VOLUME 5 □ NUMBER 2 FEBRUARY 1968

IEEE spectrum

features

Departments: *please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

Index-gradient deflectors are among the two types of electrooptic light-beam deflection devices described in the article that begins on page 45. The cover illustrates the system of alternately charged parallel rod electrodes that create the radially increasing field strengths required to implement the basic technique of this particular method.

We fear that in the past on this page we have neglected the analytical techniques in use by the up-to-date systems engineer. We hope this month's contribution, written by Stan Gery, Consultant in our Radar Techniques and Advanced Development Department, will help even things up.

The generation of correlated multivariate samples for Monte Carlo simulation

Monte Carlo simulation is often the most effective approach to the study of complex or nonlinear signal processing systems. Basically, a suitable sequence of simulated inputs is generated and the consequent response of the system model is observed. The use of psuedo-random number sequences in simulating the input process permits parametric studies to be made with identical stochastic samples. In order to provide universal applicability, psuedo-random number generators are ordinarily designed for statistical independence between numbers (references 1 and 2). However, in at least two types of simulations, the need arises for controlled correlation between elements of the psuedo-random sequences (references 3 and 4).

One type involves modeling of multiport networks such as tapped delay lines or phased arrays for which sequences of independent "looks" are required. Each "look" requires a multivariate process to be sampled. In the case of a gaussian process, the covariance matrix and means vector provide the basis for transforming a "batch" of n zero-mean, unity-variance, independent gaussian numbers into the desired multivariate sample.

To start this formulation, let $X = (x_i)$, $i = 1, 2, \ldots, n$ denote the set of independent, zero-mean normal variates having the diagonal covariance matrix $W = (w_{ij})$, where $w_{ij} = E(x_i x_j)$. Similarly, the required set is denoted by $Y = (y_i)$, $i = 1, 2, \ldots, n$ having the specified covariance matrix $R = (r_{ij})$, where $r_{ij} = E(y_i y_j)$. Letting $Y = GX$, wherein G is the desired transformation matrix, then $R = GWG'$. Since W is diagonal with unity elements, then:

$$R = GG' \qquad (1)$$

As shown by Marsaglia (reference 5), there are 2^n triangular nxn matrices, G, which solve equation 1. The following subset of these solutions is found by assuming the triangular transform $G = (g_{ij})$ with:

$$y_1 = g_{11}x_1$$

$$y_2 = g_{12}x_1 + g_{22}x_2$$

$$y_3 = g_{13}x_1 + g_{23}x_2 + g_{33}x_3 \qquad (2)$$

$$\vdots$$

$$y_k = g_{1k}x_1 + g_{2k}x_2 + \ldots\ldots + g_{kk}x_k$$

Working then with the elements of the specified covariance matrix yields:

$$E(y_1y_1) = r_{11} = g_{11}^2$$
$$g_{11} = \sqrt{r_{11}}$$

$$E(y_1y_2) = r_{12} = g_{11}g_{12}$$
$$g_{12} = r_{12}/g_{11}$$

$$E(y_2y_2) = r_{22} = g_{12}^2 + g_{22}^2$$
$$g_{22} = \sqrt{r_{22} - g_{12}^2}$$

$$E(y_1y_3) = r_{13} = g_{11}g_{13}$$
$$g_{13} = r_{13}/g_{11}$$

$$E(y_2y_3) = r_{23} = g_{12}g_{13} + g_{22}g_{23}$$
$$g_{23} = (r_{23} - g_{12}g_{13})/g_{22}$$

$$\vdots$$

etc.

The transformation matrix is thus given by (reference 6):

$$g_{11} = \sqrt{r_{11}}$$

$$g_{1j} = r_{1j}/g_{11}$$

$$g_{ii} = \sqrt{\left(r_{ii} - \sum_{m=1}^{i-1} g_{mi}^2\right)},\ i > 1 \qquad (3)$$

$$g_{ij} = \left(r_{ij} - \sum_{m=1}^{i-1} g_{mi}g_{mj}\right)/g_{ii},\ j < i$$

$$= 0,\ i > j$$

A second type of correlation requirement arises in modeling a time series such as from a filter or aperture that is excited by white gaussian noise. An indefinitely long sequence of samples is usually needed. Generally these samples are equally spaced in time. Practically they become statistically independent when they are more than n samples apart, where n is very small compared to the total number of samples. In generating this type of sequence from a psuedo-random number generator, recourse is made to the multivariate conditional density (reference 7):

$$p(y_n/y_1, y_2, \ldots, y_{n-1}) = K \exp - [(y_n - m_n)^2/2\sigma^2] \qquad (4)$$

where

$$m_n = \sum_{k=1}^{n-1} C_{nk}y_k/C_{nn}$$

$$\sigma = \sqrt{C/C_{nn}}$$

C = determinant of R, the covariance matrix

C_{ij} = ij^{th} cofactor of R

In starting this sequence the "batch" type of transformation is applied to generate the first n − 1 samples of y from an equal number of x's. Subsequent samples are formed by scaling the unity variance generator and adding the mean value, namely,

$$y_j = \sigma x_j + \sum_{k=1}^{n-1} (C_{nk}/C_{nn})y_{j+k-n},\ j \geq n. \qquad (5)$$

References

1. Muller, M. D., "A Comparison of Methods for Generating Normal Deviates on Digital Computers," *J. Assoc. Comput. Mach.*, Vol. 6, pp. 376–383; 1959.

2. Chambers, R. P., "Random-Number Generation on Digital Computers," *IEEE Spectrum*, Vol. 4, pp. 48–56; February 1967.

3. Stein, S. and Storer, J. E., "Generating a Gaussian Sample," *IRE Trans. on Information Theory*, Vol. IT-2, pp. 87–90; June 1956.

4. Levin, J. J., "Generation of a Sampled Gaussian Time Series Having a Specified Correlation Function," *IRE Trans. on Information Theory*, Vol. IT-6, pp. 545–548; December 1960.

5. Marsaglia, G., "A Note on the Construction of a Multivariate Normal Sample," *IRE Trans. on Information Theory*, Vol. IT-3, p. 149; June 1957.

6. Faddeeva, V. N., "*Computational Methods of Linear Algebra*," Dover, New York; 1959.

7. Cramer, H., "*Mathematical Methods of Statistics*," Princeton University Press; 1961.

IEEE spectrum

Transients and trends

The 38 000 engineers currently serving on active duty in the U.S. Armed Forces represents a greater number than the entire graduating class of 1967 from all the nation's engineering schools. This startling figure is taken from a late survey just reported by the Engineering Manpower Commission of the EJC.

The Air Force, with more than 14 000 engineers on duty, is the largest user of engineers in the military. The Navy has more than 10 000 on board, and the total number in the Army is not far behind that figure.

By level of academic degrees, the services have 255 engineers holding doctorates, and more than 7800 with master's degrees.

The EMC report, "Engineers in Uniform," reveals that bachelor-degree graduates in all three services are faced by something of a dilemma, since the armed forces generally do not identify engineering requirements below the master's degree level. Thus the engineer without an advanced degree must serve either in a nontechnical capacity, or fill a position intended for an officer with at least a master's degree. The services have requirements for more than 15 500 advanced-degree engineers, but have only 8500 men with these qualifications.

Electrical engineers are the largest single category in the armed services, followed by the mechanical, civil, and aerospace groups, in that order.

Electric furnaces will produce nearly half of the United States' steel output within 25 years, according to the Battelle Memorial Institute.

Computers will control the charge makeup, power input, alloy additions, and the tapping of electric furnaces of 300-tonne—and possibly larger—capacities. Most of the low- and high-alloy steels will be made in electric furnaces.

The open-hearth furnace, despite its impressive record, is obsolescent and on the way out. During the past decade, open-hearth production in the United States dropped from a high of 90 percent to 52 percent of the total output; and, it is predicted that within 15 years the open-hearth furnace will be gone. The reason for this is the speed and efficiency of the electric furnace in the steelmaking process.

In 1967, approximately 15 percent, or 20 million tonnes, of U.S. steelmaking capacity was attained by the electric furnace. Another 17 million tonnes of electric furnace capacity is presently being built.

Total sales of electric power to all categories of customers in 1967 were an estimated 1.1 trillion kWh, an increase of 6.4 percent over the 1966 total of 1.04 trillion kWh, according to statistics of the Edison Electric Institute.

About 1.3 million new customers were served by the industry in 1967, to bring the total number of customers to about 68.2 million. Of this number, some 53.5 million were served by investor-owned utilities at the end of last year, as compared with 52.6 million in 1966. About 95 percent of the new customers were in the domestic, or residential, category.

The average use of electric power in the American home rose to a record high of 5565 kWh during last year, an increase of 300 kWh over 1966. At the same time, the average price of a kilowatthour of residential electric service dropped to a record low of 2.16 cents. It was 2.20 cents in 1966.

By the end of 1972, generating capability of the total electric utility industry in the contiguous United States is expected to reach 392.2 million kW, according to the 42nd Semiannual Electric Power Survey just published by he EEI.

Previously, these semiannual surveys have presented forecasts of capabilities, peak loads, and annual energy for the current year and the following three years. But, starting with the present survey, the forecasts will cover the current year and the five following years.

The predicted 1972 capability is some 123 million kW greater than that for 1967, and represents an average annual increase of 7.9 percent for the five-year period. It will provide a margin of reserve of more than 30 percent above the expected December peak load.

The predicted summer capability of 384 million kW for 1972 will provide a 20 percent margin of reserve above the summer peak load.

The power surveys of the Edison Electric Institute have been made since 1947 in cooperation with power systems and the manufacturers of heavy power equipment.

Industry sales of all electrical products are expected to set even greater records in both 1967 and 1968 than previously reported, according to revised figures from the National Electrical Manufacturers Association (NEMA).

The new estimate for the value of industry shipments in the U.S. during 1967 is $39.58 billion—an increase of 6.1 percent above the 1966 total of $37.30 billion.

NEMA also forecasts a seventh consecutive record-breaking year for 1968, in reporting that the industry's value of shipments will hit an all-time high of $41.87 billion.

The nuclear power industry in the United States had a spectacularly good year in 1967, according to the Atomic Industrial Forum, Inc. The year ended on the upswing with news that Consumer Power Co. would build a $267 million two-reactor complex in Michigan to generate 1300 MW of electricity plus steam for industrial purposes. By mid-December 1967, U.S. utilities had placed orders for 30 nuclear units with a combined generating capacity of about 25 000 MW.

The 1967 orders were placed by utilities of all types: municipal-owned, cooperative, and investor-owned. Their service areas include all geographical regions (except the major gas- and oil-producing states): the Rocky Mountain states, the Pacific Northwest, Hawaii, and Alaska. Many of the sites chosen for the new units are in the East.

IEEE forum

Correspondence relating to activities and policies of the IEEE

About those meetings!

It is interesting to find articles on "meetingitis" and "Chapter apathy," both in one issue of IEEE SPECTRUM. Many IEEE members experience a great deal of competition for their time these days, having many commitments to the community. So the Boys Club or the Church Finance Committee or the Theatre Guild may all vie for the engineer's evenings.

But let us consider that Mr. IEEE has definitely reserved a spot for that Chapter meeting. On the day of the affair he puts in his usual eight or ten hours on the job. He then must rush to catch a train, or to fight traffic by car, to get to the premeeting dinner. After a cocktail, or two, and then a full course meal, it is close to 8 P.M. and he joins the group in the lecture hall. Settling into one of the folding chairs, he suddenly feels everything catch up with him and he has an urge to fold up for a nap.

When the speaker of the evening takes his place at the lectern, he spreads his "paper" before him and, if he is a typical engineer, reads the pages in a monotone, leaving out any trace of animation or humor. From time to time he will call for illustrations or graphs, and the lights will be dimmed and the projector turned on. Frequently, the typed figures will be too small to see and portions may be off the screen.

After what seems an interminable period, the speaker will end his reading and Mr. IEEE will shift in his chair again and look forward to getting out into the air. But alas, the question session begins—and there are always a few individuals who are unable to ask a simple question. Perhaps a half hour passes before the meeting finally slows down to a stop. Mr. IEEE then waits in line another half hour to get his car. By 11:30 he is home, weary, still a bit stuffed from dinner, and partially asphyxiated from the tobacco smog of the lecture hall. As for the paper—what did the man say? At this point, who cares!

Perhaps our illustration has been a bit overdrawn, but things perilously close to this actually *do* happen at Chapter meetings.

What can we do about it?

Somehow, the deadly dull procedure must be changed without resorting to the introduction of belly dancers or other distractions.

First and foremost, the subjects of the papers must be carefully chosen. If possible, they should be items of general interest that have novel and exciting features. The speaker should be required to present his subject in "abstract" form. Ultimate results of the project are obviously most important, but the lecturer should present sidelights, including experiences encountered in doing the work—wrong guesses, blind alleys, and curious quirks exhibited by the apparatus. In any project there are moments of frustration, uncertainty, aggravation, humor, elation, surprise. Injecting anecdotal bits from time to time will get the listener's mind off of the extra dessert he shouldn't have consumed, or the ever-present factor of that "damned hard chair."

All of this can be done without sacrificing one iota of technical information. There are times when it may even be well to eliminate the question period. Each Chapter chairman knows his audience and, if the masters of "exposition" are on hand, he will adjourn the evening's meeting as the speaker's last words die away.

There *are* brilliant technical men who can keep the attention of the listeners. In the hour or so at their disposal, they do not try to tell the audience everything they know. Most important, these men have a keen sense of the "law of diminishing returns," which exists in any audience.

Alas, there are too few such men in the professions. Most of us stuck to our required courses in school and left the behavioral arts strictly alone. As a result, engineers are, as a class, indescribably dull.

Even so, if our Chapter meetings are to be successful, in view of the competing factors mentioned earlier, we've got to make them more palatable. Otherwise, the members will simply not make the sacrifice required in our frenetic existence to come to the meetings.

E. E. Pearson
Leeds and Northrup Co.
Philadelphia, Pa.

Spectral lines

Our new editors. An invaluable asset to a society such as the IEEE is the time and effort voluntarily contributed by its members. The Institute could not function without this contribution, the value of which certainly is not measurable in dollars.

A very significant group of volunteers are the editors of Institute publications—some 40 editors for our periodicals, together with 80 associate editors and the various guest editors and editorial panels who assist us in developing special issues. In addition, more than 1000 IEEE members serve on editorial boards and as reviewers of papers submitted by prospective authors. Without their participation, the Institute would not be able to maintain appropriate technical and editorial standards for its publications.

Now, with the start of a new calendar year, we are welcoming two new editors to IEEE's general publications. After serving as editor of IEEE SPECTRUM for the past two years, C. C. Cutler of the Bell Telephone Laboratories has asked to be relieved of his editorial duties. His successor, J. J. G. McCue of M.I.T.'s Lincoln Laboratory, was appointed by the IEEE Board of Directors at its January meeting. He comes to this appointment assured of the continuing support and cooperation of the SPECTRUM staff, under its managing editor, Ron Jurgen.

Before Chap Cutler took over as editor, SPECTRUM's activities were directed by the Publication Board of the Institute. The major responsibility for SPECTRUM's development rested on Woody Gannett, and on Ron Jurgen and the other members of the SPECTRUM staff. During this period the publication began to take shape and to assume a character of its own. As SPECTRUM's first editor, Chap Cutler was able to continue this growth and to improve the magazine's content with the help of a small but active editorial board. Jerry McCue now has the opportunity to further this development so that SPECTRUM's usefulness to the membership will continue to grow.

Our second new editor is Prof. Frank Barnes, who is chairman of the Electrical Engineering Department at the University of Colorado. As the new editor of the STUDENT JOURNAL, he replaces Ted Hunter, who founded the publication in 1954 and remained its editor until 1966. The role of the STUDENT JOURNAL as an IEEE publication has been the subject of much recent study. It was originally established by the IRE at a time when the IRE PROCEEDINGS was the only publication available to the organization's student membership. Because of its high technical level, the PROCEEDINGS was not geared to student readership, and Ted Hunter, recognizing the need for a more appropriate publication for the Student members, convinced the IRE Board of Directors of this fact—which resulted in the founding of the JOURNAL. The situation in the IEEE is different, however, since SPECTRUM is written at a technical level that can be understood by most undergraduates. Realizing this, the IEEE Board of Directors voted to give Student members a choice as to which publication they wish to receive. Currently available data indicate that 55 percent of the new Student members and 42 percent of continuing Student members have selected the STUDENT JOURNAL in preference to IEEE SPECTRUM.

The student publication program of the Institute is now in a state of evolution. With the establishment of a new Educational Activities Board[1] under the chairmanship of Dr. John Shive of Bell Telephone Laboratories, the IEEE is in a better organizational position to evolve effective programs for its students, and Frank Barnes, as editor of the STUDENT JOURNAL, will play a big role in developing the publications part of such programs. He will have the help of the Publication Board, as well as the Educational Activities Board and its Student Activities Committee. Alex McKenzie of the Institute staff will continue as managing editor of the STUDENT JOURNAL.

Both Jerry McCue and Frank Barnes are undertaking significant new responsibilities for the Institute. In wishing them well, we should also remember to thank their predecessors, Chap Cutler and Ted Hunter, for the effective way in which they have completed this particular phase of their service to the Institute.

F. Karl Willenbrock

1. "IEEE educational activities—an ad hoc committee report," *IEEE Spectrum*, vol. 4, pp. 101–104, Nov. 1967.

Authors

Electrooptic light-beam deflection (page 45)

James F. Lotspeich is a research physicist and member of the technical staff of Hughes Research Laboratories, with which he has been affiliated for the past 11 years. His earlier work at Hughes was in the fields of ammonia beam maser research, ammonia chemistry, and chemisorption, in connection with a sealed-off portable maser, and in exploratory research on new materials for gas lasers. For the past four years he has worked on applications of electrooptic materials to light modulation, beam deflection, and focusing techniques. He received the B.A. degree from Princeton University in 1943, the M.S. in physics from the University of Cincinnati in 1949, and the Ph.D. in physics from Columbia University in 1958.

Prevention of power failures—The FPC report of 1967 (page 53)

Gordon D. Friedlander. A biographical sketch of Mr. Friedlander appears on page 111 of the Feb. 1965 issue.

Color television standards in Region 2 (page 62)

Charles J. Hirsch (F) received the A.B. and E.E. degrees from Columbia University in 1923 and 1925 respectively. He served as chief engineer of various companies in France, Italy, England, and the United States. In 1941 he joined Hazeltine Corp., where he worked on development of military and navigation systems and color television; he became executive vice president in 1956. In 1959 he joined RCA, where he devoted his time to the Home Instrument and Custom Aviation Divisions. He recently retired from RCA but continues as consultant. He was secretary of Panel 13 of NTSC, which developed the color television standards now in use in the U.S.A., Canada, and Japan. He was chairman of several U.S. delegations to CCIR meetings.

Human enhancement: Beyond the machine age (page 79)

Warren M. Brodey and Nilo Lindgren. Biographical sketches of Mr. Brodey and Mr. Lindgren appear in the Sept. 1967 issue (page 45) and the March 1965 issue (page 196) respectively.

Education for the engineering mission (page 94)

William G. Shepherd received the B.S. degree in electrical engineering in 1933 and the Ph.D. degree in physics in 1937, both from the University of Minnesota. From 1937 to 1947 he worked for Bell Telephone Laboratories, principally on nonlinear microwave circuits and electron tubes. He returned to the University of Minnesota in 1947 as professor of electrical engineering. He founded the university's Physical Electronics Research Laboratories and served as its director. After serving as associate dean of the university's Institute of Technology and head of the Department of Electrical Engineering, he was appointed vice president for academic administration in 1963. An active member of IEEE, he served as IEEE President in 1966 and is now Secretary of the Institute.

Silicon device technology (page 101)

D. H. Roberts received the bachelor's degree in physics from Manchester University (England) in 1953. He then joined the Caswell Research Laboratories of the Plessey Company as a research physicist. His initial work was on photoconductivity in lead selenide films and on III-V compounds. Subsequently he was responsible for setting up the first U.K. research team devoted to SIC technology. In 1961 he became technical director of Semiconductors Ltd., a Plessey subsidiary. In 1965 he became head of Solid State Research and Development at the Caswell Laboratories and in 1967 was appointed general manager of the Plessey Semiconductor Division. He is the author of a number of technical papers on photoconductivity, microelectronics, and galvanomagnetic devices.

Electrooptic light-beam deflection

Advances in laser technology have engendered new problems in deflection-control design. The solutions offered here are those that have surmounted the formidable obstacles presented by optical loss and bandwidth limitation

James F. Lotspeich *Hughes Research Laboratories*

The impetus given display, recording, computer, and space technologies by the revolutionary discovery of the laser has created an ever-increasing need for controlled optical-beam deflection techniques. Those that have been developed thus far, usually employing mechanical or acoustical methods, have had either high insertion loss or limitation on bandwidth. Electrooptic methods have not only overcome these problems, but have successfully met the demands of systems requiring aperiodic and rapid-access modes of operation. After the basic theory of optical deflection, resolution, and control are outlined, two experimental models of electrooptic beam deflectors, as well as the two major difficulties that have been encountered, are described.

Since the advent of the continuous-wave laser[1] in 1961, there has been a growing interest in the development of techniques for controlled optical-beam deflection, with application to such systems as optical displays, photographic recording, computer memory devices, and space tracking and acquisition. For these and other exciting uses, the outstanding characteristics of the laser beam, as compared with more standard sources of optical power, are its high degree of directionality, single-wavelength output, high intensity, and ability to be focused to an extremely small spot size. Within the past few years, experimenters have investigated and developed a variety of beam-deflection schemes[2] employing mechanical, acoustical, or electrooptic methods; several of these have been reduced to practical, operational devices, primarily for television-type displays.[3,4]

The electrooptic deflection techniques that I am considering here will be restricted to those involving the bending of light beams *within* an active electrooptic medium. Thus I am omitting from this presentation the very interesting and rather well-developed digital scheme that utilizes a linear chain of electrooptic binary polarization switches and passive birefringent crystals to produce a regular array of discrete beam positions.[5] This type of device has fairly high deflection-speed capabilities in a random-access mode of operation, but it possesses an appreciable built-in optical loss and is therefore best adapted to high-density optical memories or special display and photographic printing devices requiring very modest levels of optical power.

Most of the deflection techniques that have been explored to date have a somewhat restricted scope of application, either because of high optical insertion losses or because of bandwidth limitations imposed by inertial effects inherent in the deflection mechanism. In these two respects, the electrooptic devices discussed here have an advantage. Their design allows for a low optical loss, and the active material is essentially a pure dielectric having a frequency response extending well into the microwave region.[6] Thus, for example, they are well adapted to large-screen, high-brightness tactical displays that call for selective or aperiodic access at microsecond slewing rates. Their chief disadvantage is in their high capacitance load, typically 500 to 1000 picofarads, and therefore they require low driver output impedances and kilowatts of driver power for wide-band operation.

I shall begin by describing the basic processes of light-beam bending by refraction, and will then show how the angle of deflection can be varied electrically. Following a brief discussion of the concept of optical resolution and the law governing it, I shall describe two experimental embodiments of beam deflectors that employ the electrooptic principle. Finally, mention will be made of the significant technological problem areas that have been encountered, and which impose performance limitations within practical applications.

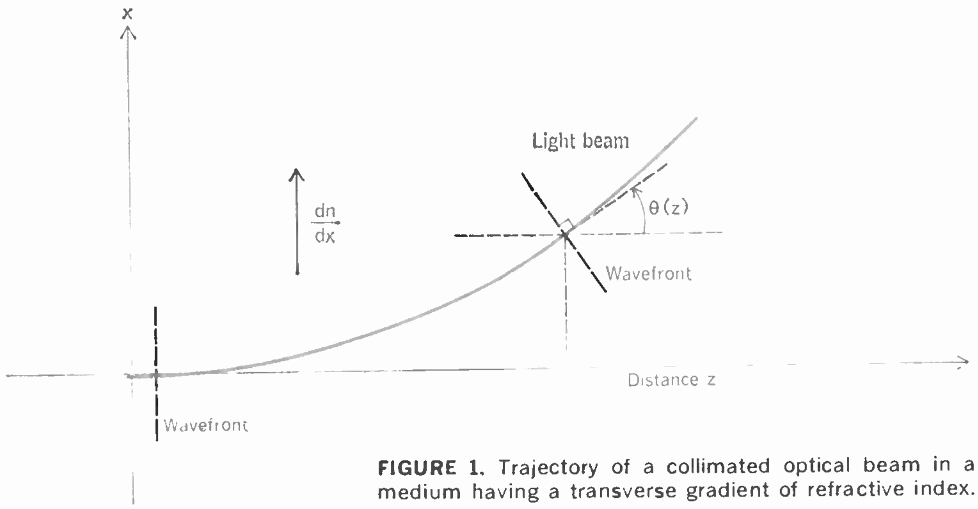


FIGURE 1. Trajectory of a collimated optical beam in a medium having a transverse gradient of refractive index.

Beam bending by refraction

The velocity of light in a transparent medium, v, is governed by a property of the medium called the index of refraction, which is usually designated by the symbol n and defined by the relation $n = c/v$, where c is the velocity of light in a vacuum. If n is everywhere constant, a plane wave of light propagates in a straight line that is normal to the wavefronts or surfaces of constant phase. If, however, there is a spatial index gradient that is transverse to the light path, the beam is bent continuously toward the region of higher index,[7] as illustrated in Fig. 1. The reason for this is that adjacent ray elements of the beam travel at slightly different velocities, with the result that the wavefronts gradually become more and more tilted, in relation to their initial orientation, as the beam propagates through the medium. A familiar example of this phenomenon is the mirage, which can occur when a warm air mass lies above a cooler mass. The cooler air mass has a greater density, resulting in a higher refractive index. Hence, light rays from an object (e.g., an oasis) which emanate more or less parallel to the earth's surface are bent toward the earth, giving a distant observer the impression that he sees the object floating in air above its true position.

The equation of a ray trajectory, giving the deflection angle θ as a function of distance z through the medium (see Fig. 1), is a solution of

$$\frac{d\theta}{dz} = \frac{1}{n}\frac{dn}{dx} \tag{1}$$

If n varies linearly with x within the region of deflection, and if its maximum change is very small relative to the value n_0 at $x = 0$, then we may let

$$\frac{1}{n}\frac{dn}{dx} = \frac{1}{n_0}\frac{dn}{dx} \equiv k = \text{constant} \tag{2}$$

and Eq. (1) integrates to

$$\theta(z) = kz \tag{3}$$

assuming $\theta(0) = 0$. An important consequence of this special case is that $\theta(z)$ is independent of the initial ray position relative to the z axis; therefore, an incident plane wave retains its plane-wave character.

Refractive beam deflection also can occur discontinuously at a sharp boundary separating two regions of different refractive index. This is demonstrated in Fig. 2(A), where a beam passes abruptly from a region of index n_1 into a second region of higher index n_2 and is instantly deflected toward the surface normal. The relationship between the angles θ_1 and θ_2 is governed by Snell's law, which takes the form[8]

$$n_1 \cos\theta_1 = n_2 \cos\theta_2$$

This law is independent of the beam direction. Anyone who has had the experience of thrusting a stick, or a canoe paddle, into water at an angle to the surface has observed that the portion of the object that is below the surface appears to be bent toward the surface. This is because light rays from the submerged object are deflected *away* from the surface normal upon emergence into air (whose index is less than that of water).

