Spectral lines

The Technical Institute. Travel and conversation provide the belief that there is developing a considerable diversification in programs, objectives, and philosophy among the engineering schools of the United States. This diversity may not be a recent development, but I cannot believe it was true thirty-five years ago. Looking back at the sample of engineering graduates in my acquaintanceship at that time, I seem to recall a considerable uniformity in educational pattern and objectives—and the rather heterogeneous group of salesmen, designers, company presidents, and technicians that has resulted from that educational pattern.

Today I see that some schools emphasize graduate work, others have more usual B.S. programs, and we now have the Technical Institute as a third and important educational orientation. The last group of schools, still small in number, is graduating possibly one-tenth of our real manpower needs for technicians. They are recognized as serving a very important purpose, and one which frees many of our four-year schools to do the job of preparing the future advanced planner and innovator.

Engineering Technical Institutes are offering two-year objective programs pointed to the preparation of Engineering Associates or Assistants. These men are often the hands and legs of the engineer; they build the models, they make the measurements and perform the tests, they make the routine calculation, and they install and operate the equipment. They can and do relieve the creative design engineer of much that is routine or repetitious—they free the more theoretically trained engineering graduate to do the thinking, planning, developing, and originating that will contribute to the new products of next year and the years after that.

A Technical Institute program, when properly publicized, can attract young men who wish to do, rather than to plan; to put their hands on equipment, rather than their minds on analysis of performance. It has been usual in the past to fill many technical jobs by men who had completed the first two years of an engineering degree program, and had then dropped out for one reason or another. It seems logical, and experience proves, that specific and terminal two-year programs can be a more fitting preparation for such work than half of a program designed for quite another purpose.

This seems true in spite of the fact that Technical Institute curricula and courses cover much of the same material taught in the first two years at the four-year schools. The difference is to be found in an objectivity in teaching not possible in the broader preparation necessary in the four-year professional schools.

Mathematics is taught more through application than as a series of exercises in proof and rigor; the design of an amplifier becomes a problem in building to specifications and proving that it works, rather than an exercise in proper location of the poles and zeros of a transfer function. At the same time, it is desirable and practicable for the Engineering Associate to understand the objectives of the engineer in such sophisticated design methods, although he may lack the higher skills utilized in preparing that design.

The leading Technical Institutes have strong programs, and their graduates are capable of carrying on many jobs previously assigned to engineers; this is recognized by industry in a pay scale starting only a little below that for some four-year B.S. graduates. Engineering Associates may be found especially valuable in some of the nonelectrical fields, where the theoretical and basic training in materials, mechanics, and fluids received by the four-year man is not immediately applicable in industry job assignments. However, the Technical Institutes seem to have received relatively more encouragement from our electrical industries than from some other areas, and as a result the number of nonelectrical technology programs is not as large as it might be; particularly does this seem true in the construction and building fields. Among ECPD-accredited Technical Institutes (and ECPD carries on this job in addition to its functions for four-year schools), one finds 39 programs in electrical and electronics technology, 31 programs in mechanical and related technology, and 18 programs in construction, surveying, and the like.

Education is needed among prospective students, since the diversity of objectives requires that they select their schools on the basis of curricular objectives which will lead them into the desired area of functional opportunity after graduation.

Education is needed among employers as well. They must be shown that a two-year technology graduate may often fit the job specification which they have written for an engineer—and they might hope that the Technical Institute graduate will be happier with the work than the B.S. man who may be dissatisfied and become a turnover statistic.

Both paths in education can lead into technical and challenging work; both paths may not be suited to the same individual; both paths are dignified and socially acceptable. Happiness in an appropriate job seems of primary importance in a man's success. Is not a man likely to be most happy in a job which fits his abilities and interests?

J. D. Ryder

Electroluminescence

Research in the field of electroluminescence is resulting in increasingly efficient light sources specifically designed for area illumination. Of particular interest at this time is the growing use of electroluminescent display devices

Lester W. Strock, Irving Greenberg *Sylvania Electric Products Inc.*

Principles and applications

Lester W. Strock

Light sources have evolved from the various types of torches and oil lamps of earliest times to the more familiar candles, kerosene lamps, and gas jets of recent memory. The era of modern lighting began well under 100 years ago as a result of the persistent efforts of Thomas A. Edison[1] to replace gas lighting by electricity "to improve the illumination to such an extent as to meet all requirements of natural, artificial and commercial conditions."

For over half a century Edison's thin-filament incandescent lamps lighted the homes, offices, and shops of a rapidly growing and complex society. The incandescent lamp, for practical purposes, was a point source. For 2½ generations people the world over were accustomed to, and satisfied with, a concentrated source of light, although it was later modulated by an infinite variety of shades and diffusers.

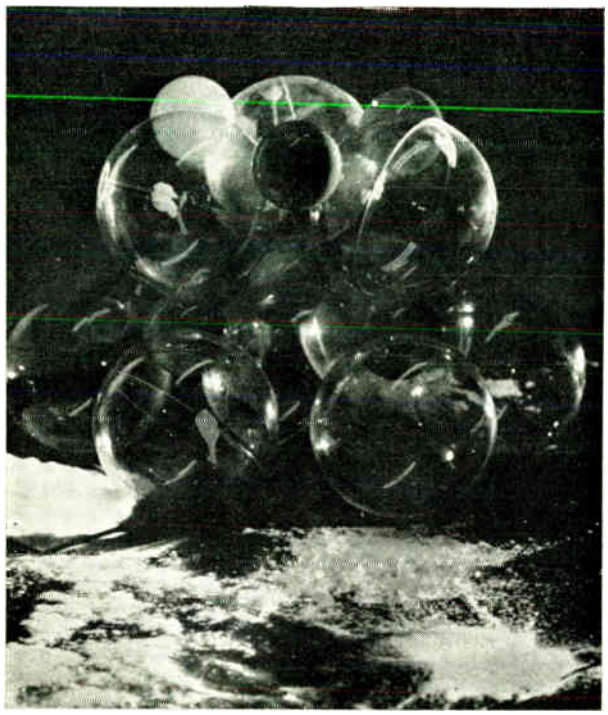
Then in the late 1930s came the commercial offering of an innovation in light sources—the fluorescent lamp. This had two essential new features: (1) The light was not derived from a hot filament but was generated by a "phosphor" when acted upon by ultraviolet radiation produced in a gas discharge within an elongated glass tube, carrying the phosphor on its inner wall. (2) This new lamp could best be made in form of long tubes to yield a line source. However, the ability of many materials to convert short-wavelength radiation to longer visible light had been known for many years prior to this lamp application.

The next step forward in lighting concepts was the demonstration in 1950 of the first practical electroluminescent lamp.[2] Electroluminescence as a scientific phenomenon had been previously reported by Destriau; however, in the experiments that he conducted the light emission had apparently been so feeble that some scientists denied its existence.

Construction of EL lamps

To create the high fields needed for electroluminescent (EL) light production, a suitable phosphor is placed between electrodes separated by only a few thousandths of an inch when the applied potential is 120 volts. The phosphor is embedded in a nonconducting dielectric material and sandwiched between the electrodes, which for most present lamps are between 1 and 100 square inches in area. Much larger experimental lamps have been made—for example, up to four square feet and up to ten feet long. These are obviously area light sources.

Since the light is generated within the very thin layer of phosphor between the electrode plates of a capacitor, one of these electrodes must be light transmissive. A well-known process for making the surface of a glass plate conducting is to heat the plate to about 600°C and coat by spraying with a tin chloride solution, then cool at a rate that will leave the glass in a properly annealed state. The resultant thin transparent layer is a good conductor of electricity. These conducting glass



plates are used in some versions of electroluminescent lamps. They may be used as a cover for a lamp built up from some other more rigid and opaque electrode, or the phosphor layer may be applied to the conducting glass, which is then covered by a second electrode in the form of an evaporated metal film or metal foil. At present there are three types of lamps commercially available as light sources. These are

1. The rugged metal–ceramic lamp, in which one electrode is a thin sheet of enameling iron onto which several layers are applied in successive steps: a thin, ceramic, reflecting layer of high dielectric constant; the phosphor + low melting glass embedment frit; conducting film; and protective glass layers. In all cases the glass is applied by being sprayed as a fine powder, carried in a liquid vehicle, which melts to a transparent layer on brief heating. In finished lamps, the total thickness between electrodes is from 3 to 6 mils, depending on the intended application.

2. A glass–ceramic lamp, which is similar to the metal–ceramic type except that a conducting glass plate is substituted for the iron electrode, in which case a ceramic ground coat may be eliminated. The glass electrode may be used as the light-escape side; then the second electrode may be opaque (evaporated film or an applied foil). These lamps are not as rugged as the metal-based ones.

3. The flexible-plastic lamp, which consists of an aluminum foil electrode coated with a thin layer of white high-dielectric material, phosphor embedded in an organic dielectric, a thin sheet of flexible light transmissive material made conductive, and a thin plastic-sheet cover. The entire assemblage is pressed into a plastic envelope of good moisture- and electrical-protection properties. The total cover-to-cover lamp thickness averages 0.025 inch. Such lamps can be made in fixed dimensions or in continuous rolls.

Standard commercial metal–ceramic lamps usually emit 1 to 1.5 foot-lamberts (fL), whereas the flexible-plastic lamp emits 5 to 7 fL, both from domestic power supplies of 120 volts, 60 c/s.

In addition to their geometrical advantages, EL lamps possess advantages derived from the phosphors themselves. For example, just as with incandescent lamps, their light output increases with increase in voltage. Their color, however, changes only little with applied voltage. Lamps for domestic use (in the United States) must be made to operate at 120 volts rms. This places a restrictive limit on EL lamp capabilities, since even for the thinner lamps (practical thickness is limited by ability to manufacture thin layers of uniform thickness) higher operating voltages can be applied, with a considerable increase in brightness.

The frequency of the exciting voltage has a more profound effect on light output than does voltage, partly because frequency may be varied over a wider range for a given lamp. To begin with, phosphors when studied in demountable laboratory lamps show that light output is composed of individual pulses of light generated at each voltage pulse, and is thus proportional to the voltage-pulse frequency. In the ideal case, approximated by some blue emitting phosphors, a tenfold light output is achieved by increasing frequency from domestic line frequency, 60 c/s, to 600 c/s. In the same ideal case, light output increases by a factor of 100 at 6000 c/s. Increases in light output by increasing frequency, although obtainable in the laboratory in demountable cells, for various reasons are not always obtainable to the same extent in commercial lamps. For example, the current drawn by an EL

lamp rises proportionately with power frequency, and thus the current may be sufficient to cause internal heating. Moreover, a lamp operating at 6000 c/s is acoustically undesirable. Depending on the phosphor used, which may be selected to yield maximum output at 60 c/s, the color of light emitted may change markedly as frequency changes. This may be desirable or undesirable, depending on the application. In addition, at high frequencies it is sometimes difficult to achieve uniform voltage distribution over a large-area lamp. Finally, and very important, the lamp life is almost inversely proportional to operating frequency.

The shift in color of emitted light with frequency is a phosphor property, the detailed study of which provides an avenue for probing into the nature of the electroluminescence mechanism. It was early recognized that the light emitted by a zinc sulfide phosphor stimulated by an electric field consisted of two major bands—blue and green. It is generally true of most phosphors that the distribution of individual wavelengths in their emitted light is intermediate between that of sunlight, or other hot-body radiation, and atomic emission spectra. In sunlight there is a smoothly varying amount of energy from one wavelength to another—i.e., a continuous spectrum whose energy gradually decreases from a peak in the infrared. In atomic spectra the light emitted (table salt in gas flame or a short-circuited copper lead wire) is concentrated into very sharply defined wavelength intervals: the spectral lines of atomic radiation or sharply defined energy peaks in the emitted light.

The most useful present-day EL phosphor is basically zinc sulfide modified by small additions of some other elements to produce various desired effects, such as color change, crystallinity, or varying frequency or voltage responses. The addition of copper is a basic and required addition, since it plays at least two roles in the phosphor. The brightest phosphors contain small amounts of a halogen. Chlorine is most common, but iodine is better if blue emission is desired. Varying degrees of substitution of zinc by cadmium (up to 12 per cent), or sulfur by selenium or oxygen, will shift the emitted light color toward yellow and red.

The field-stimulated emission of the Cu + Cl activated phosphor consists of two broad bands—one peaking at about 4600 Å (blue), and the other at about 5200 Å (green). Since each of these bands extends about 500 Å on either side of its peak wavelength position, the bands overlap to a considerable extent and contribute sufficient radiation both below blue and above green to produce a composite white-color light for certain energy ratios. Therefore, for a phosphor excited by a given frequency, the two peaks may produce a whitish color. Any imbalance in peak height will result in an obvious color mixture—i.e., a bluish green or greenish blue. However, because of the deficiency of yellow and red in the phosphor emission, a good white can be produced only by the addition of a third emission band, peaking at about 5900 Å. The technique used is to blend in a second phosphor activated with manganese, cadmium, or selenium, having a red emission band.

By selection of a phosphor having blue and green bands of proper peak height, or by mixing of two phosphors with separate blue and green bands with a yellow-emitting phosphor, lamps of a wide variety of colors and white tones may be produced. No problem arises so long as these multipeak white lamps are operated at approximately the same frequency at which the phosphor blend was selected. The color of such lamps may change with frequency, because the blue component builds up more rapidly with increasing frequency. A schematic, but typical, brightness vs. frequency relationship is shown in Fig. 1. The difference in variation for the different emission bands is given by the slope S values noted on each curve.

The blue brightness builds up almost directly with frequency ($S \approx 1.0$), while yellow builds up much more slowly; thus the blue component will rapidly swamp out the yellow. The green component presents an added problem in that after an intermediate rate of increase with frequency to about 300–400 c/s it begins to level off, and in the region of 1000–1200 c/s it reaches a plateau. It is evident that there is a difference between the emission mechanism for green and blue light and that the green process is slower, being interrupted as the time between voltage pulses gets shorter and shorter. The practical consequences are seen in the change in lamp color as frequency is increased.

Theories on electroluminescence

The scientific reader may like to know something in more detail about electroluminescence as a physical process. There are several and somewhat diverse theories on this subject. The phenomenon is a complicated overlap of two modern active scientific fields—luminescence and semiconductors. Crucial data for quantitative treatment of some aspects of the theories are not easy to obtain because the necessary experiments are difficult to carry out or interpret in spite of a considerable amount of effort in many laboratories. Present theories have gone as far as possible with present data. H. Ivey[3] recently reviewed and commented upon the diverse researches on electroluminescence that must be considered. In the present article we will consider a more specific model of the actual atom environment of a crystal at the site at which light is produced by an electric field. This will be stated after the theories developed on the basis of energy-band concepts have been outlined.

The light emitted when an alternating field operates on an EL ZnS phosphor is generally attributed to the return of an electron to a "luminescence center." This center is in the immediate vicinity of a copper atom or ion of the small ZnS phosphor crystal. It is the experience of workers on electroluminescence that the development of a high degree of crystallinity is a requisite step in activating an EL phosphor.

The return of electrons to luminescence centers is also the process by which normal ultraviolet-excited luminescence in ZnS is often explained. The centers always have electrons associated with them in an unexcited state, and when the electron is "excited" away from the center, it leaves an electron "hole." When an electron returns and recombines with the hole, light emission occurs—this emission being the radiation equivalent of the energy difference between the energy level of the "hole" level and the last stopping place of the electron. A crystalline state is necessary if this light-emitting recombination process is to occur.

From a detailed study of physical properties of solids, the concept of an average energy band that characterizes solids has evolved, and it has been applied in a spectacu-

lar manner to materials. Germanium and silicon are outstanding examples. The energy bandwidth is a measurement of the energy required to excite an electron to such an extent that it leaves its parent atom on a regular crystal lattice site (valence band) and floats about the crystal as a free charge (conduction band). In the case of ZnS (hexagonal structure), this requires approximately 3.7 eV. At the wavelength, approximately 3350 Å, of a light quantum of this energy, absorption of radiation will promote electrons to the conduction band in ZnS. The average lifetime of such free electrons is very short, so they soon become captured. In a very pure material the electron will fall down to the valence band, recombine with the hole created, and release its excitation energy as radiation. Since little energy was lost in the process, light of only slightly longer wavelength will be emitted; that is, no visible luminescence will result. The addition of small amounts of certain elements—e.g., 10–100 parts per million Cu + Cl—to the crystal will convert it to a phosphor because, in energy band terms, the added copper forms discrete or localized energy levels within the natural band gap of the ZnS crystal. These are the levels of the luminescence centers of the phosphor. The centers, containing Cu atoms, also have electrons that absorbed ultraviolet energy can excite or raise to the conduction band. Because the electrons lie above the ZnS valence band, UV radiation of lower energy (longer wavelength) will raise the "center" electrons to the conduction band. Here again no visible luminescence results when such electrons recombine with the centers by direct transition from the conduction band. Since bright luminescence is actually observed, some of the electron energy must have been given up en route. For this reason "traps" lying just below the conduction band have been postulated and are generally accepted as an essential part of the ZnS luminescence mechanism. Thus light emission results from hole-electron recombination—that is, electron transitions from "trap" to "center" levels. The energy level diagram shown in Fig. 2 permits the reader to visualize the mechanism.

In the normal UV excitation process the "center" electrons are raised to the conduction band by absorption of UV energy at about 3650 Å by the crystal. After a very short period they are temporarily trapped in the shallow traps (crystal defects, Cl atoms, etc.) from which they drop down to recombine with holes of the original center. The electrons then release their excitation energy as a broad band of radiation peaking at either green or blue, depending on the center involved.

The real difference between the theories of photoluminescence and electroluminescence lies in the method by which the field imparts sufficient energy to electrons residing in the centers to excite them—to the valence band in the case of electroluminescence. This is the subject of much controversy which is not yet resolved.

Although the electric field produced across a phosphor layer 1/1000 inch thick by the application of 120 volts to the lamp electrodes is high (approximately 44 000 volts per cm), it is still insufficient to furnish the ionization energy of approximately 3 eV that is required to excite an electron from a center into the ZnS conduction band. Since the diameter of a copper atom is nearly 3 Å, voltages are required that will produce a 1 volt per Å field in the crystal—i.e., fields of 10^8 volts per cm. A field of this strength applied across a lamp of 4 mils (100 microns) electrode spacing will require the application of one million volts. The 120 volts available on an operating lamp of 4 mils thickness is too little by a factor of 1000 to ionize the luminescence centers directly. The required voltage may be reduced, however, because of the Zener effect.

Efforts have been made by researchers to postulate mechanisms that will account for the electron source on which an electric field operates to produce light. Obviously, for the field to raise electrons from the centers, it must somehow become highly concentrated, and make all its energy available over the very short gap within which some centers lie. The required direct ionization field of 1 volt per Å thus places a maximum width of 120 Å (approximately 40 Zn + S layers of a ZnS crystal)

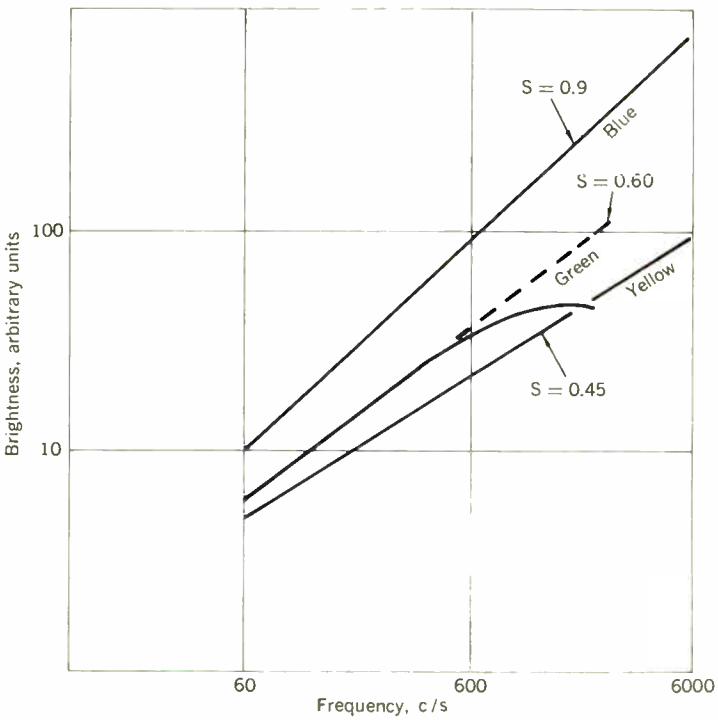
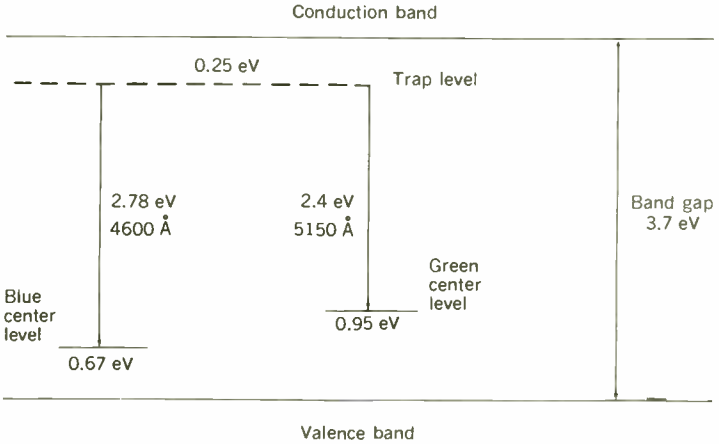


Fig. 1. Brightness vs. frequency in EL phosphors.

Fig. 2. Energy-level diagram of a ZnS phosphor (with a hexagonal structure).



for such a voltage multiplication barrier. For this reason, the remainder of the phosphor crystal and other layers between the lamp electrodes must have nearly zero resistance.

Acceleration–collision theory

To avoid the necessity of requiring such highly localized barriers of high field strength within the phosphor crystal, a theory has been developed which assumes that an initial source of a few electrons is available from shallow traps within the crystal.[4] Electrons from these traps are brought into the conduction band immediately upon the application of voltage, inasmuch as they require 10 to 100 times lower field strengths, corresponding to traps at depths of 0.3 to 0.03 eV. These few electrons are then accelerated to sufficient energies that on collision with electrons of the centers, the centers ionize and thus provide the main source of electrons for continuation of the normal luminescence process. This is the mechanism of the acceleration–collision theory of electroluminescence. This theory starts then from a few thermal electrons (residing in the crystal) which are field accelerated to energies capable of knocking other electrons out of deeper traps, and finally out of deep centers. The center electrons, once raised to the conduction band, are likewise accelerated and contribute to the center ionization process; during this time a portion of them are recombining with empty centers to produce light emission. This series of events occurs each time the ac field changes, and thus electrons are accelerated back and forth between lamp electrodes, with the result that various minor side effects are produced.

An ingenious model of this process of acceleration, trapping, and recombination of electrons, as voltage changes, has been illustrated by Ivey.[5] His latest book[3] reviews the EL phenomenon and should also be consulted for more references and details of research in this field.

There is an entirely different class of EL materials and phenomena that is dependent not on an internal supply of electrons, but rather on electrons injected from an external supply. Some ZnS crystals, for example, emit by the injection mechanism process—i.e., by low-voltage excitation in which the electrons are externally supplied. In these cases the flow of current is proportional to light output. Other materials in which p-n semiconductor junctions can form will electroluminesce; indium metal soldered to n-type germanium is an example.

Much effort has also been expended recently in the study of light emission from p-n junctions in transparent semiconductors, such as gallium arsenide (GaAs). Some of these emit intense but very small light spots. Thin films of ZnS, about 1 micron thick, will emit high light intensities over small areas.

The energy-band model described previously is a physicist's model, developed originally to account for electrical properties of solids. In the field of phosphors it provided a framework for classifying the different roles of various additives and treatments found necessary in the preparation of phosphors of various characteristics. Attempts were made to correlate composition and treatment details with various energy levels that had to be assumed in the ZnS energy-band picture to account for the observable emission and absorption bands. In all these representations, the details of the actual structure environment are not considered. Furthermore, observations have been made of several unique properties of ZnS phosphor crystals, which clearly reveal a closer structural dependence of electroluminescence than had been expected and which cannot be treated by an average statistical model.

It was discovered early by the writer that X-ray examination of individual ZnS crystals reveals a mixed-up structure. Bright EL phosphors can be made by producing a well-ordered hexagonal crystal of ZnS: Cu–Cl, which is then subjected to a lower temperature treatment after a mild mechanical working, such as light grinding or impacting. During this process the crystal partially transforms to a cubic phase. Characteristically, the cubic phase does not develop over large microscopically resolvable regions, but the new phase is mixed on so fine a scale, and in such a random manner, that its presence can be detected only by X-ray diffraction or by complex changes in index of refraction in the crystal's C-axis direction. A fine and complex color banding then appears when the crystal is viewed under a polarizing microscope.

Several workers noticed and reported these mixed structures and stacking faults in the Zn + S layering of an EL crystal, and so the idea developed that electroluminescence and stacking faults were correlated in some manner, and specifically were the source of internal barriers, which had to be assumed as a means of multiplying voltage to produce the high field strength required for direct ionization of the center. However, stacking faults are a one-directional structural disorder, and result when a stacking pattern of Zn + S layers suddenly changes with respect to the pattern shown by neighboring layers in the crystal. There is no disorder in the layering plane. Thus, although stacking faults may somehow create barriers in the C direction, they will not do so in the layering plane perpendicular to the C axis.

The writer then discovered that there is indeed a remarkable directional dependence of EL emission in single crystals of ZnS. When an electric field is applied across a crystal, viewed under a microscope, it is found that a much higher field is required to produce light when it is applied perpendicular to the layering than when applied parallel to the layering. In fact, the voltage difference is more than tenfold, and frequently much higher. Therefore, if barriers are a necessary criterion for EL emission in this "easy direction," they must exist in a direction perpendicular to the crystal C axis, and not along the C axis, where stacking faults actually exist. In repeating these observations, other workers found a correspondingly great difference in electrical conductivity in ZnS crystals. Highest conductivity is always in the direction in which EL emission occurs at lowest applied voltages. This observation is significant in the evaluation of theories of electroluminescence. If stacking faults are acting as barriers to create extremely high field strengths in narrow bands of the crystal for fields applied along the C axis, then under these conditions the acceleration–collision process of luminescence-center ionization is operating. However, orders of magnitude of higher conductivity are exhibited perpendicular to this direction in the crystal, and much higher light output results when the same field is applied in this direction. It is evident then that the injection mechanism is operating when a field is applied parallel to the ZnS layering. These observations are based upon single crystals with metallic elec-

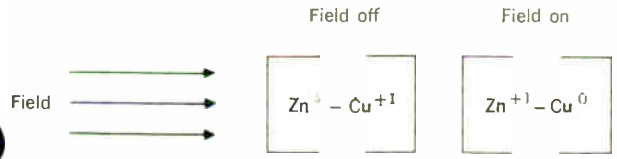


Fig. 3. Electronic state in EL center.

trodes in contact with the crystals. At least one electrode must be attached to the crystal to observe low-field electroluminescence parallel to the layering.

When embedded (in random orientation) in a solid dielectric, the particle must use its own intrinsic source of electrons, which confines the emission of light principally to the high-field mechanism wherein acceleration and collision of a few electrons bring the center electrons into action and activate the luminescence process. As far as we can predict, light emission is largely restricted to fields parallel to the C axis for such cases, because there are not sufficient electrons available for the lower field–higher current mechanism that could otherwise produce light for crystals oriented with their C axes perpendicular to the field.

In spite of this added knowledge of the details of the dependence of structure on electroluminescence in ZnS crystals, other important observations remain to be explained. For example, some beautifully birefringent banded crystals, which indicate by means of X rays, complex mixed structure and stacking faults, show no trace of electroluminescence. Further, occasional crystals are found that are strongly electroluminescent, and yet they show none of the aforementioned features. There are methods of preparing EL phosphors of cubic structure that have none of the drastic disorders detectable by polarized light or X rays. Obviously some more detailed theory had to be developed to account for the experimental observations.

One-atom displacement disorders

As a result of a detailed study of the various possible ways in which "geometrical mistakes" can be made by a growing ZnS crystal or ways in which an existing perfectly ordered one might be disordered, a "one-atom displacement" disorder was recognized as a disorder possibly related to electroluminescence.[6] Its basic difference from a stacking fault is that it can form and exist independently as an isolated defect not involving an entire layer of the crystal. The one-atom disorder does not recognizably affect the crystal's response to polarizing light or X rays, and it can occur in both cubic and hexagonal crystals. Equally important is the model it furnishes for understanding and predicting a large and otherwise confusing mass of chemical properties of EL phosphors. Finally, this model provides an important mechanism for the generation of light by the electric field, and specifically for the highly directional effect described previously.

The essence of this one-atom displacement model is that an atom–ion pair is created in the crystal as a consequence of this disorder, either during crystal growth or by a subsequent deformation. The disorder differs from a stacking fault in that the latter involves a vertical pair of Zn + S atoms. In the one-atom displacement, one atom only—e.g., a sulfur atom—is displaced from its regular lattice site. A broken vertical bond results, which is very reactive—promptly combining with any Cu^{+1} ions in the vicinity. There is just sufficient space for a Cu^{+1} ion to enter the structure at the position of the broken bond. Of course, a broken Zn bond is also created. The geometry of the crystal is such that this "foreign" Cu^{+1} ion is now placed in the approximate center of a triangle of three Zn atoms on their original lattice sites. One of these Zn atoms, however, like the displaced sulfur, has a broken vertical bond. A further geometrical property of the crystal in the immediate vicinity of this disorder—now reacted with a Cu^{+1} ion—is that there is formed an excited atom–ion pair of two different metals on a line perpendicular to the crystal C axis. Specifically, the pair is Zn^*–Cu^{+1}. The Zn^* represents the particular Zn atom with the broken vertical bond, and now in a higher energy or excited state. The Cu^{+1} is an ionized Cu atom that has lost one electron before entering this region of the ZnS crystal.

It takes energy to remove an electron from an atom. (For free atoms in a gaseous state this amounts to the ionization potential.) By the same token, any atom that has lost an electron tries to recover it. (Some atoms even in a free state attempt to obtain more electrons, a characteristic measured as electron affinity.) Therefore, when an atom–ion pair of different metals is brought together the electron transfers to and resides most of the time on that particle of the pair having the higher ionization potential. Therefore, the Zn remains an atom and the Cu^{+1} an ion. The ionization potentials of the single atoms are 9.39 eV for Zn and 7.72 eV for Cu. The Cu^{+1} ion, however, would like to take an electron from the Zn atom; it lacks only a little more than 1.67 eV energy to do so for the free-atom case.

In the case of a solid ZnS crystal, these energies are considerably modified and reduced. The electron affinity of S atoms already tends to pull electrons from the Zn atoms in the crystal, so that only 3.7 eV—instead of 9.39 eV required of free atoms—is needed to ionize Zn. The ionization energy for Cu is also lower in a solid. In the case of the (Zn^*–C^{+1}) pair of the EL ZnS crystal, the Zn^* atom is already excited above the level of normal lattice Zn atoms. This means that even less "pull" is required of the neighboring Cu^{+1} ion to obtain an electron from it. The exact amount of extra pull needed is at present unknown, but it is presumed to be much smaller than for a normal lattice Zn atom. This then is the reason that this excited atom–ion pair is able to play the prime and crucial role in electroluminescence. It is an EL center, and the effects of an electric field on it may be represented as shown in Fig. 3.

The external field has assisted the Cu^{+1} ion in capturing an electron from the excited Zn atom, with the result that the Cu^{+1} ion has become an uncharged Cu atom, and the Zn^* is converted to a Zn^{+1} ion. In other words, an electron has been transferred within this localized region, which has been termed an EL center. When the field reverses its direction (ac operation), of course, the Cu^0 atom rapidly loses its electron to the Zn^{+1} ion which, even without the reverse field, has more than sufficient electron affinity to recapture the electron. This model has strong directional properties due to the intrinsic directional pull of the Cu^{+1} ion for an electron on a neighboring Zn^* atom. In an actual crystal there will be three such directions, corresponding to three possible orienta-

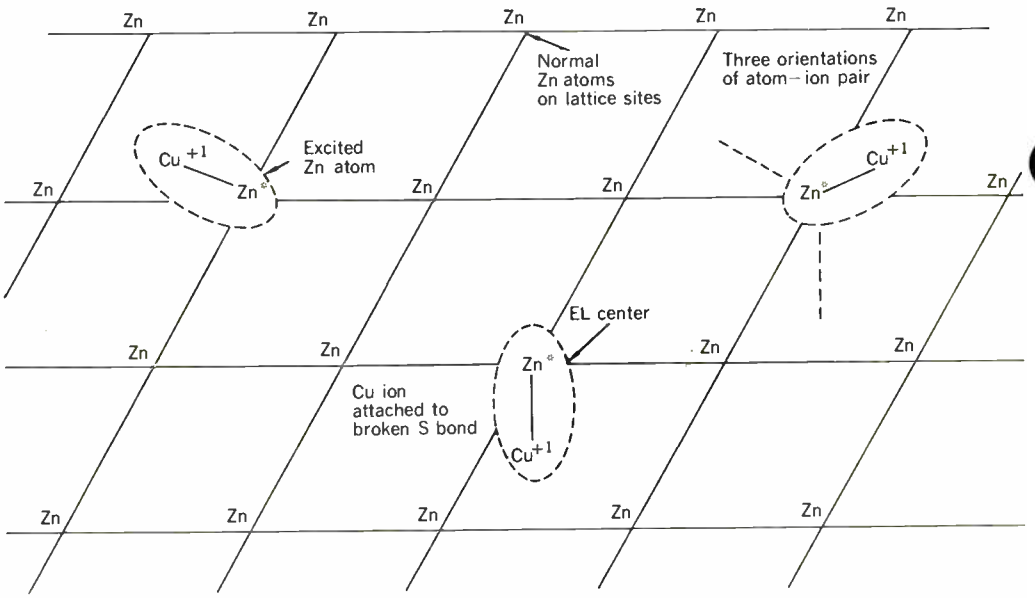


Fig. 4. Orientation of EL centers in Zn net of ZnS structure.

tions of Zn* atoms with respect to the Cu^{+1} ions.

There are several consequences of the illustrated electron-transfer process to consider with regard to electroluminescence. Since only interatomic distances are involved, the electron arrives at the Cu^{+1} ion with its excess energy provided by the field. Does the electron merely stay at some position very close to the Cu ion without actually giving up its energy, or does it actually combine with it, as implied in the schematic diagram? In this case the electron must give up its excitation energy as it falls down to lower energy levels of the resulting Cu^{0} atom. Is this a radiative recombination making some contribution to EL emission? The quasi-free state of the Cu^{+1} ion (with respect to the ZnS structure) may enable it to behave somewhat as a free atom. At any rate, as voltage returns to zero and reverses, the electron will be promptly ejected from the Cu^{0} atom and return to its parent Zn* atom. This process can throw further light on the emission from the Zn* atom.

A more direct contribution of this atom–ion pair model of the EL center to the electroluminescence mechanism problem is as a possible source of electrons for triggering off the normal luminescence process in regions of the crystal surrounding the EL center at relatively few interatomic distances. This phenomenon may be described in terms of a plot of the actual Zn positions in a horizontal plane of the ZnS structure. Figure 4 shows three EL centers, with the three possible orientations of Zn* and Cu^{+1}.

If the electric field is exactly parallel to the atom–ion axis of a center, electron–ion recombination may occur. For a field oblique to this axis, the electron—although pulled by the ion—may be pushed by the field beyond the Cu^{+1} ion and escape into the surrounding portion of the crystal. Here it may immediately encounter a normal luminescence center (Cu on Zn lattice site) during the field-on period, or eventually drift into one by its own diffusion between voltage pulses. An electron escaping an EL center in the manner mentioned may also be accelerated to an energy capable of directly ionizing a luminescence center. In this case, this portion of the disorder theory differs from the older collision–ionization theory only in that the Zn*–Cu^{+1} or EL center, rather than the luminescence centers themselves, is the major and prime source of electrons. Moreover, in this case the prime sources of electrons are those particular lattice Zn atoms lying adjacent to the Cu^{+1} ion, which has reacted with a sulfur atom whose displacement has left this Zn atom in an excited state and with a broken bond. In the collision–ionization theory a Cu activator atom on a lattice site is the major source of electrons.

Applications

The development of an area light source, based on electroluminescence, has necessitated extensive research on phosphors, as well as on other materials used in an EL lamp. At present the rugged EL lamp of ceramic–metal construction (1–1.5 fL) and the organic dielectric embedment lamp (5–7 fL) are of lower brightness than fluorescent lamps and therefore find a more restricted field of use. Today EL light sources can be used where space is limited and where low to medium brightness is needed, such as in night lights and in instrument panels of automobiles, aircraft, and space vehicles. They will withstand intense vibration and extreme temperature variations without abrupt failure.

The pliable Panelescent Tape-Lite, recently developed by Sylvania, has a higher brightness. It is suitable for use in many military, commercial, and industrial applications, as well as in decorative or safety applications around the home. It can be twisted, bent, or turned into hundreds of shapes and sizes and can assume the size and shape of any object it is placed over.

Electroluminescent units can go into curtains, which will provide light of approximately the same level as daylight from the same area of the room. They can also be installed as wall or ceiling panels or in sidewalks. A possible future use might be the paneling of the outside of buildings at street level to provide medium-level after-dark lighting in industrial areas.

Electroluminescent display devices

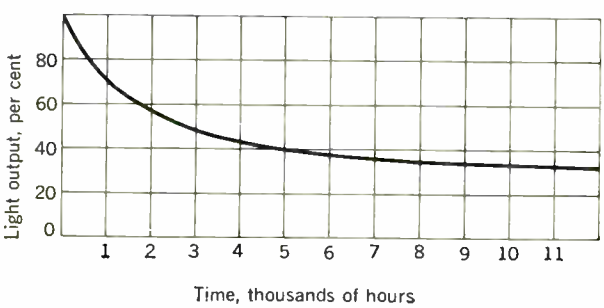
Irving Greenberg

An infinite variety of display devices is made possible with electroluminescence because of the fact that light output is usually produced only where a capacitor has been formed and voltage applied. The only restriction is that the elements comprising the display must be segmented; i.e., active lines cannot cross. In practice, display devices are manufactured on glass substrates, where the top electrode is patterned for ease of connectoring. Light output, from the EL dielectric layer, is seen through the transparent conductive bottom electrode.

Fig. 5. Basic display devices. A—7-segment numeric. B—9-segment numeric. C—14-segment alphanumeric.

Fig. 6. Type NU58A multinumeric, as employed on a Beckman Instruments Corporation frequency counter.

Fig. 7. Typical curve showing light output vs. time, for a 250-volt-rms 400-c/s source.



Numeric and alphanumeric devices

One of the most basic display devices is the numeric. In its simplest form, it contains seven segments, as shown in Fig. 5(A). Combinations of segments may be lighted to produce all the numerals, from zero through nine. The top electrode, which forms the segments of the display, is generally a vacuum vapor deposition of aluminum. The size and position of the electrodes are controlled by graphic art techniques and provide a small spacing in the gaps between the segments (typically 0.010 inch). Although the seven-segment design represents the fewest number of segments capable of forming the numeric display, it cannot center the number one. If this centering is desired, a nine-segment design can be used, as shown in Fig. 5(B).

To display the letters of the alphabet, as well as the numbers, diagonal segments must be added to the basic numeric design. The resulting display consists of 14 segments, as shown in Fig. 5(C). The appearance of the letters may be improved by utilizing more segments at the expense of greater switching complexity.

Most numerical applications require a multiplicity of digits to display the measured parameter. For these, a large variety of multinumerics are manufactured; an example is illustrated in Fig. 6. It is supplied with a special legend, which may be selectively lighted, to show the scale to which the meter is set. The multinumeric offers benefits over the use of individual numerics placed side by side, in that it provides optimized spacing from digit to digit, and eliminates the lines and framing hardware between digits. In applications where space is at a premium, it allows for the closest spacing possible between digits.

Numeric and alphanumeric display devices are generally operated under conditions of 250 volts rms and 400 c/s. These conditions represent a compromise between initial light output and light maintenance, or life. Since the light maintenance is primarily dependent upon the operating frequency, with voltage having only a secondary effect, it is desirable to keep the frequency as low as possible. At 60 c/s, the voltage required to obtain the desired light output is too high to be handled with conventional components, so the frequency of 400 c/s is chosen. At a voltage of 250 volts, the initial light output is approximately 10 to 11 fL.

The exact end of useful life cannot be specified, since the readouts do not fail catastrophically. Instead, they lose light output as shown in Fig. 7 when operated at 250 volts rms and 400 c/s. Because of this, it is seen that end of useful life is dependent upon the amount of ambient illumination and the degree of light shielding provided by the usual bezel. Viewing tests have been conducted on a display similar to a frequency counter, in which end of life was simulated by lowering the applied voltage to reduce the light output to a value below that required for adequate viewing, and then raising it to an acceptable level. The acceptable limit varied be-

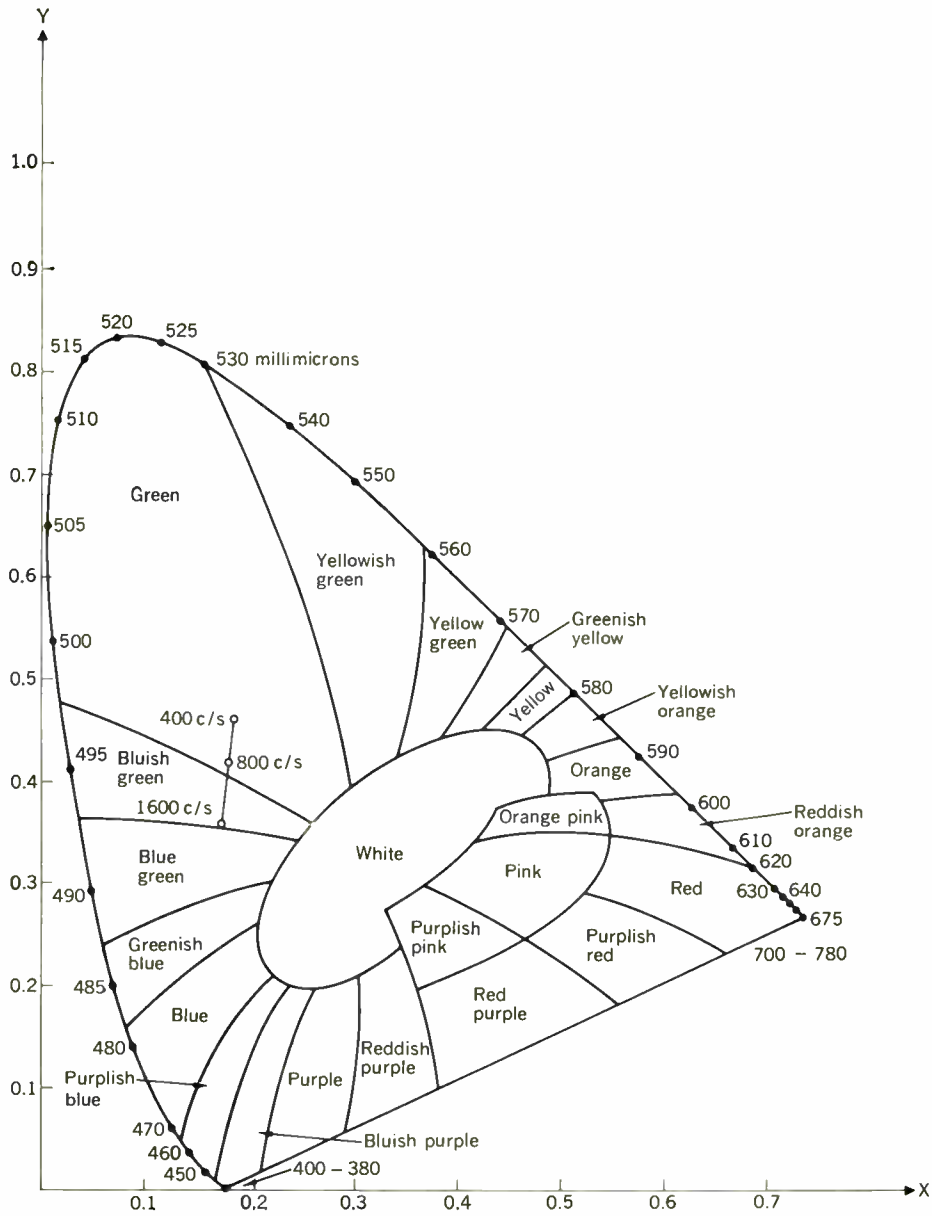


Fig. 8. C.I.E. chromaticity diagram showing ISCC–NBS designations for colored lights.

tween observers from 15 to 25 per cent of initial value. The EL display device is not affected by the number of turn-on cycles; thus, for example, in an application where the readout is used only 50 per cent of the time, the useful life will be approximately 40 000 calendar hours, which is in excess of 4.5 years.

