IEEE spectrum

features

Spectral lines

Cornucopian systems. We have been accused of satiating our young with a premature oversupply of the gratifying goodies of our civilization so that they arrive at the age of adulthood without any compelling motivation to strive as their predecessors did. Perhaps so, but young engineers are more fortunate. We have concealed from them an immense satisfaction, the inner workings of the cornucopia—manufacturing engineering. As we near the end of the first century of electrical engineering as an organized profession, the IEEE is finally forming a Manufacturing Technology Group.

From the very beginning the progress of our arts has been paced by the ingenuity (and hard work) needed to develop methods to implement economically the physical realizations of the beautiful mathematical models. It started with the primitive need to insulate long lengths of copper wire reliably. The transatlantic cable of the 1860s might have been characterized by the solution of a linear partial differential equation but it was a triumph of early electrical manufacturing engineering. Place yourself back then and try to visualize how you would prepare thousands of miles of cable with the combination of strength, corrosion resistance, insulation leakage, and reliability, using known materials.

Early electric power became possible as engineers learned to fabricate commutators and wind armatures. Progress has continued in those original arts to the present day. Reliable commutator motors for power tools and vacuum cleaners are now produced at a small fraction of the real cost of making them 50 years ago.

The universal use of electric lighting and the practical feasibility of electronics were made possible by the development of the automatic fabrication of lamp bulbs. Vacuum tube electronics was established, not by government subsidy, but by the development of methods of economical manufacture so that it could be put in many millions of homes despite the worst economic depression ever known.

With World War II, public money became available in large amounts to support military electronics, but with the exception of microwave radar the capability to engineer and produce quickly was based on the manufacturing engineering triumphs of the 1930s. Microwave radar was a joint product of the physicists who really understood electromagnetic fields and the manufacturing technologists who found methods such as the laminated magnetron and electroforming to mass produce the unusual geometries.

In the last quarter century new major industries—semiconductor electronics and data processing—have arisen. Like our older arts, they have only been possible because manufacturing engineers have solved a series of challenging problems. Much of what we need to know to design semiconductor devices in principle was published in a single paper by William Shockley in 1949. Many of the accomplishments in the 22 years since then have been the uncovering of processes physically to realize the models and their reduction to effective practice in the factory.

Computers are macroassemblies with no tolerance for malfunction. Because they comprise tens of thousands of logic circuits and millions of bits of storage, sophisticated manufacturing methods are required to make them physically and economically achievable. The methods cover the physical fabrication, automatic test, and automated manufacturing information.

Tape and disk storage for computers is based on properties known for many years. Its practical success is based on manufacturing methods that maintain uniform defect-free properties in the magnetic coating, pickup heads of incredible geometry, and complex mechanisms of great reliability.

The manufacturing engineer faces great challenges, but with success comes deep satisfaction. His product stays where he can live with it and improve it as it performs the task for which he planned it. There is no question of its "relevance" or usefulness. With the permanent presence of past success he can go on to new problems with the self-confidence required to find methods where none exist and where no algorithms are neatly packaged in textbooks. The manufacturing engineer has a unique opportunity to be a generalist. He must choose arts that will do his job. They may come from mechanical, metallurgical, chemical, or any other engineering specialty. He may use electron beams, ion beams, and lasers for machining or for interferometer position control. He starts with awareness of a wide range of technical options, combines them creatively, and then evaluates his alternatives economically, acquiring enough fundamental understanding to employ paper and computer simulation methods before trying hardware. He is a complete man because he does a complete job, from concept, through design, construction, prove-out, and use.

There are many reasons why the exciting nature of manufacturing engineering has been concealed. It is eclectic and directed at a succession of unique goals, hence it does not make a handy academic discipline. Competitive manufacturing companies must keep their most effective methods confidential. Whatever the reasons for the past invisibility of manufacturing engineering, we now have the Manufacturing Technology Group. May their meetings and publications do justice to their great works.

David DeWitt, Editor

Charge-coupled devices—A new approach to MIS device structures

Charge-coupled devices require no diffusion for their active parts and can be conceptually fabricated with two masking steps. These new devices are finding application as image sensors and memory elements

W. S. Boyle, G. E. Smith *Bell Telephone Laboratories, Inc.*

Recent advances in materials and processing have resulted in a new class of information-handling structure—the charge-coupled device. This three-layer structure creates and stores minority carriers, or their absence, in potential wells near the surface of the semiconductor. The minority carriers move from under one electrode to a closely adjacent electrode on the same substrate when a more negative voltage is applied to the adjacent electrode. Because of their high transfer efficiency, these devices have already found application as image sensors. In addition, there is every expectation that memories made by use of the stored-charge concept will be less expensive and faster, and will require less power than a magnetic counterpart now in use.

Rarely do important technological applications follow major advances in our understanding of the fundamental physics and chemistry of a phenomenon or a materials system. Rather, it is more likely that the improved understanding from fundamental studies will give rise to some rudimentary device structures that may not be used for several years before the original idea finds widespread application. In the past year, the IMPATT and transit-time oscillators, for example, have become important as sources of power for millimeter waves. These devices had their genesis in work performed by W. Shockley and W. T. Read a decade earlier. The injection laser, first demonstrated in 1962, is only now becoming a device in which the understanding of a heterojunction materials system has proceeded to a point at which continuous room temperature operation is possible.

Similar delays are apparent in the most important of all semiconductor technologies: low-level signal processing. Throughout the past decade, integrated-circuit technology has been dominated by bipolar transistor structures that comprise metal and silicon dioxide films on silicon and use a process technology of diffusion and oxide-masking. There has been little tendency to vary this system even though the many changes in materials and processing have been explored experimentally.

The reason for the delay between structural concept and application is easy to find. Although it may be possible to demonstrate the feasibility of an idea in a novel device structure, its wide application depends on a complete understanding of both the physical phenomenons involved and the materials and processes required to fabricate the structure. For these reasons, progress in major device technologies tends to await the development of a large amount of background information. Improve-

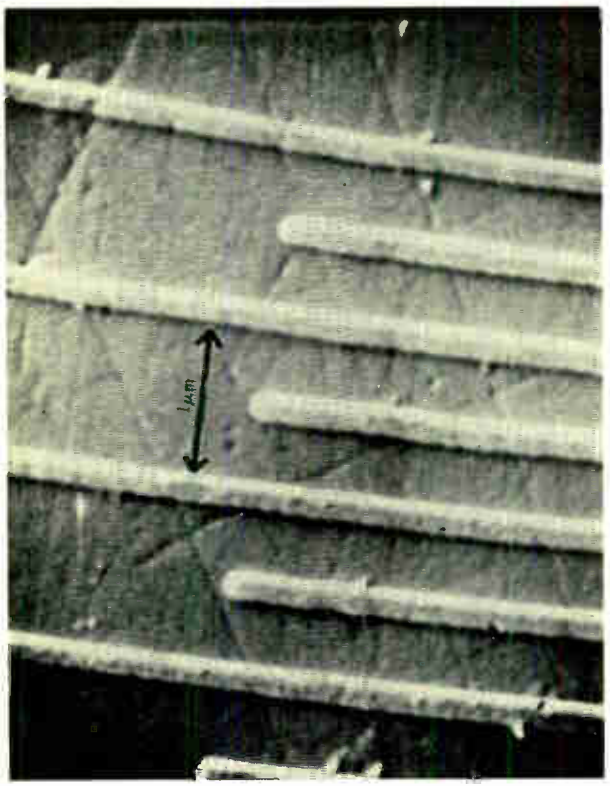
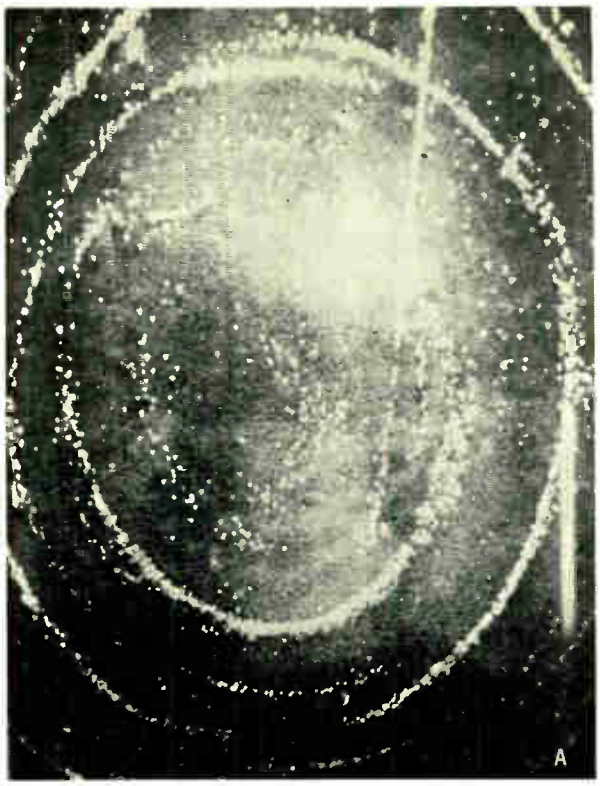


FIGURE 1. A piezoelectric surface wave transducer with metal stripes formed by electron-beam resist techniques.



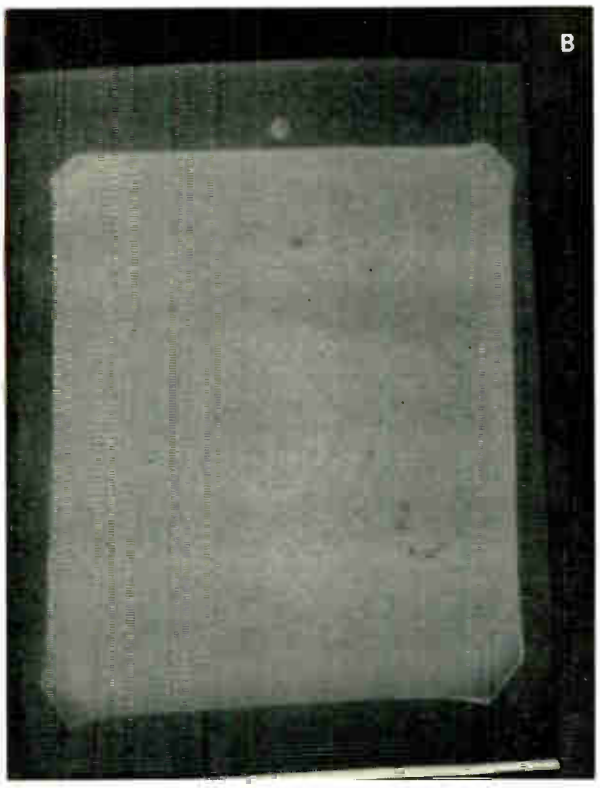


FIGURE 2. (Above right) A—Video display of excess currents from defects in a silicon-diode array camera tube. B—The same display from a defect-free array.

ments then occur in quantumlike jumps. Integrated-circuit technology is at the stage where in the next few years we are going to see just such a major change. The following areas, in which significant progress has been made, lead to this conclusion.

Semiconductor progress

First, after more than a decade of intensive study, we are beginning to understand interface states in the simplest systems—such as silicon dioxide on silicon. This understanding not only encompasses the physical description, but also includes the phenomenological knowledge of how to preserve desirable properties under severe stress. Of particular importance are dual dielectrics, which give the benefits of the low surface-state density available with silicon dioxide together with the high stability from less permeable materials, such as silicon nitride and aluminum oxide. Semiconductor devices can now be made that do not require the protection of a hermetically sealed enclosure. This single development changes the whole nature of integrated electronics by allowing a new freedom in interconnecting chips; no longer do the connections have to pass through vacuum seals. In this way a much higher order of integrated circuit is possible, with complete subsystems such as digital processors and memories taking the form of one large hybrid circuit.

The ability to fabricate structures with a thickness of less than one micrometer is a second factor. For some time it has been possible to produce insulating films of high quality in this thickness range, but only recently has ion implantation produced well-controlled doping profiles with such dimensions. An improvement in precision

of at least one order of magnitude over any other process for both the number of impurity ions and their position has been clearly demonstrated.

Furthermore, new techniques in pattern generation have decreased lateral dimensions for masking operations. Figure 1, for example, shows a device fabricated by A. Broers *et al.*[1] that consists of a metal pattern laid down on a piezoelectric material to form a transducer. The individual fingers are 0.17 μm wide and are spaced on 0.5-μm centers. He uses a fine electron beam to expose a specially developed photoresist material, poly-methyl methacrylate, and then proceeds in a conventional way to employ this material as a mask to etch the patterns. This kind of fine geometric control makes it possible to fabricate devices with improved high-frequency performance and to make integrated circuits with higher packing density. It also presents the opportunity of fabricating structures with lateral dimensions that are comparable to depletion-layer widths and oxide thickness, a factor of importance in at least one device family.

A third area of progress is in the preservation of the homogeneity of the semiconductor material throughout the fabrication of device structures. All high-temperature process steps—such as diffusion and oxide growth—degrade the crystalline perfection of starting material. Only recently, however, did the full extent of this damage become evident or even important. The problem was brought very forcefully to our attention recently when we attempted to fabricate silicon-diode-array camera tubes—arrays of 500 000 individual diodes on a single slide of silicon. In the first devices we made, defects were a serious problem. Figure 2A shows the image of the leakage current obtained from a target made three years ago with the best technology we had at that time. Each one of the white spots is a single diode with excess leakage current. A normal leakage current is about 10^{-14} ampere. The main source of the difficulty has been traced to clusters of imperfections introduced at the time that high-

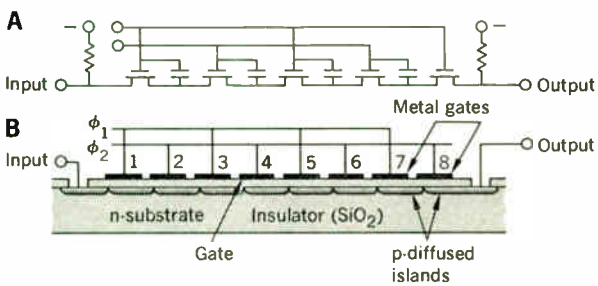
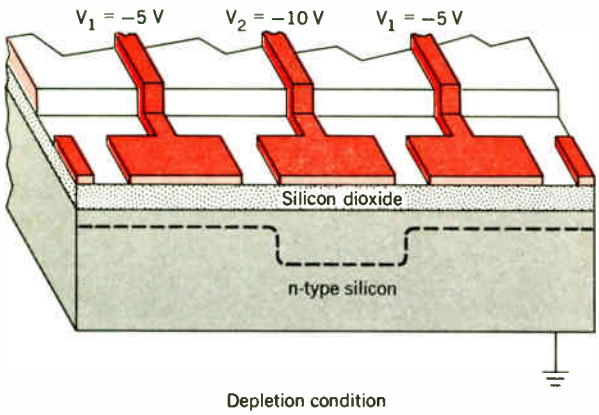


FIGURE 3. A—Circuit diagram for an IGFET bucket brigade. B—Cross section of the integrated circuit form.

FIGURE 4. Cutaway of a charge-coupled device. The dashed line represents both the edge of the depletion region and the potential distribution. Voltages are typical.



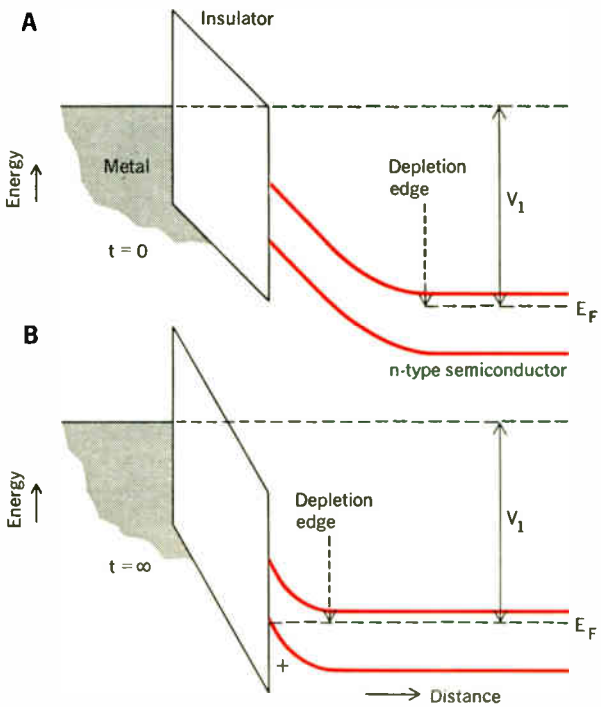
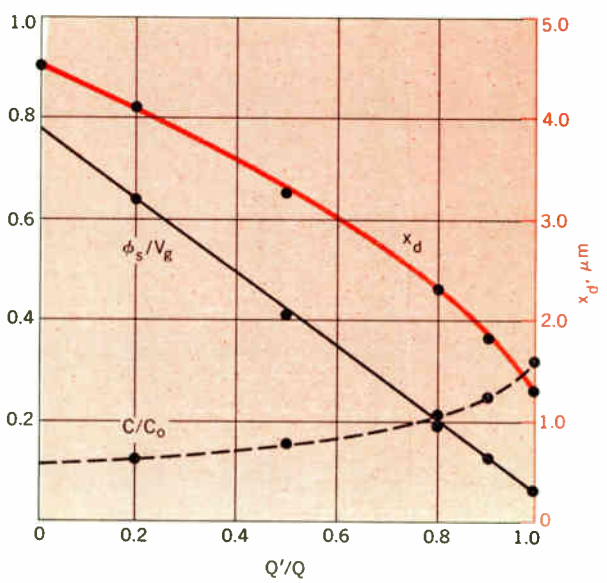


FIGURE 5. Plots of electron energy as a function of distance through an MIS structure. In (A), t = 0, no charge is stored at the surface. Charge is stored at the surface in (B), t = ∞.

FIGURE 6. A theoretical plot of depletion width (x_d), surface potential (ϕ_s), and capacitance (C) of an MIS structure as a function of charge at the interface (Q′). Q is the maximum amount of charge that can be stored under conditions described in the text.



temperature oxidation takes place. Subsequent development efforts have led to nearly complete elimination of such defects (Fig. 2B).

These advances have come about from trying to improve an existing technology. Where these possibilities will lead is uncertain; however, to be more specific, we shall discuss a class of information-handling devices that make use of most of these advances. The common factor in this device family is that information is represented by stored charge. Several such structures have been reported in the past year. One class uses stored charge in an *x-y* array of elements for a random-access memory.[2] There are many commercially available insulated-gate field-effect transistor (IGFET) dynamic memories in which advantage is taken of the large off-resistance of an IGFET to store charge on a gate capacitance.

Another element, reported by P. T. Panousis,[3] consists of a metal oxide semiconductor (MOS) capacitor that is charged and interrogated through a bipolar switch. A second class of devices includes shift registers, in which charge is passed from element to element in a linear array.[4,5] An integrated-circuit version has been reported by F. L. J. Sangster.[6,7] It consists of either bipolar or IGFET transistors that are connected in series, with capacitors connecting base to collector or source to drain. Figure 3 illustrates this circuit for IGFETs. All of these devices are conventional circuit elements connected by wires, even though they can be put in integrated-circuit form.

Charge-coupled devices

A third class, charge-coupled devices (CCD), has been reported by the writers,[8] and the principle has been confirmed experimentally.[9] These devices represent a conceptually simple form of functional charge-storage device that cannot be constructed from discrete components. The device stores a minority-carrier charge in potential wells created at the surface of a semiconductor and transports the charge along the surface by moving the potential wells. In its simplest form the structure that accomplishes this consists of an array of closely spaced metal electrodes that overlay an insulator deposited on a uniformly doped semiconductor substrate. Figure 4 shows an n-type bulk (minority carriers are holes) device in a situation wherein a sufficiently large negative bias potential has been applied to all the electrodes to produce inversion, and the center electrode has a slightly larger applied potential. There is no inversion layer, only depletion, because a negligible number of minority carriers are present. This is not a steady-state situation; nevertheless, the surface will remain depleted for times of the order of seconds before thermally generated minority carriers accumulate because the present silicon art produces material with low densities of surface and bulk-generation centers. If minority carriers are introduced through any one of several means to be described shortly, they will collect at the surface in the potential minimum defined by the excess potential on the central electrode. The potential-energy diagram for the metal-insulator-semiconductor (MIS) structure is shown in Fig. 5. The Fermi energy of the metal appears on the left; on the right the band edges of the insulator and semiconductor are shown as functions of distance perpendicular to the surface. On the far right the Fermi level in the n-type semiconductor is shown. The bias voltage V_1 is the applied potential difference between the semiconductor and metal. Figure 5A shows the potential distribution in the absence of collected minority carriers; Fig. 5B illustrates the potential distribution when a saturated number of minority carriers has been collected. In the saturation condition shown in Fig. 5B, distribution of minority carriers is such that their diffusion current away from the surface is exactly balanced by their drift toward the surface. If less than the saturation value is accumulated, the net flow is toward the surface, but if the saturation value is exceeded, there is a net flow into the undepleted bulk,

FIGURE 7. Cutaway of a charge-coupled device in (A) the storage condition and (B) the transfer condition.

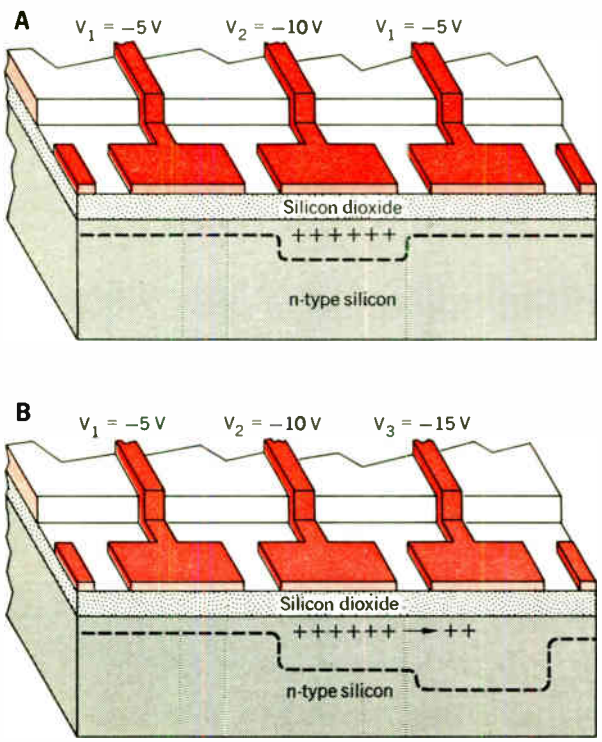
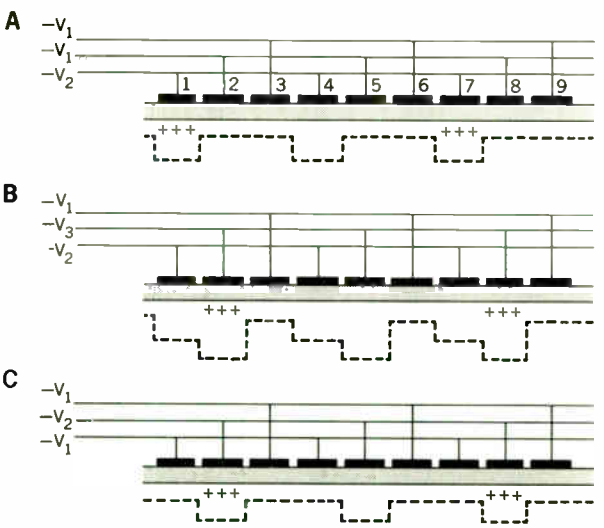


FIGURE 8. A three-phase charge-coupled device.



where the minority carriers recombine, just as in a forward-biased p-n diode. With the collection of minority carriers in the depletion well, the depletion width decreases as shown in the potential-level diagram; therefore, the differential capacity increases and, as shown, the surface potential decreases. Note that the minority-carrier accumulation held at the surface cannot change with rapid changes in electrode-to-bulk voltage. Therefore, the capacitance phenomenon consists only of mobile charge removal or addition at the depletion edge as the electrode bias is increased or decreased.

In Fig. 6, these three quantities—depletion width x_d, differential capacity C, and surface potential ϕ_s—are plotted as a function of the ratio of accumulated minority carriers to the saturation value. The plot here is for the particular values of doping density $N_d = 5 \times 10^{14}$ cm^{-3}, gate bias $V_g = 10$ volts, oxide thickness $x_o = 2000$ Å. This results in an oxide capacitance of $C_o = 1.7 \times 10^{-8}$ F/cm^2. The Debye length, which is a measure of the sharpness of the depletion edge, is 5×10^{-6} cm, which is much less than the depletion width. For convenience, ϕ_s and C are plotted as normalized values to the gate bias and oxide capacity, respectively. As minority carriers

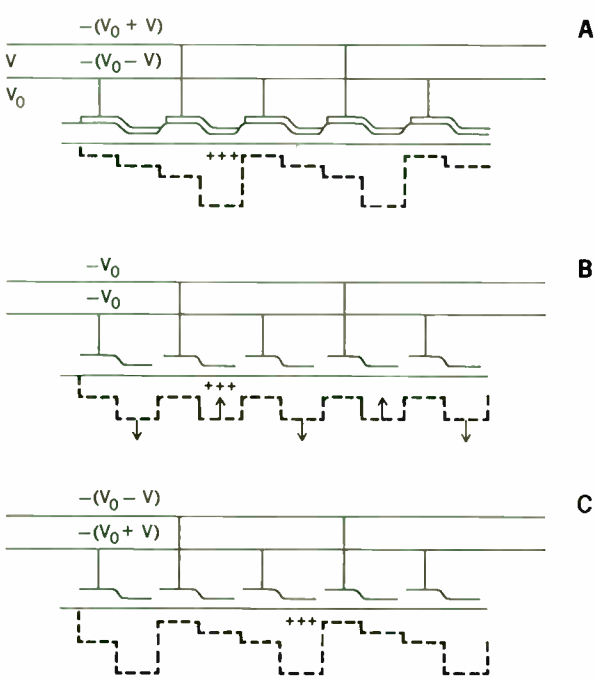
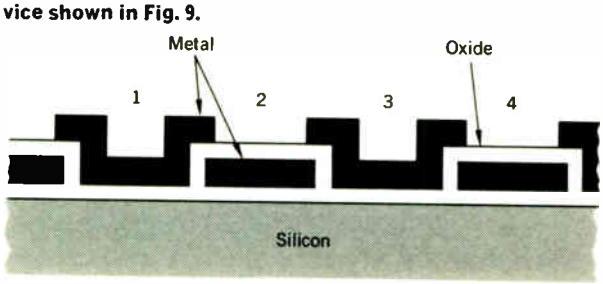


FIGURE 9. A two-phase charge-coupled device.

FIGURE 10. Basic overlapping metallization structure. The oxide under plates 1 and 3 would be made thicker and plates 1 and 2, 3, and 4 would be connected to make the device shown in Fig. 9.



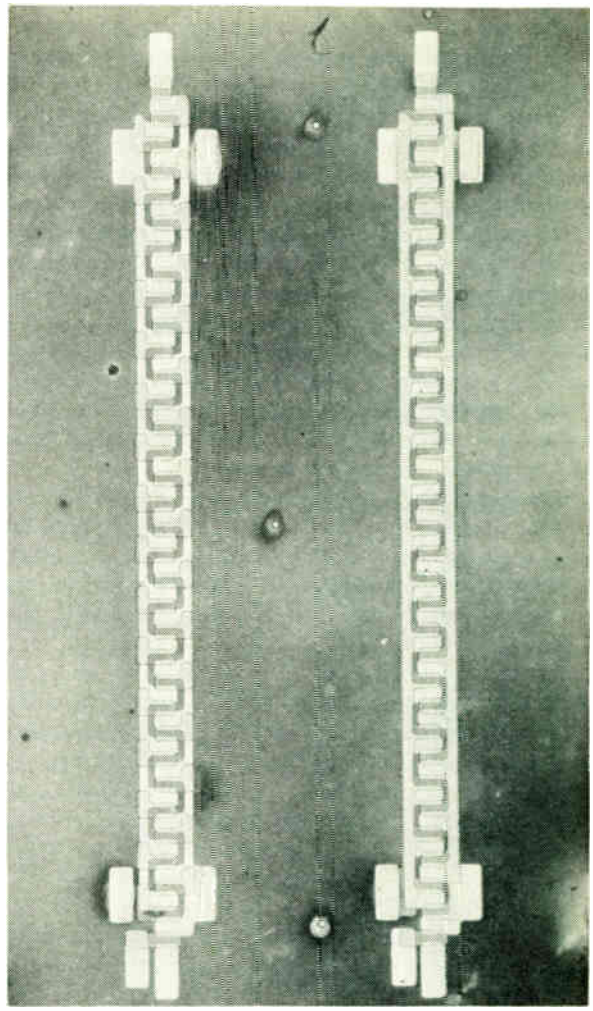
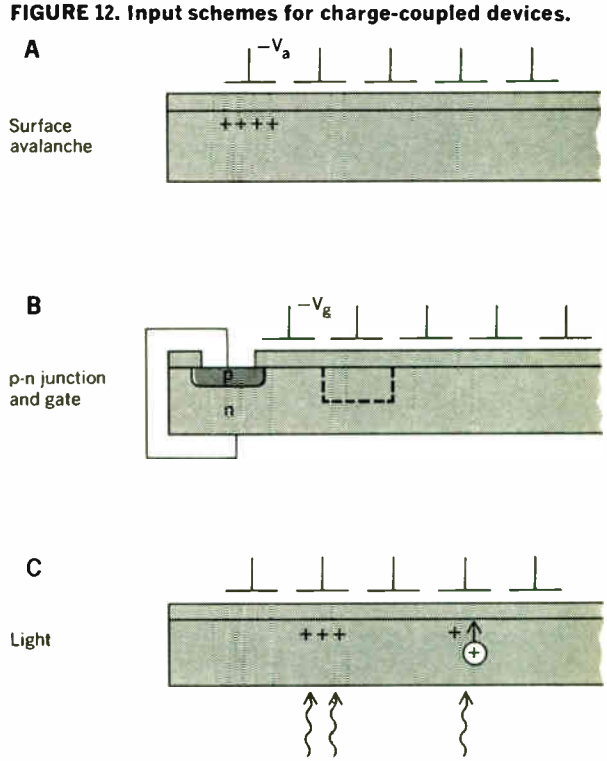


FIGURE 11. A four-phase charge-coupled device that uses silicon-gate technology.

FIGURE 12. Input schemes for charge-coupled devices.



accumulate, the differential capacity increases by a factor of 3, and the surface potential by a factor of 8. These increases provide a large signal for determining the presence or absence of charge. The physical position of this charge in the storage condition is indicated by the plus signs in Fig. 7A.