A familiar optical component is the isosceles prism, commonly made of glass ($n \approx 1.5$). Light passing through this prism is deflected twice, once at each surface. This is shown in Fig. 2(B) for the special case in which the internal beam path is parallel to the prism base. The total deflection angle α is related to the refractive indexes, n_1 and n_2, and prism apex angle β by the formula[8]

$$n_1 \sin \tfrac{1}{2}(\beta + \alpha) = n_2 \sin \tfrac{1}{2}\beta \tag{4}$$

It is obvious from Eq. (4) that if the refractive index n_2 of the prism can be varied slightly, a corresponding variation of the deflection angle α will occur. Likewise, in the case of beam bending by index gradient, Eq. (3) informs us that the angle of deflection θ can be varied by controlling the magnitude of dn/dx, or k. The electrooptic effect offers a workable means of accomplishing these controlled perturbations.

The electrooptic effect

When application of an electric field to an optical medium results in a perturbation of its refractive properties, the phenomenon is called the electrooptic effect. In liquids and some solids, where the changes in refractive indexes exhibit a quadratic dependence upon

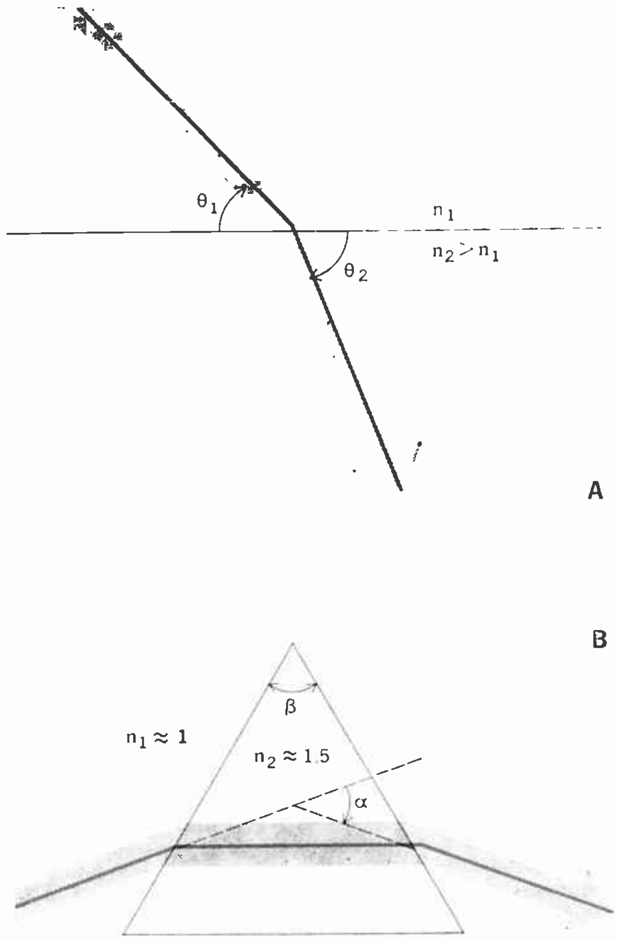


FIGURE 2. A—Abrupt deflection of a collimated light beam by refraction at a boundary surface separating two regions of different refractive index. B—Double deflection of a collimated light beam by an isosceles prism having an index of refraction greater than the surrounding medium.

the field, it is designated the Kerr effect.[8] Crystalline solids that are piezoelectric display a linear relationship known as the Pockels effect.[9]

As a result of their linear electrooptic response, low conductivity, high optical transparency throughout the visible spectrum, high optical quality, and availability in large quantities, the materials that are of most interest to investigators are the tetragonal crystals, potassium dihydrogen phosphate (KDP), and the more sensitive deuterated isomorph, potassium dideuterium phosphate (KD*P). In these crystals, there exists a unique axis that represents the preferred direction for application of the electric field. This is the so-called optic axis or c axis, designated by crystallographers as the ⟨001⟩ direction. If a beam of plane-polarized light traverses the crystal along the ⟨1$\bar{1}$0⟩ direction perpendicular to the electric field, and if its polarization is also perpendicular to the field in the ⟨110⟩ direction, it experiences a change of refractive index as the field is varied. The dependence of index on the applied electric field is expressed in the form[10,11]

$$n = n_0 \pm \tfrac{1}{2} n_0^3 r_{63} E \qquad (5)$$

where

n_0 = unperturbed refractive index
r_{63} = electrooptic modulus (Pockels coefficient)
E = applied electric field

Note that the sign of the change in n follows the polarity of the field, which is taken to be positive when directed along the positive c axis.

Both KDP and KD*P belong to a class of materials known as ferroelectrics.[12,13] Above a certain critical temperature, the Curie point, these crystals exist in a paraelectric phase in which the electrooptic modulus r_{63} depends inversely upon temperature according to the relationship

$$r_{63}(T) = r_{63}(T_r)\left(\frac{T_r - T_c}{T - T_c}\right) \qquad (6)$$

where T_r, T_c, and T are the room, Curie, and operating temperatures, respectively. The dielectric constant associated with the c axis exhibits a similar temperature dependence. The Curie point of KDP is 122°K[14]; that of fully deuterated KD*P is 222°K.[15] At room temperature, the magnitude of the refractive index change,

$$\delta \equiv |n - n_0| = \tfrac{1}{2} n_0^3 r_{63} T_r |E| \qquad (7)$$

is only of the order of 10^{-4} for acceptable limits of field strength. It is therefore advantageous, when practicable, to utilize these materials at temperatures close to their Curie points in order to obtain enhanced sensitivity, especially in the case of the deuterated species, since temperatures near the Curie point are more readily attained. An order-of-magnitude increase in the maximum value of δ can be anticipated with this technique.

Optical resolution

A direct measure of the information display capabilities of a beam-deflecting device is the number of *distinguishable* positions the beam can assume. It is important that we be familiar with the optical laws governing this performance characteristic.

A collimated optical beam emanating from a source of limited aperture cannot propagate very far without spreading because of diffraction. If this beam is brought to a focus, diffraction effects set a lower limit on the spot size. One might ask, what is the minimum angle through which the beam must be deflected such that the human eye can resolve two spots? For the case of a circular aperture and constant intensity over the area, Rayleigh's criterion[5] states that the spots are resolved when they are separated by one spot radius, and the minimum angle of resolution θ_1 is given by

$$\theta_1 = \frac{1.22\lambda}{D}$$

where λ is the wavelength and D is the aperture diameter (or diameter of the focusing lens, whichever is smaller). For a beam with a Gaussian radial intensity distribution, Rayleigh's criterion yields

$$\theta_1 = \frac{4\lambda}{\pi D} = \frac{1.27\lambda}{D}$$

if we *define* the beam radius $R = D/2$ as the radial distance from the beam center to the point where the intensity is e^{-4} of the maximum, and if the spot radius is similarly defined. (It should be noted that 98 percent of the beam power falls within this radius.) Actually, this angle is a little bit larger than is necessary. A satisfactory approximate formula for minimum angular resolution, applicable to both constant intensity and Gaussian beams, is given by

$$\theta_1 = \frac{1.2\lambda}{D}$$

If the total angular range of an optical beam deflector is γ, the number N of resolvable elements encompassed by γ is

$$N = \frac{\gamma}{\theta_1} = \frac{\gamma D}{1.2\lambda}$$

When a telescope, or any system of passive optics, is invoked to magnify the range of γ, the beam diameter is reduced in the same proportion; that is, the product

$$\gamma D = 1.2\lambda N \tag{8}$$

is invariant.[7] Thus a true measure of performance of a beam deflector is not its angular range but rather its maximum *deflection* measured in terms of *resolvable beam positions*.

Experimental models

Multiple-prism deflector. In Fig. 2(B) it was shown how a beam of light is deflected upon traversing a prism; Eq. (4) relates the deflection angle α to the prism apex angle β and refractive index n_2 for the special case in which the internal beam path runs parallel to the prism base. It was pointed out that small changes in n_2 by electrooptic means can be used to produce similar fluctuations in the angle of deflection. This technique has been demonstrated using the quadratic electrooptic effect in barium titanate[16] at elevated temperatures just above its Curie point, 404°K, and in potassium tantalate-niobate[17] at room temperature. These and other similar materials exhibit index perturbations of several parts in 10^3 when operated at acceptable voltage levels within the indicated temperature ranges. Their availability in reasonable quantities of good optical quality is extremely limited, however, and they require substantial dc bias voltages to provide an effective linear dynamic range.

The electrooptically induced change of refractive index in KD*P has a practical upper limit of about one part in 10^3 under low-temperature operation. Because the deflection capability of an optical device, as previously defined, is proportional to the product of the acceptable beam diameter and angular range [Eq. (8)], even the largest available single prism of KD*P can offer only very limited performance as a deflector element. Since there is no great materials problem with KDP-type crystals, with regard to either quantity or optical quality, a way to achieve a large enhancement of deflection capability, and at the same time eliminate the zero-field deflection of a single prism, is to use many identical prisms alternately inverted and arranged in a row as shown in Fig. 3. The prisms are cut with the c axes perpendicular to the plane defined by h and L in Fig. 3, and are oriented such that the positive c axis of each successive prism is reversed. Application of a common electric field along the c axes (in the direction of the prism thickness d) yields refractive index perturbations in the L–h plane that are alternately plus and minus in successive prisms. The beam propagates along the crystallographic ⟨110⟩ direction as required.

The effect of this arrangement on a beam polarized along h is displayed in Fig. 4. Application of the electric field produces the indicated external deflection θ_e. The total accumulated internal deflection $\theta(L)$ can be regarded as the sum of the separate deflections α resulting from the m upright prisms of index $n_0 + \delta$ that are located in a medium of index $n_0 - \delta$. (In Fig. 4, $m = 4$.) The smallness of θ permits one to approximate that the beam is parallel to each prism base throughout the total active length, so that Eq. (4) is still applicable. Likewise, the smallness of δ and α allows us to express Eq. (4) in differential form, letting

$$n_1 = n_0 - \delta \qquad n_2 = n_0 + \delta$$

$$\sin \tfrac{1}{2}(\beta + \alpha) = \sin \tfrac{1}{2}\beta + \tfrac{1}{2}\alpha \cos \tfrac{1}{2}\beta$$

Solution of α for a single prism ($m = 1$) yields

$$\alpha = \frac{4\delta}{n_0} \tan \tfrac{1}{2}\beta$$

Thus the total interior deflection $\theta(L) = m\alpha$ is

$$\theta(L) = \frac{4m\delta \tan \tfrac{1}{2}\beta}{n_0}$$

Now, since β is the prism apex angle, examination of Fig. 4 reveals that

$$L = 2mh \tan \tfrac{1}{2}\beta$$

or

$$m \tan \tfrac{1}{2}\beta = \frac{L}{2h}$$

Therefore, $\theta(L)$ is expressible in terms not involving β or m:

$$\theta(L) = \frac{2L\delta}{n_0 h} \tag{9}$$

When the beam emerges into air, it undergoes an additional deflection due to an abrupt refraction at the crystal–air interface as exemplified in Fig. 2. Applying Snell's law, the external deflection angle θ_e is obtained from the expression

$$\sin \theta_e = n_0 \sin \theta(L)$$

Since the maximum attainable deflection by electrooptic means in practical devices of the types considered here is limited to a few degrees, we can approximate $\sin\theta = \theta$, and finally derive

$$\theta_e = \frac{2L\delta}{h} \tag{10}$$

It is interesting to note the similarity between Eq. (9) and the solution of Eq. (1) for $\theta(L)$ in the case of a constant index gradient:

$$\theta(L) = \left(\frac{1}{n_0}\frac{dn}{dx}\right) L$$

In effect, the iterated-prism structure is entirely equivalent to a region of constant transverse index gradient,

$$\frac{dn}{dx} = \frac{2\delta}{h}$$

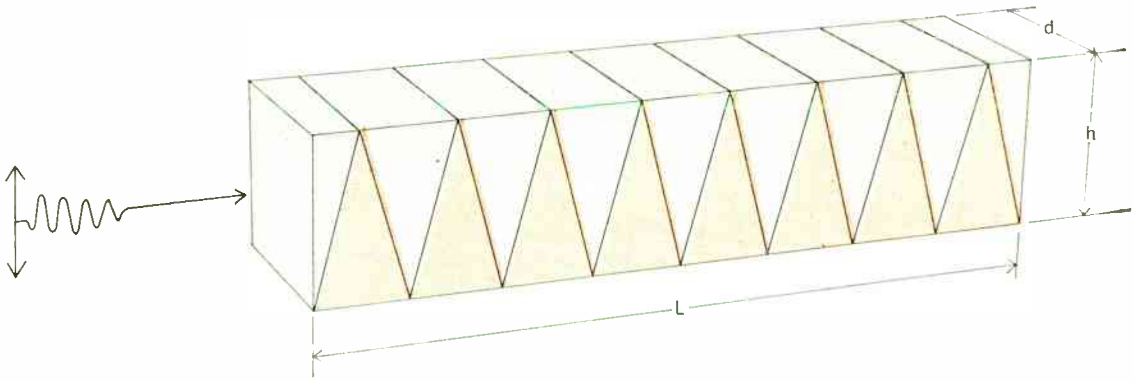


FIGURE 3. Iterated electrooptic prism deflector structure, where L = length, h = height, and d = thickness.

FIGURE 4. Trajectory of a light ray traversing an iterated electrooptic prism deflector.

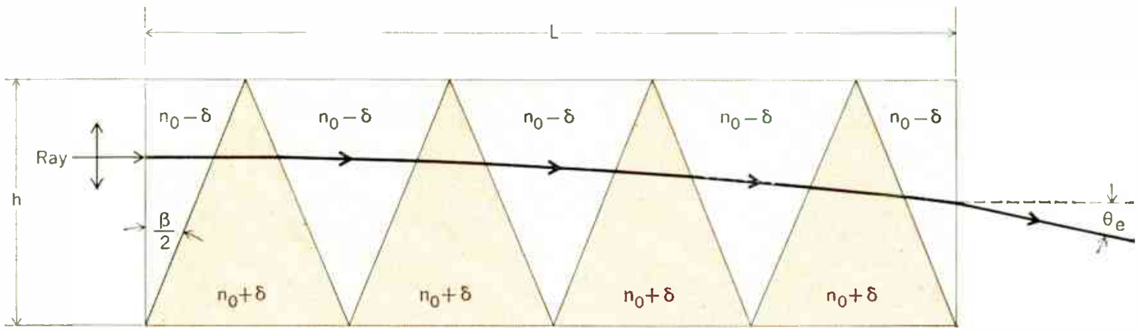
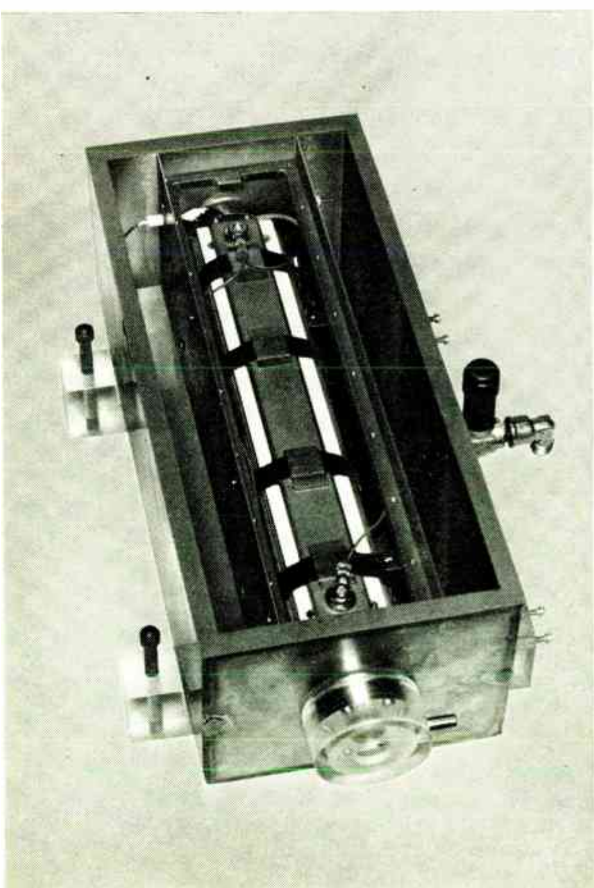


FIGURE 5. Low-temperature KD*P iterated-prism deflector.



Application of a voltage V_0 to a pair of parallel strip electrodes lying above and beneath the iterated-prism array produces a uniform electric field $E = V_0/d$ (see Fig. 3) in the desired crystalline direction. Expressing δ in terms of V_0,

$$\delta = \tfrac{1}{2} n_0^3 r_{63} E = \frac{n_0^3 r_{63} V_0}{2d}$$

Equation (10) then becomes

$$\theta_e = \left(\frac{L n_0^3 r_{63}}{hd}\right) V_0 \qquad (11)$$

The performance characteristics of an iterated KD*P prism beam deflector operated at low temperatures have been examined in considerable detail.[18] The experimental prototype contained a prism array having dimensions approximately 1 by 1 by 22 cm. Figure 5 shows the complete deflector unit, uncovered. The prism structure is contained within the interior chamber, which is in direct lateral contact with a temperature bath. The outside container, a double-walled evacuated stainless-steel cell, serves to isolate the complete interior region from room ambient conditions.

Some results of low-frequency ac deflection measurements as a function of peak-to-peak driver voltage at four selected operating temperatures are plotted in Fig. 6. The close proximity of data points to the straight-line approximations that have been drawn indicates an acceptable degree of deflection linearity with applied voltage. The measurements indicate an angular deflection sensitivity which is about 10 percent less than the calculated

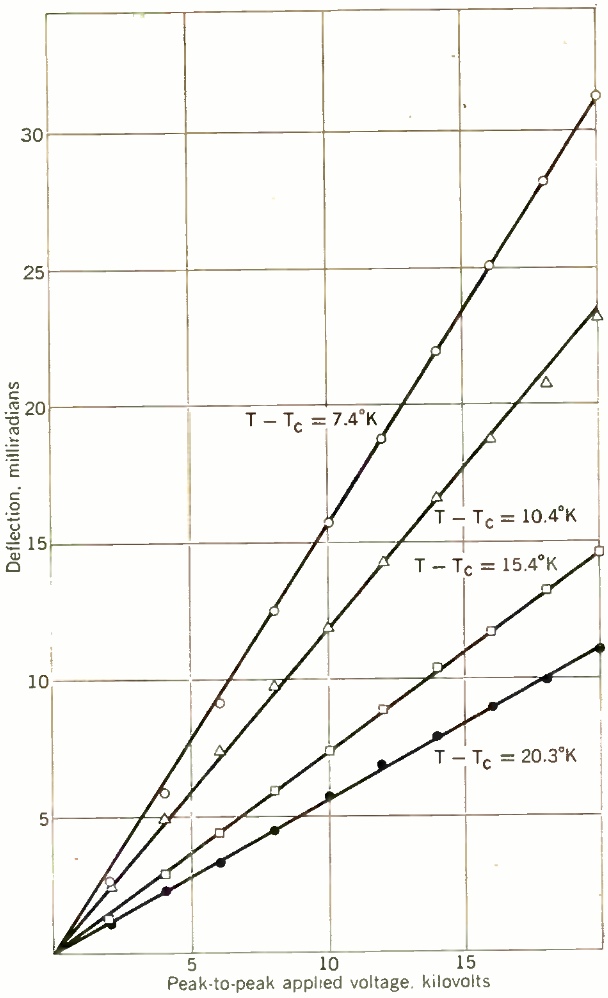


FIGURE 6. Results of 60-Hz ac deflection measurements with a KD*P iterated-prism system at selected temperatures near the crystal's Curie point.

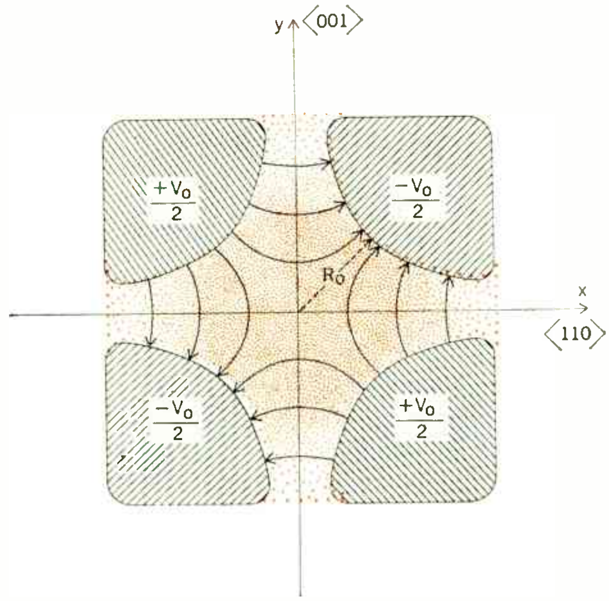


FIGURE 7. Cross-sectional schematic diagram of a quadrupolar array of parallel-rod electrodes used to produce a linear variation of refractive index in the *x* direction within a KDP-type electrooptic material. The proper crystallographic orientations of the ⟨001⟩ and ⟨110⟩ directions are indicated.

level. This is believed to be attributable to a physical "clamping" of the electrooptic crystals. This clamping causes a reduction in the effective electrooptic modulus, a portion of which is contributed by piezoelectrically induced strain[19] at frequencies below approximately 100 kHz.

Measurements of the resolving power of this deflector at a wavelength of 6328 Å gave results that were within 80 percent of the theoretical limit, indicating a moderate amount of optical distortion within the system. An upper limit on its deflection capability, with a maximum permissible peak-to-peak ac field[20] of 60 kV/cm, is estimated to be about 500 resolution elements at an operating temperature of $T_c + 10$°K.

To minimize optical losses resulting from a partial reflection at the many prism faces, a highly purified optical fluid of refractive index matching that of the KD*P was used. The measured optical transmissivity was approximately 75 percent at the low temperatures. This 1.2-dB optical insertion loss resulted principally from scattering by the cooled fluid.

Index-gradient deflector. Bending of light beams by index-gradient refraction in an electrooptic material such as KDP requires an electric field that varies linearly in a direction transverse to the beam path, as dictated by Eqs. (2), (3), and (5). Such a field distribution is achieved by use of a system containing four parallel rod electrodes arranged in a symmetric quadrupolar array.[21] The rods are embedded in the active medium and alternately charged positively and negatively from a variable high-voltage source V_0, as illustrated by the cross-sectional schematic diagram of Fig. 7. If the electrodes are carefully shaped such that their inner surfaces coincide with the hyperbolic curves

$$xy = \pm\frac{R_0^2}{2} \qquad R_0 \approx 1 \text{ cm}$$

then the potential distribution V relative to the indicated coordinate system is accurately represented by

$$V = -\frac{V_0}{R_0^2}xy \tag{12}$$

at least within the cylindrical region bounded by R_0. From the potential expression of Eq. (12), the electric-field components are immediately derived:

$$E_x = -\frac{\partial V}{\partial x} = \frac{V_0}{R_0^2}y$$

$$E_y = -\frac{\partial V}{\partial y} = \frac{V_0}{R_0^2}x \tag{13}$$

The configuration of the field lines is outlined in Fig. 7, as well as the required crystalline orientation of the electrooptic material. We see from Eq. (13) that the component E_y of the electric field along the preferred ⟨001⟩ direction exhibits the necessary linear variation in the x or ⟨110⟩ direction. No problems are encountered because of the presence of an x component of the electric field, since, in KDP-type materials, the electrooptic effect is null for fields in this direction.

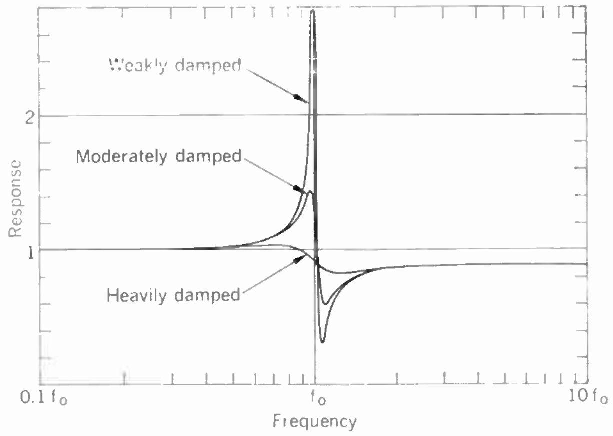


FIGURE 8. Typical frequency response of a piezoelectric electrooptic crystal showing effects of a fundamental acoustic resonance at f_0 under various conditions of damping.

If a collimated laser beam, plane-polarized in the x direction, is transmitted axially through this structure (in the z direction ⟨110⟩), it will be deflected toward positive x, as was revealed in Fig. 1. Reversal of the voltages will result in an equal but opposite deflection. An explicit expression for this deflection as a function of the applied voltage is obtained from Eqs. (2), (3), (5), and (13):

$$k = \frac{1}{n_0}\frac{dn}{dx} = \frac{n_0^2 r_{63}}{2}\frac{dE_y}{dx} = \frac{n_0^2 r_{63} V_0}{2R_0^2}$$

$$\theta(z) = kz = \left(\frac{n_0^2 r_{63} V_0}{2R_0^2}\right) z \tag{14}$$

If the total active length of material is L, then

$$\theta(L) = \left(\frac{L n_0^2 r_{63}}{2R_0^2}\right) V_0 \tag{15}$$

An additional deflection resulting from refraction at the boundary plane $z = L$ yields an external deflection $\theta_e = n_0\theta(L)$ as in our previous example. Thus we obtain

$$\theta_e = \left(\frac{L n_0^3 r_{63}}{2R_0^2}\right) V_0 \tag{16}$$

This should be compared with Eq. (11), which is the corresponding formula for the iterated-prism model.

A small 10-cm experimental index-gradient deflector, using the quadrupole electrode configuration and KDP as the electrooptic material, has been observed to yield about eight resolution elements for a 6328-Å wavelength, with a peak-to-peak signal of 20 kV at room temperature.[21] To my knowledge, no effort has been made to examine this type of device under low-temperature conditions.

It is of interest to compare the relative merits of the two types of beam deflectors that have been described—given the same requirements for deflection capability, electrooptic material, operating temperature, optical wavelength, and maximum allowable field strength. One can show that the important deflector dimensions of length and limiting transverse aperture are uniquely determined under a given complete set of operating conditions. It turns out that the iterated-prism type requires only half the length and about 70 percent the aperture of the index-gradient variety. The former, however, requires nearly twice the voltage; but its capacitance, on the other hand, is less than one third that of the quadrupole structure. Thus, both types would require about the same level of driver power. Based on considerations of relative structural complexity and quantity of electrooptic material required, the choice would seem to favor the iterated-prism model.

In the discussion of these electrooptic beam deflectors, the implication has been that the beam passes only once through the active medium and that it is of constant diameter or "collimated." Although space limitations preclude any detailed treatment of various optical tricks, I should mention in passing that the use of strategically located mirrors and focusing optics, to produce multiple passes and a slightly converging beam, can offer both enhanced deflection capability and a considerable reduction in driver voltage. This feature is important in wide-band operation, where power considerations are paramount.

Problem areas

Piezoelectric resonances. An all-important characteristic of wide-band optical deflectors, as well as of ordinary electronic signal amplifiers, is an optimally flat frequency response over the band of interest. In this respect, the piezoelectric nature of all crystals that show a linear electrooptic effect presents a fundamental problem, since the strains engendered in them by electric fields usually produce additional refractive index changes[19] having the same character as those created by the electrooptic effect itself. As a result of this "elastooptic effect," severe distortions of the desired optical transfer characteristics of these materials can occur when they are electrically excited at frequencies in the neighborhood of their natural acoustic resonances. Adequate damping techniques are therefore an important engineering consideration.