The choice of phosphors

The most efficient EL phosphor available is green at low frequencies and shifts to blue as the frequency is increased, as shown in Fig. 8. This phosphor may be fabricated into an EL lamp using either a ceramic (glass frit) or plastic dielectric. Although the plastic dielectric construction provides an initial brightness that is approximately five times greater than that of the ceramic construction, its maintenance characteristics are not as good, and it is also more susceptible to damage by moisture. The green phosphor has the advantages of matching most closely to the response curve of the human eye, and is well adapted to viewing with television cameras and to printing on photographic film. The spectral energy distribution curve of the green phosphor is shown in Fig. 9.

Other phosphor colors available include yellow, blue, and white. Red is not obtainable in the basic EL phosphor, and is provided by an energy conversion technique that employs a red phosphorescent material. The relative light output levels are as follows:

Color	*Percentage*
Green	100
Blue	45
Yellow	50
White	45
Red	20

Electroluminescent display devices generally obtain their operating power from small transistorized oscillator or inverter circuits. In most equipment only a very

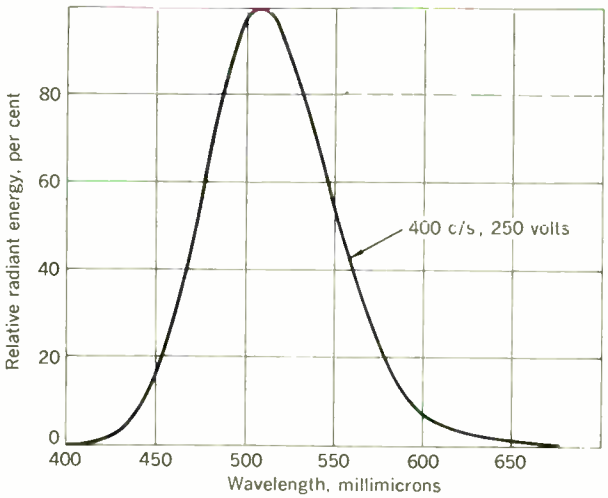


Fig. 9. Typical spectral energy emission characteristics for green phosphor.

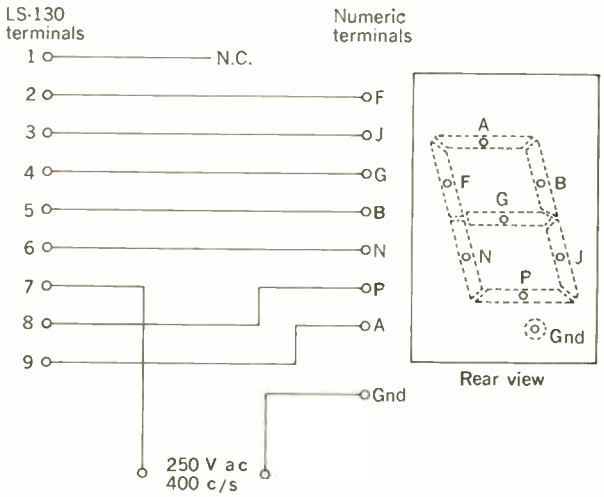


Fig. 10. Typical ganged switch for EL displays.

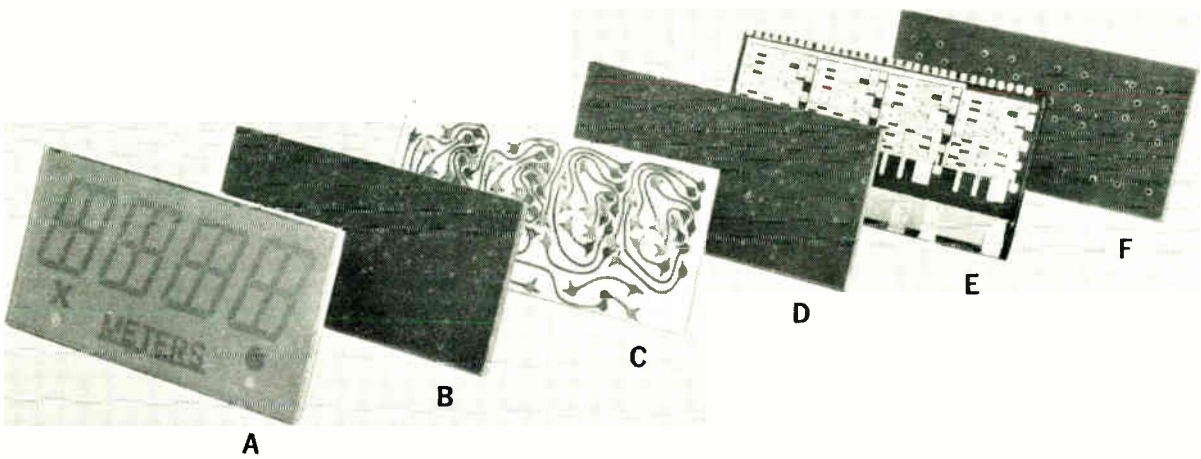
small power supply is required because of the high impedance characteristic of EL. The ½-inch-size numeric requires only 5 mW to light the number 8, which represents maximum power requirements.

There are many suitable components available for switching EL displays. Most numerical displays find application in electronic counting or scaling equipments, and electronic gating of the ac supply is required. However, some displays, such as production-line status boards, employ mechanical switching. A suitable switch, which may be ganged, is shown schematically in Fig. 10.

One of the early means of control for EL displays employed electroluminescent–photoconductive logic (EL–PC). In this technique an activated EL trigger lamp lowers the impedance of a number of PC cells, allowing current to flow to the associated segments of the display. One such application is the four-digit one-inch readout, shown in Fig. 11. This device was designed for use in a portable laser range finder for battlefield use. The EL–PC techniques are desirable because of the need for low power consumption, small volume, ruggedness, and low weight. Only ½ inch thick, the device combines the numerical readout with code conversion from binary-coded decimal input (1-2-4-8). The EL–PC switching matrix (E) is constructed on a ⅛-inch glass substrate. There are eight EL trigger lamps per digit, making a total of 32 for the entire substrate. Cadmium sulfide photoconductor material is deposited over the EL trigger lamps, and the desired output matrix is obtained by a vacuum deposition of aluminum. In operation, the proper combination of four EL trigger lamps, representing the binary number to be displayed (A), are excited from the frequency counter. These lower the impedance of the desired PC cells in the matrix and thus establish a current path to the proper EL segments in the display.

Since the outputs of the PC matrix do not line up with the EL segments on the display portion, a unique means of connectoring employing a flexible Mylar printed wiring plate (C) was devised. The Mylar is removed at each pad, exposing the copper conductors. Conductive silicone rubber cylinders are dropped into the holes in the phenolic panels, B and D. These make contact with the appro-

Fig. 11. Exploded view of four-digit one-inch readout.



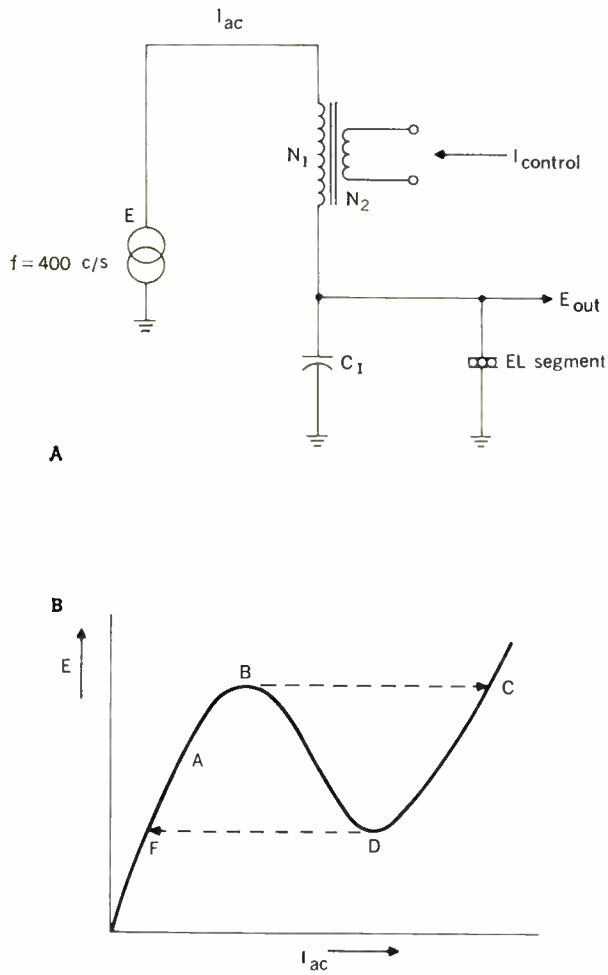
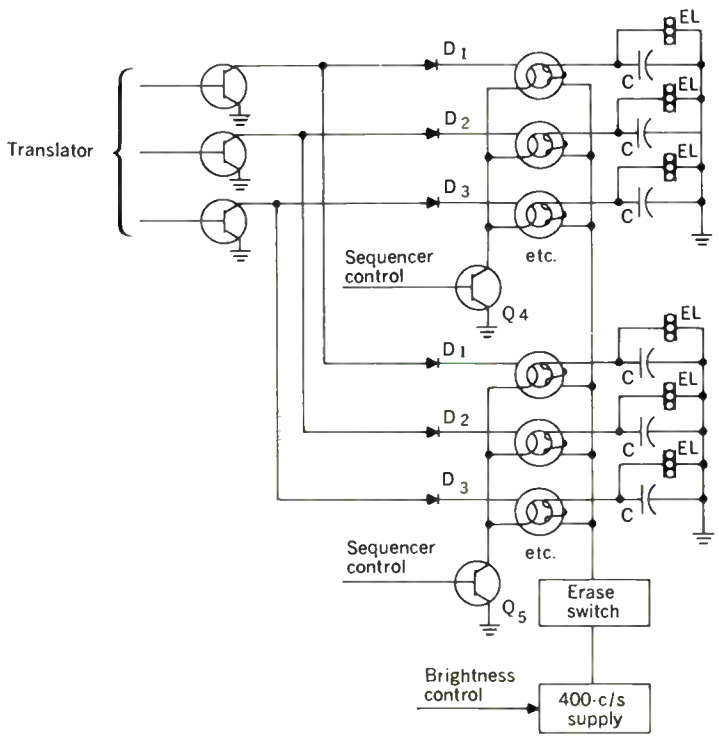


Fig. 12. A—Basic storage circuit for EL display device. B—Voltage–current curve.

Fig. 13. Magnetic gating and storage circuitry.



priate input terminals of the EL readout and output pads of the PC matrix (F). Connections to the input side of the EL–PC matrix are made with similar techniques.

The device was designed for operation at 200 volts rms sine wave at 1600 c/s to reduce the size of the core of the transformer in the power supply. The operating current for the entire display and switching matrix is less than 4 mA at the designed center operating conditions.

The input trigger lamps of the EL–PC switching matrix operate at the same voltage input as the EL display. A means of gating from the low-voltage dc logic is provided by an incandescent–photoconductive gate. This device measures less than one-tenth inch in its maximum dimension and weighs less than 10 milligrams. The dark impedance of the cadmium sulfide photoconductor is greater than 10 megohms at 1600 c/s, and the light impedance is approximately 5000 ohms.

Storage circuitry

Electroluminescent display devices sometimes employ bistable magnetic storage circuitry.[7] One magnetic switch is employed for each segment in the display. When a magnetic switch is triggered on, it remains on until power is interrupted by the erase control circuitry. During this time, the display segment associated with the magnetic switch is also on, producing part of a selected letter, number, or symbol.

The basic circuit is composed of a saturable inductor and a series capacitor, as shown in Fig. 12. This circuit exhibits two stable states, and will remain in either state until switched to the other. With the circuit assumed to be in the low-current or unsaturated state, the supply voltage is chosen so that the current through the circuit will be small and the magnetic core material will operate in a region below saturation. The circuit reactance is now mainly inductive, and a small increase of supply voltage will produce a small increase in series current. The circuit will now be operating at point A, and low output voltage will result across the output capacitor. If the voltage is raised sufficiently, enough current will flow to start saturating the core; that is, at the peak of each half cycle of the driving-voltage sine wave, the inductor will be driven into saturation, thus reducing its average impedance. At this point, enough current flows to saturate the core fully. This is a regenerative action and at point B a discontinuous jump to the high-current condition, point C, occurs. The circuit is now in the high-current or saturated stable state, and will remain there until the input voltage is reduced below point D; in this region insufficient current flows to maintain saturation and a discontinuous jump to the low-current condition, point F, results. In the high-current state, a high output voltage results across C_1 and the EL lamp is energized.

The circuit can be triggered from the "off" state to the "on" state by several methods. In this display system, a three-turn winding N_2 was added to the core and a trigger was used to drive the circuit between its stable states. The control pulse is in a form of a half cycle of the driving frequency, and is out of phase with the driving voltage.

The switching and gating circuits used in this equipment are shown in Fig. 13. Fourteen control lines are connected in parallel to one side of the control windings for each segment of the 72 characters in the display. The ground side of the control windings of each character connect through a series switch to ground. This is

used to switch the gate "on" to each character so that the drive circuits of only one character at a time are active. The diodes D_1, D_2, D_3, etc., are provided for switching circuit isolation. Erasure of the circuits is achieved by interrupting the high-voltage supply to all the display characters simultaneously. When the voltage is reapplied, the ferroresonant circuits come on in the low or "off" state, ready to display a new message.

The only known means of providing for completely solid-state display equipment involves the use of EL display devices. To complement this display capability, the silicon controlled rectifier (SCR) is starting to play an important role in switching. The usual SCR applications involve low-impedance loads and consequently high currents. When used with EL displays, they must operate into high-impedance loads at low currents.

Silicon controlled rectifiers may be used to control the EL segments in either the series or shunt mode of operation. Of the two, the shunt mode is generally preferable because the capacity of the SCR is of less concern. The switching circuit is shown in Fig. 14. The gate electrode is grounded and the cathode driven to reduce leakage currents. When the SCR is operated in this manner, its transistor characteristics are employed, since the high impedance of the series resistor does not allow the SCR to operate in its forward breakover mode of operation.

For alphanumeric displays in which completely solid-state equipment is not required, reed relays have been employed. In contrast to frequency counters, which operate the display in parallel form from the logic circuitry, this type of display is normally used with serial input. In the parallel mode of operation, one decoding matrix is required per digit. In the serial mode, switching is performed with one translator whose output is gated to the appropriate segment in which it must be stored. The reed relay provides the necessary gate and store function.

To obtain the storage feature, the reed relay is used with a permanent magnet. The reed is positioned away from the magnet so that the field is unable to close the reed. When a pulse is applied to the actuating coil, the reed closes; in this position, the field of the magnet is sufficient to hold the contacts closed.

Large-area displays

Electroluminescence is providing a new approach to the construction of large-area displays, such as traffic control boards, and situation plot displays.[8] Of interest in these applications is the crossed-grid display matrix. The basic EL crossed grid, or XY panel, consists of two sets of parallel conductors, placed at right angles to one another and separated by the EL dielectric layer. The conductors are referred to as the X (horizontal) and Y (vertical) electrodes. When an alternating or pulsed voltage is applied between X and Y electrodes, light output is produced at the intersect. Because of capacitive coupling between elements, the elements lying along each driven line receive approximately half supply voltage. This phenomenon, called "cross effect," produces a contrast ratio of approximately 10 to 1 when the light output from the intersect is compared to the light along each line. Cross effect has been found to be of value in displays in which the user requires coordinate information to be presented in a relatively static display.

For situation plot displays, a series of intersects are sequentially pulsed, on a repetitive basis, to produce the desired trace. However, in these applications cross effect cannot be tolerated, because the undesired light produced along each line would wash out the desired light from the intersects. Cross suppression is obtained by the addition of a nonlinear resistive (NLR) layer to the basic crossed grid in series with each EL element. The characteristic curve of the NLR layer is plotted in Fig. 15. The load lines represent the drop across a series EL element. Because of the nonlinear characteristic of the NLR layer, the voltage across the EL element is small when only half voltage is applied, and increases about tenfold when full voltage is applied. The brightness discrimination ratio between full voltage and half voltage becomes more than 10 000 to 1. In effect, no light output is obtained at points other than the desired intersect. In addition to providing a means for cross suppression, the NLR layer reduces the capacitance of each element in the crossed grid and thus increases the impedance.

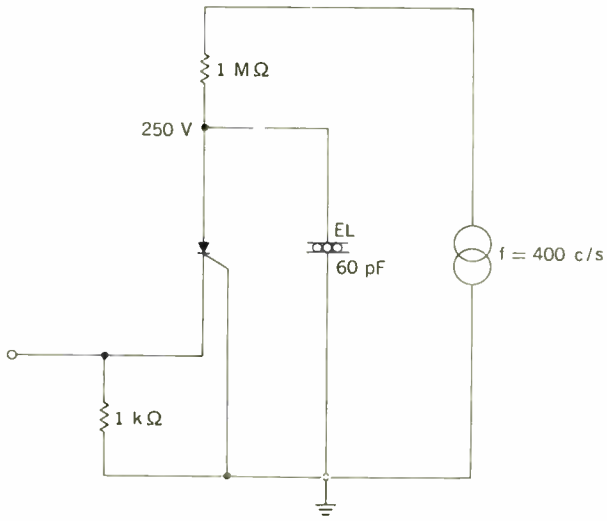


Fig. 14. SCR switching of EL segments.

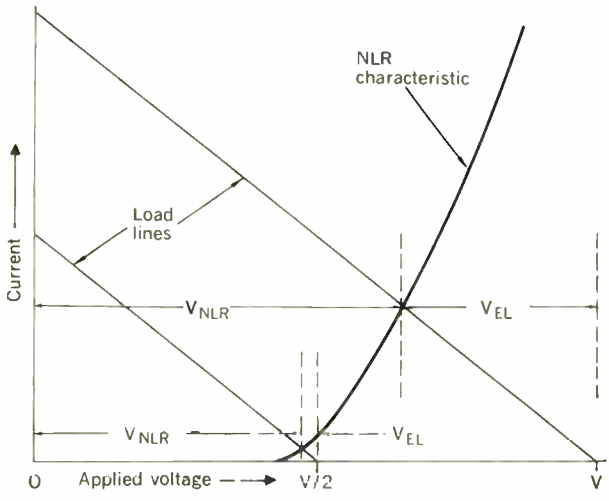


Fig. 15. Characteristics of series NLR EL device.

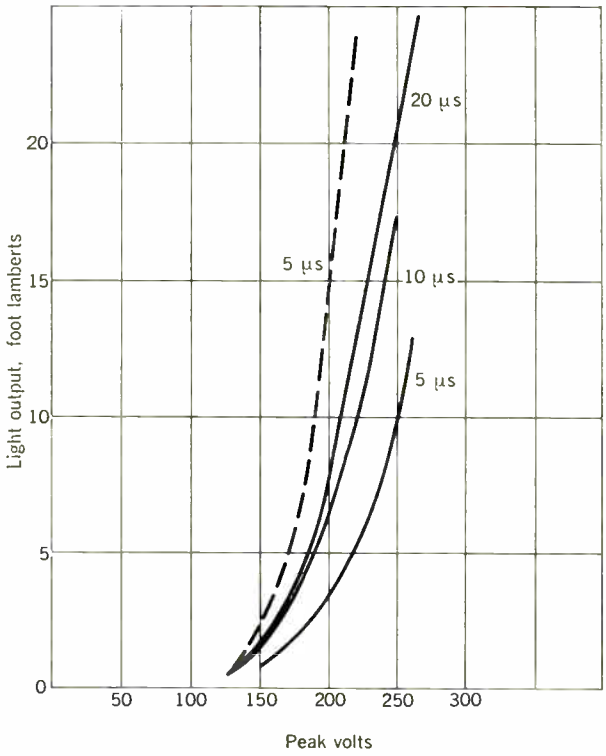


Fig. 16. Light output vs. pulse voltage for various pulse durations. Solid lines are for 10-kc/s repetition rate; dashed line is for 50-kc/s repetition rate.

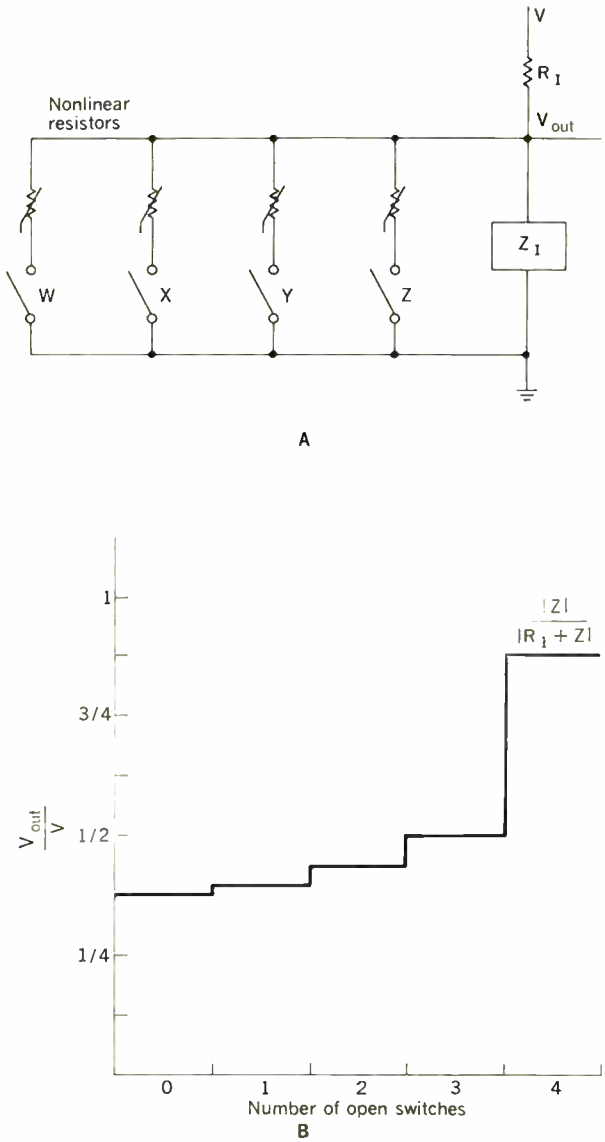


Fig. 17. A—Four-bit NLR AND circuit. B—Relative output voltage existing across impedance Z_1.

As a result, the power required from the driving source is reduced, and an economy is realized.

Crossed-grid displays are generally operated under conditions of 500 to 1000 volts peak to peak, pulse durations of 5 to 20 μs, and repetition rates of 25 to 100 kc/s. To provide a flicker-free presentation, each spot is scanned at a rate of 30 or more times per second. Figure 16 shows typical light output characteristics under pulsed excitation conditions.

For large-area displays it is necessary to manufacture modular XY panels that may be placed side by side without loss in resolution at the abutting edges. For example, a recently developed device has an active area of 4 by 8 inches in a resolution of 16 lines per inch.

Since the crossed grid is a digital form of display, it is generally operated from computer-type storage circuitry. The input drive may be controlled with magnetic circuitry similar to that discussed for numerical control, pulse transformers, or SCRs. Although the usual logic circuitry may be converted to decimal output and fed to each input line by means of shielded cable, it is convenient to perform the decoding directly at the display. By this means, the number of output cables may be reduced to that of the binary coded outputs, and the number of drivers required reduced to the same number. To accomplish this, an NLR switching matrix has been designed to provide the binary-to-decimal decoding.[9]

A typical NLR decoder is the 4-bit to 16-line type SM120. Figure 17(A) shows one of the four-bit NLR AND gates. An open switch represents an activated input. The relative output voltage across an impedance Z_1 is shown in Fig. 17(B), where the quantity $Z/(R_1 + Z_1)$ is chosen to be $^7/_8$. When none to three switches are open, the nonlinear resistors maintain the output voltage nearly constant at $V/2$. When all switches are open, the output voltage rises sharply to within $^1/_8$ of the supply voltage. Figure 18 shows a typical circuit of a crossed grid driven through SM120 decoders. Since most of the power is consumed in the unactivated AND gates of the decoders, power dissipation will not increase in direct proportion to the display area. The minimum response time for the device, which is limited only by the capacitance of the elements, is about 10 μs.

Since the crossed grid is operated by sequential scanning, the operating voltage and frequency must be high if high peak brightness is to be achieved. The average light output then becomes a function of the number of intersects that must be displayed. To overcome the low level of light output that exists when large-area displays are required, interest has been focused on storage

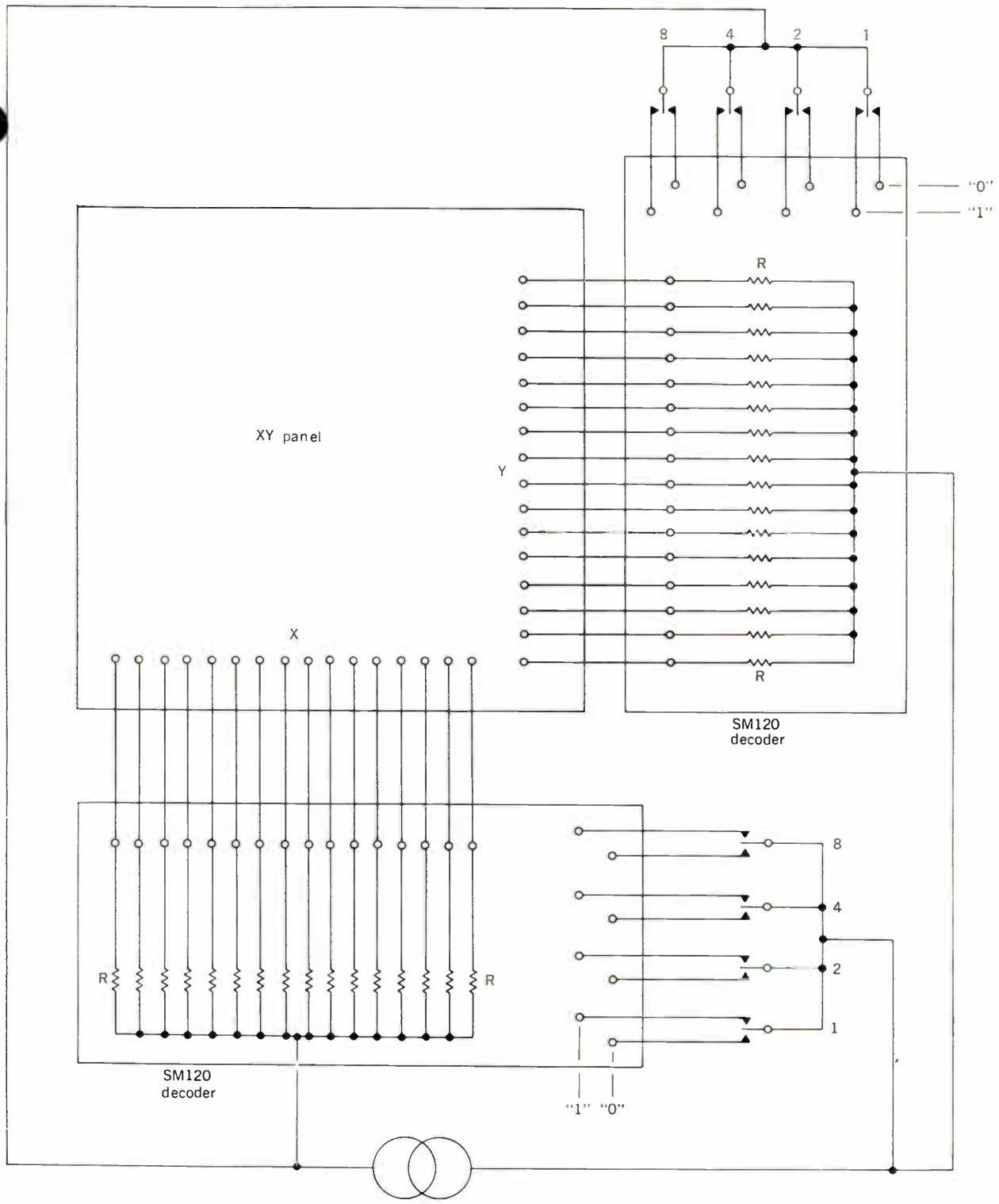


Fig. 18. Typical circuit showing crossed grid driven through decoders.

panels driven by crossed grids and other means. Each element in these panels, when excited by a source of triggering light, remains lighted until it is turned off by electric control.

The approach to the construction of a storage panel has been to utilize the self-holding capabilities of electroluminescent photoconductive elements. The basic circuit is shown in Fig. 19. Upon the application of a signal (S_1 closed), EL lamp B is lighted and lowers the impedance of its associated photocell; S_1 may then be opened and EL lamp A remains on by virtue of the path through the photocell.

Storage panels have been constructed in a resolution of five lines per inch, with an active area of 4 by 4 inches. It is believed that the basic techniques will be refined to the extent that storage panels will be available in resolutions of approximately 20 lines per inch.

The thermometer, or bar graph type of display, has

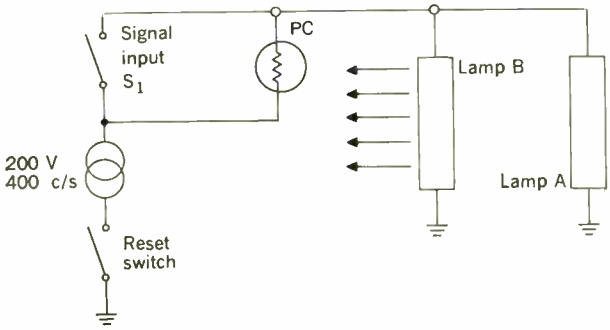


Fig. 19. Basic circuit for control of storage panel.

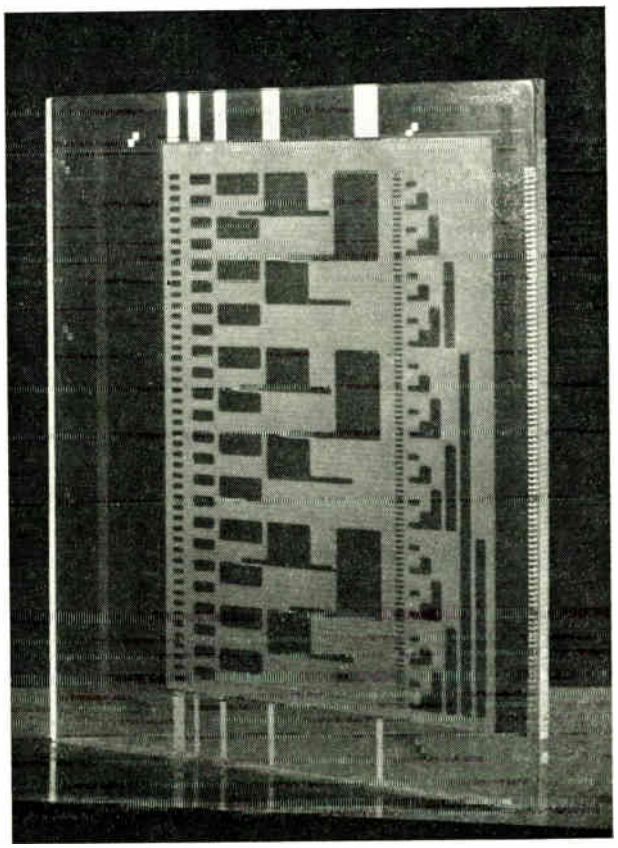
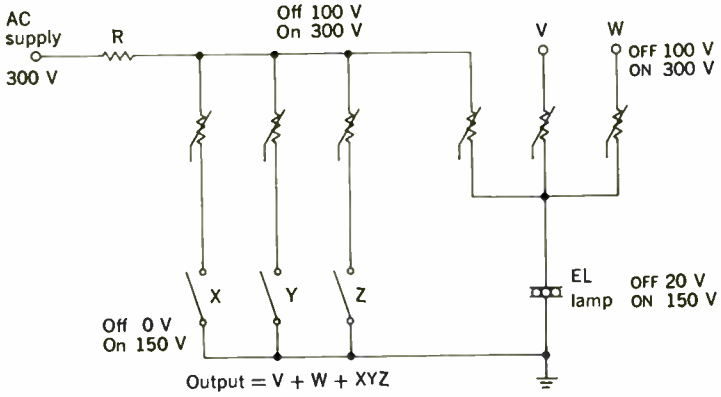


Fig. 20. Laboratory model of EL-bar graph display.

Fig. 21. Combined NLR AND-OR circuit.



been extensively used as an indicator for hydraulic systems and has many advantages over conventional dial instruments from a human engineering point of view. The EL version results in small spacing between units and thus improves the ease with which a number of displays may be compared. The present laboratory model of this solid-state display consists of a segmented EL lamp on the edge of a glass substrate (Fig. 20). The display bar thus formed is activated by a logic matrix in the form of a layer of polycrystalline NLR material. This logic matrix activates the EL display segments to form a column of light proportional to the numeric value of the binary input. Since the logic activates the exact number of display lamps, the display has no indicator error.

In operation, a column of lamps is activated which extends from the bottom of the display up to the lamp corresponding to the numeric value of the binary number input. Since the lamps corresponding to all numbers smaller than the input value must be activated, the logic need not include consideration of the zeros in the binary input code. Any lamp must be illuminated for all input codes equal to or in excess of the code corresponding to that lamp, and the logic required thus involves an AND-OR function.

The basic NLR AND circuit driving an EL lamp is shown in Fig. 17. Unless all input points, W, X, Y, and Z, are removed from ground, the voltage at the EL lamp will be held at some fixed voltage, depending upon the NLR characteristics. These characteristics are chosen so that the voltage at the lamp produces no light output. Only when all inputs are removed from ground is the output voltage sufficient to excite the lamp.

A combined AND-OR function is shown in Fig. 21. For the OR circuits, the NLR elements perform only an isolating function; the AND circuit activates the lamp in the same way as described above. In addition, an output is obtained by application of an input voltage at V or W. Representative voltage levels for the "on" and "off" states are indicated.

The AND-OR logic required for the bar graph display is implemented in the form of a continuous layer of NLR material with the matrix of elements formed by the patterning of electrodes and insulators. Figure 22 illustrates a matrix of two AND circuits. An NLR layer, represented by the vertical resistors in the figure, is continuous over the area of the substrate on which the input bus bars are deposited. An NLR element is formed by the exposed crossover of a horizontal and vertical bus bar. It is possible to make unexposed crossovers by masking the lower electrodes at the crossover.

Figure 23 illustrates a method of performing an AND-OR function for a five-bit input. The output of the AND circuit on the left side forms one of the OR inputs on the right side. In the figure, output 6 is activated by input XY or W or V. For the OR circuit, nonlinear resistors perform only an isolating function.

The NLR AND-OR circuits are adapted to the bar graph display by connecting more than one display segment to each translator output. In this way, all segments below the one corresponding to the binary input are lighted. In its present form, without additional packaging, the bar graph display panel is 5 inches high, 4½ inches deep, and ⅜ inch wide. The display column is 4 inches high and ⅛ inch wide with 95 display elements in a resolution of 25 lines per inch. The display operates from

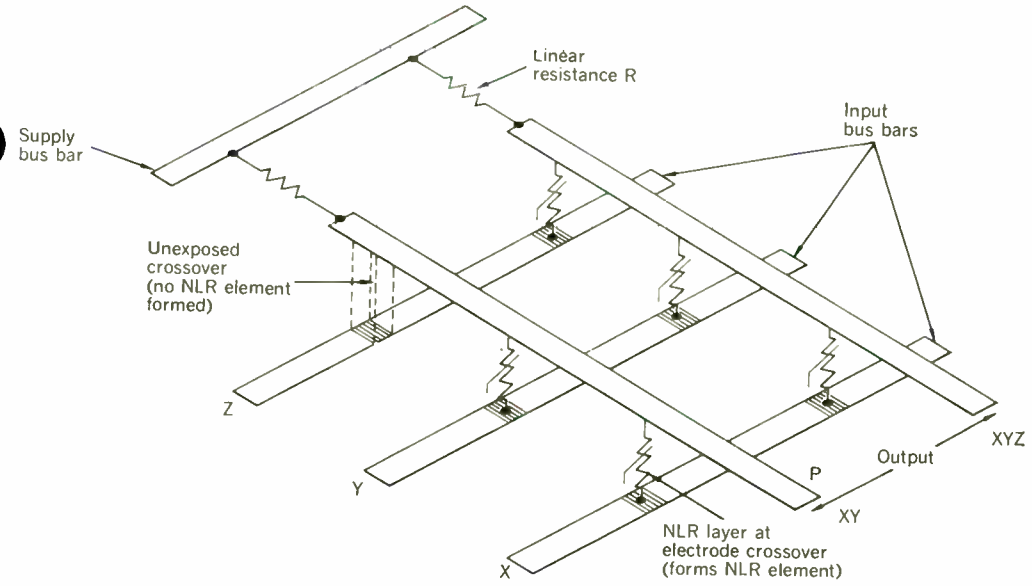
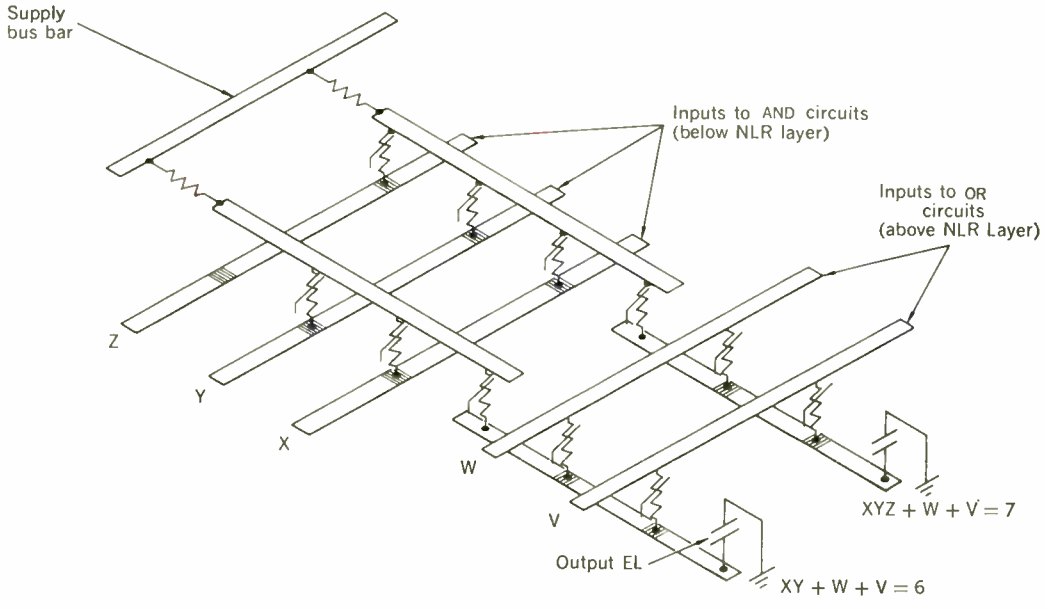


Fig. 22. Representation of a matrix of two NLR AND circuits.

Fig. 23. Setup for performing an AND-OR function for a five-bit input.



a source of 300 volts rms, 1 kc/s. The power required is approximately 2 watts, with the actual value dependent upon the input code.

Through the combination of available techniques of electroluminescence, photoconductors, nonlinear resistors, and piezoelectric materials and circuits, an almost limitless range of products may be manufactured. Many applications are unique and represent the only known way to obtain the desired display; others represent the first method devised to allow the construction of completely solid-state equipments with display features.

REFERENCES

1. Josephson, M., *The Invention of the Electric Light*. New York: McGraw-Hill Book Co., Inc., 1959.

2. Payne, E. C., Mager, E. L., and Jerome, C. W., "Electroluminescence A New Method of Producing Light," *Illum. Eng.*, vol. 45, Nov. 1950, pp. 688–693.

3. Ivey, H. F., *Electroluminescence and Related Effects* (Advances in Electronics and Electron Physics, Suppl. 1). New York: Academic Press, Inc., 1963.

4. Burns, L., "Alternative Explanation of the Waymouth-Bitter Experiments," *Phys. Rev.*, vol. 98, June 15, 1955, p. 1863.

5. Ivey, H. F., "Electroluminescence," *Sci. American*, vol. 197, Aug. 1957, pp. 40–47.

6. Strock, L. W., "Phosphors for Electroluminescent Lamps," *Illum. Eng.*, vol. 55, Jan. 1960, pp. 24–31.

7. Hallett, J., "Electroluminescent Displays," *Proc. 1st Nat'l Symp. on Information Display*, Society for Information Display, Los Angeles, Calif.

8. Greenberg, I., "Electroluminescent Display and Logic Devices," *Electron.*, vol. 34, Mar. 24, 1961, pp. 31–35.

9. O'Connell, J. A., Blank, H. G., and Wasserman, M. S., "Nonlinear Resistors Enhance Display-Panel Contrast," *Ibid.*, vol. 35, Aug. 3, 1962, pp. 33–36.

Electric heating—a 'hot line' of systems

Reports indicate that at the end of 1963 about 1.5 million homes were electrically heated. Projected estimates show that the number will increase to 6 million in 1970, and to about 19 million by 1980

Gordon D. Friedlander *Staff Writer*

Electric heating is the fast-growing baby of "great expectations" in a competitive field that long has been ruled by the fossil fuels. In two principal regional areas of the United States—the West and the South—its growth has been phenomenal since 1960. This is a tribute to the low-cost electric power available in these geographic sections. But electric heating is regarded as a prodigal infant in the Midwest and Northeast where electric power rates are generally too high to give its systems a competitive run for its money.

Usage patterns, geography, and power rates

Reference to Fig. 1 will show the comparative breakdown, by regional areas, as of January 1, 1964, of the homes that are electrically heated. Table I indicates the details of home heating rates by two methods of comparison. In the table, the utilities give an approximate rate per kilowatt-hour that applies to home heating across their system experience, rather than a complete rate schedule with steps of pricing based on total power consumption.

The rate variations by sections of the country show the Northeast to be the highest, and the next highest to be the Midwest. These are the regions where most of the electric power is furnished by private utility companies. The rates in this table are based upon 1963 prices.

The comparative annual costs indicate a high of $226 per year for the Northeast against a low of $98 per year for the South. The Electric Heating Association has computed the national average cost for electric heat per year to be $142 in 1963. This represents a $6 drop as compared to the 1962 survey.

An encouraging trend, however, is that an average of 31 per cent of the private companies, on a nation-wide basis, lowered their rates from a mean of 1.49¢ to 1.28¢ per kilowatt-hour.

Economic analysis and cost comparison

When fossil fuels were the only available source of energy, designers developed the central heating system, which, with flame fuels, was a great improvement over any other method. Generally, the central system used hot water, or steam pipes, or hot-air ducts to transmit heat to required areas.

The trend of design is now moving in another direction as electricity asserts its place in the sun as a superior

I. Cost to heat a home electrically in 1963

Section	Annual kWh Required for Heating	Public Power Companies		Private Power Companies		Average Degree Days per Year
		Heating Rate (cents per kWh)	Annual Cost	Heating Rate (cents per kWh)	Annual Cost	
Northeast	15 463	1.34	$207	1.46	$226	5588
South	12 386	0.79	98	1.25	155	2977
Midwest	16 114	1.30	209	1.33	214	6054
West	12 681	0.99	126	1.28	162	4141

(National average cost per year is $142.)

