursive filter to an uncorrelated noise input. The optimal unconstrained predictor is an nth-order time-varying recursive filter[2] which may be determined by use of the Kalman filtering equations,[3] or, more efficiently, by a generalization of the approach taken in Section VI of this paper.

If a random process cannot be modeled as the response of an nth-order recursive filter to an uncorrelated input, then the optimal unconstrained one-step linear predictor cannot be realized by an nth-order filter. Nevertheless, it is realistic to preselect the desired order,

n, of the predictor and to seek the best recursive filter of this order. In this way the structure of the predictor is conveniently specified for digital filter implementation while only the $2n$ parameter values need be supplied according to the process to be predicted. Unfortunately, with the least-mean-square error criterion, the constrained-order prediction problem is a special case of the unsolved problem of L_2 rational approximation on the unit circle.[4] No analytical solution is known and optimization search techniques are severely hampered by the multimodal nature of the error surface.*

1.2 *Projecting Filters*

In this paper we introduce the projecting filter principle of recursive prediction. Although the projecting filter is not a solution of the L_2 rational approximation problem, it has the local optimality property that at each step it forms the best linear combination of the available data. The term "projection" alludes to the geometrical interpretation of random variables as vectors in Hilbert space.[6, 7] Each prediction error of the projecting filter is a vector orthogonal to the n most recent inputs and the n previous errors. Hence the projecting filter performs a partial whitening of the input process. In this sense it approximates the action of the optimum unconstrained predictor, the error of which is a white-noise process—the innovations process of the input. If the input can be represented as the response of an nth-order filter to white noise, the nth-order projecting filter is the optimum unconstrained predictor. For any process, the mean-square error of a projecting filter is never greater than the mean-square error of the optimum nonrecursive filter of the same order. Projecting filters are stable.

1.3 *An Example*

These properties of projecting filters are observed in the example of the eighth-order process $\{x_k\}$ represented by

$$x_k = \epsilon_k - 0.8\epsilon_{k-1} + 0.5\epsilon_{k-2} + 0.25\epsilon_{k-3} - 0.6\epsilon_{k-4} - 0.2\epsilon_{k-5} + 0.1\epsilon_{k-6} + 0.4\epsilon_{k-7} - 0.08\epsilon_{k-8}$$

in which $\{\epsilon_k\}$ is a stationary white-noise process with zero mean and unit variance. The power spectral density function of $\{x_k\}$ has zeros at the 16 points in the z-plane indicated in Fig. 1. The eighth-order projecting filter for $\{x_k\}$, which is the optimum unconstrained predic-

* The complexity of the error as a function of the parameters is evidenced by the work of R. S. Phillips[5] on the corresponding continuous-time problem.

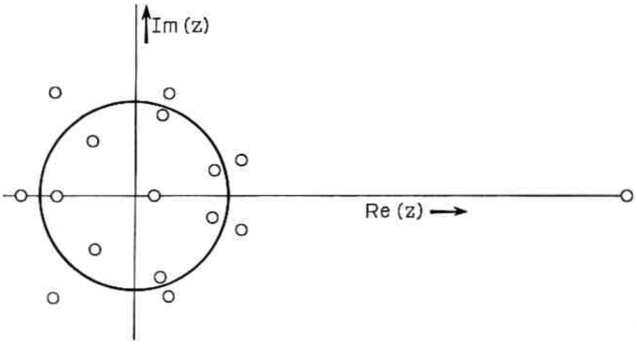


Fig. 1—Locations of zeros of the spectral density function of an eighth-order process. The eighth-order projecting filter has poles at the zero locations that are outside the unit circle.

tor, has poles at the eight locations indicated in Fig. 1 that are outside the unit circle. The pole positions of a seventh-order projecting filter are shown in Fig. 2. There are poles extremely close to all of the locations outside the unit circle indicated in Fig. 1, except the one furthest from the origin. Figures 3, 4, and 5 indicate the pole locations of the sixth-, third-, and first-order projecting filters, respectively. The poles of these filters do not coincide with zeros of the power spectral density function of $\{x_k\}$.

Figure 6 demonstrates the projecting-filter mean-square-error performance for this process. Here the horizontal base line is the optimal unconstrained prediction error. The white bars indicate errors of optimal constrained order nonrecursive predictors and the shaded bars are

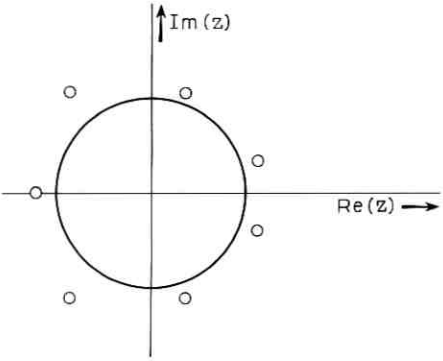


Fig. 2—Pole locations of seventh-order projecting filter.

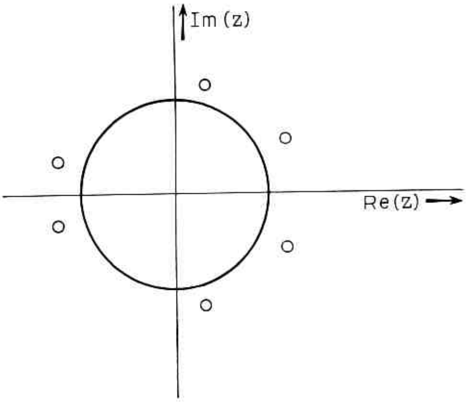


Fig. 3—Pole locations of sixth-order projecting filter.

the errors of the projecting filters. It is significant that the error of the seventh-order projecting filter is extremely close to the optimum linear-prediction error; the ratio of the two errors is approximately $1 + 10^{-7}$. By using the projecting filter approach to prediction, we have discovered a means of reducing predictor complexity with virtually no loss in accuracy. In addition, Fig. 6 shows the error resulting from low-order recursive filters and the advantages relative to nonrecursive prediction.

1.4 *Organization of the Paper*

The content of the paper falls into two categories. Some sections contain descriptive and analytic material relevant to predictors and

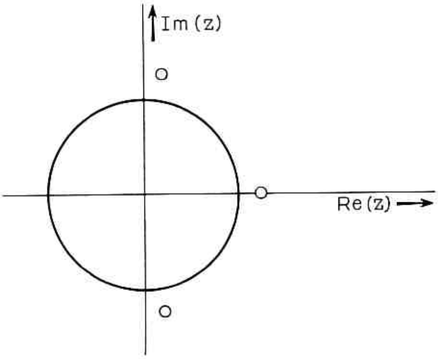


Fig. 4—Pole locations of third-order projecting filter.

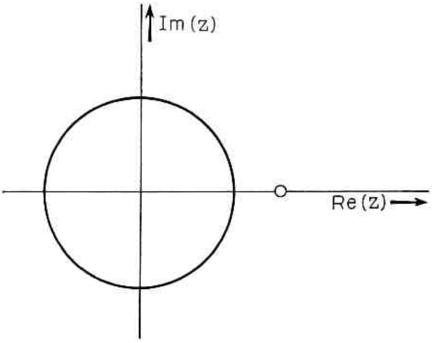


Fig. 5—Pole location of first-order projecting filter.

projecting filters in general and other sections pertain to the particular design method that has been used in synthesizing the predictors described in Section 1.3. Sections II, III and IV are in the first category; they define the prediction problem and the projecting-filter principle and focus attention on the essential properties of unconstrained predictors and projecting filters. Section V introduces the design method, an iterative scheme based upon successive projections in Hilbert space. This technique leads to a time-varying filter that asymptotically tends towards the desired projecting filter. Section VI shows that when the

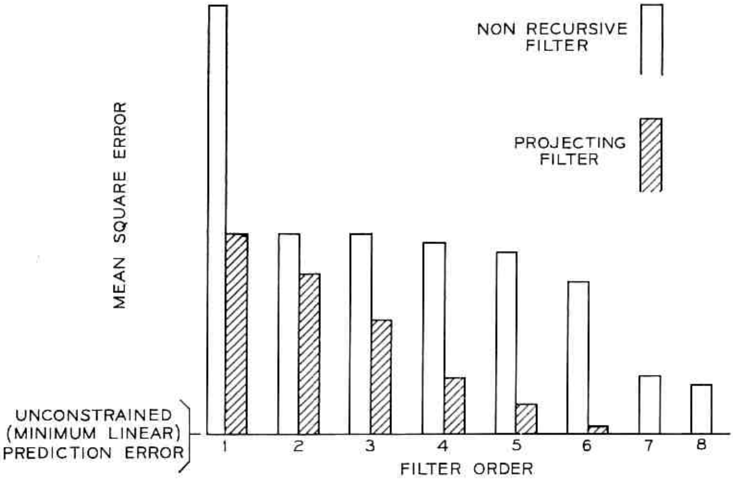


Fig. 6—Mean-square errors of projecting filters and optimal nonrecursive filters of orders 1 through 8.

order of the filter is matched to that of the process, the design algorithm converges and the projecting-filter approach results in an efficient analysis and design (equivalent to but simpler than the Kalman filtering equations) for the unconstrained optimum time-varying filter with a given initial state. Section VII presents a derivation of the design algorithm.