The interesting feature of this structure is that it is possible to place the metal electrodes close enough together and to apply voltage differences of sufficient magnitude to obtain the potential distribution shown in Fig. 7B. In this case, the bias potential V_3 exceeds V_2, which causes carriers to be transferred from one electrode to the next. Subsequently, the potentials on the electrodes can be readjusted so that the quiescent storage site is located at the third electrode. Figure 8 shows just such a sequence, which effects transfer along a linear array. The electrodes are connected in groups of three and operated with a three-phase voltage supply to give direction to the transfer operation. The left column shows the potentials applied to each electrode. The response V_1 is sufficient to provide a depletion region in the semiconductor. Although V_2 is larger than V_1 and produces the storage site previously described, V_3 is still larger and effects the transfer. In Fig. 8A, charges are stored under electrodes 1 and 7; no charge is stored under electrode 4. In Fig. 8B, the potentials on 2, 5, and 8 are increased so that the charges move over one position. In Fig. 8C, electrodes 2, 5, and 8 have become the storage sites. In this way, coded information can be shifted along the linear array. A variety of geometries other than this three-phase structure utilize the same basic concept. One (Fig. 9) is a two-phase system in which the potential on adjacent electrodes alternates between $V_0 - V$ and $V_0 + V$. The electrodes have steps along their length; that is, the oxide thickness is stepped so that a different potential appears beneath each individual electrode.[19,21] The shifting operation is analogous to that obtained with the triplet geometry previously described.

Figure 10 shows a basic structure that implements the two-phase device by connecting the electrodes either directly, capacitively, or with an external dc bias. The use of overlapping electrodes also reduces the separation between plates from a photolithography tolerance to an insulator thickness. Alternatively, the device can be used with a four-phase drive. Figure 11 shows such a device in which silicon-gate technology is employed.

At the start of an array of shifting electrodes the charge can be generated by any one of a number of means. Some of these are shown in Fig. 12. In Fig. 12A, charge is generated through the application of a large enough pulse to produce avalanche breakdown in the semiconductor. Figure 12B illustrates a means of injection that uses the left electrode to produce an inverted region adjacent to a p-n diode formed in the bulk. The p region is wired to the

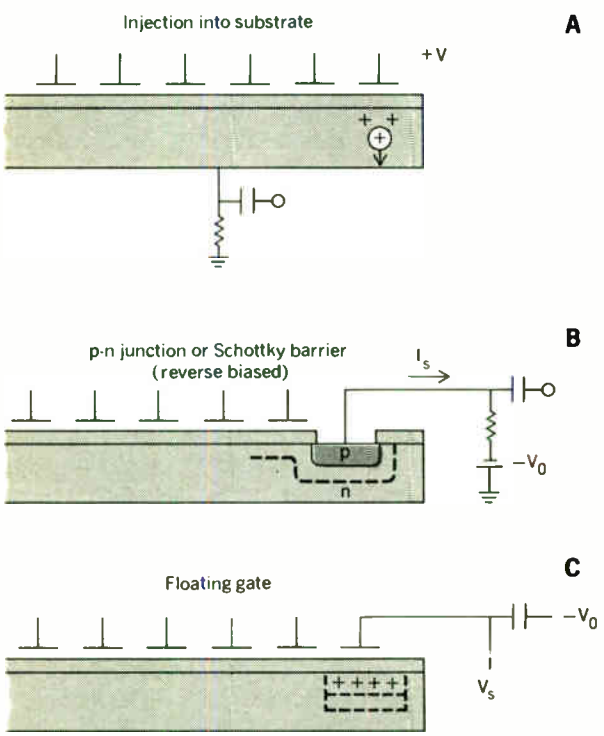


FIGURE 13. Output schemes for charge-coupled devices.

FIGURE 14. Longitudinal cross section of a charge-coupled 8-bit shift register.

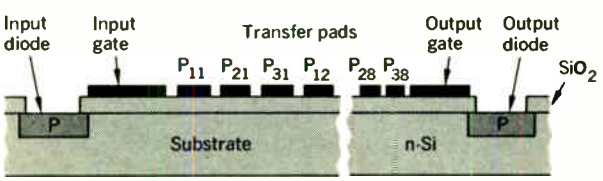
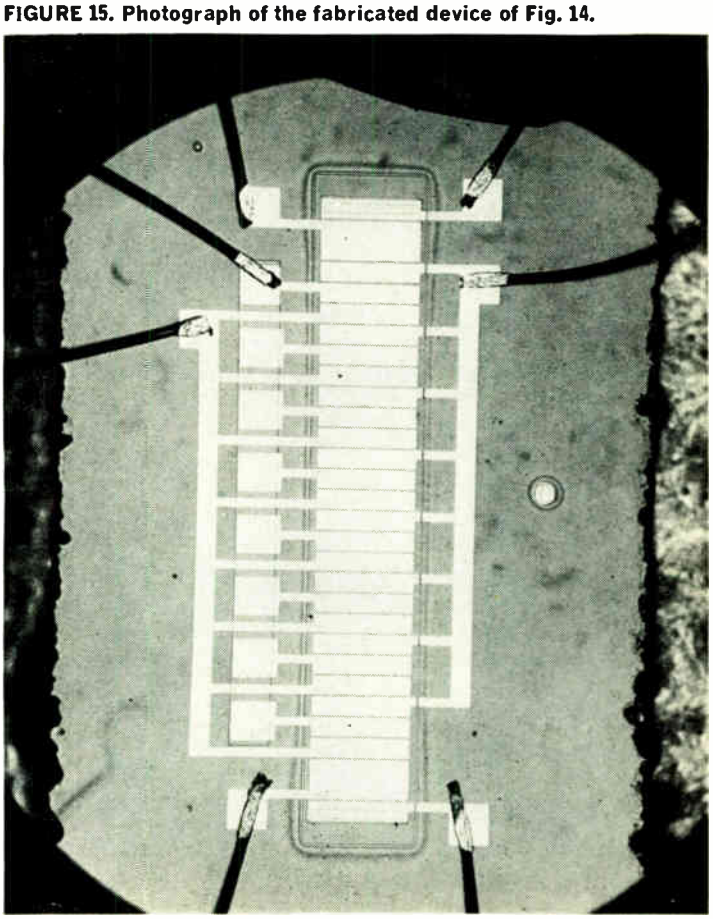


FIGURE 15. Photograph of the fabricated device of Fig. 14.



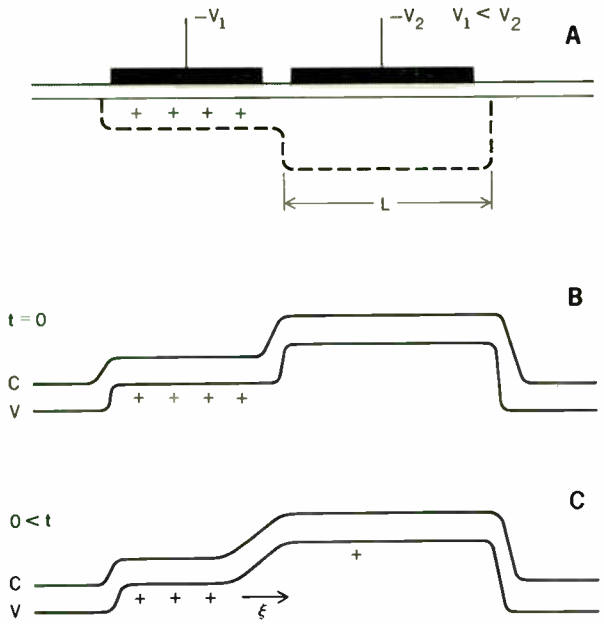
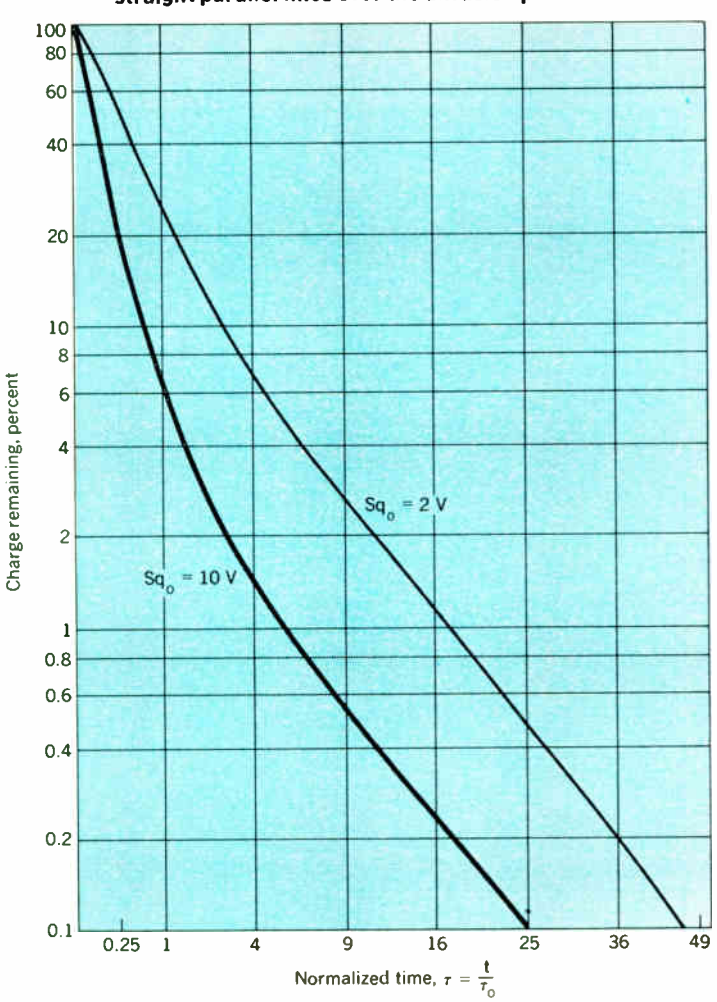


FIGURE 16. The transfer of charge from one plate to another. C and V refer to the conduction and valence bands.

FIGURE 17. The fraction of charge remaining under a plate as a function of normalized time. The abscissa is plated as $\sqrt{\tau}$, which spreads the region where $\tau < 1$, and results in straight parallel lines over the diffusion portion.



bulk and becomes a source of minority carriers to support the inversion. A minority-carrier current flows along the surface to the right and delivers a minority-carrier accumulation to the region under the next electrode, which is biased to produce a potential well. This is similar to gating charge from source to drain in an IGFET. Figure 12C shows that the minority carriers can be generated optically and collected under the electrodes. In this latter mode of operation the device operates as an imaging device and has the built-in features of image storage and scanning.

Figure 13 illustrates methods of detection. In Fig. 13A, the substrate connects to ground through a resistor. When charge is transferred to the last electrode, a positive voltage causes holes to be injected into the substrate, and a current passes through the ouput resistor. In Fig. 13B, a diode at the end of the line is reverse-biased to a voltage $-V_0$, which is more negative than any of the surface potentials used for transfer. When charge is transferred to the diode position, it produces an output current current I_s in the external circuit. In Fig. 13C, the fact that the capacitance of the MOS structure changes with charge is used in a capacitive division circuit, where the voltage V_s could be connected, for example, to the gate of an IGFET.

A cross section of one of the device structures studied[10] is shown in Fig. 14. The transfer pads are arranged for a three-phase operation, P_{11}, P_{21}, P_{31}, $\cdots$, P_{38}, with every third pad connected together as shown in Fig. 10, where the three-phase drive was illustrated. Injection is controlled by the potential on the input gate, which can induce an inversion layer that connects the first plate to a source of minority carriers at the input diode. A symmetrical arrangement is used for detection of the output signal with a reverse-biased diode.

Figure 15 illustrates this structure. In the central section, where the transfer pads are located, there is a thin (1200-Å) oxide and beyond that region a thick (5000-Å) oxide, which prevents depletion in regions other than beneath the transfer electrodes. All of the second electrodes in transfer triplets are connected together by means of a diffused crossunder. In this early device, the plates are 250 μm by 50 μm and are separated by 3-μm gaps.

Of particular interest are the mechanisms that inhibit the complete transfer of charge from one plate to the next. Our early work indicates that this is not a simple problem and depends on at least two distinct factors. The first involves the dynamics of the free carriers that undergo motion through both the mechanisms of diffusion and drift. The second involves trapping effects at interface states. Our preliminary data indicate that, with the material system and geometries we have used, both these effects are present but are not sufficiently large to preclude useful applications. We have, for example, made observations on structures where the total amount of charge left behind at each transfer is less than 0.1 percent at shifting times of 2 μs.

A qualitative feeling for the magnitude of these two effects may be obtained by first considering the dynamics of minority carriers. Figure 16A illustrates the distribution of carriers the instant after bias has been increased on the transfer electrode of length L. Figure 16B shows the corresponding potential distribution. In Fig. 16C, the potential distribution has changed because a few of the minority carriers have moved out through the action of

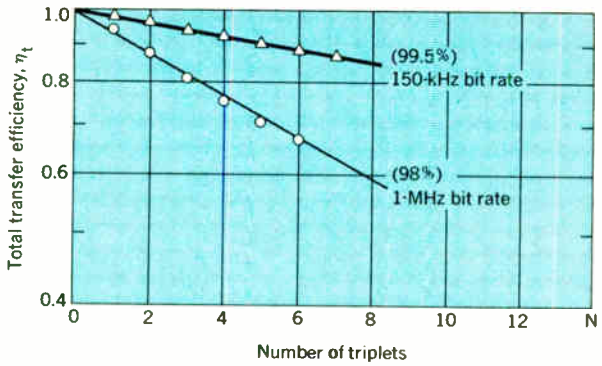
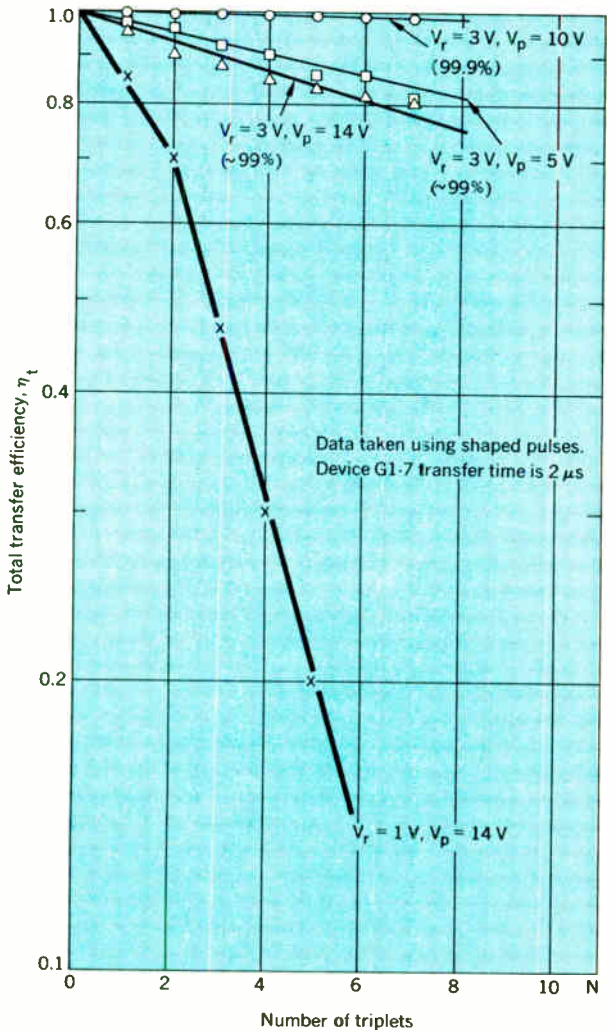


FIGURE 18. Transfer efficiency as a function of the number of bits for 150 kHz and 1 MHz.

FIGURE 19. Transfer efficiency as a function of the number of bits for various operating voltages.

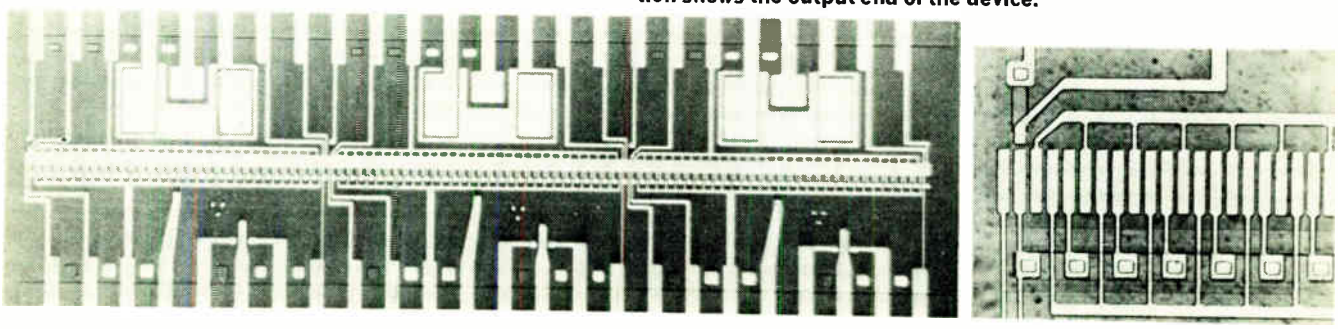


diffusion alone. At this point in time, however, the rearrangement of potential is such as to provide a drifting field for some of the remaining carriers over part of their path. This effect, as well as the enhanced transfer efficiency resulting from fringing fields that exist in finite geometries, are discussed qualitatively in Ref. 8. Mathematically, this means solving the transport equation in the presence of a potential distribution that is a function of the free carrier density. Approximate solutions to these coupled equations show that the drift component is much larger than the diffusion component except for the last fraction of charge to be transferred.[11,12,20] A plot[12] of the amount of charge remaining at a site as a function of time after a square-wave pulse is applied to an adjacent electrode is shown in Fig. 17. The normalization time is $\tau_o = L^2/\mu u_o$, where L is the length of a plate, μ is the mobility, and $u_o = 1$ volt. The quantity $V_s = Sq_o$ is a measure of the initial quantity of charge stored. There is a rapid drop initially, which is faster for a large amount of charge (storage voltage = V_s = 10 volts) than for the smaller amount (V_s = 2 volts). When only a small amount of charge remains, the slopes change as both become diffusion-limited. These calculations essentially agree with the experimentally observed charge-transfer efficiencies.

The upper limit of the amount of charge lost to surface states can be estimated as follows: The maximum number of charges stored N_m is the oxide capacity times the gate voltage divided by the electron charge. For typical values this is 2×10^{12} charges/cm^2. If the surface is kept under inversion at all times, only those states that have a time constant that corresponds to the time allowed for shifting will be effective in reducing transfer efficiency. With a density of states of 10^{11} cm^{-2} · eV^{-1} and if a band kT wide has the right time constant, the maximum fraction lost will be 2×10^{-3}.

The transfer mechanism has been studied[13] in an optical injection experiment whereby a light source, which moves along the array, provides a reproducible amount of injection at any one of the electrodes in a sequence. Figures 18 and 19 show some data that have been obtained in this way. In Fig. 18, the total efficiency plotted as an ordinate is the ratio of the amount of charge recovered at the collecting diode when the light beam is positioned at one of the electrodes divided by the charge obtained when the beam is positioned at the last electrode. The abscissa is the number of triplets along which the charge has passed. The voltages used are the same at

FIGURE 20. A 96-bit, three-phase CCD. The enlarged portion shows the output end of the device.



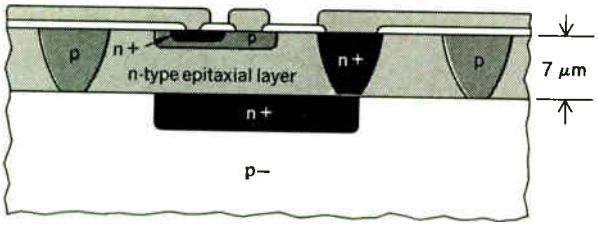


FIGURE 21. A cross section of a junction-isolated planar transistor.

both frequencies and the percentage, in parentheses, is the transfer efficiency per electrode. We do not have enough data as yet to map out in detail the frequency dependence of the transfer process; however, the early data and calculations indicate that the transfer efficiency falls off rapidly with frequency changes. The transfer phenomenon is even more complicated for the transfer efficiency as a function of resting and transfer potential. Figure 19 shows that the transfer falls off very rapidly as the resting potential is reduced from 3 volts to 1 volt. There is a further complication in the variation of the transfer with the transfer potential V_p (with a constant bias potential). An unexplained maximum occurs for a transfer potential of 10 volts. Some of the complexity may arise from the relatively large separation between electrodes, which results in poorly defined potentials in the regions of free oxide surface lying between the electrodes. Experiments under way use structures similar to Fig. 11 and will give much better control of the potential distribution over all regions through which the carriers move. In this way, some of the anomalous results may be resolved. These better-defined conditions will provide a new tool to explore experimentally and directly new facets of interface states. Not only should we be able to determine surface-state densities very close to the band edge, but we will also be able to examine the capture and emission cross sections as a function of the electric field parallel to the surface.

Advantage has been taken of these high transfer efficiencies in applying the charge-coupled-device concept to image sensors[14] (Fig. 12). Figure 20 illustrates a three-phase 96-bit linear array. Charge is transferred from right to left in this photograph, with input–output diodes located at intervals of 32 bits. This device has 10-μm plates with 3-μm spacings between them and has been run at 2 MHz with better than 98 percent efficiency.

The device structure that uses the charge-coupled process is significant because it makes use of some of the basic mechanisms in semiconductors, and it may also suggest some new experimental tools for basic measurement. However, it also typifies the direction that the technology is taking at this time. It does this in a number of ways.

Trend to simpler structures

For the past ten years the majority of integrated circuits have used a basic structure that has evolved directly from planar discrete-device technology. Figure 21 shows a cross section of a junction-isolated planar transistor. A deep p-type diffusion is formed at the periphery of the n-p-n transistor at the center of the diagram. A buried collector, shown by the n^+ region at the center, is contacted by another deep diffusion through the epitaxial layer. In total, five separate diffusion steps are required, each with its accompanying oxide masking and photolithography steps. After all the desired impurity profiles have been developed, two more masking steps are required to make ohmic contact to the transistor and delineate the metalization pattern. In the past few years, some simplification of bipolar structures has taken place. For example, by the use of much thinner epitaxial layers, it is possible to arrange for the n^+ collector diffusion to perform the dual function of providing isolation and making contact to the collector. Even further simplification has been reported recently by B. T. Murphy *et al.*[15] The field-effect devices have from their earliest days been conceptually easier to fabricate—a standard insulated-gate field-effect transistor requires only one diffusion for the establishment of the source and drain regions.

Charge-coupled devices require no such diffusion for their active parts and can be conceptually fabricated with two masking steps—one for the metal pattern, another to define the regions where thick and thin insulating film is required. In practice, IGFET-type structures are used for input–output and regeneration at the periphery of large arrays. Since these elements constitute only a small fraction of the area of a chip, overall yields should be high because of the basic simplicity of the CCD elements.

New applications for semiconductor devices

By and large, the major application for all semiconductor devices has been in the time-honored role of an active device that can provide gain or act as a switch. Indeed, it is surprising and somewhat disheartening to find that all integrated circuits are still recognizable in terms of discrete building blocks of resistors, capacitors, and transistors. At last the situation is beginning to change. This is now the case in the search for elements that will perform the memory function. The change is also of particular importance in some large digital machines, such as electronic switching systems, where there is at least an order-of-magnitude more memory elements than logic gates or switches. In the past, a bistable flip-flop circuit has been used to form a memory cell without the use of magnetic elements. Today, most of the fast memories made from semiconductors make use of this principle. A much more natural approach is to make use of charge storage. Some promising results have been obtained in MIS structures by using a dual dielectric and storing charge at the interface between the two dielectrics or in a floating metal gate sandwiched between two insulators.[16,17] Unfortunately, successful operation depends on being able to pass current in a controlled manner through one of the insulators. Three more direct approaches were mentioned earlier. There is every expectation that memories made by the use of the stored-charge principle will be less expensive and faster, and will require less power for operation than any of the magnetic memories that are now in existence. It should also be possible, of course, to make use of the shifting principles that are inherent in the charge-coupled devices to devise new ways of interrogating memories other than the coordinate address approach that is used now.

New approaches to signal processing

As measured by the fundamental physical limits, all present digital machines are extremely wasteful of power.

A figure of merit for a logic gate is equal to power times switching delay. This quantity has the units of energy required to process one bit of information. In practice, the lowest delay product of the best semiconductor devices lies in the range of 10^{-12} joule. R. Landauer[18] has shown by phenomenological reasoning that no more than a few times kT should be required for the switching function. Because kT is approximately 10^{-20} joule, there are eight orders of magnitude between today's performance and what is theoretically possible. Nowhere else, at least in the field of electronics, is there such apparent inefficiency! One does not have to search far for the reasons for such poor practical results. Because we are using junctions to obtain nonlinearities—in other words, switching action—the potential variations must be at least several kT/q volts or a fraction of a volt. The smallest structures that can be fabricated have dimensions that produce a capacitance of 10^{-12} farad. At each switching operation, CV^2 of energy must be dissipated and this is approximately equal to 10^{-12} joule.

One system for performing logic that circumvents these practical limits for junction devices is the one that uses magnetic bubbles for logic.[22]

The energy dissipated in the magnetic material per switched bit may be as small as 4×10^{-14} joule. Similarly, a highly idealized charge-coupled device in which charge is transported by a moving sine-wave potential well gives the same loss per bit (at 10 MHz) on the semiconductor slice. Both calculations neglect power-supply loss.

These charge-coupled devices had their origin in the search for semiconductors with device mechanisms similar to those obtainable from magnetic domains. Whether they will be able to remove the eight orders of magnitude between practice and theoretical limits is not yet clear. One thing, however, is evident: Much is still left to be done in getting a better understanding of the fundamental physics and chemistry of semiconductor structures. This is an extremely fruitful field for the development of devices of practical importance.

REFERENCES

1. Broers, A. N., Lean, E. G., and Hatzakis, M., "1.75 GHz acoustic-surface-wave transducer fabricated by an electron beam," *Appl. Phys. Lett.*, vol. 15, p. 98, 1969.

2. Wood, J., and Ball, R. G., "The use of insulated-gate field-effect transistors in digital storage systems," *ISSCC Dig. Tech. Papers*, pp. 82–83, 1965.

3. Panousis, P. T., "A TRIM bipolar charge storage memory," presented at International Electron Devices Meeting, Washington, D.C., October 28–39, 1970.

4. Janssen, J. M. L., "Discontinuous low-frequency delay line with continuously variable delay," *Nature*, vol. 149, pp. 148–149, Jan. 26, 1952.

5. Hannan, W. J., Schanne, J. F., and Woywood, D. J., "Automatic correction of timing errors in magnetic tape recorders," *IEEE Trans. Military Electronics*, vol. MIL-9, pp. 246–254, July/Oct. 1965.

6. Sangster, F. L. J., "Integrated MOS and bipolar analog delay lines using bucket-brigade capacitor storage," *ISSCC Dig. Tech. Papers*, pp. 74–75, 1970.

7. Sangster, F. L. J., and Teer, K., "Bucket-brigade electronics—New possibilities for delay, time-axis conversion, and scanning," *IEEE J. Solid-State Circuits*, vol. SC-4, pp. 131–136, June 1969.

8. Boyle, W. S., and Smith, G. E., "Charge-coupled semiconductor devices," *Bell Syst. Tech. J.*, vol. 49, pp. 487–493, 1970.

9. Amelio, G. F., Tompsett, M. F., and Smith, G. E., "Experimental verification of the charge-coupled device concept," *Bell. Syst. Tech. J.* vol. 49, p. 593, 1970.

10. Tompsett, M. F., Amelio, G. F., and Smith, G. E., "Charge-coupled 8-bit shift register," *Appl. Phys. Lett.*, vol. 17, p. 111, 1970.

11. Engeler, W. E., Tiemann, J. J., and Baertsch, R. D., "Surface charge transport in silicon," *Appl. Phys. Lett.*, vol. 17, p. 469, 1970.

12. Strain, R. J., and Schryer, N. L., "A nonlinear diffusion analysis of charge-coupled device transfer," *Bell. Syst. Tech. J.*, July/Aug. 1971.

13. Amelio, G. F., private communication.

14. Bertram, W. J., "Application of the charge-coupled device concept to solid-state image sensors," *1971 IEEE Internat'l Conf. Dig.*, pp. 250–251.

15. Senhouse, L. S., Kushler, D. L., and Murphy, B. T., "Base diffusion isolated transistors for low power integrated circuits," *IEEE Trans. Electron Devices*, vol. ED-18, pp. 355–358, June 1971.

16. Kahng, D., and Sze, S. M., "A floating gate and its application to memory devices," *Bell Syst. Tech. J.*, vol. 46, p. 1283, 1967.

17. Frohmann-Bentchkowsky, D., "A fully-decoded 2048-bit electrically-programmable MOS-ROM," *ISSCC Dig. Tech. Papers*, pp. 80–81, 1971.

18. Landauer, R. W., *IBM J. Res. Develop.*, vol. 5, p. 183, 1961.

19. Kosonocky, W. F., and Carnes, J. G., "Charge-coupled digital circuits," *ISSCC Dig. Tech. Papers*, pp. 162–163, 1971.

20. Kim, C. K., "Carrier transport in charge-coupled devices," *Dig. Tech. Papers*, pp. 158–159, 1971.

21. Tompsett, M. F., "A simple charge regenerator for use with charge-coupled and bucket-brigade shift registers and the design of functional logic arrays," *ISSCC Dig. Tech. Papers*, pp. 160–161, 1971.

22. Bobeck, A. H., and Scovil, H. E. D., "Magnetic bubbles," *Sci. Am.*, vol. 224, pp. 78–90, June 1971.

Reprints of this article (No. X71-071) are available to readers. Please use the order form on page 9, which gives information and prices.

Willard S. Boyle (F) is executive director of the Semiconductor Components Division at Bell Telephone Laboratories, Murray Hill, N.J. A native of Canada, he received the B.Sc., M.Sc., and Ph.D. degrees from McGill University in 1947, 1948, and 1950 respectively. Dr. Boyle joined Bell Laboratories in 1953 and was later appointed head of a department that did some of the early work on solid-state lasers. In 1962 he was named director of space science and exploratory studies at Bellcomm, Inc. He returned to the Laboratories in 1964 and assumed his present position in 1969. Dr. Boyle is the author of a number of published articles, holds 11 patents, and is a member of APS.