A typical frequency-response curve of an electrooptic element, clearly illustrating the effect of fundamental acoustic resonance under conditions of weak, moderate, and strong damping, is shown in Fig. 8. The resonant frequency is indicated by f_0. At low frequencies down to dc the response is flat; the piezoelectric strain is able to follow the electric field, and therefore the elastooptic contribution is fully effective. Near resonance, the strain amplitude builds up appreciably, similar to a driven harmonic oscillator. Slightly above resonance, the strain amplitude is still large, but out of phase with the driving field, thus accounting for the pronounced dip in the response curve. At sufficiently high frequencies, inertial impedance clamps the crystal, and only the true electrooptic response remains (at approximately 90 percent of the low-frequency response shown in the example of Fig. 8).

Heating effects. The resolving power of an optical instrument critically depends upon its ability to preserve the quality of the wavefronts of a traversing beam. In a beam-deflection device, an incident plane wave should emerge as a plane wave. Distortions of optical wavefronts result from *localized* variations of the refractive index within the cross-sectional area of the beam. One of the significant causes of such aberrations is a nonhomogeneous temperature distribution, since the refractive index of optical materials is temperature-dependent. The electrooptic deflectors we have discussed are susceptible to such nonuniformities of temperature by the very nature of the energy dissipation that exists within the crystals.

The principal sources of dissipation in electrooptic crystals are hysteresis losses, piezoelectrically induced acoustic losses, and the intrinsic electrical conductivity of the crystal. The deflector driving power must therefore be limited to maintain distortion-free operation; in effect, this sets an upper limit on the achievable resolution. A detailed theoretical anlysis, which is beyond the scope of this presentation, has predicted a maximum deflection capability of approximately 700 linear elements for an iterated KD*P prism deflector of more or less square cross section when driven with wavetrains typical of a selective access display mode. This estimate is based on a single pass of the beam. A greater capability is possible with multiple-pass techniques.

Under certain conditions, nonhomogeneous heating can also result from the laser beam itself, producing an adverse effect upon the characteristics of the transmitted beam. This is most likely to occur when the laser output is in excess of several watts, a capability currently exhibited by the blue-green argon-ion laser in CW operation.[22] In contrast to RF heating in the crystals, heating that is produced by a high-power laser is apt to show its effect, not in the crystals, but in the index-matching fluid. An optical absorption within the fluid of less than 10^{-3} cm^{-1} can produce a striking effect on the laser beam profile if the volume of fluid is appreciable. The resultant heating along the path of the beam creates a radial temperature gradient with an attendant refractive index gradient. This, in turn, produces an effective negative lens,[23] which causes the beam to diverge. It is important that care be taken, therefore, in choosing a low-loss fluid and minimizing its volume within the deflector cell.

Conclusion

The electrooptic light-beam deflection devices that have been described possess a high degree of versatility and are reasonably simple in structure. Their principal attraction, however, is for systems requiring aperiodic, rapid-access modes of operation. Admittedly, the high levels of driver power that they require in these wide-band applications is a major drawback, but this disadvantage is offset in large measure by their inherent fast-response capabilities, in addition to their low optical loss.

At the present time, state-of-the-art deflectors possess a single-pass capability of about 500 resolution elements in wide-band operation. Damping techniques have been moderately well-developed to the point where microsecond switching rates, with negligible ringing, are possible.

Further developmental work is required before systems of this type can achieve their full potential. Specifically, effort should be directed toward devising fully effective methods of suppressing piezoelectric resonances, and toward searching for a minimum-loss low-temperature index-matching fluid. Optical techniques designed to utilize more effectively the active medium (e.g., multiple passes of the beam) should also be exploited in order to relax the power burden imposed upon the signal-input drivers. Finally, we should anticipate further improvements in deflection capability as the technology of synthesizing more exotic electrooptic materials of greater sensitivity and more superior optical quality is advanced.

REFERENCES

1. Javan, A., Bennett, W. R., Jr., and Herriot, D. R., "Population inversion and continuous optical maser oscillation in a gas discharge containing a He-Ne mixture," *Phys. Rev. Letters*, vol. 6, pp. 106–110, Feb. 1961.

2. For a review, see Fowler, V. J., and Schlafer, J., "A survey of laser beam deflection techniques," *Proc. IEEE*, vol. 54, pp. 1437–1444, Oct. 1966.

3. Korpel, A., Adler, R., Desmares, P., and Watson, W., "A television display using acoustic deflection and modulation of coherent light," *Proc. IEEE*, vol. 54, pp. 1429–1437, Oct. 1966.

4. Baker, C. E., and Rugari, A. D., "The laser display—a large-screen, real-time display technique," *1965 Proc. 6th Nat'l Symp. Information Display*, pp. 85–101.

5. Kulcke, W., Kosanke, K., Max, E., Habegger, M. A., Harris, T. J., and Fleisher, H., "Digital light deflectors," *Proc. IEEE*, vol. 54, pp. 1419–1429, Oct. 1966.

6. Kaminow, I. P., "Microwave modulation of the electro-optic effect in KH_2PO_4," *Phys. Rev. Letters*, vol. 6, pp. 528–530, May 1961.

7. Born, M., and Wolf, E., *Principles of Optics*, 3rd rev. ed. New York: Pergamon, 1965, pp. 124, 167–168.

8. Jenkins, F. A., and White, H. E., *Fundamentals of Optics*, 3rd ed. New York: McGraw-Hill, 1957, pp. 5, 22, 302–303, 604–605.

9. West, C. B., and Carpenter, R. O'B., "Magneto- and electro-optics," in *American Institute of Physics Handbook*. New York: McGraw-Hill, 1957, sec. 6, pp. 94–97.

10. Billings, B. H., "The electro-optic effect in uniaxial crystals of the type XH_2PO_4. I—Theoretical," *J. Opt. Soc. Am.*, vol. 39, pp. 797–801, Oct. 1949.

11. Carpenter, R. O'B., "The electro-optic effect in uniaxial crystals of the dihydrogen phosphate type. III—Measurements of coefficients," *J. Opt. Soc. Am.*, vol. 40, pp. 225–229, Apr. 1950.

12. Zwicker, B., and Scherrer, P., "Elektrooptische eigenschaften der seignette-elektrischen Kristalle KH_2PO_4 und KD_2PO_4," *Helv. Phys. Acta*, vol. 17, pp. 346–373, Sept. 1944.

13. Känzig, W., "Ferroelectrics and antiferroelectrics," in *Solid State Physics*, F. Seitz and D. Turnbull, eds. New York: Academic Press, 1957, vol. 4, pp. 1–196.

14. Stephenson, C. C., Corbella, J. M., and Russell, L. A., "Transition temperatures in some dihydrogen and dideutero phosphates and arsenates and their solid solutions," *J. Chem. Phys.*, vol. 21, p. 1110, June 1953.

15. Sliker, T. R., and Burlage, S. R., "Some dielectric and optical properties of KD_2PO_4," *J. Appl. Phys.*, vol. 34, pp. 1837–1840, July 1963.

16. Haas, W., Johannes, R., and Cholet, P., "Light beam deflection using the Kerr effect in single crystal prisms of $BaTiO_3$," *Appl. Opt.*, vol. 3, pp. 988–989, Aug. 1964.

17. Chen, F. S., Geusic, J. E., Kurtz, S. K., Skinner, J. G., and Wemple, S. H., "Light modulation and beam deflection with potassium tantalate-niobate crystals," *J. Appl. Phys.*, vol. 37, pp. 388–398, Jan. 1966.

18. Kiefer, J. E., Lotspeich, J. F., Brown, W. P., Jr., and Senf, H. R., "Performance characteristics of an electro-optic light beam deflector," presented at the 1967 IEEE Conf. on Laser Engineering and Applications, Washington, D.C., June 6–8, 1967, Session 10; for a digest, see *IEEE J. Quantum Electronics*, vol. QE-3, p. 261, June 1967.

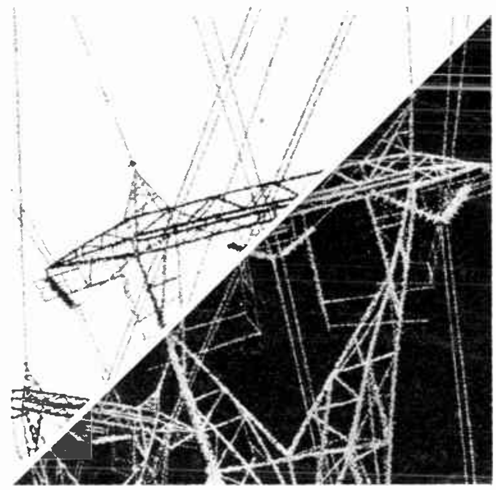
19. Cady, W. G., *Piezoelectricity*, rev. ed. New York: Dover, 1964, vol. 2, pp. 717–721.

20. Sliker, T. R., "Reference data on linear electro-optic effects," Eng. Memo. 64-10, Clevite Corp., Electronic Research Div., Cleveland, Ohio, May 15, 1964.

21. Fowler, V. J., Buhrer, C. F., and Bloom, L. R., "Electro-optic light beam deflector," *Proc. IEEE*, vol. 52, pp. 193–194, Feb. 1964.

22. Clark, P. O., "Advances in gas lasers," *1967 IEEE Internat'l Conv. Record*, part 7, pp. 69–79.

23. Gordon, J. P., Leite, R. C. C., Moore, R. S., Porto, S. P. S., and Whinnery, J. R., "Long-transient effects in lasers with inserted liquid samples," *J. Appl. Phys.*, vol. 36, pp. 3–8, Jan. 1965.



Prevention of power failures— The FPC report of 1967

Power demand in the U.S. is increasing at the rate of a geometric progression. Interconnections now cover vast geographic regional areas; hence, reliability of the bulk power supply system is the key criterion for the uninterrupted flow of electric energy

Gordon D. Friedlander *Staff Writer*

Twenty months after the Northeast blackout of November 9–10, 1965, the Federal Power Commission issued its three-volume report calling for the coordinated planning and operation of bulk power supply facilities to ensure maximum possible reliability and the prevention of future cascading tripouts and regional power failures. These and related guidelines are contained in 34 recommendations. As would be expected, the power industry, after evaluating this comprehensive report, may have some disparate reactions. A small sampling of these reactions by three investor-owned utilities—and a federal system—is herein presented.

The FPC's three-volume report, "Prevention of Power Failures," is a comprehensive and painstaking document that has been written and compiled in the best traditions of cooperation between the federal government and private enterprise. This is not the work of some government superagency dictating a set of fiat mandatory recommendations that must be followed, observed, and put into effect by private industry under the threat of punitive regulatory action for failure to comply. Instead, it is a detailed report of the recommendations and conclusions of the FPC for improving present and future electric power service; plus the Advisory Committee's own report on the reliability of electric bulk power supply; and, finally, a presentation of the studies and findings made by the various Task Groups on the Northeast power interruption of 1965 and subsequent outages that affected interconnected systems.

Significantly, the Advisory Committee and the Task Groups were composed of prominent electrical engineers and systems engineers, drawn from public and private utilities, a university, a state commission, and the manufacturers. Thus a cross section of power engineering practitioners and educators participated in the careful studies, investigations, and exhaustive legwork that went into the drafting of this important treatise.

Many improvements made—more needed

During the past two years, federal, state, and city agencies and private industry have responded well in eliminating a number of the shortcomings indicated by the massive Northeast power failure of November 9–10, 1965. Actions that have been taken on national, regional, or local levels include

1. A significant upgrading of communications and emergency lighting facilities at airports throughout the United States.

2. The development of emergency plans and procedures to ensure that teletypewriters, PBX phones, and other telecommunications systems be kept in operation and available to the public during service interruptions.

3. The initiation of a nationwide program to ensure that all civil defense and military warning equipment (red

phones) have a constant power supply, free from any possible interruption.

4. The implementation of a program to equip federal government buildings with emergency lighting units in stairwells, elevator cars, transformer and switchgear rooms, control centers, and other areas vital to national security.

5. A grant of federal funds, by the Department of Housing and Urban Development, for emergency transportation facility improvements by providing emergency power for the movement of trains; the development of emergency radio, station lighting, and alarm signal equipment, etc.

6. The improvement of standby power sources for communications and control during all critical phases of space flight missions.

7. Formulation of rules by the City of New York to ensure the continuous safe functioning of vital hospital services and facilities during power emergencies.

8. Action by the Port of New York Authority to upgrade communication, lighting, and emergency evacuation procedures for vehicular tunnel and tube train facilities under its jurisdiction.

Although these eight measures represent a significant package of improvements, there are still deficiencies in many critical areas. For example, a recent national survey of statutory provisions for standby emergency power for essential services revealed that 22 states do not have such legislative requirements on their books.

Present status in the Northeast. In the year following the Northeast outage, the major utilities affected invested $20 million for new equipment and improvements to protect existing facilities and to decrease the possibility of future cascading power failure occurrences. Additionally, more than $30 million has been earmarked for further improvements that will be made just as soon as procurement and installations will permit. One problem in this area is that the demands for equipment have temporarily exceeded the manufacturers' and suppliers' production capabilities.

Future development by Northeast utilities

The Fig. 1 map shows the principal transmission network elements in the New York State and New England areas, with interconnections to the PJM pool and to the service area of the Hydro-Electric Commission of Ontario. The map also indicates the lines added since the extensive blackout of 1965, as well as those scheduled for installation through the year 1973. In the eight-year period from November 1965 through 1973, 2250 km of 345-kV circuits and almost 500 km of 500-kV circuits either have been or are scheduled to be placed in service in the Northeast.

The peak load estimated for the Northeast systems—exclusive of Ontario Hydro—for the winter of 1973–1974 will be 31 200 MW, an increase of 35 percent above the area peak load projected for the winter of 1967–1968. Planned generating capacity in the seven-state area by the end of 1973 will be 40 300 MW, or 44 percent more than the capacity in 1967.

Provisions for load shedding

According to Volumes I and II of the report, the best insurance against a major power failure is sound planning plus a well-designed and -operated bulk power supply system. It is possible, however, that an unexpected incident can isolate a utility or power pool from the network; and, if the isolated area created is deficient in generation, loads must be reduced quickly to prevent a total collapse of the power supply system. Therefore, automatic controls are essential for this purpose because manual load shedding is usually too slow to be effective in an acute emergency situation. This fact was dramatically demonstrated during the 1965 blackout when the southeast New York and New England utilities became an "electrical island" five seconds after the disturbance began; but these isolated systems did not collapse until 12 minutes after their separation. In that brief period, system voltage and frequency fell far below normal and the capability of generators that remained in operation diminished rapidly.

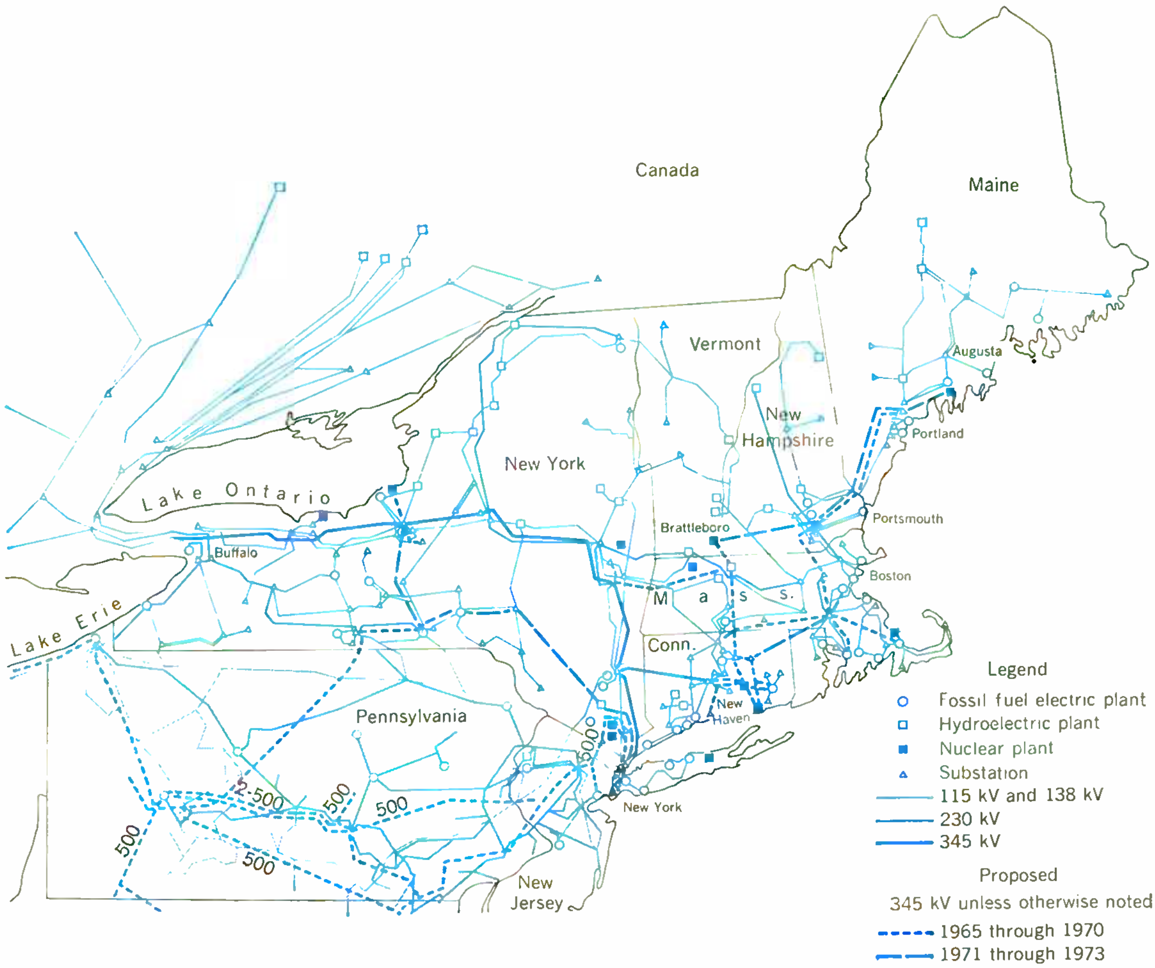
At the time of the November 9 interruption, no utility in the CANUSE network used automatic load shedding. This was in contrast to some other areas of the United States in which this procedure is utilized. At the present time, however, load-shedding procedures have been established at all the major utilities in the Northeast, largely through the efforts of the Northeast Power Coordinating Council.

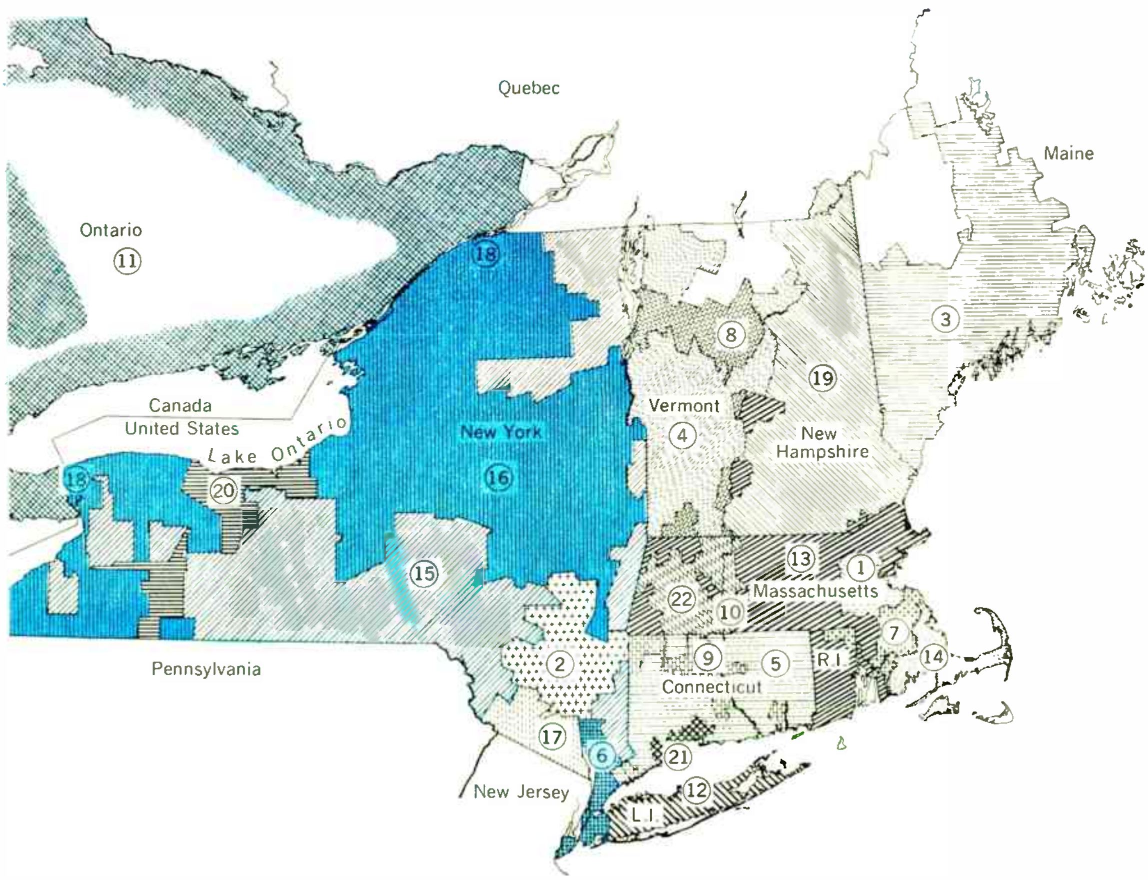
Service to New York City

The Northeast power failure hit New York City particularly hard. At the time of the massive disturbance, the Consolidated Edison Company of New York was meeting a demand of 4770 MW, but the installed capacity of the company's generating equipment was about 7580 MW. The Con Edison system collapsed because of the effects of varied and dispersed deficiencies in the interconnected system of which it was a part. These deficiencies included inadequate coordination between Ontario and

FIGURE 1. Map showing major transmission lines and generating stations, both existing and proposed, in the New York and New England power system (as of January 1967).

FIGURE 2. Map indicating the participating member systems of the Northeast Power Coordinating Council and their respective service areas.
1. Boston Edison Co.
2. Central Hudson Gas & Electric Corp.
3. Central Maine Power Co.
4. Central Vermont Public Service Corp.
5. Connecticut Light and Power Co.
6. Consolidated Edison Co. of N.Y., Inc.
7. Eastern Utilities Associates.
8. Green Mountain Power Corp.
9. Hartford Electric Light Co.
10. Holyoke Water Power Co.
11. Hydro-Electric Power Commission of Ontario
12. Long Island Lighting Co.
13. New England Electric System
14. New England Gas and Electric Association
15. New York State Electric & Gas Corp.
16. Niagara Mohawk Power Corp.
17. Orange and Rockland Utilities, Inc.
18. Power Authority of the State of N.Y.
19. Public Service Co. of New Hampshire
20. Rochester Gas and Electric Corp.
21. United Illuminating Co.
22. Western Massachusetts Electric Co.


Canada
Maine
Vermont
New Hampshire
Augusta
Portland
Portsmouth
Boston
New York
Lake Ontario
Buffalo
Brattleboro
Mass.
Conn.
New Haven
Lake Erie
Pennsylvania
New York
New Jersey
500
2-500
Legend
Fossil fuel electric plant
Hydroelectric plant
Nuclear plant
Substation
115 kV and 138 kV
230 kV
345 kV
Proposed
345 kV unless otherwise noted
1965 through 1970
1971 through 1973


Quebec
Maine
Ontario
11
18
3
8
19
Canada
United States
Lake Ontario
Vermont
4
New Hampshire
New York
16
20
15
13
1
Massachusetts
22
10
7
14
9
5
R.I.
Connecticut
2
Pennsylvania
17
6
21
12
L.I.
New Jersey

New York, inadequate transmission and interconnection in the overall network, and inadequate coordination in operation among the systems.

During the past two years, Con Edison has improved its systems of instrumentation and control by the following measures:

1. The installation of separate wide-band and narrow-band frequency meters at the control center to preclude the human error of misreading the scale.

2. The provision of alternative power sources for communication and control that can furnish continuously available and reliable power during system disturbances, including power supplied for the transmission of system performance data by telemetering and other communication equipment.

3. A system of load shedding, by means of automatic voltage reduction, that will be applied in two steps by the action of underfrequency relays.

The Northeast Power Coordinating Council

One of the top-priority actions resulting from the Northeast power failure was the formation, in January 1966, of the Northeast Power Coordinating Council to improve coordination among the 22 major utilities in the CANUSE area (see Fig. 2). The work of the Council is supplemented by task forces on system studies, system protection, load and capacity, load-shedding and spinning reserve, and computer controls.

Since its organization, many of the Council's activities have been related to the solution of problems disclosed by the big blackout. Some of the major efforts of the Council's committees have included

1. Stability studies of the Northeast network.

2. Projections of load, generation, and transmission requirements by 1973.

3. A coordinated study of load shedding.

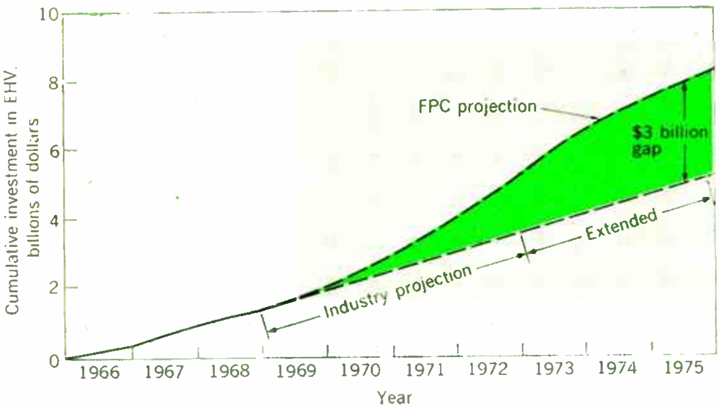


FIGURE 3. Federal Power Commission's graph of projected investment in extra-high-voltage transmission from 1966 through 1975.