II. PROBLEM STATEMENT

We consider a purely-nondeterministic* stationary process $\{x_k\}$ with known covariance function, $r_k = Ex_ix_{i+k}$. We assume that the spectral density function of the process $f(z) = \Sigma r_k z^k$ has no zeros on the unit circle, $|z| = 1$. The purpose of this paper is to describe a new approach to the design of a stable one-step predicting filter with the nth-order recursive structure

$$y_k = \sum_{i=0}^{n-1} a_i x_{k-i} + \sum_{i=1}^{n} b_i y_{k-i}. \tag{1}$$

A natural measure of the performance of the predictor is the mean-square value of the prediction error

$$e_{k+1} = x_{k+1} - y_k. \tag{2}$$

Because the determination of the optimum filter coefficients with respect to this criterion is an intractable problem of approximation theory, our design method is based on a different performance objective. Rather than synthesize the least-squares nth-order recursive filter, we seek a stable time-invariant filter with the following

> *Projecting property: With input $\{x_k\}$, the output, y_k, is, at each instant k, the least mean-square linear combination of the data $\{x_k, x_{k-1}, \cdots, x_{k-n+1}, y_{k-1}, \cdots, y_{k-n}\}$ currently in the filter memory.*

This implies that the filter coefficients a_i and b_i satisfy a set of linear equations involving the covariance functions of $\{x_k\}$ and $\{y_k\}$. The autocovariance of $\{x_k\}$ corresponds to the given data of the prediction problem but the cross-covariance between $\{x_k\}$ and $\{y_k\}$ and the autocovariance of $\{y_k\}$ are transcendental functions of a_i and b_i. It follows that an explicit solution for the coefficients from the constraints imposed by the projecting property is not possible. An algorithmic solution is presented in Section VII.

* See Ref. 1, p. 23.

III. UNCONSTRAINED PREDICTION

We refer to the problem defined in Section II as a constrained-order prediction problem because the order, n, of the predictor is prespecified. Another problem, which we refer to as unconstrained linear prediction, has received considerable attention in the literature of stochastic processes.[1,8] The optimum unconstrained prediction, $\hat{x}_{k+1}$, of x_{k+1} is the least mean-square linear combination of the entire past, $x_k, x_{k-1}, \cdots$ of $\{x_k\}$. In the terminology of the Hilbert space description of random variables, $\hat{x}_{k+1}$ is called the projection of x_{k+1} into the past of $\{x_k\}$, and we thus adopt the following convenient notation:

$$\hat{x}_{k+1} = P\{x_{k+1} \mid x_k, x_{k-1}, \cdots\}. \tag{3}$$

When $\{x_k\}$ is gaussian, the projection coincides with the conditional expectation.

3.1 *The Error Process*

The error process $\{v_k\}$, defined by

$$v_{k+1} = x_{k+1} - \hat{x}_{k+1}, \tag{4}$$

is the innovations process of $\{x_k\}$. It has the key orthogonality properties:

$$Ev_{k+1}x_{k-i} = 0, \qquad i = 0, 1, 2, \cdots; \tag{5}$$

$$Ev_{k+1}v_{k-i} = 0, \qquad i = 0, 1, 2, \cdots. \tag{6}$$

Equation (5), which characterizes the projection operation, indicates that the best linear predictor cannot make better use of the past of $\{x_k\}$. Equation (6), a direct consequence of equation (5), shows that the error process is white noise.

3.2 *Stability*

The optimal unconstrained prediction, $\hat{x}_{k+1}$, may be characterized as the limit of an infinite sequence of constrained-order nonrecursive predictions:

$$\hat{x}_{k+1} = \lim_{n\to\infty} \sum_{i=0}^{n-1} h_{in}x_{k-i} \tag{7}$$

where $h_{in}(i = 0, 1, \cdots, n - 1)$ are the coefficients of the optimum nth-order nonrecursive predictor which may be calculated by means of well-known quadratic minimization techniques. The unconstrained predictor is a stable function of the data in the sense that

$$\lim_{n\to\infty} \sum_{i=0}^{n-1} h_{in}^2 < \infty . \tag{8}$$

This is proved in Section IV.

3.3 *Process Representation*

We say $\{x_k\}$ is of nth-order if it can be represented as the response of a stable recursive nth-order filter to white noise so that

$$\sum_{i=0}^{n} \alpha_i x_{k-i} = \sum_{i=0}^{n} \beta_i \epsilon_{k-i} \tag{9}$$

in which α_n or β_n is nonzero, $\{\epsilon_k\}$ is a white-noise process, and $\Sigma\alpha_i z^i$ has no zeros in $|z| \leqq 1$. If $\{x_k\}$ is of order n, it is known that there exists an nth-order recursive filter which generates $\{\hat{x}_k\}$ in response to $\{x_k\}$. The error process of this filter is $\{v_k\}$, the innovations process of $\{x_k\}$. If $\Sigma\beta_i z^i \neq 0$ for $|z| \leqq 1$, then $v_k = \epsilon_k$.

Conversely, if $\{x_k\}$ does not possess an nth-order representation of the form of equation (9), the best unconstrained predictor cannot be realized by an nth-order filter. To prove this we assume that such a realization does exist. That is, we assume

$$\hat{x}_{k+1} = \sum_{i=0}^{n-1} d_i x_{k-i} + \sum_{i=1}^{n} c_i \hat{x}_{k+1-i}. \tag{10}$$

This combined with equation (4) implies

$$x_{k+1} - \sum_{i=0}^{n-1} (d_i + c_{i+1}) x_{k-i} = v_{k+1} + \sum_{i=0}^{n-1} c_{i+1} v_{k-i} \tag{11}$$

which shows that $\{x_k\}$ is in fact the response of an nth-order filter to the white-noise process $\{v_k\}$, which is a contradiction.

IV. PROJECTING FILTERS

4.1 *Orthogonality Properties*

We have shown that an nth-order recursive filter cannot perform optimal unconstrained linear prediction of a process of order greater than n. With such a process as input, the error process $\{e_k\}$, of an nth-order filter will necessarily have a higher mean-square value than that of the innovations process and $\{e_k\}$ will fail to meet the orthogonality conditions of equations (5) and (6). However, when the nth-order predictor possesses the projecting property defined in Section II, its error process satisfies some but not all of the orthogonality conditions met by innovations process. In particular, the projecting property requires that

$$y_k = P\{x_{k+1} \mid x_k , x_{k-1} , \cdots , x_{k-n+1} , y_{k-1} , \cdots , y_{k-n}\} \tag{12}$$

which is characterized by the orthogonality conditions

$$Ee_{k+1}x_{k-i} = 0, \qquad i = 0, 1, \cdots , n - 1; \tag{13}$$

$$Ee_{k+1}e_{k-i} = 0, \qquad i = 0, 1, \cdots , n - 1. \tag{14}$$

Note that in this case, equation (14) is not a direct consequence of equation (13). In fact equation (13) is satisfied by the error of the optimum nth-order nonrecursive filter, while equation (14) is not satisfied by this error unless $\{x_k\}$ is an nth-order autoregression, that is, an nth-order process with $\beta_i = 0$ for $i > 0$.

4.2 *Stability*

Projecting filters are inherently stable. In fact, some kind of stability property is implicit in any statement of steady-state properties of a time-invariant filter. In this paper we say that a filter is stable if its impulse response is square summable, which implies if the spectrum is rational, that the filter transfer function is analytic on and in the unit circle. We assume that the predicting filter has zero in each memory element prior to $k = 0$ at which time $\{x_k\}$ is applied to the input. The projecting property stated in Section II implies that in the limit as $k \to \infty$, y_k tends toward the projection indicated in equation (12). Thus in the limit, the orthogonality conditions of equations (13) and (14) are satisfied from which it follows that $Ee_{k+1}y_k \to 0$ and since $y_k + e_{k+1} = x_{k+1}$,

$$\lim_{k \to \infty} [Ey_k^2 + Ee_{k+1}^2] = Ex_{k+1}^2 = r_0$$

from which we infer

$$\limsup_{k \to \infty} Ey_k^2 < r_0. \tag{15}$$

We also know that the filter output for each $k \geqq 0$ is the finite sum

$$y_k = \sum_{i=0}^{k} g_i x_{k-i} \tag{16}$$

in which g_i is the filter impulse response. Equations (15) and (16) imply the existence of a positive number c, which bounds the mean-square output:

$$Ey_k^2 < c, \qquad \text{for all } k. \tag{17}$$

The existence of this bound leads to the following

Theorem: If a filter with impulse response g_i is a projecting filter, it is stable in the sense that

$$\sum_{i=0}^{\infty} g_i^2 < \infty. \tag{18}$$

Proof: In terms of $f(z)$, the power spectral density function of $\{x_k\}$, and the frequency transfer function of the filter we have

$$Ey_k^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{m=0}^{k} g_m e^{j\omega m} \right|^2 f(e^{j\omega})\, d\omega \geqq \lambda \sum_{m=0}^{k} g_m^2, \tag{19}$$

in which $\lambda = \min_{|z|=1} f(z) > 0$ according to the assumption stated in Section II. Equations (17) and (19) may be combined in the expression

$$\sum_{m=0}^{k} g_m^2 < c/\lambda, \quad \text{for all } k, \tag{20}$$

from which equation (18) follows.

The same reasoning leads to a proof of the stability of the unconstrained predictor. Replacing g_i is h_{in}, the impulse response of the nonrecursive predictor described in Section 3.3.

V. PROJECTING-FILTER DESIGN APPROACH

As we stated in Section II, an attempt to determine the filter coefficients by directly combining equation (1) and equations (13) and (14) leads to an intractable set of transcendental equations relating the coefficients and the autocovariance function of $\{x_k\}$. On the other hand, the iterative approach introduced in this paper leads to the computation of the desired coefficients by means of standard operations of arithmetic and matrix algebra.

Our design method results in a time-varying filter which, starting with zero in all memory elements, sequentially predicts x_1, x_2, $\cdots$ according to the projecting principle. At each step the filter forms the optimum linear combination of the available data.

Thus we define the process $\{x'_k\}$ such that

$$\begin{aligned} x'_k &= 0, \qquad k < 0; \\ x'_k &= x_k, \qquad k \geqq 0; \end{aligned} \tag{21}$$

and we adopt as our prediction of x_{k+1},

$$\begin{aligned} y_k &= 0, && k < 0; \\ y_k &= P\{x_{k+1} \mid x'_k, x'_{k-1}, \cdots, x'_{k-n+1}, y_{k-1}, \cdots, y_{k-n}\}, && k \geqq 0. \end{aligned} \tag{22}$$

Equation (22) uniquely defines the time-varying linear transformation which generates the nonstationary process $\{y_k\}$ from the stationary process $\{x_k\}$.