George E. Smith (SM) is head of the Interface Device Department at Bell Telephone Laboratories, Murray Hill, N.J. He received the B.A. in physics from the University of Pennsylvania in 1955 and the M.S. and Ph.D. in physics from the University of Chicago in 1956 and 1959 respectively. He joined the laboratories in 1959 and initially studied the electrical properties and band structure of semimetals. His primary interests are in the areas of new semiconductor devices and the device physics of semiconductor–insulator interfaces.

Spectrum conservation in the Land Mobile Radio Services

The assignment of a new, higher-frequency band for radiotelephone and private mobile radio systems is giving industry a fresh opportunity to develop techniques for more efficient use of one of our most precious resources—the RF spectrum

H. Staras, L. Schiff **RCA Laboratories**

This article summarizes the ideas that are being actively discussed for use in mobile radio and, in particular, how they apply to the new 900-MHz region assigned to the Land Mobile Radio Services category. Over the next decade some of these methods will prove out and others will probably fall by the wayside. It is clear, however, that some new techniques will be necessary if mobile radio is to continue its growth.

The Federal Communications Commission recently made available 115 MHz of bandwidth in the 900-MHz region for mobile radio use.[1] It is easy to pinpoint the reason for this new allocation: very serious congestion of mobile radio traffic at lower frequencies, particularly in urban areas. Some of the congestion is no doubt attributable to a rather wasteful method of assigning frequencies[2] (the so-called bloc assignment method that, for example, reserves channels in New York City for forestry service). The FCC has recognized this difficulty in its decision to establish new and more effective frequency-assignment techniques.[3] As one initial step in that direction, the Commission has authorized mobile radio systems to operate on UHF television channels 14–20 in those urban areas not using those channels for television broadcasting.[4] Despite this, it is realized that the steps taken so far are only stopgap measures[5] and that a longer-term solution is required.

The insistent request over a period of many years by the mobile radio community for relief has led to the reservation of the 115 MHz within the 900-MHz band. Part of the band will be set aside for common carriers to establish a truly high-capacity radiotelephone system and part for private mobile radio systems. In effect, the assignment of this band gives industry a fresh start. The fact that these frequencies are twice as high as the highest frequencies at which commercial mobile radio systems operate today means that commercial equipment for this band must be substantially redesigned. All this occurs at a time when engineers and the public at large are realizing that the RF spectrum is, after all, a precious, nonrefurbishable, natural resource. As a result, industry is giving increasing thought to ways of using these new frequencies more effectively and efficiently than heretofore. Our purpose in this article is to describe the various proposals that have been made for increasing spectrum efficiency. Before we do so, however, it is important to understand both how mobile radio systems operate and how they are used.

The communications channel

The most important factor influencing the types of mobile radio systems that can be efficiently utilized is the nature of the propagation phenomenons involved. This discussion is limited to propagation in urban areas because it is in these areas that spectrum congestion is most acute. More specifically, it is assumed that no line-of-sight path exists between the mobile station and the base station with which it communicates. This is at once the most common and most severe condition. The nature of propagation under these conditions has been considered by a number of investigators[6–8] over the years and is reasonably well understood.

At a sample point in a heavily built-up urban area where the various structures act as scatterers, the radio field seen by a mobile receiver may be considered to be the sum of many individual waves, each with a different phase, direction, and amplitude. The magnitude of the signal picked up on the vehicle's receiving antenna depends on whether the various signals tend to add in phase or out of phase, as shown in Fig. 1. This situation gives rise to the well-known result that the amplitude probability distribution is of the Rayleigh form. As the vehicle moves from one sample point to the next, the field changes because the relative phases change and, as just mentioned, the statistics of field strength over a "sufficiently small" area are Rayleigh.

In this "sufficiently small" area, the many waves of different phase and direction give rise to a complex standing-wave pattern. As a vehicle travels through this standing-

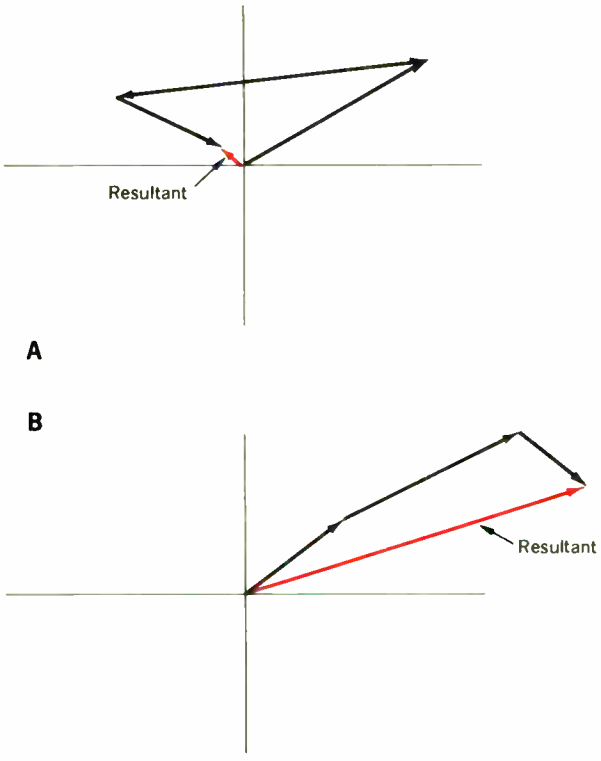
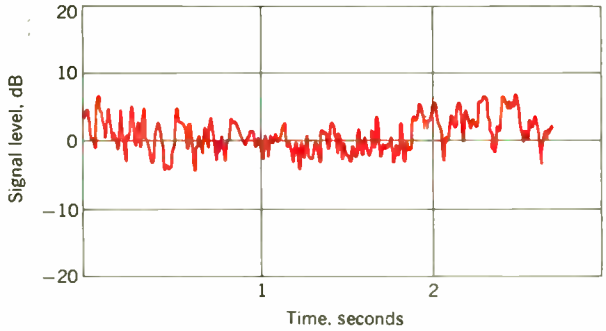


FIGURE 1. Examples of Rayleigh fading. A—Individual wave phasors add to produce near cancellation of signals. B—Individual wave phasors add to produce strong signal.

FIGURE 2. Waveform recorded at 892 MHz; speed 32 km/h.



wave pattern, the spatial variations are converted to temporal variations in field strength. The receiver in the vehicle thus experiences "fading." This fading is analogous to the fading observed in troposcatter communication in that the cause of the fading is the motion of the scatterers relative to the receiving antenna. In the troposcatter case, the scatterers move and the receiver is stationary. In the case of mobile radios, the scatterers are fixed and the receiver moves.*

For mobile radios the vehicle using a simple antenna that samples the field strength (either electric or magnetic) will experience maxima every $\lambda/2$ units at most, where λ is the wavelength. As a result, the maximum fading rate is $f = 2v/\lambda$, where v is the velocity of the vehicle. In the 900-MHz band the wavelength is about 30 cm. A vehicle traveling at 48 km/h (or 13.5 m/s) will therefore experience a fading rate of, at most, 88 Hz. A typical plot of field strength vs. time in an actual random field is shown in Fig. 2. There is another result of interest relative to the standing-wave pattern that occurs in a heavily built-up urban area. It is to be recalled that in a waveguide having a reflection at one end, the electric field is large where the magnetic field is small and vice versa. For a random field, the results are somewhat more complicated, but the essential result is analogous—namely, that the electric and magnetic fields at any location tend to be uncorrelated.[9] This fact can be exploited, as will be discussed in the next section.

In addition to the Rayleigh fading just described, the local mean of the Rayleigh distribution varies as the general topography varies. These variations are usually described as shadow losses. Measurements indicate that Rayleigh fading applies over dimensions of several hundred meters, with a fading wavelength of approximately λ. There is, however, considerable variation in the fading wavelength of shadow losses. When loss is caused primarily by small man-made obstacles, such as houses, the wavelength can be of the order of 10 meters. When it it caused chiefly by topographical features, the wavelength can be up to 1000 meters or so.

Thus, if the mobile receiver traverses a roughly circular path, with the center at the transmitter, the received signal can be described as Rayleigh fading with a time-varying mean. As already mentioned, the time variations of the mean are long term in nature compared with the Rayleigh fluctuations themselves. The variations in the Rayleigh mean can sometimes be correlated with the topography, but often this is difficult. Because of the complexity, it is easiest to resort to a statistical description. If the mean of the Rayleigh, for a circular path around the base station, is sampled and its distribution plotted, it is found that the distribution is approximately log normal with a typical standard deviation of 7 or 8 dB.[10]

As a result, the path loss in decibels between the base station and mobile can be, somewhat arbitrarily, broken into four components

$$X = X_1 + X_2 + X_3 + X_4$$

where X = path loss between stations; X_1 = free-space path loss between stations; X_2 = average excess path loss above free-space loss; X_3 = variation, about the mean, of excess path loss; and X_4 = Rayleigh variation of path loss about the local mean.

Note that X_3 and X_4 can be either positive or negative. The net combination of X_3 and X_4 is also approximately log normal with a somewhat larger standard deviation. Typical values of X_2 are 20 to 30 dB.[11] X_2 is also dependent on range and antenna height. Tentative evidence indicates that the median field at 900 MHz varies somewhere between an inverse third and inverse fourth power of distance.

As mentioned earlier, propagation is the major factor determining the types of communication systems that are most appropriate. Other factors include the effect of equipment imperfection upon the signal and also the effect of various kinds of noise. The distortion produced by equipment imperfections, notably intermodulation distortion, and the effects of thermal noise and of noise

* The other vehicles in the street can also act as scatterers. As they move, some amount of temporal fading is experienced even if the receiver is fixed. But this is a less prominent phenomenon.

added by heterodyning operations are problems usually encountered in communications systems design. Although the necessarily rugged nature of the mobile transceivers exacerbate these problems compared with, say, point-to-point communications systems, the differences are largely quantitative, not qualitative. On the other hand, the very nature of mobile communications in an urban environment makes man-made noise a major problem. This has two aspects. In the first place, the background man-made noise is much higher in built-up urban areas than in, for example, suburban locations. There is every reason to believe that this problem will be less severe at 900 MHz than at the present mobile radio frequencies since man-made noise decreases with increasing frequency.[12] However, just as in the case of fading, it is the variations about the average that create the problems. In a certain percentage of the locations the mobile unit can find itself in a man-made noise environment that will exceed the average and therefore cause problems. In fact, a potential difficulty may be the ISM (industrial, scientific, and medical) band in the middle of the new mobile radio allocation. The microwave heating units operating at this frequency may only cause difficulty in a small area, but in that area the problem could be severe. The second aspect of man-made noise is that each mobile unit carries with it a noise generator—namely, the noise from the vehicle's ignition system. Because of the system's proximity to the receiver, this is not an easy problem to overcome.

Communication techniques

This section deals with techniques that can be used to provide and improve communications in the mobile radio environment in instances wherein only a single communication link is involved. A later section deals with system techniques (that is, with mobile radio links considered in their entirety). It should be borne in mind that this separation is somewhat artificial since the two considerations are necessarily tied together. For example, it would seem that spectrum efficiency can be increased by decreasing the RF spectrum for each channel. But, all other things being equal, decreased RF bandwidth means increased noise susceptibility. This can, in turn, cause more repetitions in a conversation and lead to greater message lengths. For example, it would be counterproductive to spectrum efficiency if doubling the number of channels per unit bandwidth were coupled with a 120 percent increase in average message length.

Frequency-modulation techniques are used in land mobile radio in the United States. The reason for this choice can be inferred from a recent paper by Buesing[13] in which he compares various FM and AM systems.

Data are presented showing that reducing the frequency deviation in an FM system does, of course, reduce the RF bandwidth required, but at a price: reduction in range for a given level of intelligibility (assumed high). Single sideband (SSB) reduces the required bandwidth still further, but the effective range is also reduced further. The foregoing comparisons naturally are for equal transmitter power in all cases. The data also show that SSB is particularly vulnerable to impulse noise. The major question, then, is whether the price paid in range of intelligibility is worth the spectrum saving with these techniques. As already mentioned, to some extent the bandwidth savings can be illusory. The minimum RF bandwidth that can be allotted, even with a single-sideband system, is the baseband frequency range (nominally 3 kHz) plus enough bandwidth for the carrier instability. The present technical standards call for frequency stability of ± 5 parts per million. In the 900-MHz band this standard would imply ± 5 kHz and would mean an RF bandwidth allocation of approximately 13 kHz just for SSB. FM with a 5-kHz deviation would require approximately 26 kHz, a cost of a factor of 2 in bandwidth. For this price, however, one buys greater immunity interference and impulse noise. In fact, since cochannel interference rejection is enhanced by using greater deviation in FM, some have conjectured that wide-band FM may have a higher overall efficiency because cochannel stations then could be spaced geographically closer together. To reduce the channel bandwidth for SSB would necessitate tighter standards on frequency stability. This is likely to be quite expensive and, for the commercial market, cost is a particularly important consideration. On the whole, the question of how small the RF bandwidth for a mobile radio channel can be and still provide adequate service is a complex one even in present mobile radio bands. Nevertheless, it is almost certain that FM will remain the prescribed method of modulation in the 900-MHz band.

One new technique that operation in the 900-MHz band may provide is an opportunity for improvement of performance through the use of diversity. Both analytic investigations and experimental work have indicated that with antennas separated by approximately a wavelength, nearly uncorrelated Rayleigh fading signals can be obtained and an antenna diversity system can be constructed.[14,15] Of course, this is true in the 450-MHz band as well. But at 900 MHz both the antenna and distance between antennas shrink by a factor of 2. This may make for a practical diversity system. Figure 3 shows a comparison of diversity and nondiversity systems with regard to performance against fading. It should be pointed out that in the mobile radio field uncorrelated samples can be obtained by methods other than separated dipole antennas. The energy-density antenna[16,17] is one example and is constructed on the principle, pointed out earlier, that the electric and magnetic fields do not fade in unison but are, in fact, uncorrelated in their fading.

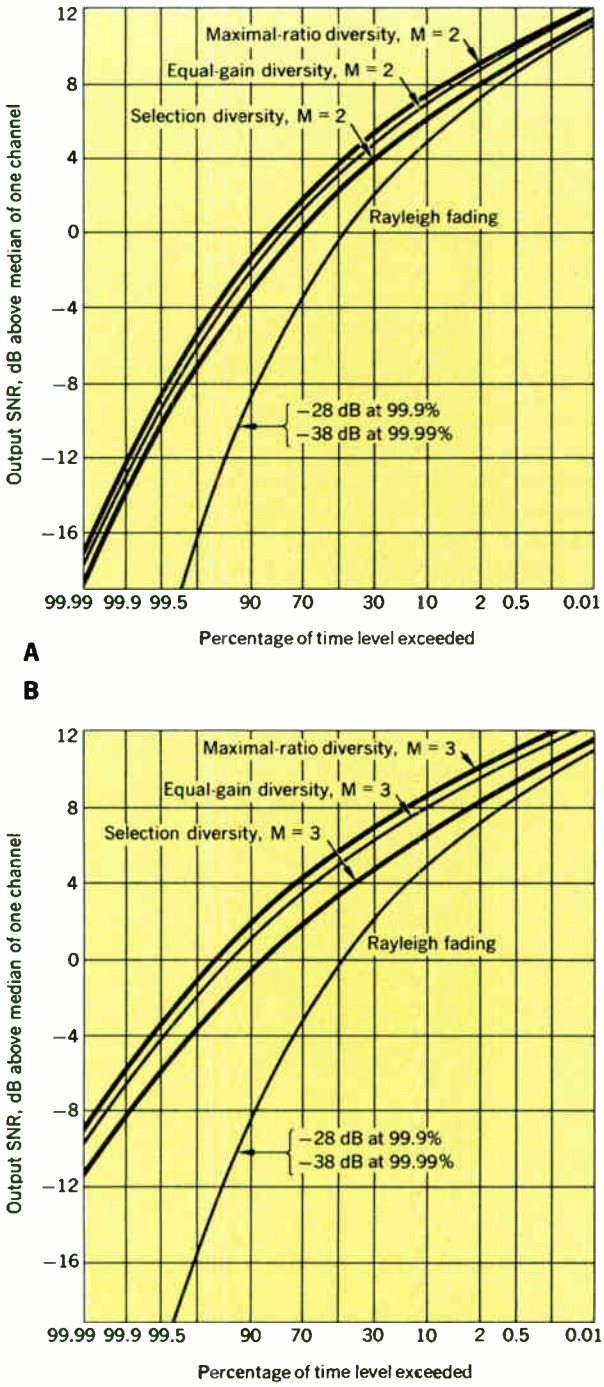
Present mobile radio operation

The discussion here shall apply only to two of the three major segments of the mobile radio business: (1) the radiotelephone service and (2) the dispatch service. We specifically refrain from discussing the third major segment of the mobile radio business, public safety mobile communications. We exclude this area because it is tailored to the specific community served and, therefore, may vary widely from place to place. Also, the requirements on reliability and survivability of equipment make it different from the other areas to be discussed.

By dispatch service we mean the typical service in which a central dispatcher communicates with one of a fixed number of vehicles in his "fleet." The fleet may consist of taxis, repair trucks, delivery vehicles, etc. The areas covered by the vehicles are quite large, typically hundreds or even thousands of square kilometers. As opposed to the radiotelephone service, it does not require interconnection with the regular telephone network nor does it require interconnection between different fleets of vehicles.

The radiotelephone service, as the name implies, is just an extension of conventional telephony to a mobile station. Aside from the obvious fact that the radiotelephone service requires interconnection with the telephone network, it differs in two major ways from dispatch-type mobile radio. The first major difference is that users are untrained in the use of their system. Because of this they cannot be expected to cooperate in making for a more efficient system. The second major difference, the average length of a conversation in each service, is partly due to the first difference and partly due to the nature of the conversations. At any rate, the average time for a conversation in radiotelephony is no different than for land telephone usage and is between 3 and 4 minutes.[18] The average length of a conversation in dispatch service is about 15 seconds.[18] These differences influence the way present-day systems are operated and the type of systems that can be considered.

FIGURE 3. Diversity improvement with Rayleigh fading for (A) M = 2 and (B) M = 3.



The operation of the dispatch service in a given urban area is a kind of organized anarchy. Each potential user tries to obtain a license to use a frequency that is as free as possible. But because of the demand and the limited spectrum allocated to mobile radio it is virtually impossible, in a major urban area, to obtain the use of a clear channel. Each user therefore shares his channel assignment with a number of other users. Since each user represents a dispatcher plus all the vehicles in his fleet, many vehicle operators and dispatchers share the same "party line." This is the anarchy. The organization is a discipline on the users that is self-imposed. They do not try to get on the air until the conversation in progress at the time is terminated, and they keep their conversations as short as possible. Although this degree of cooperation is manifested, the nonorganized nature of the method of operation hurts even more when one considers the range of each transceiver. Each user will try, all other things being equal, to maximize the range of his own equipment.

The effect of this is to squeeze more users onto the "party line," which has an adverse effect on the system as a whole. Of course, all other things are not equal and both economics and technology limit the range a given user will employ. Nevertheless, the situation today is such that when one channel is in use between a base station and a mobile, an area of several thousand square kilometers is "polluted," electromagnetically speaking—"polluted" in the sense that no other user can operate in this large area on the same channel at the same time. Note that we have used the term "traffic congestion" rather than "spectrum congestion." In terms of traffic or queuing theory, the traffic intensity is too high per server or facility (in this case one radio channel). Use of the term "spectrum congestion" implies that the only way to relieve congestion is to acquire more spectrum; this concept is not necessarily true. If the demand is too high per channel, one can increase the number of channels. This can be done by increasing the spectrum or by decreasing the RF bandwidth of each channel. As pointed out in the preceding section, however, it is not clear that this second approach will, in fact, decrease the traffic per channel since increased repetition may increase the mean length of conversations. Since usable spectrum space is very limited, the first approach is not easy to implement. We now consider other techniques to reduce traffic congestion.

One such technique used in radiotelephony is trunking. This technique is simply a pooling of facilities in such a way as to provide a more efficient way of handling traffic with the number of channels available.* Instead of having N users applied to each of n different channels, the total of nN users is given access to all n channels. Statistically this gives improved performance, basically because of the law of large numbers. Blocking occurs when the statistical fluctuations in the traffic exceed the traffic-handling capability of the system. Both average traffic and

* In analogy to the way all the trunks in a full-access trunk group are usable by a large number of telephone lines rather than each trunk being usable by a separate subset of the lines.

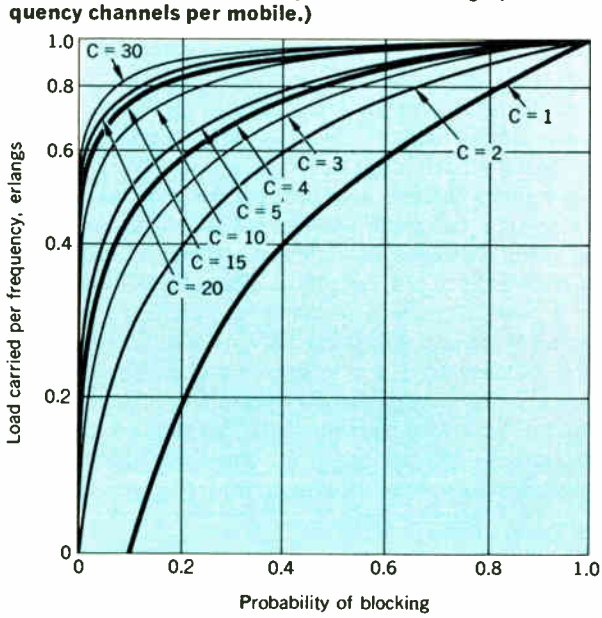
traffic-handling capability grow linearly with n, but the standard deviation of the traffic only grows as $\sqrt{n}$. Hence the probability that the traffic will exceed the traffic-handling capability decreases with n. We are thus dealing with a type of frequency diversity, although the individual probabilities of being blocked (on each channel) are not independent. Figure 4 shows the resultant traffic that can be carried per channel vs. the blocking probability when n channels are pooled in a common-access arrangement.

The blocking probability is a design parameter of a system. It can be observed from Fig. 4 that if one is trying to design a system with low blocking probability, trunking can be extremely effective in increasing the traffic-carrying capability of a system. This type of trunking has been in use for a number of years in the Bell System IMTS (Improved Mobile Telephone Service), in which eight radio channels are "trunked" in a given urban area. However, there is an economic cost to be considered for such a system; namely, each mobile must carry radio equipment that can be switched among the n channels.

On the other hand, however, when the design of a system with high blocking probability is being considered, the increase in traffic capability for trunking is not nearly so great. This is the one basic reason why trunking is not nearly so effective in a mobile dispatch service as it is in radiotelephony. It is perhaps not obvious, but careful consideration indicates that the blocking probability can be allowed to be much higher in a dispatch service than in telephony. The reason is twofold. In the first place, the radiotelephone service, as already mentioned, is an extension of land telephone service and used by unskilled operators. Although it may be impractical to try to offer a blocking probability as low as the land-line network (typically 0.01), it should at least be as low as possible. An even more basic reason relates to the average call length in the two types of operation. In telephony, it is to be recalled, the average call length is 3 to 4 minutes, but in a dispatch service it is only 15 seconds. It is to be noted that, in any traffic-carrying system, the average time for the system to become available to serve a new call, once it is blocked, is proportional to the average call duration. Therefore, the penalty for being blocked in a dispatch service is far less severe than in telephony. Hence dispatch-type mobile radio systems can function effectively at far higher blocking probabilities than telephone systems can. It is just at these higher blocking probabilities that trunking, by itself, becomes less effective.

At the risk of stating the obvious, it must be pointed out that the application of trunking is, as a practical matter, only possible when the users are connected to the RF channel via a common-user system. This situation arises naturally in telephony but has not been the historic pattern in dispatch-type mobile radio. In the spectrum-saving techniques to be discussed in the next section, common-user systems are always assumed. It is in this fundamental change in the mode of operation of dispatch systems that spectrum conservation is to be achieved.

FIGURE 4. Radio traffic carried as a function of blocking probability for various degrees of trunking. (C = frequency channels per mobile.)



Spectrum-efficient system design

To introduce the new approaches to spectrum-efficient mobile radio system design, which is the subject of this article, let us first consider the radio link* in today's mobile radio communication system. The base station, not knowing where the mobile station is, sends out a nominally omnidirectional signal, strong enough to cover all or almost all locations that the vehicle could be in. Clearly, however, only the area around the mobile station needs to be irradiated by the RF signal. If only the dispatcher had some means of knowing where the vehicle was and having an RF signal transmitted only into that area, a great deal more traffic could be handled because the same frequency could be reused simultaneously in another section of the city. This is the basic principle of a small-cell system. The urban area is divided into a number of congruent cells, or zones, as in Fig. 5. (In connection with this figure and several others in this article showing hexagonally shaped cells, it is of some interest to note that only three regular polygons—the hexagon, the square, and the triangle—can be "tight-packed" in a plane so that no gaps or overlaps occur. The hexagon's shape is closest to the nominally circular coverage pattern of the cell.) At the center of each zone is a base station that can communicate with mobile stations in the zone. Each base station is also connected by land lines (or point-to-point microwave links) to a central processing unit (CPU). The CPU is a computer-controlled switching center. In the telephony application, this CPU would be connected by trunk circuits to other switching centers and thence to the land telephone network. In the dispatch-type mobile radio application, each dispatcher belonging to this common-user system would be connected to the CPU by a land line. The communications path, in any case, proceeds by land lines from the fixed station through the switch of the CPU and from there by land lines (or microwave link) to the local cell's base station and from there by low-power radio to the mobile station. The base stations (and the mobile stations as well) require only low power because the reliable range needed (i.e., cell radius) is only a fraction of the range needed for reliable coverage of the

* There is a cable connection as well. In the simplest case it is a cable to the antenna on the roof; in other cases it is a leased telephone line to a strategically located antenna.

whole urban area (all the cells). This is especially desirable in the 900-MHz band, where it is extremely expensive to develop high-power transmission.

The advantage of this scheme, as far as traffic-carrying capacity is concerned, is that the low-power communication on a given RF channel in a certain zone allows the same channel to be reused in a number of different zones simultaneously so long as there is sufficient geographical separation between the zones. In effect, one RF channel can act as more than one server or handle more than one erlang (a unit for measuring traffic intensity).

For the communication between base and mobile in any cell to be reliable it must be recognized that the RF energy cannot be confined to that cell and will spill over to neighboring cells, as shown in Fig. 6. It will be assumed that a two-ring buffer provided around any cell using a channel is sufficient. (If more than a two-ring buffer is needed, the details are different but the principles remain the same.) If a frequency is used in a cell it may not be used simultaneously in that cell or in any of the 18 cells forming the two-ring buffer. Nevertheless it can be used in many cells simultaneously outside this forbidden region. In fact, we shall take up the question of how many cells can use the same frequency simultaneously. Before we do so, however, it should be pointed out that in one form or another a vehicle-location subsystem is needed to allow a small-cell system to function. This is an additional expense, which must be shared by the users, that has nothing to do with the communications but rather allows the communication system to function. For dispatch service, however, this subsystem in itself is very desirable since some present-day communications are involved in finding out just where the vehicle is. Another, not-so-obvious advantage is the easing of the adjacent-channel interference problem. The difficulty in mobile radio today in rejecting adjacent channels arises from the fact that the mobile stations can roam. Thus, they can sometimes be located much closer to the adjacent-channel base station than to their own base station. As a result, the adjacent-channel output is sometimes many decibels above that of the desired channel. This problem does not occur in a small-cell system since the adjacent-channel signal is coming from the same base station or from the base station in the adjacent zone.

We now return to the point under discussion: the increased traffic-handling capability of small-cell systems. There are first of all, a number of methods of operating such systems, each capable of providing increased traffic capacity. Some are more suitable for telephony, others for dispatch-type service. We shall describe two such methods here, starting with what is probably the most simple.

Fixed-frequency method. The difference between the two techniques to be discussed is the way in which the reuse of a frequency is forbidden in the zone in which it is used and in the 18 surrounding zones. In the fixed-frequency-allocation method the total number of frequencies available is divided into m separate frequency groups, each of which contains j frequency channels. A cell is assigned to one of the m groups. Each base station can communicate with as many as j mobile units in its cell simultaneously, using the j different channels in its group. The number of groups, m, is chosen as the smallest number that satisfies the interference-buffering requirements. It is easy to show that m need never be larger than 7 for 19-zone or 2-belt buffering. This is pre-

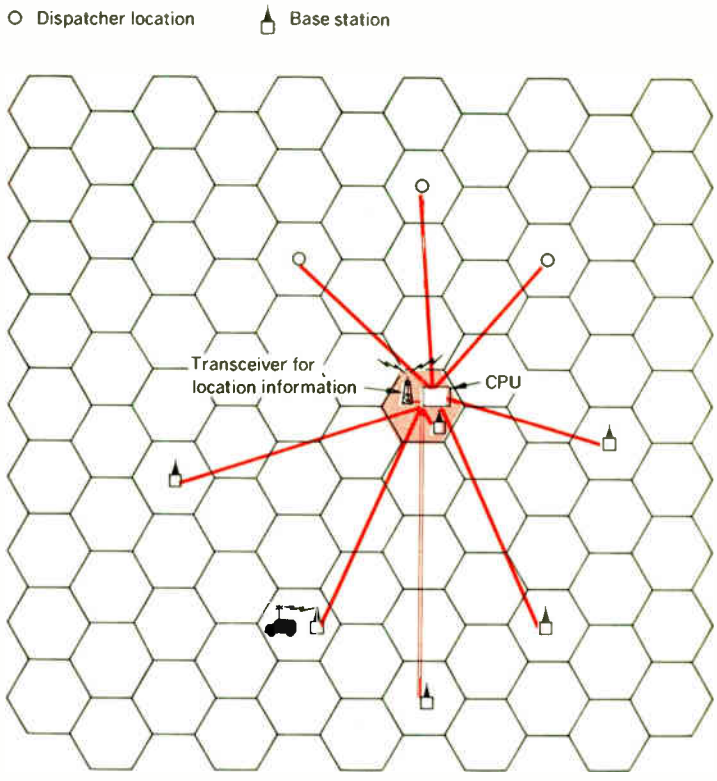
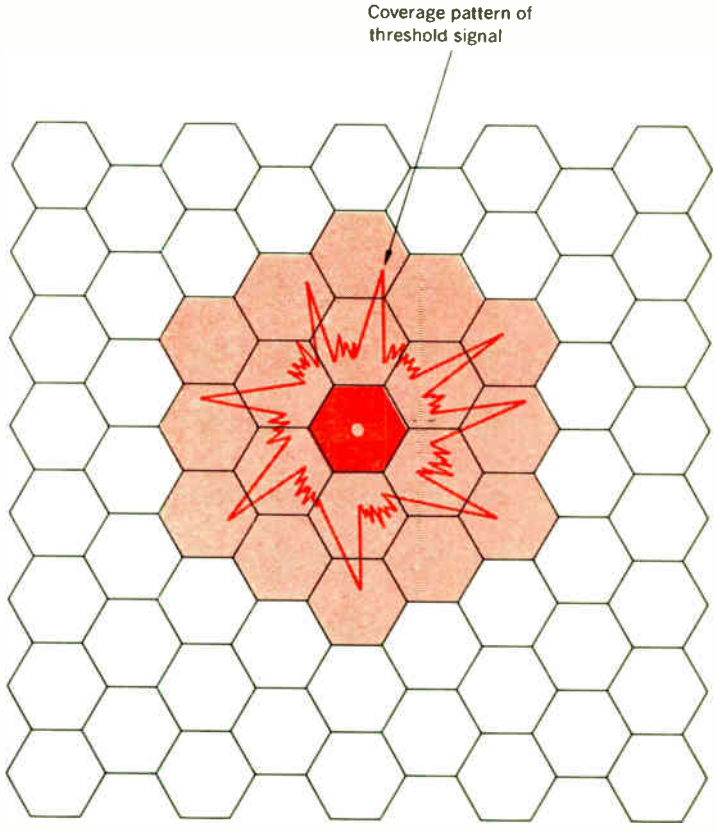


FIGURE 5. Typical layout of a metropolitan area employing a small-cell system.