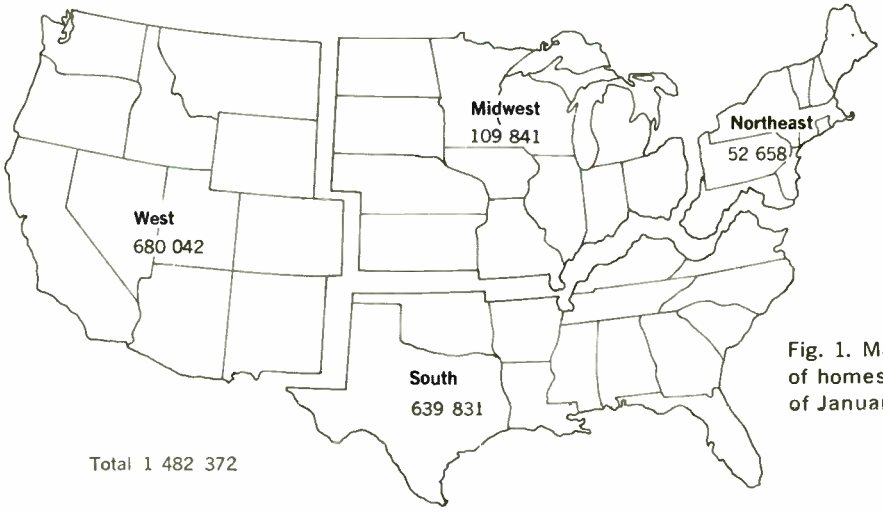


Fig. 1. Map shows the number of homes electrically heated as of January 1964.

energy source for providing heating and cooling. But it must be used properly to secure its maximum benefits.

As an example, when designers substitute an electric resistance heat source for a furnace that uses fossil fuel, many of the unique advantages of electric heat are lost, while the disadvantages of central heating systems are retained.

This situation is best illustrated by listing the principal advantages of an electric heating system over a fossil-fuel central heating system. (In this example we will assume that the electric system has individual units on each floor, either in zones or compartmented sections.)

1. The initial cost of installation is usually lower for electric heating, since this characteristic is inherent in the design of the electric system.

2. The annual energy cost for electric heating, in comparison with a fossil-fuel central system, depends on the relative incremental cost of electricity and fossil fuels as projected into the future.

3. Reliability is increased. In a central system, any trouble with the heating unit affects the entire building. If there is an electric heating unit for each floor, zone, or section, it is unlikely that more than one unit will be inoperative at any time.

4. Repairs are made faster and easier—and with less inconvenience for the building occupants.

5. The necessity for hot and cold water pipes, and associated pumps throughout the building, is eliminated.

6. The cost of chimneys, boilers, and associated motors and controls is eliminated, and the space usually occupied for this equipment and construction is available for other purposes.

7. Heating and cooling losses by transmission do not exist in an electric system, and stand-by heating losses are greatly reduced below those of a conventional central system.

II. Safety and economy factors

Electricity	Natural Gas	Oil, Low-Pressure Gas, Coal
Flameless	Flame	Flame
No storage	No storage, except for stand-by fuel when on interruptible gas rates	Storage required
No products of combustion	Products of combustion	Products of combustion
No chimney or vents	Chimney required	Chimney required
Located any place	Fire hazard precautions	Fire hazard precautions
Clean	Possible smudging	Possible smudging
Instantly available	Instantly available	Deliveries may be delayed
No carbon monoxide	Possibility of carbon monoxide	Possibility of carbon monoxide
Prices stable	Prices sharply increased in last few years	Prices sharply increased in last few years
High annual overall efficiency	Relatively low annual overall efficiency	Relatively low annual overall efficiency
No explosion	Explosion possible	Explosion possible, even for coal
Choice of central or localized installation and control	Choice is restricted in better systems to central installation and control	Choice is restricted in better systems to central installation and control

Table II lists some of the many safety and economic advantages of electric heating as compared to competitive fossil-fuel systems such as natural gas, oil, water gas, and coal.

Types of electric heating systems

To understand the fast-growing popularity and versatility of electric heating, it is necessary to review the broad range of electric heating equipment that is available for home, commercial building, and industrial applications.

Essentially, the available types include baseboard heaters, radiant heating systems, infrared heat sources, hydronic systems, wall insert heaters, central furnaces, heat pumps, ceiling cable, and the latest development—integrated environmental systems.

Baseboard heaters. In theory, baseboard heaters heat the air, which, in turn, heats the room, its objects and occupants. They place the heat where it is needed most—on the outside wall. Warm air rises to blanket the outside wall (see Fig. 2) with heat and eliminates cold drafts from the windows and walls. The rising warm air creates a gentle circulation of air in the room by natural convection, and maintains the temperature of the space nearly constant from floor to ceiling.

Baseboard heaters have practical application in residential quarters, restaurants, office buildings, hotels, and motels. These types of heaters should not be used, however, where there is a requirement for highly directional radiant heat, and the rapid heating of occasionally heated spaces such as warehouses.

Baseboard heater elements may be obtained in a wide variety of types and configurations. The most common heating element consists of closely spaced metallic fins that are carried on an element rod; see Fig. 3(A). One available model features swept-back fins to produce the high operating temperatures shown in Fig. 3(B). In another model, shown in Fig. 3(C), cells provide a large heat-exchange surface area, and the entire element assembly slides on rails to eliminate the usual expansion and contraction noises. The Fig. 3(D) elements consist of cast aluminum fins, with multiple embedded element rods to furnish large surface area and low surface operating temperatures.

Baseboard heater sections can be readily obtained in lengths of 1 to 10 feet, with ratings of 100 to 3000 watts.

In appraising the favorable operating characteristics of a baseboard heater, the "watt density"—the concentration of heat per unit area of element—in watts per square inch, is the most reliable unit of measurement. Baseboard heater accessories include inside and outside corner sections, thermostat modules, blank sections, and "heat-to-cool" selector switch modules that are designed to allow window air conditioners to be powered from baseboard receptacles.

Fig. 2. Diagram indicates natural convection heat flow from typical baseboard heater system.

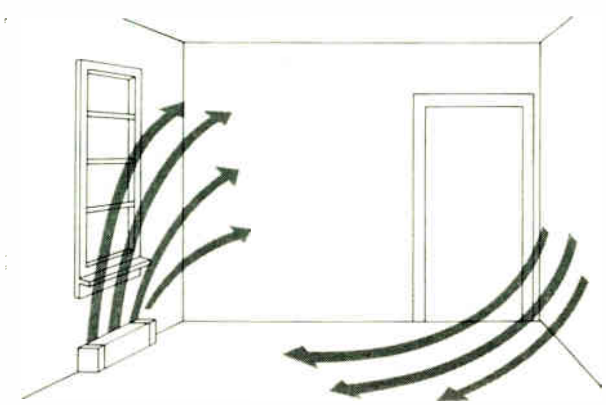
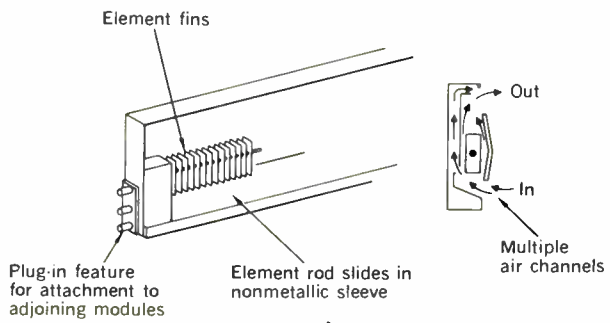
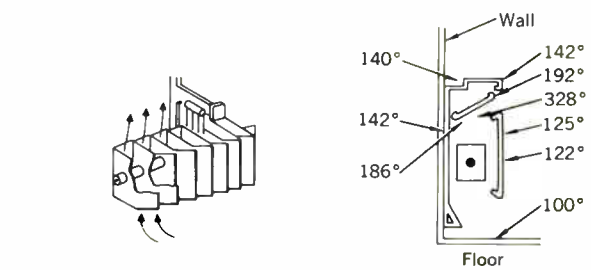


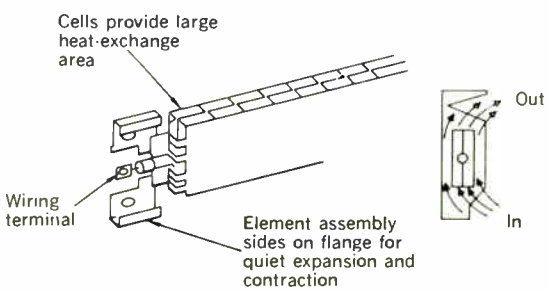
Fig. 3. A—Isometric sketch shows elements of a standard baseboard heater. B—Special type of baseboard heater features swept-back fins to increase convection flow and produce the temperature differentials indicated. C—Diagram shows a cellular type of baseboard unit. D—Multiple element rods and aluminum fins provide large surface area and low surface temperatures.



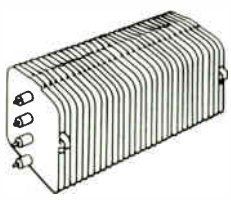
A



B



C



D

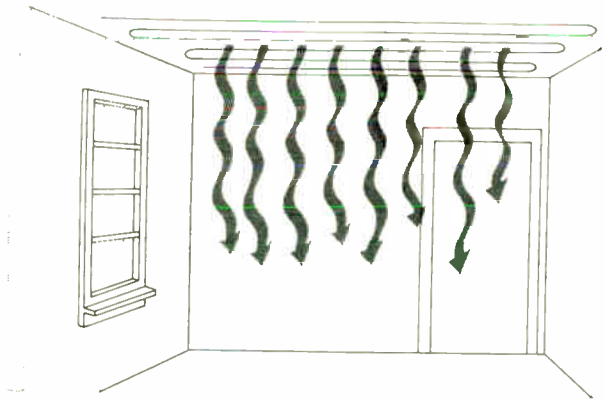


Fig. 4. Sketch indicates principle of electric heating by radiation (ceiling cable), in which occupants, and also the room furnishings, are directly heated.

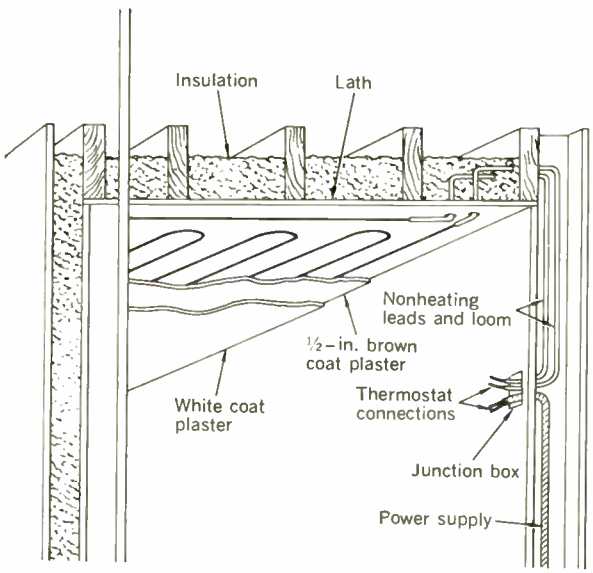


Fig. 5. Cutaway section of a typical radiant ceiling cable residential installation shows the required elements.

The principal advantages of baseboard heaters are gentle, even heat; quick response to outside changes in weather conditions; ease of installation (surface mounting does not require structural changes to building); quiet operation; and minimum required maintenance.

Baseboard heaters are most effective if installed under each window, under the largest window of a room, or along outside walls.

Radiant heating systems. A great variety of heating systems and elements may be classified as radiant systems. This category encompasses radiant glass wall heaters, ceiling cable, floor heating mats, radiant panel ceiling heaters, and ceiling radiant heaters.

The principle of radiant heating by ceiling cable is best illustrated by Fig. 4. Typical installation diagrams are shown in Figs. 5 through 7.

In this system, heat is radiated from all parts of the ceiling, and heats the walls, floor, and all objects in the room—including the occupants—directly and evenly. Ceiling cable is adaptable for continuously heated space applications similar to those for baseboard heaters. Since the cable is embedded in the ceiling, it is an ideal installation either for special commercial or industrial situations where unobstructed walls are required, or for architectural and esthetic reasons.

The first step in calculating the cable spacing for a given space is to divide the room into two equal parts (see Fig. 6), so that the dividing line is parallel to the longest outside wall. As indicated in the diagram, the half adjacent to the outside wall is the "close spacing section," and the remaining half of the room is the "open spacing section." With this method, more cable—and thus more heat—will be distributed in the section adjoining the outside wall, where the heat loss is greater.

In the next step, it is necessary to

1. Calculate the usable ceiling area of the room in square feet.
2. Determine the heat loss of the room by manufacturers' charts.
3. Divide the wattage of the cable by the usable ceiling area to obtain the required watts per square foot as indicated by the formula

Fig. 6. Diagrammatic plan view shows typical spacing layout for residential ceiling cable in a bedroom.

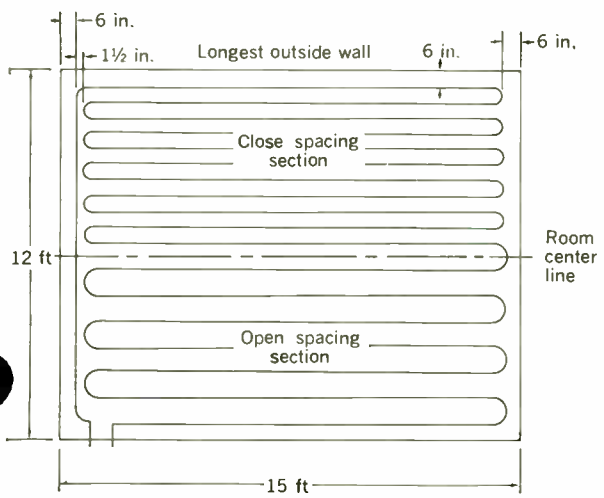
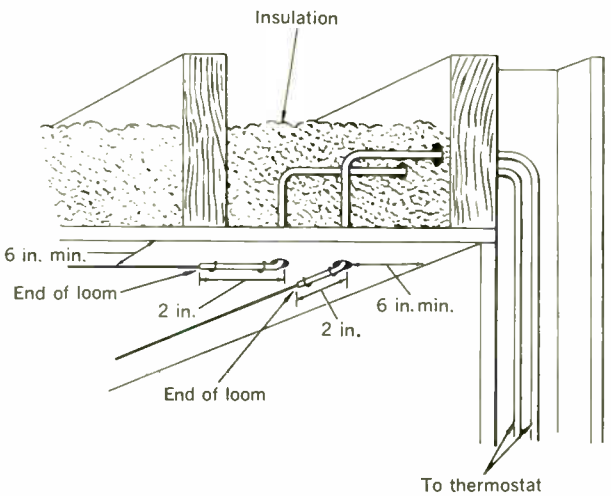


Fig. 7. Enlarged detail of Fig. 5 indicates minimum spacings required by the National Electric Code.



$$\frac{\text{Cable wattage}}{\text{Usable ceiling area (ft}^2\text{)}} = \text{Watts/ft}^2$$

When this value has been obtained, it may be located on the horizontal scale of the cable spacing chart (Fig. 8). Follow this numerical value up the ordinate scale to the point where it intersects the curve—see "arrowed" example on graph—then read the spacing required (left) for the close spacing section. To determine the spacing required for the open spacing section, move to the right from the point of intersection and read off the value on the right-hand scale. All spacing values should be read to the nearest one-eighth inch.

The salient advantages of radiant-heating ceiling cable include its invisible installation, maintenance-free qualities, broadly distributed low-temperature heat source, and completely silent operation.

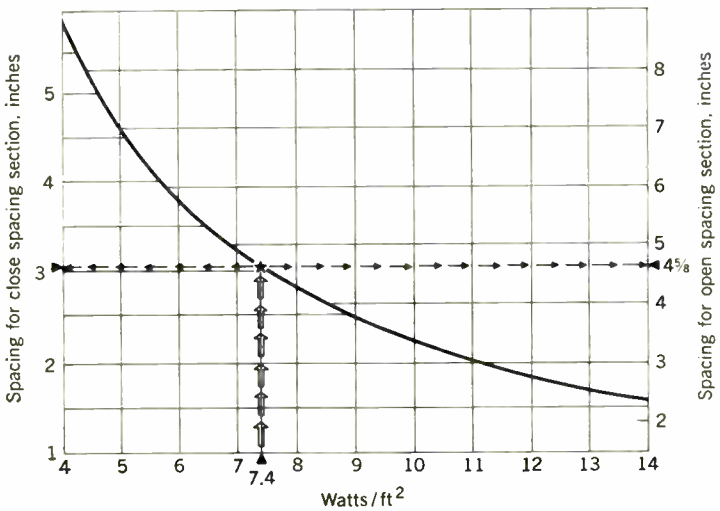
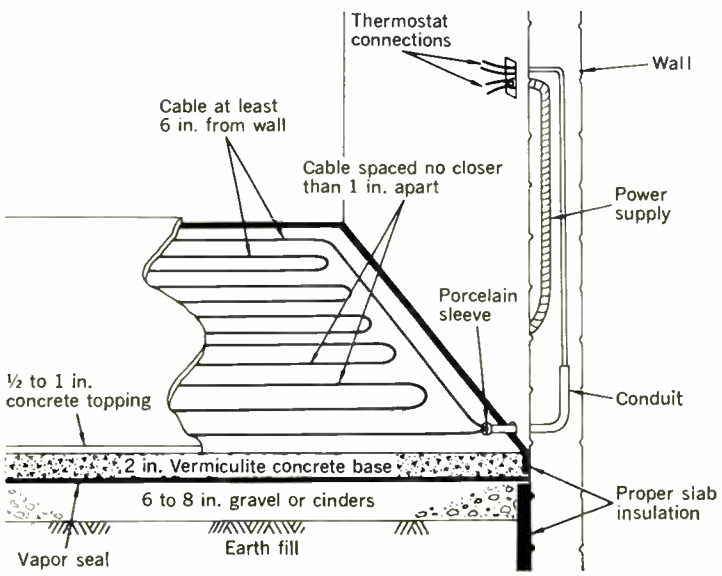


Fig. 8. Manufacturer's graph is used to determine required spacing of radiant ceiling cable for a residential space.

Fig. 9. Isometric diagram shows installation of heating cable embedded in a concrete grade slab.



Some of the disadvantages, however, are that a layer of cold air may form on the floor near large windows, and the cable must bring the mass of ceiling plaster up to room temperature before it can heat the space. Heating the ceiling plaster causes a time delay. Such a large heat sink may be ideal where the outside temperature does not vary greatly or rapidly during a normal heating day, but this condition would not be advantageous in regions of extreme outside temperature variations.

Cable in concrete. A variation of the ceiling cable application is the installation of the heating cable directly in a concrete slab floor. This method is particularly effective for one-story residential, commercial, or industrial structures where the cable is embedded in the grade slab. Figure 9 shows a typical installation.

One inch of rigid waterproof perimeter insulation between the edge of the slab and the foundation is required, and a two-inch thickness of slab insulation, carried down to a minimum depth of two feet, should be set between the foundation wall and the gravel fill under the concrete slab. National Electric Code (NEC) requirements stipulate that the embedded cable be spaced no closer than one inch apart; that there be a one-inch minimum spacing between the heating cable and any other metallic materials—such as reinforcing bars—embedded in the concrete floor slab; that nonmetallic frames or spreaders secure the cable loops; that insulating protection be used at points where nonheating leads emerge from the floor; and that excess nonheating leads be buried in the concrete.

Gypsum drywall. One of the most interesting innovations in the electric heating product line is the gypsum drywall. This material, with heating cable elements preinstalled, is available in standard modular sizes (multiples of four feet) for standard residential joist spacing. The panels may be handled and installed in the same manner as standard gypsum plaster board. And the board is available in decorator colors and textures to satisfy most architectural and esthetic requirements.

Usually, the gypsum board layout is made on the floor plan, and the various sizes required to meet the anticipated heat loss are used. Blank $^5/_8$-inch-thick board (without heating elements) is utilized to fill areas, such as closets, etc., where no heat is desired. Side and end trim boards contain heating wire in only two thirds of the board area to permit the remainder to be cut for lighting fixtures or air-conditioning louvers.

Each gypsum board usually has an 8-foot minimum pigtail (nonheating lead) of No. 14, two-conductor, NM plastic building wire, which exists two inches from the lateral edge of the board, and is taped to the board so that it does not interfere with installation. The electrician

III. Gypsum drywall heating panel specifications

Width and Length in Feet	Trim Area in Inches	Watts	Volts	Weight in Lbs
4 by 4	none	235	240	40
4 by 8	3⅜ end	450	240	80
4 by 8	32 end	330	240	80
4 by 8	15 side	330	240	80
4 by 12	3⅜ end	700	240	120
4 by 12	15 side	450	240	120

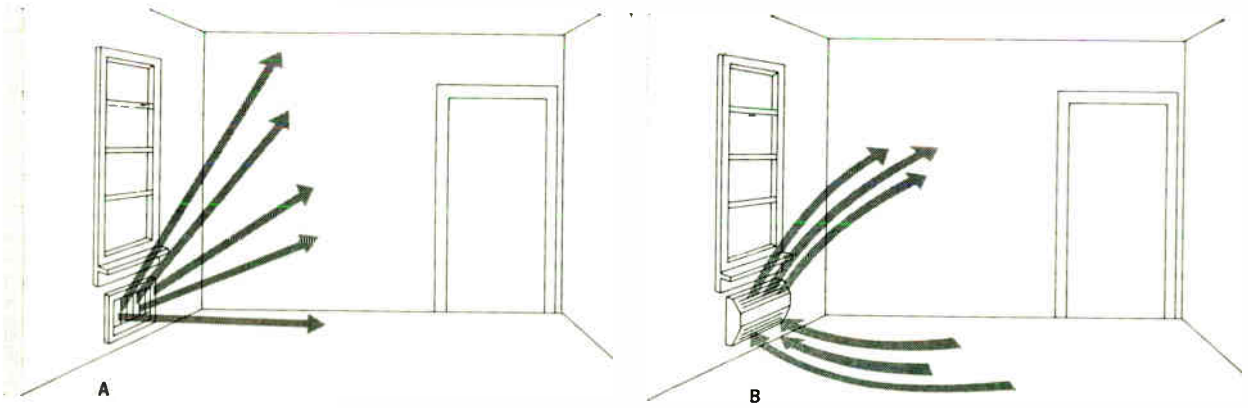


Fig. 10. A—Sketch shows principle of heating by concentrated radiation and some convection that is typical for wall insert heaters. B—Variation of this heat flow is produced by fan-forced draft in a wall heater.

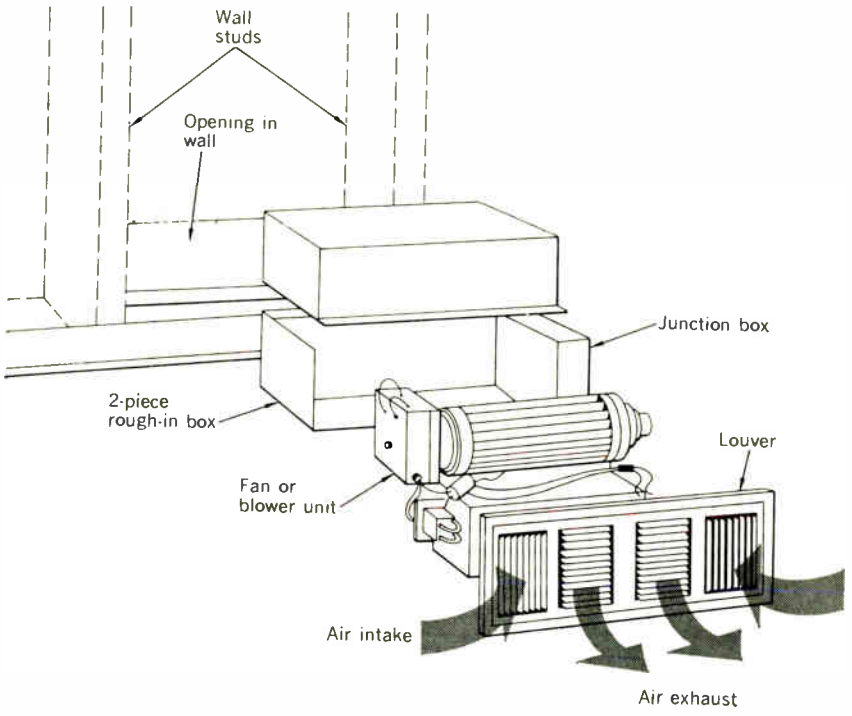


Fig. 11. Exploded view of a typical fan-forced wall heater for residential installation.

makes the hookup in parallel with other boards to the branch circuit and thermostat in a standard 4-inch junction box. In the case of multiple-story buildings, these boxes are located either in the baseboard of the room above, or in a closet on the same floor of the installation.

Table III indicates the physical specifications and electrical requirements for typical gypsum drywall heating panels.

Radiant wall heaters. The basic principle of radiant wall heaters is illustrated in Fig. 10(A). In this system, the *primary heat* is radiated toward the room in a relatively narrow heat pattern, and it produces the definite sensation of being heated upon the occupants of the space. The *secondary heat* comes from the warm air convection currents that also rise from the heater in a rather narrow pattern. This type of heater is best suited for quick space heating where air changes are frequent, in mild climates where overcoming morning and evening chill is important, and where usage does not demand a high level of comfort. Typical applications in which there would be frequent air changes include service stations, vestibules, entrance ways, and such structures as small aircraft hangers. Medium level of comfort installations would be in basements and garages, workshops, enclosed porches, and laundry rooms.

For maximum efficiency, this type of heater should be installed at least 6 inches above the floor, and 12 inches away from any side wall. Also, care should be taken to ensure that the heater does not directly radiate on furniture, draperies, etc.

An interesting new variation of the radiant wall heater is the fan-forced wall heater concept, shown in Fig. 10(B), in which air is drawn into the heater, passed over a high-capacity heater element, and then discharged back into the room by a motor-driven fan or blower. The blower produces a positive circulation of warm air in the room. Figure 11 shows an exploded diagram of the components of a typical fan-forced wall insert heater, and the structural conditions necessary for its installation. The equipment illustrated may be inserted between the studs of an existing wall through an opening only about 4 inches high. The grille itself, measuring only about 5 inches in height, will be the only visible compo-

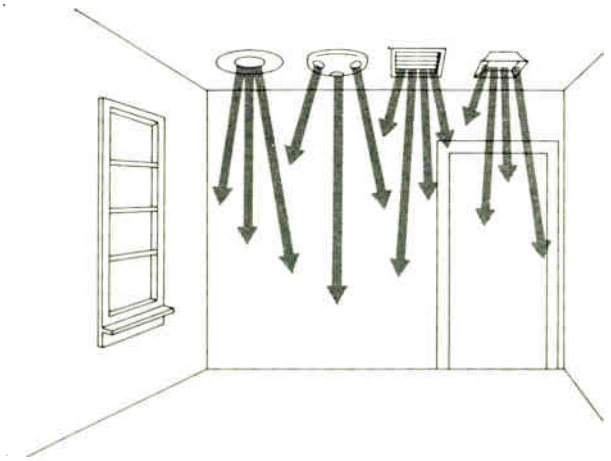


Fig. 12. Sketch shows various types of ceiling radiant heaters, some of which are combined as lighting units.

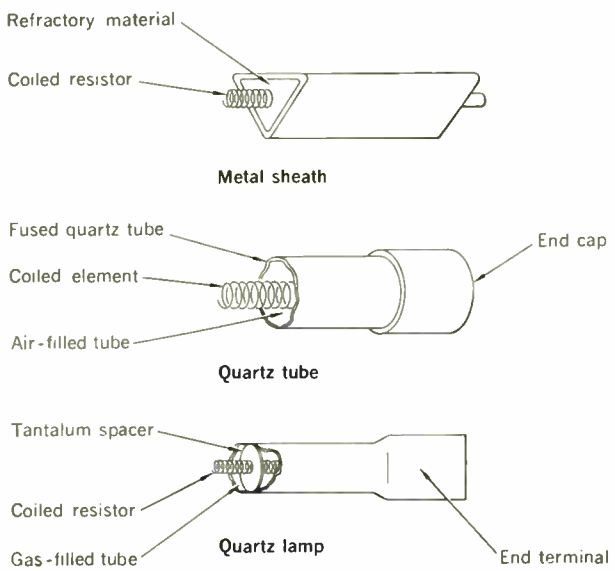
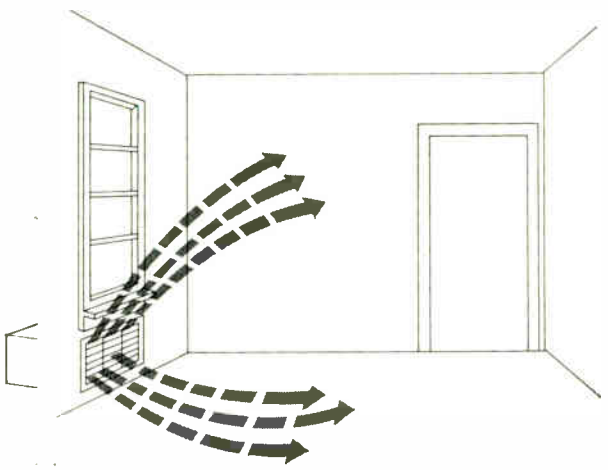


Fig. 13. Elements of three types of ceiling insert radiant heaters, one of which may be used as a combined heating and lighting unit.

Fig. 14. Diagram indicates the air movement produced for heating and cooling by the heat pump principle.



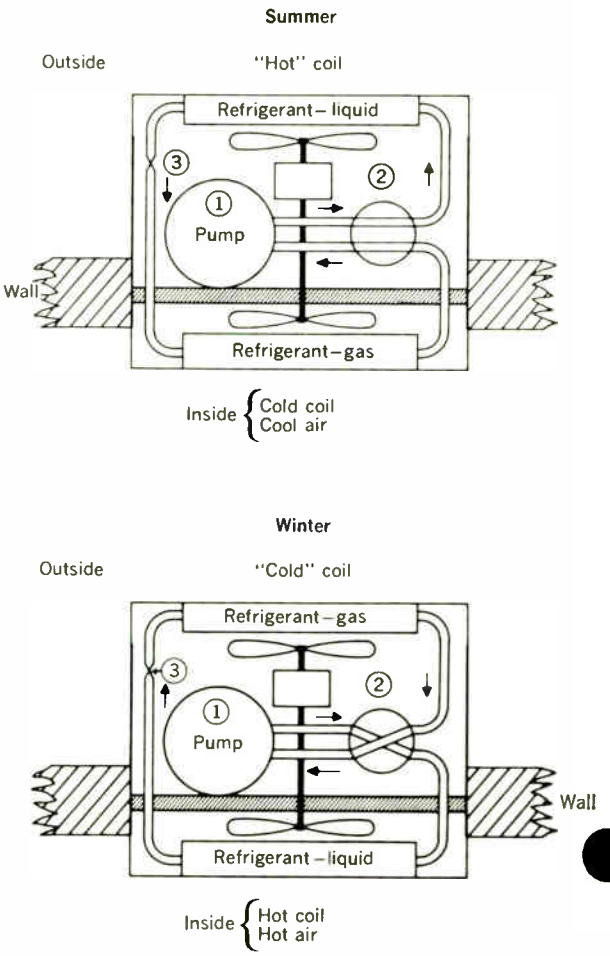
nent of the unit after installation. When these units are used in outside walls adjoining picture windows, wall streaking is eliminated, and window drapes may be carried down full length. An extra feature on some units is a removable 6-prong plug, which allows the interconnection of heating element modules up to a rating of 2000 watts.

Ceiling radiant heaters. The category of ceiling radiant heaters includes infrared bulb-type heaters, combination infrared heating–lighting units, quartz tubes, and quartz lamps. The method of heating is depicted in Fig. 12. Heat is radiated from the heater, directly heating people and objects in the room.

The infrared bulb heater, operating in the range of 250 to 750 watts, has a highly directional pattern of radiated heat that can be concentrated like a spotlight. And, as the clear bulb also produces light, the directional pattern and dual purpose make bulb heaters ideal for dressing rooms and bathrooms.

Resistance element heaters, such as the metal sheath, quartz tube, and quartz lamp (Fig. 13), operate in the range of 200 to 4000 watts, and are ideal for the localized heating of an area in an unheated space—a portion of a warehouse, garage, or pump house. The recessed combination heaters (Fig. 12) may have a built-in exhaust blower in addition to high-output fluorescent lamps.

Fig. 15. Simplified sketches show heat pump flows during summer and winter operation.



Heating and cooling by heat pumps

Heat pumps have two salient advantages over resistance heaters. As they operate on a reversing cycle principle similar to the kitchen refrigerator, they can perform the dual function of heating and cooling with the same equipment—and, if conditions for operation are good, the installation can achieve distinct cost economies.

As may be seen in the Fig. 14 diagram of the principle of a packaged heat pump, air is drawn into the unit by a blower. It is then filtered, heated or cooled, and discharged into the room. Referring to Fig. 15, we see that, in summer, the pump (or compressor) at (1) compresses the gaseous refrigerant into a hot, high-pressure gas which is cooled in the hot coil. This causes the gas to condense into a hot, high-pressure liquid that releases heat to the outside. The liquid refrigerant is then forced through an orifice (3), and expands into a gas. The expansion makes the refrigerant cold. And, as the refrigerant absorbs heat from the inside, the cold refrigerant is actually warmed in the cold coil. To complete the cycle, the gaseous refrigerant is returned to the pump to be compressed again. Thus, in summer operation, the inside cold coil absorbs heat, while the outside hot coil gives off heat, and the heat is "pumped" from the inside to the outside.

By switching the reversing valve (2) in winter (the control knob from COOL to HEAT), the flow through the coils and orifice (3) is reversed. Therefore, the heat flow is reversed, and is pumped from the outside to inside.

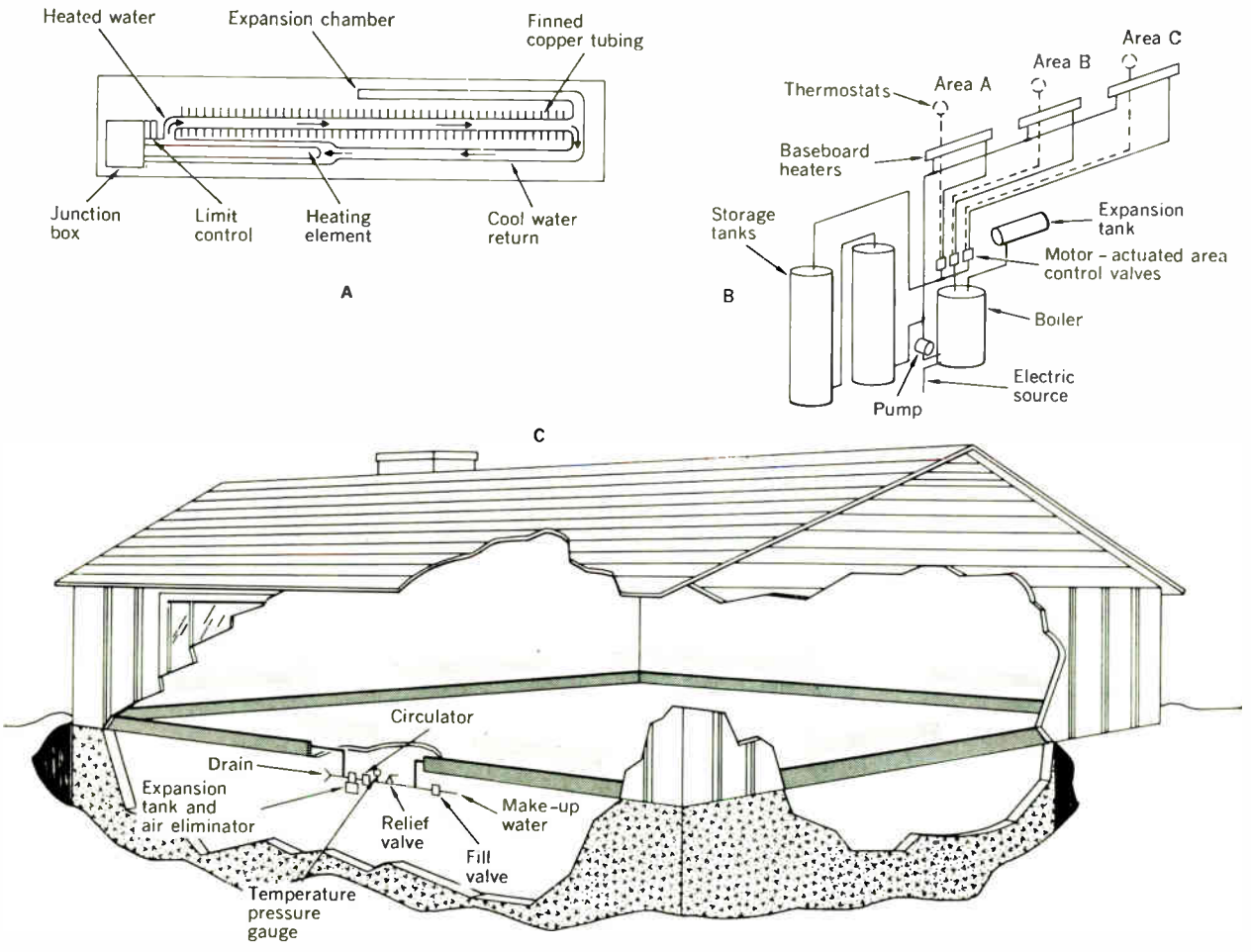
The air-to-air heat pump is usually the least expensive equipment in terms of initial cost, and it is also the most commonly used, since air is always available. The equipment may vary from small, packaged room units to large custom-built units for commercial use.

Air-to-air packaged heat pumps are recommended for full wintertime heating when normal January mean minimum outside temperature is greater than 15°F, and the installation is not in a sleeping room that must be heated during the night. Heat pumps can also be applied effectively when there is a requirement for occasional short-duration heating or supplemental heat. This would include restaurants, shops, etc., which operate only in the spring, summer, and fall months; homes in climates where the primary usage of the heat pump would be to overcome morning chill; rooms or areas which need heat in the autumn before the regular heating system is turned on. Heat pumps are used more frequently in commercial buildings, where cooling is a requirement.

Hydronic electric heating systems

Hydronic electric heating is available in a number of systems, each with a different principle of operation.

Fig. 16. Diagrams of three basic hydronic electric heating systems. A—Individual baseboard heaters, B—Electric boiler central system, and C—Continuous hot water loop system.



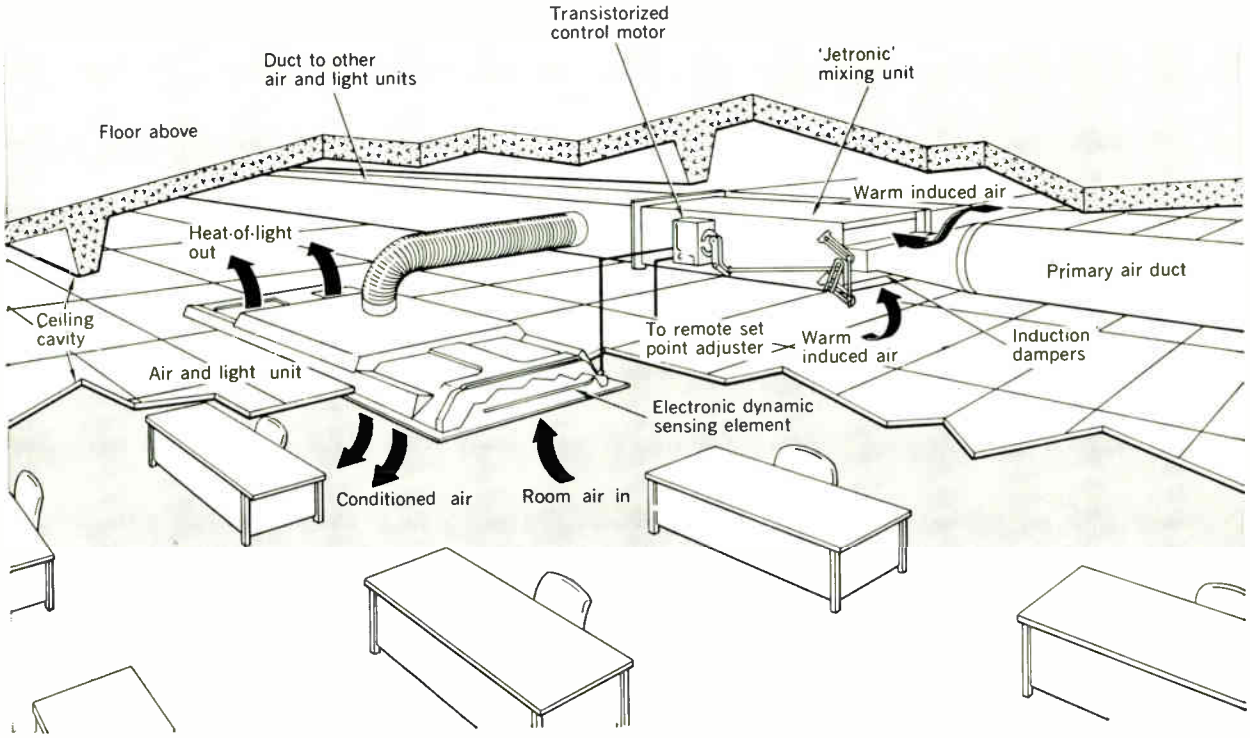


Fig. 17. Pictorial diagram shows one possible arrangement of a heat-of-light system for an office installation.

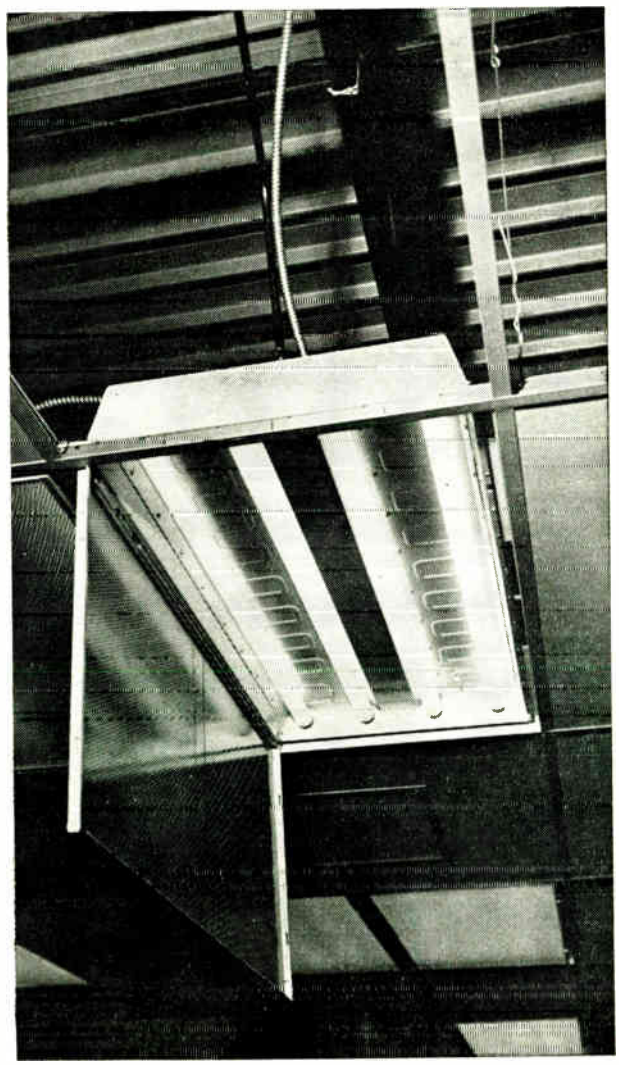


Fig. 18. A water-cooled luminaire is one of four basic components of an integrated environmental system.

Individual baseboard heaters. This system consists of completely independent baseboard heaters as shown in Fig. 16(A). An electric element located inside the copper tubing heats the permanently sealed-in water and antifreeze solution to exactly the temperature needed to balance the temperature of cold air that enters the room from windows and outside walls. The heating of the water causes it to circulate upward through the finned copper tubing where heat is removed by air circulating over the fins. The water then recirculates back to the electric element in a continuous cycle. As the water and antifreeze solution is hermetically sealed, the system never requires refilling. An expansion chamber allows for the expansion of the heated water, and an electric limit control automatically shuts off the unit if the water becomes overheated.

Electric boiler central system. Figure 16(B) shows an electric boiler central heating system with two water storage tanks. In this system, water is heated and stored during off-peak hours where advantageous power rates are available. Automatic controls limit capacity during peak load periods when higher power rates are effective. Area temperature control allows various sections of the building to be maintained at different temperature levels. This central system is actually an elaborate extension

of the individual baseboard heater principle, and capacities up to 40 kW may be attained by increasing the size and number of immersion elements.