At each step the prediction error of the time-varying filter meets the orthogonality conditions of equations (13) and (14) so that $Ee_{k+1}y_k = 0$ and therefore $Ey_k^2 < r_0$ for all k. Following the proof of the theorem in Section 4.2 we can show that with the filter output represented by

$$y_k = \sum_{i=0}^{k} g_{ik} x_{k-i}, \qquad k \geqq 0, \tag{23}$$

the time-varying filter possesses the stability property

$$\limsup_{k\to\infty} \sum_{i=0}^{k} g_{ik}^2 < \infty. \tag{24}$$

Furthermore, if this filter approaches the time-invariant projecting filter with impulse response g_i in the sense that

$$\lim_{k\to\infty} \sum_{i=0}^{k} (g_{ik} - g_i)^2 = 0, \tag{25}$$

we are assured that this filter is stable and that it has the desired nth-order recursive structure. Hence if we determine, for each k, a_{ik} and b_{ik} such that

$$y_k = \sum_{i=0}^{n-1} a_{ik} x'_{k-i} + \sum_{i=1}^{n} b_{ik} y_{k-i} \tag{26}$$

is equivalent to equation (22), then successive computation of these coefficients leads to the desired time-invariant projecting filter.

Note that although y_k is uniquely determined by equation (22), the coefficients a_{ik} and b_{ik} in the representation of equation (26) are not unique when the set of stored data is linearly dependent. This situation is analyzed in Section 7.4.

VI. MATCHED-ORDER PROCESSES

We prove in Section 6.1 that when $\{x_k\}$ is of order n, the projecting-filter design technique results in least mean-square time-varying prediction in the sense that each output y_k is the optimum linear combination of the entire observed past of $\{x_k\}$. Thus y_k is equal to the output of the optimal nonrecursive filter of order $k + 1$ as described in Section 3.2 so that

$$\lim_{k\to\infty} (y_k - \hat{x}_{k+1})^2 = 0,$$

indicating that the design algorithm converges to the optimal unconstrained predictor. In Section 6.2, we derive simple formulas for the filter coefficients generated by the design procedure.

6.1 *Optimality*

We denote by $\mathcal{H}_k$ the subspace spanned by the random variables in the filter memory at time k: $x'_k, x'_{k-1}, \cdots, x'_{k-n+1}, y_{k-1}, \cdots, y_{k-n}$; and we denote by $\mathcal{R}_k$ the subspace spanned by the observed past of $\{x_k\}$: $x_k, x_{k-1}, \cdots, x_0$. Note that another spanning set of $\mathcal{R}_k$ is e_k, $e_{k-1}, \cdots, e_1, x_0$, where $\{e_k\}$ is the error sequence of the projecting filter. This statement follows by induction since x_0 spans $\mathcal{R}_0$ and if $\{e_j, e_{j-1}, \cdots, e_1, x_0\}$ spans $\mathcal{R}_j$ then $\{e_{j+1}, e_j, \cdots, e_1, x_0\}$ spans $\mathcal{R}_{j+1}$ because $e_{j+1} = x_{j+1} - y_j$ with y_j in $\mathcal{R}_j$.

In this section we assume that $\{x_k\}$ is an nth-order process represented by equation (9) with $\alpha_0 = 1$ so that

$$x_{j+1} = u_{j+1} - \sum_{i=1}^{n} \alpha_i x_{j+1-i} \tag{27}$$

in which $\{u_k\}$ is the moving average process with

$$u_{j+1} = \sum_{i=0}^{n} \beta_i \epsilon_{j+1-i} \tag{28}$$

and $\{\epsilon_k\}$ is a unit-mean-square white-noise process. Because $\Sigma\alpha_i z^i \neq 0$ for $|z| \leqq 1$, equation (27) may be expressed in the form

$$x_{k+1} = \sum_{i=0}^{\infty} h_i \epsilon_{k+1-i}, \tag{29}$$

in which $\{h_i\}$ is square summable. Equation (29) shows that

$$E\epsilon_{k+1}x_{k-i} = 0, \qquad i \geqq 0, \tag{30}$$

and equations (28) and (30) imply

$$Eu_{k+1}x_{k-i} = 0, \qquad i \geqq n. \tag{31}$$

If we let x^*_{k+1} denote the optimal "growing-memory" prediction of x_{k+1} with the projection characteristic

$$x^*_{k+1} = P\{x_{k+1} \mid \mathcal{R}_k\},$$

we have the following

Theorem: At each instant k, the time-varying filter output defined by

$$y_k = P\{x_{k+1} \mid \mathcal{H}_k\}$$

is the optimal growing-memory predictor in the sense that

$$y_k = x^*_{k+1} . \tag{32}$$

Proof: We will show that $x^*_{k+1} \varepsilon \mathcal{H}_k$ which implies equation (32) because $\mathcal{H}_k \subset \mathcal{R}_k$. Clearly, for $0 \leqq k < n$, $\mathcal{H}_k = \mathcal{R}_k$ so that $y_k = x^*_{k+1}$. We assume $y_k = x^*_{k+1}$ for all $k < j$ and show that this implies $y_j = x^*_{j+1}$. Hence, by induction, equation (32) is valid for all k.

Let $j \geqq n$ and assume equation (32) holds for all $k < j$. Then

$$Ee_{k+1}x_{k-i} = 0, \qquad \text{for} \quad k = 0, 1, \cdots, j-1;$$
$$i = 0, 1, \cdots, k. \tag{33}$$

This implies that the vectors $e_j, e_{j-1}, \cdots, e_1, x_0$, which span $\mathcal{R}_j$ are mutually orthogonal. Thus a projection into $\mathcal{R}_j$ is the sum of the projections into each of these basis vectors. In particular

$$P\{u_{j+1} \mid \mathcal{R}_j\} = P\{u_{j+1} \mid x_0\} + \sum_{i=0}^{j-1} P\{u_{j+1} \mid e_{j-i}\}. \tag{34}$$

Now note that $e_{j-i} \varepsilon \mathcal{R}_{j-i}$ and that equation (31) states that $u_{j+1} \perp \mathcal{R}_{j-i}$ for $i \geqq n$. Thus the first term in equation (34) and all but the first n terms of the summation are zero so that

$$P\{u_{j+1} \mid \mathcal{R}_j\} = \sum_{i=0}^{n-1} P\{u_{j+1} \mid e_{j-i}\}. \tag{35}$$

We now consider x^*_{j+1} by noting that the projection operator is linear and that $P\{x_{k-i} \mid \mathcal{R}_k\} = x_{k-i}$ for $i = 0, 1, \cdots, k$. Thus equation (27) implies

$$x^*_{j+1} = P\{x_{j+1} \mid \mathcal{R}_j\} = P\{u_{j+1} \mid \mathcal{R}_j\} - \sum_{i=1}^{n} \alpha_i x_{j+1-i} \tag{36}$$

or, from equation (35)

$$x^*_{j+1} = \sum_{i=0}^{n-1} P\{u_{j+1} \mid e_{j-i}\} - \sum_{i=1}^{n} \alpha_i x_{j+1-i}. \tag{37}$$

Note that the ith term in the first summation is proportional to e_{j-i}

so that x^*_{j+1} is a linear combination of x_j, x_{j-1}, $\cdots$, x_{j-n+1}, y_{j-1}, $\cdots$, y_{j-n}, the basis vectors of $\mathcal{H}_j$. Thus $x^*_{j+1} \varepsilon \mathcal{H}_j$ and

$$x^*_{j+1} = P\{x_{j+1} \mid \mathcal{H}_j\} = y_j .$$

Hence $x^*_{k+1} = y_k$ for all k. Q.E.D.

6.2 *Filter Coefficients*

In this section we derive explicit recursions for the coefficients and mean-square error of the optimal growing-memory predictor of a stationary nth-order process. We begin with equation (37) for the optimal prediction and observe that the projections have the form

$$P\{u_{k+1} \mid e_{k-i}\} = \gamma_{ik} e_{k-i} , \qquad i = 0, 1, \cdots, n-1, \tag{38}$$

where the coefficients are ratios of two expectations,

$$\gamma_{ik} = Eu_{k+1}e_{k-i}/Ee^2_{k-i} . \tag{39}$$

These expectations may be expressed as functions of the auto-covariance coefficients,

$$\varphi_i = Eu_k u_{k-i} , \tag{40}$$

of the stationary moving average process $\{u_k\}$.

Our derivation begins with the expression of the error at step k, $e_{k+1} = x_{k+1} - x^*_{k+1}$, as the difference between equation (27) for x_{k+1} and equation (37) for x^*_{k+1}:

$$e_{k+1} = u_{k+1} - \sum_{i=0}^{n-1} \gamma_{ik} e_{k-i}. \tag{41}$$

Squaring equation (41) and taking the expectation we obtain

$$Ee^2_{k+1} = \varphi_0 - \sum_{i=0}^{n-1} \gamma^2_{ik} Ee^2_{k-i} \tag{42}$$

which gives the mean-square error at step k in terms of current filter coefficients and past errors. To find the next set of coefficients, $\gamma_{i,k+1}$, we express e_{k+1-i} as in equation (41) and we find the expected product of this random variable and u_{k+2}. Then we divide by the mean-square indicated in equation (39) with the result

$$\begin{aligned}\gamma_{n-1,k+1} &= \varphi_n / Ee^2_{k+2-n} , \\ \gamma_{i,k+1} &= \left[\varphi_{i+1} - \sum_{j=0}^{n-i-2} \gamma_{j,k-i} \gamma_{i+j+1,k+1} Ee^2_{k-i-j} \right] \bigg/ Ee^2_{k+1-i} , \\ & \qquad i = n-2, n-3, \cdots, 0,\end{aligned} \tag{43}$$

where the upper limit on the sum is a consequence of the property, $Eu_{k+2}\,e_{k-i-j} = 0$ for $j \geqq n - i - 1$. [See equation (31)].