FIGURE 6. Pattern of nonusable cells created by using one cell. Shaded area represents zones on which reuse of frequency employed in darker zone is forbidden.



sented graphically in Fig. 7. Each number corresponds to a group of j frequencies. For example, 1 might correspond to frequencies 1–10, 2 to frequencies 11–20, 3 to frequencies 21–30, etc. It is easily verified that no frequency may be reused within the area forbidden by its use in any one cell. An obvious advantage of this scheme is that each base station need only be equipped to communicate on j of the mj frequencies, rather than all mj. A clear disadvantage is that each mobile must be equipped for at least m frequencies to guarantee the ability to communicate in any zone even if no other traffic is present. Giving mobiles m frequencies will give only one channel in each cell (i.e., anywhere in the urban area). To provide lower blocking and to give equal service in all parts of the city require equipping mobile units with channels in groups of m; giving km channels to a mobile unit implies that it has k possible channels to try anywhere in the city.

This technique has been mentioned in connection with telephony.[19,20] Its ability to carry increased traffic, as compared with a standard radiotelephone system, is simple to explain. From the traffic point of view, the analysis is quite straightforward. If each vehicle is able to communicate on C channels (where C is a multiple of 7), it can communicate on $C/7$ channels in any given zone. Because of the way the frequency groups are allotted to cells, each zone can be considered independent of any other zone and can handle $1/N$ of the traffic of the entire urban area. Further, if the total traffic generated in the urban area is A erlangs, the traffic per channel in any given zone is $7A/CN$. Without the cellular structure, if C channels are used, the number of erlangs per channel is, of course, A/C. In a cellular arrangement, if C-channel capability is given to each mobile, only $C/7$ can be used at any one location. Hence, with a cellular structure of N cells, the nature of the blocking is as if the erlangs per frequency were reduced by a factor of $N/7$, whereas the *effective* number of trunked channels is divided by 7. In other words, although the effective number of trunked channels per vehicle is reduced by a factor of 7, the maximum erlang load that can be handled by any one frequency, citywide, is increased by a factor of $N/7$. At all but extremely low traffic intensities, the second effect is more important, and thus there is an increased spectrum efficiency, as shown in Fig. 8.

Dynamic-frequency assignment. The second method takes advantage of the fact that mobile-unit and base-station communication and control are effected through the central processing unit, at which signals from all dispatchers and base stations terminate. The central processing unit therefore has memory of which frequencies are in use by which base stations and mobile units, and it can assign channels on a dynamic basis. Accordingly, each base station is equipped to operate on *all* channels. A mobile unit can operate on one channel only or, for better service, with C channels for trunking. The central processing unit will permit communication to a mobile in a given cell at a given frequency only if that frequency is not currently in use in the cell or in any of the surrounding 18 cells. An advantage of this scheme is that a mobile unit need only be equipped with capability on one channel to provide communication capability anywhere in the city. Furthermore, if trunking is desired, any number of channels can be added; they need not be in multiples of some number as in the fixed-frequency-assignment scheme. On the other hand, each base station must be equipped for communication on all channels.

This method has been proposed in connection with dispatch service, largely because of the potential cost advantage of one frequency per mobile station operation.[21,22] The increase in traffic capacity is not as easily arrived at as in the fixed-frequency-assignment technique. An approximate analysis has been carried out by the writers[21] and is reproduced in Fig. 9. One can gain an

FIGURE 7. Reuse arrangement for two-ring buffer (m = 7).

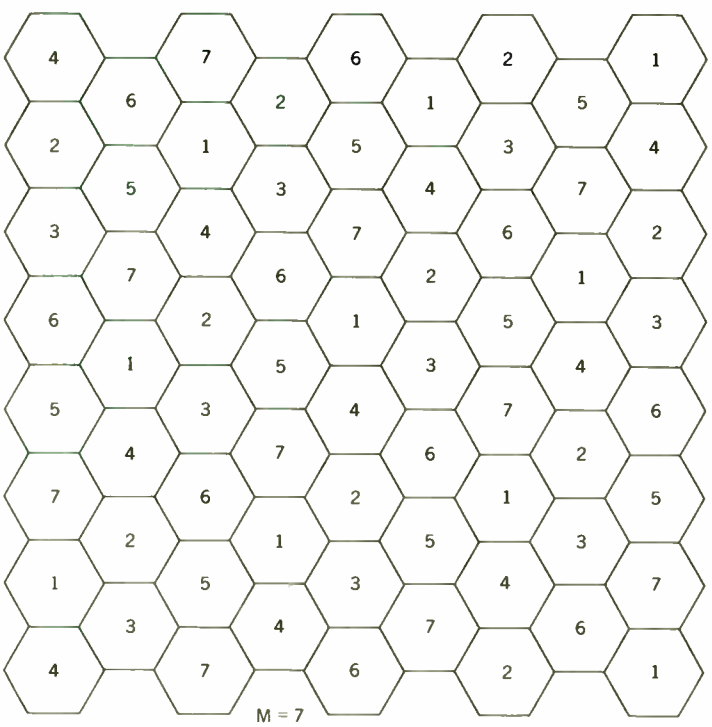
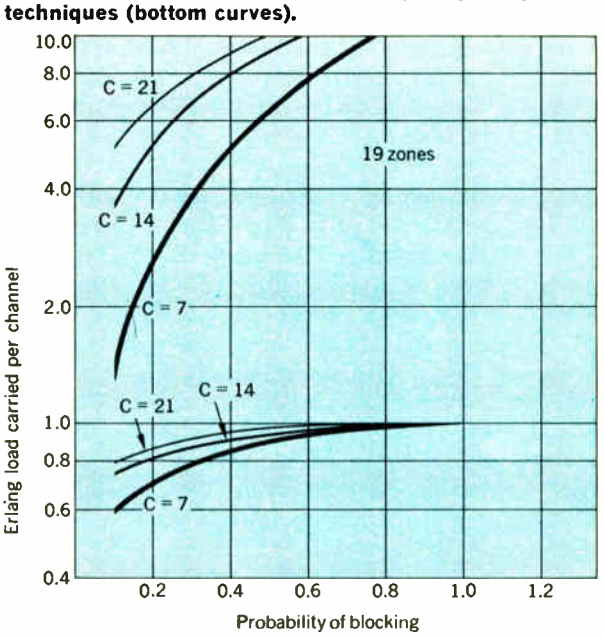


FIGURE 8. Comparative traffic improvement with trunking techniques (top curves) and fixed-frequency-assignment techniques (bottom curves).



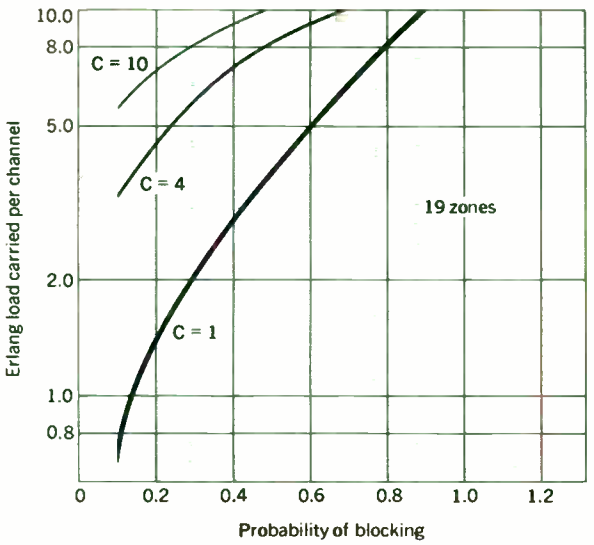


FIGURE 9. Traffic improvement through the use of dynamic-frequency assignment.

intuitive understanding of the increased traffic-carrying capacity by the following sort of argument. The system can obviously reuse the same frequency a number of times simultaneously. The tightest packing pattern is still just as shown in Fig. 7. Hence $N/7$ simultaneous uses are possible just as in the fixed-frequency scheme. But since calls arrive at random, the system may not always find itself in this tightest packing configuration. On the other hand, the number of times a frequency may be used simultaneously can certainly be lower-bounded. Each cell that uses a certain frequency makes 19 cells unusable at that frequency (that cell plus the 18 surrounding cells forming two rings). As long as 19 multiplied by the number of calls in progress on a given frequency is less than N, the total number of cells, a new call can be placed on that frequency in at least one other zone. Therefore, even in the most adverse situations a given frequency may be used about $N/19$ times; for details, see Ref. 22.

An evolutionary cellular system. Yet another type of system that has been proposed for dispatch service[18] is one that takes advantage of the zonal approach in a partial sense. In its early stages, this type of system would not use the cellular approach at all.

In the previous two systems discussed it was intrinsically assumed that the communication path between base and mobile stations consisted of a pair of frequencies—one to be used for base-to-mobile, the other for mobile-to-base communications. The spectrum savings accrue to both base-to-mobile and mobile-to-base frequencies. However, another idea is that of a common-user system in which all dispatchers are brought together to a common base station (via land lines), but in which a cellular structure is not employed. In this situation it is also possible to get some increased spectrum efficiency on the fixed-station transmission part of the channel. This is because one of the great practical difficulties, adjacent-channel interference, is substantially reduced. As already mentioned, in traditional dispatch-service mobile radio, the mobile station can roam into an area where the strength of the adjacent-channel station is much larger than the desired signal (i.e., it can be much closer to the adjacent-channel transmitter). But if all signals come from the same source the problem is much reduced and one should be able to reduce channel spacing. This naturally results in some spectrum improvement. Of course, this savings accrues only to those frequencies used for dispatcher-to-vehicle transmission. It has been estimated[18] that the traffic on these frequencies can be increased between 100 and 150 percent. However, since this is true *one way* only, the system traffic can be increased only between 50 and 75 percent. On the other hand, cellular systems can give improvements by a factor of 10 and 100 zones or so. For even larger zonal systems, the traffic intensity would increase directly as the number of cells.

In the 900-MHz band, although it might be feasible to build a fixed station with enough power to cover the urban area, it would be very difficult to have the mobile station develop enough RF power to reach a central station. Hence this common-base-station plan provides for a distributed group of receive-only stations, over the urban area, that receive from relatively low-powered mobile transmitters and relay on to the central point. This is quite similar in principle to methods already in use in mobile radio. The mobile needs to generate only enough power to transmit reliably to one of the receive-only stations. At the central point the strongest signal from the receive-only stations is selected. As mentioned earlier, the only spectrum efficiency gained is one-way and not very significant compared with what a cellular system can ideally give. However, when system load is increased and a vehicle-location system is provided, it is possible to gain spectral efficiency on the mobile-to-base frequencies. Each receive-only station can be imagined as the center of a cell or zone and different mobiles can transmit on the same frequency simultaneously if they are in cells sufficiently far removed. One can imagine either fixed- or dynamic-frequency operation.

Comparison of systems. It will be noted that all the systems ideas for spectrum conservation that have been discussed involve two essential features: small zones and common-user operation. It should be emphasized that although one can increase system spectrum efficiency by many techniques (such as trunking, decreased channel spacing, etc.), the common-user cellular system can give increases in efficiency *on top of* other increases. Of course, this improved efficiency does not necessarily come cheap. The reader will recognize that all zonal systems need some kind of vehicle-locator subsystem.

Which, if any, of the ideas we have discussed can be applied to a specific system depends on detailed economic tradeoffs that can only be hinted at here. For example, in fixed-frequency cellular systems one must have mobile stations that operate on any one of a number of channels (multiples of seven) but each base station need only be equipped for one seventh the total number of channels. In dynamic-frequency assignment, it is just the other way around. A vehicle can get by with one channel but each base station has all. The basic economics ultimately depends on the number of base stations and of mobile stations. Yet another example: Should trunking be used in a given system? In low-traffic systems, where blocking probability is lower to begin with, trunking is very effective. In systems with higher traffic it is not as effective. But in trunking each mobile is provided with many channels. It may be that once this is provided, the incremental cost of more channels may be small. Thus the economics

may indicate that systems that use trunking may be best off using fixed-frequency assignment.

Over and above the types of tradeoffs suggested by the foregoing examples, it should be pointed out that so far as cellular systems are concerned, there is a major difference between radiotelephony and dispatch service that has not yet been pointed out. In telephony, because the conversation length is on the average of 3 or 4 minutes and often much longer, provision *must* be made for transferring control to a new base station once a vehicle crosses a zone boundary. In fact, many crossings may occur during one call. On the other hand, in dispatch service with an average call length of 15 seconds, it is unlikely for a zone crossing to occur during the call and, even if it does occur, the signal is unlikely to decay drastically in so short a period of time. The fact that "handover" from zone to zone is not necessary in a dispatch service implies significant cost reductions—at least in comparison with radiotelephone service.

Details aside, the major differences between the small-cell common-user mobile radio systems and the present private-user systems are (1) substantial capital investment to implement a high-capacity common-user system; (2) reduced cost per vehicle when enough vehicles join the system; and (3) new uses for mobile radio, since the mobile equipment, requiring quite low power (a few watts at most), can be made very compact and portable.

REFERENCES

1. "First report and order and second notice of inquiry," Docket 18262, Federal Communications Commission.

2. Dayharsh, T. I., Yung, T. J., and Vincent, W. R., "A study of land mobile spectrum utilization," Stanford Research Institute, Menlo Park, Calif., July 1969.

3. Toia, M. J., "The FCC's new approach to spectrum management," presented at the 21st Annual Conference of the IEEE Vehicular Technology Group, Washington, D.C., Dec. 2–4, 1970.

4. "First report and order," Docket 18261, Federal Communications Commission.

5. "Notice of proposed rule making," Docket 19150, Federal Communications Commission.

6. Young, W. R., Jr., "Comparison of mobile radio transmission at 150, 450, 900 and 3700 Mc," *Bell System Tech. J.*, vol. 31, pp. 1068–1085, 1952.

7. Ossana, J. F., Jr., "A model for mobile radio fading due to building reflections: Theoretical and experimental fading waveform power spectra," *Bell System Tech. J.*, vol. 42, pp. 2935–2971, Nov. 1964.

8. Clarke, R. H., "A statistical theory of mobile radio reception," *Bell System Tech. J.*, vol. 47, pp. 957–1000, July/Aug. 1968.

9. Gilbert, E. N., "Energy reception for mobile radio," *Bell System Tech. J.*, vol. 44, pp. 1779–1805, Oct. 1965.

10. Okumura, *et al.*, "Propagation properties regarding mobile radiotelephone (no. 1)," Electrical Communications Laboratory, Japan, International Report, p. 3015, 1966.

11. Jakes, W. C., Jr., "New techniques for mobile radio," *Bell Lab. Record*, pp. 327–330, Dec. 1970.

12. Skomal, E. N., "The range and frequency dependence of VHF-UHF man-made radio noise in and above metropolitan areas," *IEEE Trans. Vehicular Technology*, vol. VT-19, pp. 213–222, May 1970.

13. Buesing, R. T., "Modulation methods and channel separation in the Land Mobile Service," *IEEE Trans. Vehicular Technology*, vol. VT-19, pp. 187–206, May 1970.

14. Rustako, A. J., Jr., "Evaluation of a mobile radio multiple channel diversity receiver using predetection combining," *IEEE Trans. Vehicular Technology*, vol. VT-16, pp. 46–57, Oct. 1967.

15. Lee, W. C. Y., "Level crossing rates of an equal-gain predetection diversity combiner," *IEEE Trans. Communication Technology*, vol. COM-18, pp. 417–427, Aug. 1970.

16. Lee, W. C. Y., "Statistical analysis of the level crossing and duration of fades of the signal from an energy density antenna," *Bell System Tech. J.*, vol. 46, pp. 417–448, Feb. 1967.

17. Lee, W. C. Y., "Comparison of an energy density antenna system with predetection combining systems for mobile radio," *IEEE Trans. Communication Technology*, vol. COM-17, pp. 277–284, Apr. 1969.

18. Supplemental response of Motorola Inc. to inquiry of FCC relative to Docket 18262.

19. Frenkiel, R. H., "A high capacity mobile radiotelephone system model using a coordinated small-zone approach," *1969 IEEE Commun. Conf. Record*, p. 31.

20. Araki, K., "Fundamental problems of nationwide mobile radiotelephone system," *Rev. Elec. Commun. Lab. (Japan)*, vol. 16, pp. 357–373, May–June 1968.

21. Staras, H., and Schiff, L., "A dynamic space division multiplex mobile radio system," *IEEE Trans. Vehicular Technology*, vol. VT-19, pp. 206–213, May 1970.

22. Schiff, L., "Traffic capacity of three types of common-user mobile radio communication systems," *IEEE Trans. Communication Technology*, vol. COM-18, pp. 12–21, Feb. 1970.

Harold Staras (SM) has been engaged in government and industrial research for more than 25 years. He received the B.S. degree from the City College of New York in 1944 and, after military service in World War II, returned for graduate work. He received the M.S. degree in physics from New York University in 1948. He then joined the National Bureau of Standards, where for the next six years he was engaged in radio wave propagation research. He was one of the early investigators of troposcatter propagation. In 1955 he received the Ph.D. degree from the University of Maryland. Since joining RCA in 1954 he has worked on troposcatter circuits and on a U.S. Navy classified communication project. At present he is head of radio systems research in RCA Laboratories' Communications Research Laboratory, Princeton, N.J. He is the author of more than 20 papers on radio propagation, electromagnetic theory, antennas, and communication systems.

Leonard Schiff (M) received the B.E.E. degree from the City College of New York in 1960, the M.S.E.E. degree from New York University in 1962, and the Ph.D. degree from the Polytechnic Institute of Brooklyn in 1968. From 1960 to 1966 he was employed by Bell Telephone Laboratories, Murray Hill, N.J., where he specialized in the various aspects of electronic switching systems. Since 1967 he has been on the staff of RCA Laboratories, Princeton, N.J., working in the field of communication theory. He has been particularly concerned with new mobile radio techniques and spectrum efficiency in the Land Mobile Radio Services. He has written a number of papers in this field. Dr. Schiff is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

Automation and utility system security

Continuing industry emphasis on reliability of electric utility system operation has led to widespread interest in utilizing automation technology to the best advantage for improving system security as well as economy

D. N. Ewart, L. K. Kirchmayer General Electric Company

An electric utility system is predicated to be in one of four possible states: normal, alert, emergency, or restorative. Examination of these states and the structure of the system's operating problems has led to a multilevel, multicomputer approach involving computers at various action centers interconnected by data links. Within the hierarchical computer arrangement, a four-step approach of monitor and display, contingency evaluation, corrective strategy, and automatic control is suggested. Major attention is directed toward minimizing the probability of leaving the normal state as well as the time required to return to this state. The computer functions to be undertaken at each level are described.

The electric utility industry is responding to the challenge of improving the reliability of power supply through several major steps involving both system planning and system operation. Key planning steps involve the strengthening of interconnections and more comprehensive overall planning on a regional basis.[1–4] In planning, recognition is being given to the possibility of disturbances that are more severe than the system design criterions, and automatic load shedding is being utilized to minimize the extent and duration of interruption that may result from such contingencies.

The practice of interconnecting individual power systems into large grids has resulted in economies in capital and operating expenses as well as improved reliability. The full exploitation of these benefits presents an increasingly complex problem to the power system operator; consequently, the electric utility industry is devoting greater effort to the application of automation technology to the solution of system operating problems. The relevant developments in automation technology are associated with analog and digital computers, data collection and supervisory control equipment, communication devices, and displays.[5,6]

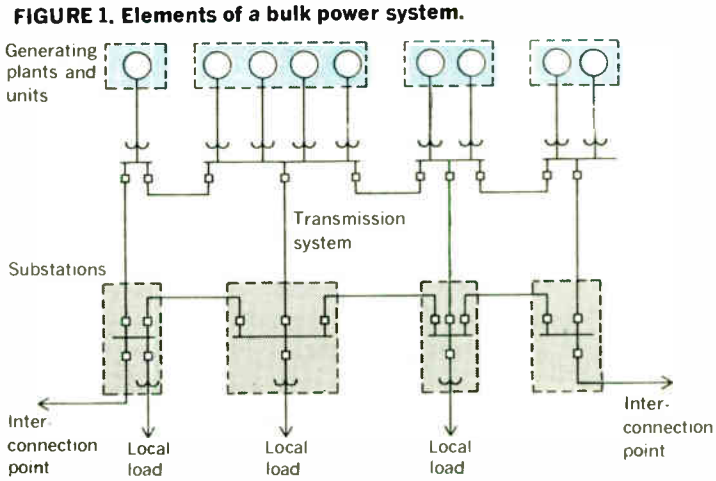
Revised text of a paper presented at the CIGRE International Conference on Large High-Tension Electric Systems, Paris, France, Aug. 24–Sept. 2, 1970.

Of particular significance has been the rapid improvement in the price-to-performance ratio of process-control digital computers. Their acceptance may be noted by the fact that as of July 1971 over 70 process-control digital computers have been installed or placed on order for company and pool dispatching offices in the United States. The computers are being applied on a time-shared basis as a valuable tool in solving the problems of operations planning, operations control, and operations accounting, all of which involve both economic and reliability considerations. The initial justification of process-control digital computers was based primarily on improvements to be achieved in fuel economy. Of even greater importance today is the application of computers to improve the reliability of system operation.

Objectives of system operation

As illustrated in Fig. 1, a bulk power system may be viewed as consisting of generating sources, loads supplied from substations, interconnecting points with neighbors, and a transmission system that interconnects these

FIGURE 1. Elements of a bulk power system.



elements. Day-to-day operation entails the use of facilities that are in place and in operation in order to meet customer requirements, within such constraints as equipment limitations, maintenance needs, fixed and variable operating costs, and interconnection agreements.

It is helpful to consider a system to be in one of the four states shown in Fig. 2. In the *normal* state, all customer demands are met, no apparatus or lines are overloaded, and there are no impending emergencies. The objective is to continue to meet customer demands, to operate the apparatus and lines within limits, to operate the system at minimum cost within the constraints, and to minimize the effects of possible future contingent events.

The *alert* state is similar to the normal state in that all customer demands are met and no apparatus or lines are overloaded. However, a potential emergency has been detected (for example, by noting that a line loading has reached a limit set below its rating or that an assumed contingency would result in a transmission overload). The objective is to impose constraints and return to the normal state in a minimum time. There will, however, be situations in which the system will reach the *emergency* state, where customer demands are not met or apparatus is overloaded. Here, the objective is to prevent the spread of the emergency. In the *restorative* state, the emergency has been stabilized, yet there may be overloaded apparatus or customer demands not served. The objective here is to return to normal in minimum time.

The normal state is the only secure state since it implies a low risk of failure. Thus, the goals of steps leading to improved system security are to minimize the probability of leaving the normal state, and to minimize the time required to return to the normal state once the system is in some other state. It is apparent that the first goal requires the steps to be taken while the system is normal.

Improved system security

A step-by-step plan for improving system security by means of computers and associated automation equipment has been proposed.[7] The following steps are involved:

1. Status monitor and display.
2. Contingency evaluation.
3. Corrective strategy.
4. Automatic control.

As a first step, it is believed that a significant improvement in system security can be obtained through increased monitoring of information. In the second step, a computer would predict the effect of contingencies and planned outages and alert operators to potential troubles on the system. As a third step, it is suggested that a computer formulate corrective strategies. At this stage, the operator could call for the execution of the strategy he chooses. In the fourth step, the computer through the communication network would execute automatically the computer-formulated strategies. This approach permits the evolutionary development of automation systems devoted to achieving increased security. In a given installation, the first two steps might be taken initially; and as experience is obtained and the required developments completed, the latter two steps could be added.

A matrix of steps and system states may be constructed as shown in Fig. 3. The columns are the four steps of development and the rows are the four system states. This matrix is most helpful in viewing the functions being undertaken by electric utility systems in meeting the goals of improved system security. Initial industry emphasis is being placed upon status monitor and display and upon contingency evaluation.

It is visualized that implementation of the four-step plan would improve data collection and communication devices and computers organized in a multilevel structure.[6-10] The lowest levels would be associated with local control and the collection of primary data. Higher levels would receive data from the levels below, process the data, and act as a source of data for the levels above. Each level acts as a coordinating agency for the levels below. A high degree of reliability is assured by placing direct control at the lowest feasible level. This multilevel computer configuration is predicted on supplementing the automatic controls already in place, such as normal protective relays, load-shedding relays, excitation controls, and load-frequency control. This structure is feasible within the framework of utility organization as described in the next section.

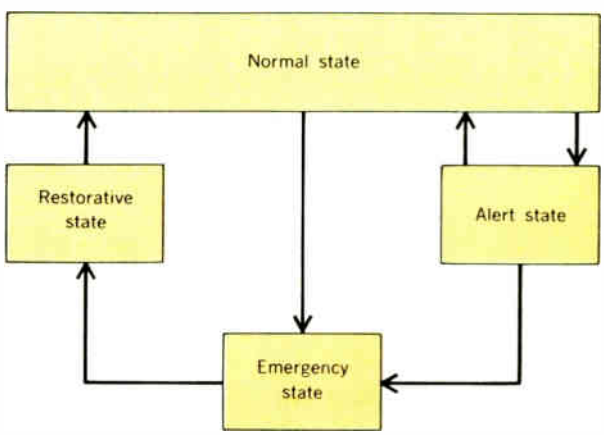
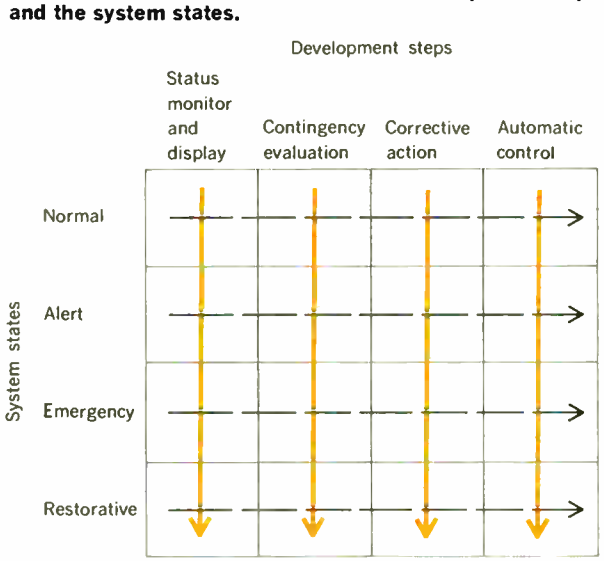


FIGURE 2. The four system states.

FIGURE 3. Interrelation between the development steps and the system states.



Hierarchy of action centers

Power companies. A hierarchy of action centers exists within a typical power company, as illustrated in the lower portion of Fig. 4. Overall responsibility for the bulk generation–transmission system resides with a company dispatching center. This center directs the second-by-second control of generation to maintain tie-line schedules and frequency, and is responsible for minimizing the cost of power generation within constraints imposed by security. Typically, such a center has access to telemetered values of all major tie-line power flows, the power output of all large units or plants, and system frequency.

Operation of the substation and transmission facilities is accomplished within transmission divisions. Division operators plan and carry out preventive and emergency maintenance and are responsible for switching operations within the division. These operators coordinate with the company dispatch center and other division centers as required.

Substations within the divisions may be manned or unmanned. Unmanned substations are monitored and controlled at a division dispatching office by means of supervisory control equipment. Manned stations may or may not have telemetry equipment installed to enable automatic monitoring by the division.

Generating plants, with the exception of remote hydro or gas-turbine installations, are staffed with operators and maintenance personnel. The operation of large thermal plants involves the local control of many plant variables and the high degree of automation achieved has contributed significantly to their reliability and economy.[11] Viewed from the company dispatch center, these plants provide a point at which generation is controlled and as a source of data on plant status.

At the dispatch center, improvements in the degree of monitoring of conditions on the bulk power supply are being sought by many utilities. Because of the limited telemetry available from major internal substations, dispatchers in the company dispatch center may wait several minutes before learning the cause of fluctuations seen on their recorders, even when the cause is within their system. Often the communication link involves several voice conversations.

Power pools and regional coordinating centers. As interconnections among utilities have become stronger, power pools have been formed in which the facilities of individual member companies are scheduled and operated in a manner that minimizes the total cost of the pool and improves its reliability. The accrued savings are shared by each member of the pool. The pool is responsible for improving overall economics, and has direct control over total generation for the purpose of maintaining net pool tie flow and frequency on schedule. For this reason, the general responsibilities and facilities of a power pool center are much like those of a company dispatch center, the difference being one of degree.

More recently, the electric utility industry has organized regional coordination centers, which have responsibilities over very large geographic areas.[2,3,12,13] These centers have been formed with the recognition that economic and reliability considerations cannot be treated completely within a single company or pool. At present they are concerned with such operating aspects as studies of future operating conditions; communicating pertinent information on current and planned status between companies, pools, and regions; coordinating emergency procedures; and analyzing disturbances. In addition, the regional center plays a role in communicating between planners and operators from a regional viewpoint.

The relationship between company, pool, and regional centers is illustrated on the upper portion of Fig. 4.