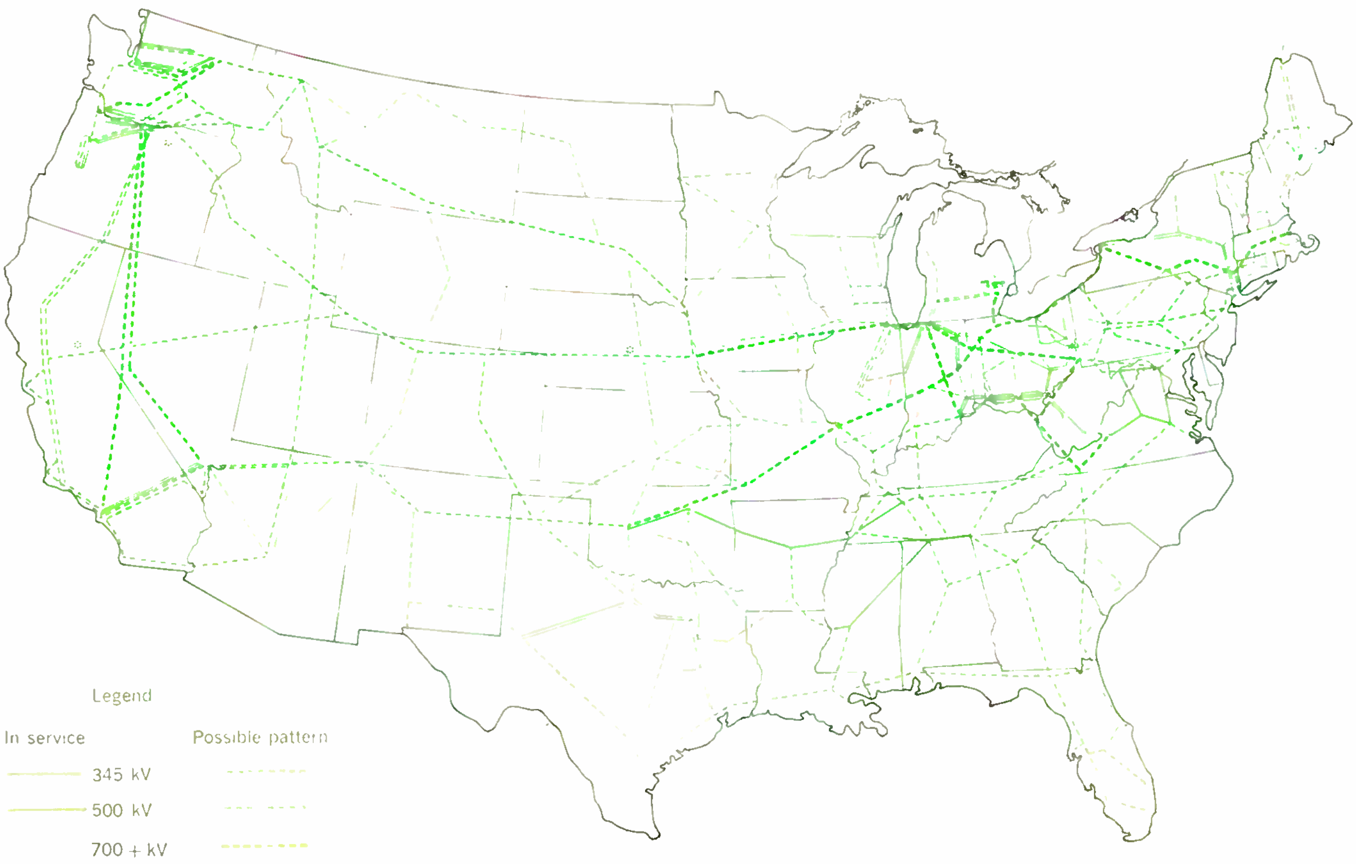


FIGURE 4. Map of possible pattern of power transmission for increased reliability by 1975. Note that none of the lines indicated on this map is to be considered as a specific recommendation of the FPC. Further, no recommendation is intended as to number of lines, level of voltage, or type of power (ac or dc) for the principal east–west ties between areas marked with an asterisk (*).

The member utilities of the Coordinating Council recently adopted an automatic load-shedding program, which provides that each system will be equipped with underfrequency relays to drop 10 percent of system load if the frequency declines to 59 Hz, and 15 percent more at 58.5 Hz. Each system will also provide for the manual dropping of an additional 25 percent of its load whenever warranted by emergency conditions.

The Northeast Council is one of several similar organizations, either established or now being established throughout the United States, for the purpose of achieving better reliability of bulk power supplies.

Possible pattern of EHV transmission

As outlined by the FPC, the possible pattern of EHV transmission lines required by 1975 would involve an investment of about $8 billion—or 12.5 percent of anticipated expenditures for all power facilities to be built during this seven-year period. This is about $3 billion more than the utilities apparently plan to spend (see Fig. 3), according to the FPC.

In the general pattern projected for 1975 (see Fig. 4 map), the additions in EHV lines beyond those put in service in 1967, include 25 800 km of 345-kV, 34 500 km of 500-kV, 9250 km of 765-kV, and 2680 km of approximately 750-kV dc transmission circuits.

The depicted scheme indicates that the utilities in the Northeast and Southeast will be more strongly integrated with the central body of utility systems in the East. An extension of the 765-kV transmission system, now being constructed in the east-central area, is shown in Fig. 4 to overlay the 345-kV network now under development in New England. This would enable major power flows to occur between the heart of New England and the central eastern section of the United States.

The Southeastern utilities are shown to be interconnected with several 500-kV loops that join an existing system of corresponding voltage that has already begun to span a major section of the Eastern Seaboard from north to south.

Strong north–south interconnections from the heavy industrial load areas of Illinois, Michigan, Ohio, Indiana, and western Pennsylvania to the utilities in the southeastern and Gulf state areas are shown extending through the TVA and CARVA regions. Here, American Electric Power (AEP) has an extensive (2900-km-long) EHV network interconnecting its six subsidiaries throughout a seven-state area from Michigan to West Virginia. A 765-kV overlay network is scheduled for initial operation by 1969 and completion by 1971.

Conclusions and recommendations of the FPC

This major section of the FPC report contains 34 subsections under nine subheads or subjects. A précis of these subsections is presented next in a numbering sequence that conforms to the full text and exact subheads as given in the original document—

Formation of coordinating organizations

1. Strong regional organizations should be established for the coordination of the planning, construction, operation, and maintenance of individual bulk power supply systems. These coordinating organizations should have financial support and representation from all participating utilities.

2. A Council on Power Coordination should be established, composed of members from each regional coordinating organization, to exchange information and to review, discuss, and assist in resolving matters that affect interregional coordination.

3. A Central Study Group, or Committee, should be established to coordinate industry efforts in investigating some of the more challenging problems of interconnected system development. This entails the early coordination of ideas, efforts, and funds required for more effective R & D in planning and operation. Much of this work could be performed under contract with qualified universities and research institutions.

Interconnected system planning

4. Early action should be taken to strengthen the Northeast transmission systems. And the peninsular relationship of New York and New England to adjoining systems demands highly coordinated planning of transmission. For example, the projected 500-kV interconnection to the PJM network is urgently needed. The FPC also recommends an early reinforcement of the Northeast network and more ties to PJM and systems in Ohio and western Pennsylvania.

5. Transmission facilities should be critically reviewed throughout the nation, and planning and construction of needed additions should be accelerated on schedules which provide sufficient capacity to meet the potential requirements of both reliability and economy. Transmission networks and interties between areas are now deficient in numerous locations. Networks should be planned and tested for their ability to remain stable under severe disturbances. The pace of construction should enable transmission capability to lead rather than lag behind emergency requirements.

6. In estimating future loads, full attention should be given to economic trends, potential weather extremes, and growth in specialized uses of electricity (electric heating, air conditioning, etc.) **in each load area.**

7. Lead times for planning and constructing major new facilities should be established that will avoid delays in meeting completion schedules and impairment of system reliability. In comparison with past practice, time extensions of one to two years may be needed for large components. In many cases, it may be necessary to develop relatively firm expansion plans not less than six years in advance of need.

8. Utilities should solicit the participation of interested parties at an early date in the resolution of problems relating to the location and environmental effects of new facilities. Utility planning, in addition to technical factors, involves careful attention to facility location, the satisfactory control of air and water pollution, and the preservation of esthetic values.

9. Special attention should be given to transmission line routing and to switching arrangements in the transmission network to provide maximum reliability in emergencies. Unusual care should be taken to prevent the excessive concentration of critical circuits, which would expose the system unnecessarily to large loss of capability.

10. The size of generating stations, magnitude of area loads, and the capability of the transmission system should be maintained in proper balance. Generating capacity that is too large in relation to the capability of the interconnecting transmission lines and area load concentrations can impair the reliability of supply.

11. Sufficient transmission should be provided to avoid excessive generating reserve margins. Limited transmission capacity tends to result in generating reserve margins larger than those justified by either economic or reliability considerations.

12. A workable number of control centers should be established in each region. The Northeast power failure dramatically emphasized the inability of operators in many centers to have significant communications with each other. Plans are under way to set up two central control points—one in New England, and the other in New York. This simplification should improve the coordination of the Northeast systems for both normal and emergency operations.

13. Relay protection should be continually updated to fit system changes and to incorporate improved relay devices. Relays are key elements in achieving reliability of bulk power supply. Since they are relatively inexpensive components, their adequacy, quality, and periodic readjustment should not be compromised for the sake of negligible economies.

14. Utilities should concentrate on opportunities to expand the effective use of computers in power system planning and operation. Many new applications are being found for digital computers in these areas. Specifically, computers may be used for the collection and printout of operating data, including warning of conditions that are approaching or exceeding safe limitations; the automatic control of generators to meet changing loads economically; increased automation of generating stations; rapid analysis of networks to determine line-load limits, etc. Further progress is needed to use computers reliably for on-line analysis and control of power systems during disturbances.

Interconnected system operating practices

15. System control centers should be equipped with display and recording apparatus to provide the operator at all times with as clear a picture of system conditions as is possible. Desirable displays include narrow- and wide-range frequency indicators, tie-line and principal line flows, lines out of service, switch positions, overload conditions, generators in service, unit and plant outputs, spinning reserve and rate of response, voltages and frequencies at key points, area control error, and appropriate alarms.

16. Communications should be supplied with continuously available power so that information on system conditions can be transmitted to control centers during disturbances. Whenever power supply for communications equipment deviates beyond specified limits, the equipment should be automatically and instantly switched to an emergency power source.

17. Control centers should be provided with a means for rapidly checking on stable and safe system capacity limits. Rapid security checks—now feasible by the use of digital computers—to determine that various elements will be operated within safe limits are essential to prevent hazardous loading.

18. Spinning reserves should respond quickly to a level which can be sustained in meeting emergency power demands. Rapid response normally requires that the reserve be distributed among many units.

19. Coordinated automatic load shedding should be established to prevent the total loss of power in an area that has separated from the main network and is deficient in generation. Load shedding should be regarded as an "insurance program," however, and *should not* be used as a substitute for adequate system design.

20. Plans should be made, and tests conducted, for the quick isolation of generating units to keep them in operation if collapse of system power is imminent. Carefully planned switching procedures should isolate appropriate units quickly during an emergency. Smaller units may be isolated primarily to ensure restarting power for larger units.

21. Emergency power should be available at all thermal plants to prevent damage to turbogenerators during rundown if system power is lost. Pressures must be maintained on bearing lubrication and hydrogen sealing systems. Emergency power should be provided to operate the turning gear and pumps of these systems, to keep the control system operable and to provide lighting in the control room.

22. Auxiliary power should be available to principal thermal plants to enable rapid restarting if system power is lost. Adequate emergency power can save hours of time in service restoration.

23. Thorough programs for operator training should be vigorously administered, with particular attention to procedures for a broad spectrum of potential contingencies. Close coordination among the planning, operating, and maintenance staffs is indispensable.

Interconnected system maintenance practices

24. Programs of system maintenance should be strongly directed toward preventive rather than remedial maintenance.

25. Manufacturers and utilities should promptly disseminate information on troubles or equipment failures. The FPC will also disseminate information on bulk power outages and selected power interruptions.

26. The isolation of any system elements for testing, repair, or replacement should be scheduled by, or receive the clearance of, the operating department.

Criteria and standards

27. Criteria and standards for planning, construction, operation, and maintenance of power systems should be formulated so that each system can be reasonably assured that its own service will not be adversely affected by its neighbors' policies.

Defense and emergency preparedness

28. Although severe damage can be inflicted upon power systems by enemy attack, cascading failures should not ensue. Steps to improve reliability will strengthen utility systems in resisting widespread failures if subjected to wartime attack. Utilities generally have adequate security programs for normal requirements, but many should increase preparedness for the contingencies of enemy attack.

29. All levels of government should establish requirements for emergency power for essential services. More than 50 percent of the states now require auxiliary power for certain critical loads, and this practice should be extended. Thus the FPC urges state, county, and local governments to encourage and direct the planning and installation of auxiliary power facilities to furnish essential and vital services for public safety and welfare.

30. Utilities should cooperate with public officials and customers in planning and maintaining customer standby facilities to ensure service to critical loads in an emergency. Typically, these services include hospitals, police and fire departments, sewage and water plants, transportation systems, communications, and emergency service in buildings that normally contain many people.

Manufacturing and testing responsibilities

31. Manufacturing capacity of electrical equipment suppliers should be expanded to meet future needs. Better liaison on projected requirements between utilities and manufacturers is needed, and this will be aided by improved planning and coordinating procedures.

32. Facilities are needed in the U.S. for more extensive testing of EHV equipment. The Commission urges the early consideration by the electric utility industry of needs, development of plans, early construction of appropriate high-voltage testing facilities, etc., so that the reliability of future power supplies will not be impaired by lack of proper testing.

Increased need for technical proficiency

33. The industry should make young people cognizant of the full challenge of modern power systems engineering. Utilities should work more closely with educational institutions to develop and sponsor appropriate research, to utilize cooperative programs for students and industry assignments of educators, and to exploit opportunities for new and sophisticated R & D.

Power system practices in other countries

34. System design and operating practices in other countries are generally similar to those in the U.S., and foreign power systems are experiencing similar problems in planning and operation. The practice of exchanging technical information in these areas with foreign countries should be continued and expanded.

Synopsis of Volume II . . .

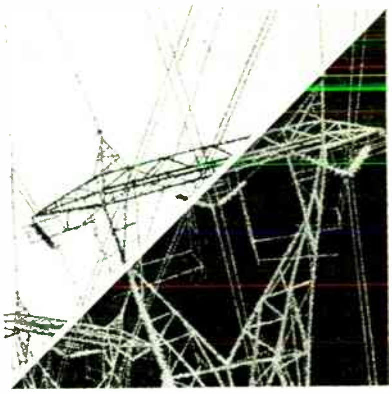
The FPC Industry Advisory Committee report on electric bulk power supply, released as Volume II of the Commission's report, concluded that the sound application of bulk power reliability principles should eliminate widespread or cascading interruptions to service from all "credible" contingencies.

This committee, established by the FPC shortly after the Northeast blackout, advised the Commission that cascading power failures result basically from inadequate transmission facilities within and between power systems, and urged the increased use of properly planned high-capacity transmission. The committee also recommended regional and interregional coordination (as already reported in this article) and proposed a nationwide council on power system coordination.

The Advisory Committee report contains considerable technical information related to various aspects of power system planning and operation—including recommended practices for dependability and reliability of electric bulk power supply systems.

. . . and Volume III

Volume III of the FPC report contains several studies of the Northeast power network, which were made following the 1965 power failure. Selected conditions were studied in detail by task groups under the direction of the Commission's Advisory Panel on the Northeast Power Interruption. The volume is composed of six sections that present the results and conclusions of the several study groups.



Reaction and comments

To provide the reader with a more comprehensive overview of the FPC report than can be presented by a recitation of the Commission's recommendations and conclusions, the writer has attempted to solicit a sampling of reactions and comments from a small segment of the power industry.

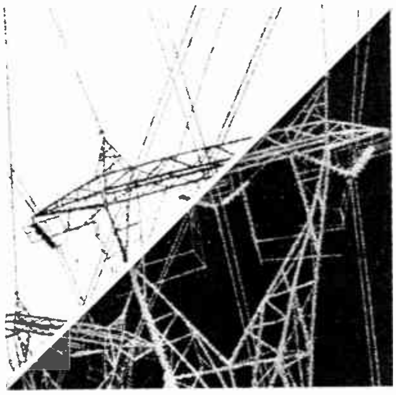
American Electric Power Company (AEP). A spokesman for AEP, the huge, fully integrated utility that provides electric power in parts or all of seven east-central states, believes that the industry can design power supply systems to prevent widespread blackouts, if the various major system components, comprising generating plants, transmission lines, and interconnections with neighboring systems, are planned as an integrated whole, with proper consideration given to their interrelated effects; in other words, if planning is carried out on a truly system basis. If this is done, instantaneous outages can be restricted to small, discrete geographical areas. But this goal cannot be achieved if coordination in system design is ignored because of either an overemphasis on immediate economies, or inadequate attention given to the mutual effects of the actions of other systems.

Although agreeing with and subscribing to the concept of regional coordination, AEP felt that the representation and functions of the regional councils, boards, and committees—as proposed in the FPC report—are too broad and diverse. These defects would result in an unmanageable organizational structure, which, in many instances, would attempt to do what the individual systems and pools should be doing and, in turn, would result in an extravagant waste of technical talent.

The regional councils should pursue reliability as their objective and not complicate this primary purpose with economic and other factors. These latter objectives can best be accomplished within the individual systems and pools. Representation on such councils should come from systems that can demonstrate that they have a substantial effect on the reliability of the bulk power network and can contribute to its achievement.

The company feels that to keep planning viable, it must be initiated within the individual utility, or pool of interconnected utilities, and then be submitted to a council for review.

It was also pointed out that some of the industry's problems could best be solved by greater integration of systems under common ownership. Although there are about 3600 individual utilities in the United States, only about 100 of these supply almost 90 percent of all electric power. Thus, to improve overall planning and operation, integration (either by merger or consolidation) to 10 to 20 major systems might be considered as an



eventual goal for greatly improved reliability and service. So far, the FPC has avoided the integration and consolidation issue and, instead, has concentrated on system pooling and coordination.

In regard to future construction of EHV transmission lines, the company believes that the projected FPC estimate is open to question, since it fails to take into account the inevitable time lag between the development of industry plans and their public announcement. Also, the report's simple extrapolation of past EHV construction into the future ignores the rapidly growing role of EHV in transmission as indicated by Table I of the Advisory Committee's report.

On load shedding, AEP considers that the prime function of a utility is to provide power, not to curtail it. Therefore it believes that power systems should be planned without the need to resort to load shedding, and that load shedding should be viewed strictly as "insurance," not as a planning tool.

Commenting on the other FPC recommendations for improvements in planning and operating practices, the AEP spokesman voiced general overall agreement and noted that the vast majority of these conclusions were also items listed in the Advisory Committee report.

Bonneville Power Administration. In a brief conversation with a consulting engineer of the BPA, the writer was informed that this federal system is essentially in agreement with the findings of the FPC report. Like AEP, BPA—although conceding there is a variance of opinion in the industry—regards load shedding as a "last-ditch insurance," but considers load shedding definitely preferable to loss of generation due to sinking frequency. Also, BPA believes that voltage regulators should be kept in service during severe disturbances.

Regarding the EHV "transmission gap," this spokesman agrees with AEP that the projected deficiency (Fig. 3) indicates too wide a margin. Sophisticated design and control is an effective substitute for some aspects of brute capital expenditure.

Finally, maximum credible-incident studies and maximum potential-disaster studies should be simultaneously undertaken by all networks to ensure an absolute minimum probability of a cascading outage.

Pacific Gas and Electric Company. This utility submitted comments on ten of the FPC's recommendations:

On Recommendations 1 and 2, PG&E feels that carefully executed studies by competent and knowledgeable engineers are essential for the proper understanding and evaluation of large interconnected systems. Organizations have been formed, or are evolving, to conduct such studies on a continuing basis in a number of regions.

To function effectively, however, these organizations should be of manageable size, and include engineering representatives from the principal bulk power transmitting and generating agencies. The studies require the best available techniques in system representation, computer programming, and the best computer facilities.

Regarding Recommendation 5, which, in effect, suggests the application of criteria on a national basis for the design of transmission systems, the company believes this suggestion to be of questionable value because it does not adequately recognize the widely divergent conditions (climate, geography, etc.) in different regions of the United States.

In reference to Recommendation 7, PG&E suggests *adequate* lead time for planning and building new facilities. Although it is easy to ask for longer lead time, sound engineering practice also suggests that lead times be kept *as short as possible* to assure that decisions are based upon the latest trends in load growth and state of the art in equipment and facilities. Lead time for major projects should be predicated upon the necessary planning, engineering design, manufacturing, and construction time requirements. Any factors that increase this time may militate against the project in terms of cost and latest available technologies.

The company feels that Recommendation 9, which calls attention to the importance of transmission-line routing and switching arrangements from the viewpoint of reliability in emergencies, is generally sound—but again it does not consider the tremendous range of topographic and climatic conditions encountered in various sections of the country, which are, necessarily, a determining factor in situating facilities.

Commenting upon Recommendation 14, which relates to the use of computers in power system planning and operation, PG&E considers that the stated objectives are sound in principle, and many of them have been, or are being, adopted by progressive utilities. Computer applications presents one of the most challenging technical areas facing the contemporary power system engineer; however, the industry recognizes that some of the more sophisticated proposed applications require considerable R&D to ensure the fulfillment of overall system reliability criteria.

On Recommendation 18, it is felt that emphasis could be added to the thought on spinning reserve. The time-rate of response is vital to effective action of spinning reserve in an emergency situation. More information, based upon field tests, on the pickup rates of actual machines would be desirable, and this can now be factored into system studies.

On Recommendation 19, urging the establishment of automatic load shedding as a backup to other emergency measures, PG&E's views are similar to those of Commonwealth Edison (presented in the next section). The effectiveness of load shedding as a means of providing better service in an emergency has been demonstrated and documented on a number of occasions for systems in California and elsewhere. Nevertheless, system planners generally agree that this practice is not a substitute for adequate system design.

In discussing the last two recommendations, nos. 33 and 34, the utility thinks these are "seemingly self-evident," but many aspects of large regional system analysis, design, and operation involve engineering concepts that are relatively new to the industry. The shifts to higher voltages, direct current, and larger generating unit sizes offer many challenging areas of technological development to the young engineer. Thus the industry must make every effort to present and explain these challenges. Article no. 34 calls attention to the obvious fact that we in the United States do not live in a technologically isolated nation;

other countries also have excellent engineering talent and innovative ingenuity. By active participation in organizations such as IEEE and CIGRE, in contacts with foreign engineers, and in our own travels, we should be alert to these developments in other countries.

Commonwealth Edison Company. This large, Chicago-based utility believes that the Northeast blackout focused attention on the need to improve coordination of planning and operation of the vast bulk power systems that are being built. The advent of EHV transmission (345 kV and higher) and very large generating units (500–1100 MW) has produced a marked change in electric power systems, and a new dimension in power system planning has been introduced. Although the basic principles for designing these systems have been well established, practical experience has been lacking.

Several important lessons were learned from the Northeast blackout and the studies that followed. These were in the areas of—

Regional planning. Working with today's large generators and EHV lines, the planning of an individual utility's needs, without regard to the effects on its neighbors, can lead to dangerous conditions. This situation was recognized before the blackout when regional organizations such as MAIN, MAPP, CAPCO, etc., were organized to coordinate the planning efforts of the bulk power systems in the regions covered by these groups. The Northeast power failure indicated the need for such groups throughout the U.S., and their establishment is now well under way.

Criteria for transmission planning. Although it is axiomatic that a strong transmission system is a basic requirement for protection against cascading outages, the problem is to determine what constitutes an adequate transmission system—after considering all economic factors and system reliability.

Volume II of the FPC report presents guidelines for the tests to be applied for determining the adequacy of a bulk power system, and these are a good starting point in studying the needs for transmission in a particular area. However, conditions affecting bulk power reliability vary greatly throughout the country. Thus it would be unwise to establish uniform planning criteria for transmission design on a national basis. Such variables as load density, climatic conditions, area geography (see PG&E comments), the availability of hydro power, transmission-line concentration on a single right of way, etc., must affect final decisions regarding transmission system design. Further, the need for flexibility in planning criteria also applies to the size of generating units and margin of reserve required.

Although the FPC report indicates that the utilities should invest an additional $3 billion above their present plans for transmission over the next ten years, the Commission has not documented its argument to prove the necessity for this additional transmission increment. Commonwealth Edison believes that utilities constructing the bulk power systems can be relied upon to provide adequate facilities; anything beyond the requirements determined by the utilities will represent an uneconomic investment. Parenthetically, it is entirely possible—based on past history—that the utilities will install even more transmission than that recommended by the FPC. This will depend upon the results of studies to determine the future need for EHV lines.

Load shedding. The utility feels that the use of load shedding to check a condition of frequency reduction is a powerful tool for preventing large-scale blackouts. The basic cause of an area-wide blackout is the shortage of generating capacity, resulting from system separation, which leaves a portion of the system with load in excess of power supply. In such an emergency, the simplest method for restoring stability is to cut off load either by manual switching or the automatic opening of switches by the use of underfrequency relays.

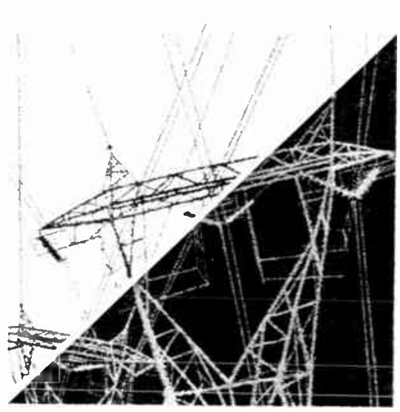
Commonwealth Edison adopted the principle of load shedding many years ago and, in 1963, installed automatic relays for this purpose. On one occasion, a widespread blackout was prevented by manually disconnecting about 10 percent of the system load. In less than one hour's time, the system was stabilized and the load was returned to normal.

The Northeast blackout convinced many power companies that the principle of load shedding—if properly applied—provides good insurance against system failure.

Computers. After the 1965 power failure, a school of thought developed in support of the thesis that bulk power systems had become so complex that it was necessary to automate their operation completely. The utilities generally have been major users of computers, both for system planning and system operation. But experience has indicated that the changeover to automation must evolve gradually over a period of time to "prove in" and coordinate the various automated components. The industry consensus is that available computers do not have the capability and reliability to perform correctly *all* of the necessary functions in the operation of a bulk power system. Also, the communications systems required to provide the input data to the computer are presently inadequate for the successful automated operation of a power system.

The utilities, however, *are interested* in expanding the role of the computer in system operation, and many studies are now under way to develop plans for greater usage of the computer in bulk power systems. The next stage of expanded use of the computer should probably be in assisting system operators, but the final decision on action to be taken in various emergency situations should be left to the operators.

The Commonwealth Edison Company concludes that the FPC report has been helpful to the power industry in forcing discussions and studies of the various factors that affect the reliability of the bulk power systems in the U.S. The company believes that the power industry is well organized to use the tools available for planning and operation, and can be depended upon to provide the adequate electric service demanded in this country, both in quantity and quality.



BIBLIOGRAPHY

"Prevention of power failures, a report to the President by the Federal Power Commission," (in three volumes), U.S. Government Printing Office, Washington, D.C., 1967.

In the March 1967 issue, SPECTRUM published an authoritative discussion on "International Standards for Color Television" and some not so authoritative remarks by the editor. Correspondence stimulated by this has been most interesting, and has revealed that there is much more heat than light left in the subject. Now we have an article by an outstanding spokesman for the NTSC system, followed by discussions solicited from a number of European experts, including one with a broad experience in the maintenance business.

Hopefully, this exchange of views will clarify some points on this still-controversial subject, and help our South American members to reach a sensible conclusion for their area. Which is the best system? See if you can decide.

C. C. Cutler
Editor

Color television

Charles J. Hirsch
RCA Consultant

The countries of Latin America, all of which are in Region 2, will shortly adopt standards for color television.

Three color television systems have been proposed for Region 2. These are: (1) NTSC, which is in wide use in the United States, Canada, and Japan; (2) PAL, which has been adopted by most of Western Europe; and (3) SECAM, which has been adopted by France and the Communist countries of Europe. However, PAL or SECAM standards suitable for Region 2 differ fundamentally from PAL or SECAM standards suitable for Region 1, which includes Europe. This is because, by international treaties, Region 2 operates with 6-MHz television channels, instead of 7.0- and 8.0-MHz, which are standard for Region 1.

Direct transmission of color television by orbiting satellites

The transmission of color television over very long distances by way of orbiting synchronous satellites[1] is eminently successful. It is expected that such transmission directly into the home can be achieved in the proximate future. It is therefore important that common transmission standards be established that will permit and encourage this to happen. Unfortunately, this standardization is no longer possible on a worldwide basis, but it is still possible in Region 2, which includes all the Americas. It is especially desirable in this region because it is divided into few, but contiguous, time zones that permit live broadcasting in real time.

Exchange of programs

There are two television scanning standards in use by the countries of North, Central, and South America. The larger number of countries use the 525/30 standards (CCIR System M) comprising 525 lines per picture and 30 pictures per second. The other countries use 625/25 standards (CCIR System N) comprising 625 lines per picture and 25 pictures per second. The line frequency for both these systems is substantially the same; consequently, 625/25 and 525/30 sets can receive each other's pictures if the remaining standards are the same. All that is required is that the vertical oscillators be capable of operating over the range of 50–60 Hz.