The filter coefficients a_{ik} and b_{ik} of equation (26) are related to the projection coefficients γ_{ik} and the autoregressive coefficients, α_i, of the process representation by

$$\begin{aligned} a_{ik} &= \gamma_{ik} - \alpha_{i+1} , \\ b_{ik} &= -\gamma_{i-1,k} , \end{aligned} \tag{44}$$

because equations (37) and (38) combine to form

$$x^*_{k+1} = \sum_{i=0}^{n-1} (\gamma_{ik} - \alpha_{i+1})x_{k-i} - \sum_{i=1}^{n} \gamma_{i-1,k}x^*_{k+1-i} . \tag{45}$$

Our recursive technique for finding the characteristics of the optimal nth-order growing memory predictor thus consists of alternately performing the calculations of equations (42) and (43) and of obtaining the filter coefficients at each step by means of equation (45).

6.3 *Convergence of Filter Coefficients*

Since the time-varying filter output y_k converges to the optimal unconstrained predictor $\hat{x}_{k+1}$, one would expect that the time-varying coefficients a_{ik} and b_{ik} will converge to constant coefficients a_i and b_i. Since we have excluded processes with zeros on the unit circle, an nth-order recursive structure for the optimal predictor is known to exist.[1] But this is not sufficient. It is also necessary to exclude the possibility that the intrinsic order of the process is less than n. Then the coefficients of the nth-order recursive equation for the optimal predictor are unique and the time-varying coefficients a_{ik} and b_{ik} will in fact converge to these constant coefficients.

6.4 *Relation to Kalman Filtering*

In addition to proving convergence of our design approach, we have shown for the matched order case that the time-varying filter generated by the design procedure is the optimal growing-memory predictor. At each instant, k, the $2n$ stored data samples contain all the needed information about the observed past of the process, $x_0 , x_1 , \cdots , x_k$. It follows that the time-varying filter must be identical to the Kalman predictor[3] which is obtained by expressing the process model in state equation form. However, the Kalman development is computationally less efficient as may be seen by comparing the Ricatti equations with the simpler recursions given in Section 6.2.

In recent months recursions similar to ours have been published in various contexts. They appear in a paper by J. Rissanen and L. Barbosa[9] as steps in the factorization of the covariance matrix of $\{u_k\}$, the nth-order moving average, and Kailath[10] has indicated that such recursions follow from an innovations approach to prediction. Related formulas also appear in R. L. Kashyap's[11] derivation of predictor characteristics in terms of the parameters α_i and β_i of the process representation. In our derivation, as in Refs. 9 and 10, the basic data are the set of α_i and the autocovariance function of $\{u_k\}$. In contrast, the new design algorithm presented in Section VII uses only the covariances of the process to be predicted, quantities that are often more accessible in practice than the process parameters.

VII. SYNTHESIS TECHNIQUE

In this section we apply the projecting-filter design approach of Section V to obtain a computational algorithm for the general case in which the order of the process may differ from the order of the filter. The basic idea of the approach is to compute successive sets of weighting coefficients for an nth-order *time-varying* recursive filter which asymptotically approaches the desired *time-invariant* projecting filter.

As discussed in Section V, the time-varying projecting filter of interest is characterized by the input-output relationship

$$y_k = P\{x_{k+1} \mid \mathcal{H}_k\} \tag{46}$$

where $\mathcal{H}_k$ denotes the subspace spanned by the $2n$ variates

$$x'_k , x'_{k-1} , \cdots , x'_{k-n+1} , y_{k-1} , y_{k-2} , \cdots , y_{k-n} .$$

Equation (46) uniquely defines y_k as the projection of x_{k+1} into $\mathcal{H}_k$. This projection can be expressed explicitly as a linear combination of the $2n$ variates; that is,

$$y_k = \sum_{i=0}^{n-1} a_{ik} x'_{k-i} + \sum_{i=1}^{n} b_{ik} y_{k-i} . \tag{47}$$

Let $d(\mathcal{H}_k)$ denote the dimension of the subspace $\mathcal{H}_k$, i.e., $d(\mathcal{H}_k)$ is the minimum number of variates needed to span $\mathcal{H}_k$. If $d(\mathcal{H}_k) = 2n$ then the $2n$ spanning variates are linearly independent and the coefficient set used in equation (47) is unique. On the other hand if $d(\mathcal{H}_k) < 2n$, the $2n$ spanning variates are linearly dependent and consequently there is an infinite number of possible choices for the coefficient set. This situation always occurs in the first $2n - 1$ iterations ($0 \leqq k <$

$2n - 1$) and it may occur as well in subsequent iterations. To overcome this difficulty, we adopt a consistent procedure for selecting a linearly independent subset of the $2n$ spanning variates for each k. Variates are eliminated by setting appropriate coefficients to zero in equation (47). The remaining coefficients are then uniquely determined from the covariance matrix of the remaining variates and the cross-covariances between the remaining variates and x_{k+1}.

The algorithm is initialized with $y_0 = a_{00}x_0$ and all of the other coefficients a_{i0} and b_{i0} ($i \neq 0$) set to zero. Then each iteration consists of the following steps: (*i*) solving for the appropriate coefficient values, (*ii*) computing the needed covariances for the following iteration, (*iii*) determining an independent set of variates for the next prediction.

7.1 *Reduced Representation*

The procedure for eliminating dependent variates from the set of available data at time k leads to the following expression [equivalent to equation (47)] for the kth prediction

$$y_k = \sum_{i=0}^{p-1} a_{ik}x_{k-i} + \sum_{i=1}^{q} b_{ik}y_{k-i} \tag{48}$$

with $p \leqq n$, $q \leqq n$.* The coefficients that do not appear in equation (48) are all set to zero in the process of eliminating dependent variates; that is,

$$a_{ik} = 0, \qquad i = p, p+1, \cdots, n-1;$$
$$b_{ik} = 0, \qquad i = q+1, q+2, \cdots, n.$$

Note that x_{k-i} rather than x'_{k-i} appears in equation (48). This is so because $x'_{k-i} = 0$ for $i > k$ so that any set containing this variate is necessarily dependent. Hence, in the initial n steps, $p \leqq k - 1$. Section 7.4 presents the general method by which a set of independent variates is determined.

7.2 *The Filter Equations*

With the prediction error defined as $e_{k+1} = x_{k+1} - y_k$, the projecting property implies the following orthogonality conditions

$$\begin{aligned} Ee_{k+1}x_{k-i} &= 0, \qquad i = 0, 1, \cdots, p-1; \\ Ee_{k+1}y_{k-i} &= 0, \qquad i = 0, 1, \cdots, q. \end{aligned} \tag{49}$$

* Note that p and q depend on k. They will be denoted $p(k)$ and $q(k)$ when ambiguity might otherwise arise.

By substituting equation (48) for y_k into equation (49), we obtain the following set of $d(\mathcal{H}_k)$ linear equations in the $d(\mathcal{H}_k)$ coefficients:

$$r_{j+1} = \sum_{i=0}^{p-1} a_{ik} r_{i-j} + \sum_{i=1}^{q} b_{ik} w(k-j, k-i), \qquad j = 0, 1, \cdots, p-1;$$

$$w(k+1, k-j) = \sum_{i=0}^{p-1} a_{ik} w(k-i, k-j) + \sum_{i=1}^{q} b_{ik} v(k-i, k-j), \qquad j = 1, 2, \cdots, q; \tag{50}$$

in which we have adopted the notation:

$$\begin{aligned} r_i &= Ex_k x_{k-i} = r_{-i}\ , \\ w(k, j) &= Ex_k y_j\ , \\ v(k, j) &= Ey_k y_j\ . \end{aligned}$$

The function r_i comprises the given statistical information ot the prediction problem and w and v must be expressed as functions of r_i and previously computed filter coefficients.

Equations (50) have the following partitioned matrix form

$$\begin{bmatrix} T_p & X_k \\ X_k' & V_k \end{bmatrix} \begin{bmatrix} A_k \\ B_k \end{bmatrix} = \begin{bmatrix} R_p \\ W_k \end{bmatrix} \tag{51}$$

with

T_p the $p \times p$ autocovariance matrix of $\{x_k\}$,
X_k the $p \times q$ cross-covariance matrix of $\{x_k\}$ and $\{y_k\}$,
V_k the $p \times p$ autocovariance matrix of $\{y_k\}$,
$A_k = [a_{0k}\ , a_{1k}\ , \cdots , a_{p-1,k}]'$,
$B_k = [b_{1k}\ , b_{2k}\ , \cdots , b_{qk}]'$,
$R_p = [r_1\ , r_2\ , \cdots , r_p]'$,
$W_k = [w(k+1, k-1), \cdots , w(k+1, k-q)]'$.

Note that T_p and R_p depend only on the given autocovariance function r_i and on p, the number of forward coefficients to be computed. They are independent of previously computed coefficients.

If we perform the multiplication indicated in equation (51) and then solve for A_k and B_k we derive

$$\begin{aligned} B_k &= [V_k - X_k' U_p X_k]^{-1} [W_k - X_k' C_p], \\ A_k &= C_p - U_p X_k B_k\ , \end{aligned} \tag{52}$$

where $U_p = T_p^{-1}$ and $C_p = U_pR_p$, the column matrix of weights corresponding to the optimum pth-order nonrecursive predictor. By using efficient algorithms developed for the analysis of nonrecursive predictors,[12,13] one may successively calculate $U_0, U_1, \cdots, U_{n-1}, C_0, C_1, \cdots, C_{n-1}$ before the start of the synthesis procedure so that at the kth step, only a $q \times q$ matrix inversion [rather than one of order $(p + q)$] is required. We are assured that the matrix to be inverted is nonsingular because we have eliminated dependent variates by reducing the number of unknowns from $2n$ to $p + q$. Note that A_k consists of the coefficients of the optimum pth-order nonrecursive predictor modified by $U_pX_kB_k$ which indicates the effect of the feedback section of the predicting filter.