Continuous loop system. This is a system that employs a single continuous loop of hot water, with baseboard heating elements installed along the entire inside perimeter of the house; see Fig. 16(C). Note thecom pact simplicity of the system, in which there is no bulky heating plant to occupy valuable floor space.

The central furnace system with heat storage

One of the most recent developments in electric heating systems employs a central furnace in which a chemical compound is cycled through a temperature range which includes its melting point. This process takes advantage of the *heat of fusion* principle,[1] and it uses sodium hydroxide, modified with corrosion inhibitors, as a heat storage medium. The heat storage unit, or furnace, is operated in conjunction with a conventional hot-air heating system. The overall dimensions of the compact central unit are approximately 55 inches in height by 48 in width and 41 in depth. The unit is guaranteed for a trouble-free life expectancy of 20 years.

'Heat-of-light' systems

Until quite recently, lighting and air conditioning were considered to be separate engineering design problems.[2] And even though the dissipation of the heat caused by lighting has always been a part of air-conditioning cooling requirements, it did not become a major problem until lighting levels of 200 foot-candles or more became generally accepted in offices and drafting rooms. The heat that results from such high illumination levels often places demands upon the air-conditioning systems that cannot be met by conventional methods.

When the heat-of-light system is used, the foot-candles can be more than doubled without increasing the amount of air needed for cooling. Figure 17 shows a typical heat-of-light equipment arrangement for an office floor space.

For example, if a fixture is lighted to 50 fc, and requires 100 ft^3/min for cooling, 35 ft^3/min is for lighting and 65 ft^3/min is for the heat input from people, office machines, and other extra-system sources. When the lighting level is raised to 100 fc, a total of 135 ft^3/min is needed—70 for lighting and 65 for people and other sources.

By using the heat-of-light system, 65 per cent of the heat from lighting is excluded from the office space, and the original 70 ft^3/min is reduced to 20 ft^3/min. When this amount is added to the 65 ft^3/min required for people and machines, the total air requirement is only 85 ft^3/min. Thus, the heat-of-light system makes it possible to double the lighting level and still use less air than a conventional system.

The heat transfer fixtures deposit large amounts of warm air in the suspended ceiling cavity, and this air may be up to 15°F warmer than room air. It can be used readily for zone tempering and reheat purposes without the necessity for reheat coils, piping, or the hot duct of a dual duct system.

The Jetronic mixing unit (see Fig. 17) is installed in the ceiling cavity directly above the zone to be served. This piece of equipment was developed by the Barber-Colman Company to utilize the warm air cavity by the application of the induction principle. A single cold air supply duct is all that is required to induce the warm cavity air for tempering purposes as needed for comfort conditions in the occupied space. The Jetronic delivers a constant volume of air, either cold or a mixture of cold and warm induced air. The number and size of the Jetronic units needed is determined by the flexibility requirements of the building. The units utilize only a maximum of 50 per cent of the return air for induction purposes, and leave at least 50 per cent of the warm air cavity available for return to the main system, where it can be used to heat partially or totally the exterior of the building during the winter.

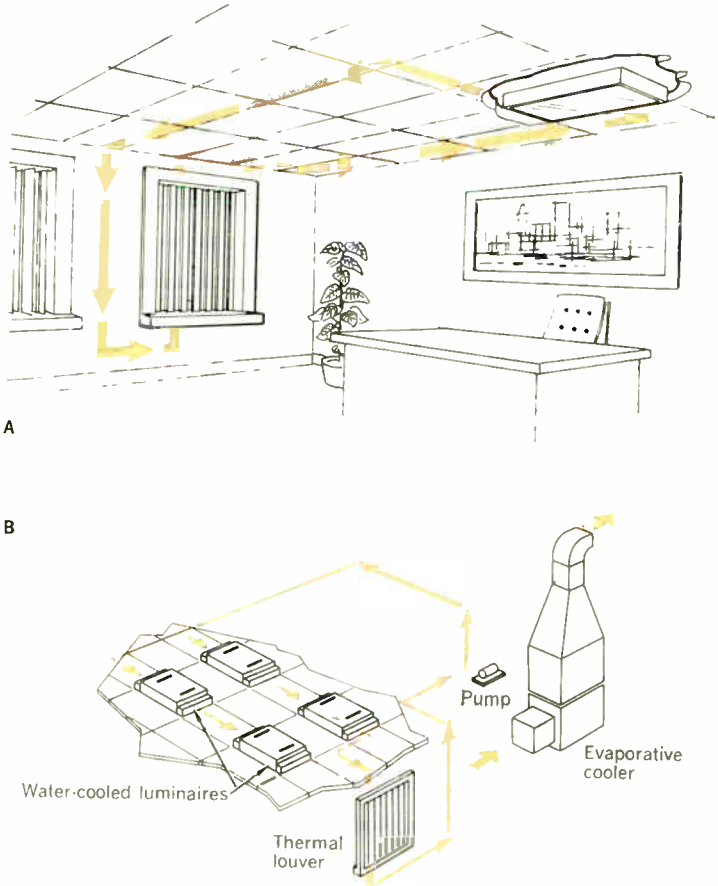
As building designs and types vary, as well as climatic conditions, lighting levels, and other factors, some buildings can be heated and cooled by an all-electric system, while others will need fossil fuels as an auxiliary heat source.

Integrated environmental systems

An interesting variation of the heat-of-light system wherein an air plenum and air ducts are utilized, is the all-electric environmental system[3,4] which uses non-refrigerated water in conjunction with thermal louvers and the lighting fixtures for integrating and balancing the lighting, heating, and cooling systems.

In simplified form, the system encompasses four basic

Fig. 19. A—Flow diagram of the winter operation of an integrated environmental system. B—Winter operation is shown in more complex diagrammatic form, in which four luminaires are operated in tandem.



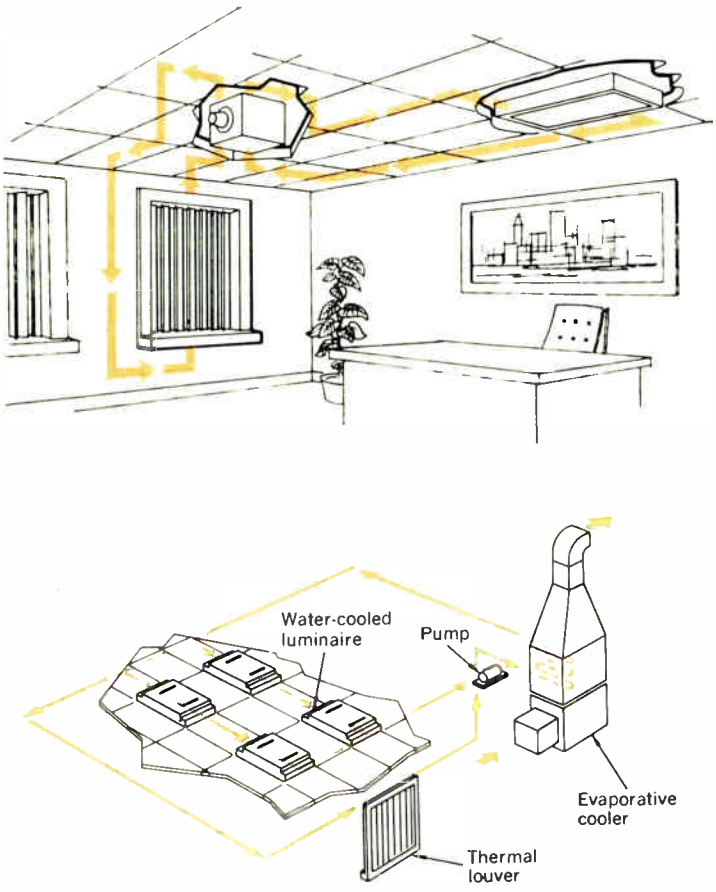


Fig. 20. A—Flow diagram indicates summer operation hook up of the Fig. 19 system. B—Note that an evaporative cooler is connected to the circuit for summer cooling.

components: water-cooled luminaires (Fig. 18), water-cooled thermal louvers that are similar to vertical Venetian blinds at the windows, an evaporative cooler, and a circulatory system that connects these elements with appropriate automatic-operating control valves. A fifth element, in the form of a supplementary heat pump, is usually required to adapt the basic system to specific climatic or operating conditions.

Figure 19(A) shows the winter operation of the system in which the water circuit picks up the heat from the lighting fixtures and delivers it to the louvers to offset heat losses at the fenestration area. Figure 19(B) indicates this circuit in a more complex diagrammatic form in which four luminaires are operated in tandem. During the summer, the water circuit, as shown in Fig. 20(A), picks up the heat from the lighting fixtures and from the thermal louvers (which now intercept solar heat gains at the fenestration area), and circulates it through an evaporative cooler to dissipate the heat. In addition to their primary purpose, the thermal louvers help to control the level of lighting, and, as they are controlled by a photoelectric cell within the lighted space, the vanes automatically change position as the amount of outside light varies with the position of the sun, passing clouds, and other natural occurrences.

The operating water temperature level in the thermal louvers is sufficient to offset transmission heat losses

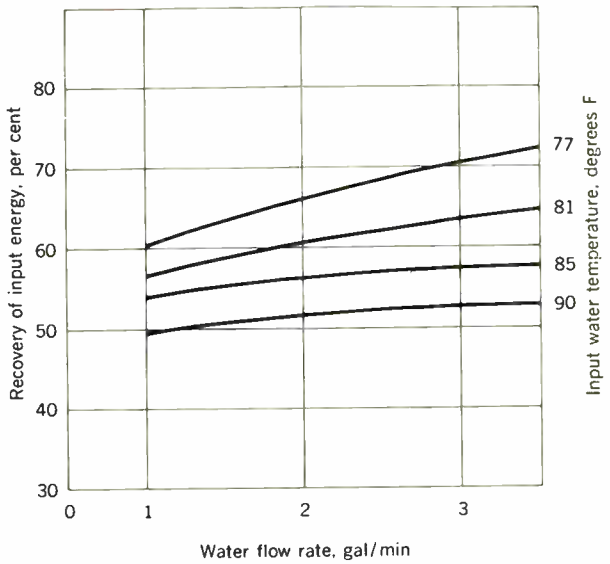


Fig. 21. Graph plots the thermal performance of a water-cooled luminaire.

through the perimeter fenestration areas during winter operation, and to remove the solar heat gains during the summer. Thus, relatively neutral temperatures are maintained at the wall.

From the viewpoint of the lighting fixture operation, the 77°F (average temperature) water serves to remove heat gains generated by the lamp sources and ballast. The heated water may be cooled in the evaporative cooler, or may be supplied to the thermal louvers if there is a need for heat at the perimeter of the system. The luminaires and the louvers are automatically connected in series during the winter—as in Fig. 20(B)—and are connected in parallel to the evaporative cooler during summer operation.

The total heating need of exterior office space is provided by the lighting system through the redistribution of interior heat gain to the perimeter thermal louvers and tempered air supply. The total heating requirement of interior offices is furnished directly by the lighting system.

Figure 21 illustrates the thermal performance of a water-cooled luminaire. The testing was performed in a specially constructed test room, and the top of the luminaire was insulated to maximize the energy transferred to the circulating water.

The 'living effects laboratory'

The Emerson Electric Company has established in St. Louis, Mo., an Electric Living Effects Laboratory that is unique in design and construction. The roof is suspended from rigid-frame steel bents so that the interior and exterior walls and partitions can be relocated, removed, or replaced. This flexibility permits changes in building materials, dimensional changes, and advancements in construction methods as they are developed. Thus testing conditions can be set up as required for effect studies.

Test room no. 1 (Fig. 22 floor plan) projects into a vault, or plenum, in which the "artificial" outside tem-

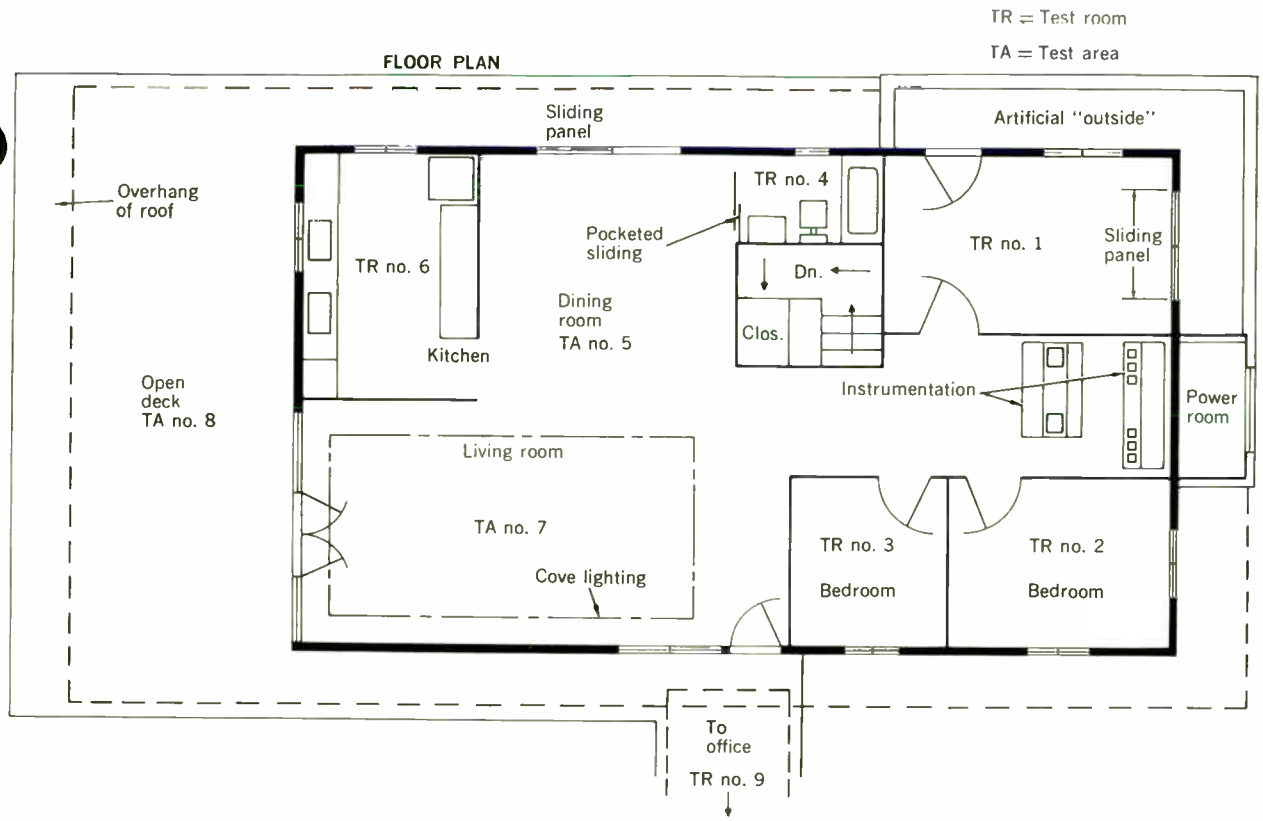
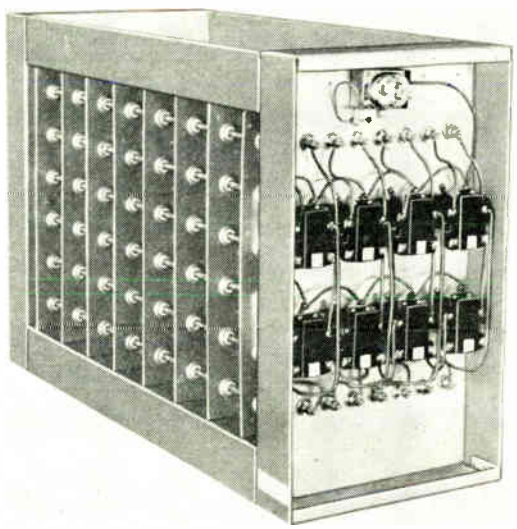


Fig. 22. Architectural floor plan shows testing rooms and areas in living effects laboratory.

Fig. 23. Blast coil heater, consisting of resistance coils carried by metal frame, is for industrial and commercial installations.



perature is controllable. Thus, the air surrounding the room on two sides—as well as above and below—can be chilled or heated to simulate the effects of winter or summer as desired. Much of the investigation conducted in this room deals with the effectiveness of electric heating. Electric baseboard heat undergoes extensive study to increase its usefulness as a principal heat source. And comparative analyses are made to determine methods of offsetting heat losses, caused by large fenestration areas, through the effective use of electric heat.

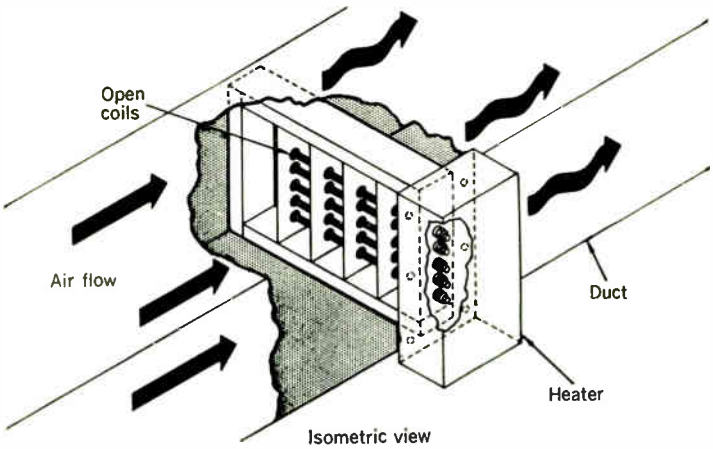
A large console in the instrumentation area functions as the control center of the laboratory. Through 12 test stations in the house and one on the deck, 24 different temperature readings can be recorded simultaneously, and up to 144 different test data can be monitored at the instrument panel.

The effects of humidity on comfort are evaluated in test room no. 2. Electrically operated devices are studied to determine economical and efficient methods of dehumidification.

Industrial space heating systems

Thus far we have primarily discussed residential and commercial space heating systems. Industrial space heating, however, encompasses a specialized area of space heating problems, equipment, and systems.

Historically, the advent of the modern electric duct heater for industrial space heating applications occurred about 20 years ago.[5] At that time electric space heat was used only for highly specialized purposes. A few units of this type were included with package air conditioners. The heaters were either used for reheat or for total heat, and the usual practice was to put several electric strip heaters in a frame near the outlet of the cooling apparatus, or in a duct. In those days thermostats were occasionally used for air temperature control, but thermal cutouts, or similar safety devices, were unknown. It was not until a few years after the pioneer duct heaters were



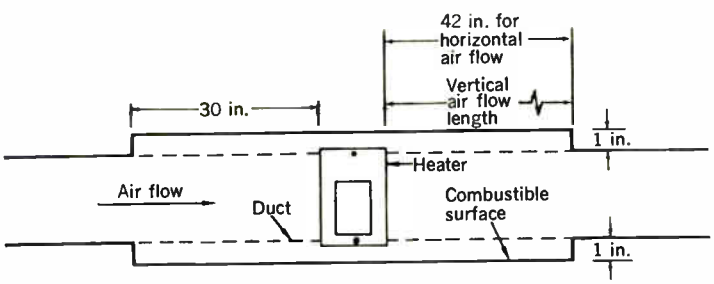
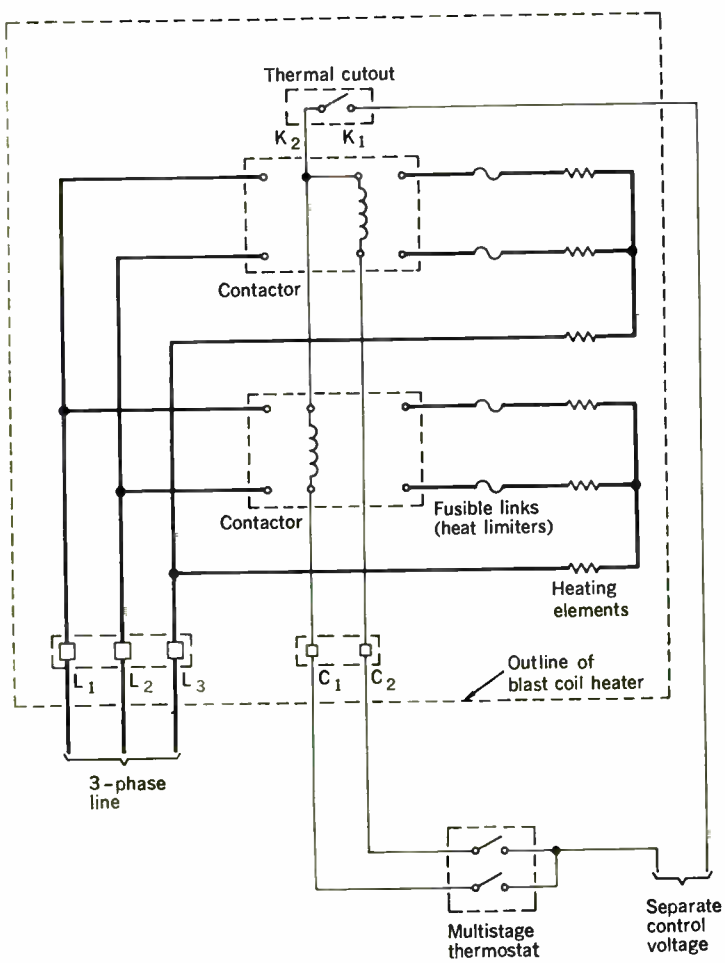


Fig. 24. Diagram shows typical installation of blast coil heater unit in an air-conditioning duct.

Fig. 25. Blast coil wiring diagram indicates the circuitry and safety features of the system.



produced that consideration was given to the potential hazards of an air failure during operation. Today, safety devices such as fusible links are readily available.

Blast coil heaters. In principle, the blast coil heater is merely the "big brother" of the residential fan-forced resistance heater, but the primary application of the blast coil heater is for large commercial and industrial installations. The heater (Fig. 23) consists of resistance coils which are located and insulated in metal frames. With this construction, the coil itself is exposed to the air stream when the unit is installed in the duct (see Fig. 24). Also, the exposed coil permits lower coil operating temperatures, and, hence, assures longer life of the heater. This type of construction affords an immediate transfer of heat to the air stream when the current is turned on. And conversely, when the current is turned off, the heat stops immediately. Overshooting and undershooting of control instruments is reduced considerably, and an even temperature curve is the result. Based upon actual performance in the field of thousands of installations, it is the consensus among experts that, except where there is a possibility of someone touching the heater or where there is an explosion hazard, the open coil elements are preferable to the enclosed sheath. Figure 25 is a typical blast coil wiring diagram that indicates the circuitry, etc.

Although the blast coil type of space heater is the most widely used for large industrial and commercial applications, the pipe flange immersion heater has found considerable favor for heating boilers in hot water or conventional steam systems. These heaters are available in ratings up to 432 kW.

The overall evaluation

From all indications, electric space heating will command an increasing percentage of the total heating equipment market in the three principal areas—residential, commercial, and industrial installations. And as new developments in more economical and practical systems evolve, the rate of consumer acceptance will accelerate. It will be up to the public utilities in certain regional sections of the country, however, to do their share, through the establishment of preferential power rates, to make electric heating installations universally feasible and attractive.

The author wishes to acknowledge the courtesy of the Dallas Power & Light Co. in making available Figs. 19(A) and 20(A); the Illuminating Engineering Society for furnishing Figs. 18, 19(B), 20(B), and 21; the Emerson Electric Co. for Fig. 22; the Industrial Engineering & Equipment Co. for Figs. 23 through 25; and Paul C. Greiner, of the Electric Heating Association, Inc., for his cooperation in making background data available for this article.

REFERENCES

1. Bary, C. W., Rice, R. E., and Paquette, J. F., Jr., "Heat Storage for Electric House Heating," *IEEE Spectrum*, vol. 1, no. 4, Apr. 1964, pp. 109–112.

2. Darling, R. B., "Use—Don't Waste—Heat of Lighting," Barber-Colman Co., New York, N.Y.

3. Humphreville, T. N., Folsom, W. E., and Meckler, G., "Installation and Operation of Integrated All-Electric Total Energy Environmental Systems," paper presented at the Nat'l Tech. Conf. of the IES, Sept. 1964, Miami, Fla.; scheduled to be published in *Illum. Eng.*

4. Tatum, C. A., Jr., "Electric Heating—a Market of Tremendous Proportions," *EEI Bulletin*, vol. 32, no. 6, July 1964, pp. 172–178.

5. Epstein, M., "The Design and Application of Electric Duct Heaters," *IEEE Conf. Paper CP63-1245.*

Instrumenting the sea floor

The oceanic abyss two to four miles below sea level is more forbidding than outer space. Nevertheless, experimentation is under way to explore and exploit this region, using both unmanned instruments and manned vehicles

R. A. Frosch *Advanced Research Projects Agency, Department of Defense*
Victor C. Anderson *University of California*
Hugh Bradner *University of California*

Introduction: some general problems

R. A. Frosch *Advanced Research Projects Agency, Department of Defense*

Although more than 70 per cent of the earth's crust is beneath the oceans, this portion of the earth is little known and scarcely used. Man has throughly explored most of the land areas of the earth—sometimes with great difficulty and at great risk—and has sailed for centuries on the surface of the oceans, but has yet to learn very much about that land directly beneath the keel of his ship.

Three facets of sea-floor exploration are here introduced: general features of the ocean bottom; some possible uses for the ocean bottom; and some techniques useful for the exploration and exploitation of this region of the earth.

Let us look at Fig. 1, which shows the so-called hypsographic ("hypso" is Greek for height) curve of the earth's solid surface. The left and right edges read depth and elevation below and above sea level in meters, while the bottom edge is in percentage of the earth's surface above a given depth. Approximately 30 per cent of the surface is above sea level. The mean depth of the solid sphere of the earth is some 2440 meters below sea level, and most of the bottom of the sea is to be found between 3000 and 6000 meters.

Topography of ocean floor

The ocean bottom may be divided into a number of separate provinces that differ from each other in certain important respects. These differences determine both the uses to which they lend themselves and the engineering problems of exploitation.

Many coasts are bordered by a continental shelf. Generally a flat, sedimentary region with slopes less than 1° from the horizontal, it may be up to 100, or more, miles wide. The west coast of the United States generally lacks a continental shelf, but both the Gulf and east coasts have splendid examples. In the New York area the distance from the beach to the hundred-fathom curve is about 100 miles. (As a fathom is 6 feet, 1000 fathoms equal about one nautical mile, corresponding to one degree of latitude.) This shelf region is generally covered with fine sand. Since the water on the continental shelf is relatively shallow, the pressure at the bottom of the shelf goes only from one atmosphere to 20 atmospheres, and some light filters through the water to the bottom.

At the edge of the continental shelf is the continental slope. The slope of the bottom changes abruptly from

less than 1° to an average of 3–6°, although in some places it is as high as 15 or 20°. This slope usually runs from a depth of 100 fathoms to 1500 fathoms. While the continental shelves are, on the whole, relatively smooth, the slopes are generally rather rough and deeply cut by numerous canyons. These canyons are found at the edge of the continental slope of the east coast, and run out nearly from the shore on the west coast.

At the foot of the continental slope is usually an abrupt turn into a gently sloping (about 1°, or less) region called the continental rise. This ordinarily runs from about 1500 fathoms to 2000 or 2500 fathoms and merges rather gently into the deep abyssal plains.

The deep plains are generally covered with organic sediment, and may be extremely flat and level for hundreds of miles. In some regions (such as south and southeast of Bermuda in the Atlantic), rather than a flat plain there is a flat region covered with gently rolling hills.

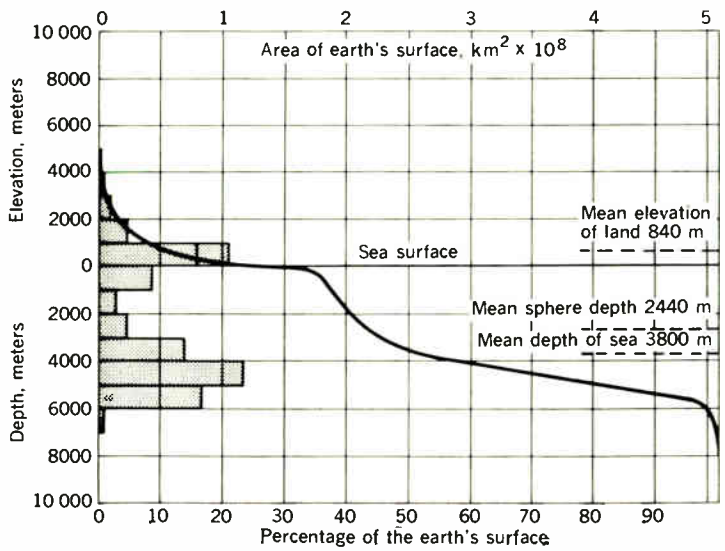
At a depth of 2000 fathoms, the pressure is 400 atmospheres, and no light filters down from the surface. It is not completely dark, however, since many deep organisms have luminescence of their own. The water temperature is near 0°C.

In many places, particularly surrounding island arcs such as the Antilles, Aleutians, Kurils, and Marianas, deep trenches are found in the sea floor. Less than 50 to 100 miles wide they form clefts as deep as 2000 fathoms, for a total depth of 4500 or more fathoms. The bottoms of the trenches are filled with sediment.

The earth is girdled by a system of underwater ridges through the centers of the major oceans. These narrow, steep ridges are frequently high enough to break the surface in islands such as the Azores group in the Atlantic. The ridges appear to have a central rift valley running down their middle.

In addition to these general features, there are numerous seamounts, ranging in size from small hills on the bottom to mountains that break the surface to form islands. They can occur as relatively isolated features in the abyssal plains or as chains of mountains, and, particularly in the Pacific, many are terminated by flat tops at depths of several hundred fathoms.

Fig. 1. Hypsographic curve showing the area of the earth's solid surface above any given level of elevation or depth. At the left is the frequency distribution of elevations and depths for a 1000-meter interval.



Knowledge about the features just described is derived almost entirely from observations made from the surface. It is only in the shallow portions of the continental shelves, and in a few isolated special cases, that anything corresponding to direct human observation has been possible. Depth information comes principally from acoustic echo sounders, operated routinely by many ships, which reflect sound pulses from the bottom. Although the resolution of these instruments has not been sufficient to give us fine-grained detail about the bottom, they have been satisfactory for study of many of the general features. Additional information has been provided by photographs made by cameras lowered to the sea floor. The precise locations being photographed generally are unknown because the camera, hanging at the end of a long cable, may be a mile or more away from the ship and is very difficult to control. Nevertheless, such photographs have been extremely illuminating with regard to conditions to be found on the sea floor. A certain amount of information also has come from bottom samplers and dredges, and perhaps even more from cores taken from the bottom by instruments lowered on cables.

We know that the bottom of the sea in general is covered with fine organic sediment, although the steeper slopes, the ridges and seamounts, are frequently rocky. In regions near the bottoms of slopes the ordered layering of the sediments is often disturbed by what appear to be the results of turbidity currents: giant slides in which a slurry of sediments flows down the slope carrying everything before it.

Uses of the ocean bottom

Simply because it is there the ocean deserves detailed exploration as a relatively unknown portion of the earth. It is not easily accessible, but requires study so that we may gain a more complete understanding of our home planet.

The geophysicist is, of course, interested in the rest of the earth's crust. He wants to know the structure that lies beneath the surface sediments and wants to study it with controlled seismic energy, and by means of measurements of magnetic and gravitational fields. This is all facilitated if the instruments are at or near the bottom.

The seismologist, who needs a good geometric distribution of stations to study earthquakes, is attracted to the bottom of the sea as a place to measure the motions of the earth. It is almost essential for detection of underground nuclear blasts, since many of the regions of the world (such as the Kamchatka-Kuril region) prone to earthquakes that might mimic nuclear tests are bordered by deep ocean. Good system geometries can only be achieved with ocean-bottom instruments.

The oceanography, biology, and chemistry of the bottom of the sea have scarcely been explored because of the difficulty of doing suitable sampling with techniques operated from the surface.

The most important uses for the bottom of the sea probably are still unsuspected, because we know so little about what is down there. This provides our major task of exploration. We want to know the detailed shapes

and textures of the bottom, the way in which the sub-bottom layers were formed and distributed, the ecology and the forms of life on and in the bottom. Many of these things are quite easy to investigate on land but are very difficult to study at the bottom of the ocean.

The sea floor provides a fixed place upon which to stand and place instruments. The normal experience at sea is continual motion with no fixed points available, and so this is an extremely valuable attribute. It suggests that the bottom would be useful for the emplacement of navigational equipment. Since the oceans are essentially opaque to virtually all forms of energy transmission except sound, that is the form of navigational information that can most easily be provided from bottom-fixed installations.

In addition to furnishing a floor on which to stand for navigational research equipment, the bottom may be used as a fixed anchoring point from which instruments and devices may be floated up to any point in the water or to the surface. This is, of course, the commonest use of the bottom—for anchoring ships and for the emplacement of ordinary shallow-water navigation buoys.

From a commercial point of view, the extraction of oil from pools beneath the bottom has received the most attention. Most of this work has been done on continental shelves and a good deal of the technique has differed little from drilling in marshy areas. As oil fields are discovered farther from the coast in deeper water, however, technology which truly exploits the ocean bottom is becoming more necessary. The most extreme experiments in this regard have been connected with the preliminary mohole drilling.

Very little mining has been attempted on the ocean floor although sulfur and salt are extracted in relatively shallow regions off the Gulf coast of the United States. One successful enterprise of this type is the dredging of diamond-bearing gravel from deep waters off the coast of South Africa. The material is apparently carried down by river out-flow from continental deposits. It has also been suggested that the manganese nodules that appear to litter the floor of the ocean in many places would be an attractive mining possibility. We have essentially no idea what else lies on or beneath the sedimentary floors of the ocean that might be of commercial value.

Aside from the shallow-water cultivation of the oyster, when man goes to sea it is as a hunter rather than a farmer. There has been discussion of the farming of other sea-bottom creatures—lobsters, crabs, and some forms of fish—but very little has been done about it. We have come to the point where technology will allow extensive enterprises of this sort. Although it is not yet clear that they would be commercially useful, the possibility deserves study and experimentation.

Techniques of exploration and exploitation

There is in marine science, as in space science, an argument between those who want to study the bottom with unmanned instruments and those who want to go and have a look themselves; but fortunately the disparity in costs will not be nearly as great. Almost all of the data now available have been taken with instruments hung from the surface, usually without any real control or knowledge of their exact position. Thus the sampling has had a large component of randomness, and it has been difficult or impossible to take samples of various parameters in a coordinated and correlated manner. Trying to make measurements at the bottom of the sea while standing on the deck of a ship can be rather like sitting in a balloon 10 000 feet up at night in a dense fog, while trying to learn something about the forest that is expected to be underneath by hanging instruments down on bell wire. With the use of more elaborate modern technology, it is becoming possible to operate multi-sensor packages at the ends of long cables, including observation with television and acoustic means of locating the instruments relative to the ship. This enables us to know what we are doing, but we have not yet achieved sufficient control of the instruments from the surface to be able to do precisely what we want to do on the bottom.

A possible way to improve the situation is to go down directly. The bathyscaph, Cousteau's saucer, and deep experimental submarines like *Alvin* make it possible for us to go down ourselves and observe, move around to a limited extent, and manipulate instruments. To be fair to those who prefer the cable route, it must be admitted that the man who goes down is confined to a steel chamber with limited vision ports and only indirect control of the tools and equipment outside of the sphere. Artificial light must be provided for him. It is frequently suggested that this amount of control of the environment could be achieved from the surface via cable or pipe. From my own experience I believe that once we really try to use the deep-diving submarine, we will find it much easier and cheaper to explore and use the bottom of the ocean in that way than to sit on the surface and fish for the information at long distance. I do not mean that the cable method will be discarded, but only that new knowledge can be gained more easily by those who go to the bottom.

Exploration and observation for general understanding and the search for new phenomena require a manned submersible. We cannot yet telemeter the observing qualities of a man, and as long as it is relatively easy for him to go himself he will do so. Nevertheless, once specific new measurements and observations have been chosen, many useful jobs may be done more easily from the surface. Eventually we will learn a suitable division of scientific and engineering tasks between equipment operated from the surface and that operated at the bottom.

The problems of exploration in the various provinces described at the beginning of the paper vary somewhat. The continental shelves are almost certainly completely accessible to the human diver, and recent experiments by U.S. Navy personnel and by Captain Cousteau and his group have demonstrated that it will be feasible to live at depths of several hundred feet for prolonged periods of time. This will make real exploration of the bottom in these depth ranges relatively easy.

The continental slope region presents considerably more difficulty because of the roughness and the frequently complicated current and water patterns to be found there. These factors may hinder the use of manned submersibles; they certainly have not simplified exploration of this region with cable-hung equipment.

The deep plains and the continental rise, which constitute the main area of the oceans, are relatively accessible to cable-hung equipment and probably also to practical deep submersibles; exploration of this

region will be complicated principally by the large areas involved and the difficulties of logistics.

The problems posed by the trench regions are perhaps somewhat greater for manned submersibles than for cable-hung equipment, although the dive of the bathyscaph *Trieste* to the deep trench off Guam and the recent dives of the *Archimede* off Puerto Rico demonstrate that this is not impossible. The problems posed by the ridges and seamounts arise largely because of the roughness of the terrain. As in the case of the continental slopes, considerable maneuverability on the part of a submersible, or excellent control of a cable, are required.

From an electrical and electronic engineering point of view, only a few characteristics of the sea-bottom environment indicate special problems or advantages. The great pressure can be a problem, but many modern electronic devices are unaffected by it, and simple measures such as pinholes for pressure equalization through transistor shields for components in oil bags are often effective. In any case, the technology of pressure cases is already good and is improving continually.

The presence of the conducting salt water itself can be a nuisance, since it provides an intrusive and corrosive surrounding medium. It also must be considered in design because of its effect on impedances to ground, etc.

One advantage of the bottom of the sea as an environment for electronic devices is the good heat-dissipation and constant-temperature-bath properties of sea water. Once the device is in place, the environment surrounding it may be expected to be constant, and the temperature unlikely to change by more than a small fraction of a degree.

Thus equipment designed for and tested under simulated ocean-bottom conditions is often extremely successful. Like any other branch of technology, specialized knowledge and experience are essential, but the path is being eased by the rapid accumulation of useful information.

Vehicles and stations for installation and maintenance of sea-floor equipment

Victor C. Anderson University of California

It is quite a blow to man's pride to admit that he will never be able to encounter the environment of more than half of this home planet's crust with anywhere near the intimacy with which he will some day experience outer space or even the crust of other planets. The oceanic abyss, lying from two to four miles below sea level, accounts for more than 50 per cent of the crustal area of the earth, and the hydrostatic pressure of up to 10 000 psi encountered at these depths creates a far more hostile environment than does the nearly perfect vacuum of outer space. Even if some day the hope generated by the wet lung experiments performed in the last few years on dogs gels into a capability that would permit man's body to be exposed directly to these extreme hydrostatic pressures, the seemingly mundane problem of maintaining body heat in the near-freezing deep ocean environment would be orders of magnitude more difficult than in the near absolute zero temperatures of outer space. At these high pressures, the efficient cellular insulation materials, which are available in a profusion of types, are ineffective, and the normally simple matter of interposing a thermal barrier between the human body and the surrounding water becomes a formidable problem.

Design considerations

Once we accept the fact that the human body is denied access to the environment, it is apparent that the engineering methods used for the installation and maintenance of equipment on land are not directly applicable to the sea floor.

The design of equipment for the sea floor of the oceanic abyss must take into account not only the fact of the exclusion of the human worker but also the hostility of the environment to the equipment itself. The equipment must operate at 10 000-psi ambient pressure, and at near-freezing ambient temperature. Ocean water with its high salt content (3½ per cent) is a corrosive medium and a rather good electric conductor. One of the major problems of sea-floor instrumentation is to maintain a suitable barrier between this corrosive conducting environment and sensitive electric circuits and delicate mechanisms. To this end, the selection of suitable corrosion-resistant structural materials and coatings for prolonged submergence is crucial. In addition to the salt content, the corrosion and fouling by marine life at great depths, although not a severe problem, must be considered.

Present-day technology in oceanographic work on the sea floor relies heavily on electric circuits protected from the environment by pressure cases. Scientists have, in general, contented themselves with cable-connected instrumentation lowered from a ship. However, recent work with instruments that can be dropped to the ocean floor and brought back by either a time-program or a call-up command system have shown good promise for carrying out extensive observations on the sea floor.

There is good reason to conjecture that the technology of the future will not be satisfied with this type of instrumentation but may require more intensive installations on the sea floor. The operations associated with undersea mining or the operation of nuclear power plants at great depths are the more striking examples of the types of intensive equipment installations that will require advances in underwater technology beyond its present state.

An encouraging note in the development of sea-floor equipment lies in the realization that the deep ocean is not a hostile environment for all of today's technology. Hydraulic systems, for example, can cope with operation in a 10 000-psi ambient environment. Many electrical components of themselves are essentially unaffected by ambient pressures of this magnitude. Several of the commercially available plastic materials are immune to both salt water and biological corrosion effects.

Careful selection of compatible components and ma-

terials permits the use of an oil bath instead of a pressure case as an environmental barrier. This has two beneficial results: (1) the increased reliability of the system because of the elimination of high-pressure seals (the prime source of leakage failures) and the extra protection afforded by the environmental oil barrier surrounding each component; (2) the elimination of pressure locks to transport material across the interface between the sea water and the protected environment. A free oil–water interface is stable and can be provided in any required configuration to meet the needs of material transfer. By permitting access to the equipment for servicing and maintenance at the installed depth through the use of an ambient-pressure environmental barrier, such as an oil bath, the way may open towards greater technological flexibility of deep-sea-floor instrumentation.

Implicit in a sea-floor installation designed to permit access to its interior equipment is an underwater vehicle that can perform the general functions of logistic support, observation, and manipulation. Logistic support implies the ability to transport material across the ship-to-vehicle interface at the surface, down from the surface to the sea floor, and finally across the vehicle-to-installation interface at the sea floor. The observation function provides an operator with the ability to identify objects in the environment, and determine their position and orientation with respect to their surroundings. Manipulation follows the observation function. It provides a means by which the operator may change the position and orientation of objects in the environment at will.

Level of present technology

The functions just defined, although perhaps deceptively oversimplifying the vehicle requirements by their generality, encompass three distinct aspects of the problem of support for an accessible ambient-pressure deep-ocean installation. What follows is a review of the extent to which the existing vehicle technology meets the requirements of an accessible installation with respect to these functions.

A variety of underwater vehicles have appeared on the scene over the past few years. They may be characterized by features such as depth capability or maneuverability; they may be manned or unmanned, free falling or free swimming, cable tethered or cable supported, battery or cable powered. The vehicles to be described do not constitute a complete list, but they do cover the broad spectrum of types that have been developed and thus rather clearly delineate the state of the art.

Webster defines a vehicle as "that in or on which a person or thing is or may be carried." We are concerned here in the broadest sense with vehicles that will carry persons or things to and from the sea bottom and the surface, and also vehicles that will serve the limited range functions of observation and manipulation.

Among the simplest vehicles in the context of this definition are the free-falling instrumentation packages that are being used for measurements on the sea floor. These represent a class of simple, unmanned, nonmaneuverable, free-falling vehicles.

The spherical seismometer package of Bradner and Snodgrass at Scripps is an example of this type of vehicle. The spherical shell forms a pressure case which protects the electronic and sensitive mechanical instrumentation enclosed within it, and at the same time provides sufficient positive buoyancy for motive power in recovery. An anchor weight is released when the package is to be recovered.

Another example, also from Scripps, is the magnetometer package of Vacquier. This free-falling instrument package houses a recording magnetometer. It uses soluble salt blocks for disposable ballast, which may be seen clustered near the base, and gasoline-filled rubber storage drums for buoyancy. A cylindrical pressure case is incorporated for the instrumentation. Both of these vehicles use negative and positive buoyancy for motive power. Only a very simple acoustic communication and program-control link connects them with the surface. This type of vehicle may have application for logistic support of a sea-floor installation if sufficient dropping accuracy can be achieved, and provided an additional support vehicle is available at the sea bottom. The vehicle is attractive by virtue of its simplicity; however, it lacks the communication link and mobility required to perform the functions of observation and manipulation. In larger versions, in a deep sea barge configuration, vehicles of this class may find application in the transportation of ore in sea-floor mining operations.