Television in Region 2 is confined to 6.0-MHz channels, resulting in sound and color-intercarrier frequencies that are respectively 4.5 and 3.58 MHz. In Europe, all color channels are 8.0 MHz wide, with a color-intercarrier frequency of 4.43 MHz and sound-intercarrier frequencies of 5.5, 6.0, or 6.5 MHz, depending on the country. Therefore, PAL or SECAM signals suitable for Region 2 differ from the PAL or SECAM signals now designed for Region 1, so that direct exchange of programs between the two regions is not possible, even if both use the PAL or SECAM systems. On the other hand, direct exchange of monochrome and color programs is possible between all American countries using the NTSC system, because the receivers can operate on each other's scanning standards with relative ease.

Standard converters to be used by local broadcasting stations for converting received or taped PAL or SECAM signals from Europe for NTSC retransmission in Region 2 are no more complex than they would be for converting to PAL or SECAM suitable for Region 2.

The adoption of PAL and/or SECAM by Latin America will not permit the direct exchange of color programs with Europe, but will prevent direct exchange within the Americas. Adoption of NTSC, however, will permit direct exchange of monochrome and color programs throughout the Americas. This is true whether the programs are rebroadcast from local stations or are received directly into the home from synchronous satellites. The system will not prejudice the exchange of rebroadcast programs with Europe.

The ability to exchange programs directly is especially important for the countries of Region 2, because of the essentially common time zones.

Comparison of NTSC, PAL, and SECAM

Before comparing experimental systems, such as PAL and SECAM, with a system in such wide commercial operation as NTSC, it is desirable to review briefly the present status of NTSC color television in the United States.

Broadcasting in the United States is largely commercial; the broadcasters compete with each other for the advertising dollar, and therefore seek the largest possible audience for the lowest cost. Several years ago, when the color audience was small, not all broadcasters found it profitable to spend the funds necessary for equipment and supervision to insure high technical quality of color broadcasts, and quality suffered accordingly.

In the last three years the situation has reversed itself

standards for Region 2

Color television systems are rapidly gaining worldwide popularity; but each country seems to have a different system. Latin America is now approaching the same problem, hopefully with a regionally compatible solution

dramatically. By April 1967 there were over 10 million color receivers in the United States. Of these, 2.7 million were made in 1965 and 4.7 million in 1966. The production rate for the first three months of 1967 was 50 percent higher than for the same period of 1966; thus, it is apparent that the number of color receivers is increasing rapidly.

Essentially all U.S. stations can now broadcast in color, and the proportion of color programs is high and increasing rapidly. For example, the majority of programs from New York City's nine commercial stations are in color; the same is true for Philadelphia's five stations. The competition for the audience is making the broadcasters adhere more closely to the transmission standards, so that the quality of received pictures today is good, and steadily improving. The rapid growth of color sets in the U.S.A. shows that the NTSC system is satisfying the public. We wish the same pleasant fate to our good friends in Latin America.

Performance of NTSC. Foreign engineers have often seen poor reception of color television in the United States, usually in hotels with poor installations, or in locations with strong echoes where the black-and-white image was also unsatisfactory. Even in good installations, the quality is often bad because of poor lighting at the camera, as in outdoor sports events in late afternoon or night games, or because of the transmission of programs that were recorded on now-obsolete tape machines. In these cases the poor quality results because wrong values of R, G, and B (red, green, and blue) are presented to the encoder by the camera or recorder. This would result in poor color from PAL, SECAM IIIA, SECAM IV, or any other system.

Princeton, N.J., the home of the RCA Laboratories, is about 80 km from both New York City and Philadelphia. In spite of the distance, normal television reception in homes is good, but not exceptionally good, as the signal strength is below that considered desirable by European broadcasters. Princeton homes are convenient places in which to compare the performance of New York and Philadelphia stations, and to study the effects of the distribution links and propagation. It is possible to compare NBC programs as they are emitted simultaneously from New York City on Channel 4 and Philadelphia on Channel 3. The same is possible for CBS programs on Channels 2 and 10 respectively. The Philadelphia stations are often at the end of a 3000-km network link that originates in New York City.

Color resolution is eminently satisfactory. (All sets in the U.S.A. use a bandwidth of about 0.6 MHz for both components of chrominance, instead of 1.5 MHz for the I and 0.5 MHz for the Q components, which the signal provides.) The old problems of differential gain and differential phase have disappeared, except on some afternoon programs that, unfortunately, still use old tape machines. No adjustment of the color controls is necessary when receiving any one channel, except to satisfy individual taste. The improved registration of color dis-

I. Main characteristics of television signals in Regions 1 and 2*

	Region 2	Region 1			
		PAL		SECAM	
		Western Europe	Great Britain	France	Eastern Europe
Channel bandwidth, MHz	6.0	8.0	8.0	8.0	8.0
Sound					
Intercarrier frequency, MHz	4.5	5.5	6.0	6.5	6.5
Modulation	—	FM	FM	AM	FM
Color subcarrier					
Intercarrier frequency, MHz	3.57+	4.43+	4.43+	4.43+	4.43+
Modulation	—	QUAM†	QUAM†	FM	FM
Video					
Bandwidth	4.2	5.0	5.5	6.0	6.0
Polarity	—	—	—	+	—

*Note: Because of different values of color subcarrier and sound frequencies, signals or tapes cannot be interchanged directly between Regions 1 and 2. †Quadrature amplitude modulaton.

plays, made possible by the automatic degausser, has eliminated most unwanted colored edges, for which quadrature crosstalk was often erroneously blamed.

Because of the large number of color stations available, it is desirable that sets be equipped with automatic chroma controls to keep the saturation constant when changing stations. For example, when comparing a program as is is received from Channel 2 in New York City and from Channel 10 in Philadelphia, it is often necessary to readjust the saturation control, generally because of the different characteristics of the receiver antenna and transmission line at the two frequencies.

In viewing old color films, which were not processed for color television, it is sometimes desirable to change the "saturation" control and occasionally the "hue" control.

When it is not possible for the broadcasters to illuminate the scene adequately, as was the case in an indoor fashion show, and is often the case for outdoor sports events in the late afternoon, it is sometimes desirable to adjust receiver saturation and hue controls. In changing rapidly between stations in New York, color differences can be perceived between the stations, due to the impossibility of standardizing the performance of nine sets of cameras from nine independently operating stations.

Maintenance of NTSC receivers. NTSC color television receivers do not require appreciably more maintenance than black-and-white receivers. The average number of service calls per set in 1964 was 1.8 for black-and-white, and 3.0 for color. The average yearly service cost per set was about $17 for black-and-white, and $35 for color. The need for service of color receivers is steadily decreasing.

Why PAL or SECAM?

PAL and SECAM were developed to prevent errors in hue due to unwanted phase shifts of the color subcarrier. The most important of these phase shifts is "differential phase distortion," produced by circuits that, through poor design or misadjustment, cause the hue to change as the brightness changes.

Studio equipment, distribution links, transmitters, and receivers can produce differential phase distortion, but these have been largely brought under control. Today this form of distortion is of little importance, except in old video-tape machines, which are being replaced as fast as new ones become available.

Poor gamma tracking of color cameras, to which NTSC, PAL, and SECAM are equally susceptible, produces an effect resembling differential phase distortion.

At NBC, comparison is made routinely of color programs as they enter the network and as they return after traveling more than 3000 km over the distribution links. Usually no difference can be seen between the two pictures, demonstrating that differential distortions have been eliminated. When there is a difference, means are available for correction.

Cost of PAL and SECAM receivers

PAL and SECAM cure the effects of differential phase distortion by the addition of delay-line circuits; these increase the cost of receivers appreciably. As shown by H. O. Wood and C. F. Otis, the relative complexities of decoders, using NTSC = 1.0 for comparison are PAL = 1.74 and SECAM = 1.46.[2] For this reason the PAL receiver is expected to cost the final purchaser about $30 more than an NTSC receiver. This extra cost does not include the cost of a "hue" control, which is believed to be essential for reasons to be given later.

In 1952, the NTSC standards included color-phase alternation (CPA), which was an early form of PAL. It was deleted from the final NTSC standards adopted in 1953 because the results did not warrant the additional complexity and cost. The writer feels that it is fundamentally wrong to introduce costly complications in receivers to cure correctable errors in the studio or transmitting equipment.

Color controls

A hue control is necessary for all systems,[3] but is much more costly in PAL or SECAM receivers than in NTSC receivers. For that reason it is usually omitted, with the excuse that PAL and SECAM do not need such a control. PAL and SECAM proponents claim that they can adjust the hue by operating differentially on the $R-Y$ and $B-Y$ components of the chrominance signal, after demodulation to video frequencies. This method is also available to NTSC receivers; but it is the only one available to PAL and SECAM receivers. It is the video equivalent of shifting the phase of the color subcarrier, and is subject to the same limitations.

The manual phase control in NTSC receivers, while not providing exact compensation for all hues, is a convenient, economical, and adequate means of adjusting the important hues, especially flesh tones, to the viewer's preference.

Since television pictures differ from reality in brightness, contrast, size, and perspective, it is likely that they should differ also in color. It is well known in photography that color reproduction should be "pleasing and believable," rather than exact. Exact reproduction, especially of flesh tones on the television screen, is often not "pleasing and believable."

It has been said, with some justification, that the viewer should see the color intended by the studio director as seen by him on the studio monitor. However, this is not always possible because the viewing conditions in the home differ from those in the studio control room.

Color controls are needed to compensate for:

1. *Differences in home ambient lighting conditions.* Programs intended to be seen in a home with artificial lighting will be seen during the daytime with ambient light, having a color temperature higher than 11 000°K, and at night with tungsten light at 3000°K. The eye adapts to these illuminations, and views the programs as having different colors during day and night. In the case of photographic color transparencies, a related problem is solved by the use of one film type for daylight and another for artificial light. Also, the color adaptation of the eye depends on the color of the viewer's surrounding such as the color of walls, which differs from home to home.

2. *Deficiencies in the original picture.* It is not always possible to illuminate a scene adequately without causing discomfort to spectators. In fashion shows, for example, the television camera is only an intruder. In outdoor scenes, such as sports events, daylight in the later afternoon is often inadequate in both quantity and color for color television.

Color balance occasionally varies from camera to camera. In addition, cameras do not make use of negative lobes of color mixture curves. The lack of these intro-

II. Subjective improvement, on a scale extending from same as NTSC (0%) to slightly better than NTSC (100%)

Location		Percent
Germany[4] (Average of 405 observations in mixed terrain)		26
Switzerland[5] (896 observations, often without line of sight)		50
Italy[6] (Mountainous terrain, Piemonte and Val d'Aosta, 1260 observations)		
NTSC rating	2.79	
PAL rating	2.03	
Difference:	0.76	76

duces color shifts. Sometimes there is also a difference in the color balance of color films, processed by different laboratories, which are cut into a continuous sequence, and in many old color films not processed for color television.

3. *Differences in the color perception of individual viewers.*[3] It is known that the perception of color varies greatly from observer to observer. For example, tests of 52 observers with normal color vision showed that they differed appreciably from each other in their sensitivity to the luminance $\bar{y}$ of monochromatic light. The proportional standard deviation $\sigma/\bar{y}$ was ± 18.4 percent for the luminosity $\bar{y}$ of the red primary at 0.61 μm and ± 30.0 percent for the blue primary at 0.47 μm. Assuming a Gaussian distribution of observers, this means that 32 percent of all observers, not necessarily the same ones, vary by ± 18.4 percent and ± 30.0 percent in their response to the red and blue primaries respectively. Reference 3 shows how this affects their color vision.

The need for a hue control is therefore well established. In any case, the viewer should be able to adjust the set to see the color he wants if he does not like the color he gets, while selecting between independently operated stations.

Effects of multipath (echoes)

Echoes affect both the luminance and chrominance components of the signal, and the damage to the luminance components is far more serious. At best, PAL (and SECAM to an even lesser degree) can correct only for the effect on the chrominance, leaving the more important luminance unimproved. Even if the colors of an echo were important, this color should be one that blends into the background rather than that of the original object. Using the effect of echoes on NTSC receivers for comparison, tests conducted under extreme conditions of multipaths in Italy, Germany, and Switzerland have shown that (1) when the NTSC picture is "excellent" (grade 1), the NTSC picture is superior to the PAL or SECAM picture because the subcarrier is less visible; (2) when the NTSC picture is bad (grade 4 or higher), the black-and-white picture is also bad; and the improvement due to PAL, while noticeable, does not make the picture acceptable; and (3) for pictures whose quality is "good" to "passable" (grades 2 and 3), the improvement due to PAL was as shown in Table II.

Note that the greatest improvement is less than 76 percent of "slightly better than NTSC." This is faint praise indeed! The improvement with SECAM was even less. It should be noted that even this very slight improvement is achieved only under the worst possible mountainous environment, and does not reflect the viewing conditions of the average viewer.[7]

It is much more economical for individual receivers or groups of receivers plagued with multipath reception to use highly directional antennas. Otherwise, all viewers, even those troubled with few echoes, will pay about $30 more per receiver; thus, a few viewers will receive slightly improved pictures.

The tests performed on SECAM in the mountainous northern part of Italy showed about the same performance concerning echoes as NTSC.

Quadrature crosstalk

Quadrature crosstalk can produce unwanted colored edges. SECAM is free of it and PAL cancels it by averaging the color on successive lines. It was one of the reasons for the use of CPA in the original NTSC standards. As stated earlier, CPA was deleted from the final NTSC standards because its benefits did not warrant the additional complexity and cost.

In a correctly designed NTSC receiver, quadrature crosstalk results from mistuning and, very rarely, from echoes that unbalance the chrominance sidebands. With either detent or push-button tuning, the stability of solid-state local oscillators can be made high enough for the receiver to remain stable once the station is accurately tuned in. This has been shown to be the case at UHF in the United States. Moreover, it is likely that better receivers on UHF will be equipped with some form of automatic frequency control, thereby making quadrature crosstalk of little consequence. The influence of echoes on quadrature crosstalk is not believed to be serious.

The colored edges due to misconverged picture tubes or to poorly registered color cameras are often confused with quadrature crosstalk. The recent introductions of automatic degaussing of picture tubes and of "separate-luminance" cameras have reduced greatly the number of complaints wrongly attributed to quadrature crosstalk.

Ultimate performance

Under good conditions all three systems can produce acceptable pictures; however, NTSC pictures are superior to those of SECAM and PAL, because the color subcarrier in its receivers integrates out more completely, and the spurious patterns it can create are less visible.

PAL and SECAM use a 1-H(one-horizontal-line) delay line to cure phase errors in equipment and links, which are no longer important. All receivers that claim to effect this cure require the delay line. NTSC can use a 1-H delay line to much better advantage to eliminate "cross color" or its inverse, rather than to eliminate troubles that should not exist.[8] Cross-color is present in all high-resolution, band-shared systems. However, only NTSC can eliminate it at no greater additional cost than that of a PAL receiver. Whether this improvement is used or not is entirely the option of the receiver manufacturer, who can use it in his more costly models.

The main objections to PAL as a broadcast signal are that it is not worth the additional receiver cost, cannot produce as good pictures as NTSC under good conditions, and is not as capable of improvement.

SECAM is open to major criticisms partly because of its use of FM for the chrominance subcarrier. When the SECAM signal is weak and below the limiting level of the

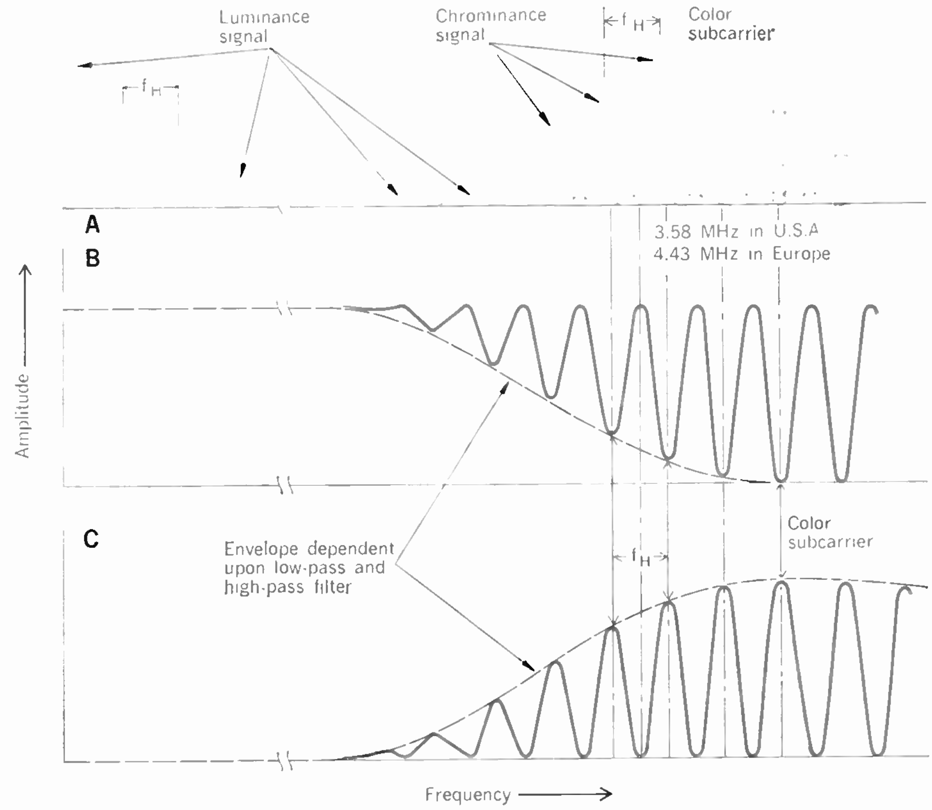


FIGURE 1.(A) The chrominance frequencies are interleaved between the luminance frequencies, thus sharing the luminance passband. (B) Luminance frequencies are passed by the "teeth" of the comb filter, and the chrominance frequencies are attenuated by the "valleys" of the filter. (C) In a second path, the comb filter passes chrominance and attenuates luminance frequencies. The comb-filter effect is obtained by using a 1-H delay line, and will correct cross-color and its inverse in NTSC receivers.

FM detector, as occurs when it is transmitted over long distances with equipment and lines which introduce differential gain effects, noise is stretched in time by the time constants of the limiters. Because of the 1-H delay, this stretched noise occurs, adjacently to itself, on successive scanning lines. This results in long, coarse, objectionable noise patterns, having a width of three scanning lines, called "silverfish." Patterns that contain much horizontal detail are often treated adversely by SECAM by introducing wild moiré effects.

Because the luminance is transmitted as AM of the main carrier and the chrominance as FM of the color subcarrier, it is not possible to fade one picture into another at the studio without decoding the signal and recoding it. In NTSC and PAL this is accomplished by attenuating the composite signal. This is not possible in SECAM because the chrominance does not depend on the amplitude of the subcarrier but mainly on its frequency.

In France SECAM is used with AM sound; and in Eastern Europe it is used with FM sound. In both cases, the frequency separation between the undeviated FM color subcarrier and sound carrier is about 2.1 MHz. In Region 2, FM sound must be used for compatibility with existing monochrome receivers. However, the frequency separation between the SECAM FM color subcarrier and FM sound is much less than in Region 1. This combination can produce annoying low-frequency patterns in existing monochrome receivers, and their presence should be investigated in field tests before the adoption of SECAM.

Cross-color[8]

In the NTSC and PAL systems, the frequency components of the luminance and chrominance signals consist respectively of line spectra whose component frequencies are separated by harmonics of the line frequency. The chrominance frequencies share the luminance passband, by being interleaved between the luminance frequencies, near its high-frequency end as shown in Figure 1(A). The decoder of the receiver detects the luminance frequencies as if they were chrominance frequencies. The detected luminance then produces spurious effects, such as making checkered patterns scintillate in color, which are called cross-color. Inversely, the chrominance signal adds to the luminance signal to produce crawling dots and edges, and also affects the brightness. These effects are quite objectionable, and are inherent in all band-sharing systems that use high-resolution receivers without comb filters.

Cross-color, and its inverse, can be almost completely eliminated from NTSC receivers by using a 1-H delay line to obtain the "comb-filter" characteristic [Fig. 1(B) and (C)]. The comb filter separates the luminance and chrominance signals so that they cannot contaminate each other, by providing two paths for the signal. In one path [Fig. 1(B)] the "teeth" of the filter pass the luminance frequencies, while the valleys of the filter attenuate the chrominance frequencies. The other path [Fig. 1(C)] passes the chrominance and attenuates the luminance frequencies. The two signals now behave as if they were transmitted by separate channels. Receivers can then use the full luminance resolution transmitted with no need for roll-off or notch filters.

The cost of a 1-H delay line used as a comb filter in an NTSC receiver is lower than that of a PAL delay line, because no switching is required. To add a comb filter to a PAL receiver requires a more costly line, with twice the delay (2H), because the PAL subcarrier is offset by one-quarter of the line frequency instead of one-half of the line frequency, as is the case for the NTSC subcarrier.

NTSC pictures using this technique are the best that have been produced by any system. The addition of a comb filter to an NTSC receiver requires no change in standards, and can be done at any time by any receiver manufacturer at his option.

Satellite communication[1]

Direct broadcasts from satellite to the home will be achieved in the proximate future. Such broadcasts may well cover several continents. It is therefore desirable that standards of transmission be as nearly alike as possible within Region 2, since this situation is no longer possible on a worldwide basis.

Because the horizontal line frequency of the 625-and 525-line systems is almost the same, and because the vertical scanning of receivers can easily be made to scan both 50- and 60-field pictures, receivers can easily be adapted to operate with both scanning standards. Bandwidth considerations and the location of subcarriers present a problem. This is a good reason for adopting one common standard for all the Americas which uses common bandwidth and subcarrier frequencies.

Transcoders[9,10]

Devices known as transcoders, which make use of phase alternation and subcarrier modifiers, have been developed that result in almost 100 percent cancellation of spurious phase shifts. Such devices can be used to radiate an essentially perfect NTSC signal. For example, a television program can be generated by means of PAL, transmitted by various phase distorting links, and then be "purified" and converted to NTSC by a transcoder.

This procedure satisfies the broadcaster by giving him greater phase tolerance and the means of radiating an essentially perfect signal. In addition, it provides the viewer with NTSC receivers that are less costly and can provide better pictures. It is thereby possible to achieve the best of the PAL and NTSC systems.

Conclusions

The NTSC color television system has been proved both successful and economical by field test and public acceptance to an extent that has not been matched by any other system. The principal reasons for the development of the more complex PAL, SECAM, and other variants of NTSC involve equipment limitations, which are considered to be of only temporary duration. If, to the contrary, there is a belief by others that these limitations are a cause for concern, it is recommended that the radiated signal be of the NTSC type obtained by transcoding from a PAL signal generated in the studio because of the following considerations:

1. The PAL–NTSC transcoder combination can generate an essentially perfect NTSC signal by purifying the original signal from spurious phase effects acquired during its generation and distribution.

2. The NTSC signal, once radiated, even in very mountainous regions, is deteriorated by echoes only very slightly more than the PAL signals. The improvement of PAL can be detected only when the monochrome picture is also badly affected by echoes. Even then, superior means, such as directional antennas, can be used by NTSC to cope with echoes.

3. The NTSC signal uses receivers of minimum complexity that, in the great majority of locations, can provide the best performance at lowest cost for the largest number of viewers. Known and proven improvements, such as comb filters, can be added to NTSC by any receiver manufacturer at his own option and without changes of standards.

4. The NTSC signal has the experience of many years of broadcasting and mass production, and presents no unknown problems. This experience is immediately available to manufacturers in Latin America. The transcoding from PAL to NTSC will (1) satisfy the need of the broadcaster for a signal that is immune to spurious phase shifts, and in addition, give him the freedom to purchase studio, distribution, and transmitting equipment best suited for his purpose; (2) satisfy the need of the viewer for a simple NTSC receiver of lowest cost; (3) give the receiver manufacturer the flexibility to supply receivers covering a continuous price range from the lowest cost with moderate performance to the highest performance at higher cost with such improvements as delay-line comb filters.

5. It can provide direct exchange of programs within the Americas, and it will make possible the direct reception into the home of programs transmitted by satellites operating for this purpose.

6. It will not make more difficult the exchange of programs on PAL or SECAM originating in Europe.

7. The introduction of PAL or SECAM as the signal to be radiated into Region 2 will prevent the direct exchange of color programs within the Americas, without facilitating exchange with the countries of Europe. In addition, SECAM presents special problems under conditions of poor propagation and in the studio. Its version for Region 2 presents problems, especially of compatibility, and therefore should be thoroughly field tested if its introduction is contemplated.

REFERENCES

1. Kiesling, J. D., "Spaceborne transmitter for direct color television transmission for satellites," *IEEE Trans. Broadcast and Television Receivers*, vol BTR-12, pp. 73–78, May 1966.

2. Wood, H. O., and Otis, C. F., "Relative receiver complexity of NTSC, PAL, and SECAM systems," *IEEE Trans. Broadcast and Television Receivers*, vol. BTR-12, pp. 79–82, May 1966.

3. Hirsch, C. J., "A study of the need for color controls on color TV receivers in a color TV system operating perfectly," *IEEE Trans. Broadcast and Television Receivers*, vol. BTR-10, pp. 71–96, Nov. 1964.

4. Dietz, Von V., Gileivitz, G., and Hopf, H., "Comparative observation NTSC and PAL colour television systems in the Summer of 1964," *Rundfunktech. Mitt. (Germany)*, vol. 9, p. 43, 1965.

5. "Colour television field trials with domestic receivers at normal receiving sites in fairly flat, hilly, and mountainous terrain," Res. and Testing Dept., Swiss PTT, Berne, Switzerland, COM. T (E) 166.

6. "Field test on domestic colour television receivers in Piedmonte and Val d'Aosta," RAI, Radiotelevisione Italiana Research Lab., COM. T (E) 154, Nov. 1964.

7. "The effect of multipath propagation on the reception of monochrome and colour television signals," British Broadcasting Corp. Tech. Memo. T 1075 Com. T (E) 175.

8. Parker, N., "The cost of using PAL or SECAM and possible improvements in NTSC receivers," *IEEE Trans. Broadcast and Television Receivers*, vol. BTR-12, pp. 159–161, July 1966.

9. Watson, S. N., "A survey of color transcoding," *IEEE Trans. Broadcast and Television Receivers*, vol. BTR-12, pp. 62–63, May 1966.

10. Castelli, E., "Standards conversion and chrominance transcoding problems in the exchange of television programs," *IEEE Trans. Broadcast and Television Receivers*, vol. BTR-12, pp. 43–45, May 1966.

The European viewpoint

Sir Francis McLean
Technical Director
British Broadcasting Corporation
London, England

The examination of the factors affecting the choice of a color television system carried out in Europe in the years 1962 66 was so exhaustive and well documented that it is difficult to add to it at the present time—nor would I disagree with the abstracts from this documentation quoted by Mr. Hirsch in his article. At the time the documentation was prepared, most people in Europe had very little information or experience about what was involved in actually running a color service; and the research was almost wholly derived from laboratory experimentation. Our experience in actual operation is still very limited but it does show that, although there are many differences in the performance of various types of coders and decoders, the problems associated with these are small as compared with the overall problem of maintaining consistent quality. Quality is enormously dependent on factors ahead of the encoder at the originating end and following the decoder at the receiver end. Observations in both Europe and the United States confirm this view.