7.3 *Obtaining Successive Covariance Statistics*

The nature of $w(k, j)$ depends on which time index is the greater. If $j \geqq k$ we observe that the projection property of the jth estimate implies that $Ex_ke_j = 0$ for $k = j - 1, j - 2, \cdots, j - n$. Thus if we substitute $x_{j+1} - e_{j+1}$ for y_j in the definition of $w(k, j)$, we obtain

$$\begin{aligned} w(k, j) &= E[(x_{j+1} - e_{j+1})x_k], \\ &= r_{j+1-k}, \qquad j = k, k + 1, \cdots, k + n - 1. \end{aligned} \tag{53}$$

For $j < k$, we substitute equation (26) for y_j in the definition of $w(k, j)$, with the result

$$w(k, j) = \sum_{i=0}^{n-1} a_{ij}r_{j-k-i} + \sum_{i=1}^{n} b_{ij}w(k, j - i), \qquad j = 0, 1, \cdots, k - 1. \tag{54}$$

Equation (54) indicates that $\{w(k, 0), w(k, 1), \cdots, w(k, k - 1)\}$ is the sequence of filter outputs when $\{r_{-k}, r_{-k+1}, \cdots, r_{-1}\}$ is the sequence of inputs. This is an example of the property of linear filters that the cross-covariance between input and output is the correlation of the filter impulse response with the input autocovariance function. Using the initial conditions $w(k, j) = 0$ for $j < 0$, we may iteratively apply equation (54) in order to compute the required values of $w(k, j)$ for $j < k$.

The autocovariance coefficients of $\{y_k\}$ may be determined from the orthogonality conditions. With $k - n \leqq j \leqq k$, we have $Ee_{k+1}y_j = 0$ so that

$$v(k, j) = E[(x_{k+1} - e_{k+1})y_j] = w(k + 1, j), \qquad j = k - n, \cdots, k, \tag{55}$$

and of course $v(j, k) = v(k, j)$. Thus, equations (53), (54), and (55) express, in terms of known quantities, the covariance coefficients that appear in equation (51).

7.4 *The Number of Independent Variates*

In Section 7.2 we have assumed that p and q, the number of forward coefficients and the number of feedback coefficients to be computed at time k are determined in a manner that assures the linear independence of the $p + q$ variates that appear in equation (48) and therefore, the existence of the inverse matrix of equation (52). In many instances $p = q = n$ so that all of the data in the predictor memory are linearly independent. On the other hand, there are two conditions under which the data are dependent. The first is called an initialization condition and this arises in the course of every synthesis procedure because the predictor begins to operate at $k = 0$ with zero in all memory elements except one. The initialization condition obtains for the first $2n - 2$ iterations of the design procedure during which $d(\mathcal{H}_k) \leqq k + 1 < 2n$ because $\mathcal{H}_k \subset \mathcal{R}_k$ and $d(\mathcal{R}_k) = k + 1$. The other condition under which $d(\mathcal{H}_k) < 2n$ is called a reduced order condition, which arises when certain of the final feedback coefficients and/or final forward coefficients are zero. A reduced-order condition arises for all processes of order less than n.

7.4.1 *Initialization*

In this section we assume that no reduced order condition arises during the first $2n - 1$ steps of the predictor synthesis. This implies that $d(\mathcal{H}_k) = k + 1$ so that $p + q$, the number of coefficients determined by orthogonality conditions, increases by one at each iteration. At $k = 0$, the predictor estimates x_1 given x_0 which implies $p = 1$, $q = 0$. For increasing k, we alternately increase q and p by one so that for $0 \leqq k \leqq 2n - 2$

$$\begin{aligned} p &= 1 + \tfrac{1}{2}k, \qquad q = \tfrac{1}{2}k, && k \text{ even;} \\ p &= \tfrac{1}{2}(k + 1) = q, && k \text{ odd;} \end{aligned} \tag{56}$$

when no reduced order condition arises. Table I shows the variates that appear in equation (48) during the initial design stages of a second-order predictor.

7.4.2 *Reduced-Order Condition*

At time $k + 1$, the dependency of the data in storage can be deduced by observation of the coefficients computed at time k. In this

TABLE I—STEPS IN PREDICTOR DESIGN

Time	Predicted Variate	Independent Data				Projection
0	x_1	x_0				y_0
1	x_2	x_1		y_0		y_1
2	x_3	x_2	x_1	y_1		y_2
3	x_4	x_3	x_2	y_2	y_1	y_3
k	x_{k+1}	x_k	x_{k-1}	y_{k-1}	y_{k-2}	y_k

section we show how the values of certain coefficients, in particular whether or not they are zero, determine the relationship between $d(\mathcal{H}_k)$ and $d(\mathcal{H}_{k+1})$, the numbers of linearly independent variates in storage at time k and at time $k + 1$. In the next section we present the algorithm for determining the number of forward coefficients and the number of feedback coefficients to be computed at each step of the design.

The following theorem states that there is a dependence among the variates in storage at time $k + 1$ if and only if the coefficients determined at time k correspond to a filter of order less than n.

Theorem: With $d(\mathcal{H}_k) = 2n$, $d(\mathcal{H}_{k+1}) = 2n - 1$ *if and only if* $a_{n-1,k} = b_{n,k} = 0$. *Otherwise* $d(\mathcal{H}_k) = 2n$.

Proof: Assume $a_{n-1,k} = b_{n,k} = 0$. Then

$$y_k = \sum_{i=0}^{n-2} a_{ik}x_{k-i} + \sum_{i=1}^{n-1} b_{ik}y_{k-i}$$

which shows the linear dependency of the following variates in storage at time $k + 1$: $x_k, x_{k-1}, \cdots, x_{k-n+2}, y_k, \cdots, y_{k-n+1}$. Thus $d(\mathcal{H}_{k+1}) < 2n$. On the other hand, the $2n - 1$ variates: $x_{k+1}, x_k, \cdots, x_{k-n+2}, y_{k-1}, \cdots, y_{k-n}$ are linearly independent. All except x_{k+1} are independent because they are in storage at time k and $d(\mathcal{H}_k) = 2n$. In addition, the assumption that $\{x_k\}$ is nondeterministic implies that x_{k+1} cannot be expressed as a linear combination of the other stored variates because each of these is in $\mathcal{R}_k$. It follows that $d(\mathcal{H}_{k+1}) = 2n - 1$.

To prove the converse, assume $d(\mathcal{H}_{k+1}) = 2n - 1$. It follows that there exists a linearly dependent set of stored data. By the reasoning given above this set does not include x_{k+1} because all of the other stored variates are in $\mathcal{R}_k$. However the set does include y_k because all of the other variates are independent. Hence y_k can be represented as a linear combination of $x_k, x_{k-1}, \cdots, x_{k-n+2}, y_{k-1}, \cdots, y_{k-n+1}$. But the data in storage at time k also includes x_{k-n+1} and y_{k-n} and the fact that

$d(\mathcal{H}_k) = 2n$ implies that the representation of y_k is unique. Therefore we have the coefficients of x_{k-n+1} and y_{k-n}, $a_{n-1,k} = b_{n,k} = 0$. Q.E.D.

By reasoning similar to that used to prove this theorem we may establish the dimensionality of the data in storage at time $k + 1$ when $d(\mathcal{H}_k) < 2n$. Thus we have the following corollaries which apply for all k including the initial steps of the predictor design.

Corollary 1: With $d(\mathcal{H}_k) = p + q$ *and* $p = q < n$, $d(\mathcal{H}_{k+1}) = p + q - 1$ *if and only if* $a_{p-1,k} = b_{q,k} = 0$. *Otherwise* $d(\mathcal{H}_{k+1}) = p + q + 1$.
Corollary 2: With $d(\mathcal{H}_k) = p + q$ *and* $n \geqq p = q + 1$, $d(\mathcal{H}_{k+1}) = p + q$ *if and only if* $a_{p-1,k} = 0$. *Otherwise* $d(\mathcal{H}_{k+1}) = p + q + 1$.
Corollary 3: With $d(\mathcal{H}_k) = p + q$ *and* $p = q - 1 < n$, $d(\mathcal{H}_{k+1}) = p + q$ *if and only if* $b_{qk} = 0$. *Otherwise* $d(\mathcal{H}_{k+1}) = p + q + 1$.

7.4.3 *The Number of Computed Coefficients*

On the basis of the theorem and corollaries of Section 7.4.2, we establish the procedure shown in Table II for determining the numbers of forward and feedback coefficients $p(k + 1)$ and $q(k + 1)$ to be computed at time $k + 1$. The table indicates that $p(k + 1)$ and $q(k + 1)$ may be determined from $p = p(k)$ and $q = q(k)$ (shown in the left column) and from the final two feedback coefficients and the final

TABLE II—THE NUMBER OF COEFFICIENTS COMPUTED

Number of Coefficients Computed at Time k		Final Coefficients Computed at Time k				Number of Coefficients Computed at Time $k+1$	
		$b_{q,k}$	$b_{q-1,k}$	$a_{p-1,k}$	$a_{p-2,k}$	$p(k+1)$	$q(k+1)$
1	$p = q = n$	$\neq 0$				n	n
2				$\neq 0$		n	n
3	$p = q$	0	$\neq 0$	0		p	$q - 1$
4		0		0	$\neq 0$	$p - 1$	q
5	$p = q < n$	$\neq 0$				$p + 1$	q
6				$\neq 0$		p	$q + 1$
7	$p > q$			$\neq 0$		p	$q + 1$
8		$\neq 0$		0		p	q
9		0		0	$\neq 0$	$p - 1$	$q + 1$
10	$p < q$	$\neq 0$				$p + 1$	p
11		0		$\neq 0$		p	q
12		0	$\neq 0$	0		$p + 1$	$q - 1$
13	any p, q	0	0	0	0	irregular	

two forward coefficients (shown in the central four columns) computed at time k. If there is no entry for one of the coefficients, the indicated relationship between $p(k+1)$, $q(k+1)$ and p, q is independent of that coefficient. The other symbols indicate that a coefficient must necessarily be zero or nonzero for a relationship to be valid.