Implementation of the automation plan

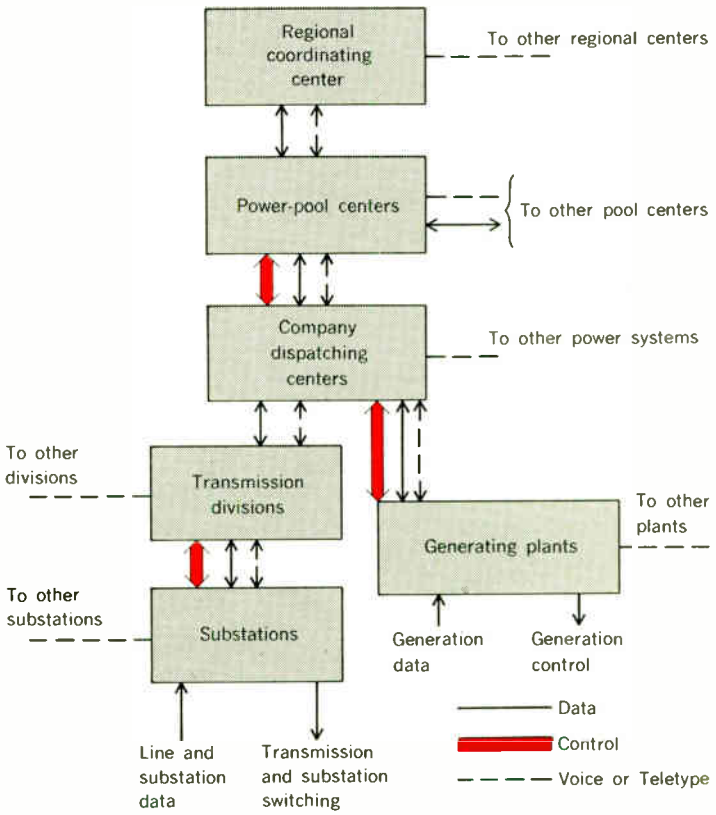
In a preceding section a four-step plan for automation of system operation was suggested. This plan entails the application of computers in a multilevel configuration. Next, the hierarchy of electric utility organization was described, in which direct control was placed at the lowest levels and system-wide coordination at the highest. In this section, recommendations are made for specific functions to be carried out at each level.

Substation level. The following functions are suggested for the substation level:

1. Data collection and logging.
2. Apparatus monitoring.
3. Evaluation of designated system conditions.
4. Switching operations.
5. Normal relaying.
6. Load shedding.

The degree to which these functions are carried out depends upon the size of the substation and upon its significance to the rest of the system. Smaller substations may be served with hard-wired logic elements and a supervisory control remote station connected to a master station or computer located in the division

FIGURE 4. Hierarchy of action centers.



office. The remote equipment should permit switching to be performed from the division level and serve as an input for primary data. Larger substations, on the other hand, even though unattended, may require small digital computers to perform some of the functions required.[14] In the case of the unattended substations, the data would be collected and transmitted to the division office, and there logged and utilized for division functions.

With a computer installed at the substation, data may be temporarily stored in the event of a failure of the communication link to the division and sent when the failure has been repaired. The computer may also perform various logical functions based upon relay actions.[15]

More complex functions, such as the evaluation of contingencies, would be done by the use of distribution factors computed at the division or company level and transmitted to the substation. The on-line calculation of substation availability, or probability of remaining in the present state, gives division operators valuable information regarding the current status of the system and provides insight into the selection of contingencies.

It is not recommended at this time that the computer undertake the primary relaying function, but consideration should be given to applying the computer for back-up relaying and the checking and modification of relay settings. Load shedding would be undertaken with conventional relay techniques and the computer would be assigned to change shedding patterns to distribute the risk among customers.

At several attended 500-kV substations, small computers are in place to perform data logging and sequence-of-events recording and to maintain a short history of key variables, which are printed out following a disturbance.[16]

Division level. The division level acts as a focal point for decisions involving substation and transmission facilities within its territory. Likewise, it serves as a concentration point for data collected at the substations. Recommended functions to be performed at this important level are as follows:

1. Data collection and logging.
2. Maintenance scheduling.
3. Evaluation of designated system conditions.
4. Switching operations.
5. Circuit protection.

In addition to requirements for data collection and processing, there is a need for analysis programs (such as load-flow and short-circuit programs) and for programs to compute distribution factors for use in contingency evaluations by substation computers and within the division center itself. Depending upon the size of the division, these programs may have modest capabilities in terms of number of busses and lines. Thus, the computer of the division office may be of small to moderate size. However, because of the critical role played by the division office in overall system security, it is recommended that a backup computer be provided, at least for all data-collection functions.

The computer functions visualized under "switching operations" include calling upon stored sequences of switching operations, preanalyzing sequences prior to initiation to ensure against undesirable consequences, and maintaining records of all clearances. As part of the circuit-protection function, the computer would maintain data on present relay settings and change relay settings as required.

Company and pool levels. The company and pool levels are responsible for operation of a complete bulk generation–transmission system and for coordinating energy and capacity interchanges with neighboring companies and pools.[17–26] At this level, economic considerations become significant. The following functions are recommended for performance at the company and pool level:

1. Data collection and logging.
2. Load forecasting.
3. Unit commitment.
4. Maintenance scheduling.
5. Evaluation of designated system conditions.
6. Load-frequency control, economic dispatch, and energy interchange.
7. System-voltage control.

A recent joint study by engineers of the Commonwealth Edison Company and the General Electric Company evaluated the role of computers and associated data collection and display devices in improving power system security.[7]

Data requiring most frequent updating are those associated with load-frequency control; namely, generator outputs, tie-line flows, and frequency. It is recommended that these be scanned at least every 2 seconds. System frequency at several locations should be included in this fast scan to minimize delay in detecting system separation and to enable immediate suspension of normal load-frequency control should separation occur.

The variables not used directly for control may be updated less frequently. An update interval of 20 seconds is short enough to enable tracking the system when in the normal state but is not so short as to place undue speed requirements on the data-collection system or to overload the computer with data. Changes in status, such as circuit-breaker operation during normal or emergency states, need to be sent to the computer only when changes occur or when specificially requested for purposes of initialization.

There are also a number of items that may be computed from the telemetered quantities. Some of these items, such as zone loads, line current, transformer kVA, and spinning reserve, require very little calculation and may be updated every 20 seconds, along with the bulk-data scan. These may then be treated like scanned quantities for alarming, logging, and similar functions. Other derived quantities, such as phase angle and reserve response, require more computation and might be updated every 5 or 10 minutes.

The study concluded that a system diagram board that portrays the major transmission circuits should be utilized, even though its size places a significant constraint on the physical arrangement of the operating center. It provides the dispatchers with an overall view of the physical network so they may readily visualize all the alternative paths for power. This plays a particularly important role during emergency conditions, but it is also important during normal conditions and periods of restoration as well. It is desirable that the board be animated to the extent that the status of the major lines is indicated automatically.

Conventional recording instruments are considered desirable for quantities such as plant generation, total

system generation, the loading of key tie lines, and system frequency. Alternatives to the use of recorders were explored, but none were found to have the same suitability for portraying a continuous and immediate record of past history so that dispatchers coming on shift and others have a ready means of becoming familiar with the current conditions on the system.

Consoles must be provided for control of display and communication equipment as well as for entering data associated with generation–transmission system.

The application of cathode-ray-tube devices for displaying information of interest to the dispatcher is sound and technically feasible. These devices are quiet and can quickly display a large amount of pertinent data on request. A need was seen for this type of display for detailed substation information, generation information, interchange data, and one for public or management viewing. They are useful for displaying one-line diagrams and the associated telemetered and computed data.

The functions to be performed at the company and pool levels involve a significant amount of computation.[27–31] Load-flow, stability, and short-circuit programs used by pool centers might typically be sized to enable studies of 1000-bus networks to be made. Computer requirements are, therefore, greater than at the division level and it is not unusual to find large-size scientific computers being suggested. Backup is recommended for all major computer functions through the use of dual processors.

Various arrangements may be considered with respect to the load-frequency control and economic dispatch function between the company and pool center.[9] A configuration such as shown in Fig. 5 is recommended from the viewpoint of flexibility and reliability of operation. At the pool level, company area-control errors are formed that recognize both economics and regulation requirements. In the event the pool computer is out of service, the company computer can form an area control error in the usual manner.

Regional coordination centers. The regional centers are less involved in hour-by-hour operations and more concerned with operating plans for the weeks and months ahead. As such, the functions performed do not require the use of large amounts of on-line data pertaining to the current status of the system. Recommended functions are as follows:

1. Role as communication center.
2. Daily and weekly generation and load forecasts.
3. Reserve analysis.
4. Coordination between planning and operation.

Computer requirements are mainly for the analysis of large networks; thus a large-scale scientific computer is appropriate. Backup is not required because the loss of computation capability for the time required to repair the computer is not serious.

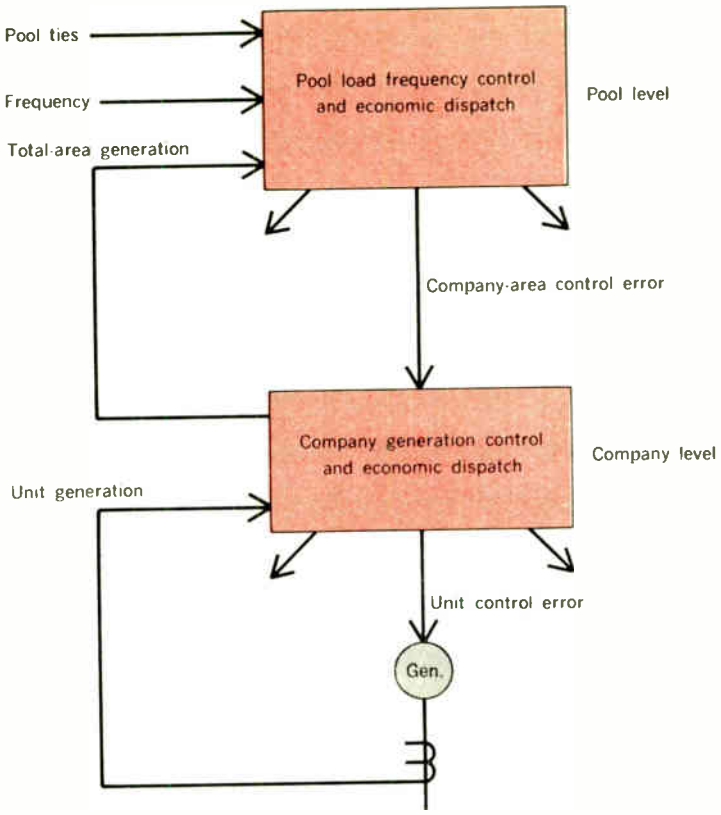


FIGURE 5. Relationship between company and pool computer centers.

Conclusions

The rapid developments in automation technology and the continuing industry emphasis on reliability of operation have led to widespread interest in utilizing computer technology to its best advantage for the improvement of system security as well as economy of operation. Industry trends toward large units, extra-high voltage, and increased interconnections increase interdependence. Moreover, the increasing number of system elements and variables adds to the complexity of the problems facing the system operator.

Examination of the structure of the system operation functions and problems has led to a multilevel and multicomputer approach involving computers at various action centers interconnected by data links. These centers include the power plants, substations, transmission division offices, company operating centers, and pool and regional operating centers. This multilevel configuration leads to improved reliability and speed, and reduces the overall data communication requirements as well.

Within the hierarchical configuration, a four-step approach of monitor and display, contingency evaluation, corrective strategy, and automatic control is suggested. Initial industry emphasis has been placed upon the first two steps, capitalizing primarily on improvements in data collection and communication and display techniques as well as digital process computer technology. Future efforts will be directed toward further improvements in these first two steps as well as significant new developments in corrective strategy and automatic control.

REFERENCES.

1. Concordia, C., "Design of electric power systems for maximum service reliability," presented at the CIGRE International Conference on Large High-Tension Electric Systems, Paris, Aug. 24–Sept. 2, 1970.
2. Bleiweis, J., "Northeast Power Coordinating Council," presented at the IEEE/ASME Joint Power Generation Conference,

Charlotte, N.C., Sept. 21–25, 1969.

3. Milner, B. H., Nagel, T. J., and Lentz, O. A.: "Coordination is ECAR's road to system reliability," *Elec. World*, pp. 64–66, Aug. 16, 1968.

4. Federal Power Commission, "Prevention of power failures. I—Report of the Commission (July 1967). II—Advisory Committee report: Reliability of electric bulk power supply (June 1967). III—Studies of the Task Groups on the Northeast Power Interruption (June 1967)," U.S. Government Printing Office, Washington, D.C.

5. Feidler, H. J., and Kirchmayer, L. K., "Automation developments in the control of interconnected electric utility systems," presented at the IFAC/IFIP Symposium on Digital Control of Large Industrial Systems, Toronto, June 17–19, 1968.

6. Fiedler, H. J., Robinson, P. B., and Tandy, G. E., "Application of a remote information and control system," *Proc. PICA Conf.*, pp. 107–116, 1969.

7. Cihlar, T. C., Wear, C. H., Ewart, D. N., and Kirchmayer, L. K., "Electric utility system security," presented at the American Power Conference, Chicago, Apr. 22–24, 1969.

8. Dy Liacco, T. E., "The adaptive reliability control system," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-86, pp. 517–531, May 1967.

9. Happ, H. H., "Multicomputer configurations and diakoptics: Dispatch of real power in power pools," *PICA Conf. Record*, pp. 95–108, 1967.

10. Happ, H. H., and Undrill, J. M., "Multicomputer configurations and diakoptics: Linear power flow in power pools," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-88, pp. 789–796, June 1969.

11. May, W. O., and Krings, F. C., "Control computer system's success prompts repetition," *Elec. World*, pp. 22–24, Jan. 27, 1969.

12. Marks, J. A., "Regional control key for EHV expansion (special report)," *Elec. Light Power*, pp. 75–82, Mar. 1969.

13. Benkusky, A. W., "MAPP's coordination center," *Power Eng.*, pp. 30–33, June 1968.

14. Marks, J. A., "Fully automated supervisory control: inevitable?" *Elec. Light Power*, pp. 69–75, June 1969.

15. Dy Liacco, T. C., and Kraynak, T. J., "Processing by logic programming of circuit-breaker and protective-relaying information," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-88, pp. 171–175, Feb. 1969.

16. Coulter, J. C., and Russell, J. C., "Application of computers in EHV substations," Pacific Coast Electric Association, Los Angeles, Mar. 9, 1967.

17. Oprea, G. W., Jr., "Control center computerizes energy dispatch operation," *Elec. World*, pp. 73–76, Aug. 26, 1968.

18. Ku, W. S., and Van Olinda, P., "Security and voltage applications of the public service dispatch computer," *Proc. PICA Conf.*, pp. 201–207, 1969.

19. Limmer, H., "Techniques and applications of security calculations applied to dispatching computers," presented at the Third Power Systems Computation Conference, Rome, June 23–27, 1969.

20. Sweeny, S. J., "Master-satellite dispatch activated by REMVEC," *Elec. World*, pp. 81–84, June 14, 1969.

21. Pence, W. K., "New Michigan pool center stresses system security," *Elec. World*, pp. 71–73, Oct. 21, 1968.

22. Vogel, J. R., Jr., "A power control center for the New York Power Pool," presented at the IEEE/ASME Joint Power Generation Conference, Charlotte, N.C., Sept. 21–25, 1969.

23. Mochon, H. H., Jr., "New England Power Exchange control center," presented at the IEEE/ASME Joint Power Conference, Charlotte, N.C., Sept. 21–25, 1969.

24. Stewart, H. G., "Computer moves onstage as part of system operation," *Elec. World*, pp. 92–94, July 14, 1969.

25. Knight, U. G., "The use of computers in system operation and control of Central Electricity Generating Board's system," presented at the Third Power Systems Computation Conference, Rome, June 23–27, 1969.

26. Siroux, J., "The national dispatching of Electricité de France," *Rev. Gen. Elec.*, vol. 76, ppl 509–522, Mar./Apr. 1967.

27. Brewer, G. L., and Tam, C., "The computational aspects of the CEGB new national control center," presented at the Third Power Systems Computation Conference, Rome, June 23–27, 1969.

28. Happ, H. H., Johnson, R. C., Newman, G. A., and Wright, W. J., "Large-scale hydrothermal unit commitment," *Proc. PICA Conf.*, pp. 273–294, 1969.

29. Denison, O. J., Hayward, D., and Undrill, J. M., "Dispatchers' load flow for the REMVEC Dispatch Center," presented at the IEEE Summer Power Meeting, Dallas, Tex., June 22–27, 1969.

30. Wollenberg, B. F., "Bulk power load flow is system operator-oriented," *Elec. World*, pp. 79–80, July 14, 1969.

31. Stagg, G. W., Dopazo, J. R., Elitin, O. A., and Van Slysk, L. S., "Techniques for the real-time monitoring of power system operations," presented at the Third Power Systems Computation Conference, Rome, June 23–27, 1969.

Reprints of this article (No. X71-073) are available to readers. Please use the order form on page 9 , which gives information and prices.

D. N. Ewart (M) received the B.E.E. degree from Cornell University in 1954 and the M.S.E. degree from Union College in 1963. From 1954 to 1956 he served in the U.S. Air Force at Wright Air Development Center. In 1956 he joined General Electric in Pittsfield, Mass., where he spent two years as a laboratory engineer at the High Voltage Laboratory and two years as an insulation development engineer in the Power Transformer Department. In 1960 he transferred to the Electric Utility Engineering Operation at Schenectady, where he is now manager of the System Control Subsection. During the past few years he has been concerned with the analysis of power system dynamics and control. He has developed dynamic models of excitation systems, prime movers, governors, and synchronous machines. He also has been actively involved with the application of process control computers for improving the economics and security of power system operation and served as a project engineer on a joint study of power system security with the Commonwealth Edison Company. Mr. Ewart has written a number of technical papers. He is a member of Eta Kappa Nu.

L. K. Kirchmayer (F), manager of General Electric's System Planning and Control Section, received the B.S.E.E. degree from Marquette University in 1945, and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin in 1947 and 1950 respectively. He is also a graduate of GE's Advanced Management and Modern Engineering Courses. After graduating from Marquette he joined Cutler-Hammer, Inc., as an experimental research engineer in control circuits and apparatus. He later taught electrical engineering at the University of Wisconsin. He joined GE's Analytical Engineering Section in 1948. He served as manager of Power Systems Operational Investigations from 1956 to 1958 and as manager of System Generation Analytical Engineering from 1958 to 1963. Dr. Kirchmayer has received several patents related to computer control of power systems and is the author of approximately 70 papers and two books, "Economic Operation of Power Systems" and "Economic Control of Interconnected Systems," both published by John Wiley. He is a Fellow of the American Society of Mechanical Engineers.

Impediments to societal problem solving

What must happen before we can succeed?

If we are to make orderly and effective progress in the domestic societal area, we must shift our focus from past successes to trying to determine how best to reorder our priorities for the future

Gabor Strasser *Office of Science and Technology*

Historically the national goals of the United States have depended to a large extent on its science and technology. Now the character of this dependence is changing in light of the increasing emphasis on societal problems. The nature of these problems is discussed and needed actions delineated. It is concluded that the most critical areas in which progress is necessary are in the formulation of objectives and translation of these objectives into implementable action items, in the orchestration of resources, and in institutional reforms. It is pointed out that the engineer has the choice of playing a critical leadership role in the decade ahead or of being left to perform routine technical tasks set for him by others.

During the 1950s and 1960s, many of the national priorities of the United States in such fields as defense, space exploration, and nuclear energy depended heavily—and often critically—on science and technology. Because of this extensive past dependence, it has been difficult to tell at times whether scientific/technological efforts have been supported for the sake of knowledge building—in conjunction with the training of scientific manpower at universities—or for the contribution that they were expected to make to some national priority. There was no clear line of demarcation among these rationales, nor was there need for one. The nation needed the scientific/technical inputs and the scientific/technical community was anxious to provide them.

It is safe to say, however, that the support for scientific/technical endeavors has primarily depended on the requirements of our national priorities. Had this not been the case, astronomy, for example, should have fared as well as electronics did. That it did not is attributable to the fact that what the United States considered important during the past two decades depended much more on electronics than on astronomy.

Revised text of a paper presented at the IEEE International Convention and Exposition, New York, N.Y., March 24, 1971.

We must realize that the bulk of the federal monies that underwrite scientific/technological efforts has gone, and probably will continue to be going, for the support of those efforts that, in turn, support the nation's currently perceived needs or priorities. However, we also must realize that the nature and extent of the dependence of national priorities upon science and technology are changing. This fact should be of paramount concern to the scientific/technical community. We should not attempt to preserve the *modus operandi* of the past but try to determine how to contribute constructively to the solution of the problems of the future. This thought takes on added significance in view of the following.

Recently in some quarters, the utility, and even the honesty and morality, of science and technology has come into question. This is anything but fair since our past progress as a nation has demonstrably and extensively depended on science and technology. If anything, our continuing dependence, even if in a different way, should increase with the many complex issues waiting to be addressed, and with the many difficult problems that will have to be solved.

The source of this criticism of science and technology is varied. Some of our national purposes, which science and technology have been supporting in the past, have now grown "unpopular," such as the "arms race." (Incidentally, what is often overlooked is that it takes at least two to have a race. Those who tend to come to simplistic conclusions in this regard may find the readily available, extensive literature on the subject enlightening.)

At other times, deleterious side effects of technology-dependent programs have been blamed on technology, such as various modes of environmental pollution. And, in other instances, the imbalance of our technological know-how has been criticized: "If we can go to the moon why can't we cure cancer or clean up the environment?"

The truth of the matter is that science/technology represents but one of the enabling (though often critical) mechanisms that allow us to achieve our aspirations. It has been, is, and will continue to be the sociopolitical system that decides what we should go after and how we

should use science and technology in the process. This has been the practice in the past two decades, and most assuredly will continue to be the practice in the decades to come.

Let us now consider the nature of our emerging national purposes and examine the roles technology can be expected to play in the future.

Some speculations

The decade of the 1970s is emerging as a period of profound change for U.S. society. We as a people seem to be caught up in an extensive reexamination of our values and a massive reordering of our priorities. The focus of our attention is shifting toward a host of domestic societal problems concerning social, environmental, and esthetic values as contrasted with material values, measured by the average per capita Gross National Product and its various derivatives. This shift may warrant a change in the primary focus of our attention from a standard-of-living measure or index (based primarily on material goods) to a more general measure, such as quality of life, encompassing a much broader set of considerations. The latter might include environmental quality, personal safety, educational opportunity, health care, and many others.

Incidentally, a goodly portion of our objectives under this quality-of-life concept are nonmaterialistic in nature. Nevertheless, to achieve them still will cost money and hence will depend on our economic and "materialistic" capabilities and capacities. Therefore, the rate at which we as a nation will be able to improve our quality of life will be inexorably tied to the rate of increase in our efficiency and productivity in providing goods and services. This extremely important fact all too often is overlooked by many of our quality-of-life proponents. (It may also be worth noting that the revenue-raising capacity of the federal government through taxation is not independent of the manner in which the tax dollar is spent. The federal budget represents almost a quarter of the GNP and hence has a great effect on the nation's economy, on which in turn tax revenues depend. Also, it is one thing to decide to shift priorities; it is another to bring these shifts about in a positive, nondisruptive manner.)

What would a quality-of-life focus mean to the engineer whose main forte is either the advancement of the state of the art of some sophisticated field or the solving of practical problems that depend primarily on technology and other, reasonably relatable disciplines?

Perhaps the most profound manifestation of this change, from the engineer's viewpoint, is the likelihood that even though the generation and application of new knowledge will continue to play important and often critical roles, the solution of most of our domestic societal problems will primarily depend on the proper orchestration of a host of disciplines and other resources. Much of what will need orchestrating will be "off-the-shelf" knowledge, procedures, or hardware, where the source may be as varied and often as incommensurable as hard physical and natural sciences, technologies, political and social sciences, economics, management and institutional arrangements, alternative sources of funds, government practices, and so on. Such orchestration will be difficult as well as critical to accomplish. The engineer, and especially the systems engineer, may be as well suited as any to attempt this orchestrating role.

I hasten to add that the last thing I want to do is to play down the importance of R&D. Quite the contrary! Research and development has been responsible for many highly effective, and initially often unimaginable, solutions and has provided a way to help keep our future options open at relatively modest cost. What I am talking about is a shift in the market for scientific/engineering talent, both in kind and in number. For example, it is unrealistic to expect that DOD and NASA (perhaps the biggest users of the scientific community in the 50s and 60s) will be able to support science and R&D at the ever-increasing rates to which we have become accustomed.

Let me turn now to the nature of societal problems—to what we must do to approach them constructively and to the impediments that we must overcome if we are to succeed.

The problem of choices

Our natural national wealth, however enviable, is still finite. Fortunately, our ingenuity to make the most of this wealth is not. However, our appetite can easily outpace any feasible menu that our wealth, combined with our ingenuity, can serve up. Hence more difficult choices, greater efficiency, increased productivity, and better management will have to be the way of life in the future.

Examples of "more difficult choices" are all too evident as we look at the myriad of domestic problems demanding attention. Most of these demands are justifiable, such as in housing, health, education, environment, etc. It is when we put them all together that it becomes clear that the sum of our individual aspirations, however defensible they may be on their own merits, far exceeds our available resources.

This problem, of course, is related to that of setting priorities, to the value judgments of the nation as a whole and of people as individuals, and to the questions: How do we set goals that enjoy the support of the majority? How do we balance societal benefits against individual affluence?

For example, the upgrading of our physical environment will compete for the same funds that people use to acquire personal possessions. The notion of "let the government clean up the environment" does not work, since the government must pass the cost on through taxes, and thus reduce disposable personal income. The notion of "let industry, which does the polluting, pay for it" does not work either, since the cost ultimately shows up in the price of consumer goods and services.

The citizen of the 70s will be faced with such questions as:

. . . the support for scientific/technical endeavors has primarily depended on the requirements of our national priorities.

The rate at which we will be able to improve our quality of life will be tied to the rate of increase in our efficiency . . .

1. I don't want to give up my car but at the same time I don't want to pollute the air. How much should I pay for emission-control devices?
2. How much sewerage tax assessment should I vote for (and thus forgo the purchase of a color television set or a vacation) to improve the quality of the river in my backyard?
3. Should I turn off the air conditioner in order to reduce the demand on our energy resources, as well as alleviate some of the concomitant pollution?
4. More generally, how should I trade off social/environmental/esthetic benefits for personal/economic/materialistic ones as an individual?

The systems approach

Systems engineers have encountered three major problems when attempting to confront societal issues in, say, urban or environmental areas. The first concerns their role in the decision-making process. The second concerns the "indicator problem," or how to measure success. The third problem is about inadequate long-range planning.

The decision-making process. Decision making in the sphere of societal problem solving, like any type of decision making, normally progresses through three phases: the objective phase—deciding what should be done; the approach phase—deciding how it should be done; and the implementation phase—taking the necessary action to get it done. Contemporary systems engineers who have been engaged in the solution of complex problems gained most of their experience in the approach phase of military and aerospace problems. Their claim to fame has been the development of a host of "cost-effective" approaches to problem solutions.

Urban decision makers need help most in the objective and implementation phases rather than in the approach phase. In other words, public action is most often stymied because no consensus can be developed relative to objectives and implementation measures. At a summer conference in 1969, the Deputy Mayor of New York City, Timothy Costello, described this problem somewhat as follows: "The trouble with systems engineers is that they arrive on the scene too late and leave it too soon. They fail to help decision makers with policy and objective formulation or with problems of implementation. The engineers tend to take it for granted that 'what should be done' is known, and once they (the engineers) figure out how to do it effectively, the program in fact could be implemented without much further ado."

The indicator problem. Although with military or aerospace problems most of the relevant variables are both qualifiable and quantifiable, in urban or environmental problems analysis is difficult because the most relevant variables are often not only hard to measure but also hard to define.

Many have recognized this fact, which has led to the ever-widening search for social, urban, or environmental indicators as output measures to serve in roles analogous to performance specifications for weapons or aerospace systems. Unfortunately, appropriate indicators with three essential qualities have eluded us so far. These qualities are relevance, operational utility, and reasonably wide acceptance.

Lack of long-range planning. Long-range plans, however tentative, provide an essential context in which the systems engineer solves complex, current problems. Such long-range plans by and large are lacking in the societal area. Why is this so?

First, the rewards and punishments of our political system, especially at the mayoral level, are such that they favor short-term "fire-fighting"-type actions. A mayor has to conceive, plan, and implement his program in two years, if he wants to be reelected. He does not, as a rule, get credit for trying to make his successor more successful. The lead times afforded within a two-year tenure are too short to address most of the societal problems confronting us today. In addition, public pressure is most often focused on problems that are current rather than long term, direct rather than indirect, tangible rather than intangible. Emphasis is on "fix it now," favoring a short-term expediency-type operation, taking problems one at a time, without due consideration for their interrelationships. This approach often proves counterproductive later on. Most of our elected representatives and leaders recognize the need for some comprehensive long-term approach, while being pressured for short-term results by their constituents. Many of the subjects of John F. Kennedy's *Profiles in Courage*[1] recognized this dichotomy. They took stands for essential but unpopular causes rather than go along with currently popular positions and quick-fix actions if, in their opinion, these positions and actions did not serve the best interest of the nation in the long run. Most of this decade's mayors have a different problem. They are forced to expend a large part of their energies in trying to prevent a further deterioration of the status quo, which, in most of our cities, is unacceptable to begin with.

Second, the chronic shortage of funds at the state and local government levels has hindered, and often precluded, any comprehensive planning.

Third, the complexities of our world today have made planning increasingly difficult. The pace of our lives has been accelerating. Our goals, expectations, and objectives have also increased, in number and complexity, as have the secondary and tertiary effects of our undertakings.

Effects of obsolete institutional frameworks

The lack of long-range planning is but one of the manifestations of a much broader problem, namely, institutional obsolescence.

Toffler, in *Future Shock*,[2] said that the rate at which things are changing all around us has increased to the point where we have difficulty adapting to it. Others say it is not the rate of change but our apparent inability to affect or control the change itself that is bothering us, giving us the sensation that we are no longer the masters of our fate. Whichever, if either, is the case, the fact remains that

How do we set goals that enjoy the support of the majority? How do we balance societal benefits against individual affluence?

1. Our technology and management techniques provide ever-increasing leverage, showing up in more severe consequences, shorter lead times, and greater impacts.

2. Our goals, as well as the approaches to the achievement of these goals, are more complex and nonhomogeneous, calling for interdisciplinary approaches.

3. The cause–effect relationships, as a result, are also more complex.

4. Our potential mistakes are bigger and more costly.

5. There is an increase in the irreversibility of many of the actions that we take.

6. There is less damping. Our environment is becoming less forgiving because of the use we have made of it.

These difficulties have combined to form what appears to be a gap between the old ways of doing things and some necessary new ways, which are yet to be identified. The fact that we have not bridged this gap has caused discouragement and, on occasion, a sense of futility.