Program interchange

The importance of program exchange grows continuously; and advances in communications, particularly by satellite, increase the demand for compatibility between all areas of the world and between all systems. Experience has shown that it is possible to convert in either direction between NTSC and PAL with a negligible loss of quality. The quality obtained by converting to and from SECAM is much less. It is to be expected that a somewhat greater degradation of picture quality in both black and white and color will result, but the extent of this depends on whether the SECAM system will be operated with the same degree of line frequency stability as is essential for the NTSC and PAL systems. The biggest problem in program exchange has been, however, not the change between the coding systems but between the 50-field and 60-field standards of the various areas of the world. This problem has now been solved, and, although I think it is obviously true that it would be better not to have to make the coding system conversion, it cannot be thought that this additional process really complicates the situation. I cannot add to what has already been said about the SECAM system, but I can say something about PAL, based on about one year of experimental operation of PAL and about one-half year in regular service, during which color has been transmitted over the BBC's second television program network.

Factors in broadcasting

From the point of view of the Broadcasting Authority, the PAL signal, as compared with NTSC, has two distinct advantages: (1) It is very tolerant of phase and differential-phase distortions. (2) The lack of response symmetry of the chrominance channel, with respect to the color subcarrier, causes very little degradation to the quality of the picture as displayed on a delay-line receiver.

The greater tolerance of the PAL signal to phase and differential-phase distortions is of considerable importance in the operation of a large and complex broadcasting system. Such distortions can arise at many points in the broadcasting chain, including video tape recorders (normal broadcasting techniques often demand two recording processes, and frequently three), picture-mixing equipment, long-distance link transmission, and transmitters. In our experience it is not easy to hold differential-phase distortion to a limit lower than about 20°, taking into account all the processes that a color signal can be put through—from leaving the studio camera to actual broadcast by the most remote transmitter on the network. Such an amount of distortion, according to tests carried out by the EBU (European Broadcasting Union), yields a quality of picture on the NTSC system that lies between fairly good and rather poor in the presence of this distortion alone. This kind of distortion with the PAL system has a negligible effect on the quality of the picture.

The greater tolerance of PAL in the presence of asymmetry of the frequency band carrying the chrominance signal can be quite important in the case of a system that has a relatively small frequency spacing between the color subcarrier and the sound carrier. Although the principal advantage of this tolerance lies in the design and operation of the receiver, it also can be a factor in the design of the broadcasting system. On the other hand, the use of PAL undoubtedly adds complexity in waveform generation and in picture-mixing equipment.

It appears that on balance PAL is advantageous from the point of view of the broadcasting authorities, particularly for very long, land-based transmission systems in which the signal is repeatedly demodulated to video frequencies.

Receiving PAL

On the receiver's side the biggest advantage of PAL lies in its resistance to effects caused by lack of chrominance channel response symmetry, provided that a delay-line-type receiver is used. Accordingly, the quality of the picture produced by the receiver is much less affected by mistuning, by insufficient or misshaped bandwidths, and by effects of misalignment of the receiver with time. There are grounds for hoping that a PAL receiver will be less sensitive to drift with time than will be an NTSC receiver. In addition, a PAL delay-line receiver is less sensitive to any changes in the phase of the locally regenerated color subcarrier that is used for demodulating the chrominance signal.

Cost factors

I believe that, overall, it is possible to run a high-quality color service on either NTSC or the PAL system. In some important aspects more care is required with NTSC than with PAL, but this additional care does not impose too difficult a requirement. On the other hand, the PAL receiver with delay line is more complex than the NTSC receiver and so must cost more. It is very difficult to assess just how much more this cost will be but, at the present time, I think it would be only a few percent greater than that of an NTSC receiver. The cost of the color receiver is, however, so dependent on the cost of the color tube that if this cost itself were brought down then the increased cost of the PAL receiver could be relatively greater. It is possible, of course, to reduce the cost of the receiver on the PAL system by use of the so-called simple PAL receiver, that is, a receiver with alternate line switching, but with the eye rather than the delay-line circuit making the average between the successive lines. The performances of this type of receiver, as shown in the EBU documentation, were appreciably worse than those of the delay-line receiver.

Color quality

There are many factors that can contribute to poor performance at the originating end of a color service, such as incorrect lighting, incorrect setting up of cameras and picture monitors, and lack of suitable characteristics of, and of adequate identity, between the *R-G-B* chains. Means and methods of adjusting and operating the origination and monitoring equipment are not yet fully defined anywhere, and obtaining consistent color quality still contains an appreciable "art" component rather than being purely scientifically based.

Video tape recorders equipped with the appropriate color-correcting devices, in the best state of performance, give approximately equal results when used on either system. The PAL signal is more tolerant of the inevitable misadjustment that occurs under operational conditions. Moreover, the apparatus necessary with NTSC for correcting the mean velocity along a scanning line is not required for PAL. Film equipment is adequate but is, of course, entirely dependent on the quality of the color film put into it. The TARIF (Television Apparatus for Rectifying Indifferent Film) device can improve a film considerably, by separately amplifying and correcting the separate *R-G-B* gains before encoding.

Possibly the most difficult area in which to maintain adequate picture quality is the receiver. The design and manufacture of a receiver that will initially have good quality and will maintain good quality over a period of time, in a competitive market, is not an easy problem to solve. The area of the receiver probably needing most attention is the *R-G-B* side, where it is essential to maintain a satisfactory white balance that must compromise to give "reasonable and pleasing reproduction of both artificial light and day pictures," and to maintain this identity of balance at all picture amplitudes. Further accurate decoding of chrominance and accurate matrixing of luminance and chrominance are essential if the receiver is to give a reasonably faithful portrayal of the transmitted pictures. Factors such as these represent a very serious problem for the receiver designer but are, of course, independent of the choice of any coding system used at the present time.

Claude Mercier

Technical Director, Engineering and Operations
Office de Radiodiffusion-Télévision Francaise
Paris, France

In his article, Mr. Hirsch has analyzed the various problems relating to color television systems. Since the editors of IEEE SPECTRUM were kind enough to ask me for my comments, I am presenting some remarks on the operation of a color television service. My article does not pretend to be on the same level as the detailed report of Mr. Hirsch; my comments are simply inspired by recent experiences during the introduction of color television in France, and are those of a director of operations and not of a laboratory man.

Inevitably and rapidly, the development of television leads to color. European television stations have already started to fit it into their routine operations. For example, the Federal Republic of Germany, the ORTF, the BBC, and Dutch Television now have a regular color television service, and the RAI is about to start it. There are still other broadcasting organizations ready to enter this field. Black-and-white television is beginning to recede into the background, since the viewer who has seen color television wants even more of it. The trend is unavoidable, and suggests that in countries where television is to be introduced in the near future, it would be just as well to avoid starting with black-and-white television and to start, instead, with color. The subject chosen is, therefore, quite timely.

Coding system

When television entered the planning stage for the transition from black and white to color, it seemed quite a hazardous undertaking; and elaborate studies were carried out over a span of several years by individual companies, as well as within the European Broadcasting Union, whereby the emphasis was put on the overall subject of coding systems. The coding system is indeed a vital factor in the equipment of color television, and its choice is a matter of great importance.

Quite obviously, the intrinsic quality of the NTSC system has never been contested, nor has the ingenuity of the American technicians who overcame the difficulties that beset the path to color television, at a time when most of the broadcasting enterprises in the world dealt only with black-and-white television. Just the same, the tolerances that must be constantly observed for color transmission are considerably more exacting than those for black and white. In order to achieve a more effective use of the transmitters and the transmission channels, the characteristics of which are relatively heterogeneous in Europe, European technicians worked out some systems that have many points in common with the NTSC. But certain equipment is involved that renders these systems more resistant to the unavoidable distortions in the transmission channels.

Color in France

To my personal knowledge, such improvements in the operational use have been achieved in France, where the ORTF started a television service on October 1, 1967, which had been preceded by a period of experimentation, so that production as well as transmission have been in operation for about one year. From the start, the color transmissions were broadcast over the entire second network, consisting of 50 transmitters on the IV and V bands (19 of which are of insignificant power, namely, 50 kW), distributed over France and interconnected. Production is centered in Paris, and broadcast is effected over a microwave system of 6000 km. The networks used are those that had been set up for black-and-white television some years ago. Aside from some special adjustments and maintenance work, the color programs could be carried over the transmitters without any expense whatsoever for transformation. At the rate of almost 35 hours of programs and color demonstrations per week since the start of the regular service, no distortion or deterioration in quality has been observed so far in the reception control centers installed anywhere in France under conditions existing at the viewers' respective locations. Finally, during difficult transmissions (for example, from Paris to Algiers with its tropospheric links) or during transmission over satellite (from the ORTF studios at Montreal over Mill Village and Pleumeur-Bodou), the colors showed no appreciable change in quality. As far as production equipment is concerned, it is, of course, necessary to acquire cameras and film television scanners specialized for color. The ORTF has thus equipped within a few months four color studios, with two mobile units, 12 film television scanners (16 mm and 35 mm), one laboratory for color film development, coders, monitors, etc. We have, of course, also sought to keep expenses down by using for the color television process certain materials we already had for black-and-white television. For our French standards, we have, for example, been able to retain the sync generators (for which a precise accuracy of frequency is not needed under French standards) and the video tape recorders on hand, which are at present used for black and white as well as for color.

Program exchange

An important point raised by Mr. Hirsch in his article is the exchange of programs. It is regrettable that there exists a multiplicity of television standards throughout the world, already divided as far as black-and-white television standards are concerned, which aggravates the situation of the standard color transmission. But should we regard this aggravation as critical? Upon analyzing the question, one realizes that the differences in coding in Europe are not very important, and are not the source of serious complications. As a matter of fact, a satisfactory solution to the transcoding problems of PAL and SECAM was found within a few months, involving only impairments of a second order, since it is generally the quality of the receiving installation that largely determines the quality of the image at the viewer's end. By television standards it is the number of lines and rasters that are of prime importance for the signal transformations. We are fortunate that we have in Europe the same field frequency; the power distribution networks all operate at a frequency of 50 Hz.

In 1948, France chose for its first very-high-frequency network a standard of 819 lines; but the ORTF adopted the generally accepted television standard of 625 lines in 1960 for the ultrahigh-frequency band in order to facilitate the European standardization of the channels in bands IV and V. It thus uses the 625-line norm for color transmission over its second network, and omits color transmissions over the first network (VHF). It was in this manner that the PAL-SECAM transcoding problem was solved, since it was found to be impossible to establish complete agreement for all the parameters of color standards.

The development of television will inevitably lead, sooner or later, to ultrahigh frequency. It might well be that in the latter frequency range it would be advisable to study the problem of UHF standards, so that an exchange with European networks will be facilitated if the need arises. Obviously, this in particular concerns those who operate the 50-Hz power-supply networks. When the field frequencies are different, signal conversion becomes difficult. The BBC must be credited with having built such a converter utilizable for color, but it is by no means certain that the solution of the problem is as simple as it would be if two color systems having the same standard of 625 lines were to be transcoded. It is very possible, however, that future international standards for program exchanges over satellite will adapt the 625-line system; this would be quite feasible for the United States. Under such circumstances, the Latin-American countries, whose cultural ties with Europe are of particular importance, would benefit from the 625-line system, since it would let them become members in the network of a future international system embracing a steadily growing audience, a network, moreover, in which the United States could easily participate.

Different coding standards

Mr. Hirsch suggests the use of different coding standards in production and transmission. However, there seems to be no merit in this suggestion. In any case, for a network such as that of ORTF, it would present disadvantages, since the number of picture-producing centers, as well as the need for regional broadcasting, would make it necessary to install a transcoder in *each* transmitter. Such a solution would be burdensome; it would complicate the engineering system and, in addition, it would be necessary to train television engineers to handle two types of signal. Of course, experience has taught us that training programs for technicians are important and exacting.

In my opinion, there are very few arguments that could convincingly support the proposal to use different coding standards in production and broadcasting.

I shall now briefly discuss the problems connected with receiver prices. No doubt improvements could be made to lessen the effect of a crosstalk quadrature, obviously to the detriment of the prices of the receivers, but they might revive certain arguments about the price differences relating to European coding decisions. I could also discuss the rather small price differences in receivers, as well as the fluctuation in prices relating to improvements in transmission equipment. The calculations, in practice, are found to be quite different; and there are few countries that would consider granting television enterprises the additional credits necessary for improving a network if this can be avoided. Finally, when the total price of a receiver is considered, including the cost of the coding circuits, the defects in the delay line are of no importance in the sales development of receivers. I should also like to point out that in France we greatly appreciate not having to adjust the color reception of the receiver. We have endeavored to obtain receivers which, once they are adjusted, function so reliably that the television viewer need not try to adjust them. In fact, we believe it advisable, from our French point of view, that the images created by the television system are neither intentionally nor unintentionally altered in the receiver. Once the color image is produced in the studio, or for television film, it should be reproduced precisely, and not be modified more or less arbitrarily. When a scene is shown on the screen it should be viewed as if there were no intermediary of television transmission. In the case of artistic creations that are not to be disclosed or of documents—educational material, for example—that should not be distorted, this condition of a precise reproduction appears self-evident; all responsibility for a true image transmission and reception rests with the professional photographer. Experience has shown that in France these exigencies of steadiness of operation of the receivers and true reproduction are fully met by the system used by the ORTF.

In closing, I would like to express my thanks to the editors of IEEE SPECTRUM for giving me permission to comment on Mr. Hirsch's excellent article. Seeking the opinion of a European on a study concerning the American continent reflects a spirit of direct and constructive collaboration, which makes the contact with our American colleagues so pleasant. As I am not very familiar with the peculiarities of Region 2, I might have underestimated certain aspects that my American colleagues might consider to be of prime importance. In any event, I have limited myself to certain operational aspects of color television from the point of view of my personal experience with the ORTF system.

Dr. W. Gerber
Technical Director
Schweizerische Post-, Telephon-, and Telegraphenbetriebe
Bern, Switzerland

In nearly all European newspapers today a distinction is made between the so-called "American," "French," "German," and "Russian" color television standards. Such a simplification is certainly not correct. The names of the more or less different standards proposed or in use should be denationalized. As a matter of fact, in 1938 G. Valensi, former director of the CCIR, stipulated the discrimination of luminance and chrominance in a signal to be transmitted. Fifteen years later, the FCC decided in favor of the NTSC system, a very ingenious system and probably one of the most important technical achievements since World War II. What could we do today with our 200 000 000 monochrome receiving stations throughout the world without the philosophy of compatibility?

It is well known that "weighting factors" are not the same all over the world. As a matter of fact, in Europe we are relatively sensitive to the so-called "differential-phase gain," in connection with distortions of the hue. That is why Henri de France proposed, in 1956, a variant of the NTSC system, in which the disproportion of horizontal and vertical resolution is reduced by omitting half the vertical chrominance information in order for the system to become independent of the aforementioned phase distortions. In his procedure only one of the two chrominance signals is transmitted, alternating line by line. In the receiver, the other one is delayed and then correctly added. In 1962 a further step foward was made by W. Bruch, who proposed to alternate, line by line, the polarity of one of the two chrominance signals; as a result, the phase-distortion problem is more or less out of discussion. This system has the further advantage that it can be considered a straightforward variant of the classic NTSC color system.

Two systems

Actually, two different types of systems must be distinguished: the 525-line system, with 60 fields per second, and the 625-line system, with 50 fields per second. The respective line frequencies of these two systems are about the same, namely, 15 734 Hz and 15 625 Hz. A further coordination has been considered by international bodies, with the future of color television in mind, but no decision has yet been made. Concerning the field frequencies it is unanimously agreed today that a television system can operate independently of the mains frequency if certain precautions are taken. Furthermore, it is known from classical experiments with a Mechau projector that, for the reproduction of movement, a picture rate of about 12 Hz would be sufficient. Flicker problems do not exist in nature. Our so-called technical solutions are still far away from the natural facts of human physiology. Therefore, in actual use the unification of the two systems does not appear to be justified.

Proposals for Latin America

1. The channel width in Region 2 has already been unified at 6 MHz, and this situation should be maintained throughout all the frequency bands.

2. Since the bandwidth is relatively small, the choice of a color standard can be restricted to a version of the NTSC system or its PAL variant.

3. The existing 50- and 60-field systems in Latin America may be further coordinated in order to facilitate program exchange. Solutions such as common line frequencies, common subcarrier frequencies, or so-called dual-standard receivers are envisaged.

4. Finally, it is foreseen that full electronic standards conversion will be possible in the near future, as a result of recent developments of the BBC.

Henri de France

Compagnie Francaise de Télévision
Ansieres, France

In the utilization of a television system, especially in the case of color television, the reception problems are certainly among the most important ones. The multiplicity of receivers in use and the economical incidence thereof, the propagation and covering difficulties in large cities as well as in wide rural areas, and the lack of experience of the users are many direct factors involved in the choice of means required and compromises to be accepted.

Economical aspects

It is not deemed necessary to treat again the questions of production costs that have been the object of numerous comparative studies. These show that the SECAM delay line by no means substantially increases the price of the receiver, which, for equal quality and production, can be manufactured at a lower cost than the NTSC receiver (CCIR 1963 66 Doct. XI/177). The reason for this is that both SECAM and NTSC are in fact line-or dot-sequential systems, but the former contains the necessary time references within the conventional scanning signals, whereas the latter has to produce them from the burst with a much higher precision. This is also the source of the SECAM advantages in the reliability and simplicity of use of the receiver.

Technical requirements

The NTSC system. The differences between NTSC and SECAM result exclusively from the treatment of the chrominance signals; the other parts of the receiver are identical. Consequently, one must examine, on the one hand, the synchronous detectors and the production of the subcarrier, and, on the other, the frequency discriminators and the synchronization of colors. The ultrasonic delay line is a passive component whose manufacture was a problem in the past, but whose life duration and performance stability are now perfectly acknowledged, particularly on the SECAM, which allows for very wide utilitzation tolerances.

The NTSC synchronous detectors must be fed by signals having very strict phase relationships (a few degrees at 3.58 MHz). This requirement is equivalent to a precision of the order of 8×10^{-9} second, to be maintained in all circumstances during the entire life of the receiver. This rather severe specification is required for the separation of the two quadrature components. The same difficultly appears in PAL regarding the time factor of the delay line, which requires an equivalent precision. In this matter, SECAM is ten times less demanding.

The restoration of the subcarrier, required for the synchronization of the colors, makes use of a burst to provide the necessary phase reference. Here again, there is the need for excellent precision, which cannot be obtained without a manual control at the user's disposal; moreover, this control, acting upon the hue, must be skillfully effected using basic colors.

The SECAM system. The SECAM demodulating methods are more direct, and are comprised of two frequency discriminators assigned to the two chromatic components respectively. In each discriminator both the linearity, which depends on the bandwidth, and the tuning frequency must have good stability. In particular, the tuning frequency generally defines the dc component of the chrominance signals and, in order to avoid hue distortions, it should not differ by more than approximately 20 kHz from its nominal value. This condition, analogous to that of audio-frequency discriminators, can be easily satisfied. It is absolutely unnecessary to put this control at the user's disposal.

Owing to the use of frequency modulation, it is likewise unnecessary to provide a manual or automatic gain control for the chrominance (or hue saturation). The presence of amplitude limiters ensures a perfect level constancy. It is, therefore, the duty of the broadcaster to fix the luminance/chrominance ratio at the correct value desired by the program producer, since this ratio will be maintained without distortion during the entire transmission.

Thus it is seen that the SECAM receiver is capable of ensuring, in an easy and reliable way, the decoding of the composite signals, while leaving only the usual control knobs for tuning, brilliance, contrast, and sound volume.

As far as color synchronization is concerned, there is also an essential difference in comparison with the NTSC system: the SECAM decoder switch can have only two different states alternating at each scanning line. Therefore, it is sufficient merely to effect one checking operation, but not the phase control of the switching device. Whenever the switch is in the wrong state, a phase-correcting operation is initiated rather than a continuous control action. It is readily seen that this operation is ensured in a more reliable way than in the NTSC.

Consequently, as far as the color signals are concerned, the SECAM receiver is entirely automatic. This is an advantage, as automatic control is obtained without any complication or degradation of receiver performance.

Inserting a delay line of 63.5 microseconds into the receiver has proved to be an excellent operation, a further proof of which lies in the fact that the NTSC is considering its use for "cross-color" error correction. However, it is thought that the way in which SECAM first used the delay line is by far the most efficient one for providing sturdiness and automaticity.

Since the entire chain still comprises many sources of distortion, (for instance, those resulting from the instabilities of the three-color picture tubes), use should be made of reliable means to ensure exceptional fidelity in the transmission of signals as delicate as color.

Friedrich G. Weyl

Bonn, Germany

In a world without common standards in electrical engineering (frequencies, voltages, etc.) there is no international standard for television—monochromatic or color. The problem of standardization goes all the way back to the discovery of the electrodynamic principle by Werner von Siemens. The first standard was 50 Hz, 220/380 volts, in Germany. Why did the United States not adopt this first standard?

Germany can claim priority for itself with the first completely electronic television; Walter Bruch constructed the camera and operated it during the Olympics of 1936. The ingeniousness of the NTSC color television system, a pioneering feat of American engineers, is neither disputed nor deprecated by most German engineers. This statement does not imply, however, that German engineers would adopt a new system (for example, the NTSC color system) for Germany without any criticism.

American engineers introduced the NTSC system prematurely—meaning that it was introduced before it was completely perfected, when it still contained a considerable number of defects (such as phase sensitivity). Various engineers immediately proposed improvements (such as multiburst). During the first year or two after the introduction of the NTSC system in the United States there still remained the possibility of a considerable improvement of the system due to the small number of color receivers. However, the NTSC system was not improved, and is still as imperfect as it was in the beginning—and there are now over ten million color receivers in the United States.

Under these circumstances, the NTSC system will not be able to attain as qualitative a standard as PAL. CCIR represents Europe, with its many nations; NTSC represents only two nations, the United States and Canada. The known shortcomings of NTSC made it impossible to adopt the system in Europe, with its many mountainous regions, without modifications. The correct reproduction of colors in the color receiver is an important feature in Europe. The negligence of American engineers prevented a common (international) standard for the United States and Europe. It is useless now to complain about the absence of a common standard, but this does not change the facts.

Telefunken in Germany has improved the NTSC system to the extent that it is complete and gives the correct colors at the receiver. The total development cost for the resulting PAL, amounting to over DM 100 million (about $25 million), has not been spent in vain, as we already knew only three months after its initiation. The PAL system is flawless, with the correct colors, up to a noise level at which even a monochromatic picture is of poor quality. The adoption of the PAL standard in the European Broadcast Union was accepted after long and exhaustive tests with the participation of industry. The results are recorded in the "Grey Book" and the "Orange Book."

The essential difference between NTSC (purely quadrature modulation) and PAL (with phase inversion) is clearly expressed in the following equations:

Addition:

$$[U + g(t)V + U - g(t)V] \cos \omega_0 = 2U \cos \omega_0 t$$

Subtraction:

$$[U + g(t)V - U + g(t)V] \cos \omega_0 = 2g(t)V \cos \omega_0 t$$

The phase inversion $\pm g(t)$ is canceled in the PAL switch ($\pm jV$), so that no phase errors are visible in the *R-G-B* matrix.

In the NTSC system, chrominance and luminance ($F + Y$) alternate in the spectrum. In the PAL system the color components, red and blue, are on both sides with respect to luminance at half the distance ($U + Y + V$) as compared with NTSC. The segregated filtering out of a spectrum group each (U/V) lets the PAL addition and subtraction circuit act in the fashion of a comb filter. A comb filter, however, makes it possible to improve the signal-to-noise ratio in the color channel.

Herbstreit and Pouliquen (March 1967 SPECTRUM) speak *only* of NTSC and SECAM and mention PAL only in passing. Why such disrespect of this thorough German endeavor, which, through PAL, has created a really perfect color television system? SECAM does not reproduce all color points on the picture screen as correctly as NTSC (without phase error) and PAL; SECAM leaves it up to the eye to find a subjective middle value.

Service for standard PAL receivers

For the well-trained television service technician, the servicing of a PAL receiver is no problem if he has had adequate monochromic experience. The German radio and television industry has developed and designed (with well-known German thoroughness) all PAL elements in the color receiver so well that during the first few months of operation (July to October) practically only convergence corrections have been necessary.

The defects in the PAL receiver are practically the same as in monochromic television receivers: tubes and transistors. The location of such standard defects is relatively simple as a result of the excellent and detailed service manuals. I have not yet encountered defects in specific PAL elements.

Conclusion

American engineers should terminate and bury the discussions, which are useless for their NTSC system. The discussions improve neither NTSC colors in the presence of phase errors nor the system itself. If the vigor of these discussions had been applied to the improvement of NTSC color television at the right time, the results certainly would have been a better system. The million NTSC color receivers await a real improvement in color reproduction in the presence of phase errors.

Comments by the author

It is always a great pleasure for me to renew discussion with such old and good friends as Sir Francis McLean, M. Claude Mercier, Dr. Walter Gerber, and M. Henri de France. I value their comments, even when I don't agree, and I respect their opinions. However they have not answered my main argument for the adoption of NTSC by all the countries of Region 2, which is at the beginning of my article, and which I shall paraphrase in the following.

All the countries of Region 2 (Western Hemisphere) are limited by treaty to television channels that are only 6.0 MHz wide, and the European countries use 8.0-MHz channels. Programs originating in one region cannot be used in another without major signal processing, because the intercarrier frequency spacings of the sound and color subcarriers are different for the two regions. These spacings, and other characteristics, are shown in Table I of my article. On the other hand, programs can be exchanged between all countries of Region 2 without such signal processing, because they use the same, or very nearly the same, sound and color intercarrier frequencies. This lack of need for signal processing is always desirable; but it will become essential, in the proximate future, when reception from synchronous satellites directly into the home will be achieved.

The fact that some countries in Region 2 use 625/25 and others use 525/30 scanning standards is of little consequence, because receivers can easily scan both standards by an adjustment of the manual speed control. This was demonstrated at recent tests of NTSC in Buenos Aires, using receivers made in the United States.

The adoption of NTSC by all the nations of the Western Hemisphere will permit the direct exchange of programs among themselves without the need for standards conversion, and without making more difficult the exchange of programs with Europe.

The adoption of PAL and SECAM, modified for Region 2, would create a new standard that would be incompatible with European PAL or SECAM and also with NTSC. I believe that we have too many standards now, without creating useless new ones.

Another advantage to Latin-American set manufacturers is that they can use United States designs with little, if any, change. The adoption of PAL or SECAM will not allow the use of existing European designs without fundamental modifications.

Now that Europe has begun to broadcast color, it finds that factors that are common to all systems affect the performance much more than the details of the system. I am happy that Sir Francis and M. Mercier agree on this point. Such factors influencing the transmitted signal are: operating experience; adequacy of lighting, especially of outdoor sport events in the late afternoon; uniformity of film processing by different laboratories; tape quality; and camera performance, which, in turn, requires that all cameras have the same color balance, adequate registration, and good gamma tracking. Disturbing factors affecting the receiver include poor registration, wrong white balance, and poor color tracking. These disturbing factors affect PAL and SECAM as well as NTSC, and can no longer be blamed on NTSC alone.

Henri de France and Dr. Bruch each invented a different and very clever cure for "differential-phase distortion," an illness that affected NTSC in its early days but no longer exists, except on rare occasions. These cures became the SECAM and PAL systems—but Europe could not agree on which was the better, and adopted both. PAL and SECAM operators will find effects that resemble differential-phase distortion, but will call it "poor gamma tracking of cameras." In the case of SECAM some side effects of the medication are worse than the illness, and can produce "silverfish" and poor compatibility.

Finally, I disagree with the proponents of PAL and SECAM who claim that a hue control is not necessary. I suspect that they are making a virtue out of a necessity, because such a control is difficult and costly to implement in PAL and SECAM receivers. In my article I give reasons why hue and chroma controls are necessary, even when the system works perfectly. Moreover, in New York City, where nine simultaneous color programs can be compared, and each is produced by an independent organization with different lighting, different makes of cameras, and differently processed films, it is good to be able to get the color you like if the various studios transmit colors differently.