If, at time k, $p + q = d(\mathcal{H}_k)$, the variates $x_k, x_{k-1}, \cdots, x_{k-p+1}$, $y_{k-1}, \cdots, y_{k-q}$ are independent. This condition and the theorem and corollaries imply that the set $\{x_{k+1}, x_k, \cdots, x_{k-p(k+1)+2}, y_k, \cdots, y_{k-q(k+1)+1}\}$ is independent and spans $\mathcal{H}_{k+1}$. Thus lines 1 and 2 of Table II follow from the theorem; lines 3 through 6, from the theorem and Corollary 1; lines 7 through 9, from Corollary 2; and lines 10 through 12 from Corollary 3.

The table accounts for all possible combinations of computed coefficient values except those in which the last two forward coefficients and the last two feedback coefficients are all zero. This situation arises during the initial design stages whenever the input process is partially decorrelated. The manner in which independent variates are chosen for such a process is described in Section 7.4.5. When the irregularity arises in the design of predictors for other processes, there is no independent basis of $\mathcal{H}_{k+1}$ that is the union of consecutive members of $\{x_k\}$ beginning with x_{k+1} and consecutive members of $\{y_k\}$ beginning with y_k. Thus it is impossible to represent $P\{y_k \mid \mathcal{H}_k\}$ in the concise form of equation (48). Nor is it possible in general to determine at all times subsequent to k an independent set of stored data solely by considering p, q and the previously computed coefficients. All this serves to complicate quite substantially the representation of y_k, the equations which determine the coefficients, and the algorithm for determining the numbers of coefficients to be computed after the occurrence of the irregular condition indicated on the last line of Table II.

Rather than add substantially to the size of this paper by presenting a general technique for treating this situation, we simply note that except for partially decorrelated processes, it has never arisen in our experience of designing projecting filters and that in fact it appears to represent a pathological case. We have not discovered an example of a process for which four projection coefficients are simultaneously zero after one or more of their counterparts is nonzero at the previous time instant.

7.4.4 *Low Order Processes*

When $\{x_k\}$ is the response of an mth-order filter to white noise and m is no greater than n, the order of the predictor, the synthesis method leads to the mth-order form of the optimum unconstrained predictor. Section 6.1 contains a proof of this statement for $m = n$ and in this section we show that if $m < n$, a reduced-order situation arises and the effective order of the predictor does not grow beyond n.

Let a_{ik} and b_{ik} be the coefficients of the optimal growing-memory mth-order predictor, determined in the manner indicated in Section 6.2. Thus

$$x_{k+1}^* = \sum_{i=0}^{m-1} a_{ik}x_{k-i} + \sum_{i=1}^{m} b_{ik}x_{k+1-i}^* . \tag{57}$$

Note that for all $k \leqq 2m - 1$, y_k, the output of the nth-order predictor is identical to x_{k+1}^* because the design proceeds as for a predictor of order m.

Equation (56) indicates that at step $2m$ the initialization procedure leads to $p = m + 1$, $q = m$ and

$$y_{2m} = \sum_{i=0}^{m} a'_{i,2m}x_{2m-i} + \sum_{i=1}^{m} b'_{i,2m}y_{2m-i} \tag{58}$$

where $a'_{i,2m}$ and $b'_{i,2m}$ are determined uniquely by the orthogonality conditions. Hence it follows from the optimality of equation (57) that $a'_{m,2m} = 0$ and that the other coefficients are equal to the ones in equation (57) with $k = 2m$. Line 8 of Table I indicates that $p(2m + 1) = m + 1$ and $q(2m + 1) = m$ and once again we have $a'_{m,2m+1} = 0$ and the other coefficients equal to those in equation (57) for the optimal mth-order predictor. It is clear that for all $k \geqq 2m$ this sequence is repeated with $p(k) = m + 1$, $q(k) = m$ and $a'_{m,k} = 0$. Hence the algorithm converges to the unique mth-order form of the unconstrained optimal predictor.

7.4.5 *Partially Decorrelated Input Process*

A partially decorrelated process is a nonwhite process for which every set of $j + 1$ $(j > 0)$ adjacent samples is uncorrelated. In other words, $\{x_k\}$ is partially decorrelated if for some $j > 0$, $r_1 = r_2 = \cdots = r_j = 0$ and $r_{j+1} \neq 0$. For example the error process of an nth-order projecting filter is partially decorrelated with $j = n$.

Note that with a partially decorrelated input, the initial j generating filter outputs (corresponding to optimal nonrecursive predictions) are zero. Thus

$$y_k = x_{k+1}^* = 0 = a_{ik} = b_{ik}, \text{ for } 0 \leqq k < j \text{ and all } i. \tag{59}$$

This is a reduced-order situation conforming to line 13 of Table II (if we assume $b_{0k} = 0$ and $a_{ik} = b_{ik} = 0$ for $i < 0$). For this irregular case we adopt the following initialization procedure as an alternative to equation (56).

(*i*) All coefficients are 0 for $k < j$.
(*ii*) $p(j) = j + 1$, $q(j) = 0$.
(*iii*) $p(k)$, $q(k)$ according to Table II for $k > j$.

VIII. CONCLUSIONS

This paper introduces the projecting-filter principle of constrained-order recursive prediction and presents one technique of projecting filter synthesis. This technique has led to the design of the predictors described in Section 1.3 and to several other successful designs for a variety of random processes. However, the class of processes for which the technique is valid (that is, for which the algorithm converges to a time-invariant filter) and indeed the class for which a projecting filter of a given order exists have not as yet been determined. These questions are the subject of current research. Another important area of investigation involves the numerical aspect of the synthesis—the study of the sensitivity of this or any other design method to round-off in the calculation of coefficients.

Our studies to date indicate that the projecting filter is valuable in that it predicts many processes more accurately than other known devices of equal complexity. Our results are readily extended to vector-valued processes. Finally, we note that the projecting filter principle is applicable to a large class of estimation problems of which prediction one unit of time in the future is but a single example.

REFERENCES

1. Whittle, P., *Prediction and Regulation by Linear Least-Square Methods,* Princeton, New Jersey: D. Van Nostrand Company, Inc., 1963, pp. 31–35.
2. Whittle, P., "Recursive Relations for Predictors of Nonstationary Processes," J. Roy. Statist. Soc., Ser. B, *27,* No. 3 (1965), pp. 523–532.
3. Kalman, R. E., "A New Approach to Linear Filtering and Prediction Problems," Trans. ASME, J. Basic Engrg., Ser. D, *82,* No. 1 (March 1960), pp. 35–45.
4. Walsh, J. L., *Interpolation and Approximation by Rational Functions in the Complex Domain,* Amer. Math. Soc. Colloquium Publ. *XX,* Providence, 1960.
5. Phillips, R. S., "RMS-error Criterion in Servomechanism Design," Chap. 7 in H. M. James, N. B. Nichols, and R. S. Phillips, *Theory of Servomechanisms,* MIT Rad. Lab. Series, *25,* McGraw-Hill Book Co., New York 1947.
6. Kolmogorov, A. N., "Interpolation and Extrapolation of Stationary Random Sequences," Bull. Acad. Sci., USSR, Ser. Math., *5* (1941); transl.: Memorandum RM-3090-PR, Rand Corp., Santa Monica, California, April 1962.

7. Luenberger, D. G., *Optimization by Vector Space Methods,* New York: John Wiley and Sons, Inc., 1969, Chapter 4.
8. Doob, J. L., *Stochastic Processes,* New York: John Wiley and Sons, Inc., 1969, pp. 569–581.
9. Rissanen, J., and Barbosa, L., "Properties of Infinite Covariance Matrices and Stability of Optimum Predictors," Information Sciences, *1,* No. 2 (April 1969), pp. 221–236.
10. Kailath, T., "The Innovations Approach to Detection and Estimation Theory," Proc. IEEE, *58,* No. 5 (May 1970), pp. 680–695.
11. Kashyap, R. L., "A New Method of Recursive Estimation in Discrete Linear Systems," IEEE Trans. Aut. Cont., *AC-15,* No. 1 (February 1970), pp. 18–24.
12. Trench, W. F., "An Algorithm for the Inversion of Finite Toeplitz Matrices," J. Soc. Indust. Appl. Math., *12,* No. 3 (September 1964), pp. 515–522.
13. Trench, W. F., "Weighting Coefficients for the Prediction of Stationary Time Series from the Finite Past," J. Soc. Indust. Appl. Math., *15,* No. 6 (November 1967), pp. 1502–1510.

Contributors to This Issue

David S. Alles, B.S.M.E., 1962, Clarkson College, Potsdam, N. Y.; M.S., 1963, and Sc.D., 1968 (mechanical engineering), Massachusetts Institute of Technology; Bell Telephone Laboratories, 1968—. Mr. Alles has worked on the development of photolithographic equipment. Member, Tau Beta Pi, Pi Tau Sigma, Sigma Xi.

Jacques A. Arnaud, Dipl. Ing., 1953, Ecole Supérieure d'Electricité, Paris, France; Docteur Ing., 1963, University of Paris; Assistant at E.S.E., 1953–1955; C.S.F., Centre de Recherche de Corbeville, Orsay, France, 1955–1966; Warnecke Elec. Tubes, Des Plaines, Illinois, 1966–1967; Bell Telephone Laboratories, 1967—. At C.S.F., Mr. Arnaud was engaged in research on high-power travelling wave tubes and supervised a group working on noise generators. He is currently studying optical amplifiers and millimeter wave focusers. Senior Member, I.E.E.E.; Member, Optical Society of America.

F. R. Ashley, B.S., 1960, University of Arizona; MEE, 1962, New York University; Bell Telephone Laboratories, 1960—. Mr. Ashley is currently engaged in design of computer-controlled measurement systems and data acquisition systems. Member, IEEE, Tau Beta Pi, Phi Kappa Phi.