The family of small submersibles, which has appeared on the scene in recent years, represents another class of vehicle. These are manned, free-swimming vehicles with various mobility and depth capabilities. This class of vehicle has received the greatest amount of attention from both the public and the scientific community, perhaps because a manned vehicle having a deep submergence capability will take man physically close to the deep sea floor and thus satisfy his exploratory longings to the fullest extent possible.

The *Soucoupe* designed by Costeau and operated by Westinghouse is an example of a moderate depth submersible. It has an operating depth of 1000 feet with a design safety factor of 2.9. It carries a crew of two, cruising at a speed of 1 knot for a range of 3 miles. Its overall weight is 7700 pounds including batteries and safety ballast. The *Soucoupe* carried out a series of dives in the submarine canyons off La Jolla in the spring of 1964.

A photograph of the bottom with the shadow of the *Soucoupe's* sampling arm poised for a sample is shown in Fig. 2. This illustrates the rudimentary capability of manipulation existing in this vehicle. It should be noted that sophisticated manipulation from a free-swimming vehicle of this type is severely complicated by the motion of the vehicle.

Further complications arise when the observation capability is degraded by sea-floor operations. For example, a sediment cloud stirred up by the sampling arm reduced visual observation to virtually zero. As a result of the prevalence of this type of disturbance, short-range high-resolution sonar equipment may well be a required augmentation of the observation capability of any vehicle used for manipulation on the sea floor.

A small submersible of greater depth capability is illustrated by Woods Hole's new two-man submersible *Alvin* (Fig. 3). Operating depth of the vehicle is 6000 feet with a design safety factor of 1.8. Its maximum design speed is 6 to 8 knots with a 10-hour endurance at 2.5 knots. It provides life support of 24 hours for the two-man crew. The total instrumentation payload is

Fig. 2. Photo of the sea bottom with the shadow of sampling arm poised for a sample.

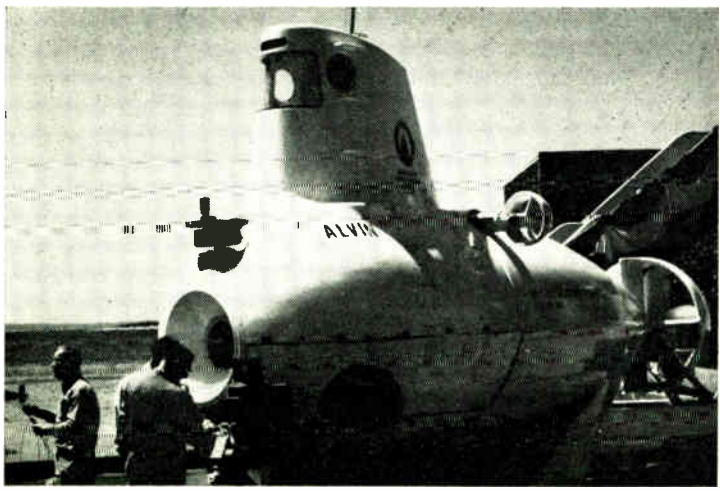


Fig. 3. Woods Holes' new two-man submersible Alvin.

1200 pounds, with 18 kW of electric power provided for scientific purposes other than propulsion and life support. The interior diameter of the steel pressure hull is 6½ feet and the vehicle weighs 10 tons in air. It will be outfitted with sonar, echo sounder, underwater telephone, underwater television, and a manipulator arm, in addition to cameras and miscellaneous recorders and instruments. *Alvin* is presently undergoing tests at Woods Hole, Mass., and will be carrying out research dives in the near future.

The most recent arrival on the scene of deep submersibles is Reynolds *Aluminaut*, recently launched at the General Dynamics Electric Boat Division dock in Groton, Conn. The *Aluminaut* is constructed of forged aluminum cylindrical sections, which are precision-machined and bolted together.

The interior view of the hull shown in Fig. 4 was taken during outfitting, and shows the bolted cylindrical section construction. The design depth of the *Aluminaut* is 15 000 feet. It carries a normal crew of three—one pilot and two scientists. With a design cruising range of 80 miles at a speed of 3.8 knots, its gross displacement is 75 tons, considerably larger than that of *Alvin*. Outfitting will include basically the same features as those of *Alvin*.

The bathyscaphes represent another class of deep submersibles—free-falling vehicles which use buoyancy as the primary motive power. They have limited mobility because of the large size of the gasoline flotation hull. An illustration of this class is the bathyscaphe *USS Trieste* operated by the U.S. Navy Electronics Laboratory (Fig. 5). The *Trieste* has made a dive to the deepest part of the ocean, to the bottom of the Marianas trench (35 800 feet). It has a range of 2 to 3 miles at a speed of 2 knots, and carries a normal complement of two, although three were used in the *USS Thresher* search last year. A new salt-water hydraulic manipulator arm was added in a recent conversion. The manipulator

Fig. 4. Interior view of hull of Reynolds Aluminaut.

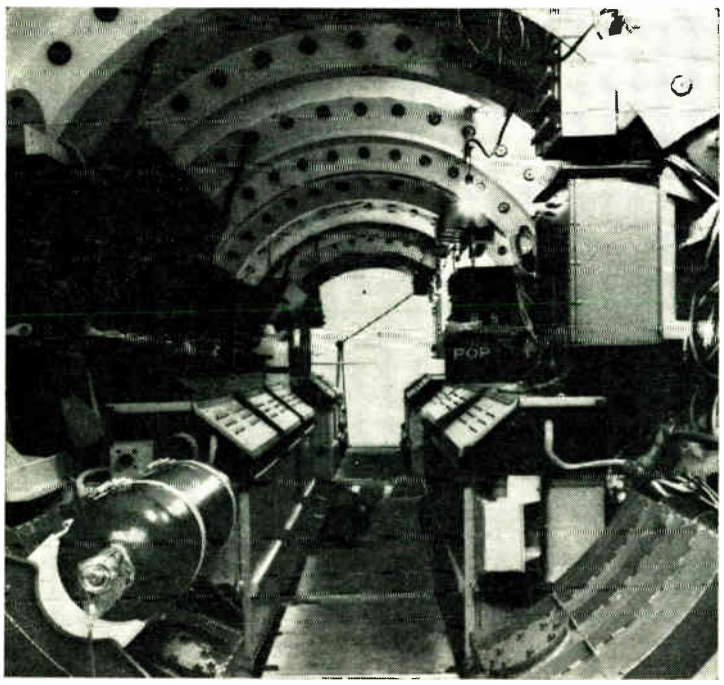
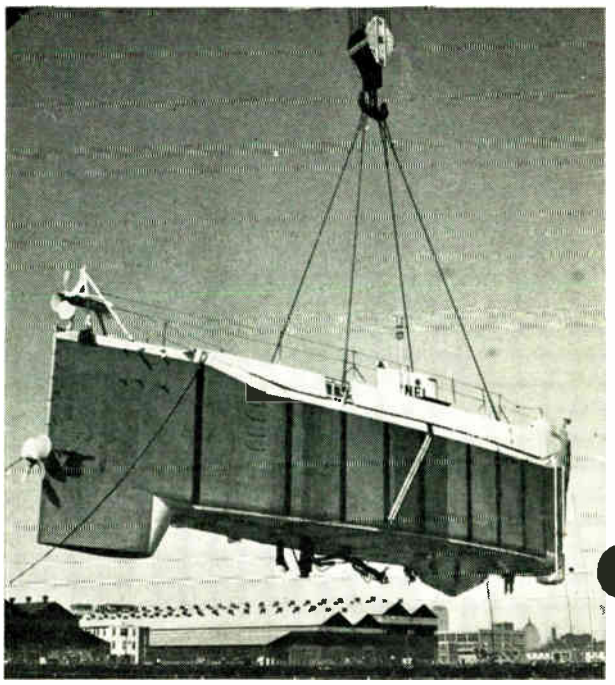


Fig. 5. The bathyscaphe Trieste operated by the U.S. Navy.



is shown in Fig. 6 in the retracted position, most of the arm being withdrawn into the gasoline float.

These last-named vehicles, the free-swimming manned submersibles, can fulfill all three of the required functions. However, as all suffer from a finite endurance as limited by the life support systems and battery power, they may be more applicable for occasional maintenance to an accessible sea-floor installation than for more extended tasks associated with construction of the installation.

Cable-supported manned vehicles have also been used. The bathysphere in which Beebee and Barton descended to a depth of 3028 feet in 1934 is most famous of this type. This type of equipment is commercially available today from an Italian firm. It is used primarily for observation in salvage work. These observation vehicles do not suffer from the power limitation of the battery-operated submersibles but they do suffer from lack of mobility and thus have only limited application.

A more versatile vehicle is the cable-connected bell-shaped device. This vehicle, operated by Conrad Industries of Long Beach, Calif., possesses motive power by virtue of powered tracks. It has portholes for observation and possesses an elementary manipulator capability. Power and air are supplied from the surface, and thus there is no inherent limit to endurance other than operator fatigue. The maximum operating depth is 1000 feet.

In all of these vehicles, the environmental barrier protecting the operator consists of a heavy pressure hull. Another genus of vehicle is represented by the remotely operated vehicles where the environmental barrier protecting the operator becomes the column of water between the vehicle and the surface. In these vehicles the complications of life support and human safety associated with the pressure hull compartment is exchanged for a more complex communication link which can effectively telemeter the operator's senses and commands from the surface station to the vehicles. This type of vehicle also has been developed in a wide variety of shapes.

The next step up in the vehicle spectrum is again illustrated by a Scripps instrument, the Marine Physical Laboratory's Deep Tow (Fig. 7). Here, mobility is achieved by towing with a length of armored electric cable. Acoustic depth-sounding information is telemetered back up the armored cable.

Several search and recovery vehicles, operating over short tether cables, have been developed for the Navy. These, in general, incorporate propulsion motors, television and sonar observation equipment, and some grappling capability for recovery operations. They represent a class of vehicles that may have particular application as auxiliary support vehicles used primarily for observation in the vicinity of a sea-floor installation where only a limited range capability of a few hundred feet would be required. The limited range capability imposed by the tethering cable can be augmented by terminating a larger electric strain cable with an anchor clump and tethering the vehicle to the clump.

Rounding out the family of remotely operated vehicles is the class of track layers of which the Marine Physical Laboratory's Remote Underwater Manipulator (RUM) Fig. 8 is an example. This experimental version was operated over a 5-mile length of ½-inch-diameter coaxial cable in 1960. This type of bottom-oriented vehicle

Fig. 6. New salt-water hydraulic manipulator arm installed in the Trieste in recent conversion.

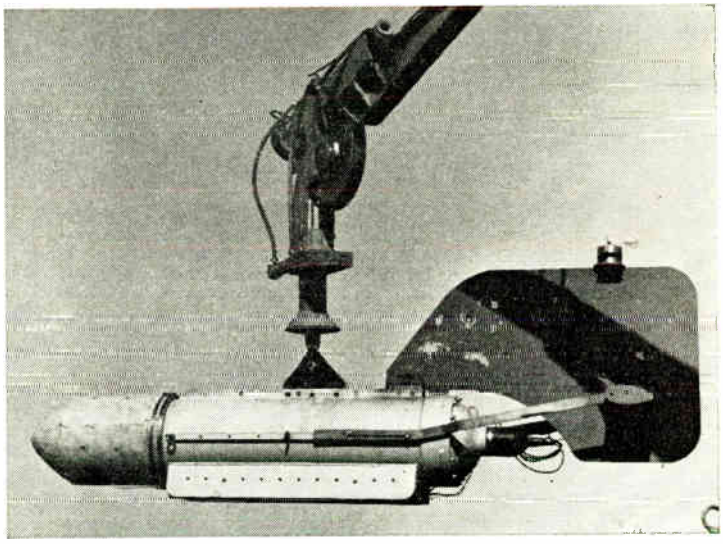
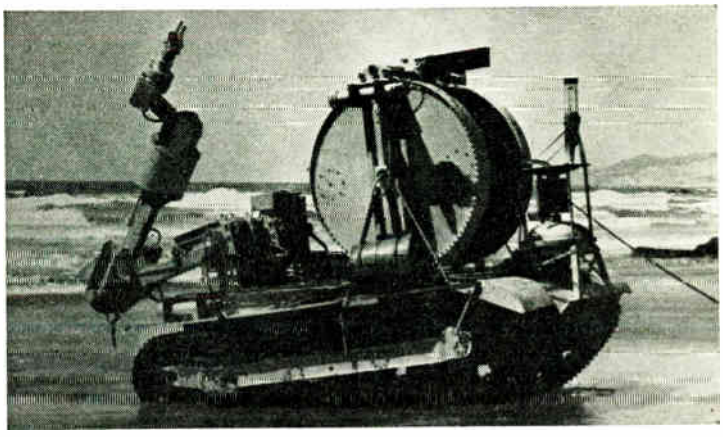


Fig. 7. The Deep Tow vehicle. Mobility is achieved by towing with a length of armored electric cable.

Fig. 8. The remotely operated submersible, RUM. Vehicle was operated over 5-mile length of coaxial cable in 1960.



offers a stable platform for manipulation, freeing the operator from the complex six-coordinate dynamic vehicle control requirement. Offsetting the advantage of platform stability of bottom operation are, of course, the limitations of operation imposed by the nature of the terrain and by the increased turbidity, and consequent reduction of visibility, caused by the operation of the track layer in the bottom sediments.

It is safe to assume that, from this profusion of state-of-the-art vehicle types, the suitable one or ones may be selected that will meet the support requirements of an accessible sea-floor installation once these requirements are definitely delineated.

The Benthic laboratory

The accessible sea-floor installation itself is in an embryonic state—not nearly as far advanced as the vehicle art. In fact, the state-of-the-art may be easily presented by describing the design of the first experimental installation in the Benthic Laboratory program presently underway at the Marine Physical Laboratory. The Benthic Laboratory is a remotely operated instrumentation terminal to be located in deep water off the coast of San Diego and operated over a coaxial shore cable link. Its purpose is to develop a capability for carrying out intensive oceanographic and acoustic measurements in a restricted area on and near the sea floor. The accessible features of the Benthic Laboratory will permit the installation and maintenance of a large number of sensors *in situ* at precisely determined positions. The initial phase of this program is to install an experimental station embodying the concepts and features of the Benthic Laboratory in moderate depth for evaluation.

Physically the station resembles an overgrown bee hive. It is nine feet high and five feet in diameter, and constructed of plastic-coated reinforced concrete. The inverted dome is filled with acid-washed kerosene which forms the environmental barrier for the interior components. A hydraulic, piston-operated door is mounted at the bottom to retain the oil during installation.

The interior arrangement has electronic modules distributed around the circumference of the hive. Two television camera units are installed, one on each side for manipulator control. A hybrid, electric-hydraulic manipulator operating in a cylindrical coordinate system is mounted in the center of the hive. The forward section of the manipulator base plate is cut away to allow access to the free oil–water interface by the manipulator.

A sealed, oil-filled container holding replacement parts may be brought through the interface from underneath the hive, by using a suitable manipulator-equipped vehicle. Once the lid of the container is well above the free interface, the hive manipulator can be used to remove the lid and extract the replacement parts. The compatible design of modular components to match the manipulator characteristics contributes measurably to the efficiency with which operations can be carried out.

It is hoped that this experimental installation will provide a good insight into the technology of the accessible ambient-pressure sea-floor installation. With this insight it should be possible to project the technique to other applications and other configurations. It will be necessary to consider the vehicle station complex as a system and carefully design for the greatest efficiency of operation. The cost of an hour's worth of manipulation time on the sea floor may well be measured in thousands of dollars and thus every effort spent in optimizing all aspects of the system will yield rewarding economic dividends.

Geophysical measurements at ocean bottom

Hugh Bradner University of California

Essentially all the geophysical questions that are asked concerning land masses have counterparts on the ocean floor, and the same variables must be studied: the area which must be explored and mapped; the geology which must be learned; and temperature, heat flow, magnetic field, electric conductivity, gravity, winds (or water currents), response to pressure variations, seismic disturbances, etc., all of which must be measured.

Some of the properties can be measured from ships on the water surface or from submarines just below the surface. Other properties can be measured only by placing instruments directly on the bottom, three or four miles down. Some recent work in this category will be described with special emphasis on seismic measurements.

For many years, oil companies have used geophones (insensitive, high-frequency seismometers) in shallow water for geophysical prospecting. The usual technique is to lower the geophone on a cable from an anchored ship and record the seismic information on shipboard.

The task becomes much more difficult in the unattended environment of the deep ocean bottom. It is not economically feasible to predetermine either the instrument coupling to the bottom or its reposing angle. Seismograph records must be examined critically for environmental effects, such as oscillation of the instrument by water currents. In ordinary land-based seismometry, the effects of a 10-mile-an-hour wind may shake the instrument many millimicrons, even though it is mounted on a concrete pedestal and buried in an underground vault. In water those same forces would be produced by currents of a few tenths of a mile per hour. Unfortunately, oceanographers cannot state even the order of magnitude of deep ocean-bottom currents in most of the places of seismic interest.

The weight and volume of the components are severely limited by the difficulties of handling the equipment on shipboard. In turn, these impose restrictions on the power consumption and length of record that may be obtained. If the seismometer is connected to the surface by a recovery cable, it is likely to be dragged or jerked by the pitching and drifting of the ship—and seamanship, plus luck, may be required to keep from tangling and breaking the cable. Even though the cable hangs vertically downward for 20000 feet, the large heaving and dragging at the upper end shakes the seismometer an intolerable amount, unless the cable can be decoupled from it. The usual way of decoupling is to terminate

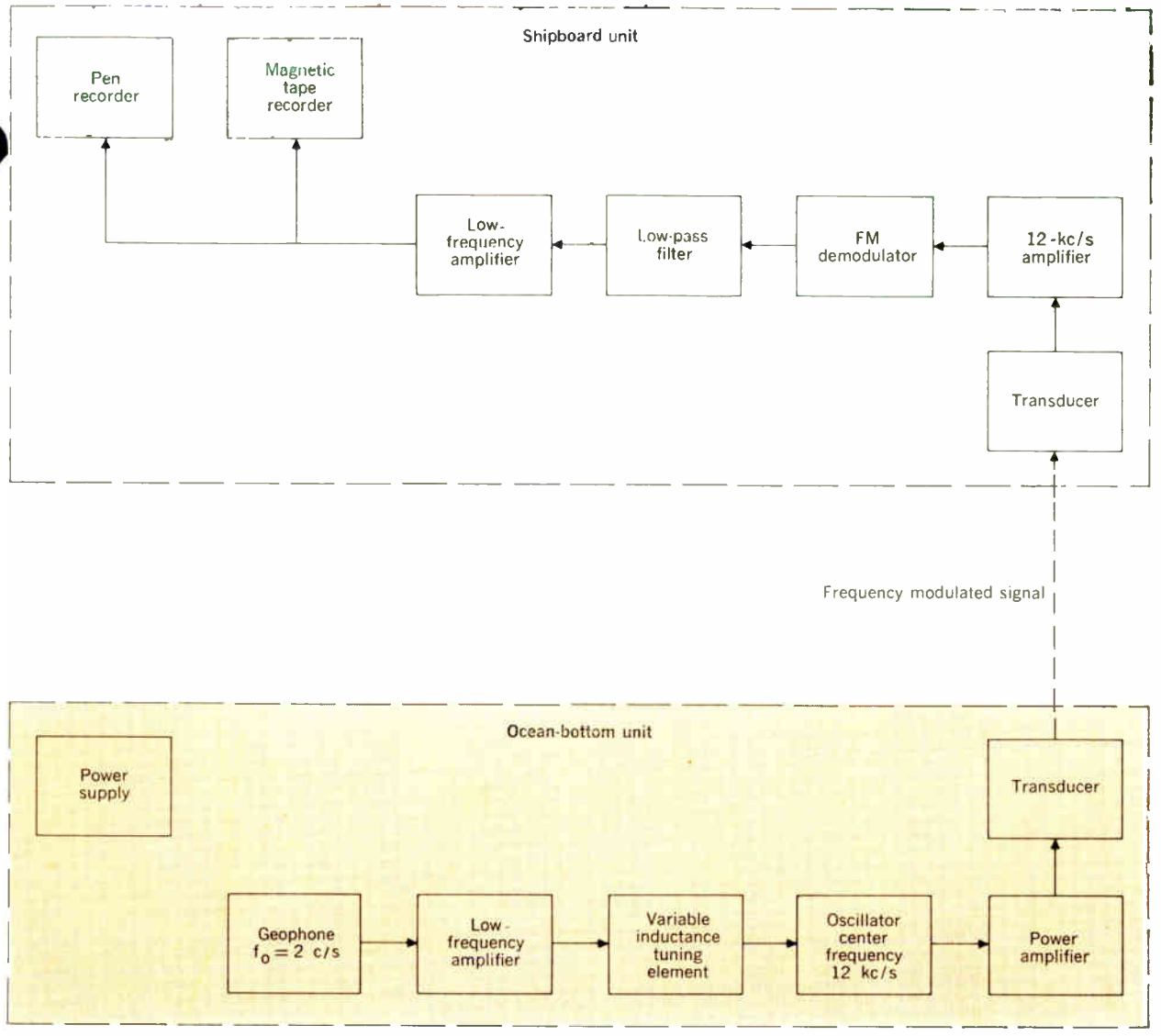


Fig. 9. Block diagram of Columbia University's telemetering ocean-bottom seismograph. Signal from ocean-bottom unit can be telemetered acoustically to the shipboard unit from depths of 20 000 feet. (Courtesy of Lamont Geological Observatory.)

the cable in as heavy an anchor as the ship and cable can stand, and to separate the seismometer from the anchor by a few hundred yards of flexible line. Even so, there is some danger that the heaving on the anchor will transmit vibrations to the seismometer, or that the ship will drift far enough in a few minutes to drag the anchor and instrument. If the instrument is intended to sink freely to the bottom without tether and rise to the surface after making a record, the apparatus must be conservatively designed, and much care and good luck must attend the seagoing operation. The situation is bad if the instrument rises, but fails to transmit a radio signal on reaching the surface. The ocean surface is a lonely place to search for a small, drifting object that can be located only by visual inspection. The situation is worse if it fails to rise, since recovery of a small object from the ocean bottom is beyond the state of today's level of technology.

If a free-falling, untethered seismometer is to telemeter the seismic information back to a ship on the surface, there are a number of problems to consider, including signal strength, Doppler modulation of the signal when the receiving hydrophone rises and falls in the waves, and difficulty of operating in heavy weather.

All of the above problems may be circumvented by connecting the seismometer to the nearest land by electric cable, using extremely reliable apparatus. However, economic factors limit the distance that cables with repeater stations can be run from shore.

Maurice Ewing, of Columbia University's Lamont Geological Observatory, pioneered deep-ocean seismic work. The Texas Instruments Company (TI) and the University of California (UC) at La Jolla undertook measurements in 1961 as a result of the United States Government's interest in underground bomb-test detection. The oceans give seismic access to many areas of the world; but up to 1961 less than a half-dozen measurements of ocean-bottom seismic background had been made, and information on S/N ratios was essentially nonexistent.

The prevailing view was that the ocean basins might provide a very quiet environment for monitoring explosion signals. An opposing view was that the ocean basins might be more noisy than land if the microseism background is generated at sea, far from shore. Today, there is much more information on the ocean-bottom background and signals—but neither the origin of the microseisms nor the magnitudes of the disturbances are fully understood.

Telemetering seismometer

Most of Columbia University's recent work has been carried out with acoustically telemetering instruments, planted and recovered by cable from a ship. A block diagram of such an instrument is shown in Fig. 9.

The amplified output from the geophone drives a variable inductance element which frequency-modulates a 12-kc/s oscillator of about 1 watt power. The oscillator drives a small, free-flooding magnetostrictive acoustic radiator. The signal is then amplified, recorded with a lower carrier frequency on a tape recorder, and simultaneously displayed in demodulated form on a chart recorder. This method of operation allows the experimenter to monitor his equipment. That advantage is partially offset by the difficulties previously mentioned.

Implacement method. In the Atlantic near Bermuda, a seismometer package was coupled to a 1500-pound weight by 600 feet of manila line. This was lowered to the bottom on the end of a steel cable. Measurements were obtained until the ship had drifted far enough to drag the weight. By slacking the wire at the proper rate, the seismometer was decoupled from the ship for as long as seven minutes. In the Arctic, measurements were made from an ice island which was drifting about one mile per day. In that environment the 1500-pound weight was omitted, and the seismograph remained quiet for periods of as long as one hour.

Data reduction. Two analog methods have been used by the Columbia group to obtain seismic spectra. In the first method, predominant periods and amplitudes are measured on broadband visible recordings of the demodulated seismometer signals. A smooth line is then drawn through these data points. In the second method, the demodulated seismograph records are passed through standard one-octave bandpass filters and then recorded on a chart. An envelope is drawn to include 90 per cent of the peak-to-peak amplitude for the record of five minutes' duration. These methods of analysis are fast and easy, but they do not give the high spectral resolution or quantitative results that can be obtained from numerical methods of time series analysis with digital computers.

Seismicity in the Atlantic and Arctic Oceans. Ewing's group reports five measurements in the Atlantic, three in the Arctic, and one in the Gulf of Mexico. Fig. 10 shows these locations along with all other deep ocean-bottom seismic measurements that are reported in the literature. On a global scale, the number of sample points is very small, even though the amount of data has increased thirtyfold in the past five years.

The Columbia group observes that microseism background noise ranges from below the amplitude of a quiet continental site, to well above the average continental limits. They conclude that the S/N ratios for seismic body waves, on at least some parts of the ocean floor, will be as good as those at average and quiet land stations. These data are qualitatively different from the Pacific Ocean results of Texas Instruments Company and University of California, which are discussed later. Columbia measurements from a drifting ice island in the Arctic Ocean during the spring season indicated very low noise levels. The ice surface was found to be as quiet as the ocean bottom at that time.

The limited data from Columbia do not lead to final conclusions on the origin or mode of propagation of ambient seismic energy on the ocean floor. They suggest that the absolute amplitude of the ambient noise may be as much a function of local geology, bottom structure, and location in the ocean basin as it is of the meteorological conditions at the surface.

The shore-cabled seismometer. Recent unpublished work by Columbia has been directed toward modifying a 15-second-period 3-component lunar seismometer for ocean-bottom operation. This long-period seismometer, together with a short-period vertical seismometer and recorders of temperature, tide, and water current, will be installed about 90 miles off Point Reyes, along the northern California coast. Data will be transmitted to shore by coaxial cable.

Tethered seismometer

The Texas Instruments Company has made a large number of measurements with a tape recording seismometer that is planted and recovered by cable from a ship.

The completed instrument is more than 10 feet high and weighs 1700 pounds. The Danforth anchor weighs an additional 800 pounds. In spite of the weight and awkwardness of this assembly, TI has recorded far more seismic data in the deep ocean than anyone else. The company also has made simultaneous land seismic records in most cases.

The original TI instrument used digital recording of the multiplexed signal from three seismic transducers and one pressure transducer. The dynamic range was 36 dB. An improved instrument makes dual-level wide-band FM recordings from the same transducers.

Digital recording has the advantage of permitting very large dynamic range, with values as high as 100–120 dB readily obtained. However, the recording time is limited, since the digital format does use the magnetic tape efficiently. Wide-band FM recordings, limited to about one third of that dynamic range, make somewhat more efficient use of the magnetic tape. The highest tape-packing density can be obtained by direct recording of the subaudio signal at very low tape speeds. The dynamic ranges is again limited to about 30 or 40 dB; and there are very severe requirements on amplifier stability and tape speed uniformity.

Data reduction. The TI data are transcribed in analog strip-chart form for quick appraisal; then selected parts are analyzed digitally on the TI special-purpose computer.

Noise samples and events selected for analysis are transcribed to computer–compatible digital format with a twentyfold speed-up. The digital sampling rate is 500 per second, and hence the maximum permissible digital analysis frequency is 250 c/s. Low-pass filters are used in the FM–digital transcription to suppress energy above the Nyquist frequency of 250 c/s (12.5-c/s real time).

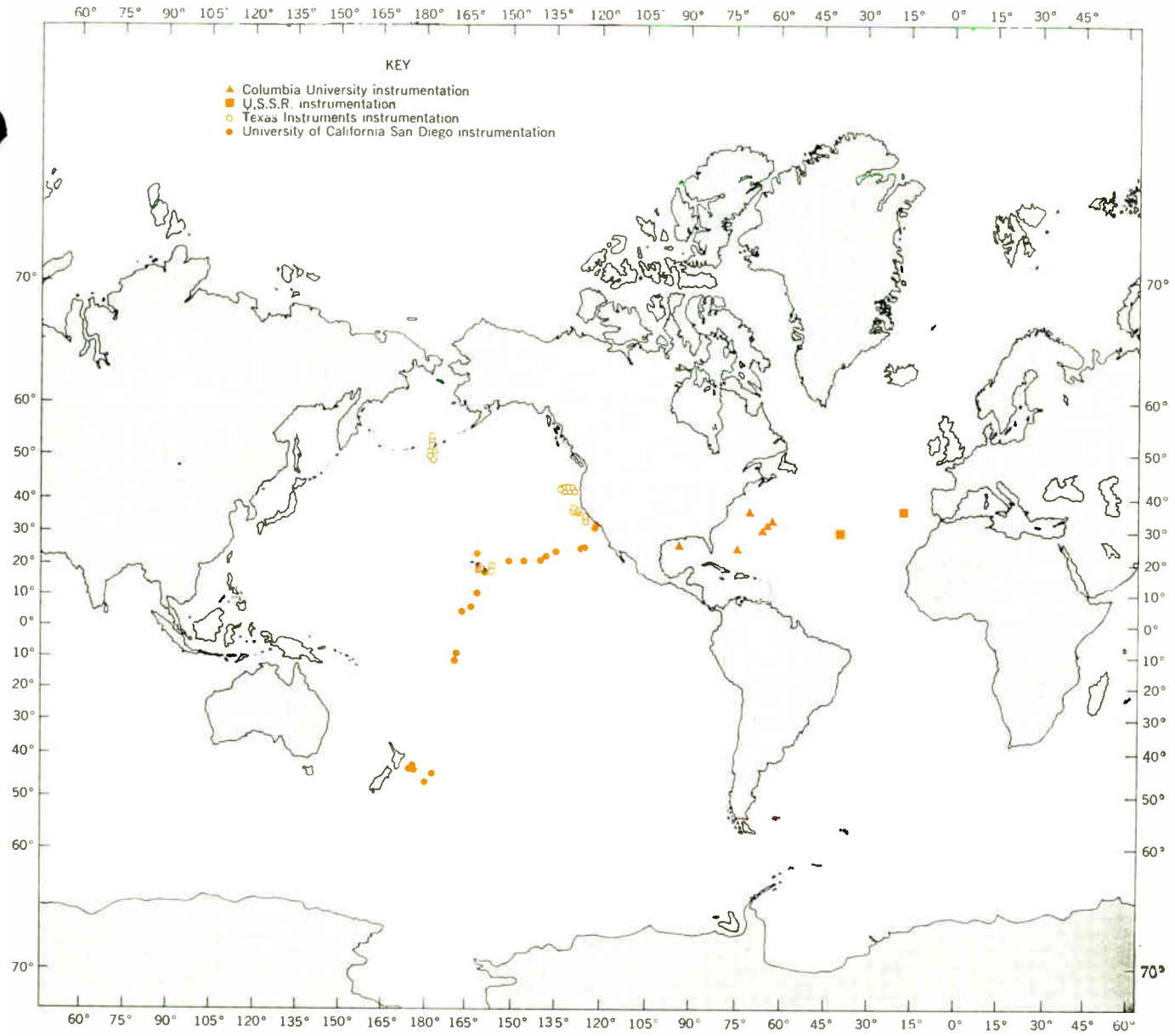


Fig. 10. Locations of deep-ocean seismic measurements.

Seismic noise power spectra are obtained from the digital data by correlating noise samples, approximately three minutes in length. The correlations are computed to ±5 seconds, and Fourier-transformed to give spectral estimates with 0.1-c/s resolution. The spectral estimates are smoothed to reduce the correlation truncation effects. The resolution after smoothing is approximately 0.2 c/s. This is in sharp contrast with the resolution of one octave that is normally obtained with analog filters.

Texas Instruments is now completing a group of ten new instruments with 30-day direct recording on low-speed magnetic tape. The handling and recovery will be similar to the UC pop-up instruments, described in the next section. The 30-day TI instruments will be used for an extensive set of measurements of natural signals, explosions, and background in the Aleutian area of the north Pacific this fall.

Pop-up seismometer

Interest at UC's Institute of Geophysics and Planetary Physics in La Jolla is focused on studying the origin and propagation of microseisms. Measurements were planned at a variety of locations throughout the central and south Pacific. It was felt that simultaneous ocean-bottom and land seismic measurements of one-hour duration would be sufficient. A requirement was that the instruments be small enough to be handled by 60-foot boats, with a minimum of special rigging. These considerations led us to put a tape-recording seismometer into a hollow, buoyant aluminum sphere which could be rigidly attached to a 150-pound lead-spike anchor, lifted over the ship rail, and allowed to sink freely to the bottom. There is no link between the instrument and the surface during the recording. When the hour-long record is completed, the lead anchor detaches from the

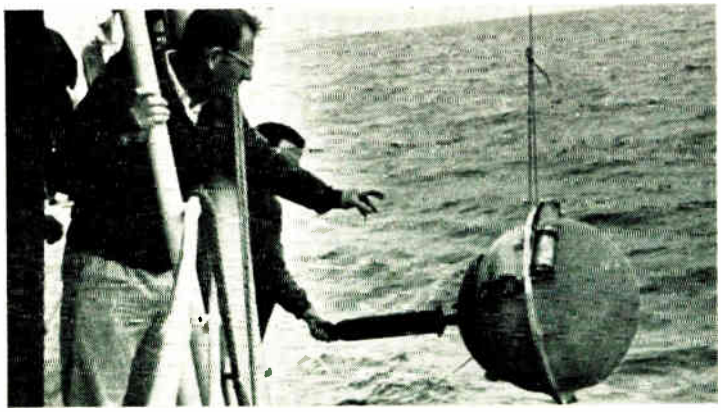
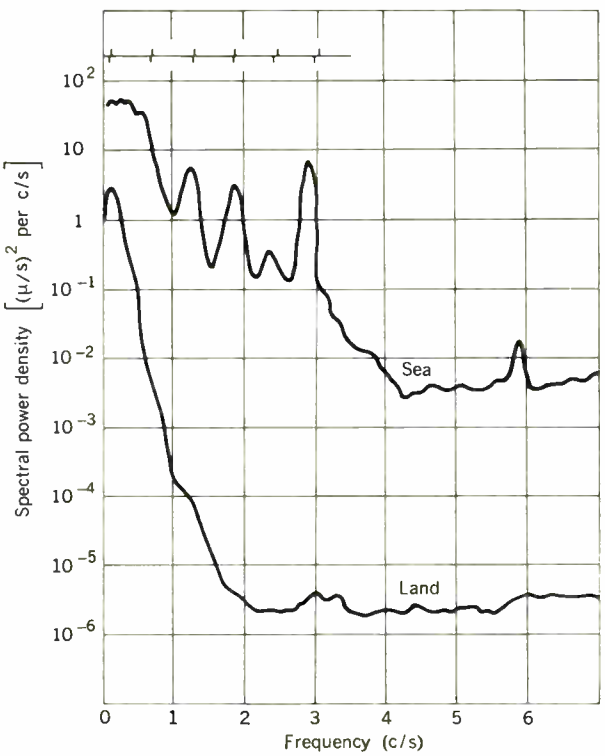


Fig. 11. Launching the UC seismometer. The polypropylene support line is cut as soon as the instrument is clear of the ship. Seismometer then sinks to bottom, and makes a record on internal FM tape recorder. Anchor is then automatically jettisoned, and sphere floats to the surface.

Fig. 12. Representative microseism background spectra on the ocean bottom and in Hawaii, Feb. 8, 1963.



instrument, which then floats to the surface. To facilitate retrieval, a citizens band radio transmitter turns on when the sphere reaches the surface. Fig. 11 shows the instrument being launched.

The seismometer is a modified version of the lunar seismometer developed at the California Institute of Technology. Seismic signals are amplified, passed through a voltage-to-frequency converter, and recorded on ½-inch magnetic tape. Timing circuits turn on the amplifier and tape recorder after a preset delay, and turn it off after one hour of recording.

The seismometer can function from vertical to within 10° of horizontal by virtue of a motor-driven screw that adjusts spring tension to center the mass at the start of the recording cycle. The tilt angle and compass orientation of the instrument on the ocean bottom are recorded by a tilt meter mounted inside the pressure sphere. The tilt indicator is a flash bulb in a pendulum, with a fiducial aperture to expose Eastman photo-resist coating when the bulk fires. At the same time, the orientation is photographically recorded by the light through a small slit in the compass card.

The pressure vessel containing the seismographic equipment is composed of two 22-inch-ID deep-drawn hemispheres of 7178 T-6 aluminum. The spheres are assembled by clamping, with six angle brackets, onto an O-ringed aluminum center plate, which carries all the instruments in the sphere. Electric leads can be brought out through the center plate or through the surface of the sphere by Mecca plugs or by similar high-pressure bulkhead connectors. A Mecca plug mounted through the top of the hemisphere serves as antenna feedthrough for the citizens band radio recovery beacon.

The anchor is a lead spike, 4½ inches in diameter and 2 feet long, with a 30° conical tip. The top of the spike terminates in a dish-shaped flange, which carries three cable eyes for attachment to the buoyant seismometer sphere. The dimensions were chosen so that the spike will penetrate 16 to 24 inches into mud-clay sediment of the rigidity thought to be characteristic of the broad, flat Pacific Ocean basin. If the anchor strikes soft mud, the flange will keep the package from penetrating so far that the Van Dorn magnesium release could be obstructed.

The package returns to the surface in about six hours, and the radio begins to transmit as soon as the antenna is above water. The radio beacon signal is detected on shipboard by an ordinary citizens band receiver. A commercial three-element yagi antenna, mounted at a high point on the ship, serves for homing on the floating sphere, which is retrieved by a short length of floating polypropylene tag line, and easily lifted back on board the ship. More than 40 successful drops and recoveries have been made with the untethered pop-up seismometers.

Data reduction

The TI data and the magnetic tapes of UC are examined in similar ways. The records can be played back for visual observation in analog display, and they can be fed into a computer for detailed analysis. The FM signals are fed directly from the tape to an electronic counter whose sampling rate is controlled by the 582-c/s reference track on the tape. The output of this counter is processed in a CDC-1604 computer with a general-purpose time-series analysis program.

Normally, each spectrum represents a time series eight minutes long. The Tukey method of spectral analysis is used, with 200 lags. The sample rate is 18 per second, resulting in 90 degrees of freedom.

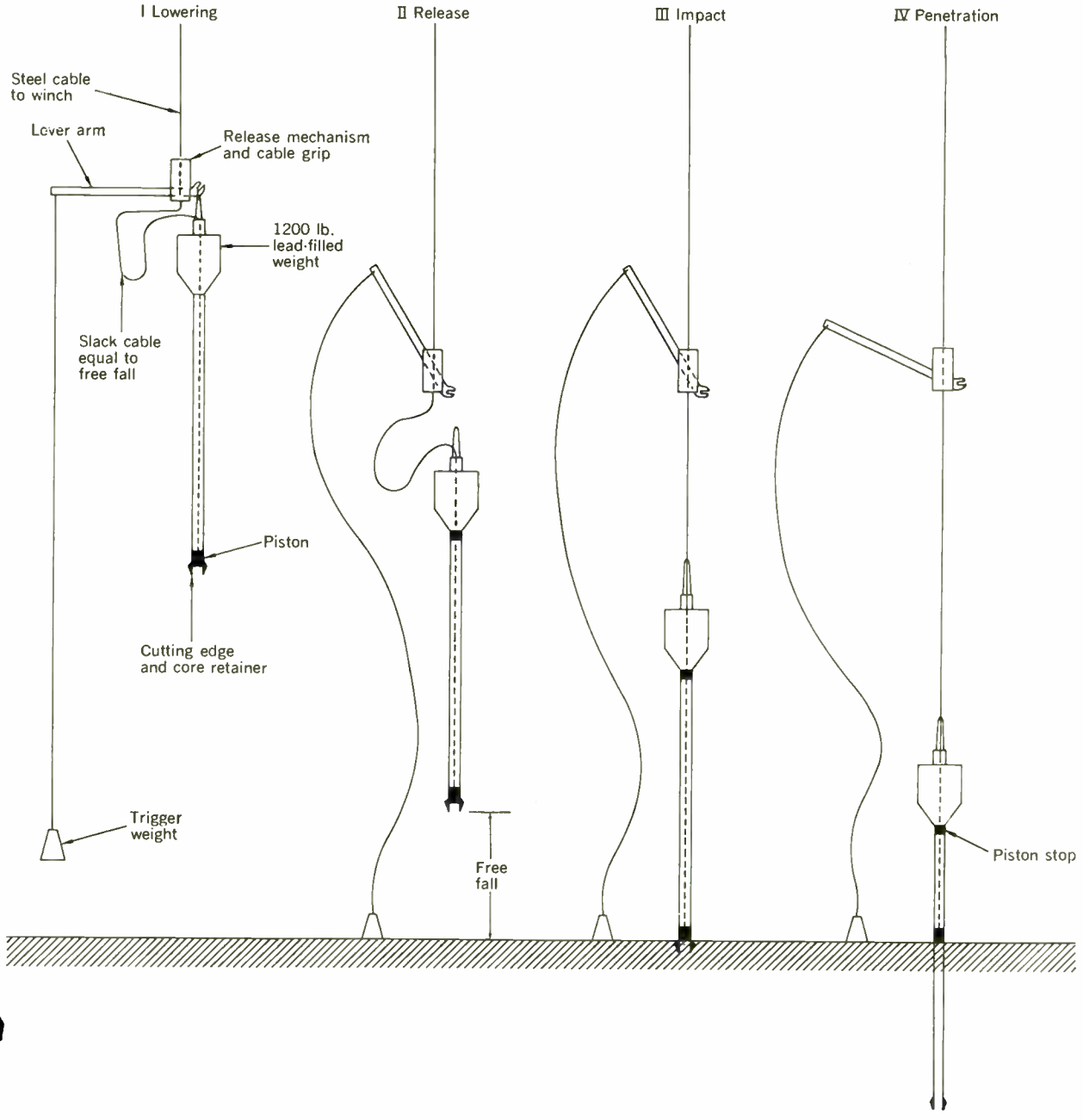
Representative seismic background spectra on land and on ocean bottom are shown in Fig. 12. The calibrations of the instruments were taken into account during the analysis; therefore, the computed spectra represent actual earth noise power. Instrument noise is below the minimum signal level for the spectra. The land instrument was a vertical seismometer on bedrock

in a cave on Oahu, Hawaii. The ocean-bottom instrument was a three-component equiaxial seismometer 500 miles east, in 18 000-foot-deep water. Only one component is shown in the figure, the other two components were similar. The relatively smooth monotonic decrease in power density of the land spectrum at frequencies above the microseism peak is characteristic of all our island records. Continental United States spectra usually show more peaked structure, presumably due to the large amount of man-made noise, and the large area for meteorological effects. The evenly spaced peaks on the ocean-bottom record are a common occurrence. Their frequencies are compatible with a regular set of organ-pipe modes in the water. In this example, every fourth mode appears to be strongly excited.

Data assembled by TI and UC show the energy of the deep sea-bottom noise to be one or more orders of magnitude higher than that measured on land in the lower frequencies, and three or more orders of magnitude higher in the upper frequencies. This indicates that a great deal of the microseismic energy may be generated at sea and that the higher frequencies are not transmitted to land.

On land, a significant part of the microseism energy is carried in surface waves by two types: retrograde vertically polarized waves, called Rayleigh waves; and horizontally polarized shear waves, called Love waves. A three-component seismometer can distinguish these waves in spectral peaks if the energy comes from a localized source. Pure Love waves, for example, would have no vertical component and a 0° or 180° phase difference between the

Fig. 13. Columbia University piston corer (courtesy Lamont Geological Observatory.)



two coherent, horizontal components. Pure Rayleigh or Love waves are rarely found, since more than one source is usually active at any time; and energy is further mixed up by the inhomogeneous earth. The energy in the microseism peak, it was found, cannot be regularly associated with either the vertically or horizontally polarized surface waves that are commonly recognized on land stations, although either form of wave may contribute much of the energy on a given day. Near-vertical compressional waves in the water, called "organ-pipe" waves, can be responsible for some marked peaks.