At first, most new organizations are product- or objective-oriented. As time goes on the process by which the objective is sought becomes entrenched, and if management does not watch out, the process becomes an "end" unto itself rather than a "means."

As we as individuals and as a nation reexamine our values and reorder our priorities, it behooves us also to reexamine our institutions to ascertain that they help rather than hinder us in the pursuit of our objectives. If they don't help, we should take the necessary steps to change them. All too often, progress is stifled because many of our institutions, established to cater to past needs, cannot respond to the challenges of today and of tomorrow. Obsolete building codes, counterproductive union practices, the plight of the problem-solving, interdisciplinary-oriented professor at most universities, obsolete counterproductive laws on the books are but a few examples.

Recently President Nixon was quoted as saying "...good people cannot do good things with bad mechanisms...bad mechanisms can frustrate even the noblest aims...public officials cannot be patient with outmoded forms when the people have grown so impatient with government..." Unfortunately, we have a tendency to "end-run" the system. That is, when we really want to get something done and lack faith in the current mechanisms we set up additional new ones to do the job, rather than ascertain what is wrong with the existing system, and then fix it. This type of institutional proliferation results in the ever-decreasing effectiveness of many of our organizations that have been end-run, but nevertheless continue to exist. To some measure, such organizations become overhead items. As a consequence, even though some recently created mechanisms may be superbly efficient (until they themselves may get end-run), the overall result is that the efficiency and productivity of our system progressively decline.

The need for an orderly upgrading of our institutional structures to enable them to respond to what we as a nation want is all too evident. President Nixon, in his 1970 State of the Union message, said "...new knowledge and hard experience argue persuasively that both our programs and institutions in America need to be reformed...[we should develop] better ways of managing what we have..."

Since then, several communications have been sent to the Congress to bring about this streamlining, to have departments and agencies become more product- or objective-oriented. As a result, we now have the Council on Environmental Quality, the Environmental Protection Agency, and the National Oceanographic and Atmospheric Administration.

In his 1971 State of the Union message, the President said, in part: "The sixth great goal is a complete reform of the Federal Government itself. Based on a long and intensive study with the aid of the best advice obtainable, I have concluded that a sweeping reorganization of the Executive Branch is needed if the Government is to keep up with the times and with the needs of the people. I propose, therefore, that we reduce the present 12 Cabinet Departments to eight. I propose that the Departments of State, Treasury, Defense, and Justice remain, but that all the other Departments be consolidated into four: Human Resources, Community Development, Natural Resources, and Economic Development."

These new Departments would be directly relatable to primary national concerns, giving added impetus to still more product or objective, as opposed to process, orientation.

Needed actions

To come to grips with our societal problems in a systematic and effective manner, we first must be able to do a number of things that we cannot do adequately today. We should be able

1. To do a better job, to articulate qualitatively and quantitatively (wherever possible) the nature of our societal problems, as well as of our resulting goals, and to translate these goals into implementable action items.

2. To provide viable alternatives based not only on dollar but on social, political, institutional, and other "costs."

3. To relate sets of value judgments to different alternatives so that we will better understand the reasons for our decisions.

4. To trade off qualitative versus quantitative, deferred versus immediate, and indirect versus direct benefits and consequences.

5. To modernize those institutional frameworks that today preclude success, irrespective of what else we might do.

6. To recognize that what we want may be more than what we can have, and therefore the more efficiently we utilize our resources, the more we can reduce this unpleasant disparity.

7. To set some priorities, especially in light of the last statement.

Here then are some of the things we should be able to do but cannot. Why is this so? There are a number of reasons.

The impediments

The reasons for our inability to act, the impediments, include the following:

1. As mentioned before, the solutions to most of our current problems depend on a host of disciplines, such as sociology, politics, economics, and institutional arrangements, in addition to technology. Superposition, that is, separately and sequentially looking at, say, the economic considerations, then the political ones, then the social aspects, etc., does not work. They must be integrated. Furthermore, this integration must take place throughout the continuum of objective setting, approach formulation, and implementation, forming the basis for relevant cause-and-effect relationships. That is, we must establish the necessary parameters, in terms of appropriate inputs from all of the relevant disciplines, to which end-objectives are sensitive. *We do not know how to establish such parameters today.*

2. The output of a given program can manifest itself in a host of widely differing noncommensurable impacts—for example, cleaner air and better education, housing, health, transportation, etc. We do not know how to define, let alone trade off, various benefits against one another. We need better social, urban, and environmental indexes. *We do not have such indexes.*

3. Even if we had such indexes in a qualitative sense, we would still have trouble using them unless we could quantify them, or at least order them according to some criterion. This situation is akin to the need for end-objective rather than process-oriented standards. Hence we need to develop more standards that are broad in scope as well as performance- rather than design-oriented. *We do not have enough such standards today.*

4. We need to include in our current plans more rational and realistic ways to treat discounted future environmental costs and benefits when these are interpreted in the broadest sense. *We cannot do this now.*

5. We must consider in our plans potential good and bad side effects or impacts on other programs. Although probabilistic in nature, these considerations could be critical. Hence our deliberations and proposed alternatives heavily depend on notions of uncertainty, risks, expected values, etc. *We don't quite know how to popularize these notions,* as we must, so as to include them in the deliberations of decision makers and make them understandable to their constituencies.

6. We must reorganize our existing institutions to the extent necessary to enable them to be more responsive to the achievement of the goals and aspirations that we as a nation, through our government, have set for ourselves. Today we have trouble articulating many of these aspirations in an implementable fashion. Even where we can do so, and *even where we do see the called-for change, institutional inertia all too often makes the change impossible.*

Constructive activities

I don't want to leave the impression that all we have is problems and that nothing constructive is being done about them. A number of positive actions are being taken at various levels.

Universities. Many universities have recognized the need for new approaches. They have set up numerous urban, environmental, and other types of problem-solving institutes and groups on their campuses. However, such groups often continue to be multidisciplinary rather than interdisciplinary in their nature; the distinction is that the former represent a lot of different talent under one roof whereas the latter have learned to work together in an integrated fashion toward the solution of relevant, complex problems. In interdisciplinary groups, people use their respective talents in concert, primarily as means rather than as ends in themselves. Advancing the state of the art of any given specialty in an interdisciplinary group is not a primary objective, although if it should happen in the process, so much the better.

I offer the following suggestions for those universities sincerely interested in establishing interdisciplinary activities.

1. *Department status.* Give departmental status to interdisciplinary problem-solving groups.

2. *Appointments with tenure.* Offer appointments with tenure in such departments, giving them the necessary independence, power, and prestige within the university community. Institutes, centers, and *ad hoc* arrangements don't seem to work too well in most cases.

3. *Consultation with other departments.* Encourage the interdisciplinary department to rely on other departments (for expertise in the various specialties) on a part-time or consulting basis to supplement its own capabilities. Incidentally, this would provide a two-way street since, in return, the other departments would receive an additional outlet for their talents.

4. *Industry and government liaison.* Have the interdisciplinary department set up close working relationships with industry and various governments in order to expose itself to existing real problems.

5. *Think-tank consulting firm liaison.* Have the interdisciplinary department also establish close working relationships with think-tank and management consulting firms to share ideas on the techniques of problem solving.

6. *University administration support.* Support the interdisciplinary department intellectually and financially.

7. *Problem-solving orientation.* Make sure that the interdisciplinary department does not allow any one discipline to emerge as the "leader" (not even engineering) with the other disciplines relegated to subservient roles.

Government. There are also many places within the government where progress is being made; for example:

1. The National Science Foundation has a brand-new program called Research Applied to National Needs (RANN), with a problem-solving orientation. Its charter includes, if only implicitly, most of the activities that are listed in this article as prerequisites for needed progress.

2. The Office of Science and Technology has been carrying out a methodology study of technology assessment, defined as "a systematic planning and forecasting

Public pressure is most often focused on problems that are current rather than long term . . . tangible rather than intangible.

process that delineates options and costs, encompassing economic as well as environmental and social considerations, that are both external and internal, with special focus on technology-related 'bad' as well as 'good' effects."

The objective of the study is to delineate where existing analytical or institutional processes can be applied or could be made to apply if modified, to ascertain where new processes must be developed, and actually to develop such new processes to the extent that time and funds permit. It is being carried out in conjunction with five pilot case studies: Automotive Emission Control; Certain Aspects of the Impact of Computers on Society; Introduction of Industrial Enzymes; Ocean Farming; and Pollution Abatement with Emphasis on Recycling.

Other groups. There are, of course, other groups actively working on these problems as well. For example, the National Academy of Engineering last summer sponsored a Symposium on Social Directions for Technology, and recently had another—Benefit-Risk Relationships for Decision Making—with special focus on the societal area. Another example is a meeting on "Systems Problems of the 70s" planned for this fall under the joint sponsorship of the IEEE and the Operations Research Society of America, at which the writer will chair a session entitled "Orchestrating Technology with Other Means to Solve Societal Problems."

The problem thus is not lack of interest, or even of activity. It is one of integration, for their mutual benefit, of the many ongoing, varied, yet related actions so that ultimately they may be focused on their common objective—to cope with societal problems.

Conclusions

Context. This article focused on societal domestic problems and on the future role of the engineer in society. Since many other problems compete for our attention today, it was felt that at least two particular comments need to be made to put in proper context the foregoing observations and recommendations. (1) Arguments about national defense vs. progress in the domestic societal area per se are irrelevant. We need both. Without adequate security we could not proceed on the domestic front, and given that we have security, we would not be fully exploiting its benefits if we did not simultaneously make needed progress in the domestic societal area. The relevant questions concern balance and appropriate means. (2) The current plight of unemployed scientists and engineers is of the utmost importance not only to those who are personally affected but also to the profession as a whole. The fact that this article addressed the long-range focus of our profession in no way detracts from the importance of the current, hopefully temporary, problem of unemployed engineers and scientists.

What must be done. If we are to make needed, orderly, and effective progress in the domestic societal area, we must direct our attention, on a continuing basis, to these questions: What are the things that we want to do, and can afford to do? As a result, what important things will remain undone? How do we orchestrate our many and varied resources to achieve most effectively our selected objectives? How do we bring about clearly needed institutional changes to enhance the achievement of these objectives?

In other words, unless we can do a better job of deciding where we want to go and how we are going to get there, and, at the same time, can create the kind of organizations that can help rather than hinder us in our efforts, we are not going to get very far!

What the engineer can do. The decade of the 1970s will be a period of change and challenge as well as opportunity. The engineer can become a leader in this exciting new venture, provided he is willing to: (1) consciously and diligently seek a better understanding of the society that he must serve, and within which he must function; (2) intentionally and credibly join forces with representatives of other disciplines with whom he will have to work; and (3) "advertise" and "market" his willingness to undergo this renaissance in an aggressive, yet nonobjectionable manner.

Should he be unwilling to do these things, he will be left behind to perform isolated, often mundane, technical tasks that others will have prescribed for him.

The choice is up to him!

REFERENCES

1. Kennedy, J. F., *Profiles in Courage*. New York: Harper & Row, 1964.
2. Toffler, A., *Future Shock*. New York: Random House, 1969.

Gabor Strasser has served in the Executive Office of the President, Office of Science and Technology, since 1969, assisting the Science Advisor with development of policy, technology assessment, criterions for research and development, and program review and planning. He also serves as Executive Secretary of the President's Science Advisory Committee Panel on Science and Technology Policy.

Mr. Strasser received the B.C.E. degree from the City College of New York in 1954 and the M.S. degree in mechanical engineering from the University of Buffalo in 1959, and is an alumnus of the Harvard Graduate School of Business Administration Program for Management Development. He spent the period between 1956 and 1962 in the aerospace industry, first with Bell and then with Boeing, starting as a research engineer and later becoming project engineer. From 1962 to 1968 he filled various positions with the Mitre Corporation in Bedford, Mass., and Washington, D.C. During this period he was head of the company's National Military Command Systems Design Department and later headed the Systems Analysis Department. Prior to his present assignment, he was vice president for planning of the Urban Institute in Washington, D.C.

Mr. Strasser is a registered professional engineer in the states of New York, Washington, and Massachusetts, and is a member of the American Institute of Aeronautics and Astronautics. He is the author of a number of papers reflecting both his initial work in aerospace and his current interest in social, urban, and environmental problems.

Test signals for music reproduction systems

Despite a search that has ranged over the most sophisticated fields in modern engineering, attempts to replace the century-old theories of Lord Rayleigh have repeatedly failed

J. Robert Ashley, Thomas A. Saponas, Randolph C. Matson
University of Colorado, Colorado Springs

Experiments with both long, continuous tones and short, transient tones have shown that the human ear is insensitive to relative phase differences between fundamentals and overtones. Short notes from a piano and from a drum are studied here to show that the phase requirement for music reproduction is that the group velocity of the system be constant in the relatively narrow regions surrounding the fundamentals and overtones. The phase velocity across the audio spectrum does not have to be constant, thus easing the requirements on loudspeakers, crossover networks, and tape recorders. Pulse-testing schemes require minimum-phase behavior of a system for simple interpretation and therefore tend to overtest a music reproduction system. Random noise can be used as a test signal if elaborate processing equipment is available. However, the sinusoid is still the best test signal for determining distortion and relating device performance to theory.

The sine wave has been used as an audio test signal from the beginning of the electronic era. Undoubtedly, the elegant work of Lord Rayleigh[1] in the mathematical physics of sound was the most significant single factor in determining this choice. As documented in the two volumes of Ref. 1, Rayleigh either found or extended solutions by Helmholtz and others for propagation of sound through air, generation of sound by strings, organ pipes, membranes, etc. With amazing regularity, the eigenfunctions for the partial differential equations turned out to be sines and cosines. In common with electric circuit theory, Lord Rayleigh found that many of the differential equations could be solved only with sinusoidal excitation. Finally, he fully appreciated and promoted the ideas of Fourier in representing complex periodic functions as the sum of a set of sinusoids—the universally well-known Fourier series.

Thus, when early acoustic engineers had need for a test signal, they chose the signal that had the most to offer in confirming their designs with existing mathematical theory. This was the sine wave and we note that generation with "beat-frequency oscillators" was not especially convenient, nor were measurement techniques without the oscilloscopes we now take for granted. At the same time, and for analogous mathematical reasons, electric circuit theory and transmission line theory were emphasizing the use of sinusoids, and this meant that an engineer developing an electroacoustic device found the sine wave to be the best-understood test signal for both the electrical and the acoustical portions of his device. With this strong background in the history of audio, the sinusoid has definitely emerged the standard of comparison.

Why, then, should we be looking for other test signals for music systems? For one thing, the relatively wide audible spectrum requires broadband devices and the taking of many points of data to determine the device

Adapted, with permission, from the original version published in the *Journal of the Audio Engineering Society*, vol. 19, pp. 294–305, Apr. 1971.

performance. This has been partially automated with swept-frequency signal generators, good microphones and test meters, and automatic plotting equipment. This automation is expensive, however, even though the basic test signal is simple. If some other test signal could give the same information as a swept-sinusoid setup, there might be significant reduction of test-equipment cost. As an example of what might be hoped for, the testing of oscilloscope rise time with a square-wave test signal is a less expensive method of evaluation than taking the point-by-point high-frequency response. There is also a nagging suspicion that the sinusoid does not evaluate the transient response of a music reproduction system.

Finally, the testing of loudspeakers is greatly confused by room acoustics. Construction of anechoic chambers is both expensive and of questionable significance when the loudspeaker is to be used in a fairly reverberant room. If some test signal could remove room acoustics from the evaluation of loudspeakers, this would greatly reduce the complexity of testing and some of the witchcraft of loudspeaker design and evaluation.

Without a doubt, we have adequate motivation to seek a test signal other than the sinusoid. Moreover, there have been great advances in system theory, information theory, communication theory, and computational machines since Lord Rayleigh laid the foundations for audio engineering. It seems worthwhile, therefore, to search these newer theories for a better test signal.

Requirements for a test signal

Before proceeding with the search for a better test signal, we must carefully define the object of our investigation. First and foremost, we want to evaluate music reproduction systems; therefore, our test signal must be a fairly accurate simulation of music signals. The test signal should subject the system to the same kind of excitation as music signals but not grossly greater. The test signal must also be easy to generate. Furthermore, we will consider the level of cost allowable for the generator to be about equal to the cost of a swept-sinusoidal signal generator for the audio band. In addition, the results of applying such a test signal to an audio system must be easy to measure and interpret.

Finally, as a most important requirement, we think that the test and its interpretation must be free from booby traps. As a concrete example of such a booby trap, consider testing a complete audio system using a two-way loudspeaker with a square wave. Let us assume this is an unusually good system and that the loudspeaker designer has allowed proper overlap between woofer and tweeter, and has been careful regarding mounting the loudspeaker driver units in a sufficiently large cabinet, phasing the system, and the design and sine-wave testing of a third-order Butterworth crossover network.

For the sake of argument, we will further assume that the driver units in this loudspeaker are blissfully ignorant of such trivia as standing waves along the cone, falloff of radiation at high frequencies, etc. Visualize with us the care required in carefully baffling this loudspeaker and arranging it to radiate vertically from the roof of the laboratory. The loudspeaker is connected to the dc-to-100-kHz solid-state amplifier and the frequency of the square generator carefully set at 100 Hz to test the system with a 30-Hz lower corner frequency. Can you imagine the sick feeling of this engineer when his "kilobuck" capacitor microphone and oscilloscope indicate that the signal in Fig. 1 is coming out of this great loudspeaker?

Those who have read the paper by Ashley and Henne[2] may realize that this weird square-wave response is

FIGURE 1. Output waveform from a hypothetical two-way loudspeaker using third-order Butterworth filters in a crossover network.

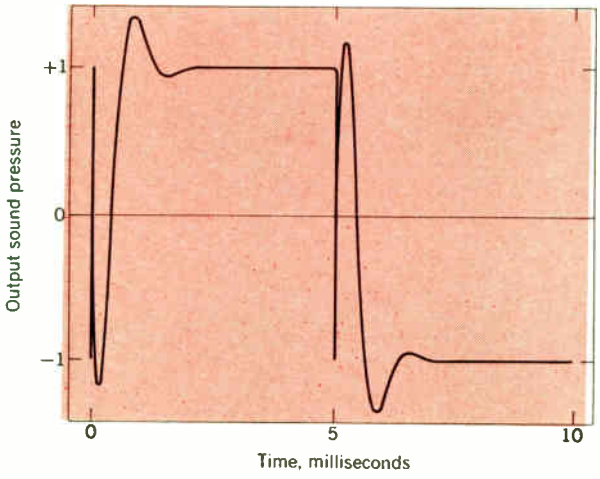
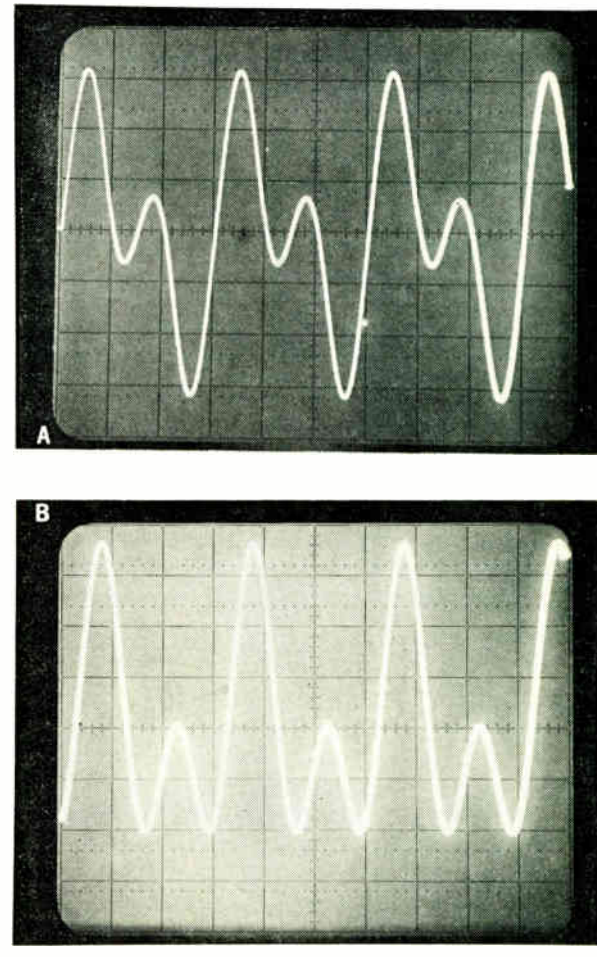


FIGURE 2. A fundamental and second harmonic with two different phase shifts. A—$f(t) = \sin(4\pi f_0 t)$. B—$f(t) = \sin(2\pi f_0 t) + \sin(4\pi f_0 t - 90°)$.



actually caused by the third-order Butterworth filters used as a crossover network. Before you think such a booby trap would not catch you, consider that the senior author of that paper nearly "wrote off" third-order Butterworth filters on the basis of an analog computer simulation that obeyed all of the assumptions just spelled out. In spite of an education in both academia and the school of hard knocks, this professor actually had to hear that the weird transient response made no difference when a music signal went through a loudspeaker system before he would help his coauthor discover that the cause was a gradual phase shift rather than a "hole" in the frequency response.

This same square wave illustrates another feature of audio systems that must be kept in mind when avoiding booby traps. Again for the sake of argument, assume that this same loudspeaker system had the constant-voltage crossover network described by Small[3] (see also Ref. 4), which does not have this defect in the transient response. Consider too that at least the tweeter knows all the facts of life that cause peaks and valleys in the frequency response. Could we then use the leading edge of a square wave to estimate the high-frequency response, as is done in video amplifier testing?

The answer is no! Buried in the theory that makes this work for the video amplifier are two little theoretical gems of knowledge: (1) the amplifier must be a minimum-phase system, and (2) the high-frequency response must be nearly Gaussian in shape before it can be inferred from a measurement of rise time.[5] Our loudspeaker with its peaks and valleys in the frequency response is not a minimum-phase device nor is the frequency response anything close to Gaussian.

Regretfully, in this discourse on booby traps we have demolished the usefulness of a square wave as a test signal for complete audio systems (only). A final warning about booby traps—we had originally thought the square wave to be one of the most promising candidates.

Time description of music signals

Since we have carefully restricted this study to music reproduction systems and stated that a test signal must be a fair simulation of a music signal, we must delve into the mathematical description of music signals.

There is no question, practically or mathematically, that a long, continuous, single tone from most musical instruments can be considered periodic, nearly infinite in duration, and therefore completely describable by a Fourier series. The question that arises here is whether the phase of the harmonic terms in the series is or is not important. In Fig. 2, we have shown the result of mixing a fundamental and second harmonic with differing phase. We set up this experiment in our laboratory using a trigger phase lock to maintain frequency synchronization between the two sine waves. This trigger phase lock could be adjusted to vary phase between fundamental and harmonic without losing frequency synchronization. The mixed signal was amplified and supplied to a single-driver loudspeaker adequate for the frequencies being used. We "tested" every willing student who ventured near our laboratory and found none who could detect the changes in phase angle between these two signals.

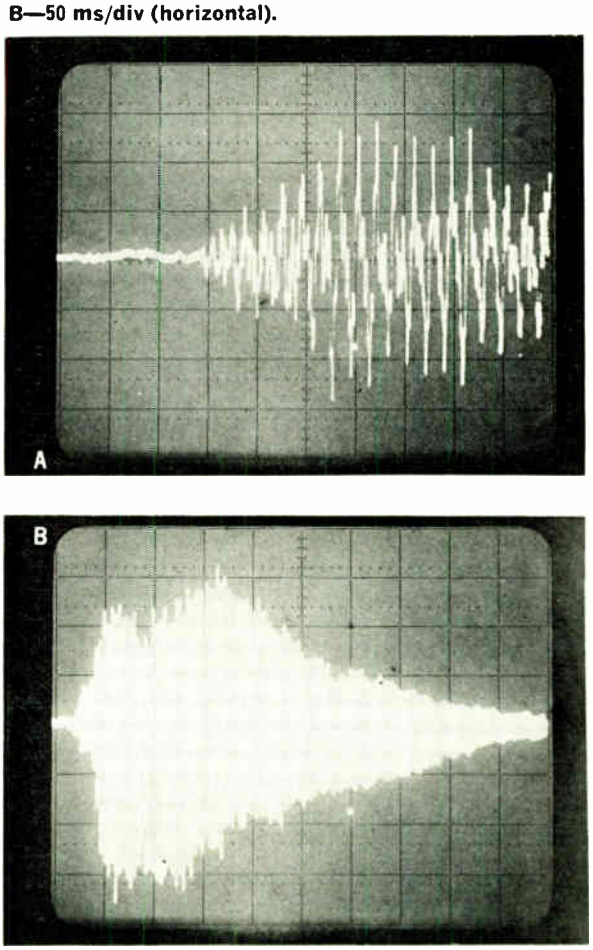
This is not a surprising result when we recall that neither pipe nor electronic organ builders really worry about phase synchronization of harmonic tones with fundamental tones. It is only the magnitude of the harmonics that determines the musical timbre of a sound. Thus, we do not have to worry about phase shift between widely spaced regions of the audio spectrum. This is a benefit in generating the signal but a potential booby trap in understanding the test results.

A continuous-wave tone from a single instrument is not the one causing present concern regarding the suitability of music reproductive systems; the nagging suspicions are about transient tones such as staccato notes on a piano, pizzicato tones from stringed instruments, drum beats, and cymbal clashes.

First, we observe that it is sufficient to consider these tones one at a time. In musical passages, pizzicato or staccato notes frequently occur in runs up or down the scale, and the performer must move his fingers from one position to another on the instrument. During the finger transit time, the preceding tone has time to die out (or be damped in the piano) before the next tone builds up. The total of several tones from several instruments playing simultaneously must be added according to the laws of probability theory. This requires careful description of each tone but nothing about the absence or presence of phase or frequency synchronization of the individual notes.

Second, the consequence of describing only one note is that we must use Fourier-integral transform theory as

FIGURE 3. Sound pressure waveform for staccato middle-C note on a grand piano. A—10 ms/div (horizontal). B—50 ms/div (horizontal).



the tie-in between the time and frequency domains. We are not dealing with harmonic functions; therefore, we must abandon the Fourier series as our mathematical mainstay.

Third, the staccato notes are actually the result of starting and stopping an oscillating mechanism. The best mathematical model for this situation is the pulse amplitude (mostly) modulation of a carrier. In Fig. 3, we see a storage oscilloscope trace that resulted from the most rapid possible (after a bit of practice) single stroke on a piano key. Notice that the signal builds up over a sizable number of carrier cycles and damps out over an even longer number of cycles. The interesting thing is that similar tests for notes up and down the piano keyboard indicated that the buildup and decay at the low end of the keyboard takes a longer time, whereas the buildup and decay at the high end takes a shorter time. The somewhat flat region across the top of the "pulse" is invariant with keyboard position.

All this makes sense when one thinks about the piano mechanism. The string is a rather high-Q mechanical resonator that is excited by the stroke of a padded hammer. The energy imparted by the felt hammer cannot cause an instantaneous buildup of the string vibration any more than the string vibration causes an instantaneous buildup of a vibration in the sounding board. When you think of the length of the strings, the "tapered" shape of the sounding board, and the amount of energy that must be transferred, it becomes quite reasonable to expert the buildup to require a constant number of cycles of string oscillation rather than a constant time.

The same kind of reasoning explains why the decay portion of piano tone also requires a constant number of cycles. The start of the damping interval occurs when the felt damping hammer first touches the string. The damping hammer cannot instantaneously stop the vibration of the string any more than the instantaneous cessation of string vibration could cause an immediate stop of soundboard vibration.

The constant time interval from the end of the buildup to the beginning of the decay is explained by the construction of the linkage between keyboard and hammers. To make the "touch" even over the keyboard, the linkage mechanism and hammer inertia are the same for all the keys. Thus, there is an invariant minimum time between the application of the striking hammer and the damper.

The mathematical description of the single staccato piano note is

$$f(t) = A(t)\cos \omega t + a_2 A(t)\cos(2\omega t + \phi_2) + a_3 A(t)\cos(3\omega t + \phi_3) + \cdots \quad (1)$$

where the $a_2, a_3, \cdots, a_m$ are a converging set of weighting factors; that is, the Fourier series coefficients of the musical tone that is being modulated by $A(t)$.

Finally, observe that Fig. 3 looks nothing at all like the usual tone burst of a few cycles that has been used as a test signal. Even without a formal mathematical proof, one should suspect that a short (less than 20 cycles) tone burst is a poor simulation of a musical note.

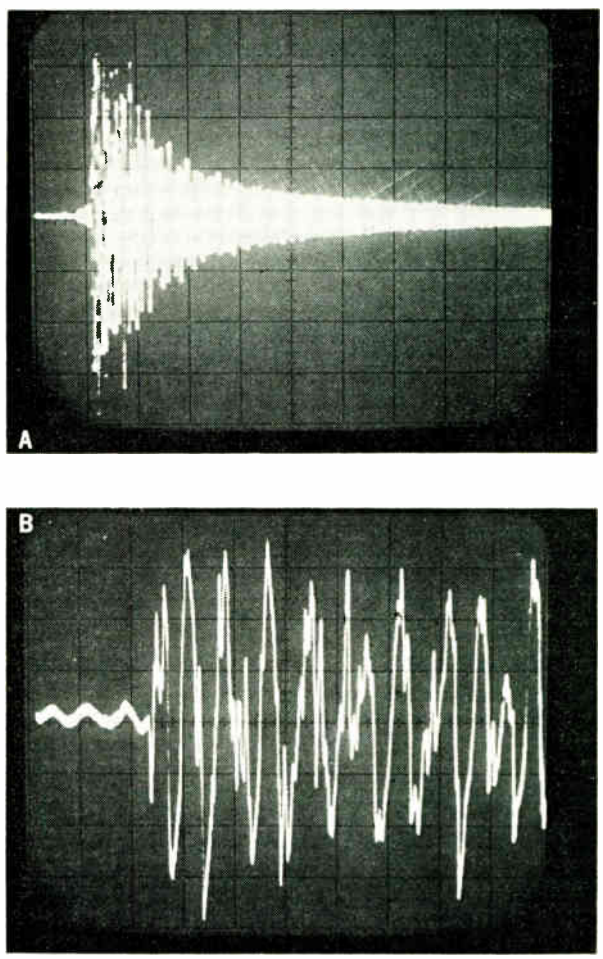
We also tested the pizzicato notes from a violoncello played by a competent musician. In going from the low to the high end of the range, we found the same behavior of rise and decay being described by a constant number of cycles of the vibration. The buildup of violoncello tones is more gradual than the piano notes, and we suppose this is because the body of the instrument has higher-Q resonances than the sounding board of a piano.