Franklin H. Blecher, B.E.E., 1949, M.E.E., 1950, and D.E.E., 1955, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1952—. Mr. Blecher's early work concerned the design of transistor circuits for application in analog and digital computers; design of wideband feedback amplifiers for application in carrier systems; and development of active filters, IF amplifiers, and wideband video amplifiers. He later headed a group in developing solid-state short-haul carrier circuits and millimeter wave networks. From 1961 to 1967, Mr. Blecher was Director of the Carrier Transmission Laboratory and was responsible for the development of short-haul and long-haul carrier systems using wire pair and coaxial cable transmission media. In May 1968 he was appointed Director of the Electron Device Laboratory. Fellow, IEEE; member, Tau Beta Pi, Eta Kappa Nu.

Mrs. Judith G. Brinsfield, B.S., technical writing, 1963, and B.S.E.E., 1964, Carnegie Mellon University; M.S., mathematics, 1967, Stevens Institute of Technology; Bell Telephone Laboratories, 1965—. Mrs. Brinsfield has been engaged in report writing on the Nike-X project, programming on machine aids projects and design and development of the Mask Shop Information System. She is presently supervisor of the Engineering Applications Group, working on a computer-based information system for a new integrated circuit mask-making facility and a computer system for the automatic generation of program flowcharts.

Barret Broyde, B.A. (magna cum laude), Yeshiva College, 1955; Ph.D. (Chemistry), Polytechnic Institute of Brooklyn, 1960; Western Electric Engineering Research Center, 1967–. Mr. Broyde was engaged originally in investigations on more sensitive electron beam recording media. He is now Research Leader of the Materials and Analysis and Characterization Organization where new methods, techniques and instruments are being developed. Member, American Chemical Society, The Chemical Society (London), The American Institute of Physics, IEEE, New York Academy of Science, AAAS.

Ta-Mu Chien, B.S.E.E., 1959, National Taiwan University; M.S.E.E., 1963, University of Kansas; Ph.D., Electrophysics, 1969, Polytechnic Institute of Brooklyn; Western Electric Company, 1962–1966; Bell Telephone Laboratories, 1968—. Since joining Bell Labs, Mr. Chien has been studying various problems related to PCM transmission systems. Member, IEEE.

M. J. Cowan, B.S., 1955, University of Maryland; Ph.D. (Physics), 1959, Duke University; National Science Foundation Postdoctoral Fellow, 1960; Assistant Professor, Duke University, 1961; Bell Telephone Laboratories, 1962—. Mr. Cowan did research in the field of submillimeter wavelength microwave spectroscopy prior to joining Bell Laboratories. Since 1962, he has worked on parametric amplifiers, piezoelectric devices and integrated circuits. He is presently involved with the evaluation of mask-shop processes. Member, Phi Kappa Phi, Phi Beta Kappa, Sigma Xi.

Patrick G. Dowd, B.A. (Math), 1957, St. Michael's College; Western Electric Company, 1957-1963; Bell Telephone Laboratories, 1963—. Mr. Dowd worked on the S.A.G.E. Air Defense System while at Western Electric. Since joining Bell Telephone Laboratories, he has been involved in computer software development for such things as automated transformer design, automated miniature-wire spring-relay design and graphics-terminal development as well as the primary pattern generator. He is presently concerned with the development of graphical time sharing terminals.

J. W. ELEK, B.S.M.E., 1957, Case Institute of Technology; M.S. (engineering mechanics), 1961, Lehigh University; Bell Telephone Laboratories, 1958—. Mr. Elek worked on shock and vibration problems as well as on stress analyis and semiconductor device processing. He is Supervisor of a mechanical engineering group in the Materials and Process Technology Laboratory responsible for work on photolithographic masks for semiconductor and thin-film devices, with a strong emphasis on computer control of mask-making equipment. Member, Tau Beta Pi, Engineering Club of the Lehigh Valley.

ALLEN GERSHO, B.S., 1960, Massachusetts Institute of Technology; M.S., 1961, and Ph.D., 1963, Cornell University; Bell Telephone Laboratories, 1963—. During the 1966–67 academic year, Mr. Gersho was Assistant Professor of Electrical Engineering at the City University of New York. He has performed research in time varying and nonlinear signal processing, synchronization, adaptive filtering and the statistical approach to digital filter design.

DAVID J. GOODMAN, B.E.E., 1960, Rensselaer Polytechnic Institute; M.E.E., 1962, New York University; Ph.D., 1967, University of London; Bell Telephone Laboratories, 1960–62, 1967—. A member of the Systems Theory Research Department, Mr. Goodman has studied principles of digital signal processing including analog-to-digital conversion and the statistical approach to digital filter design. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

ARTHUR G. GROSS, B.E.E., 1956, M.S., 1959, and Ph.D., 1964, Rensselaer Polytechnic Institute; Bell Telephone Laboratories, 1964—. Mr. Gross has been mainly concerned with the development of computer aids for integrated circuit design and artwork generation. He is presently supervisor of the Computer Graphics Applications Group in the Computer Graphics Development Department. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi, Association for Computing Machinery, SIAM, AAAS.

Donald R. Herriott, studied undergraduate physics at Duke University, optics at the University of Rochester and electrical engineering at Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1956—. Mr. Herriott has worked on the optical design of the flying spot store for E.S.S., photoelectric lens evaluation, the development of the helium-neon laser and interferometry with and applications of lasers. He is currently Head of the Optical Device Department and is responsible for the development of new optical devices and systems. Fellow and director, Optical Society of America.

E. Y. Ho, B.S.E.E., 1964, The National Taiwan University; Ph.D., 1969, University of Pennsylvania; Bell Telephone Laboratories, 1969—. Mr. Ho has been engaged in developing and analyzing automatic equalizers for data transmission systems. Member, IEEE.

Frank L. Howland, B.S. (civil engineering), 1950, Rutgers University; M.S. (structural engineering), 1952, and Ph.D. (structural engineering), 1955, University of Illinois; Bell Telephone Laboratories, 1955—. Mr. Howland worked on the mechanical design of traveling-wave and millimeter-wave tubes and the development of metal-ceramic seal techniques for electron devices. As Head of the Applied Mechanics Department, he is involved in silicon integrated circuits and the associated packaging concepts of the Materials and Process Technology Laboratory. Member, IEEE, AAAS, SESA, Sigma Xi, Tau Beta Pi.

A. M. Johnson, American Telephone and Telegraph Company, 1950–1952; Bell Telephone Laboratories, 1956—. At Bell Labs, Mr. Johnson first worked on the development of special purpose cathode ray tubes. He then was engaged in developing high-speed photomultiplier tubes and gas lasers. At present he is involved in mask-making development work in the Optical Device Department.

Robert E. Kerwin, B.S., 1954, Boston College; M.S., 1958, Massachusetts Institute of Technology; Ph.D., 1964, University of Pittsburgh; Mellon Institute, 1958–1964; Bell Telephone Laboratories, 1964—. Mr. Kerwin worked in the field of polymer science at Mellon Institute and studied the structure of water at the University of Pittsburgh. At Bell Laboratories he is a member of the Photochemical Materials and Processes group and has been concerned with semiconductor device processing, silicon gate technology, photolithography, and unconventional imaging techniques. Fellow, American Institute of Chemists; Member, Sigma Xi, American Chemical Society, Society of Photographic Scientists and Engineers.

G. J.-W. KOSSYK, Associate Degree (Mechanical Engineering), 1955, Erie County Technical Institute, Buffalo, New York; B. A. (Mathematics), 1964, Rutgers University; Bell Telephone Laboratories, 1955—. Mr. Kossyk's early work was in electron tube development. Later he participated in the mechanical design and testing of the *Telstar*® Satellite. He has more recently been involved in the design and construction of the primary pattern generator. Member, Institute of Environmental Sciences.

JOSEPH P. LAICO, M.E., 1933, Brooklyn Polytechnic Institute; Bell Telephone Laboratories, 1929–1970. Mr. Laico specialized in mechanical design and development work, including work on various electron tubes from early amplifiers to magnetrons and klystrons. He also worked on traveling-wave tubes for radar, coaxial cable, radio relay, defense systems, and the *Telstar*® communications satellite project. Before retirement, Mr. Laico was Supervisor of a mechanical design group in the electron device laboratory. He has been granted 23 patents on electron devices.

ELLEN B. MURPHY, B.A. (Mathematics), 1959, Marywood College; Bell Telephone Laboratories, 1961—. Miss Murphy has been engaged in numerical analysis work and computer-controlled systems operations. She is presently working on a computer-controlled wavefront-measuring program.

BENJAMIN E. NEVIS, B.S.M.E., 1955, M.S.M.E., 1962, and Ph.D.M.E., 1965, Lehigh University; Assistant Professor, Lehigh University, 1965–68; Bell Telephone Laboratories, 1968—. Mr. Nevis has worked principally in heat transfer and fluid mechanics. Member, AIAA, Sigma Xi, Tau Beta Pi, Pi Tau Sigma; Associate Member, ASME.

R. J. NIELSEN, University of Idaho and Fairleigh-Dickinson University; Bell Telephone Laboratories, 1941—. Mr. Nielsen was first engaged in design drafting automatic central office equipment. He has been concerned with the study of metal-ceramic sealing and other problems in the mechanical design of electron tubes. He has worked on the mechanical design of the solar power plant for the *Telstar®* communications satellite, microwave oscillator tubes and a high power travelling wave tube for a radar system.

JAMES F. OBERST, B.E.E., 1964, Manhattan College; M.S.(E.E.), 1966, and Ph.D.(E.E.), 1969, Polytechnic Institute of Brooklyn; Assistant Professor of Electrical Engineering, 1968–1969, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1969—. At Brooklyn Polytechnic Institute, Mr. Oberst worked on digital phase-shift keyed communication systems. Since joining Bell Laboratories, he has been concerned with various aspects of PCM transmission over cable. Member, IEEE, Eta Kappa Nu.