The shape of the spectrum does not change markedly as one travels to mid-ocean, far from storms or shore lines. From that observation, it can be concluded that high-frequency, seismic background noise is generated locally, throughout the station areas in the Pacific Ocean. The most likely mechanism for generating the high-frequency energy is a statistical superposition of oppositely traveling surface water waves. These standing waves will exert forces on the bottom, at twice the wave frequency.

Good examples of all of the currently accepted mechanisms of microseism generations were found. Also found was evidence for generation near shore, and in storms far from shore. It is suggested that the normal, continual, microseism background is a superposition of energy from many sources. The peak at a 6–8-second period may be accentuated by the resonance of the 15000-foot-thick ocean waveguide, as well as by the predominant frequency of the surface waves.

To summarize, several laboratories have developed apparatus for ocean-bottom seismic measurements and have made a total of about 100 records in deep oceans. They have observed earthquakes, measured propagation velocities, studied microseism noise, and determined that S/N ratios may allow them to monitor bomb tests. But many more measurements will be needed before definitive statements can be made about any of these phenomena.

Other measurements

Coring tool. At the opposite extreme from the recent development of seismometers, there are modifications to a simple tool that is 100 years old: one used for getting cores from the ocean bottom.

Cores are used for studying geology, sedimentation rate, geochemistry, paleo temperatures, paleoecology, mechanical properties, length of the day and year in early earth-history, wandering and flip of magnetic poles, geothermal gradients, etc. Although most are six to ten feet long, Columbia University has obtained 60-foot cores in soft sediments with the conventional piston device shown in Fig. 13.

Expensive, tapered, steel cables are used to hold 1500-pound weights in deep-ocean work. And good seamanship is needed, to keep from dragging and bending the core tube.

David Moore at the Naval Electronics Laboratory, San Diego, recently developed coring tools that can be free-dropped from shipboard. They plummet into the bottom, drop their weights, and float back to the surface for later recovery.

Willard Bascom, of Ocean Science and Engineering, Inc., is testing a vibratory corer, which is oscillated longitudinally by an eccentric flywheel motor. Long cores, in hard compacted sediments, should be obtained with this tool.

The well-publicized Mohole project is still some distance away from realization.

Telluric fields. Two electrodes stuck at different places into the earth will exhibit a potential difference of about 8-mV/mi separation. The potential is caused by atmospheric electricity, by currents in the earth's interior, and by conductors moving in the magnetic field of the earth.

With this technique, M. S. Longuet-Higgins is now measuring the rock resistivity across the English Channel. D. Cartwright used potential differences across the English Channel to find the water flow there, when he calculated the sea-level difference between France and England. And the Gulf Stream current has been determined by measuring the potential difference between Key West and Havana.

C. S. Cox recently tried measuring potentials on the ocean bottom to look for internal water waves. He used Ag–AgCl electrodes, separated three miles. He is now trying to reduce the spacing to 30 feet, although that requires keeping electrochemical potentials far below the 50-μV level of the real signals. Cox says that the main problem is knowing the separation and orientation of the electrodes.

Working on the ocean bottom has one advantage: the 20 000-foot-deep ocean is one skin depth for electromagnetic waves of $^1/_4$ cycle per second. Thus, the bottom is well-shielded for higher frequencies.

Heat flow. Some puzzling new results have been observed concerning the temperature gradient a few feet from the ocean bottom.

Heat flow through the bottom is usually measured by recording the thermal conductivity of the sediment and the temperature difference between two probes separated vertically about six feet. The conductivity is about the same as water, 0.4 $Btu \cdot hr^{-1} \cdot ft^{-2} \cdot {}^{\circ}F^{-1} \cdot ft$, or 1 per cent of the conductivity of steel. The temperature gradient is about 0.05°F per inch, or 100 times that in the body of the ocean where convection establishes the adiabatic gradient. Recent measurements, made by hanging probes just above the sediments rather than in the bottom, showed the gradient to be characteristic of the sediments: 100 times as large as expected.

The flow can be treated theoretically by considering circulation between two horizontal plates, with the lower plate heated. The circulation is convective unless the spacing of the plates is too small. The critical distance, however, is inches, rather than feet. The critical distance can be increased by roll stability if the water is made to flow horizontally between the plates, but six feet is still surprisingly large.

This example is not unique. The ocean bottom gives surprises every time new observations are made. Hence it is clear that, in the alien environment of the ocean bottom, apparatus and observations must be exceptionally reliable if real effects are to be differentiated from the instrumental ones.

Victor C. Anderson's paper represents work sponsored by the Office of Naval Research under contract Nonr 2216(05).

Hugh Bradner expresses his thanks to the researchers cited in his paper for permission to quote their data.

This article in its entirety is based on three of the four papers presented at the session on "Instrumenting the Sea-Floor—Why and How," WESCON, Los Angeles, Calif., Aug. 26, 1964.

Array antennas:

new applications for an old technique

Antenna arrays exist in a wide variety of configurations, with the number of radiating elements ranging from two to several thousand. However, all are governed by the same fundamental principles, which are described in this article

John L. Allen *Lincoln Laboratory, M.I.T.*

Contemporary antenna problems have stimulated renewed interest in the use of array-type antennas, in which a number of individual radiators are grouped together and coherently excited in some controlled manner. This type of antenna predates the application of reflector types in radio and radar applications; but the new interest, with emphasis on arrays of large numbers of elements, has provided the impetus for advances in our knowledge of the principles of arrays and the invention of new arraying techniques. It is the aim of this article to provide an exposition of the current state of the field.

The primary consideration in the choice of an antenna composed of many discretely located sources over a single, continuous structure—such as a parabolic reflector—is usually reducible to the fact that in continuous-structure antennas the problem of achieving and maintaining a desired electromagnetic field amplitude and phase distribution across the antenna is primarily mechanical, whereas in an array antenna it can be made primarily electrical. In some cases, an electrical problem may be preferable to a mechanical one, as in the following instances:

1. When very narrow antenna beams (e.g., $\ll 1°$) are desired but the targets of interest vary in character so slowly that the antenna may be individually adjusted for maximum response from each target. In these cases, electrical alignment may be easier to implement than equivalent mechanical alignment. For example, there are several arrays of parabolic dishes used for radio astronomy in which the "target" is a celestial radio source.

2. When beams of more modest widths are desired (e.g., of the order of 1°), and either rapid movement of a single beam or the formation of many simultaneous beams is required for high-speed angular coverage, as in radar for missile defense. These arrays usually consist of many electrically small radiators (having dimensions comparable to or less than the operating wavelength) such as the

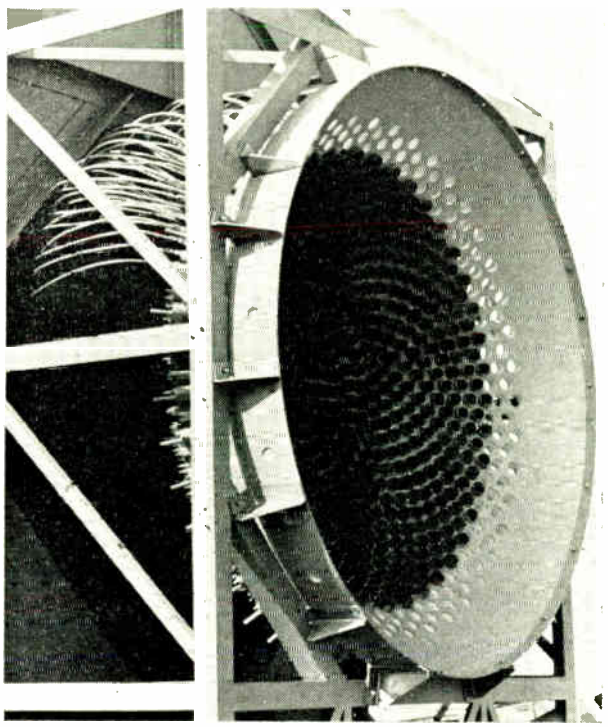


Fig. 1. Array of many electrically small antennas for radar applications. (Courtesy of and developed by Hughes Aircraft Co. under Contract No. AF30(602)2415, sponsored by Rome Air Development Center.)

one indicated in Fig. 1, in which several simultaneous beams are formed and scanned by electronic variation of the interconnection of the elements, without mechanical motion.

3. When mechanical motion of an antenna is prohibited because of environmental considerations, as in the case of directional antennas for satellite applications, such as the 16-element array used on the SYNCOM communications satellite.

As these examples illustrate, arrays come in a wide variety of configurations with as few as two and as many as several thousand radiating elements. Nevertheless, all are governed by a common set of fundamental principles, the more important of which will be surveyed in the following pages.

We will first briefly review the theory of ideal arrays—that is, error-free arrays of noninteracting, nondirectional radiators. The applicability of and modifications to this idealized theory for arrays of interacting, directional radiators will then be explored and the effects of errors in the array excitation examined. The article will conclude with a survey of some representative types of networks for exciting array antennas.*

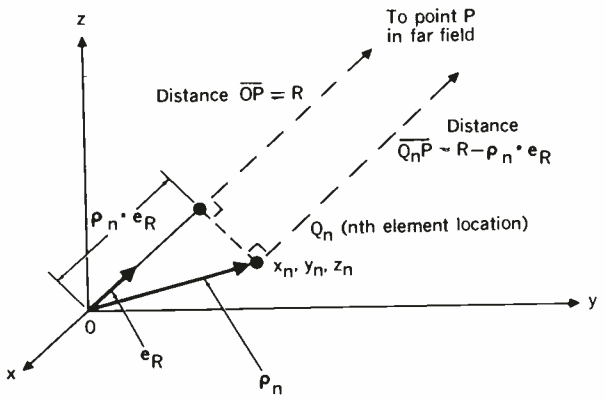
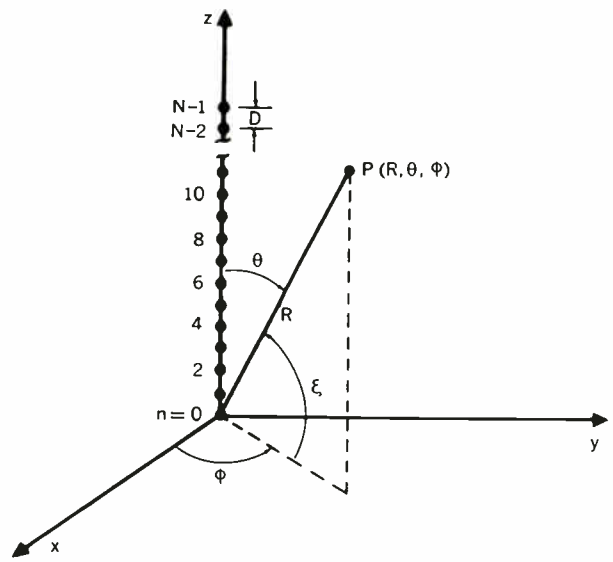


Fig. 2. Geometric construction for determining differences in path length from the element locations to a point in the far field, so that paths from all element locations to P will be approximately parallel.

Fig. 3. N-element linear-array geometry.



The error-free arrays approach

A useful and simplified point from which to develop the basic theory of arrays is the consideration of a highly idealized case: all the radiators (elements) of the array are assumed to radiate isotropically and to experience no interaction effects (mutual coupling) between elements. In spite of its artificiality, this approach provides useful results in many cases. For example, the field-strength patterns calculated on this basis can be applied with very little modification to many types of real arrays. This approach also lends useful insight and a starting point for some of the more difficult analyses of other types of real arrays.

Broadside array far-field pattern shaping. If we place N isotropic radiators at arbitrary positions specified by a set of vectors ϱ_n drawn from the origin of a coordinate system to each element location and excite the nth radiator with a sinusoidally varying current i_n of frequency ω (free-space wavelength λ), then the contribution† of the nth element to the field at some point a distance R from the origin in a direction specified by a unit vector $\mathbf{e}_R$ approaches the form

$$A_n(\mathbf{e}_R) \approx i_n e^{jk\varrho_n \cdot \mathbf{e}_R} \qquad k = \frac{2\pi}{\lambda} \tag{1}$$

as R becomes very large relative to the magnitude of the largest ϱ_n. The point is then said to be in the "far field" of the array. (A rigorous derivation may be found in most texts on antennas.) The factor $\varrho_n \cdot \mathbf{e}_R$ is the perpendicular distance of the nth element from a plane through the origin perpendicular to the direction $\mathbf{e}_R$ as indicated in Fig. 2. This distance, when multiplied by the wave number k, gives the phase shift of the signal from an element at that location relative to the phase it would have contributed if the element were located at the origin.

The total field strength at the observation point is the vector sum of the contributions of each of the N elements:

$$A(\mathbf{e}_R) = \sum_n i_n e^{jk\varrho_n \cdot \mathbf{e}_R} \tag{2}$$

This is the fundamental expression for the far-field pattern of an array of isotropic radiators, and is usually referred to as the "array factor" or "space factor" of the array. Note that, as written, it has only an angular dependence (the direction $\mathbf{e}_R$), since we tacitly assumed that the pattern is determined at a constant distance from the origin R, and all proportionality constants are suppressed as being unnecessary to this discussion.

* The reader wishing a similar but more detailed description is referred to Ref. 1. Many of the results and approximations given in this article are justified therein (essentially the same material is to be published by McGraw-Hill as Chapter 15 of *Elements of High Power Radar*, edited by J. Freedman and L. D. Smullin). References 2 through 5 also treat the subject in varying degrees of generality and depth.

† Relative to some reference, such as the contribution of an element with unit excitation at the origin. Factors of $e^{j\omega t}$ are suppressed throughout this article.

While elegant in its generality, Eq. (2) leaves something to be desired as far as providing insight into the questions of importance. In particular, the array designer would like answers to the questions:

1. How many elements N should be used?
2. Where should they be placed (ϱ_n's)?
3. How should the exciting currents i_n be chosen?

To indicate the general nature of the answers to these questions, let us examine a special, simple case: an array of radiators equally spaced along a line (linear in the geometric sense) as indicated in Fig. 3. For this geometry, (2) simplifies to

$$A(\xi) = \sum_{n=0}^{N-1} i_n e^{jknD \sin \xi} \tag{3}$$

As a matter of convenience we choose to express linear array factors in terms of the angle ξ, which is the complement of the usual spherical angle θ. The term $\varrho_n \cdot \mathbf{e}_R$ of (2) is $nD \cos \theta$, and hence $nD \sin \xi$ in the geometry shown in Fig. 3.

Figure 4 shows the resulting normalized magnitude of the array factor plotted against $\sin \xi$ for $N = 16$ for $i_n \equiv 1$, $D/\lambda > 1$. The pattern is of the form $\sin Nx/(N \sin x)$, where $x = (\pi D/\lambda) \sin \xi$, and has, of course, no ϕ dependence.

The maximum at the origin of Fig. 4 is usually referred to as the "main lobe" or "main beam" of the pattern and the other principal maxima as "grating lobes." The smaller lobes are the "side lobes." Because the main lobe is at $\xi = 0$, the array is referred to as a "broadside" array, and we will temporarily concentrate on such arrays. The pattern in the region near the main lobe and first few side lobes is similar to the pattern of a continuous antenna of length ND, if the exciting currents i_n of the array are samples of the current distribution $i(z)$ along the continuous antenna; that is, $i_n = i(nD)$.

Using the pattern of Fig. 4 and Eq. (3) as a guide, we can make several general observations about the dependence of linear array patterns upon the parameters available for array design.

The array factor is of the form of a Fourier series in the variable $\sin \xi$, with a period equal to $1/(D/\lambda)$. The grating lobes are a manifestation of this periodicity. When $D/\lambda = 1/2$, exactly one period of the array factor fills "visible space," for $-90° < \xi < 90°$. If D/λ is increased, the pattern contracts and the main beam becomes narrower. For $D/\lambda \geq 1$, visible grating lobes exist and the antenna angular information becomes ambiguous. However, the occurrence of grating lobes can be avoided by unequally spacing the elements, thereby removing the periodicity from (3). The result is a pattern qualitatively similar to that of Fig. 5. The grating lobes are suppressed, but the energy that would be in the grating lobes of a regularly spaced array is now spread out over the pattern regions well removed from the antenna main lobe, raising the side lobes in that region.[6-9] This may be a useful trade in some instances, and will be examined in more detail later.

From the properties of Fourier transforms, Eq. (3) indicates that the width of the far-field main lobe becomes narrower as the extent of the array is increased, by increasing either N or D/λ. Either possibility is included in the statement that the broadside beam width Θ depends on the length L of the array as shown in the following relation:

$$\Theta = \frac{K}{L/\lambda} \tag{4}$$

where K is a "fudge factor," the numerical value of which depends on (1) how we define "beam width" (we will use the half-power beam width in numerical examples unless explicitly stated otherwise); (2) whether we wish Θ to be in radians or degrees; (3) the choice of the i_n's. Numerically, it ranges upwards of about 50 (degrees) in cases of practical interest.

A measure of the efficiency of the antenna in concentrating its sensitivity into a small angle is given by the antenna directivity, which is defined as the ratio of the power density per unit solid angle on the peak of the main beam to the average power radiated over all space. (This definition is based on the assumption that the antenna is transmitting; a corresponding definition can be framed in receiving terminology, and the resulting numbers are identical, as reciprocity indicates they should be.) The directivity U depends on the number of elements used, the element spacing, and the choice of element currents. If no grating lobes are in visible space, the dependence is of the form

$$U = 2\eta \frac{ND}{\lambda} \tag{5}$$

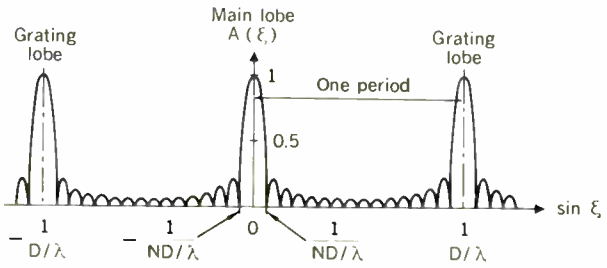
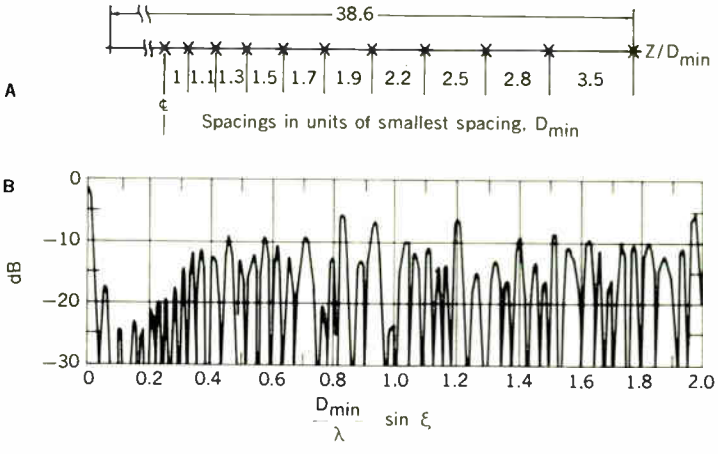


Fig. 4. Normalized array factor for a 16-element uniformly illuminated array vs. sin ξ for $D/\lambda > 1$.

Fig. 5. Spacing (A) and far-field pattern (B) of 21-element array of isotropic radiators.[7]



The factor η expresses the dependence of the directivity on the i_n's and will be referred to as the antenna "illumination efficiency." For a broadside array η is related to the i_n's by

$$\eta = \frac{|\Sigma i_n|^2}{N\Sigma|i_n|^2}$$

It follows that $\eta \leq 1$, with equality applying for uniform illumination. For some combinations of i_n's and spacings, η may depend on spacing, but for most applications the dependence is negligible.[1] If more than one principal maximum is visible ($D/\lambda > 1$), the directivity is reduced approximately by the number of principal lobes in visible space. Thus, visible grating lobes represent not only potential directional ambiguities, but a reduction in antenna sensitivity as well.

We generally wish to choose the i_n's in such a manner as to achieve a "desirable" pattern; however, that which is desirable may change significantly from one task to another. For many applications the principal requirement is the achievement of patterns with a low side-lobe level, high directivity (high value of η), and narrow beam width (small K).* The most common method of achieving this end for a broadside main beam consists of "tapering" the amplitudes of the element currents while exciting all the elements in phase; for example, if we write

$$i_n = a_n e^{j\psi_n}$$

with a_n and ψ_n real, we set all the ψ's equal and vary the a_n's. The amplitude tapering qualitatively takes the form indicated in Fig. 6, in which one excites the outer elements less strongly than those near the center. The height of the close-in side lobes decreases as the "edge illumination" or "pedestal" height h is decreased. The exact shape of the illumination funtions has been the subject of considerable study, and several different techniques exist.[2, 15–19] Quantitatively, the results of the different techniques do not differ dramatically for low side-lobe patterns, and one can formulate rules of thumb that, for a linear array, the illumination efficiency and half-power beam-width coefficient are related to the highest side-lobe levels S (in dB) by

$$\eta \approx 1 - \frac{S - 13}{100} \qquad S \geq 13$$

$$K \approx 50 + (S - 13) \text{ degrees}$$

($S = 13$ dB is achieved for uniform illumination). Thus, for example, if a one-degree beam with a 30-dB highest side-lobe level is required, approximately a 67λ array length is necessary, and the loss in illumination efficiency to achieve the side-lobe level (over the same array with uniform excitation) is about 17 per cent. For a single period of the pattern in visible space ($D/\lambda = 1/2$), about 140 elements are therefore required.

Many antenna applications require "pencil beams" that are narrow in all angular dimensions and require the use of elements distributed in two or more dimensions. The most common geometric configuration is that in which the elements are spread out on a plane surface (a planar array), as illustrated in Fig. 7. Qualitatively, the behavior of a planar array can be inferred from our preceding remarks on linear arrays, but the addition of another dimension gives rise to some quantitative changes. It may be noted that although inference about arrays on arbitrarily shaped surfaces can be roughly drawn by analogy to planar arrays, detailed analysis requires the application of (2).

If the planar array is regularly spaced as indicated in Fig. 7, the resulting array factor is periodic in two dimensions, and a two-dimensional spatial distribution of grating lobes results. The closest grating lobes will be in the spatial planes (specified by the value of angle ϕ in Fig. 7) in which the *projected* element spacing (which would result if the element locations were all projected onto the line formed by the intersection of the spatial plane and the array plane) is widest. This results in essentially the same spacing criterion as for linear arrays; that is, spacings of a half wavelength in the x and y direction permit approximately one period of the array factor in visible space.

By direct analogy to linear arrays, the beam width of a planar array in a specified spatial plane depends on the width of the array in the spatial plane in a manner expressed by (4). For a regularly spaced array with no grating lobes in visible space, radiating into half space—e.g., backed by a ground plane—the directivity for a broadside main beam with no visible grating lobes is

* In other cases, generation of beams of a prescribed main beam shape is desired.[10–14]

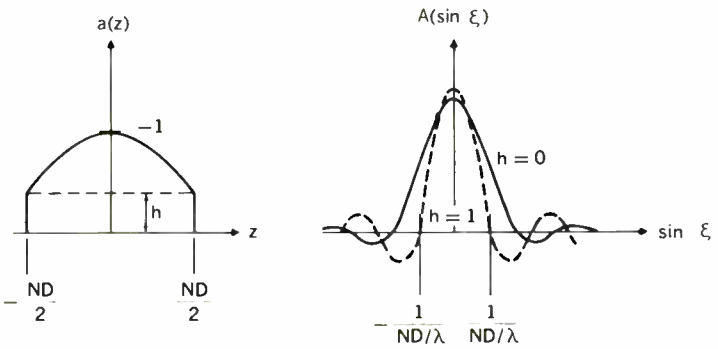
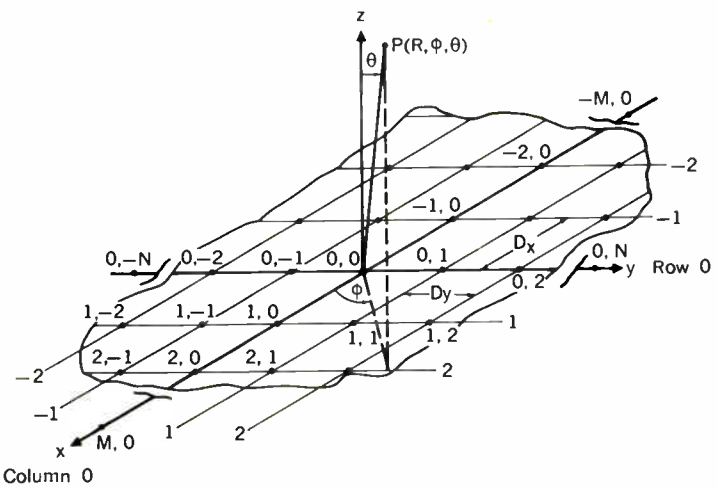


Fig. 6. Typical amplitude-taper envelope, with $a_n = a(nD)$, and resulting dependence of main beam and near-in side lobes on edge illumination h (pedestal height).

Fig. 7. A typical planar-array geometry.



$$U = 4\pi\eta N \frac{a}{\lambda^2} \tag{6}$$

where a is the area allotted each element (the element cell area). The numerical relationships between K, η, and S are somewhat different than for linear arrays. For example, for a circularly symmetrical amplitude taper

$$\left.\begin{aligned} K &\approx 58 + 0.8(S - 17) \\ \eta &\approx 1 - 1.5\,\frac{S - 17}{100} \end{aligned}\right\} \quad \text{degrees} \qquad S > 17$$

($S = 17$ for a uniformly illuminated circular antenna.) Thus, to produce a 30-dB first side-lobe pattern with a one-degree beam width requires about a 60λ-diameter array and produces a loss in directive efficiency (over uniform illumination) of about 20 per cent. For no more than a single period of the pattern in the $\phi = 0$ and $\phi = 90°$ planes of visible space, about 10 000 elements are required.

In some applications, it is inconvenient to taper the amplitudes of the element currents; for example, in high-power transmitting arrays in which each element is fed by a separate transmitter, efficient operation requires that each transmitter be operated in a saturated mode. We have previously mentioned the use of unequal element spacing to smear out grating lobes. This technique can also be used to shape the main beam and close-in side lobes with no amplitude tapering, and is referred to as density tapering or space tapering.[20,8]

The far-field pattern is shaped by the use of a variation in element density that follows a curve similar to that of Fig. 6. The main-beam and close-in side-lobe shapes of a density-tapered array are essentially indistinguishable from those produced by an array of the same size using an amplitude taper generated by the same curve. However, the average side-lobe level far out in the pattern of a density-tapered array is approximately

$$S_A \approx 10 \log_{10} \frac{1}{N} \tag{7}$$

for low first side-lobe tapers. Note that the far-out average side-lobe level essentially depends only on the number of elements used; only the uniformity of the side-lobe level—staying close to the average everywhere—depends upon the element placement. By contrast, the far-out side-lobe level of a perfect amplitude-tapered regularly spaced array is approximately $20 \log_{10} 1/N$. The difference in practice is not so great as these idealized results would indicate—for example, −30-dB average side-lobe level for a 1000-element density-tapered array and −60 dB for the same array with amplitude taper only—because of error effects.

For practical reasons—such as the control of interaction between real elements, ease of design of feed networks, and the recognition of the finite physical dimensions of real antennas—the elements cannot be placed in completely arbitrary locations, and a compromise technique is frequently used in which the elements are placed on a regularly spaced grid whose spacing was chosen to prevent grating lobe occurrence in visible space. Figure 8(A) shows a grid with space for 4000 elements in which only 900 have been placed, as indicated by the darkened squares. Figure 8(B) shows a cut of the resulting pattern with the predicted average side-lobe level of (7) indicated. Ideally, all peaks would be 3 dB above S_A if all of the peaks were equal; the performance that is indicated comes close to this ideal.

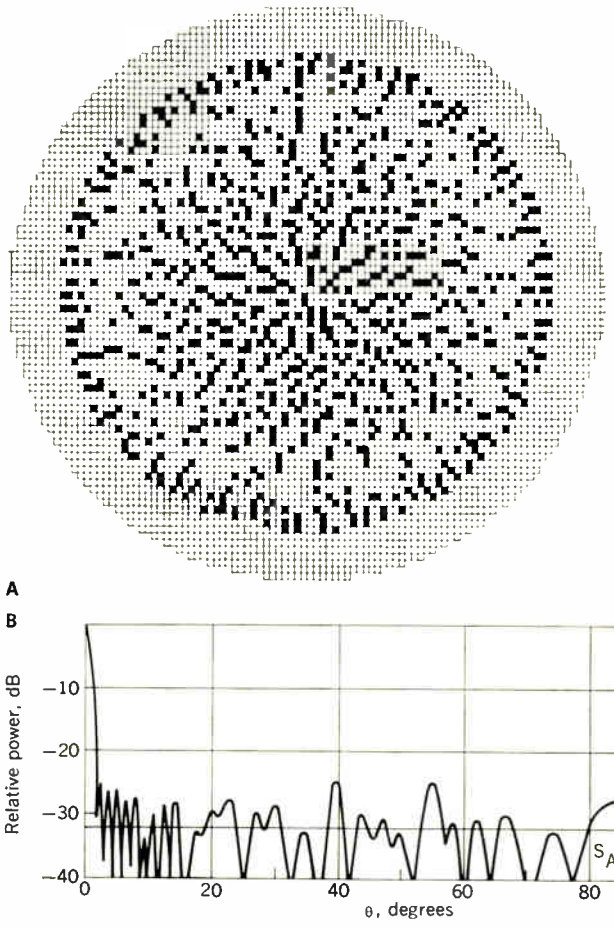
The ratio of the directivity U_d of a density-tapered array of N_a elements to the directivity of an equal-size amplitude-tapered array of N_T elements and amplitude taper efficiency η, when both arrays have the same basic element spacing (i.e., there were N_T available grid locations for the density-tapered array of which N_a were used), is

$$\frac{U_d}{U_a} = \frac{N_a}{\eta N_T} \tag{8}$$

We can interpret this ratio as indicative of the relative performance of the same basic array with the same tapering curve used as either a density plot or an amplitude plot. Subscripts are necessary on the N's because with arrays of real elements all the antennas are left in the density-tapered array, but only some are excited (the N_a active elements out of N_T total elements). For example, the array of Fig. 8 actually has 4000 antennas, but only 900 transmitters, and therefore there are only 900 active elements.

The maximum value of N_a/N_T achievable for any density-taper shape is sufficiently less than unity that a density-tapered array will have a directivity at least 2 or 3

Fig. 8. A—Density-tapered array with 4000-element grid containing 900 active elements (darkened squares). B—Typical pattern cut for density-tapered array.[20]



dB below that of an amplitude-tapered array using the same weighting function for equal N_T and element grid spacing. However, for equal N_a, the density-tapered array will have a higher directivity by the same ratio (the elements now being spread over a larger area). Thus the choice of most economic techniques hinges upon the details of the relative expense of array area and of active elements.

A rule of thumb for planar arrays with circular symmetry is that

$$\left(\frac{N_a}{N_T}\right)_{\max} \approx 1 - \frac{S - 17}{30}$$

which indicates a value of approximately 0.6 for −30-dB first side lobes.

In some applications, very narrow beams are desired, but with only modest side-lobe requirements. These can be achieved by using arrays with very large N_T, and reducing N_a until N_a/N_T is very small, keeping the same shape to the density curves. The beam width and the side-lobe levels close to the main beam will be preserved, but a price will be paid for this "thinning" in directivity and far-out side-lobe level. For example, we can achieve a circularly symmetric beam width of 0.10° with a 1000-active-element planar array on a 10^6-element grid if we will settle for an average side-lobe level of −30 dB.

Beam steering by element phasing. One of the attractive features of the array configuration is that the beam of the antenna can be pointed without moving the antenna itself. Mathematically, the technique is the last word in simplicity; practically, it is often the last word in complexity. Equation (3) indicates that if a set of element currents has been selected to achieve some desired broadside pattern, the substitution of a new set i_n' related to the old set i_n by

$$i_n' = i_n e^{jn\alpha}$$

where α is related to the angle ξ_0 to which it is desired to point the beam by

$$\alpha = -kD \sin \xi_0 \tag{9}$$

produces a change of variable (in terms of sin ξ) in (3) as

$$A(\xi, \xi_0) = \sum_n i_n e^{jknD(\sin \xi - \sin \xi_0)} \tag{10}$$

and translates the point originally at $\xi = 0$ to $\xi = \xi_0$.

The phase term of (9) represents a progressive or "linear" phase across the array and the principal result is to move the maximum of the beam from broadside to some new pointing angle. The movement is approximately one beam width for each wavelength of phase differential between opposite sides of the array.

It should be pointed out that regardless of the array geometry, some set of element phases can be found to scan the beam; however, they may not be as simply related as for flat arrays. In terms of (2), to point the beam in any direction specified by a unit vector $\mathbf{e}_{Ro}$, choose

$$i_n' = i_n e^{-jk\mathbf{e}_n \cdot \mathbf{e}_{Ro}}$$

Several second-order effects occur in linear and planar arrays that distort the pattern somewhat from its shape at broadside. The distortion can be avoided if the elements are arrayed on a surface of revolution such as a cylinder (for one-dimensional scanning) or sphere (for scanning in two dimensions). However, as flat structures still form the majority of array antennas in present use, we will briefly outline the effects of beam steering on the patterns of flat arrays.

1. Steering the beam can bring into visible space grating lobes that were not there when the beam was pointed broadside, as indicated in Fig. 9. If scanning the beam to some maximum angle θ_M from broadside is desired, the spacing between elements of a linear array and the projected spacing between elements of a planar array must satisfy (replace θ by ξ for linear arrays in our notation)

$$\frac{D}{\lambda} < \frac{1}{1 + \sin |\theta_M|} \tag{11}$$

2. As the beam scans, it broadens approximately[21] as $1/\cos \theta$. As the beam is scanned very near "end fire" ($\theta_0 = \pi/2$), the broadening ratio approaches a limiting value so that the end-fire beam width is not infinite, but inversely proportional to $\sqrt{L/\lambda}$ rather than to L/λ. In addition to

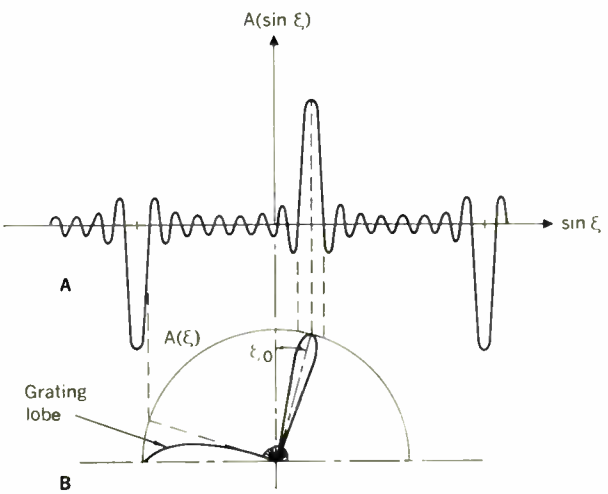
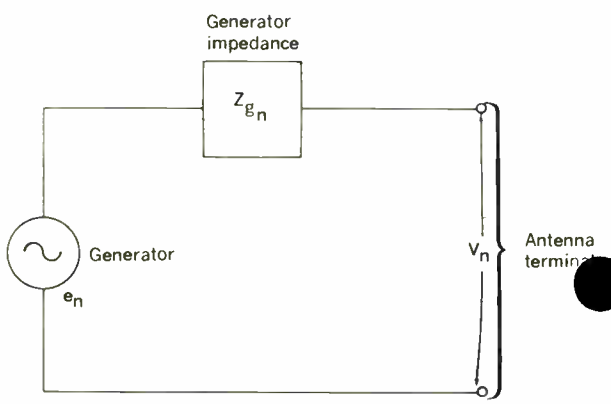


Fig. 9. Scanned-array factor vs. sin ξ, and array factor in visible space. A—Rectangular plot. B—Polar plot.

Fig. 10. Equivalent circuit for array elements that are independently excited.



becoming broader, the beam becomes somewhat asymmetrical.

3. The directivity of linear arrays of isotropic radiators is invariant to scan angle, but for a planar array as the beam is scanned off broadside, the directivity as given in (5) is modified by a cos θ_0 factor:

$$U(\theta_0) = 4\pi\eta \frac{Na}{\lambda^2} \cos\theta_0 \qquad (12)$$

4. Equation (10) is strictly applicable to only CW (continuous wave) operation of an array, and it was under these conditions that it was indicated that phasing the elements scans the beam. However, from a receiving point of view, when transients such as short pulses are incident on the array from some angle other than broadside, the transient requires a finite time to traverse the array and, hence, reaches the different antennas at different times. The result is that the output of the array may be a distorted reproduction of the original transient, even if the array is properly phased to "look" in the direction of the incident signal for the center frequency. Put another way, inspection of the phasing term of (9) required to steer the beam shows that it is frequency dependent, since $k = \omega/c$. A nondispersive delay line has the proper phase change with frequency to accommodate wide-bandwidth signals, and insertion of such a device behind each antenna will provide the correct phase variation with frequency so the beam will point in the same direction for all frequencies. The resultant configuration is usually called a "delayed array," as distinguished from a "phased array." The distinction becomes important in the handling of signal bandwidths that are approximately equal to the reciprocal of the transit time of a signal across the projected depth of the array; for example, a 10 per cent instantaneous bandwidth signal at 1000 Mc/s begins to suffer distortion at wide scan angles in phased arrays of 10 feet or greater.

Arrays of real radiators

In actual practice, arrays must be constructed of directional, interacting antennas. It is therefore necessary to consider not only the extent to which the patterns of an array of real radiators differ from the patterns of an array of noninteracting isotropic radiators, but also the impedance-matching problems that the circuitry must accommodate.

Far-field pattern considerations. To assess the pattern effect of real radiators, a specialized but practically important type of array will be considered: an array excited by a network that permits no coupling between the elements via the feed network, so that the only interaction existing between elements must take place via the antennas. Although this class is not all-inclusive, it is the type frequently used when good control of patterns is required.

We can model the excitation of a typical element of such an array by a circuit such as that of Fig. 10. Each element is excited by a voltage generator of controlled open-circuit voltage e_n for the nth element and fixed internal impedance Z_g. We can then turn on each generator individually with the others turned off ($e_m = 0$, $m \neq n$) so that the others are effectively terminated in an impedance Z_g. With only $e_n \neq 0$, a current i_n will flow past some reference point on the nth antenna, exciting the nth antenna directly and parasitically exciting the rest of the array through

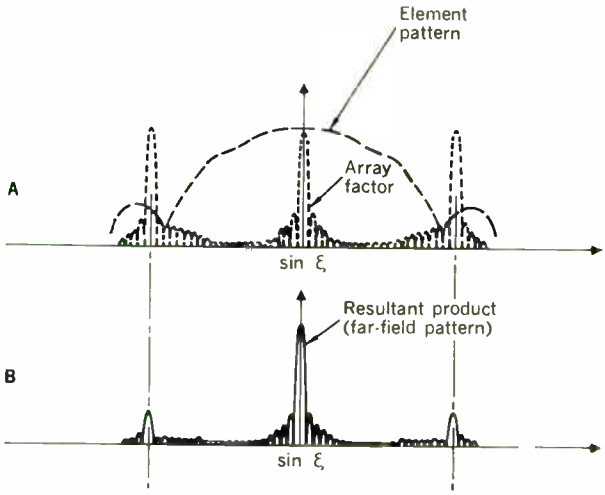


Fig. 11. A—Comparative structure of element pattern and array factor of an array having a large number of elements. B—The resultant far-field pattern.

mutual coupling. A measurement of the far field under these conditions produces some relative field strength which we will call $\mathbf{f}_n(\phi, \theta)$, since it is due to the nth-element generator. The experiment can be repeated for each element and the far field can be found by taking the vector sum of the individual element contributions, as in (1), properly weighted by a phase term dependent upon the element location:

$$\mathbf{F}(\phi, \theta) = \sum_n \mathbf{f}_n(\phi, \theta)\, i_n e^{jk\mathbf{e}_n \cdot \mathbf{e}_R} \qquad (13)$$

For completely arbitrary distributions of elements, little more can be said. However, for the important case in which (a) the array is large compared to the region over which element interaction is strong so that edge effects can be neglected (dependent upon the type of element used*), (b) the elements and their generator impedances are nominally identical, and (c) the elements are regularly spaced on a flat† surface so that all interior elements "see" the same interaction environment, the pattern of each element measured as indicated above will be essentially the same as the pattern of any other element. We can then factor (13) and write

$$\mathbf{F}(\phi, \theta) = \mathbf{f}(\phi, \theta) \sum_n i_n e^{jk\mathbf{e}_n \cdot \mathbf{e}_R} \qquad (14)$$

where $\mathbf{f}(\phi, \theta)$ is a typical $\mathbf{f}_n(\phi, \theta)$.

The factor $\mathbf{f}(\phi, \theta)$ multiplying the summation in (14) is the pattern of a typical (central) element in the presence of all the other elements when they are terminated in the impedance from which they are normally excited, and is usually referred to as the element factor. The i_n of (14)

* For dipoles above a ground plane, approximately the outer two rows of elements differ significantly in patterns from the interior elements. Elements with broader "free space" patterns, such as dipoles without a ground plane or thin slots, may require four or five rows to "settle down" to a repetitive pattern.

† The modification for large, regularly spaced arrays on curved surfaces with large radii of curvature (compared to λ) is messy, but not conceptually difficult.

represents the current flow in the nth element due solely to the excitation of its own generator. For all elements (neglecting edge effects), i_n is related to e_n of Fig. 10 by the same proportionality independent of the other element excitations, and we could, in fact, replace i_n by e_n in (14).

The practical significance of this development lies in the fact that the summation of (14) is exactly the array factor of an array of isotropic radiators as in (2), the properties of which were extensively discussed previously. The element factor contains essentially all the pattern effects of the element type and the interaction between elements. Since it is the pattern of the entire array when only one element is excited and the others terminated, it is much broader than the pattern when the entire array is excited; hence, $f(\phi, \theta)$ is a function whose variation with angle is slow compared to any reasonably directive array factor in large arrays. The patterns are similar to those of Fig. 11. The element factor acts as a window of a varying degree of opaqueness through which the array factor "looks." We can qualitatively conclude that under normal circumstances in large arrays the element factor structure will have little effect on the relative structure of the main beam and close-in side lobes of a large array. However, in arrays in which the beam is scanned by element phasing, the absolute strength of the pattern will be varied with angle to conform to the element factor weighting, since the scanning moves the array factor in space while the element factor is stationary.

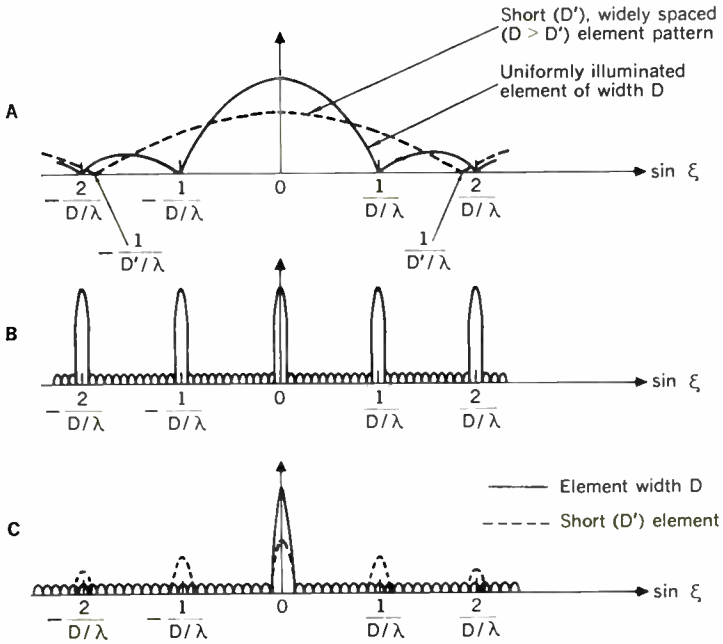
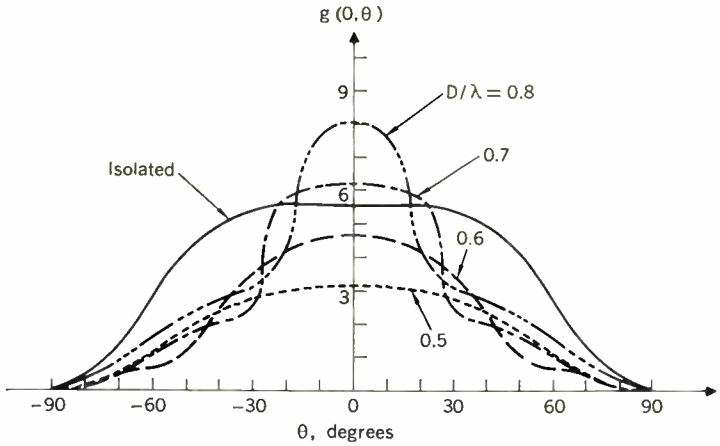


Fig. 12. Relationship between element illumination and grating lobe suppression capability. A—Element pattern. B—Array factor for element spacing D. C—The resultant far-field pattern.