String, woodwind, and brass instruments would be expected to show similar behavior because all use some kind of acoustic amplifier (columns of air in woodwind and brass), which has appreciable energy storage capacity as well as high-Q resonant behavior. Thus, the model for notes from nonpercussion instruments is a complex periodic carrier amplitude modulated by an "envelope" function that is characterized by a relatively slow (in terms of cycles of the carrier function) buildup and decay. The Fourier-integral transform is the mathematical tool par excellence for the study of these notes.

From a mathematical standpoint, percussion instruments are the most challenging, and their reputation for "tearing up" audio systems is equal to the mathematical challenge. Yet, going back to fundamentals shows that the mathematical model must be based on modulation of a "carrier" by a pulse-envelope function.

As a specific example of a percussion instrument, consider a drum. A drum has a stretched membrane over a sealed volume of air that is a resonant chamber. Lord Rayleigh again set the stage for this discussion by thoroughly investigating vibrations of membranes (see Ref. 1, Chap. 9). Because of the circular geometry, Rayleigh

FIGURE 4. Sound pressure waveform for one stroke on a 40-cm floor tom-tom drum. A—100 ms/div (horizontal). B—10 ms/div (horizontal).



found the eigenfunctions to be Bessel functions rather than the sines and cosines that characterize organ pipes and strings. Fortunately, Bessel functions have the proper orthogonal properties to allow their use in a generalized Fourier-series expansion. Furthermore, the $J_0(z)$ and, in general, $J_n(z)$ Bessel functions have the zeros needed to match the clamped-edge boundary condition of a drum head. The essential difference between the Bessel-function series and the trigonometric-function series is that the zeros of a trigonometric function are harmonically related and the Bessel-function zeros are not.

In Fig. 4 we see oscillograms of one stroke on a drum. The fundamental resonance is 112 Hz; and from Ref. 1, Sect. 206, we expect resonances at 257, 404, 550, $\cdots$ Hz. The total time description of the drum tone is

$$d(t) = A(t)[\sin(224\pi t) + h_1 \sin(514\pi t) + h_2 \sin(808\pi t) + h_3 \sin(1100\pi t) + \cdots] \quad (2)$$

where $A(t)$ is the envelope seen in Fig. 4A. The coefficients $h_1, \cdots, h_n$ depend on the construction of the drum and how the drumstick excites the drumhead. Observe that the ratios between the drumhead resonances are 2.29, 3.59, 4.90, $\cdots$, and this is not a set of integers. In Fig. 4B, notice that the noninteger relationships of the higher tones to the lower give the appearance of a harmonic "moving" through the fundamental. For noninteger relationships, the concept of phase between lower and higher tones has no meaning. Notice, however, that the various resonances are spaced at relatively wide intervals, which means that we can consider them mathematically one at a time and add the results to get the total answer.

This brief excursion into the mathematics of musical instruments and sounds indicates that all the tones can be studied in terms of single-pulse amplitude modulation of a single sinusoidal term. The harmonic structure of the carrier waveform can be handled by individually studying the Fourier components of the carrier signal.

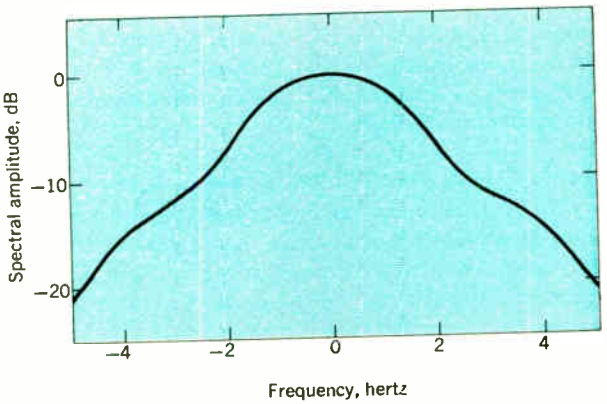
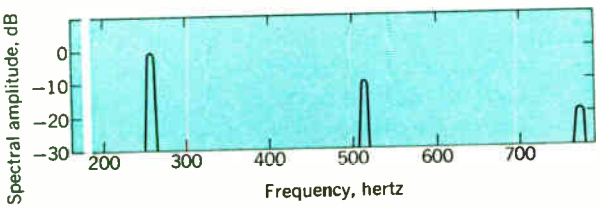


FIGURE 5. Computed spectrum for the envelope of the piano tone of Fig. 3.

FIGURE 6. Theoretically derived spectrum of a single staccato middle-C note on a grand piano.



Frequency description of musical tones

The time descriptions of musical tones are converted to frequency-domain descriptions by use of the Fourier-integral transform and simple modulation theory.

First, we will go through the details for the piano tone of Fig. 3. The amplitude function $A(t)$ does not fit any simple mathematical description; consequently, the peak height of the sound pressure waveform in Fig. 3B was sampled at 25-ms intervals across the waveform. These data were used in a Simpson's rule integral to numerically determine the Fourier transform of the amplitude function, $F_a(f)$. As seen in Fig. 5, the width of $F_a(f)$ is about 10 Hz, which is small compared with the frequency of the lowest component, 256 Hz. Simple amplitude-modulation theory tells us that the actual spectrum of the piano note will be the spectrum of the envelope moved out to the carrier frequencies (256 Hz and associated harmonics) and adjusted in height according to the magnitude of the harmonic tones. Thus, the spectrum of the entire piano note is

$$F(\omega) = F_a(2\pi f - 512\pi) + a_2F_a(2\pi f - 1024\pi) + a_3F_a(2\pi f - 3072\pi) + \cdots \quad (3)$$

which is illustrated in Fig. 6. Observe that we have not considered the phase relationships ($\phi_2, \phi_3, \cdots$) between the various harmonics because, as previously justified, the ear cannot detect the (possible) changes in angle. This point has a most important bearing on the suitability of test signals.

One qualifying remark is in order. This discussion has assumed no phase or frequency modulation of the carrier signal; i.e., f_c is a constant. One of the writers[6] has documented the derivation of more complicated integrals as well as the computer evaluation of these integrals when ω_c is not a constant. The process of staccato-tone generation by a piano will cause a small amount of instantaneous phase deviation in the carrier. From the computer work given by this same writer, we can estimate that the phase deviation in this case will not influence the spectrum in the range from its peak to 20 dB below. It seems quite reasonable to neglect phase or frequency modulation of the musical carrier at this stage of the theoretical development.

The result of long application of intuition in audio engineering is an expectation that the very steep leading edge of the drum waveform in Fig. 4 will cause a very wide spectral distribution in the frequency domain. To get an accurate idea of what happens in this case, consider an envelope function

$$A(t) = Ae^{-\alpha t}U(t) \quad (4)$$

modulating a single sinusoid $\sin(\omega_0 t)$. Here the step function $U(t)$ causes the leading edge of the modulated waveform to be essentially the first quarter cycle of the sinusoid of radian frequency ω_0. This amplitude function has a steeper leading edge than the actual waveform and is used to estimate the limiting behavior of drum notes. For a single sinusoid amplitude-modulated by Eq. (4), Papoulis gives the Fourier transform

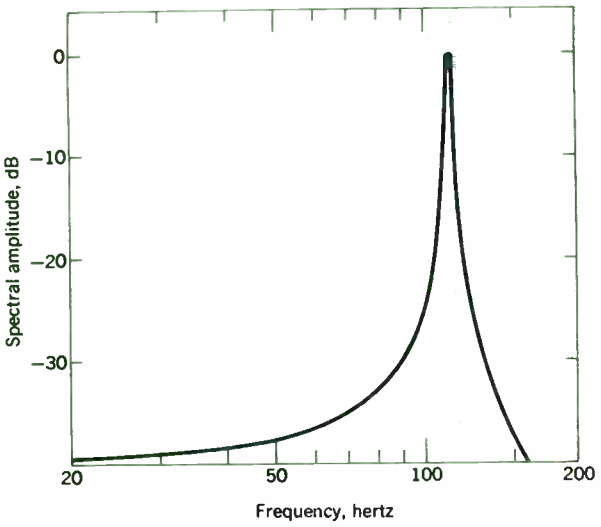


FIGURE 7. Computed spectrum from Eq. (5) corresponding to the gravest tone of the drum beat of Fig. 4.

$$F_D(\omega) = \frac{A\omega_0}{(\alpha + j\omega)^2 + \omega_0^2} \qquad (5)$$

For the gravest tone in the drumbeat of Fig. 4, $\alpha = 3.57$ and $\omega_0 = 2\pi(112)$. Equation (5) was programmed for the GE time-sharing service with these values of α and ω_0. The result of this computation is plotted on logarithmic scales (Bode amplitude coordinates) in Fig. 7. For all practical purposes, this drumbeat spectrum is contained within a band of 10-Hz total width centered at 112 Hz.

This result may be contrary to intuitive expectation, but results quite logically from the value of α compared with ω_0. In this and any other kind of modulated waveform, it is not the rise time that determines the spectral width; rather, it is the total width, in terms of the time for one cycle of the carrier, that defines spectral width.

These two spectrums make it clear that requirements for music reproduction are the same for short-duration notes as for long, continuous tones. The spectrum of short tones is confined to narrow "packets" with much blank space between each of the packets. In the local region of one of these packets, phase effects are important, since Bode[7] (see also Ref. 8) has shown that the slope of the amplitude response determines the phase shift at a given frequency. Thus, if the system amplitude response has a sharp peak or valley located in the range of our packet of spectrum, it will also have a high rate of change of phase shift in the same range. The result will be "frequency distortion" because of both the amplitude and phase shift of the system. But a gradual phase shift (such as that caused by using third-order Butterworth filters as crossover networks) will not be accompanied by an amplitude change in the region of each of our packets of spectrum; therefore, this kind of phase shift will not be audible. *In terms of transmission theory,[8] we require only that the group velocity for each harmonic of the musical tone be a constant.* There is no requirement on the system phase velocity over the entire audio spectrum.

In the section on time description of music signals, we noted that all musical tones have time descriptions similar to those we have been discussing. Applying exactly the same kind of reasoning will show that individual notes of all instruments will cause "packets" of spectrum located in the region of the harmonics that will exist for the steady (a long sustained note) production of the same note. With 88 keys on a piano and some 100 instruments in a symphony orchestra, it is not hard to visualize the addition of tones to achieve a frequency spectrum that generally covers the audio band. In simulating this spectrum for system testing, *it is necessary to simulate only the magnitude.*

We suggest that this idea of each musical note—even the shortest possible ones—occupying a group of limited-width spectral packets is quite reasonable when the physical system of musical instruments is considered. The device that produces each individual tone is some form of high-Q resonator. The justification for calling it high Q is simple: musicians demand that their tones be in tune, which is just another way of saying that the instrument must have frequency stability. Frequency stability of an oscillator demands high Q. In electronic instruments (organs, synthesizers, etc.), which use steady-running oscillators and "switches," the biggest construction problem is limiting the rise time of the switching signal so that the tone does not have what amateur radio operators call "key clicks." The minimum length of the tones is set by inertia in the keyboard switching mechanisms; thus the width of the spectrum packets can be expected to be comparable to the width illustrated for staccato piano notes. The high Q of a piano string or organ pipe implies a very-narrow-band filtering effect, and this gives a general explanation of the existence of spectral packets that occupy a relatively narrow bandwidth centered at each of the harmonics of the steady tone.

As usual, Lord Rayleigh was right when he put the emphasis on determining the harmonic content of the steady tone from an instrument. Modern theory tells us how to use the laws of probability in both the time and frequency domains to obtain a total description. Fortunately, the ear's lack of sensitivity to gradual phase shift across the audio spectrum means that we have no need to constrain audio systems into being of the minimum-phase or constant-phase velocity type; therefore, we must reject any test signal that requires a minimum-phase system for interpretation of results.

Evaluation of possible test signals

With a good understanding of both the time- and frequency-domain description of a music signal, it is possible to evaluate how useful a signal is for testing a music reproduction system. In the section on test-signal requirements, we eliminated the square wave. It was pointed out that the problem with a square wave is that it demands more of the phase characteristic of the system than is necessary for music reproduction; that is, evaluation of the results (which appear in the time domain) is meaningful only when testing a minimum-phase system.

Currently, there is great interest in using pulses for testing audio systems. There is considerable merit in testing audio amplifiers with pulses but much caution must be exercised in regard to transducers and recording equipment. To illustrate some of the pitfalls, we have used our analog and digital computers to draw Fig. 8.

A single rectangular pulse, or rectangular pulses repeated slowly enough for all effects to die out during the time between pulses, has been proposed as a test signal; mainly because it is very easy to generate with a pair of

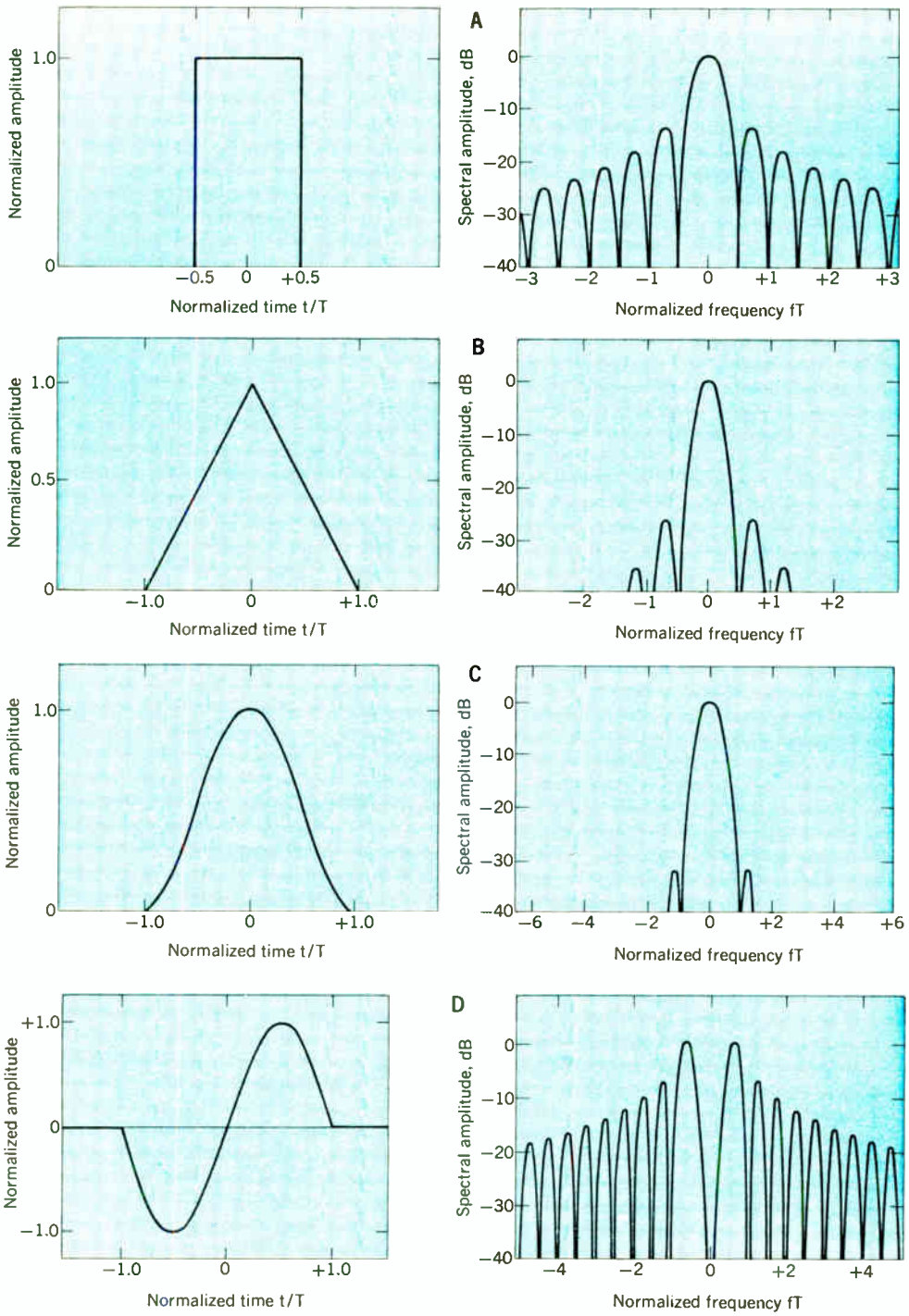


FIGURE 8. Time- and frequency-domain plots. A—Rectangular pulse. B—Triangular pulse. C—Cosine-squared pulse. D—Single-cycle sine pulse.

transistors in a multivibrator circuit. By setting the pulse width at 50 μs, the first zero in the spectrum (Fig. 8A) will occur at 20 kHz; at first glance, this might seem equivalent to sweeping the range from dc to 20 kHz. Two problems make this an unsatisfactory test signal.

First, the spectral shape is not at all compatible with the audio range. The lobe in the spectrum between 20 and 40 kHz contains a significant amount of energy and yet there is no need for music systems to reproduce this range. Furthermore, the shape of the spectrum in the 1–5-kHz range is higher than in a typical music spectrum. Adjusting the pulse width will move these problems around in such a way as to improve one while making something else worse (such as bringing the zero in the spectrum to the middle of the audio band).

Second, interpretation of the test results is nearly impossible with this test signal. Since an audio system is not a minimum-phase system, it is nearly impossible to look at the output from a test microphone and say if this is a good or bad result. As an example, assume that a two-way loudspeaker is being tested and that the test microphone is a significantly different distance from the two

drivers. This would be natural if the drivers were arranged horizontally and the microphone at 45 to 90 degrees away from a normal to the speaker baffle. The microphone would pick up two distinct pulses of different shape and separated in time by the difference in distance from the drivers. Does this tell anything about the sound of the system?

Again, the answer is no! The reason for the separation of the pulses is a difference in phase delay between the low and high portions of the spectrum, and this difference in phase is immaterial in music reproduction.

The sharp corners of the rectangular pulse can be thought of as causing the extra lobes above the first lobe of the spectrum. Rounding the corners will reduce the magnitude of the higher lobes. One pulse that has proved very useful for television testing is the cosine-squared pulse of Fig. 8. Notice that the lobes other than the first are appreciably reduced in magnitude as compared with the rectangular pulse. This reduction of the higher lobes makes the cosine-squared pulse beautifully adapted for television-system testing (because the video amplifiers are minimum phase), but does not help much for audio testing. Again, the problem is the permissible phase shift in an audio system that makes interpretation of time-domain results nearly impossible. A frequency-domain presentation of results would be useful, but the cost of the spectrum-analysis equipment would be prohibitive.

Generalization, by holding "area" constant while reducing width, of either of these two pulses to a true impulse (delta function) is not worthwhile because the widths quoted are sufficiently small. Going to a true impulse would simply overload the system and generate spectral components above the frequency range of interest.

All the pulses described have a constant spectral amplitude at the low-frequency end and this causes a fair amount of trouble because audio systems are bandpass rather than low-pass. Schaumberger[9] has suggested the use of a single cycle of a sine wave (Fig. 8D) as a pulse that avoids this problem. We note that this pulse is actually the result of modulating a sine wave with a rectangular pulse—a one-cycle tone burst. Proper adjustment of the width will place the main lobe at any desired point in the audio spectrum. This helps but still does not alleviate the problem—the presence of a wide spectral bandwidth and the need for a minimum-phase system for reasonable ease in interpreting results. We must conclude, therefore, that this signal is not an appreciable improvement over the two previous signals.

All of these pulses fail for one essential reason: they are not adequate simulations of music signals and consequently demand far too much of a system. Since music signals are essentially modulated sinusoids, adequate simulation should be a sinusoid modulated by a pulse. The first candidate that comes to mind is the tone burst. Here we have a sinusoid modulated by a square pulse and the spectrum is essentially that of the modulating pulse shifted out and symmetrical about the frequency of the sinusoid.

To ease the formidable burden of understanding the results, the tone burst should be separated from the following burst by at least ten times its width. The spectral distribution of a relatively long tone burst is found by application of modulation theory. To keep the width of the spectral distribution compatible with music signals, the width of the modulating pulse must be several hundred cycles of the sinusoid being modulated. (This is a far cry from the usual 5 to 20 cycles of tone bursts shown in recent literature. This short a tone burst has the same wide spectral distribution that ruled out the short pulses.) With a pulse several hundred cycles long, we have no chance for room acoustics to be eliminated in loudspeaker testing; actually, we see no reason to spend money generating this long tone burst when the sinusoid itself will get the same information.

A good simulation of a short-time music signal is the use of a cosine-squared pulse to modulate a sinusoid. The width of the cosine-squared pulse should be adjusted to cover about 50 cycles of the sinusoid. We can "easily" generate this kind of pulse with our $10 000 analog computer and we leave it to the reader to decide whether a specialized implementation is worth the expense. Since this pulse covers (as does a music signal) only a narrow portion of the spectrum, the frequency of the modulated sinusoid (and synchronously the width of the pulse) would have to be moved through the audio band to get complete data. We feel that a swept sinusoid signal will give more understandable data at considerably less cost.

At this point, it appears that we must invoke some kind of devil for all the deterministic test signals we consider suitable candidates for supplanting the sine wave. Is there no alternative to a signal that has reigned as king for most of the past century?

The answer is a qualified yes! If we relax our restrictions on the education of the users of the test and the cost of test equipment, then noise can be a useful test signal for music reproduction systems. As shown recently by Ashley and Henne,[2] the probability distribution function (in the time domain) for music is close to Gaussian. This simple fact is of fundamental value here. First, most commercial noise generators suitable for the audio band are either Gaussian or pseudo-Gaussian, which is a more than adequate approximation for audio testing. Second, audio systems are engineered for low harmonic distortion, which means that they are linear systems in the classic sense. Therefore, we can use linear system theory. Third, linear system theory tells us that the probability distribution function at the output of an audio system driven by Gaussian noise will remain Gaussian regardless of the frequency response. Fourth, the spectrum (power spectral density) at the output of a linear system is the product of the input spectrum and the frequency response (magnitude only) of the system.

To use noise as a test signal, it is necessary only to shape the input spectrum to simulate a music signal and apply this signal to the audio system. Presently, our best guess for the shape of the spectrum would be pink (sloping downward at 3 dB per octave with increasing frequency). The output of the system is checked with a wave or spectrum analyzer.

Where much education and judgment are required is in the selection of the spectrum-analyzer bandwidth. Narrow bandwidths increase the resolution of the test and the test time. Broad bandwidths or the use of overlapping filter banks speed the testing at the expense of resolution. It is possible to obtain the equivalent of phase information when using noise as a test signal. The recently developed correlators give good information that can be processed mathematically in order to find the phase of the system transfer function.

One final word about noise testing. All the noise theory useful here depends on linear or low-distortion systems. Noise testing is not sensitive enough to detect the level of distortion that must be maintained in audio equipment. In regard to loudspeaker testing, the noise test signal cannot yield the results given by the intermodulation test devised by Klipsch.[10] It seems that a sinusoidal signal will have to be used along with the noise test to determine if distortion is adequately low. Only then will the noise test give some insight into the performance of the system in an actual listening environment, because the absence of phase coherence in the signal will lessen the effects of room resonances. In the final analysis, however, it is a moot question as to whether noise testing is worth the price in either education or capital-equipment expenditure.

Conclusion

The theoretical background for the performance of both musical and electroacoustic devices was established by Lord Rayleigh and based on the use of sinusoids. The sinusoid remains unchallenged for relating device performance to design theory and for making very sensitive distortion tests. All of the short-duration test signals are unsuitable for audio testing because they are easily applied only to minimum-phase systems. Noise testing might give insight into system performance in an actual listening environment but requires supplementary sine-wave testing for distortion measurement.

We conclude that the sine wave is still very much the king of test signals. Even our excursion into communication theory did not locate a reasonable challenger!

R. H. Small and P. W. Klipsch gave us many useful comments after reading our first manuscript. These comments have aided us in clarifying several points. The analog and digital computations reported herein were made in the laboratories of the Colorado Springs Center of the University of Colorado.

REFERENCES

1. Rayleigh, J. W. S., *The Theory of Sound*, vols. 1 and 2, 1877; reprinted New York: Dover, 1945.

2. Ashley, J. R., and Henne, L. M., "Operational amplifier implementation of ideal electronic crossover networks," *J. Audio Eng. Soc.*, vol. 19, Jan. 1971.

3. Small, R. H., "Constant-voltage crossover network design," *Proc. IREE Australia*, vol. 31, p. 66, Mar. 1970; also *J. Audio Eng. Soc.*, vol. 19, p. 12, Jan. 1971.

4. Ashley, J. R., and Kaminsky, A. L., "Active and passive filters as loudspeaker crossover networks," presented at 39th Conv. of Audio Eng. Soc., New York, Oct. 1970.

5. Battjes, C., "Short pulse techniques of adjusting wideband amplifiers," *Tekscope*, vol. 3, Tektronix, Inc., P.O. Box 500, Beaverton, Oreg. 97005, Jan. 1971.

6. Saponas, T. A., "Computation of the microwave spectrum of Gaussian envelope pulses with phase modulation," 2nd prize paper presented at 1970 Internat'l Symp. on Electromagnetic Compatibility, Anaheim, Calif., July 15, 1970.

7. Bode, H. W., *Network Analysis and Feedback Amplifier Design*. Princeton, N. J.: Van Nostrand, 1940.

8. Papoulis, A., *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

9. Schaumberger, A., "Impulse measurement techniques for quality determination in hi-fi equipment, with special emphasis on loudspeakers," *J. Audio Eng. Soc.*, vol. 19, p. 101, Feb. 1971.

10. Klipsch, P. W., "Modulation distortion in loudspeakers; Part II," *J. Audio Eng. Soc.*, vol. 18, p. 29, Feb. 1970.

J. Robert Ashley (SM) has been interested in audio systems since his high-school days in the early 1940s. Two years in the U.S. Navy electronics technician program enabled him to attend the University of Kansas; he received the B.S.E.E. degree in 1952. After working on klystrons at Sperry for a year, he returned to the University of Kansas as an instructor and part-time graduate student, where he was first introduced to "equivalent circuits" for audio transducers while teaching an electroacoustics course. Upon receiving the M.S.E.E. degree in 1956, he became an R&D engineer for the Sperry Electronic Tube Division, during which time he authored "Introduction to Analog Computation." In 1965, under an NSF engineering traineeship, he attended the University of Florida, receiving the Ph.D. degree in EE in 1967.

At present, Dr. Ashley is associate professor of electrical engineering at the Colorado Springs Center of the University of Colorado, where his main teaching interest involves integrating a meaningful experience with computers into the undergraduate engineering program. Every few semesters he teaches an electroacoustics course that is considerably improved over his 1955 version because of the availability of modern computers. Dr. Ashley is a member of the editorial boards of the Journal of the AES and the IEEE MTT Transactions, a senior member of SCI, and a member of the AES and seven honorary and fraternal societies.

Thomas A. Saponas (S) is currently a senior in electrical engineering at the Colorado Springs Center of the University of Colorado under a Boeing Company scholarship. Although his experience in electroacoustics prior to his collaboration with Dr. Ashley was that of a hobbyist, he has presented both the present article and another paper (also written with Dr. Ashley) at two recent AES conventions. In addition, a student paper that he presented at the IEEE International Symposium on Electromagnetic Compatibility in 1970 won second prize. More recently, he won first prize in the student paper competition at the IEEE International Microwave Symposium, Washington, D.C., May 16–20 of this year, for a paper on spectrum conservation. He is a member of Tau Beta Pi and a student member of ACM.

Randolph C. Matson received the B.S.E.E. degree from Kansas State University in 1961, and the Ph.D. degree from the University of Texas, Austin, in 1967. From July 1967 until September 1968, he was a research engineer with the Boeing Company at Seattle, Wash. In 1968, he joined the electrical engineering faculty of the University of Colorado at Colorado Springs. Dr. Matson is a member of both Eta Kappa Nu and Sigma Tau.

The environment and the electrical engineering curriculum

If electrical engineers have an important role to play in dealing with environmental issues, then engineering school curriculums must be modified to familiarize students with the problems involved

Warren L. Flock University of Colorado

According to the ECPD definition, engineering involves the use of the forces and materials of nature for the benefit of mankind. Unfortunately, in the course of their daily work engineers may forget to question whether their technological developments really will be a boon to society, and experience has shown that too often quite the reverse proves to be true. If the engineer of the future is to be more than a technical specialist and is to assume an effective and active role in dealing with the environment, then it would seem that electrical engineering curriculums should be adapted to familiarize the student with environmental problems and their relationship to his chosen profession. This article discusses the beginning that has been made at one university with the introduction of an experimental course on environmental electromagnetics.

Belatedly, the environment has become a popular cause—and as a result electrical engineers have begun to wonder what relation their profession has, or should have, to the world around them. This question is answered in part by Gordon Friedlander in the November and December 1970 issues of SPECTRUM.[1] After reading his articles, one can hardly avoid being impressed by the importance of electrical engineering's role in environmental issues. The electric power field has declined in popularity over the years, however, and electrical engineers in other specialities may wonder how their training fits in. Some suggested subject areas in this respect are electromagnetic radiation, instrumentation, systems theory, acoustical noise, and transportation.

If it is accepted that electrical engineering does have an important role in dealing with the environment, or even if it is only acknowledged that electrical engineers should be informed about it, the question arises as to what changes, if any, should be made in electrical engineering curriculums to help prepare students to cope with, or at least understand, the problem. The purpose of this article is to discuss this subject and to describe an experimental course on environmental electromagnetics that was given in the fall of 1970 at the University of Colorado.

Introducing the environment

There are various ways in which the environment can be introduced into electrical engineering curriculums and it would seem advantageous to utilize several, rather than just one or two, of these approaches. One possibility is to introduce comments about the environment and examples of applications into standard electrical engineering courses. Energy-conversion classes can discuss the environmental problems of the electric power industry, computer classes can discuss computer simulation of ecological systems, etc. This approach is certainly to be encouraged. It may be surprising to find out how much is already being done in this vein in one's own department. However, most courses have very full agendas, and not all instructors will be inclined in this direction.

An obvious procedure is to encourage students to take suitable environmental courses in other university departments. Perhaps the present sociohumanistic elective requirements can be modified or enlarged to encompass the environment better. To facilitate the choice of suitable electives, a list of such courses can be prepared for use by students and advisors. A course in ecology would be highly desirable for engineers, but many biology departments are being swamped with students as a result of the widespread interest in the subject. Also, a course that does not require prerequisites may not be available.

Special courses of various kinds—all-university seminars on the environment, interdepartmental courses within the engineering college, short summer courses—can be extremely valuable. For example, Prof. Samuel Maley of the Department of Electrical Engineering at the University of Colorado is arranging a short course on atmospheric monitoring for the summer of 1972.