STEPHEN PARDEE, B.S.E.E. and M.S.E.E., 1952, California Institute of Technology; Bell Telephone Laboratories, 1952–1955, 1957–1960, 1967—. At Bell Labs, Mr. Pardee has worked on the development of military radio relay systems and the design of digital systems. Since 1967, he has been involved in the development of computer systems to assist in the design, documentation, and manufacture of electronic systems. He is presently Head of the Machine Aids Development Department.

PETER D. PARRY, A.B., 1963, Hamilton College; M.A., 1965, and Ph.D., 1968, Princeton University; Western Electric Engineering Research Center, Princeton, New Jersey, 1968—. Mr. Parry has worked on the electron beam pattern generator with special emphasis on magnetic shielding and radiation effects. Member, Sigma Xi, American Physical Society.

K. M. POOLE, B.A. (physics), 1948, and D. Phil., 1951, Clarendon Laboratory, Oxford; Bell Telephone Laboratories, 1953—. Mr. Poole worked initially in the fields of electron optics, physical electronics, and ferrite devices. Later he was responsible for programs in microwave and other electron tubes, microwave circuits and optics. He subsequently was responsible for coordinating the development of new mask-making systems. As Head of the Solid State Device Electronics Department, he is presently responsible for development of microwave circuits and for device and subsystem development programs using the single-wall, magnetic domain technology. Member, IEEE, Optical Society of America.

M. E. POULSEN, Bell Telephone Laboratories, 1939—. Mr. Poulsen was first involved in electron-tube development, primarily glass and glass-to-metal seal problems. He has worked on submarine cable tubes and devices, design of thermal controls for the Telstar® project, and high-powered radar traveling tube development. At present he is engaged in scanning devices for high-speed data transmission.

JEAN RAAMOT, B.S.E.E., 1957, and M.S.E.E., 1963, Columbia University; Western Electric Company, 1957—. Mr. Raamot worked initially on telephone test sets. Since 1964, he has been a member of the Research Staff at the Western Electric Engineering Research Center, Princeton, New Jersey. His work is presently in small computer applications and high-precision graphics. He has contributed to digital-to-analog conversion, integer arithmetic, and pattern generation. Member, IEEE.

LAWRENCE R. RABINER, S.B. and S.M., 1964, and Ph.D. (E.E.), 1967, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1962–1964, 1967—. Mr. Rabiner has worked on digital circuitry, military communications problems, and problems in binaural hearing. Since 1967, he has been engaged in research on speech communication, signal analysis, digital filtering, and techniques for waveform processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, IEEE, Acoustical Society of America.

CHARLES M. RADER, B.E.E., 1960, and M.E.E., 1961, Brooklyn Polytechnic Institute; Lincoln Laboratory, Massachusetts Institute of Technology, 1961—. Mr. Rader has worked in the areas of speech compression, system simulation, digital signal processing, optics and educational technology. He is coauthor of a book on modern techniques for signal processing. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, Acoustical Society of America, IEEE.

ERIC G. RAWSON, B.A., 1959; M.A., 1960, University of Saskatchewan; Ph.D., 1966, University of Toronto; Bell Telephone Laboratories, 1966—. Mr. Rawson has been concerned with exploratory studies in optics, including three-dimensional computer graphics display using vibrating varifocal mirrors, glass GRIN rods which utilize a graded refractive index for imaging and light guidance, and the computer-automated design of complex lenses. He is currently also involved with the design and alignment of high-performance photolithographic cameras. Member, Optical Society of America.

Stephen O. Rice, B.S., 1929, Oregon State College; Graduate Studies, 1929–30 and 1934–35, California Institute of Technology; D. Sc. (Honorary), 1961, Oregon State College; Bell Telephone Laboratories, 1930—. In his first years at Bell Labs, Mr. Rice was concerned with nonlinear circuit theory, especially methods of computing modulation products. Since 1935, he has served as a consultant on mathematical problems and in investigation of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He is Head of the Communications Analysis Research Department. In the spring term of 1958, he was a Gordon McKay Visiting Lecturer in applied physics at Harvard University. Fellow, I.R.E.

Gordon I. Robertson, B.Sc., 1963, and Ph.D., 1967, University College, London; Standard Telecommunications Laboratory, Essex, England, 1967–1969; Western Electric Engineering Research Center, 1969 —. Mr. Robertson has worked on electron-phonon interactions in semiconductors. He is currently working on computor controlled electron beams. Member, I.E.E.E.

L. Rongved, B.S., 1950, M.S., 1951, and Ph.D., 1954, theoretical mechanics, Columbia University; Bell Telephone Laboratories, 1956—. Mr. Rongved has worked on ballistic missile guidance, the Telstar® project, and Apollo project. Member, Industrial Professional Advisory Council at Pennsylvania State University.

Peter E. Rosenfeld, ScB (EE), 1957, Brown University; ScM (EE), 1959, Harvard University; Bell Telephone Laboratories, 1959—. From 1959 to 1966 Mr. Rosenfeld worked in the Transmission Measuring Systems Department on the automation of transmission test sets. In 1966, he joined the Computer Graphics Development Department and worked on the Graphic 2 hardware development. He was responsible for the design of the primary pattern generator computer interface as well as the Mask Shop Information System interfaces.
Mr. Rosenfeld is presently Supervisor of the Computer Graphics Design Group which is responsible for the development of the Graphics 101 terminal and a new STARE hard copy system.

WINSTON R. SAMAROO, B.S.E.E., 1959, McGill University, Montreal; M.S.E.E., 1961, and Ph.D., 1965, University of Ottawa; Western Electric Engineering Research Center, Princeton, New Jersey, 1965—. Mr. Samaroo has worked on the development of electron beam pattern generators and has supervised a group in that activity. He is at present Assistant Director responsible for electron beam, ion implantation and gallium phosphide activities. Member, IEEE.

HASSEL J. SAVARD, JR., Graduate, 1962, Union County Technical Institute; Bell Telephone Laboratories, 1962–1968, 1969—. Initially Mr. Savard worked on high-power microwave tube development. More recently he has been engaged in interfacing the small computer to laboratory test equipment. He is presently developing methods for converting video information into useful computer inputs.

RONALD W. SCHAFER, B.S. (E.E.), 1961, and M.S. (E.E.), 1962, University of Nebraska; Ph.D., 1968, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1968—. Mr. Schafer has been engaged in research on digital waveform processing techniques and speech communication. Member, Phi Eta Sigma, Eta Kappa Nu, Sigma Xi, IEEE, Acoustical Society of America.

W. A. SCHLEGEL, B.S. (agricultural engineering), 1950, Pennsylvania State University; M.S. (mechanical engineering), 1956, Lehigh University; Bell Telephone Laboratories, 1958—. Mr. Schlegel has worked on the mechanical design and stress analysis of transistor and electron tube components and development work on subcable amplifier tube, heat transfer analysis, and experimental verification for single component and composite transistor header developments. He is engaged in the development of computer-controlled photolithographic equipment.

JOHN G. SKINNER, H.N.C. (mechanics), 1948, H.N.C. (physics), 1950, Northampton Polytechnic, London; M.S. (physics), 1958, and Ph.D. (physics), 1962, Oregon State University; Bell Telephone Laboratories, 1961—. Mr. Skinner was engaged with solid-state lasers, electro-optic material studies and electro-optical deflection schemes. His later work involved Raman spectroscopy for the study of stimulated Raman scattering and lattice dynamics.

JOHN W. STAFFORD, B.S. (aeronautical engineering), 1954, Massachusetts Institute of Technology; M.S. (applied mechanics), 1959, and M.S. (electrical engineering), 1970, Brooklyn Polytechnic Institute; Bell Telephone Laboratories, 1961—. Mr. Stafford has worked on the mechanical design and testing of the *Telstar*® communications satellite, development studies of an orientation system for communications satellites, and the mechanical design of plated wire memory and the primary pattern generator and reduction cameras. He is supervisor of the mechanical design group engaged in developing high-speed computer-controlled bonding systems. Member, American Institute of Aeronautics and Astronautics.

C. E. STOUT, JR., B.S.M.E., 1952, Carnegie-Mellon University; Western Electric Company, 1958—. Mr. Stout is a Senior Engineer in the Machine Design Department. His design experience varies from semiautomatic machinery such as the switchboard lamp stem mount machine to the mechanical aspects of his current project, the step-and-repeat camera.

MRS. SUZANNE B. WATKINS, B.A., Economics, 1967, Douglas College; Western Electric Engineering Research Center, Princeton, New Jersey, 1967—. Mrs. Watkins is an Information Systems Designer in the Computer Department. Her work has involved programming in the areas of hydrostatic metal forming, optimum allocation of inspection efforts, and maskmaking. She is currently on a leave of absence and residing in England.

YU S. YEH, B.S.E.E., 1961, National Taiwan University; M.S.E.E., 1964, and Ph.D., 1966, University of California, Berkeley; Harvard University, 1967; Bell Telephone Laboratories, 1967—. Mr. Yeh is a member of the Radio Transmission Research Department and is doing research work concerning Mobile Radio communication.

ALFRED ZACHARIAS, B.E.E., 1953, Cooper Union School of Engineering; S.M., 1955, and E.E., 1959, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1959—. Mr. Zacharias has investigated the microwave noise behavior of electron beams for use in traveling wave tubes. He has also investigated the performance of semiconductor diodes for use in high-power microwave switches and phase shifters. Most recently he has contributed to the development of a computer-controlled artwork generator for a mask-making system. He is currently Supervisor of the Applied Optics Group. Member, Tau Beta Pi, Sigma Xi.

H. Zucker, Dipl. Ing., 1950, Technische Hochschule, Munich, Germany; M.S.E.E., 1954, Ph.D., 1959, Illinois Institute of Technology; Bell Telephone Laboratories, 1964—. Mr. Zucker has been concerned with satellite communication antennas, optical resonators and problems related to physical and geometrical optics. Member, IEEE, Eta Kappa Nu, Sigma Xi.