Fig. 13. Computed H-plane gain functions for dipole λ/4 above ground isolated and imbedded in array with square grid spacing D. The generator impedance has been chosen to maximize g(0, 0) for each value of D.



For element types in which mutual impedances can be analytically determined (at the present state of knowledge, only dipoles and slots[2,22,23]), the computation of the element factor is straightforward. Perhaps more important, in the frequent case of elements for which computational formulas do not exist, the element factor can be determined experimentally by building an array only large enough to render edge effects negligible on the pattern of the central element. Finally, we can also rationalize some generally valid, quantitative conclusions about element factor shapes.

Of course, if the antenna elements are separated far enough that the interactions are negligible, the $f_n(\phi, \theta)$ become the patterns of isolated elements (which are given in the literature for a wide variety of radiators). However, only when the elements are separated by distances that are much greater than their largest dimension is it safe for us to assume that interaction effects are negligible. Unfortunately, achieving freedom from interaction in this way results in spacing the elements sufficiently far apart that several grating lobes of the array factor appear in visible space, as indicated in Fig. 12. The grating lobes can now be "smeared" by irregular spacing of the N available radiators, but a close-in pattern that is the same as predicted by isotropic theory will be achieved only if essentially all the element separations are such that interactions are negligible.[24]

The more interesting case for most array applications is that of closely spaced elements—in particular, the case of only slightly directive elements spaced less than a wavelength on centers. Here some element factors shapes can be inferred from a consideration of the gain and directivity variation of electronically scanned arrays.

The array gain G is a measure of efficiency of the array in using all the energy available, and involves questions of array impedance matching and ohmic losses. (We have previously concerned ourselves with an examination of array directivity, which deals with the efficiency of the array relative to the energy that it radiates and therefore is completely specified by the far-field pattern. Although mismatch effects are not always included as a factor in the gain of an antenna, they will be included in this discussion.) Such effects are accounted for by defining the gain as the ratio of power density per unit solid angle at the peak of the main beam to the power density that would be achieved by radiating all the available power isotropically. For electronically scanned arrays, the variation of gain with beam pointing angle ϕ, θ is specified for a large, flat, equally spaced array of identical radiators by

$$G(\phi, \theta) = g(\phi, \theta)\eta N_a \qquad (15)$$

where, taking into account the possibility of density tapering, N_a is the number of active elements and η is the amplitude taper efficiency. The factor $g(\phi, \theta)$ is referred to as the element gain function and its spatial variation with the beam pointing angle ϕ and θ is just $|f(\phi, \theta)|^2$. Equation (15) is the quantitative statement of the fact that the element factor acts as a weighting function on the field strength of the array.

We have previously established that the directivity of a planar array of N_a active radiators with no visible grating lobes is, from (12),

$$U(\theta) = 4\pi\eta \frac{N_a}{\lambda^2} a \cos\theta$$

where θ is the angle from the array broadside direction (see Fig. 7). For independently excited radiators with no ohmic loss, which can support only a single polarization, the array directivity and gain can differ only as the result of mismatch losses. These losses can be accounted for, in terms of the reflection coefficient $\Gamma(\phi, \theta)$ seen when looking into a typical element when the entire array is excited, by writing

$$G(\phi, \theta) = U(\theta)\,(1 - |\Gamma(\phi, \theta)|^2) \qquad (16)$$

Substitution of (12) and (15) into (16) sheds some light on the element factor behavior to be expected as a function of element spacing

$$g(\phi, \theta) = 4\pi \frac{a}{\lambda^2} \cos\theta\,(1 - |\Gamma(\phi, \theta)|^2) \qquad (17)$$

The only place the type of radiator enters the above expression is implicitly in $\Gamma(\phi, \theta)$. Regardless of type, we can choose the generator impedance to match the element at some angle of scan of our choice. If, for the sake of illustration, we match the elements when the array is phased for a broadside beam, we can make the following generalization about the element pattern for small (less than λ in dimension) elements from Eq. (17): As we space the elements further apart, the value of $g(0, 0)$ can be made to increase directly with the area per element if the right generator impedance is chosen for each spacing. However, as the element spacing is increased, the angles at which grating lobes becomes visible decrease. At these angles, the array directivity must drop rapidly by approximately 3 dB and the gain function must behave in a similar fashion (we are not matched at these angles, so only approximate statements are justified). That the element pattern does behave in such a manner has been verified analytically for some elements and experimentally for others.[25,26] Figure 13 shows element patterns of dipoles above a ground plane for spacings illustrating the effects stated. If the elements are not matched to maximize $g(0, 0)$ for each spacing, the result is a decrease in broadside gain and, usually, a slight increase in the width of the element factor main lobe.

For arrays on curved surfaces, if the curvature is shallow compared to the extent over which appreciable coupling takes place, the element gain function shape relative to each element's own broadside angle will be similar to those of an equivalently spaced planar array. Thus, even though (14) fails to apply, the element patterns can usually be taken into account in a straightforward and relatively simple manner.

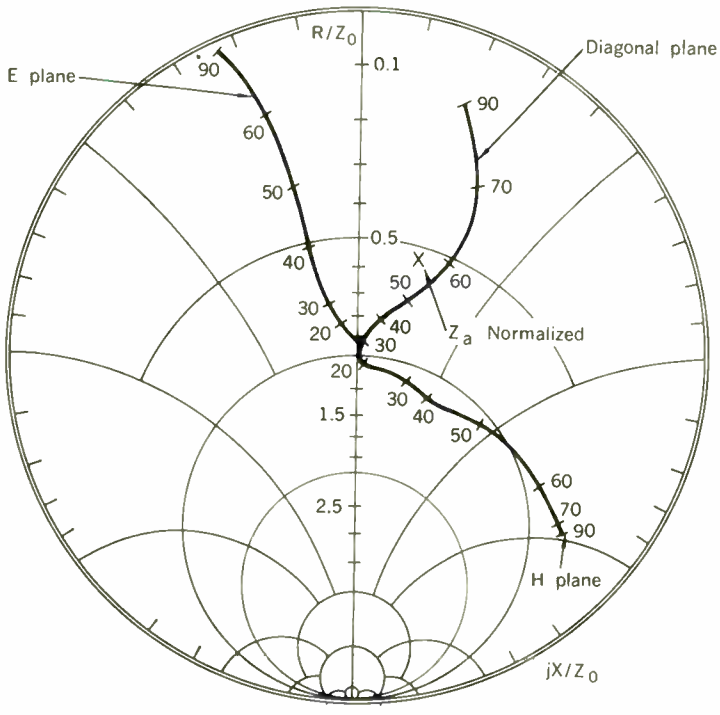
When array elements are unequally spaced, the $f_n(\phi, \theta)$ are not, in general, alike, and (14) is not applicable.[24] Furthermore, since the element patterns differ in a changing manner with the angles ϕ, θ, the patterns of such arrays may vary significantly with scan angle.[27]

In antenna arrays having elements that are capable of supporting more than one polarization, the effects of coupling also may cause the polarization of the antenna to be altered.[28]

Interaction effects on element impedance. If two antennas are individually matched to their respective generators, so that no reflected waves exist in the feeds, and are then excited and brought into proximity, part of the energy radiated by antenna 1 is picked up by antenna 2 and manifests itself as a wave traveling towards the generator of antenna 2 and vice versa. If the excitations are coherent, then as the relative phases of the excitation of the two antennas are changed, the phase of the forward wave (representing the energy traveling from the generator to be radiated) and the reverse wave (representing the energy from the other antenna as received) will change. The result at the generator terminals looks exactly as though the impedance of the antennas has changed from its value when only a single antenna was present and is continuing to change as the relative phase of the excitations is changing, producing the reflection coefficient $\Gamma(\phi, \theta)$ of (16) and (17).

This apparent impedance is precisely the impedance that any transmitter or receiver should match to achieve maximum system efficiency; hence, we will call it the ele-

Fig. 14. Calculated variation in driving impedance vs. angle of scan for a typical element of an array of dipoles $\lambda/4$ above ground on a square $\lambda/2$ grid. The numbers represent degrees of scan from broadside in indicated plane.



Z_D (0, 0) = 150.3 + j42.3
Z_a = 85.7 + j72.5

ment driving impedance $Z_D(\phi, \theta)$. Furthermore, because of the variation in impedance with scan angle in scanning arrays, the transmitters and receivers must work into a varying mismatch most of the time. The usual effects of such mismatches may occur: loss of radiated power and breakdown effects caused by standing waves in transmitting arrays; degradation of noise figure in receiving arrays; and possible instabilities in both transmitters and receivers because of source and load mismatches.

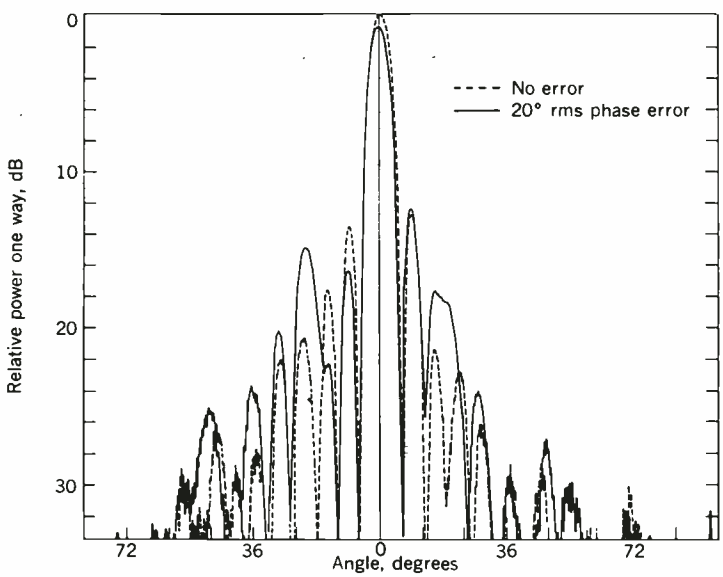
What sort of impedance behavior can the array designer expect with each possible type of antenna and how can he utilize that antenna most effectively to minimize the mismatch effects? This knowledge would intrinsically provide an answer to another question: Of all the elements satisfying a designer's requirements with respect to such properties as physical size, polarization, and ease of feeding, which would yield the minimum change in impedance with scan angle in an array?

Unfortunately, the present state of knowledge does not permit quantitative answers to either of the foregoing questions. As a compromise we will attempt to present a summary of known facts and qualitatively extrapolate this information as far as discretion permits.

The analytical determination of Z_D requires analytical expressions for the mutual impedance, and such results are limited to simple elements. Moreover, only if the array is large, regularly spaced, and smoothly amplitude- and phase-tapered is $Z_D(\phi, \theta)$ the same for essentially all elements. Fortunately, these restrictions apply to a large fraction of actual arrays, so the case is of practical interest. It also sheds some qualitative light on the impedance behavior of arrays of other types.

An example of the behavior of a central element of a planar array of dipoles mounted $\lambda/4$ above ground on a $\lambda/2$ square grid is shown in Fig. 14.[8, 26] The plot has been normalized to the impedance $Z_D(0, 0)$ of the element in the array phased for broadside radiation. Shown also is the impedance Z_a of a single, isolated dipole similarly mounted above a large ground plane. The impedance mismatches indicated are quite severe (e.g., a voltage standing-wave ratio of 2.5 to 1 at a 50° scan) and can be made even worse by failure to use the elements optimally. For at least some elements, design parameters can be found that minimize the mismatch with scan angle—e.g., the height of dipoles above a ground plane which is optimized in Fig. 14 for the spacing.[26] Furthermore, the choice of element spacing changes the mismatch with scan, and while individual interactions generally weaken as the elements are separated, the total interaction (the vector sum of the individual interactions) may actually be smaller for closer spacings for the typical scanning array spacings $(0.5 < D/\lambda < 1.0)$.[29]

Fig. 15. Typical effects of random errors on the pattern of 16-element uniformly illuminated array.



On the bright side, two effects frequently contribute to lessening the mismatch achieved in practice. First, any loss in the element feed reduces the VSWR, at the expense of array gain, since the incident wave must traverse the loss only once while the energy in the reverse wave travels through it twice. Second, if the elements will support more than one polarization, the coupled energy may be converted to the polarization sense orthogonal to that from which it was excited,[28] in which case the coupling does not show as a mismatch to the generator. However, it still represents lost available energy (gain) if absorbed in a load, and a polarization error if reradiated.

General interaction effects. Information about more complex elements than dipoles must be inferred from experiment or by indirect analytical means. Such data are rather sparse and qualitative.

There are indications that the gain-function behavior of dipoles—that is, the patterns are smooth and reproduce well from element to element—is typical of simple antennas such as dipoles, slots, and open waveguides. However, many radiators have appreciable metallic or dielectric structure, and Bates[30] has shown that this can result in an increase in the index of refraction in the immediate vicinity of the antennas, causing the electrical spacing between the antennas to be greater than that predicted on the basis of the free-space wavelength. The result is a tendency toward the "sharpening" of the element factor as indicated in Fig. 13, even for quite close spacings (measured in free-space wavelengths), and possibly increased impedance variation with scan for a given element spacing. Thus, we can only offer the observation that the rather sparse and fragmentary experimental data seem to indicate the advisability of using geometrically simple antennas.

Numerous attempts[27, 31] have been made to infer the impedance variation for arbitrary element types, without ultimate success. Some insight has been gained in the process, however. For example, Hannan[31] has demonstrated a constraint on the average reflection loss over all possible array phasings for a planar array, and he indicates that it would not be inconsistent with the constraint to have a zero reflection coefficient over an extended region of scan and a total reflection outside this region. The possibility of using compensating networks in the array feed[32] to accomplish this goal has been investigated. A search for an element radiation behavior that could result in such reflection coefficients without compensating interconnections has thus far led to theoretically informative but practically useless results—for example, a linear array of elements of such a type that the magnitude of the coupling between two elements is independent of the separation. Unfortunately, this form of mutual cou-

pling is representative of a surface wave that does not radiate.

Effects of illumination errors

We have confined our discussion up to this point to the case of "error-free" arrays. However, knowing what we want is one thing, but getting it is something else. There are obviously many sources of errors, both mechanical and electrical, in an array. These can be resolved at any angle into equivalent errors in the amplitude and phase of the illumination currents, and we will confine our explicit discussion to such errors.

It is convenient to divide the illumination errors into two categories: systematic errors, to which some definite pattern can be ascribed; and random errors, which are those defying an a priori description of any pattern. The effects of systematic errors are usually easily determined once the pattern of the error is recognized, but a separate analysis must be performed for each type of systematic error. Random errors can be treated by statistical methods, and one analysis suffices, with the reservation that the result of the statistical analysis is a statistical answer. In this respect, the "neatness" with which one can analyze random errors is somewhat illusory.

Random-error effects. Figure 15 illustrates the deleterious effects of random illumination errors: the relative side-lobe levels are increased (on the average), the gain and directivity are decreased, and the beam pointing direction changes.

Small errors usually are most noticeable as an increase in side-lobe level in the regions of the pattern where the "no-error" side lobes are extremely low; in fact, in most arrays, the pattern structure in these regions is attributable almost entirely to errors. It can be shown[33,34] that for phase errors and amplitude errors that are independent of each other and independent from element to element, the side-lobe level at any angle to be expected from an ensemble of antennas with identical error statistics* is distributed so that in the regions where error effects predominate and the element factor is near its peak value, the ratio R of the main beam power density to that of the average side-lobe power density is given by

$$R = \frac{\eta(PN_a)}{\epsilon^2} \tag{18}$$

where ϵ^2 is related to the mean-square phase error σ_ϕ^2 (in radians), the mean-square fractional amplitude error σ_A^2, and the fraction P of elements actually operating, by

$$\epsilon^2 \approx (1 - P) + \sigma_\phi^2 + \sigma_A^2 \tag{19}$$

Since the side-lobe level fluctuates about this average value, it is to be expected that side-lobe peaks of several times the average will occur.

Equation (18) can also be rewritten to make explicit the dependence of the allowable error on the no-error directivity of the array by noting that for a planar array with nominal $\lambda/2$ spacing, substituting (6) with $\alpha = \lambda^2/4$ into (18) and solving for the allowable error gives

$$\epsilon^2 \approx \frac{P}{\pi}\left(\frac{U_0}{R}\right) \tag{20}$$

* That is, we conceptually construct a large number of identical antennas, and then insert into each errors from the same probability distribution.

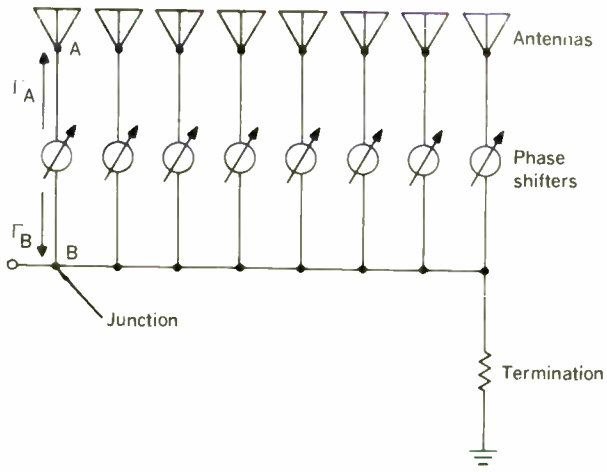


Fig. 16. Serial feed with branch-line phase shifters.

Thus, the mean-square error determines how far below the no-error directivity of the antenna one can expect to find the far-out side lobes. For example, to hold the mean-square side-lobe level 10 dB below the no-error directivity ($U_0/R = 0.1$) requires $\epsilon^2/P \approx 1/10\pi$. Thus, in the optimistic case of *no* amplitude error and *no* failed elements, an rms phase error of less than 0.2 radian ($\approx 12°$) is allowable. Therein lies the difficulty in achieving average side-lobe ratios well below the no-error directivity in the region of an array pattern where the element factor is not small.

The effect of errors on the array directivity and an optimistic estimate of the gain decrease (since random impedance mismatches may add to the gain degradation) can be expressed as[5]

$$\frac{U}{U_0} \approx \frac{1}{1 + \epsilon^2/P} \tag{21}$$

This equation, in conjunction with the previous one, makes obvious the fact that if the average side-lobe level becomes comparable to the no-error directivity, the actual directivity of the antenna will begin to decrease rapidly with increasing ϵ^2. For smaller values of ϵ^2, the effect on directivity is trivial.

To a first order, the beam pointing of the array is independent of amplitude errors and depends only upon the mean-square phase error. An approximate relationship for the error in fractional beam widths is[5]

$$\frac{\delta\theta}{\theta} \approx 0.6\,\frac{\sigma_\phi}{\sqrt{N_a P}}$$

where the numerical value of the constant actually depends on the no-error amplitude taper, but it is usually in the range of 0.5 to 1.

A sampling of systematic errors and their effects. To impart some feeling for the effects of systematic errors in array feeds, a representative sampling is given by the following examples:

1. One frequently troublesome class of systematic errors arises from the inevitable mismatches in arrays, particularly the systematic mismatches due to antenna impedance variations with scan. For example, Fig. 16

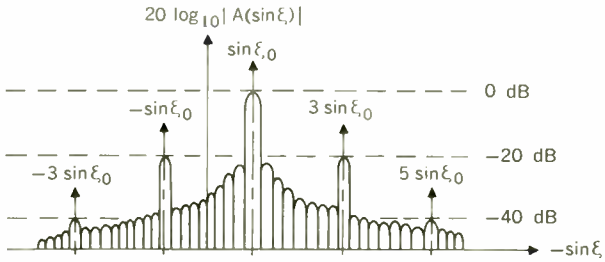


Fig. 17. Typical pattern of an array utilizing serial feed, with $|\Gamma_A\Gamma_B| = 0.1$.

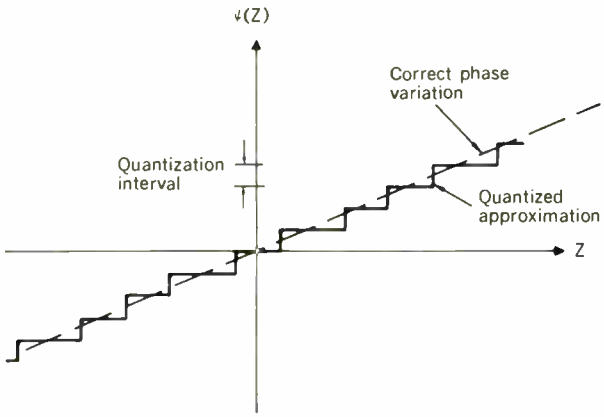


Fig. 18. Typical stepped phase distribution resulting from the use of discrete phase shifters.

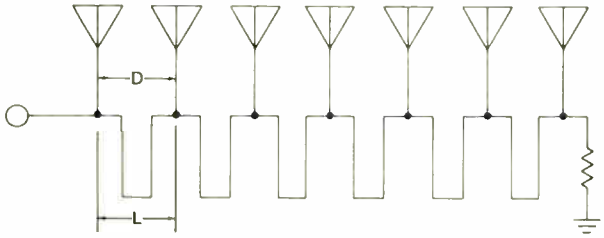
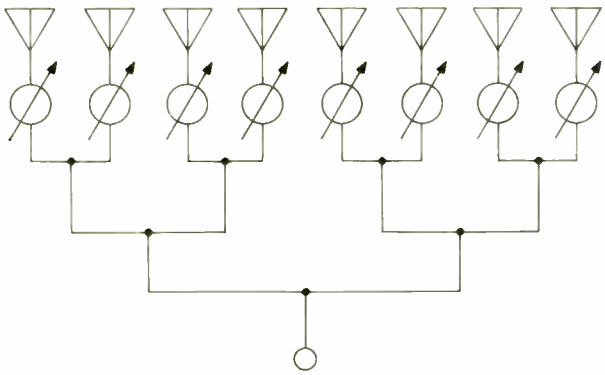


Fig. 19. Frequency scan feed.

Fig. 20. Parallel (corporate) feed.



shows a schematic of a common feed structure for arrays in which energy is fed down a main line and tapped off at junctions to be fed through phase shifters to antennas. If points A and B represent reflections Γ_A and Γ_B, respectively, to an observer looking at these points from the phase shifter, energy will traverse the phase shifter many times as the result of multiple reflections. If the phase shifter is a reciprocal device (the phase shift is the same for energy traveling in either direction), the result[35] for a phase shifter setting of α is an antenna excitation consisting of a first wave of phase α, a second of phase 3α and complex amplitude $\Gamma_A\Gamma_B$ relative to the first wave, a third of 5α and amplitude $|\Gamma_A\Gamma_B|^2$, etc. The effect of these reflections on the antenna pattern is shown in Fig. 17. The spurious side lobes are produced at odd multiples of $\sin\xi_0$ and are suppressed by powers of the product $\Gamma_A\Gamma_B$. The numerical results shown in the figure are for $|\Gamma_A\Gamma_B| = 0.1$. Since reflection coefficients of up to 0.3 may be likely at point A because of mutual coupling, it is apparent that extremely good impedance matches must be held at point B to prevent such pattern distortion. If nonreciprocal phase shifters are used, however, the sum of the phase shift to a forward wave and to a backward wave is a constant, and the reflections represent only a variation in gain with scan without relative pattern distortion.

2. An observer looking down the main feed line of Fig. 16 may see a different sort of impedance problem because each of the junctions will inevitably represent a small mismatch. If all the mismatches are identical or have a component that is identical from junction to junction, it is possible in the case of some spacings between the junctions for all reflections to add in phase, thus presenting a high standing-wave ratio at the feed point of the main line. This phenomenon, which is termed feed "resonance" because of its critical dependence on the phase of the reflections, can be particularly troublesome if the electrical length between junctions is varied, as in frequency-scanned arrays. The total reflection coefficient at resonance may be of the order of N times the reflection coefficient at each junction for an N-element array, and thus a substantial mismatch is presented to any transmitter or receiver.

3. In view of hardware considerations, it is sometimes necessary to make the phasing of the elements take on only discrete values. The two most common causes are the use of step (discrete) phase shifters, or grouping of small numbers of elements together and driving them from a common phase shifter—using a group of small elements as a single, more directive element, a technique commonly called "subarraying." This results in the phase front across the array being stepped, as indicated in Fig. 18. Qualitatively, the resulting pattern distortion is similar to that indicated in Fig. 17. If it is caused by discrete phase shifters, a useful rule of thumb[36, 8] is that the amplitude of the first spurious lobe will be reduced in voltage by approximately $1/n$ (or in power by $1/n^2$) for phase shifters that have a smallest quantization interval of λ/n. The location of the first spurious lobe will be nB beam widths from the main beam, when the latter is pointed B beam widths from broadside. If the step phase front is caused by grouping the N elements into M subarrays, the first spurious lobe will be down approximately B/M in voltage and will be M beam widths removed from the main beam.

Some philosophy with respect to error effects. The

most obvious effect of illumination errors is upon the side lobes of the antenna. Generally speaking, systematic errors give rise to a low side-lobe level over most of space, but with a few relatively high side lobes. Nonsystematic errors—those that are more nearly random in nature—tend to raise the side-lobe level everywhere. There is a strong tendency in the antenna business to specify side-lobe performance goals in terms of the highest permissible side-lobe level; consequently, attention has been directed[37] to the possibility of lowering the few high side lobes accompanying systematic errors, at a cost of a higher average side-lobe level, by deliberately introducing additional pseudorandom errors into the array illumination. The additional error effects can then be offset by increasing the number of active elements.

The wisdom of such a move is perhaps questionable. Since the array beam width depends principally on the array dimensions and not the number of elements (if we allow the possibility of density tapering or "thinning"), the only factors for determining the number of active elements to be used in an antenna are the desired gain and the side-lobe level. If both gain and low side lobes are important, increasing the number of elements is certainly a valid approach to lessening error effects, in that a double payoff of decreased side lobes and increased gain is achieved. However, in some situations (e.g., radar receiving antennas in a jamming environment), main beam gain per se is of little importance but the side-lobe levels are extremely important in discriminating against side-lobe jamming. In these cases, decreasing the errors by a factor of two is equivalent in side-lobe level effect to increasing the number of elements by a factor of four. Therefore, in such applications it may be more economical to tighten the error tolerance on components than to increase the number of elements to achieve required performance.

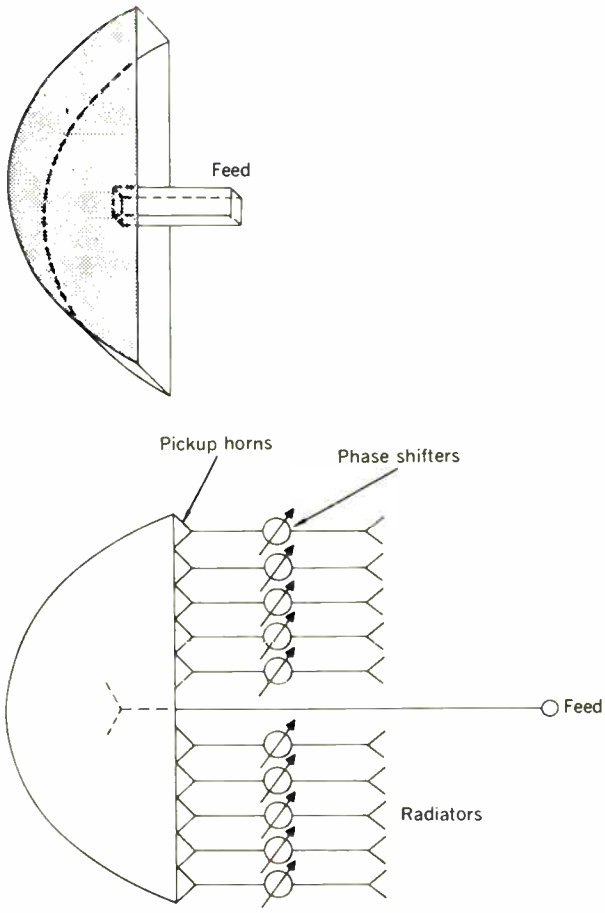


Fig. 21. Example of array feeding by optical techniques. Top—Basic pillbox antenna. Bottom—Pillbox used as an array feed.

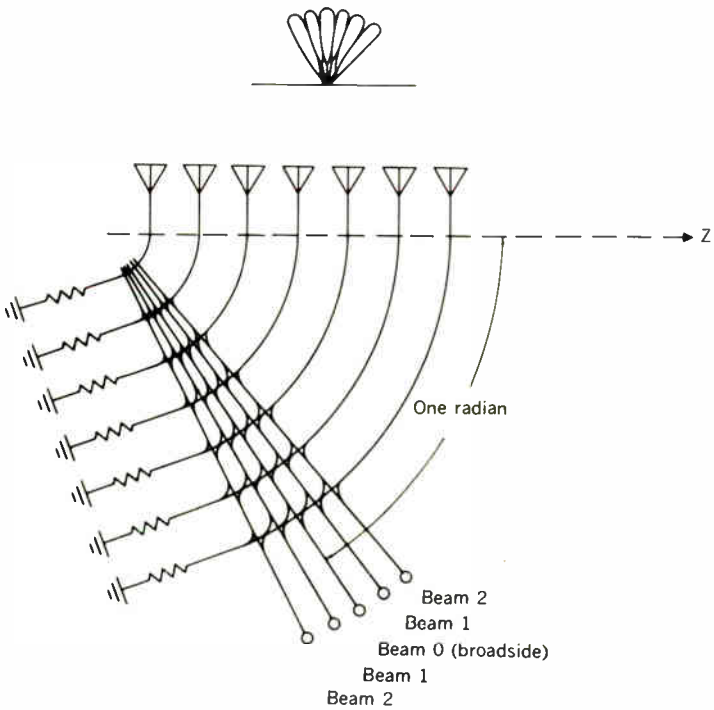
Representative feed networks

We have still left unanswered the question of how the desired array excitations can be obtained. In this section, some representative types of feed networks will be surveyed and some of their principal advantages, disadvantages, and limitations will be indicated.

First of all, a distinction should be made between active-element and passive-element arrays. The former use a transmitter or receiver, or both, at each antenna element; the latter use only one transmitter and receiver for the entire array. Between the two extremes, there is, of course, a spectrum of hybrid arrangements. The active-element array alleviates some feed network problems, such as requirements for high power-handling capability in transmitting arrays and low losses in receiving arrays. Furthermore, through the use of mixers at each element, feed networks can be constructed at low frequencies, at which lumped circuit techniques can sometimes be employed to advantage. However, this freedom is purchased at a cost of great redundancies of active electronics. Rather than indulge here in a discussion of the relative merits of the two approaches, we will concentrate on techniques for passive-element arrays and preserve neutrality with the observation that most active-element feed techniques have close parallels in techniques for passive-element array feeds.

Figure 16 shows a schematic representation of arrays in which the energy to excite each antenna is tapped off

Fig. 22. Serially fed multiple-beam feed.



serially from a main feed line. Such arrays have the advantages of simplicity and ease of manufacture, but require care in the design of the junctions to avoid multiple-reflection problems and resonance effects. As pictured in the figure, they are also narrow-band feeds for broadside arrays (but can be broadband feeds for end-fire arrays), since the beam pointing direction varies with frequency, unless compensation by the phase shifter is feasible and is used. (Wide instantaneous bandwidth signals represent an example where it is not feasible.) This frequency sensitivity is often a curse, but it can be a blessing for some narrow-band applications. If the length of line L between each element is increased without increasing the element spacing D, as indicated in Fig. 19, an array that scans simply by change of the radiated frequency can be built. Such a configuration probably represents the simplest method of achieving an electronically scanned array, though a narrow-band one.

For wider bandwidth applications, one can resort to a parallel feed technique by use of either transmission-line devices, such as the corporate feed of Fig. 20, or what might be termed optical devices, such as the pillbox feed of Fig. 21. The bandwidth of such feeds is limited only by the bandwidth of the network components and the bandwidth limitation of any steering (phasing or delaying) technique used. The transmission-line feeds are more complex to manufacture in any quantity and are not as compact as the serially fed type. The optical feeds largely overcome these disadvantages, but at a cost of decreased control over the illumination. Although "resonance" effects are not a problem, the multiple-reflection problem exists in parallel as well as in serial feeds and forces careful control of junction mismatches.

The foregoing techniques are designed for the generation of a single beam and, with the addition of phase shifters or delays, scanning this beam in space. A distinct class of devices is that of passive multiple-beam antennas which form a large number of simultaneous beams, each beam having a gain essentially equal to that attainable from the same array arranged to form only a single beam.

The earliest technique in this class is shown, in a form slightly modified from its original conception,[38] in Fig. 22, and is related to the serial feed of Fig. 16. Each antenna is connected to a transmission line that is terminated at the other end by a load. Between the antennas and the terminations are placed crossbars with directional couplers at each junction. The crossbars are also terminated. If one does not attempt to get the beams too close together, the transmission coefficients of the couplers can be chosen so that beams are formed with little interaction between the feeds and little power is absorbed in the terminations. In particular, if the couplers are designed to give a uniform array illumination, it has been shown[39] that the minimum fractional power lost in the loads is only $1/(N + 1)$ for an N-element array feed that is producing N simultaneous beams. The curvature of the antenna bars shown in the figure allows the primary paths (those changing direction only once) to be made equal from the broadside beam terminal to all antennas, and if only primary paths contributed to the operation of the device, it would be a time-delay device with beam point-

Fig. 23. Four-element parallel multiple-beam-forming feed.

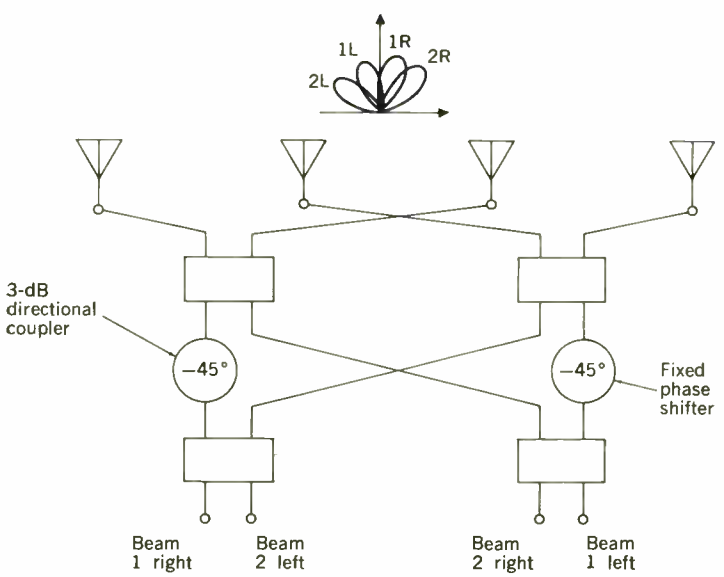
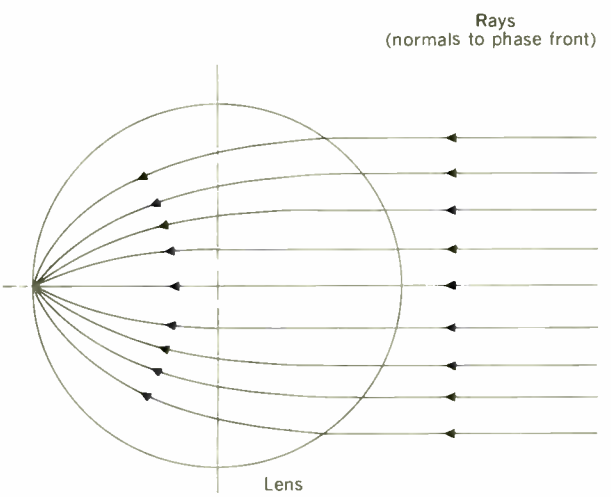
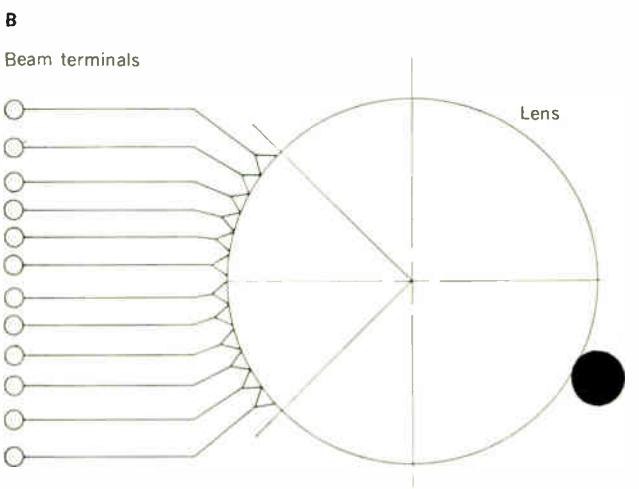


Fig. 24. Multibeam array using Luneburg lens. A—Focusing action of lens. B—Array.





ing directions independent of frequency. However, when the device is designed for maximum efficiency, the higher-order paths contribute substantially and the device is not really very broadband. Only when the array is designed for lower efficiencies can larger bandwidth be obtained. The previously referenced study[39] indicates that, for example, when 90 per cent of the available power is absorbed in the loads, a bandwidth of the order of 50 per cent is achievable from moderately large ($N < 100$) arrays. The larger the array for a specified efficiency and the larger the number of beams formed, the less the bandwidth.

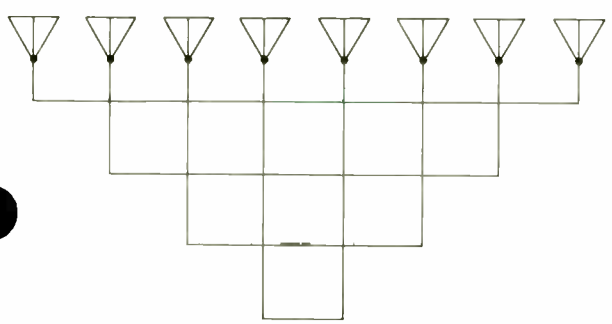
The serial-fed multiple-beam matrix can be conveniently packaged in a waveguide using cross-guide couplers if the array is large enough that only light coupling is required. For small high-efficiency arrays, it becomes difficult to achieve the tight coupling in a reasonable layout.

A parallel multiple-beam configuration,[40,41] as shown in Fig. 23, can be designed for any binary number of elements. This matrix is ideally lossless; that is, the beam terminals are decoupled from one another so that all power put into the device is radiated. The illumination is intrinsically uniform and the device produces N beams of the sin Nx/sin x array factor shape, crossing over at about 4 dB below the beam peaks. All the couplers in the device are 3-dB couplers and, for small arrays where the layout problem is manageable, the device can be conveniently built of strip transmission line. For large arrays or high-power applications, layout again becomes a problem. Beams of shapes other than sin Nx/sin x can be synthesized by combining the intrinsic beams from such a matrix.[36] However, if the lossless features are to be maintained, the crossover levels must be decreased as the side-lobe levels are reduced.

A different type of multiple-beam-forming antenna is achieved by use of multiple feeds in an optical-type system, such as the Luneburg lens configuration of Fig. 24. Achieving good patterns with this type of technique makes it necessary to cope with the mutual coupling effects in the same way as in a conventional array. The primary (feed) pattern that is transformed by the optical system is approximately the same as that of the excited element in the presence of the other elements, not the pattern of a single isolated feed.[42] An optical feed is often realized more simply than is a transmission-line feed, but some control over the illumination is sacrificed.

All multiple-beam antennas have restrictions on the patterns that are realizable without loss.[43-45] First, the beams must cross over at fixed levels that are dependent only on beam shape but are always 4 dB or more below the peaks for low side-lobe tapers; as the illumination is tapered, the crossover levels must go down farther, or a loss must be accepted. Second, a device with constant beam-pointing direction independent of frequency cannot be completely lossless, which is actually a corollary of the first restriction.

Fig. 25. Schematic representation of a Van Atta array. All of the interconnecting lines are the same length.



For applications such as satellite repeater stations, a useful antenna array is one that can make maximum use of all of its intrinsic gain in receiving a signal from a particular direction and returning the signal in the same direction over wide angles. These arrays, referred to as retrodirective arrays, come in essentially two classes: passive retrodirective arrays and active, or self-focusing, arrays. The Van Atta array[46] shown in Fig. 25 is an array which passively redirects the received power in the direction from which it was received with a gain, on both transmit and receive, equivalent to the projected aperture of the antenna. Self-focusing arrays[47] make use of active electronics behind each element to receive a signal, amplify it, and automatically phase it and return it in the direction from which it was received by use of conjugate networks. (These networks operate on a signal $Ae^{j(\omega t + \theta)}$ to produce $Ae^{j(\omega t - \theta)}$; a double mixing process can perform such an operation.[47])

Present and future uses of array antennas

It seems likely at this point that the revival of interest in arrays will last and that such techniques will be even more widely used. There are now several arrays of small numbers of large parabolic dishes for radio astronomy. There is also consideration of arrays of large numbers of parabolic dishes for radio astronomy applications, perhaps with randomized spacing to smear out the grating lobes. There are several low-frequency dipole arrays for radio and radar astronomy that operate in the region of a few tens of megacycles and occupy areas measured in terms of acres.[48,49]

The use of arrays in radars is extensive at the experimental level and some are operational. An experimental active-element array with several hundred active elements (ESAR) has been in operation for several years and others are presently being constructed. Passive-element arrays are operational in the Navy.[50] In the meantime, new developments in high-power electronic phasing devices have raised the performance capability of passive-element arrays with their inherent economies to the point where it may be feasible to consider arrays for radar tasks for which they would have previously been presumed uneconomical.

Medium-size arrays are finding considerable application in communications, particularly as satellite antennas and in direction-finding equipment.

As the use of arrays has spread, more time and effort have been devoted to the understanding of the principles of such devices, and recent years have witnessed significant contributions in such areas as density tapering, multiple-beam arrays, and components for arrays. There remain areas not well understood, however. For example, we still do not understand the generalities of mutual coupling sufficiently to answer questions about optimum types of elements, if any, or even questions about the optimum use of most types of elements.

There also remains a great need for invention and innovation in components for arrays of large numbers of

elements, where the existence of great numbers of duplicate components justifies, and in fact makes mandatory, careful and thorough component engineering. The present state of the art in this area is about as adequate for realizations of the large-array designers' dreams as the electromechanical relay would be as the sole logic element in modern digital computers.

The author acknowledges the contributions of the other members of the Array Radars Group of Lincoln Laboratory.

Lincoln Laboratory is operated with support from the U.S. Air Force.

REFERENCES

1. Allen, J. L., "The Theory of Array Antennas (with Emphasis on Radar Applications)," Tech. Rept. 323(U), Lincoln Laboratory, M.I.T., Lexington, Mass., July 25, 1963.

2. Kraus, J. D., *Antennas*. New York: McGraw-Hill Book Co., Inc., 1950.

3. Von Aulock, W. H., "Properties of Phased Arrays," *Proc. IRE*, vol. 48, Oct. 1960, pp. 1715–1727.

4. Skolnik, M. I., *Introduction to Radar Systems*. New York: McGraw-Hill Book Co., Inc., 1962.

5. Ogg, F., Jr., "Steerable Array Radars," *IRE Trans. on Military Electronics*, vol. MIL-5, no. 2, Apr. 1961, pp. 80–94.

6. Unz, H., "Linear Array with Arbitrarily Distributed Elements," Rept. No. 56, Electronics Research Laboratory, University of California, Berkeley, Nov. 2, 1956.