I regret that I cannot answer Herr Weyl's comments. I do not know why the United States did not adopt the 50-Hz (220/380-volt) power system advocated by W. Von Siemens before he died in 1892; perhaps we are just too stubborn.

Again I wish to thank Sir Francis, Dr. Gerber, and Messrs. Mercier and de France. Perhaps we can renew the discussion "avec l'inspiration d'un vieil Armagnac à Paris." In that case I would hope that the discussion never gets resolved.

Human enhancement: Beyond the machine age

Self-organizing systems of hardware, computer simulations of evolutionary logics, artificial intelligence, novel conceptions of time and language—all these could herald the design of evolutionary systems that work for human enhancement

Warren M. Brodey *Massachusetts Institute of Technology*
Nilo Lindgren *Staff Writer*

Our previous article proposed a very broad evolving and ecological "think," urging designers to begin considering new types of systems, some to incorporate artificial intelligence, to work for the enhancement of human powers. Here we describe some bases of that think, and attempt to join researches that, as they evolve, could form the ingredients of evolutionary systems. In the broadest sense, we attempt to interpret aspects of today's science and technology within the evolutionary framework, hoping to stimulate a dialogue among all kinds of system designers.

> *"Nothing at first can appear more difficult to believe than that the more complex organs and instincts should have been perfected, not by means superior to, though analogous with, human reason, but by the accumulation of innumerable slight variations, each good for the individual possessor."*[1]
>
> —*Charles Darwin*

In our first article, we stressed the need for an evolutionary technology, citing the urgency of our present environmental situation, owing to the increasing disruption of our natural ecology. We tried to bring forward the important central role of "dialogue" in using the evolutionary process, especially as it might be applied to human enhancement.[2]

In this present article, we describe certain "building blocks" now becoming available that, as they accumulate in number and variety, could be used in evolutionary systems. The computer is the herald of major change. It is the one ingredient connecting hardware and software advances that could contribute to the design of evolutionary systems.

Our society has been willing to develop weapons systems, space systems, highway systems, telephone systems, but it has manifested no coherent, coalescing drive to make devices and systems to satisfy human users in the unique ways we have described; namely, that they should be responsive and interactive in an intelligent way with their individual users. Such systems are possible, although perhaps not probable, unless we are willing to create them.

On artificial intelligence

The crucial ingredient in evolutionary systems, we believe, will be artificial intelligence, which itself may come about through computer simulation of evolutionary processes. By artificial intelligence, we mean the acquisition by machines of capacities of pattern recognition, problem solving, self-improvement through an "understanding" of their own problem-solving processes, "creative" thought, and so on—capacities normally associated with the mentality of man as his distinction.

In our previous article, we perhaps failed many readers by supposing them familiar with the vast recent literature on artificial intelligence in the technical journals.[3] That machines can *think* has become as manifest to us as that machines could *work* became to our forefathers. The differences between kinds of hardware and software, between dry and wet thinking, are beside the point in this article.

We were more remiss as to *distributed intelligence*, which is related to the efforts directed toward making available computer capacity (say, by time-sharing) to the many, in the same way that linked utilities supply electric power to almost any place. Project MAC (for multiaccess computer or machine-aided cognition) is the most publicized example of the on-line computer time-shared system. Naturally we jumped from such distributed computational power to the day when artificial intelligence will be a common facility, like the mail and the telephone.

By lumping together small units of computation such as tiny chips of integrated circuits with the global tool envisioned by some scientists, we undoubtedly irritated many readers and led others astray. But we sought a philosophical unity. By juxtaposing large and small systems, ordinarily held separate, we sought to focus a perspective of evolutionary kinship. Designers of toys, furniture, homes, cities, transportation and communication media, political and even symbol systems, can capitalize on the new possibilities.[4,5]

Artificial intelligence will not come simply or all at once, but designers should feel some sense of urgency since parts of it are already here. We take M. Minsky's point seriously when he writes: "Once we have devised programs with a genuine capacity for self-improvement a rapid evolutionary process will begin."[6] Those who wish details on progress in artificial intelligence will want to read M. Minsky's now classic summary[7] and R. J. Solomonoff's follow-up review.[8]

It should be clear to designers that the melding of artificial intelligence functions (information manipulators) and physical parts (energy manipulators) will create new entities. For instance, the designing of integrated circuits gradually forced circuit designers to abandon their concepts of discrete components (resistors, capacitors, diodes, etc.). In integrated circuits, these functions are no longer discrete, but are blended and distributed throughout the chip. Thus, the circuit components do not exist in their old form, either in their physical formation or in the designer's conception of them. That transition has taken nearly two decades. It will undoubtedly take us much longer to acclimate ourselves to the roughly comparable concept of sharing intelligence functions with our man-made environment in any large measure, even when it will be to our advantage to do so.

On mind and hand, noise, and novelty

Artificial intelligence is a term that usually embraces purely intellectual functions. For applications aiming at human enhancement, however, intellectual power alone will not suffice. Multichanneled, multileveled, richly coupled input–output faculties and languages will also be required.

Our point of view is that in the whole animal, intellectual and physical functions are really inseparable—that is, thinking cannot be separated from the context of the living animal anymore than the animal can be separated from the context of his real environment. The goals of the living creature only make sense in terms of the specific world he inhabits. This distinction may seem trivial, but we believe that much research proceeds on the "antievolutionary" assumption that living functions can be *meaningfully* studied outside of context. Such context-free studies spring out of, and automatically reinforce, the view of man as a mechanistic system of being, and such studies encourage the design of machine systems that militate against man's growth and well-being. In the design of evolutionary systems for human enhancement, it must be recognized that man's changeability, and the evolving nature of his responsiveness to changes in his environment, his variability in learning, must be taken into account from the very beginning of the design.

Even Shannon's information theory, as great a step as it is, has not been extended to systems in which there is learning (i.e., in which the receiver's behavior is affected by what he has received, and in which the sender is affected by the way his message is received). Shannon did not preclude such an extension, and his collaborator, Weaver, specifically states that such extensions of the theory should be attempted.

We recognize that in our discussion we make a fatal error by misusing the term "noise." The Shannon use, though it presents an easily managed mathematical conception, does not attempt to model the real-world problem of noise. It does give an intuitively right approach by placing novelty, noise, ambiguity, and information in the same bag and forcing the discrimination between these. Because we merely highlight the need for such discriminations, and because we know no other recourse, we have used the Shannon relationships and terms out of context. In the *evolutionary* approach, we do not accept for our definition, as Shannon does, that there is a limited vocabulary of symbols that both sender and receiver know. Instead, we see the sender and receiver linked in a communication loop, seeking to maintain a level of communication appropriate to their purpose by changing code, channel capacity, use of the message, etc. *The whole message system is evolving at once in all its parts.* This is a far cry from Shannon's description, yet it is the world that now needs to be modeled.

In the development of evolutionary hardware and software, particularly as it is integrated in man–machine evolutionary systems, we must seek those situations that are sufficiently transparent and controlled to help provide us with the quantification necessary for resolving some of these dark problems.

Self-organizing control system

Traditionally, engineers have been conditioned to avoid the introduction of "noise" in the design of systems. Noise is usually considered by engineers to be any unwanted, indeterminate signal; but, more generally, it is a signal that is uncorrelated with other (information-bearing) signals within the system. If noise cannot be eliminated, it is at least minimized and looked upon with ill favor. Historically, too, controls, as in airplanes and spacecraft, have been treated as being linear, and if they have not been linear, control engineers have tended to look the other way.

Now, however, hardware has been developed for use in control systems that deliberately makes use of noise

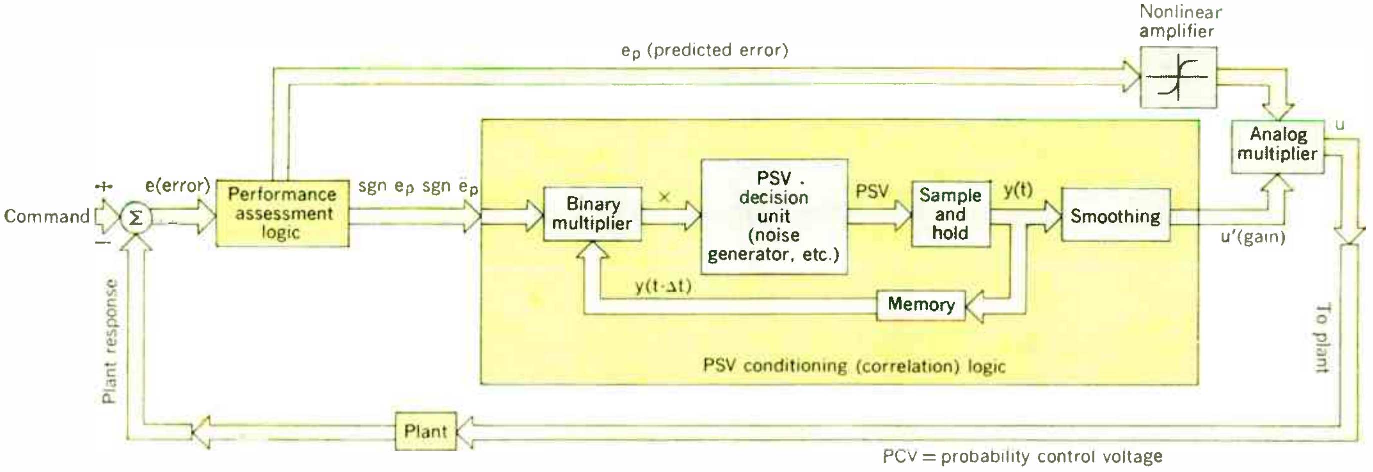


FIGURE 1. Basic elements of a self-organizing controller.

injection to resolve interactive nonlinear control situations. The hardware uses the noise to generate rapid random predictions of control response to unforeseen disturbances. Such control elements, called self-organizing controllers (or SOCs), are an advanced form of an adaptive control that gathers information about the aircraft plant while the craft is in operation and uses this information to improve performance of the system. Adaptation by this self-organizing controller is so fast that it can readapt many times within the closed-loop response period of the plant. When the plant characteristics change, owing to the plane's changing environment, this SOC system compensates for such changes and keeps the plane stable. That is, the learning is done by the plant itself, thus easing the pilot's control function.

There are now many types of adaptive and learning systems, developed experimentally over the past half-dozen years,[9] and going under many names. One might distinguish two general types: self-organizing systems, which are those with short-term memory, and learning systems, which are those with long-term memory.[10] Learning systems take a long time to adjust to new conditions, and long training periods are associated with their use. However, a self-organizing system adjusts to new conditions within a small fraction of the plant time constant.

One feature of particular importance in the SOC system developed by L. Gilstrap, R. L. Barron, *et al.*, at Adaptronics, Inc., is the use of random search techniques that allow rapid convergence in multiple-parameter spaces. For SOC applications, random search methods have proved superior in both speed and memory requirements to both systematic and gradient hill-climbing techniques. Interest in the random search methods may be traced to the original contributions of S. H. Brooks[11] in 1958–1959. The work of the Soviet scientist L. A. Rastrigin[12] and the Czechoslovakian J. Matyas[13] have stimulated wide attention to the potentialities of the random search method. In the United States, the application of random search techniques is exemplified by the PSV (probability state variable) control systems and in the development of *accelerated* random searches[14–15] for application to transformational automata (pattern recognizers, adaptive computers) and to systems that infer the properties of incompletely known dynamic processes. PSV random search makes it possible for the SOC to do without explicit identifications of all plant parameters.[16] In a sense, the SOC is engaged in a dialogue with the plant, rapidly postulating models of its performance and comparing these against its actual performance.

The SOC is the outcome of a neuron analog developed theoretically in the 1950s, which has led to a whole family of PSV devices that could be used for many problems that can be reduced to a problem of a search in nonlinear spaces.[12, 17]

Figure 1 shows the basic elements of a self-organizing controller. It is described as follows[17]: There is (1) a goal circuit (performance assessment logic), which is a means for evaluating current performance; (2) a conditioning logic for computing and effecting suitable changes of the controller parameters and/or output signals; and (3) a memory for storing information concerning past parameter states. The memory exhibits an "exponential forgetting," important in control applications because experiences in the remote past usually have less pertinence to present actions than do relatively recent experiences.[18]

It should be noted that the performance criterion in the conditioning subsystem may be defined in somewhat abstract terms. Many criteria of system performance are possible because there are many uses for adaptive control systems. The important thing is that the performance criterion should provide an unequivocal indication as to whether or not the experiments of the SOC are leading to betterment or worsening of system performance. Future SOCs may employ variable performance criteria, which themselves are subject to learning processes guided by one or more supreme goals.

The principal characteristic of the operation of the functional elements in the SOC is the closed-loop nature of the system; namely, a change in internal parameters must be fed back through the (unknown) plant and environment to the performance assessment logic to permit correlation between change and overall effect.

A comparison must be made between current performance relative to stored information regarding recent parameter experiments.

From the evolutionary point of view, one of the features of the SOC is its use of noise to create variety of action in the face of variety in the environment.[19] The noise generator, which sits in the PSV decision unit shown in Fig. 1, generates a random sequence of outputs that rapidly converge to the correct values for the actuator signals. The statistical source (random pulse generator with bistable statistics) fed by a noise generator is especially advantageous in problems of multiple plant identifications, where the correct signals for many actuators must be found rapidly simultaneously.

The use of SOCs could be significant for solving problems in which there are many interacting response variables. Multiple component systems, which theroretically could be built up to great size, can exploit coupling effects to achieve stable, simultaneous control of all variables. Barron and Gilstrap point out that "multiple-input, multiple-actuator SOC connectivity is an elegant expression of Ashby's Law of Requisite Variety[19] and may also have some interesting biological parallels as in, for example, the reticular formation. . .when one has a complex plant the crucial problem can be that of sorting out the cause–effect relationships between SOC experiments and plant responses. The PSV logic in the 'actuation–correlation logic module' (ACL) appears to be well-suited to the achievement of *simultaneous* correlations. These simultaneous correlations are enhanced by the mutual competition between multiple correlation processes. In the early days of SOC development, one or two prophets of doom forecast that multiple correlation processes would bog down if run simultaneously. In fact, while there is no doubt that the correlation processes become slower as their number multiplies, the competition between them is the *sine qua non* for obtaining correct identifications of multiple cause effect relationships."[20]

"One of the many fascinating aspects of self-organizing control," R. L. Barron says, "is the possibility that we may some day see systems that not only adapt to the plant or process and its inputs and disturbances, but which also adjust system behavior in response to the varying needs and desires of human users. . . .It is expected that future SOC systems will adapt their responses to individual human traits. Achievement of this capability will require development of performance assessment structures that provide the SOC with pertinent value-function information. This information should be obtained *without communication between the SOC and the human on a verbal level*. That is, the 'how goes it' question should be answered by evaluation of the human's automatic reactions."[21]

Artificial intelligence through evolutionary programming

We have seen how the use of noise has been turned to a very useful end in an unusual piece of hardware. Now, let us consider how noise has been made the essential ingredient in a computer program that attempts to replicate a fundamental aspect of the processes of natural evolution. Although such simulation of evolution by computer, which has been studied by L. J. Fogel, A. J. Owens, and M. J. Walsh, has thus far dealt with only relatively simple problems, the method appears to offer an extremely powerful tool.[22,24]

In his attack on the problems of artificial intelligence, Fogel stresses the role of prediction. In his argument, he views intelligent behavior as a composite of ability to predict one's environment coupled with a translation of each prediction into a suitable response in the light of the given goal.[23]

The "organism" in the evolutionary program is not a physical device, but is rather a mathematical entity that describes a particular logic for transforming a sequence of input symbols into a sequence of output symbols. Nonetheless, as the organism evolves, it becomes a rationale for prediction, reflects the pertinent logic underlying the data base, and thus provides a first approximation for solving similar problems in the future.

Fogel describes the computer simulation of evolutionary problem solving somewhat as follows: The computer is instructed to evaluate an arbitrary logic that describes the "stimulus-response" behavior of an initial "organism" in terms of its appropriateness with respect to a given goal in the context of the observed environment. That is, the computer measures the suitability for "survival" of the organism, namely, its demonstrated ability to solve the given prediction problem. This organism is then mutated by means of the introduction of randomness or noise so that there is produced an offspring different from its parent. The ability of this offspring to solve the given prediction problem—that is, its ability to survive in the given environment—is then measured. If the offspring does better than the parent, the parent is discarded, but if the parent does better, the offspring is discarded. The survivor becomes the parent of the next generation. This process is iterated until a cost criterion has been reached.[25]

This simulation of nonregressive evolution is carried out in fast time so that many "generations" come to pass in minutes. Since this is all done by computer simulation, the organisms produced are simply descriptions of desired stimulus-response behavior in the given environment. The randomness of mutation makes the process not wholly predictable. In fact, the evolved logic may well be a surprise to the programmer.

Although we cannot go into detail here (the book[23] and papers[22,24] are readily available), there are many interesting aspects that should be mentioned.

For instance, as in natural evolution of species, a great variety is produced and preserved, so the evolutionary program can also preserve some of the less-than-best organisms or species of logic. This allows, among other things, simulation across species, giving each species a different length of recall (memory) of the environment. Thus, a logic that was previously less than best might win a superior position. Likewise, through the operation of selection in natural evolution, certain species, out of the great variety nature produces, find themselves best adapted for differing environmental conditions. There are, as it were, ecological niches in which they fit, and the species that fall between these environmental "sets" tend to be eliminated. This Darwinian idea can be extended to species of machines.

Of great interest in Fogel's experiments are his efforts to develop successive models of the relationship among sensed variables in a way that reflects the nature of the goal of the investigator. The evolutionary modeling of the

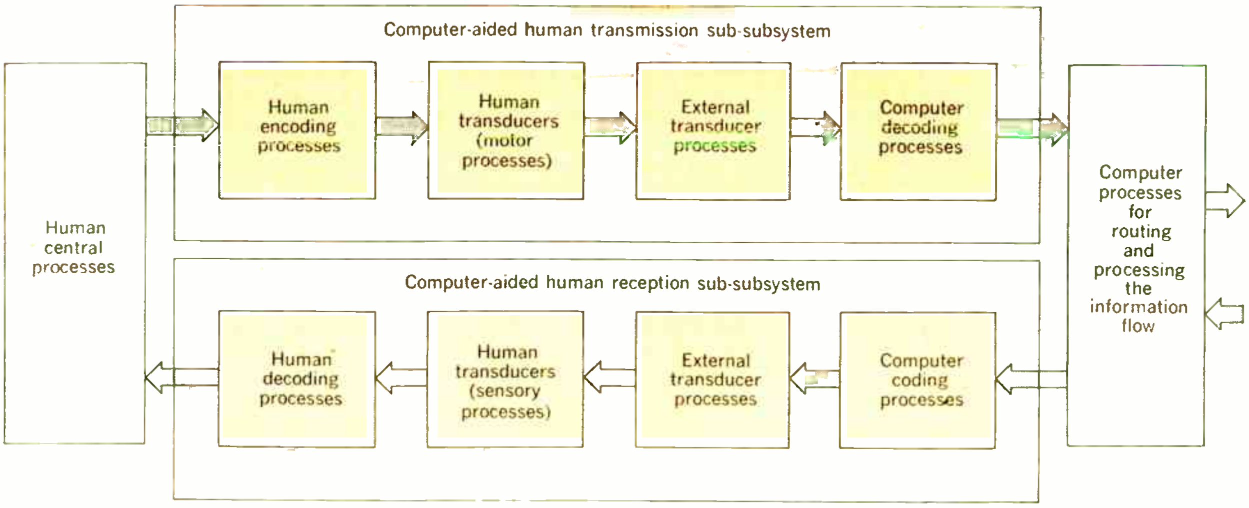


FIGURE 2. Conceptually, it is useful to breakdown man-and-computer communications in these subsystems, so as to make clear where direct computer aid can be given to each subsystem. Breakdown scheme is by D. C. Engelbart of Stanford Research Institute.

self may well provide a foundation for achieving higher levels of artificial intelligence. At a rudimentary level, this endeavor means the random mutation of models; at higher levels, it means the random combination of selected models into new logics. In point of fact, evolutionary programming will probably be needed to operate simultaneously at various levels of abstraction (each level interacting with each other level although on different time scales) if machines are to determine the "logical" structure or behavior of unknown and changeable interactive environments as complex as a man. (Insights into how man himself actively models and predicts his environment appear in the work of K. Craik and D. M. MacKay.[26])

In any event, here in generic form, through the preservation of those aspects of randomness that appear worthwhile to the organism, is the scientific method (induction being performed through a nondeterministic manipulation of the data base comprised of previous observations of consistencies and commonalities built into relationships or models). Fogel goes on to recognize that the scientific method is itself an evolutionary process that *can* be simulated. Thus, high-speed computers may find new logics for addressing old problems.

Augmenting human intellect

In any truly sophisticated man-machine system, we cannot imagine that the "two" traditional categories of human behavior (mind and hand) would be separated. Rather, they are viewed as reinforcing one another, so that what happens on the intellectual level affects what can happen on the psychomotor level, and vice versa. In the work of D. C. Engelbart, such interactions are considered to exemplify the "reverberation principle." He says: "In improving a system, an innovation at one level often leads to reverberating waves of (1) possibilities for other innovations, and (2) needs for other innovations. Waves of possibilities tend to propagate upward with an increasingly broad effect—new gains from an innovation (or possibility) at a lower level provide a new innovation possibility (or perhaps several) at a higher level. Gains from these possible innovations are added to the original gains, to stimulate possibilities at still higher levels, etc. . . . Each new innovation arising in either the upward wave of possibilities or the downward avalanche of needs is the potential source of a new wave in the opposite direction; it reverberates."

For nearly ten years, Engelbart has been pushing a broadly evolutionary program of research for "augmenting human intellect," aimed at using new tools such as computers, and incorporating a whole systems approach, to find ways in which men's capabilities for solving complex problems could be enhanced or, to use his word, augmented.

Although Engelbart's research does seem to stress the augmentation of intellectual activities, his writings make it clear that he does not regard learning and problem solving as going on at the mental level alone. For instance, he postulates a human-communication subsystem, as shown in Fig. 2, in terms of which he considers how computer-aided lower-level systems can be integrated into a higher-level system. The basic aid given this subsystem by computer processing he sees as the providing of versatile feedback through the two open-ended subsystem channels, thereby enabling more effective use of each channel through cooperation with the other channel. The augmentation research, then, looks at the chief design factors—the information characteristics of the messages, the signal forms and the information encoding at the interface, and the computer decoding process—which must be compatible with human ability to learn and to perform.

Engelbart's first philosophical statement of his program was laid out in 1962.[27] He set forward a conceptual framework that broke down the means of extending human capabilities into four basic classes. These included: (1) *artifacts*, the physical objects designed for the manipulation of things and materials, for the manipulation of symbols, for human comfort; (2) *language*, with which an individual models his world into concepts, and the symbols he attaches to those concepts for his thinking and reasoning about them; (3) *methodology*, the methods, procedures, strategies, plans, both small and large, through which an individual organizes his goal-seeking or problem-solving activity; (4) *training*, the conditioning and attainment of human skills for using effectively the artifacts, languages, and methodologies available. Any changes wrought in any one of these categories would reverberate and cause changes in the other categories. In reviewing the spectrum of research activities in the man–computer community, in terms of the "total system" comprised of these four classes, Engelbart rightly notes that the major share of interest thus far has been centered on artifacts.

In taking an engineering-like approach to all the elements of the man–computer system, rather than, for instance, focusing on the man or the computer alone, Engelbart makes it clear why a new conceptual framework is needed, and what the consequences of a new approach must be.[27–30]

For purposes of identification, he relates his ideas to the "Whorfian hypothesis, which states that 'the world view of a culture is limited by the structure of the language which that culture uses.' But there seems to be another factor to consider in the evolution of language and human reasoning ability. We offer the following hypothesis, which is related to the Whorfian hypothesis: Both the language used by a culture, and the capability for effective intellectual activity, are directly affected during their evolution by the means by which individuals control the external manipulation of symbols." Engelbart refers to this as the neo-Whorfian hypothesis.[27]

Individuals, he writes, who operate effectively in our culture have already been considerably augmented. For instance, an aborigine who possesses all of our basic sensory–mental–motor capabilities, but does not possess our background of indirect knowledge and procedure, cannot organize the proper direct actions necessary to drive a car through traffic, request a book from the library, call a committee meeting to discuss a tentative plan, call someone on the telephone, or compose a letter on the typewriter.

On the other hand, the aborigine can see, hear, smell, and interpret the meaning of events in the wilderness that would completely bypass our awareness.[31] He is not trained in our system of augmentation, and we are not trained in his. The restoration of our keenness in some of the aborigine's faculties would be meaningful to consider if we could create an environment that was *worth* seeing, hearing, and smelling. Who wants to see only square boxes of buildings and rooms, who wants to see fluorescent lights, who wants to hear cars, trucks, and planes continually, who wants to smell poisonous air? To sense such aspects of our environment is only to awaken one's helpless rage, and under that (at a deeper level, since we are talking of levels) one's sorrow at the grievous things we men have inflicted on ourselves. It is difficult to value, in such terms, the price we are paying by hanging too long onto a primitive, machine-age technology whose chief virtue is mass-produced consistency, that notorious "hobgoblin of little minds." The machine age is over; let us hasten on to the creation of an environment whose novelty, variety, and intelligence could nourish our growth instead of stunting it.

Engelbart's major project has been with computer-aided text manipulation. This program has many facets that are well reported in the literature.[27–30] The text-editing systems allow users to see their text on a scope, to compose on it, to modify text, to study it, to store it, and, to retrieve it in any order they wish; thus, there is already a great power and ease, and flexibility in keeping "plastic" working records. The system has both off-line and on-line text-manipulating capabilities; recent reports have been produced entirely through the use of these computer aids, and a 12-terminal time-shared system has recently been set up.

In addition, Engelbart has developed a unique chord handset that allows improved human display to the computer. He uses five-finger chords for transmitting English text to the computer, which, he says, allows the human user to achieve a more intimate sense of communication with the computer.

Automated psychomotor skill training

Another facet of Engelbart's broad program for the augmentation of human intellect is a study for automating psychomotor skill training, an area that is potentially very important for human enhancement applications since teaching and learning go on at a nonverbal or nonsymbolic level.[32] In such applications, automatic display devices would present direct stimuli to the trainee and automatic sensing devices would monitor his behavioral and physiological responses, both in real time. By using a computer to analyze the performance of the trainee, and through the accumulated insight into how to guide him toward a desired response (this would be the evolutionary dialogue situation) through nonverbal feedback as well as through symbolic information, it should be possible to bring enhancement to many kinds of training situations, to add enriched experience to the training of the young, and so on. The range of possibilities could cover the simplest tasks, such as operating keyboards, on out to piloting high-speed vehicles, and so on. Engelbart's proposal was to experiment with simple tactile cueing stimuli that were to intervene between the primary stimulus and the completion of the desired response. These cueing stimuli were to guide the subject through coordinated sequences of elemental actions. It was intended that through automatic monitoring, the cueing stimuli would be modified (either automatically or assisted by a human coach) to resonate with the changes in the subject's performance. Engelbart's notion was that by relieving the load on the subject's "higher" faculties, by cueing signals applied cutaneously at points on his body, the subject would not need to memorize the *details* of the patterns through which he was learning to weave his way. The specific project was to train people to use the five-key chording device for transmitting English text mentioned earlier. Both visual (lights) and tactile (air jets) stimuli were used. The instruction to the trainees was, in effect, "Push the key with the finger on which you feel the sensation of the air jet." The conclusion of the study was that automated sensorimotor skill training was certainly feasible, and that the potential of computer-based training systems of this sort appeared great.[33]

These were certainly not the first experiments in this area. One of the pioneers in dynamic and adaptive automatic teaching machines was the English cybernetician, Gordon Pask,[34–36] whose investigations have ranged widely from highly theoretical to highly practical projects, that is, to the development of many varieties of machines. One of the central ideas involved in his machines that teach is that their control mechanisms be adaptive (having a degree of artificial intelligence) so that they change their characteristics as a function of the subject's performance, thus maintaining a degree of novelty so that the subject does not become bored.