Another way to deal with the environment is to develop new environmental or societal engineering curriculums or departments. Although this approach may have some merit, it can be argued that all engineering curriculums should be involved with the environment to some degree. Concern for the environment should not and cannot be delegated to one curriculum or department in large universities. However, it need not preclude concentration on a strong environmental program in some institutions.

With regard to electrical engineering curriculums, reliance should not be placed on electives given outside the department or on special courses scheduled irregularly or featuring lectures by large numbers of specialists. If students hear about the environment primarily only in courses outside their department, they hardly will tend to feel very seriously involved. Persons outside the department, furthermore, are not apt to be highly qualified to discuss the application of electrical engineering to the environment. Special courses taught by several lecturers tend to lack continuity and to be difficult to administer

on a regular basis. One of the difficulties in the past has been that each aspect of the environment has been handled by someone who was expert in it but likely to be relatively uninformed about other aspects. There is a need for some individuals to develop a broad overall knowledge of the environment, preferably in addition to expertise and depth in a particular area. Thus it is suggested that an electrical engineering department should develop a small number of regularly scheduled courses that are especially oriented to the environment.

Such courses need not be taught by one instructor alone and can make use of some guest lecturers. Student interest and participation can be encouraged by having each student prepare a paper and present it orally to the class. The writer feels that it is preferable, however, for one or two instructors to assume major responsibility for presenting material rather than to rely on using a large number of lecturers. It is recognized that the environment is an interdisciplinary subject and that the use of interdisciplinary teams for research and teaching is commendable and should be utilized in many situations. It is also desirable, however, for individual faculty members, students, and departments to develop some interdisciplinary characteristics and interests within themselves; this is one reason for suggesting that one instructor assume responsibility for an environmental course of relatively broad scope rather than restrict his preparation and comments to his own specialty.

An experimental course, based on the foregoing philosophy, was given as an elective course for seniors in the fall of 1970 at the University of Colorado. This course, entitled "Environmental Electromagnetics," could be used as a technical elective in electromagnetic fields, as far as meeting departmental elective requirements was concerned. It was probably the first, and very possibly the last, environmental course in their undergraduate career for most of the students who took it. For this reason, and also because general knowledge about the environment as well as specialized knowledge is needed, the course included a general introduction to environmental problems.

The environment, conservation, and natural resources

The term "environment" came into widespread use fairly recently. Formerly, one tended to speak of conservation or natural resources. The definitions of the three terms are not identical, but any topic that falls under one of these headings is a legitimate one for consideration in an environmental course. It is suggested that an engineer's knowledge of the environment, conservation, and resources should include basic information about population, gross national product and economic growth, ecology, the history of the conservation and environmental movements, political and legal developments in pollution control, etc. In fact, it seems obvious that if engineering is concerned with utilizing the materials and forces of nature for the benefit of mankind, as the Engineers Council for Professional Development's definition states, such information is just as important to engineering as is information about circuits, structures, devices, or man-made systems. Furthermore, if engineering students do not learn about these topics in their engineering courses, many will not hear of them at all in their academic program.

A very important part of the course on environmental electromagnetics was devoted to the topics suggested here. Various facts and figures were presented. It was pointed out, for example, that the population of the United States would rise from its 1970 level of 205 million to 275 million in 2037 even if no more than two children per family were born from 1970 on. It was mentioned that whereas the population of the United States grew by 13.6 percent between 1960 and 1970 the gross national product increased by 30 percent. The history of the conservation movement was outlined, and ecosystems were defined. However, the intention was primarily to initiate thought and discussion. The roles of engineers and scientists in dealing with the environment were discussed and an effort was made to present various views on questions such as population and growth. It was not meant to use the course as a forum for promoting personal viewpoints, but no effort was made to hide a sense of urgency. Although the material of this section was presented at the beginning of the course, it was not restricted to that period but permeated the entire course.

References found to be useful and cited in notes prepared for this portion of the course were concerned with the resources essential for man's survival,[2,3] the environment[4] and natural resources[5] in general, population,[6-8] economic growth,[9,10] the history of the conservation movement,[11] current news of conservation and the environment,[12-14] ecology,[15-19] shortcomings in the application of technology and science,[20] and the roles of engineers and scientists in applying their training and know-how and in conducting long-range analysis and research on environmental problems.[21,22] The beautiful and classical presentation of a conservation viewpoint, *Sand County Almanac* by Aldo Leopold,[23] was recommended to the class. The literature on the environment is extremely voluminous, and no effort is made to present a complete bibliography here; only a few representative references are listed, particularly those that present viewpoints that might otherwise be overlooked.

Environmental electromagnetics

The principal themes of the course were electromagnetic radiation, energy resources and electric power, and associated environmental considerations.

The electromagnetic spectrum covering a range of frequencies from near zero to as high as 10^{23} or more was considered a major natural resource that is essential to life and valuable to man in various other ways as well. Electromagnetic radiation was a major factor in all four technical areas of the course: the electromagnetic spectrum and its utilization by man for telecommunications, the atmosphere and thermal radiation, energy resources and utilization, and remote sensing of the environment.

The electromagnetic spectrum. The first technical section—The Electromagnetic Spectrum and Its Utilization—was concerned with telecommunications systems, the radio spectrum, radio noise, and interference and compatibility problems. The history of telecommunications, from the first telegraph systems to modern satellite and PCM systems currently under development, is fascinating and a good background for this section. This topic might be assigned or suggested for reading in a text such as Martin's,[24] which also considers possible future developments in telecommunications. Publications of the Joint Technical Advisory Committee[25,26] and of

the IEEE Group on Electromagnetic Compatibility are useful for considering the use of the radio spectrum and its pollution. Radio noise is a basic phenomenon that is very important from the environmental viewpoint and is suitable for quantitative treatment and for the assignment of problems. Radio noise generated internally and externally with respect to the receiving system and both natural and man-made noise need to be considered. Many references on communication theory include basic material on noise, but others that are system-oriented, such as "Transmission Systems for Communications," [27] are perhaps more suitable at this point. The text on radio astronomy by Kraus[28] is among those that treat noise of external origin and pertinent related concepts.

Potential interference between surface and satellite microwave links and microwave radars, protection of frequencies for radio astronomy, spectrum congestion in urban areas, and problems that arise when pacemaker heart devices are used in the presence of microwave ovens are among the interesting topics that can be considered in this section. Interference between microwave telecommunication systems is the particular subject the writer chose to emphasize. At this stage the reader may have noticed that many of these topics could be considered in conventional electrical engineering courses. One might think that all of the material of an environmental course must be obviously related to a pressing environmental problem and distinct from conventional subject matter, but this limitation is neither desirable nor possible. On the contrary, many conventional topics can be approached from an environmental viewpoint.

Whether or not one thinks of telecommunications as an environmental or natural resource topic, it is obviously one in which the application of engineering know-how is directly important to society. Future developments in telecommunications may have important effects on society and on the environment. For example, some business and educational travel might be reduced through greater reliance on picturephones, cable TV, etc.

Another logical subject for future coverage is that of biological effects of electromagnetic radiation.

The atmosphere and thermal radiation. The section on the atmosphere and thermal radiation began with a treatment of the structure of the atmosphere (variation of pressure, density, temperature, composition, and ionization with height, etc.).[29–31] Much attention was devoted to atmospheric effects on electromagnetic radiation. The subject of the atmosphere as a communications medium, which is an extension of the material of the previous section, was considered. Expressions for the indexes of refraction in the troposphere[29] and ionosphere[31] were developed. (A more advanced and thorough treatment of such topics is available in another departmental course—Atmospheric Effects on Electromagnetic Waves. This course might well be considered environmental by definition and could be made more obviously so in the popular sense with little additional effort.)

The effects of the atmosphere on thermal radiation from the sun and from the earth also were treated. This topic involves the important subject of the radiation balance of the earth and possible effects on this balance of increases of atmospheric carbon dioxide and particulate matter.[32,33] To present such subjects adequately it was considered necessary to discuss blackbody radiation laws and the mathematics of absorption and emission within the atmosphere. Kraus' text on radio astronomy[28] again was a suitable reference. Because of the considerable attention given to thermal radiation, this section is called "The Atmosphere and Thermal Radiation."

Any discussion of the atmosphere as part of the environment must consider the major problem of air pollution.[34,35] The sources of pollution, control measures, natural processes of dispersion and degradation or conversion of pollutants, and effects of pollution were considered. An important aspect of air pollution to which electrical engineering is readily applicable is that of atmospheric monitoring as applied to the detection and measurement of temperature inversions and other atmospheric characteristics and gaseous and particulate pollutants. Discussion of some of these techniques, however, was deferred to the section on remote sensing. Biogeochemical cycles involving carbon, oxygen, nitrogen, sulfur, etc., were discussed in the section on the atmosphere; this important topic was well covered in the special *Scientific American* issue on the biosphere.[36]

The section on the atmosphere was a large and important one, and portions of it (the atmosphere as a communications medium, blackbody radiation, etc.) were suitable for quantitative treatment and problem assignment.

Energy resources and utilization. Energy is crucial to modern civilization because of the high and increasing rate at which it is consumed and because it is convertible into most of the other requirements of life.[37] With unlimited energy, fresh water could be produced from salt water, metals could be refined from the seas and from low-grade ores, food could be manufactured in sufficient quantity to meet foreseeable needs. These considerations illustrate the significance of the statement that energy is "convertible into most of the other requirements of life," or can be used to produce the means to satisfy the other requirements. Conversely, our present high consumption of our resources and the convenience of using electricity for air conditioning, space heating, television, and various household gadgets results in the consumption of large amounts of energy.

The study of energy resources was begun by considering the energy received at the earth's surface from the sun and its utilization in the biosphere.[38] Incident solar energy is utilized in the process of photosynthesis in plants to synthesize organic carbon compounds, which are man's major sources of food and energy. In the photosynthesis process plants utilize carbon dioxide, water, and sunlight, and give off oxygen. It is considered that the atmospheric oxygen has been formed in this way.[32] The efficiency of photosynthesis is low—only 0.01 of the light energy absorbed may be converted—but it has been responsible for producing the fossil fuels as well as for maintaining the biosphere.[38] Man's food is provided by plants or by animals that feed upon plants. For his minimum daily needs man uses an average of 2.4 kWh per day, the energy content of his food. Primitive man used very little more energy than this altogether, and in some parts of the world per capita consumption of energy is still not much more than this amount. In contrast, per capita consumption in the U.S. is about 240 kWh.

The direct use of solar energy, fossil fuel and nuclear energy resources, and miscellaneous sources such as water power and geothermal, tidal, and wind energy also were discussed. Daniels' fine treatment of solar energy[39]

and Hubbert's thorough coverage[40] of fossil fuels and nuclear energy were particularly useful here.

Probably no environmental problem is more crucial than that of energy resources and their utilization. Almost any development of new electric power facilities, as well as the processes of extracting and transporting oil and coal, etc., has detrimental environmental effects that are objectionable to many. All of man's ingenuity and initiative will be required to supply essential needs and still protect the environment. In attacking this problem consideration must be given to just what the essential needs are as distinguished from frills and waste. In this respect it is common to speak of demand for electricity, but one wonders to what extent the "demand" is generated by promotion and advertising techniques and to what extent it is a true demand. In any case it is obvious that the consumption of electric energy cannot continue to increase indefinitely. To attempt to discourage the use of electricity in any way may seem unrealistic to many, but one wonders if failure to take such measures is not more unrealistic. The problems of electric power production are described in the aforementioned SPECTRUM articles[1] and the reader is referred to them for further discussion of this subject. It might be added that hydroelectric installations may also create problems, as they require the construction of dams and the flooding of valleys. Dams for the storage of cooling water for thermal power plants may be no less objectionable. In addition to the fact that most good hydro sites in the United States are already in use, there is increasing reluctance to have the last relatively few natural unspoiled river valleys flooded. Elgerd[41] has proposed a moratorium on the construction of new hydropower installations.

Consideration of the topic of energy utilization leads to a large area of environmental problems that might not initially appear to be associated with energy. Wasteful or extravagant use of anything, however, probably means wasteful or extravagant use of energy, although there will be disagreement as to what is wasteful or extravagant. Mammoth newspapers and trash mail involve the consumption of energy. Mechanized recreation—snowmobiles, trail bikes, power boats—involve the consumption of energy. Extreme examples of mechanized recreation, such as a camper towing a trailer with a boat mounted on top and a trail bike or two strapped on the sides, are commonly seen on our highways. The use of high-powered automobiles to carry single commuters over medium to long distances is commonly cited as an example of extravagant use of energy. Whether or not the products are used wisely, industries consume large amounts of energy—about 41 percent of the total production of electric energy in the United States. The aluminum industry is a particularly large consumer of electric energy. The automobile industry is another large consumer, but its consumption might be lowered somewhat if automobiles were designed for long life and ease of reclamation.

In treating the vital complex and controversial subject of energy resources and utilization it is essential that the various viewpoints be presented and examined carefully. An effort must be made to avoid excessive personal bias and adherence to preconceived ideas. Penetrating questions should be asked of oneself as well as of others. The case-study method seems to be appropriate for considering electric power installations and associated environmental problems. The techniques for reducing gaseous and particulate emissions and for disposing of waste heat, or using it beneficially, should also be considered.

Remote sensing of the environment. To understand the environment and to manage it wisely require information about its characteristics and condition. The necessary data sometimes can be obtained by on-the-spot observations and measurements but in other situations it may be desirable or essential to sense the condition of the environment remotely. Before the advent of rockets and satellites, determination of properties of the upper atmosphere of necessity involved remote sensing, and geophysical exploration usually involves remote sensing of conditions below the earth's surface. However, a common application of the term is to observations of the earth from aircraft and spacecraft. Remote sensing from satellites has certain advantages over sensing from aircraft, and conversely. Information from spacecraft, aircraft, and from the ground may all be required to do a satisfactory job of, for example, surveying a certain natural resource. Useful remote-sensing techniques can be carried out from the ground, and, even if attention is concentrated on airborne or satellite remote-sensing observations, some ground observations must be made to interpret and confirm the airborne or satellite measurements. In the terminology of remote sensing, the observations on the ground provide ground truth.

Aerial photography is a time-honored and valuable remote-sensing technique but was mentioned only briefly in the course on environmental electromagnetics, which was prepared for electrical engineers. Attention was given to infrared[42] and microwave radiometry[43] and imaging systems and to various types of radar,[44] including conventional and ultrasensitive pulse radar,[45] FM continuous-wave radar,[46] chirp or pulse compression radar,[47] Doppler radar, and side-looking or synthetic-aperture radar.[48] The material on thermal radiation introduced in earlier sections is applicable to radiometry. Radiometers might be used, for example, for applications such as detecting illegal discharges of oil by ships at sea and for studies of thermal pollution. Conventional pulse radars such as those used for aircraft control and surveillance can also be used for obtaining data about weather, bird movements,[49] sea ice, etc. Multipurpose use of radars should be more actively considered than in the past, and environmental limitations on radar performance[47] also need attention. Frequency-modulated continuous-wave radar has been used successfully for high-resolution studies of the troposphere.[46] Chirp radars utilize frequency modulation within the transmitted pulse, with subsequent pulse compression on reception, to obtain improved range resolution without reduction in pulse power. Side-looking radars can provide images of the terrain traversed in darkness and overcast weather, but at the expense of considerable complexity. For example, the Darien region of Panama, which is almost always overcast, was mapped by side-looking radar. Azimuth resolution in side-looking radar is based on the fact that the Doppler effect produces linear frequency modulation of the received echo, somewhat as in a pulse-compression radar. The elementary theory of various types of radar was an interesting and pertinent quantitative subject that was treated in the course.

Optical and acoustic radar techniques also have considerable utility for remote sensing of the atmosphere.

It has been proposed that worldwide monitoring of particulate matter be undertaken by using four lidar, or optical radar, sites in conjunction with a larger number of solar radiation monitors.[33] The interest in the particulate matter lies in its possible effects on the radiation balance of the earth, and lidars could make measurements as a function of height in the troposphere and lower stratosphere, as well as monitor particulate matter in the atmosphere on a local scale. A discussion of the various optical scattering mechanisms (Mie, Rayleigh, Raman, etc.) is pertinent to such applications of lidar. An important paper by Derr and Little[50] compares optical, radio, and acoustic radar techniques and their applications.

Techniques for monitoring and identifying atmospheric gaseous pollutants are of much interest at present, and a recent paper considers the use of tunable semiconductor diode lasers for this purpose.[51] Techniques for monitoring SO_2, NO_2, and iodine gases from aircraft or satellite by comparing observed absorption spectrums with theoretical spectrums have been described by Barringer. One of the papers in which he has commented on this subject was a useful summary paper that treated various airborne techniques, including aerogeophysical techniques for discovering magnetic and conductive anomalies related to commercial mineral deposits.[52] Also among the topics discussed was a method for observing fluorescence of minerals and chlorophyll in daylight by measuring radiation intensity in the Fraunhofer lines, which are deep absorption lines in the solar spectrum. The use of this technique for observing luminescence is also discussed by Hemphill and Stoertz.[53]

Colwell has described the application of remote sensing to the inventory and monitoring of natural resources[54]; the University of Michigan has sponsored a series of conferences on remote sensing[55]; the April 1969 issue of the PROCEEDINGS OF THE IEEE was devoted to remote sensing; and NASA has an ERTS (Earth Resources Technology Satellite) program, which will provide data that will be analyzed by a number of teams of investigators. Alaska is a natural location for the application of remote sensing because of its large size and the fact that it is relatively inaccessible and undeveloped. In December 1969, Alaska and the U.S. Department of the Interior sponsored a symposium entitled "The Use of Remote Sensing in Conservation, Development, and Management of Natural Resources of the State of Alaska,"[55] and the University of Alaska has an active interest in remote sensing and hopes to participate in the ERTS program.

Conclusion

From this discussion, it is obvious that environmental electromagnetics could easily involve more than one semester's work. The present course might be extended to more than one semester, or portions might be taken away and set up as separate courses. As it stands it is necessary to select certain subjects for thorough treatment and mention others only briefly. It was not feasible to introduce a series of new courses at once, however, and although students may fit one environmental course into their schedule easily they may not be able to accommodate several. The present course can best be considered as introductory, and could be followed by more specialized treatments of telecommunications (economic, social, and environmental aspects as well as strictly technical ones), atmospheric effects on electromagnetic waves, energy resources and utilization, or remote sensing. In addition, such topics as acoustical noise, transportation (including substitutes for the internal-combustion engine, high-speed surface transportation, and air traffic control), and systems theory are suitable for consideration in electrical engineering courses that are environmentally oriented.

Environmental electromagnetics should be regarded as a first step in introducing the environment into an electrical engineering curriculum, and further effort will be needed to work out optimum overall plans. One of the important features of the course was the material on the environment, conservation, and natural resources in general, but this need not be repeated more than once in a sequence of courses, and perhaps would not be included at all, depending on the circumstances, in courses given elsewhere. However it seems essential that such material be kept very much in mind and that it permeate all environmental courses to some degree whether or not a separate section is set aside for it.

Engineers are commonly considered to be technical specialists, and to be a good specialist is a worthwhile accomplishment in itself. Perhaps most engineers will continue to specialize, but there is opportunity and need for persons with broader training as well—and the best generalists may well be those who have depth in some area as well as general knowledge. Thus some engineers should make good environmentalists, or generalists, especially if the engineering curriculum is modified to encourage their interest. In this respect, the discussion presented here has been most directly applicable to the undergraduate curriculum, and environmental electromagnetics was developed for seniors, but attention should be directed to graduate programs as well. These have tended to be more restricted and specialized than undergraduate programs but a way must be found to give the engineering graduate student additional opportunity, encouragement, and credit for work outside of one specialized field.

According to the ECPD definition engineering involves the use of the forces and materials of nature for the benefit of mankind. Most engineers presumably accept this definition, but it often seems to be forgotten in the course of our daily activities. Perhaps the difficulty has been that we have not questioned as frequently or as deeply as we should the social implications of our work. It seems to have been blindly accepted that all technological, industrial, or commercial developments have benefited mankind but accumulated experience now shows that such is not the case. To help insure that technology henceforth will be applied beneficially, and to help convince society to support technology and science, engineers must no longer be restricted to their former roles of technical specialists.

REFERENCES

1. Friedlander, G. D., "Power, pollution, and the imperiled environment," *IEEE Spectrum*, vol. 7, pp. 40–50, Nov. 1970; pp. 65–75, Dec. 1970.

2. National Academy of Sciences–National Research Council, *Resources and Man*, San Francisco: Freeman, 1969.

3. Brown, H., Bonner, J., and Weir, J., *The Next Hundred Years*. New York: Viking, 1963.

4. Revelle, R., and Landsberg, H. H. (eds.), *America's Changing Environment*. Boston: Houghton Mifflin, 1967.

5. Huberty, M. R., and Flock, W. L. (eds.), *Natural Resources*. New York: McGraw-Hill, 1959.

6. Ehrlich, P. R., and Ehrlich, A. H., *Population Resources Environment*. San Francisco: Freeman, 1970.

7. Hardin, G. (ed.), *Population, Evolution, and Birth Control*. San Francisco: Freeman, 1969.

8. Hauser, P. M., *The Population Dilemma* (2nd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1969.

9. Fuller, V., "Natural and human resources" in *Natural Resources*, M. R. Huberty and W. L. Flock, eds. New York: McGraw-Hill, 1959.

10. Boulding, K. E., "The economics of the coming spaceship earth" in *Environmental Quality in a Growing Economy*, H. Jarrett, ed. Baltimore: The Johns Hopkins Press, 1965.

11. Udall, S., *The Quiet Crisis*. New York: Holt, Rinehart and Winston, 1965.

12. *Audubon, The Magazine of the National Audubon Society*, New York, N.Y.

13. *Sierra Club Bull.*, San Francisco, Calif.

14. *Environment*, St. Louis, Mo.

15. Buchsbaum, R., and Buchsbaum, M., *Basic Ecology*. Pittsburgh: Boxwood Press, 1958.

16. Kormondy, E., *Concepts of Ecology*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

17. Odum, E. P., *Ecology*. New York: Holt, Rinehart and Winston, 1963.

18. "Diversity and stability in ecological systems," Rept. of Symp. Brookhaven National Laboratory, Upton, N.Y., May 26–28, 1969; available from Clearinghouse, Springfield, Va.

19. Ehrenfeld, D. W., *Biological Conservation*. New York: Holt, Rinehart and Winston, 1970.

20. Commoner, B., *Science and Survival*. New York: Viking, 1967.

21. Reinecke, E., "Pollution, political expediency, and technological competence," *Mech. Eng.*, vol. 92, pp. 35–37, July 1970.

22. Brooks, H., and Bowers, R., "The assessment of technology," *Sci. Am.*, vol. 222, pp. 13–21, Feb. 1970.

23. Leopold, A., *Sand County Almanac*. New York: Oxford, 1949.

24. Martin, J., *Telecommunications and the Computer*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

25. "Radio spectrum utilization," Rept. Joint Technical Advisory Committee, 1965.

26. "Spectrum engineering—the key to progress," Rept. Joint Technical Advisory Committee, 1968.

27. "Transmission systems for communications," Bell Telephone Laboratories Technical Staff, 1964.

28. Kraus, J. D., *Radio Astronomy*. New York: McGraw-Hill, 1966.

29. Bean, B. R., and Dutton, E. J., *Radio Meteorology*. New York: Dover, 1966.

30. Rishbeth, H., and Garriott, O. K., *Introduction to Ionospheric Physics*. New York: Academic, 1969.

31. Flock, W. L., "Introduction to atmospheric effects on electromagnetic waves" (unpublished notes).

32. Johnson, F. S., "The balance of atmospheric oxygen and carbon dioxide," *Biol. Conserv.*, vol. 2, pp. 83–89, Jan. 1970.

33. *Study of Critical Environmental Problems, Man's Impact on the Global Environment*. Cambridge, Mass.: M.I.T. Press, 1970.

34. "Cleaning our environment," American Chemical Society, 1969.

35. "Air conservation," American Association for the Advancement of Science, 1965.

36. "Special issue on the biosphere," *Sci. Am.*, vol. 223, Sept. 1970.

37. Weinberg, A. M., and Hammond, R. P., "Limits to the use of energy," *Am. Scientist*, vol. 58, pp. 412–418, July/Aug. 1970.

38. Rabinowitch, E., and Govindjee, *Photosynthesis*. New York: Wiley, 1970.

39. Daniels, F., *Direct Use of the Sun's Energy*. New Haven: Yale University Press, 1964.

40. Hubbert, M. K., "Energy resources," in *Resources and Man*, National Academy of Sciences–National Research Council. San Francisco: Freeman, 1969.

41. Elgerd, O. I., "Moratorium on hydro plants?" *IEEE Spectrum* (*Forum*), vol. 7, p. 12, Jan. 1970.

42. Hudson, R. D., *Infrared System Engineering*. New York: Wiley-Interscience, 1969.

43. Staelin, D. H., "Passive remote sensing at microwave wavelengths," *Proc. IEEE*, vol. 57, pp. 427–439, Apr. 1969.

44. Skolnik, M. I., *Introduction to Radar Systems*. New York: McGraw-Hill, 1962.

45. Hardy, K. R., and Katz, I., "Probing the clear atmosphere with high power, high resolution radars," *Proc. IEEE*, vol. 57, pp. 468–480, Apr. 1969.

46. Richter, J. H., "High resolution tropospheric radar sounding," *Radio Sci.*, vol. 4, pp. 1261–1268, Dec. 1969.

47. Nathanson, F. E., *Radar Design Principles*. New York: McGraw-Hill, 1969.

48. Harger, R. O., *Synthetic Aperture Radar Systems*. New York: Academic, 1970.

49. Flock, W. L., "Monitoring bird movements by radar," *IEEE Spectrum*, vol. 5, pp. 62–66, June 1968.

50. Derr, V. E., and Little, C. G., "A comparison of remote sensing of the clear atmosphere by optical, radio, and acoustic radar techniques," *Appl. Opt.*, vol. 9, pp. 1976–1992, Sept. 1970.

51. Hinkley, E. D., and Kelley, P. L., "Detection of air pollutants with tunable diode lasers," *Science*, vol. 171, pp. 635–639, Feb. 19, 1971.

52. Barringer, A. R., "Remote-sensing techniques for mineral discovery," Paper 20, Ninth Commonwealth Mining and Metallurgical Congr., The Institution of Mining and Metallurgy, London, 1969.

53. Hemphill, W. R. and Stoertz, G. E., "Remote sensing of luminescent materials," *Proc. 6th Symp. Remote Sensing of the Environment*, University of Michigan, Oct. 1969, pp. 565–585. (See also other papers in this proceedings and proceedings for other symposiums in the series.)

54. Colwell, R. N., "Remote sensing of natural resources," *Sci. Am.*, vol. 218, pp. 54–69, Jan. 1968.

55. *Proc. Conf. on the Use of Remote Sensing in Conservation, Development, and Management of the Natural Resources of the State of Alaska*, Juneau, Alaska, 1970.

Reprints of this article (No. X71-076) are available to readers. Please use the order form on page 9, which gives information and prices.

Warren L. Flock (M) is a professor of electrical engineering at the University of Colorado. He received the B.S. degree from the University of Washington in 1942 and the M.S. degree from the University of California, Berkeley, in 1948, both in electrical engineering. He was awarded the Ph.D. in engineering by the University of California, Los Angeles, in 1960. Prior to his present position, he was a professor of geophysics at the Geophysical Institute of the University of Alaska in 1960–64. Earlier he was an associate engineer and lecturer in the U.C.L.A. Department of Engineering and from 1942 to 1945 he was a staff member with the Radiation Laboratory at the Massachusetts Institute of Technology. In the summer of 1970 he returned to the University of Alaska as NSF Visiting Professor of Geophysics.

Dr. Flock's special interests in research and teaching involve atmospheric effects on electromagnetic waves; the use of electromagnetic waves in investigations of the ionosphere, troposphere, and solid earth; radar ornithology; and the role of electrical engineering and geophysics in dealing with the environment. In the fall of 1970 he taught a special environmental course for electrical engineers at the University of Colorado. A member of Tau Beta Pi and Sigma Xi, his other membership affiliations include the American Geophysical Union, the American Association for the Advancement of Science, the Sierra Club, and the National Audubon Society.

New product applications

Complementary-symmetry MOS circuits are described in a new manual

Intended primarily for circuit and system designers, the new COS/MOS Integrated Circuits Manual (CMS-270) will also be useful for hobbyists and others using semiconductor devices and circuits. Information on device physics, construction, theory of operation, and ratings for monolithic circuits containing p- and n-channel MOS transistors are included.

Noise immunity, power supply considerations, and methods of interfacing these devices with other logic forms are detailed in 160 pages.

The integrated circuits discussed are monolithic types batch-fabricated on silicon wafers. The number of chips produced, depending on the circuit complexity, varies from about 1000 to 250. By integrating compatible p-channel and n-channel enhancement-type field-effect transistors on a monolithic substrate, the complementary-symmetry circuit advantages can be used in IC design.

One section of the manual provides charts that display the logic diagram and package choice for each gate, flip-flop, and MSI device in the CD4000 and CD4000A series. The former group is designed to operate from a supply voltage of 5 volts minimum to 15 volts maximum. The CD4000A group operates from a supply voltage of 3 volts to 15 volts.

Included with other applications information is that showing how to limit extraneous voltages to safe levels under all operating conditions.

The input circuit designs are shown in Fig. 1. Because of the low *RC* time constants of these circuits, they have no noticeable effect on circuit speed and do not interfere with logic operation. For circuits that contain gate-protection circuits, the power-supply voltage V_{DD} (the most positive dc supply voltage) should not be turned off while a signal from a low-impedance pulse generator is applied to any of the inputs to the integrated circuit.

The reason for the restriction becomes apparent by inspection of Fig. 1(A) where the V_{DD} line is essentially grounded so that a positive voltage input from a pulse generator is impressed across the diodes D_2. The voltage, between 3 and 15 volts, could cause permanent damage to the diodes or could burn out the V_{DD} metalization.

In any system design, therefore, if any input excursion is expected to exceed $+V_{DD}$ or fall below $-V_{SS}$, the current through the input diodes should be limited to 50 milliamperes to assure safe operation. The symbol V_{SS} is the most negative dc supply.

Application notes listed at the back of the manual, covering specific subjects, can be obtained on request. These publications include papers on noise immunity, counter and register design, gate-oxide protection circuits, astable and monostable oscillators, power-supply considerations, interfacing with other logic families, and arithmetic arrays.

The manual is available at $2.50 from RCA, Electronic Components, Harrison, N.J. 07029.

Circle No. 60 on Reader Service Card

FIGURE 1. Design of input circuits that limit extraneous voltages to safe levels.