7. King, D. D., *et al.*, "Unequally Spaced Broadband Antenna Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-8, July 1960, pp. 380–384.

8. Allen, J. L., *et al.*, "Phased Array Radar Studies, 1 July 1960 to 1 July 1961," Tech. Rept. 236(U), Lincoln Laboratory, M.I.T., Nov. 13, 1961.

9. Lo, Y. T., "A Probabilistic Approach to the Design of Large Antenna Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-11, no. 1, Jan. 1963, pp. 95–97.

10. Wolff, I., "Determination of the Radiating System Which Will Produce a Specified Directional Characteristic," *Proc. IRE*, vol. 25, May 1937, pp. 630–643.

11. Woodward, P. M., "A Method of Calculating the Field Over a Plane Aperture Required to Produce a Given Polar Diagram," *J. Inst. Elec. Engrs. (London)*, vol. 93, pt. IIIA, 1947, p. 1554.

12. Dunbar, A. S., "On the Theory of Beam Shaping," *J. Appl. Phys.*, vol. 23, 1952, pp. 847–853.

13. Shanks, H. E., "A Geometrical Optics Method of Pattern Synthesis for Linear Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-8, Sept. 1960, pp. 485–490.

14. Ksienski, A., "Maximally Flat and Quasi-Smooth Sector Beams," *Ibid.*, pp. 476–484.

15. Silver, S., *Microwave Antenna Theory and Design*, M.I.T. Radiation Laboratory Series. New York: McGraw-Hill Book Co., Inc., 1949.

16. Taylor, T. T., "Design of Line-Source Antennas for Narrow Beamwidth and Low Side Lobes," *IRE Trans. on Antennas and Propagation*, vol. AP-3, Jan. 1955, pp. 16–28.

17. Taylor, T. T., "One Parameter Family of Line Sources Producing Modified sin $\pi u/\pi u$ Patterns," Tech. Memo. 324, Microwave Laboratory, Hughes Aircraft Co., Culver City, Calif., Sept. 4, 1953.

18. Dolph, C. L., "A Current Distribution for Broadside Arrays Which Optimizes the Relationship Between Beam Width and Sidelobe Level," *Proc. IRE*, vol. 34, June 1946, pp. 335–348.

19. Taylor, T. T., "Design of Circular Apertures for Narrow Beamwidth and Low Sidelobes," *IRE Trans. on Antennas and Propagation*, vol. AP-8, Jan. 1960, pp. 17–23.

20. Willey, R. E., "Space Tapering of Linear and Planar Arrays," *Ibid.*, vol. AP-10, July 1962, pp. 369–377.

21. Bickmore, R. W., "A Note on the Effective Aperture of Electronically Scanned Arrays," *Ibid.*, vol. AP-6, Apr. 1958, pp. 194–196.

22. King, R. W. P., *The Theory of Linear Antennas*. Cambridge, Mass.: Harvard University Press, 1949.

23. Rabinowitz, S. J., "The Conductance of a Slot in an Array Antenna," Tech. Rept. 192, Lincoln Laboratory, M.I.T., Dec. 31, 1958.

24. Allen, J. L., and Delaney, W. P., "On the Effect of Mutual Coupling on Unequally Spaced Dipole Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-10, Nov. 1962, pp. 784–785.

25. Tang, R., Unpublished work, Hughes Aircraft Co., Ground Systems Div., Fullerton, Calif.

26. Allen, J. L., "Gain and Impedance Variation in Scanned Dipole Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-10, Sept. 1962, pp. 566–572.

27. Allen, J. L., *et al.*, "Phased Array Radar Studies, 1 July 1961 to 1 December 1962," Tech. Rept. 299(U), Lincoln Laboratory, M.I.T., Feb. 20, 1963.

28. Parad, L. I., and Kreutel, R. W., "Mutual Effects Between Circularly Polarized Elements," *Abstr. 12th Ann. Symp. on U.S.A.F Antenna Research and Develop. Program*, Antenna Arrays Section, University of Illinois, Urbana, Oct. 1962.

29. Allen, J. L., "On Array Element Impedance Variation with Spacing," *IEEE Trans. on Antennas and Propagation*, vol. AP-12, May 1964, pp. 371–372.

30. Bates, R. H. T., "Mode Theory Approach to Arrays," submitted for publication in *IEEE Trans. on Antennas and Propagation*, vol. AP-13, Mar. 1965.

31. Hannan, P. W., "The Element-Gain Paradox for a Phased Array Antenna," *IEEE Trans. on Antennas and Propagation*, vol. AP-12, July 1964, pp. 423–424.

32. Hannan, P. W., Lerner, D. S., and Knittel, G. H., "Wide Angle Impedance Matching Calculated for a Phased Array Antenna," *Digest 1963 IEEE Internat'l Symp. on Antennas and Propagation*, pp. 228–233.

33. Ruze, J., "Physical Limitations on Antennas," Tech. Rept. 248, Research Laboratory of Electronics, M.I.T., Cambridge, Mass., Oct. 30, 1952; also "The Effect of Aperture Errors on the Antenna Radiation Pattern," *Nuovo Cimento (Suppl.)*, vol. 9, no. 3, 1952, pp. 364–380.

34. Elliott, R. S., "Mechanical and Electrical Tolerances for Two-Dimensional Scannning Antenna Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-6, Jan. 1958, pp. 114–120.

35. Kurtz, L. A., and Elliott, R. S., "Systematic Errors Caused by the Scanning of Antenna Arrays: Phase Shifters in the Branch Lines," Tech. Memo. 359, Hughes Aircraft Co., Culver City, Calif., May 1, 1963.

36. Allen, J. L., *et al.*, "Phased Array Radar Studies, 1 July 1959 to 1 July 1960," Tech. Rept. 228(U), Lincoln Laboratory, M.I.T., Aug. 12, 1960, pp. 162–170.

37. Miller, C. J., "Minimizing the Effects of Phase Quantization Errors in an Electronically Scanned Array," *Proc. Symp. on Electronically Scanned Array Techniques and Applications*, RADC-TDR-64-225, vol. 1, Rome Air Development Center, Griffiss Air Force Base, N.Y., July 1964.

38. Blass, J., "The Multidirectional Antenna: A New Approach to Stacked Beams," *1960 IRE Internat'l Conv. Record*, vol. 8, pt. 1, pp. 48–50.

39. Bernella, D. M., and Pratt, H. J., Jr., in "Phased Array Radar Studies, 1 Jan. 1963 to 1 July 1964," Tech. Rept., Lincoln Laboratory, M.I.T., to be published.

40. Butler, J., and Lowe, R., "Beam Forming Matrix Simplifies Design of Electronically Scanned Antennas," *Electron. Design*, vol. 9, Apr. 12, 1961, pp. 170–173.

41. Shelton, J. P., and Kelleher, K. S., "Multiple Beams from Linear Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-9, Mar. 1961, pp. 154–161.

42. Allen, J. L., in "Phased Array Radar Studies, 1 Jan. 1963 to 1 July 1964," Tech. Rept., Lincoln Laboratory, M.I.T., to be published.

43. Allen, J. L., "A Theoretical Limitation on the Formation of Lossless Multiple Beams in Linear Arrays," *IRE Trans. on Antennas and Propagation*, vol. AP-9, July 1961, pp. 350–352.

44. Stein, S., "On Cross Coupling in Multiple-Beam Antennas," *Ibid.*, vol. AP-10, Sept. 1962, pp. 548–557.

45. White, W. D., "Pattern Limitations in Multiple-Beam Antennas," *Ibid.*, July 1962, pp. 430–436.

46. Sharp, E. D., and Diab, M. A., "Van Atta Reflector Array," *Ibid.*, vol. AP-8, July 1960, pp. 436–438.

47. *IEEE Trans. on Antennas and Propagation*, vol. AP-12, Mar. 1964 (Special Issue on Active and Adaptive Antennas).

48. Devane, M. E., and Dion, A. R., "The El Campo Solar Radar Antenna," Tech. Rept. 276, Lincoln Laboratory, M.I.T., Aug. 17, 1962.

49. Ochs, G. R., "The Large 50 Mc/s Dipole Array at Jicamarca Radio Observatory," *Digest 1963 IEEE Internat'l Symp. on Antennas and Propagation*, pp. 237–241.

50. Mroz, E., "Some Aspects of Integrated Technology," *Microwave J.*, vol. 7, July 1964.

PERT for the engineer

Engineers may associate PERT with large military programs and computers, but they should know that PERT can also be applied effectively to modest engineering tasks

Jordan Kadet, Bruce H. Frank *Sylvania Electric Products Inc.*

Although PERT (Program Evaluation and Review Technique) was successfully introduced in 1958 on the Polaris Weapons System Program and subsequently it and similar networking techniques were enthusiastically accepted in the management of the development phase of weapons systems programs, as well as within industries such as construction, we now observe a curious phenomenon. Many engineers, employed as project engineers either by large corporations on major programs or by small laboratories engaged in component or subsystem development, still question the effectiveness of PERT for their own use. Much of the cynicism that greets the technique can be traced, in part, to a lack of conviction on the part of engineers that PERT can be an effective planning and control technique for modest task applications as well as for large programs necessitating elaborate, computer-based applications.

When should PERT be used?

This article is directed at the individual engineer or project engineer engaged in an engineering task or small project, whether oriented toward defense or commercial products. The size of the task does not affect the requirement for planning nor need it influence the decision to use PERT techniques. PERT techniques for planning and control, involving graphic methods and network analysis to depict and analyze a project, can be applied to large or small projects, formally or informally. The nature of the project does, however, influence this decision. The networking technique is applicable primarily to the "once-through" type of effort typically associated with the development of a system or subsystem, that is, one where

- The objective of the effort is to realize an engineering model or prototype model of a hitherto nonexistent item or items of equipment.
- The achievement of the objective involves some degree of uncertainty because of the lack of directly applicable experience on which to base estimates.
- The achievement of the objective is subject to significant changes in direction, intensity, and scope as the work progresses.

Efforts of a repetitive nature and predictable course, such as volume production runs, are normally better planned and managed through the use of other techniques (for example, Line-of-Balance). There is, however, no clear cutoff point between the once-through process and the repetitive process when development programs include the production of initial or limited quantities of end items.

Basic research programs and endeavors aimed at breakthroughs in the state-of-the-art are less amenable to PERT even though they, too, are normally once-through processes. *PERT becomes less meaningful when objectives cannot be clearly defined.*

Within these limits, networking may be successfully applied. This does not imply that PERT and other networking techniques such as CPM (Critical Path Method) apply only to developmental engineering tasks. The applications are as varied as industry itself. These techniques have been applied and found useful in the following areas:

Government research and development
Commercial product development
Installation of equipment
Construction
First production runs
Maintenance
Systems and procedures installation
Training programs
Movie and stage productions

Military electronics, at one time the predominant user of PERT techniques, can no longer make this claim. Booz, Allen and Hamilton, Management Consultants, reported that, "By 1963, 85 per cent of the companies surveyed were using PERT either exclusively or partially for private commercial work."[1] Although diverse applications have bred many variations of the techniques, the basic concepts have not changed.

What PERT is and is not

Before PERT methods of planning and control are discussed, several myths shrouding these techniques must be dispelled. Conventional, nonprobabilistic, bar chart planning led to the popular belief, reinforced by many engineers, that research and development work cannot aptly be measured in terms of time and cost progress. "You can't schedule invention," "You can't expect us to design by the calendar," "Mental processes can't be planned" are frequent statements heard. PERT, with a probabilistic approach to time planning, has helped to dispel this myth.

Whenever a new technique such as PERT is developed, it tends to introduce a new breed of specialist, and a new language is promulgated which is understood, presumably, only by the "privileged few." When discussing their work, PERT specialists ("PERTniks") use terms such as slack, critical path, expected time, latest allowable time, and a host of others. Because PERT specialists in staff and line positions are needed for large-scale PERT applications, some people are led to believe that PERT is a mystical system, complicated to use and one which will not prove durable. In practice, however, PERT is surviving, and the techniques do not necessarily require specialist personnel for intelligent application. It has been suggested by some writers[2] that a PERT analysis staff be eliminated and that project engineers perform this function. The suggestion has merit and certainly is correct in its implication that the engineer or project engineer plays a key role in a successful PERT system.

The reason that PERT applications continue to expand is that the technique provides many advantages. Among these are

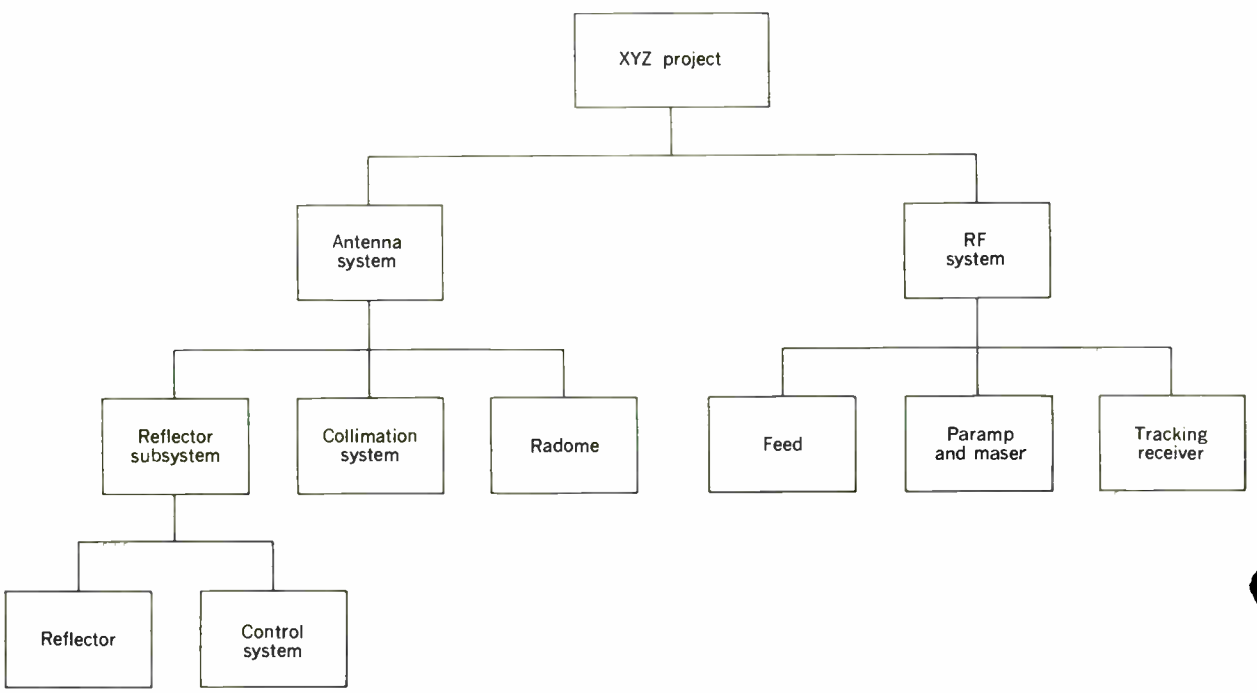
- A disciplined approach to planning.
- A method of visualizing the work and of communicating plans.
- A plan which can reflect uncertainties but which can also be easily used for calculating the time to perform the project.
- A means for appraising progress and forecasting problem areas.

The subsequent sections define the steps that should be followed in planning a project, large or small, and the networking techniques that should be applied. Many of these techniques are not unique to networking but have been in use for a long time. This is why we refer to PERT as being both evolutionary and revolutionary.

Before delving into the how and why of PERT, let us define what it is. PERT is a planning and control discipline which employs a specific set of principles, methods, and techniques for effective planning. The key elements of this discipline are

- A *work breakdown structure*, which begins with the objectives and subdivides them into successively smaller elements of work.
- A *network*, comprising all the work which must be accomplished to reach the objectives, and depicting the planned sequence of accomplishment of this work as well as the interdependencies and interrelationships.
- Elapsed time *estimates* of work to be performed and *schedules* which also consider the availability of resources.

Fig. 1. Typical engineering equipment breakdown or "family tree."



■ *Analysis* of the interrelated networks and schedules as a basis for continued evaluation of performance and identification of problem areas.

PERT is also a discipline for organizing data, documenting the plan, and manipulating the plan to effect a successful conclusion to the project. The statement that PERT is a discipline and a tool for planning and controlling a job is significant to an understanding of the technique. Although the technique is formalized it is no different from what any of us must ordinarily do to plan and control our work properly. As will be seen, elements of this technique had their origin in engineering applications. It must be remembered that PERT is not a panacea or substitute for the decision-making process; it is only a tool to aid the decision makers.

Defining the job

Before any project or task, large or small, can be undertaken it must be defined. The engineer must define the objectives in terms such as final hardware. The hardware, perhaps, can be broken down into various levels such as assemblies, subassemblies, and components. This exercise is generally expressed on paper by the preparation of a "family tree" or "Christmas tree" equipment breakdown. This breakdown, dividing the piece or pieces successively into component parts (see Fig. 1), is the "road map" for the designer or project engineer. In any project there may be other end-item requirements such as engineering data and ancillary items. It is possible and practical to expand the family tree to include these items. The expanded family tree that evolves is known in the language of PERT as a *work breakdown structure*. The work breakdown structure does not stop at the level of end-item hardware, data, and services. As can be seen in Fig. 2, the breakdown continues to a definition of the required tasks such as design, fabrication, and test of each piece of hardware.

The resultant work tasks or *work packages* are likely to be assigned as separate responsibilities to organizations or individuals and are thus defined separately. As the family tree provided a road map for designing, so the work breakdown structure provides a road map for planning. The work breakdown structure establishes the basis for

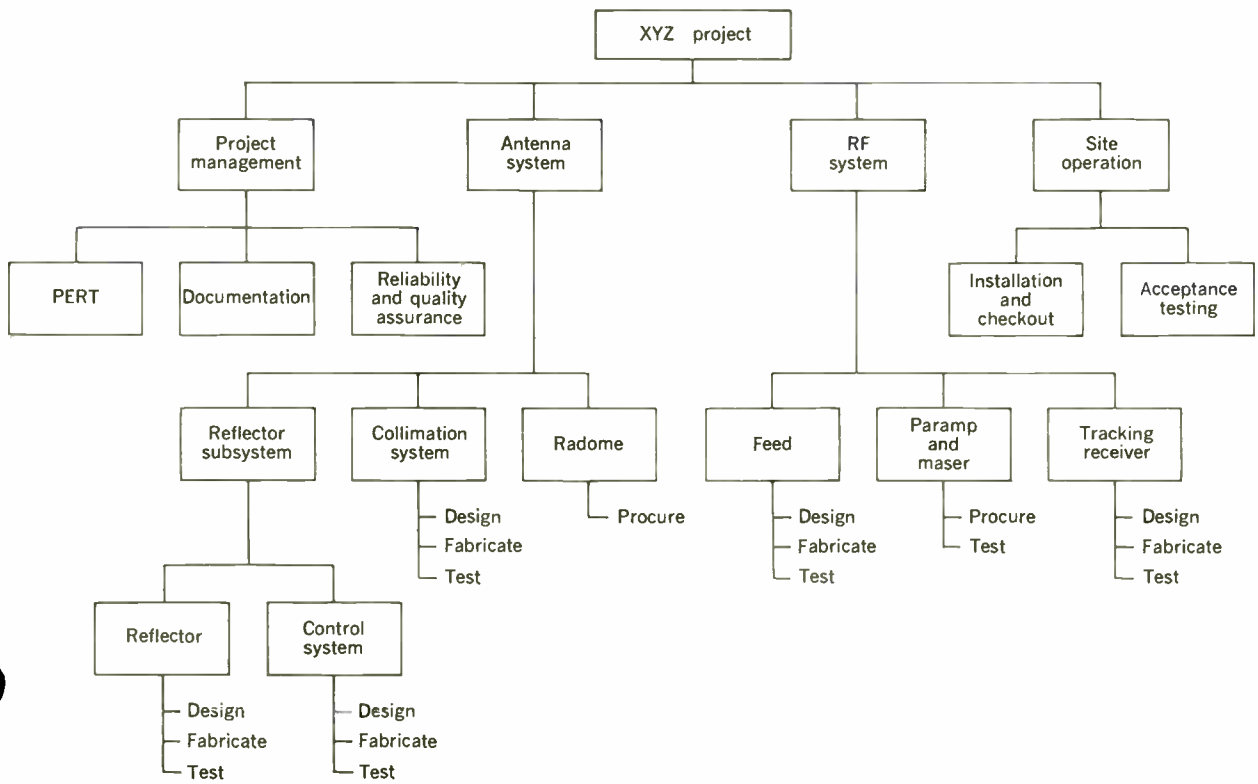
- Defining and relating the work to be performed and the program objectives.
- Identifying the responsibility for accomplishment of the work packages and monitoring at higher levels.
- Detailed PERT planning and control.
- Cost planning and control.

In a project, the work packages must be defined in detail. This is usually done either by simple written descriptions on a small project, or on a larger, complex program by more formal task authorizations. These should define the specifications, performance parameters, and the budget and schedule for the work packages. While these define the work in terms the engineer can understand, the plan for performance of this work remains to be prepared.

Network planning

Network planning or PERT was developed primarily to fill the need for a tool that would allow the engineer to depict what really was going to happen in his relatively

Fig. 2. In PERT language, this is a work breakdown structure.



complex R&D task or project. The Gantt or milestone charts which were in use for planning had been adapted from production planning methods and were found wanting. Networking techniques were derived from those used in electrical engineering and from flow charting used in computer programming. Networking is a logical discipline for planning a task. It is not the *network* which is prepared; it is the *plan* which is prepared using *network discipline*. This distinction is basic for an appreciation of networking as a planning and control tool for the engineer.

The discipline of planning can be defined by a series of questions which must be answered:

- What work must be performed?
- What is needed to perform this work?
- How will completion be identified?

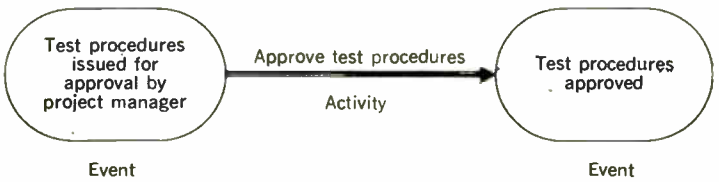
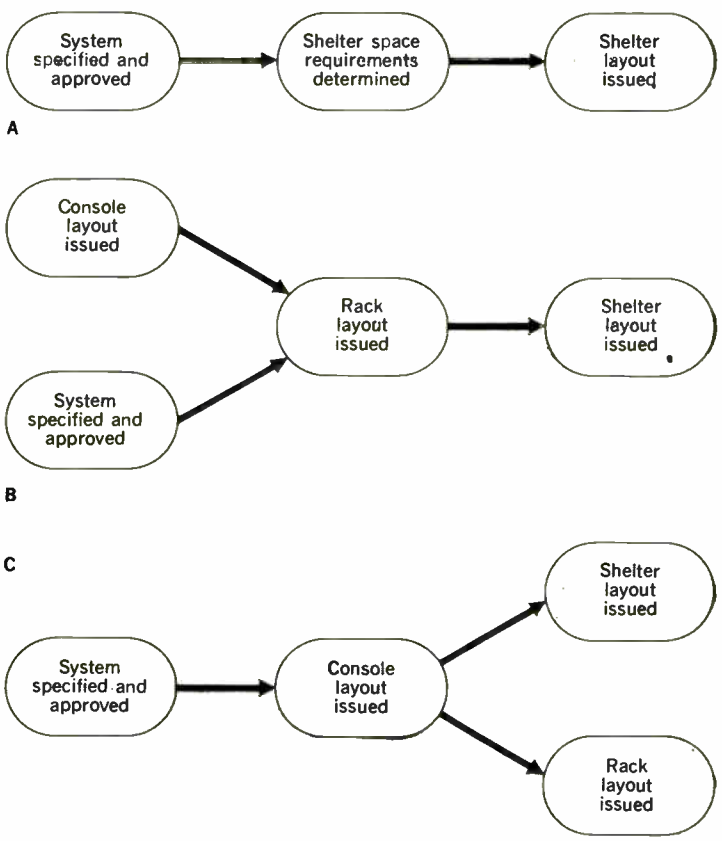


Fig. 3. Basic elements of a PERT network are events and activities.

Fig. 4. Network relationships: serial activities (A), an activity dependent upon parallel activities (B), and parallel activities dependent upon one restraining activity (C).



These questions can only be answered accurately by the responsible engineer or project engineer. When he thinks of his task in terms of these questions, he can plan in detail using a network approach. Once he has delineated the work and its sequence in detail, he may consider another very important question:

- How long will it take to perform the work?

The basic advantage of planning by this discipline is that it requires the engineer to think of his work in its entirety and to consider all of the individual parts.

Simply, a PERT network or chart is made up of two elements: events and activities. Activities are the work to be performed and are signified by arrows. Events are specific definable achievements—either the beginning or completion of one or more activities—and are represented on the network by circles, ellipses, rectangles, or other geometric figures. Every activity is bounded by an event at the beginning and an event at the end (see Fig. 3).

The method of preparing a network simply involves connecting activities and events to show their relationship. If one activity cannot be performed until another has been completed, it should be drawn as shown in Fig. 4(A). If, on the other hand, the activity cannot start until more than one activity is complete, it should be drawn as in Fig. 4(B).

Another type of network interrelationship exists in which more than one activity may start when a single previous activity has been completed. This type of constraint and parallel start is shown in Fig. 4(C). It is important to remember that the network represents logical restraints, *not* time sequencing. The network shows which activities *must* be accomplished before another may start. It should not be drawn to show all activities that may, because of preconceived ideas, be occurring prior to another in time only. If the activity can be performed regardless of the status of other activities, no constraint should be shown. Figure 5 shows an example of an abbreviated network.

Once the above principles of networking are understood, many other questions regarding rules of preparation are raised. One of the first is: How do we prepare it—do we start at the beginning and work toward the end or do we start at the end and work toward the beginning—what is the "standard" method? Although some articles in the past have extolled the virtues of one or the other method, we have found that *there is no "standard" method*. One can start at the front, back, or even in the middle, in preparing a network; it depends on the individual or group planning the job. If, as is frequently possible, a more effective plan can be developed by working from the end towards the beginning, then the network should be drawn accordingly. If you prefer, start at the beginning; once the network is prepared, however, it must represent the complete plan for accomplishing the job.

Another frequently asked question is: How large is a good network? A quick answer might be: Large enough to be a complete plan of the work to be performed. This answer only provokes another request for some guidelines for measurement. We have seen networks with thousands of activities (on a major weapons systems program) and with as few as 20 to 25 (on a small task). These were both good for their uses. Conversely, very small networks have been developed for very large projects and some

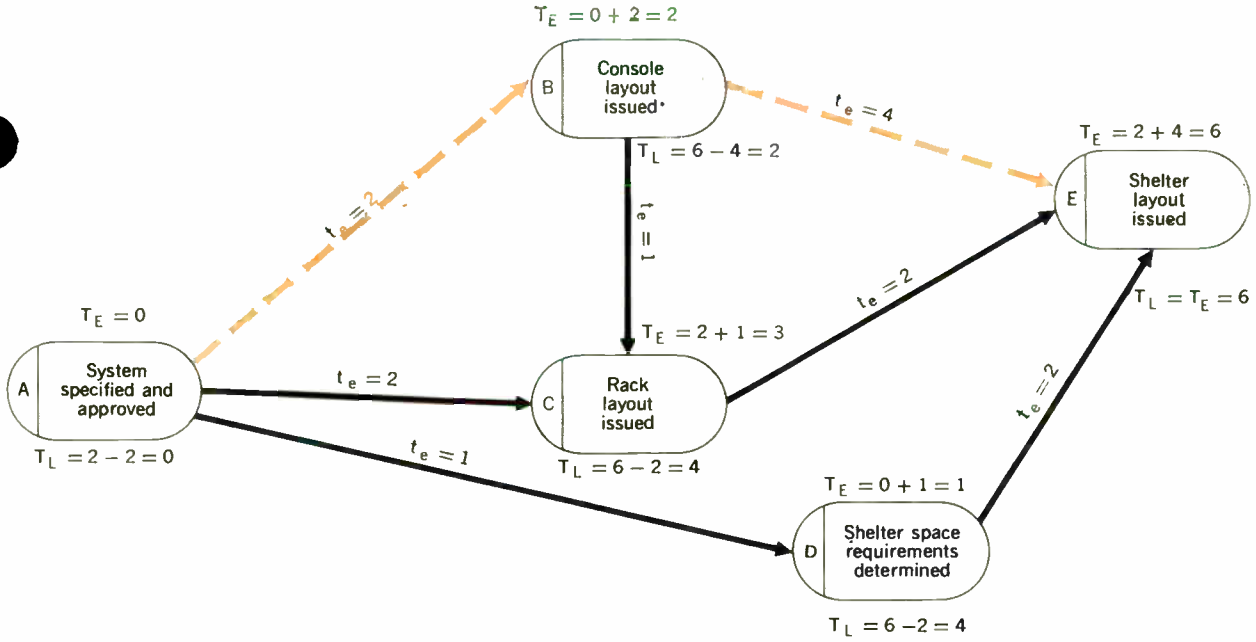


Fig. 5. Example of a PERT network structure, where t_e is statistically expected time of activity, T_E is earliest expected date, T_L is latest allowable date, and dashed line is critical path.

very large networks developed for small projects, none of which constituted a good plan. The basic problem with these networks was that preconceived artificial limits were set on network size. Such limits—"no fewer than 150 activities," "no greater than 100 activities," "on a sheet of paper no larger than one foot by two feet"—are usually artificial and should be avoided whenever possible. The first rule is to prepare the network without regard for size or neatness and then, when it is complete, worry about the size limitations that may be imposed. As a rule of thumb, with full regard to the differences in projects in their scope, technical sophistication, and duration, we have found that there should be an average elapsed time for the majority of the activities approximately equal to the period between progress reviews. As an example, if the network is reviewed weekly (as might be the case in a small engineering job), the activities should average 1–1½ weeks. If the network is reviewed monthly, the average should be 3–4 weeks. Although this is not a hard and fast rule, we have found it to hold true in most cases. This may seem to contradict our comment that the network should be prepared first and the duration estimated second, but network size is not really checked until time estimating is completed. This procedure is not illogical for it has been found that many networks undergo major modifications and improvements as time estimates are made.

Definitions

Optimistic time (a) = The time the activity will take if everything to be done is done successfully the first time.

Most likely time (m) = The time required to accomplish an activity under normal circumstances with some success and some failure.

Pessimistic time (b) = The time the activity will take with extremely bad luck.

Time estimating

Once the network has been developed, portraying the relationships and interdependencies among the activities, the time needed for each activity should be estimated. This should be done by the engineers who are going to do the work.

There are two basic methods of time estimating, probabilistic and deterministic, and the use of one or the other is optional.

Probabilistic time estimating, which was developed as part of the PERT system, requires three estimates for the duration of an activity (see Definitions). These three estimates are plotted on a beta distribution to find statistically expected time $t_e = (a + 4m + b)\ 6$, where t_e is expected time, a is optimistic time, m is most likely time and b is pessimistic time. (For a detailed discussion of these formulas, see Ref. 3.) This expected time, the weighted mean, is then used in all PERT calculations; it represents a probability of 50 per cent. A distinct advantage of this system is that the engineer is able to express his uncertainty in terms of these three estimates rather than having to express a more certain single time estimate. The three estimates also provide a means of assessing the risk being taken in performing the job. The three-time-estimate system is specifically designed for uncertain, nonrepetitive types of work.

Deterministic, or single time estimating, on the other hand, does not utilize probability weighting. The engineer can make one estimate for the length of the activity which is the time he needs to accomplish it. The single time estimate has been used mainly in process industries and construction where it has been found valuable. CPM, which is a networking system employing a slight vari-

ation of PERT in terms and method of network preparation, also uses one time estimate.

Once the time for the activities has been estimated and in the case of the three-time-estimate system, the expected time for the activity calculated, the expected length of the project can be calculated. Also, the expected dates for each event can be determined. The expected length of the project (the sum of the t_e's on the longest network path) is referred to as the earliest expected date (T_E). It is calculated by summing the expected times of the activities from the start through the completion of each network path (see Fig. 5). Once the T_E is calculated, the T_L, or latest allowable date, is calculated by starting with a predetermined date for the end event and subtracting the expected elapsed times, moving "backward" through the various network paths. The predetermined date for the end event might be the date set by management or the date imposed by a customer. The T_L calculated for the first event on the network indicates the latest date the task(s)/project(s) can be started without causing the end event to slip beyond the predetermined target date (see Fig. 5).

After both the earliest expected date (T_E) and the latest allowable date (T_L) have been computed for each of the events, slack may be determined for each path in the network. Slack is the time difference between the earliest expected date and latest allowable date: slack $= T_L - T_E$.

The amount of time the expected date can slip before it equals the latest allowable date (T_L) can also be used as the definition of slack. Slack can be positive, negative, or zero. When the latest allowable date (T_L) is later than the earliest expected date (T_E), positive slack exists. Positive slack is "time-to-spare."

The longest time path or sequence of activities through a network is called the critical path. This path controls the completion date for the task(s)/project(s) represented by the network, since all other paths are shorter. Should the length of time for the critical path be the same as the established completion date for the network, there will be no time-to-spare. This time-to-spare relative to the established completion date is "slack time." In Fig. 5, the slack time for the critical path would be zero. (Note that until a schedule date or directed date is applied to the network end event, the slack value of the critical path is *always necessarily zero*.) Sometimes the length of time for the critical path exceeds the established date for completion. In this case the critical path is said to have negative slack. There is *no* time-to-spare; in fact, a projected time slippage condition beyond the established completion date exists. In like manner, there can be "positive slack" where the duration of the critical path is less than the scheduled time-to-perform from the established start-to-completion dates, indicating that the network activities will be completed with time-to-spare.

Each series path in the network will have a slack value in relation to the established completion date for the project. In this respect each and every series path has a measurable "degree of criticality" which can be positive, zero, or negative. By definition, the longest time path is the most critical in relation to the established completion date and hence is called the critical path. All other paths, which are shorter than the critical path, are therefore called slack paths. Two or more paths may have the same duration and so will have the same slack value. Should these equal paths be the longest in the network, there will be two or more critical paths for the same network.

It should be emphasized that the slack time for a series path pertains to *the entire path* and not to any one event or activity in particular. Any change in the activity time for any one activity in the path series will change the slack value for the entire path.

In Fig. 5, the value of slack is zero for the path through events A, B, and E. Since this is the longest time path through the network, it is the critical path. Other paths through this network have positive slack up to three weeks (A, D, E).

Scheduling

Once the network has been prepared and the time estimates made, the calculations of expected time, latest time, slack, and the critical path reveal whether or not the plans are acceptable. Usually, the plans must be reconsidered owing to the length of time they take and the considerations of the resources (people, equipment, and materials) required. The plan must be converted to a schedule. Scheduling is the conversion of the plan into a set of specific dates that govern the start and completion of work and involves the allocation of resources required to achieve the planned objectives.

Many times, when the initial plan is completed, negative slack will exist. This occurs because, in conforming with the basic rule of first laying out the network without regard for time, the engineer planned the job the way he would like to do it. The scheduling job now involves balancing the work and resources to achieve a schedule which can be met. By first examining the paths with negative slack and by reallocating resources and adjusting activities, the planner can draw up a work plan that is realistically phased with scheduled requirements. The activities with positive slack should be considered second since they provide a latitude for scheduling resources. Once established, the schedule should not be considered changeable at will. It should only be changed when objectives are changed, which essentially means that the plans to achieve these objectives require modification.

Day-to-day control

As the work progresses, there are changes that affect the PERT network: slippages occur, key personnel become unavailable, unexpected breakthroughs are made, activities are accomplished faster than anticipated. This information must be accumulated for analysis, to determine the effect on all interrelated parts of the project, so that the plan can be updated as the schedule demands.

The network can also be used during this phase (or at any time) to assess the effect of contemplated changes on the schedule. This process, called simulation, can provide a quick determination of the impact of a proposed change in the plan.

In engineering, as well as other areas, the network can (by being hung on the wall and information posted on it) serve as a visual reminder of the work to be done, who is to do it, and when it is to be done. As activities are completed, the completion dates should be recorded on the network and the activity checked off. As more information becomes available on the expected time needed to accomplish an activity, the new estimate should be written on the network. This day-to-day use of the network provides the latest information available. On a routine basis, new T_E and T_L calculations can be made and the effect on

schedule measured. When slippages affect the schedule, an analysis must first be made into the cause and possible solution of the problem. Only then should changes be made in the plans and the schedule adjusted to assure successful completion. If the network is not updated and continuously used as a working tool, PERT becomes merely an extra exercise. The value of the technique lies in its use as a dynamic reflection of the work which must be done. As the project progresses, the network serves another valuable purpose—it can be used as a communications tool. It is a basic reference document which can be used graphically in discussing the project with other engineers and management.

One of its limitations, however, is its use as a visual aid in top management conferences or summary level presentations to the customer, where detailed network information is not required. We have found, more often than not, that the network must be translated onto a time base for presentation (by using a Gantt chart, for example).

Task plan vs. project plan

This article has discussed PERT in terms of the engineer's role, whether he is working on a single task or as a member of a project team. If he is on a project team, the network he prepares is only part of the total project network plan. His network cannot be prepared in a vacuum; it must be prepared with the engineers responsible for other parts of the project. The network system depends upon planning of all interrelationships; interfaces between one engineer and another must be shown. We have found that many times two or more engineers have different conceptions of what they are going to do for each other. The network points out such discrepancies because the plans cannot mesh. Whether or not the engineer prepares the network for part of a larger project or for a task for which he is solely responsible, the basic rules of network preparation are the same.

PERT/COST estimating

Many engineers, as they acquire experience and understanding of the PERT (Time) techniques, ask "Why not include cost estimating?" In June 1962, the government (DOD and NASA) introduced PERT/COST. The principles are relatively simple although the terminology is sometimes confusing. The basic principle of PERT/COST is that the network can be used for both time and cost planning. The work breakdown structure, as previously described, serves as the coupling device with cost estimating done on a "work package" level. Thus the engineer would estimate his costs by time-phased work package.

There are also proponents of other methods, including those who prefer to estimate for each activity. On large programs, involving many labor skills and people, this has been found too cumbersome. In smaller applications it may be possible. In process industries and in the construction industry, where the application of manpower has a linear effect on time, they have been using a cost extension of CPM, where one objective is to optimize cost and time to determine the most applicable schedule and cost for the project. PERT/COST, on the other hand, is simply a discipline for estimating costs and collecting actuals versus the estimate. (To delve into the details of these systems in this limited article is not practical. The interested reader should refer to the publications listed in the bibliography.)

Use of computers

In attempting to convey a basic understanding of the concepts of PERT as it relates to and benefits the engineer, the application of computers in processing PERT has not been discussed. When PERT is applied to large projects, the use of a computer becomes essential in order that the information on a multitude of activities can be processed in a reasonable amount of time. Most computer manufacturers have PERT programs available and many companies and government agencies have prepared their own. In addition, there are programs available for producing networks using plotting or drafting machines tied in with the computer.

It must be remembered that the use of the computer does not vary the technique; the computer is simply the mechanism for rapid processing of a large volume of data. Possibly the emphasis placed on computer usage with PERT has tended to overshadow the advantages to be gained by applying the PERT technique manually and has discouraged its use by engineers in the thought process applied to planning and controlling modest tasks.

Conclusions

In conclusion, it must be restated that PERT is neither a panacea nor solely an automated means for planning. PERT planning must be done competently, with an understanding of both the tool and the project. It can only be done effectively by the engineer or project engineer, however odious planning may seem to him.

When used competently, PERT yields many advantages. It offers a plan which is well developed and is easy to communicate, showing clearly its objectives and interrelationships. It is an improved method of time and cost measurement of tasks, aiding in improved assessment of accomplishment. It is a predictive tool to highlight potential problem areas. It provides the capability of examining alternative plans by simulating changes.

It must be emphasized that proper employment of PERT as an effective working tool necessarily requires a full understanding of the technique and its assimilation by the engineer as part of his normal discipline in the planning and control of assigned tasks or projects.

REFERENCES

1. *New Uses and Management Implications of PERT.* Chicago, Ill.: Booz, Allen and Hamilton, Inc., 1964.
2. Boverie, R. T., "The Practicalities of PERT," *IEEE Trans. on Engineering Management*, vol. EM-10, no. 1, Mar. 1963.
3. "PERT," Summary Phase Report, Special Projects Office, Bureau of Naval Weapons, Department of the Navy, Washington, D.C., July 1958.

BIBLIOGRAPHY

Note: This list is abbreviated. The interested reader should refer to the third item for an excellent compendium of references on the subject.

"USAF PERT," vol. I-V, Air Force Systems Command, Andrews AFB, Washington, D. C., 1963–1964.

Miller, R. W., *Schedule, Cost and Profit Control with PERT.* New York: McGraw-Hill Book Co., Inc., 1963.

Dooley, A. R., "Interpretations of PERT," *Harvard Bus. Rev.*, Mar.–Apr. 1964.

Martino, R., *Project Management and Control.* New York: American Management Association.

"PERT Guide for Management Use," PERT Coordinating Group, Washington, D. C., June 1963.

"DOD/NASA Guide—PERT/COST," U.S. Government Printing Office, June 1962.

"General Information Manual—PERT," International Business Machines Corp., White Plains, N. Y.

Authors

Lester W. Strock received the B.S. degree in 1927 from the Philadelphia College of Pharmacy and Science, and the M.S. degree in 1929 and the Ph.D. degree in 1939 from the University of Pennsylvania. From 1931 to 1932 he was a Fellow in biophysics at the Rockefeller Institute for Medical Research in New York City. He did postgraduate research work in various European research centers from 1932 to 1936. For the next two years he was employed by the Norwegian Department of Commerce to develop a spectrochemical laboratory and to train personnel at the University of Oslo's Geological Museum. From 1939 to 1952 he served as geochemist for the State of New York's Conservation Department. During this time he served as a consultant to companies producing beryllium for the Manhattan Project. He joined Sylvania in 1952 as an engineering specialist at the company's Bayside, N.Y., research laboratories. He was transferred to the Lighting Division's engineering laboratory in Salem, Mass., in 1958. His work there has been chiefly in the field of crystal defects, particularly as applied to electroluminescence. In addition, he currently is conducting research on tungsten and the effects of crystal imperfections on its performance in lamps.

Irving Greenberg (M) received the B.S.E.E. degree in 1950 from Syracuse University. After graduation he joined Sylvania Electric Products in Seneca Falls, N.Y., as a project engineer on television picture tubes. Three years later he became an applications and field engineer, specializing in hydrogen thyratron development. In 1955 he was transferred to Sylvania's Woburn, Mass., facility, where as applications engineer he worked on counter tubes, trigger tubes, and pencil tubes. He later served as a product specialist in the field of semiconductors. After a two-year period as a semiconductor sales engineer for Sylvania at Camillus, N.Y., he returned to Woburn as a senior design engineer, where he was involved in the design of diode and transistor test equipment. From 1960 until recently he was product manager for electroluminescent display devices, at Seneca Falls. In this position, he was in charge of field engineering on these devices and of liaison between customers, sales engineers, and the manufacturing operation. He is now project manager, aerospace/microelectronics, for Sylvania at Waltham, Mass.

Mr. Greenberg is the author of two articles in technical publications, and has lectured to IEEE Sections on the subject of electroluminescent display devices. He is a member of the Society for Information Display.

R. A. Frosch (SM) was born in New York, N.Y., on May 28, 1928. He received the A.B. degree at Columbia University in 1947. He also pursued his graduate studies at Columbia, from which he received the M.A. degree in 1949 and the Ph.D. degree (in theoretical physics) in 1952.

Dr. Frosch is currently director for nuclear test detection at the Advanced Research Projects Agency, Washington, D.C. Prior to September 1963 he served as director of Columbia University's Hudson Laboratories, Dobbs Ferry, N.Y. He has served on various committees, including the Acoustical Society of America's Technical Committee on Underwater Acoustics; the U.S. Navy Underwater Sound Group; the U.S. Navy Undersea Warfare Research and Development Planning Council, of which he was chairman; the Office of Naval Research Deep Water Propagation Committee; and the Technical Advisory Group for Thresher Search Operations.

He is a Fellow of the American Association for the Advancement of Science and the Acoustical Society of America and a member of the American Physical Society, the American Geophysical Union, the Seismological Society of America, Phi Beta Kappa, and Sigma Xi.