Computer-assisted instruction, on many levels of sophistication from the evolutionary dialogue point of view, is a growing field of interest, as evidenced in a recent special issue of IEEE TRANSACTIONS ON HUMAN FACTORS IN ELECTRONICS.[37] Some of Pask's recent work is reported there, as is the work of W. Feurzeig of BBN, who has been working on machines that teach medical students approaches to diagnostics; there is also a discussion of the well-known PLATO system at the University of Illinois by D. L. Bitzer *et al.* J. Weizenbaum's ELIZA system, being developed for teaching by Edwin Taylor at M.I.T.'s Educational Research Center, is also of much interest because of its efforts to mimic realistically the psychiatric situation between patient and psychiatrist.[38] Insights from such behavioral studies should certainly be relevant to evolutionary man-machine dialogue. Also, recently, Oliver Selfridge of the M.I.T. Lincoln Laboratory has been evolving methods whereby computers can teach their own language and usage to a user who otherwise has no "outside" guidance. Such "responsiveness" by the machine to its users' problems is certainly significant.

Enriched coupling

In 1961, W. A. Rosenblith wrote: "We may in ten or fifteen years have derived some useful generalizations from the widespread experience with man–machine (including man–computer) systems and from exposure to novel sensory environments. We shall certainly be able to telemeter converging data (electrophysiological, neurochemical, behavioral) from organisms while they interact with their environment. It should also be possible to sharpen the formulation of issues that are critical to our understanding of sensory function by incorporating the experimental subject into a closed loop—an arrangement wherein stimulus sequences are made contingent upon the physiological and psychological responses that the subject emits."[39]

Investigations of sensory communication in such closed-loop situations have not yet been conducted on the scale Rosenblith imagined. There is, however, an enormous literature relating to the monitoring of physiological and behavioral variables. When such information grows from the closed-loop situation,[39–45] it could be utilized in richly coupled, multimodal man–machine dialogue, employing a redundancy of channels in which each channel acts as a metaphor of the other.

To indicate something of the unexpected and even bizarre findings that might turn up in monitoring human dynamic functioning, one might consider one kind of relation between stimuli and cerebral activity discovered by W. Grey Walter of the Burden Neurological Institute, England.[40] He noted that there were certain brain waves that were set up in anticipation of some external happening, waves that were extinguished when the anticipation has been satisfied. The clear determination of such measurable brain events might be used in automatic training situations, where a suitable monitor would tell the computer control that a subject had already received a message he was expecting.

In a similar vein, E. M. Dewan of the Air Force Cambridge Research Laboratory, and Belmont Farley (who has made studies in the computer simulation of neuron nets[41,42]) have experimented with the modification of brain-wave activity. Both men determined that they could consciously alter their brain waves in such a way that through the use of electrodes on the scalp they could control external devices. Dewan thus activates an external switch to turn a light off or on and he postulates that a kind of communication by such electroencephalography might have a number of useful applications.[43]

A study carried out by Hecker, Stevens, *et al.*, at BBN, on the effects of task-induced stress on speech indicates that the range of bearable stress can be measured and serve as an indication of what different individuals will tolerate in task loading.[44] The monitoring of such stress could be another channel in a multichannel man–machine communication situation.

A proposal for a massive and coherent monitoring of physiological variables (through implanted probes, etc.) has been put forward by Dr. Charles Ray of Johns Hopkins College of Medicine.[45] By monitoring *transitions* (chemical changes) rather than absolute levels of the living system and through using appropriate sophisticated displays, Ray visualizes, for instance, that it would be possible to track the changes, and thus interrupt or control changes that would otherwise become dangerous to a patient. This kind of sensitive control of a patient's responses to drugs, etc., are not now possible.

Computer model of reticular system

There are many approaches being taken to artificial intelligence that could then evolve to higher-order, stable, intelligent systems. As we have seen, in the approaches of Gilstrap, Barron, and Fogel, there are, in a sense, "searching" ways of evolving small models, which then may grow into larger models. There are also profoundly theoretical approaches of how models may gain in problem-solving power, as in the work of Amarel, Newell, Simon, Minsky, Papert, Pask, and others, who make no special effort to relate their models to how intelligent living systems function. There are combinations of such approaches. And then there are specifically approaches that attempt to mimic or caricature, to a fairly realistic extent, the functions of the vertebrate nervous system.

In this last camp belongs some work of a deep order, like that of W. Kilmer, principally inspired through the lifelong work of W. S. McCulloch in his studies on the functioning of the brain.[46] In recent years, McCulloch has devoted his efforts to unraveling the mysteries of the operation of the core of the reticular formation, which runs through the spinal cord and the brain stem. From it all other parts of the brain have evolved; it is the "begetter" to which they all report and whose commands they obey. Figure 3 shows a diagram of the reticular formation in a cat. The "retic" has evolved so little that, if you do not know the magnification, you cannot tell whether a cross section from the retic came from a man, a mouse, an elephant, or a frog. Its crucial business is to commit the whole animal to one of a dozen or more incompatible modes of total behavior, such as eating, mating, sleeping, fleeing, fighting, hunting, or hiding.

Its million or so of relatively undifferentiated neurons sample at random all sensory channels and all tracts ascending and descending in the nervous system and talk to each other as well as to all other parts of the brain and sense organs. A single one of its neurons may respond to clicks, to a touch of the nose, and to a shaking of the left

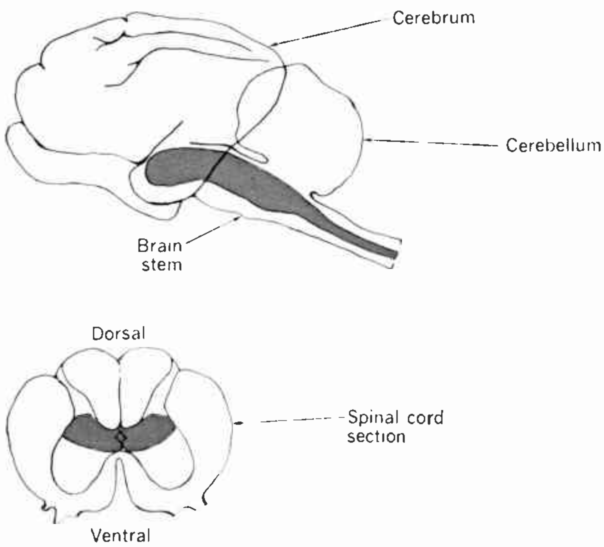


FIGURE 3. Drawing of cat brain and section of spinal cord shows the relatively small volume of the reticular formation (in color) from which the "higher" levels of the brain evolved. In vertebrates, it runs the length of the spinal cord up to the head of the brain stem. Phylogenetically the oldest structure, the "retic" is the crucial control center that governs the gross behavior modes (running, fighting, sleeping, eating, mating, etc.) of the whole animal.

hind leg; but, if the stimulus is repeated, it responds less and less and may quit altogether unless that stimulation—say a click—is associated with a signal to which it must respond—say a strong electric shock to the left hind leg. Then the response to the click returns. When the animal sleeps that same neuron may respond only to respiration or a bubble in the gut.

Thanks to V. Amasian, A. and M. Scheibel, and many others, the activity of the retic components is well known. The Scheibels, D. Fortuyn, Valverde, and W. Nauta have worked out the anatomy of its internal connections. Much is known of its paths to receptors and to other parts of the brain, and much of its actions upon them. The problem confronting the neurophysiologist is its intrinsic functional organization, which must account for the ability of the reticular core to reach a working consensus of its many components in a fraction of a second so as to commit the whole animal to the proper mode of behavior.

Because neurons are nonlinear oscillators the natural way to analyze this circuit action is in terms of the behavior of a vast number of coupled nonlinear oscillators. Unfortunately, the mathematics of such systems remained where N. Wiener and B. Vander Pol had left it, until the recent work of E. Caianiello, which may be of future help. It was simply inadequate to the task; neither McCulloch nor Kilmer see how to use Caianiello's theory today. Next, since the reticular core is a net iterated throughout the length of the nervous system, the theory of iterated nets was investigated, but F. Hennie and W. Kilmer were able to show that the questions they thought worth asking could be proved to be recursively insoluble problems.

The model they finally attempted was based on the flow of information of the modern naval fleet in which every important ship has its one center to which come signals from many other ships, and from all sensors, to detect friends and foes in the air, on the sea, and under the sea. The admiral has disciplined his fleet in a number of maneuvers and remains in titular command, but what the fleet does in games or in battles is actually determined by that ship having the requisite variety of information at the moment; the real command moves from one ship to another as the engagement goes on. In McCulloch's words, "It enjoys a redundancy of potential command in which information constitutes authority." The admiral is but its mouthpiece.

It is such a system that W. Kilmer, with the assistance of J. Blum, has been computer-simulating at the Instrumentation Laboratory of M.I.T. for L. Sutro's group, which has been working on a sophisticated artificial visual system for unmanned exploration of Mars.[47] The simulation is designed so that it can be realized in hardware and miniaturized. Kilmer's model consists of a dozen hybrid probabilistic computing modules coupled much as the specialists in a diagnostic team who, having each examined the patient, must come to a good working accord as to proper mode of treatment. The proper mode might be, say, one out of 16. In the present model, it is only one out of four. Figure 4 shows the system organization.

Each of the computing modules receives a random sample of the inputs and makes the best guess it can at the relative appropriateness of each mode of behavior; and each is informed of the random samples of the other modules. The coupling of the modules is fairly weak so that the computers exchange their preliminary guesses for several rounds, usually less than 15, before reaching their consensus. The dissimilarity of random inputs ensures the requisite variety of guesses. The nonlinear skewing of the estimates of probability and the high weighting of those that peak on a preferred mode prevent a "hung jury." The model can be made to decide fast enough to command the whole system in response to a real environment. The redundancy of potential command insures that if a fraction of the modules lock in any position, oscillate wildly, or are shot out, the remainder can override them and continue to command reliably. The coupling of the modules leading to a given modal decision gives the system an inertial property, stabilizing the mode in the face of small or insignificant changes in inputs so that it is less distractable than, say, a monkey, which darts about responding to each slight stimulus. The system would be as stubborn as a pig, however, but for a second trick. Confronted with a drastic environmental change requiring a new mode, its modules decouple to an extent determined by the significance of the change and for a duration directly proportional to the system's degree of entrenchment in a past mode.[48,49]

The monkey-pig imagery highlights the fact that there is a great variability in the inertias of entrenched modes among animals; and one can easily see such a variability among one's friends. That is, in the human species, there are a variety of behavioral styles of organizing response, and certain deeply entrenched modes have long been the source of interesting "characters" in the tradition of the novel.

In the Kilmer system, this kind of inertial response of the overall system is balanced through the cooperative decision behavior of the computers.

The limitations of the present model are due to oversimplification purposely made to keep its programming transparent and flexible. The number of modules can be increased, yielding ever-better action. The number of modes can be increased to a dozen or more as the problems require. The time between a conditioning stimulus and the unconditional stimulus can be extended by shift registers that are easily miniaturized. In hardware, its complexity is today limited by the technical difficulties of making the vast number of required connections.

Ordinary algorithms for designing systems to handle information require some regularity and even distribution of input, whereas this model of the reticular core has to work, like the real reticular core, when the crucial information is presented partially to few and scattered modules. This, and the necessity of adjusting its nonlinearities, made its design extremely difficult, and only the necessity of inventing a computer to command computers and controllers (it could be used as a command center for a robotic device or a community of robots) pushed the work forward.

In fact, Kilmer tells us, the system is so complex to build and to test that its designers have had to play with it, to engage in a dialogue with it, to feel out the best way of using it, to feel out the best way of even describing it.[50] Only through dialogue of the kind we have described[2] has the system been evolved to its present status. In this, Kilmer and his associates have been following a principle of McCulloch's, namely, that to think about such systems in new terms, it was necessary to evolve a system that specifically one could think *with* rather than just think *about*. In this sense, the system incorporates noise, novelty, a sense of the importance of relative factors, the capacity to shift attention, and so on, none of which would have been possible without dialogue (Kilmer's statement, not ours). The system thus fits clearly into the camp of evolutionary machines.

Other appropriate models

There are many other important efforts at modeling intelligence that ought to be described at equal length. We should at least consider briefly the works of men like Papert, Minsky, Amarel, Pask, Arbib,[51] and others.

Without hesitation, we should cite the models being developed by S. Papert in his cooperative endeavor with M. Minsky to evolve a robotic system with powers of artificial intelligence. Papert[52] has described three levels of computer-based models that he has labelled tau, theta, and rho models. The first level of models, the tau's, are really toys, small idea embodiments, with which to develop a feeling for how they might operate. From these, there arise a second series of models, the theta's, that are theoretical constructs, or higher-level models based on the experience with the toys. On the last level come the rho's, the real-world models, machine constructs that could operate in real time in the real world, carry out intelligent tasks, form the basis for robotic intelligence (e.g., carrying out construction tasks on other planets).

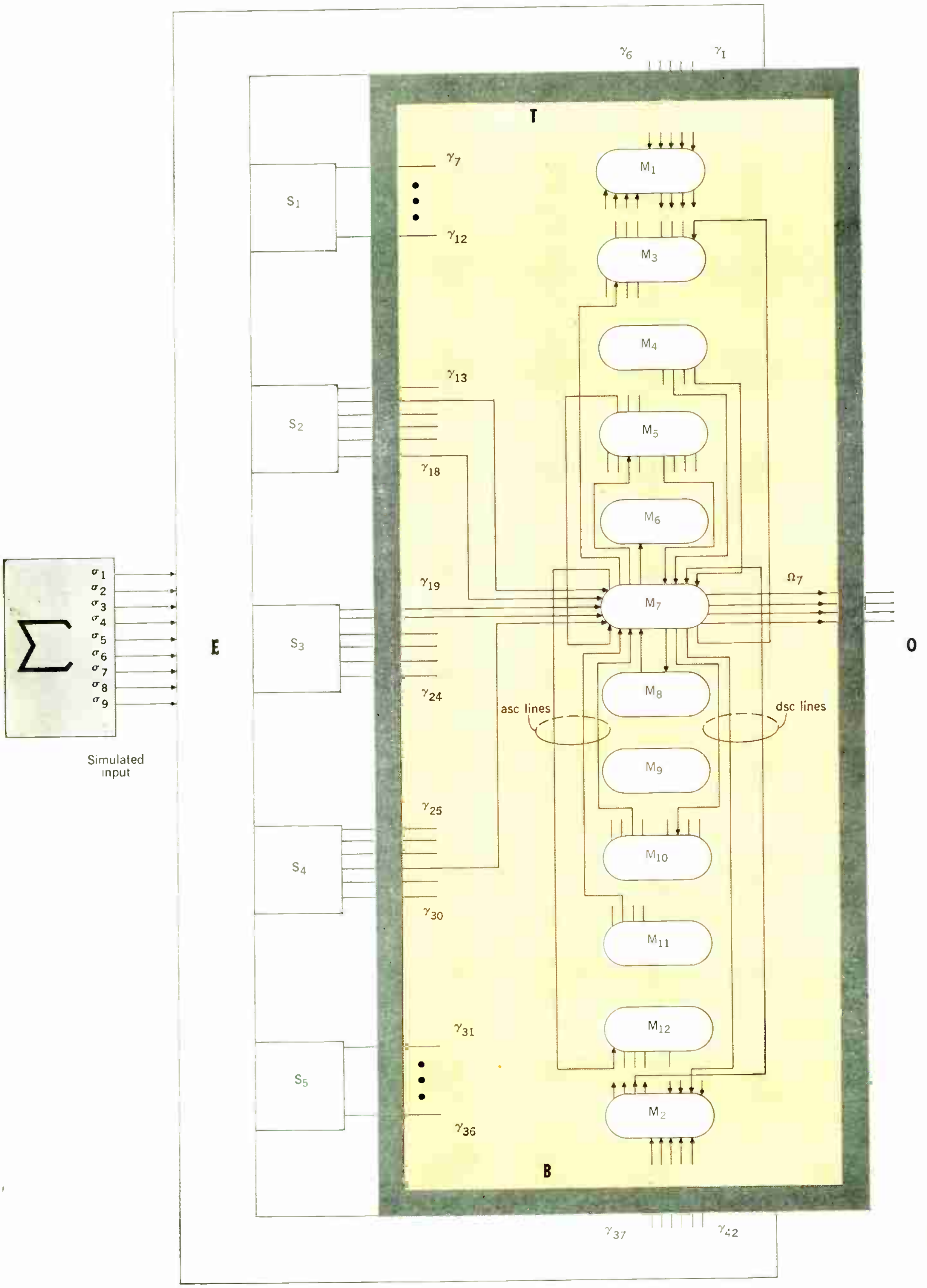
Another man who has produced many important ideas on problems of representation and modeling, especially in realms where there is very little a priori knowledge about the structure of the problem territory, is S. Amarel.[53-57] He shows in his work that a change in problem representation has a profound effect on the power of a problem-solving system. The problems of mechanizing such evolutionary transitions, Amarel sees as being related to the problem of mechanizing certain creative processes.[57]

G. Pask, whose research on adaptive teaching machines was mentioned earlier, in his most recent paper on communications between men and machines,[5] makes certain most relevant points about the organization of the levels and languages of evolutionary machines. He says, for instance, that a stable tutorial discourse must include higher-level components, in the sense that the student can propose modifications of the educational goals and of the mode of instruction just as he would in a real-life tutorial. Pask says: The machine must be able to accept or reject these proposals according to whether or not they foster the learning process and it must, in some sense, discuss its acceptance or rejection with the student. The mechanical language used to mediate the man–machine interaction must be rich enough to accommodate *this* sort of repartee.

The trouble is, says Pask, the present design of machines (including procedural machines and institutions) precludes an adequate man–machine rapport, because we design mechanical languages (and consequently design machines) in terms that are rigorous (as in logic textbooks) rather than natural to our way of thinking. And, Pask contends, the solution to the dilemma appears to rest upon a radical reappraisal of the character of human and mechanical systems. Unlike the control systems we traditionally have considered, man, he goes on, shows a propensity for seeking novel goals at all levels in the hierarchy (for posing problems as well as solving them). And he argues, very much as we have, that new machines must possess internal evolutionary processes; the goals and the concepts they generate cannot be completely described in the working language of the system at the moment of their inception, but rather become describable as selection occurs. If two systems (men or machines) are to innovate jointly, the language used for the discourse must be able to express ambiguous or "incompletely described" concepts, as they might in natural language, which might be augmented by gesture and pictorial displays. He says that if there is to be a coupling between the evolutionary processes in the two systems (men or machines), then the two systems must be able to interpret ambiguous expressions. And the trouble, of course, he says, is that existing mechanical languages differ from natural languages in that they do not accommodate ambiguous expressions.

In his work, Pask proposes an organizational or cybernetic model to replace our present existing models. Perhaps the highlights to remember here are that different languages must be evolved for each level of the hierarchy in the evolutionary machine, that ambiguity must exist in each language at each level, that discourse or dialogue should go on at many levels (in parallel, or simultaneously), and that either machine or man will compensate for a heightened ambiguity (and hence novelty) by checking information inputs on one channel against information inputs on another channel.

There has been implicit in our examples and arguments up to this point a crucial shift in viewpoint on how it is that a living creature such as man "reads" his environment and maintains his stability as a whole organism within it. This shift in viewpoint, which has to do basically with man's perceptual system,[58] is explicated in a paper of D. M. MacKay.[26] The traditional view of perception took it to be a kind of passive two-way mechanical process in which the organism is delivered a stimulus and *then* executes a response, as though there were two sequential



γ6
γ1
T
M1
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M2
S1
S2
S3
S4
S5
γ7
γ12
γ13
γ18
γ19
γ24
γ25
γ30
γ31
γ36
γ37
γ42
Ω7
O
E
B
asc lines
dsc lines
σ1
σ2
σ3
σ4
σ5
σ6
σ7
σ8
σ9
Σ
Simulated input

FIGURE 4. Simulation model of the reticular formation (RF) called the S-RETIC, a caricature of the living brain stem control center, uses 12 nonlinear, probabilistic, hybrid computers. The M_i are logic modules; the S_i correspond to various humoral, chemoreceptive, exteroceptive, and interoceptive sensory and internuncial systems that feed inputs directly into the RF; all σ_i and γ_i lines are binary; all Ω_i (only Ω_i shown) are modular mode-indicating output (0) lines; Σ and E simulate an RF environment that engenders input signals (sights, sounds, etc.); T and B are the top and bottom boundaries of the RF; the asc and dsc are the ascending and descending "nerve" bundles. For clarity, the connections that recur on all M modules are shown only on M_7.

processes involved. MacKay's research suggests that perception is the activity of *keeping up to date* the internal organizing system, that represents the external world, by an internal matching response to the *current* impact of the world upon the receptive system. In a sense, the internal organizing system is continually making fast-time predictions on what is going on *out there* and what is likely to happen *out there*, and it takes anticipatory action to counter the small errors that might threaten its overall stability mode.

One simple way of looking at this model of perception is to perform a simple experiment that shows how you bring your own meaning to the environment through action on your own part to heighten your sense of interaction. Take a small object, close your eyes, and strike it or drag it across your open palm. Then grasp the object with your whole hand, and manipulate it so that you feel its shape. At some point, you will instinctively *grasp mentally* the nature or character of the object. That is, your positive action of searching finally illuminates a whole image in your mind, whereas the passive reception does not. This analogy is very rough and ready, of course, and does not convey the heightened sense of the anticipatory updating action, but it may at least suggest how it is that perception is an active process. The fact that you can see the object as well as feel it, etc., heightens and reinforces your sense of its significance to your own purposes.

Studies of speech perception exhibit similar findings, showing that we tend to anticipate what another person is going to say, so that when the other person says something surprising, something unexpected, our whole receptive system tends to become more alert. The unexpected words, sentences, tones, etc., may excite, threaten, and so on, depending on the context, the relation to our internal mapping of the world, and so on. Speech simulators that utter sentences without intonation may leave the listener with the feeling that he has heard the words without being able to determine their meaning.

A. R. Johnson has also written about this view of sensory-motor behavior acting as an active perceptual buffer between the real world and the person's stored cognitive model of that world.[59,60] In short, he says, our understanding is now this: our ability to be passive observers of incoming data is a learned and highly developed ability, and such learning may only ensue through active interaction with the external world. Like us, machines may learn best by active participation.

And, certainly, your "response" and interaction with a machine that actively "looks out" at you through many channels, and that actively anticipates your moves (as another person does), is going to be quantitatively and qualitatively different.

H. T. Hermann and J. C. Kotelly, in studying psychiatric situations, have been the first authors who have really tried to deal with "context" formally.[61]

Finally, to go afield, we should realize that Soviet cyberneticians are well advanced in their models of thinking and of the mind.[62,63]

Temporal coupling

In the design of evolutionary systems, it will be crucial to recognize that different time cycles will be manifested on different levels of such systems. In general, this will mean that different forms of dialogue will occur. Dialogue at a microbehavioral level occurs, for instance, when both man and machine have short time constants (the machine might monitor a man's eye pupil dilation, finger movement, etc.); very large systems, such as whole societies of men and machines, are characterized by very slow change, measured in years, as for instance in the dialogue of politics. In general, a man or machine, one acting slightly slower or faster than the other, might be used to evolve a kind of "time driving" or "entrainment," that would intentionally drive systems out of their habitual timing. For instance, a fast machine can push a slower man, as in speed-reading training; on the social level, things such as telephones and television have speeded up political responsiveness and created a tighter dialogue within the body politic.

Machine–man investigation of the use of such phenomena to enhance or disrupt information exchange is yet to be carried out in any coherent way. We are hard pressed to omit consideration of such time questions here, but we must stake out a few signposts in this dark and difficult territory so that those who take evolutionary systems seriously will know where to begin looking for relevant research.[64–70]

A recent report by A. Iberall and W. McCulloch shows something of how a physicist looks at temporal events in biosystems, from the cell or atomistic level on up to the whole organism.[64] Figure 5, for instance, from that report, shows a kind of rough schematic of the time for complete event cycles at various levels of man's physiological behavior.

The concept of different kinds of time, usually associated with different levels of abstraction, called "time graining,"[65] apparently has a near counterpart in cybernetic thinking in the Soviet Union. N. M. Amosov describes levels of time storage shown schematically in Fig. 6, in his book, *Modeling of Thinking and the Mind*.[62]

The use of transmission time, in links in self-organizing systems, for information storage, is opened up clearly by D. M. MacKay.[66] In effect, MacKay considers what aspects of systems are being neglected, or not doing their share. For information storage and handling, he concludes that time itself is a medium for such storage. For instance, the length of delay in a person's response tells his interlocuter (man or machine) information he might otherwise miss. It is information that can be sensed on a nonverbal and nonvisual level. Moreover, time information, MacKay notes, may be transmitted on many levels simultaneously, and the event cycles on these different levels would, of course, be of different duration; which means, too, that interactions between different levels would be of a complicated kind.[66]

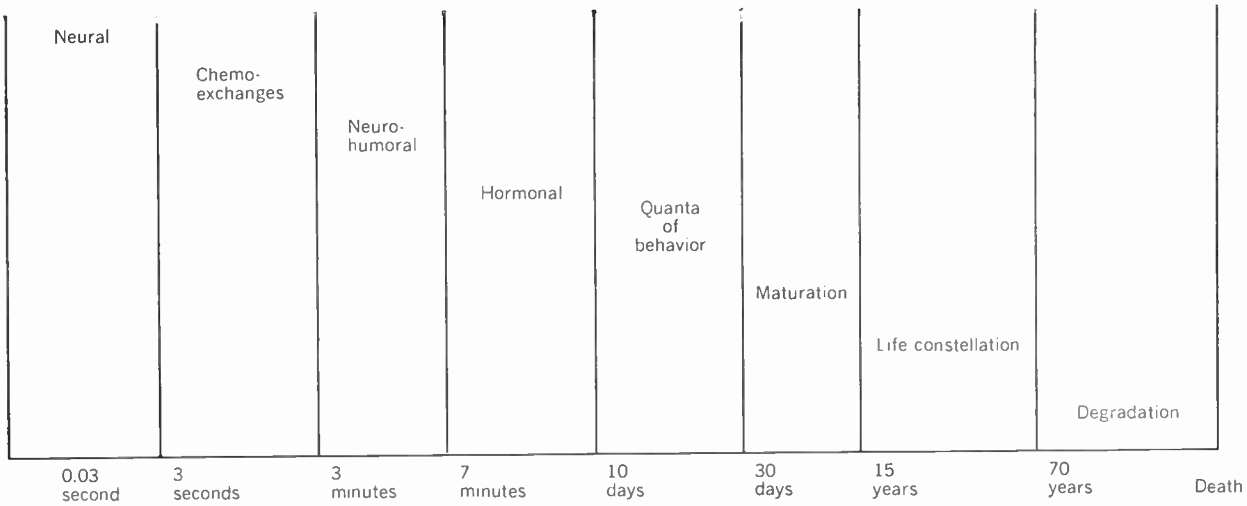


FIGURE 5. The dynamic spectrum for man, involving behavioral response, for the various levels of his organization as a system.

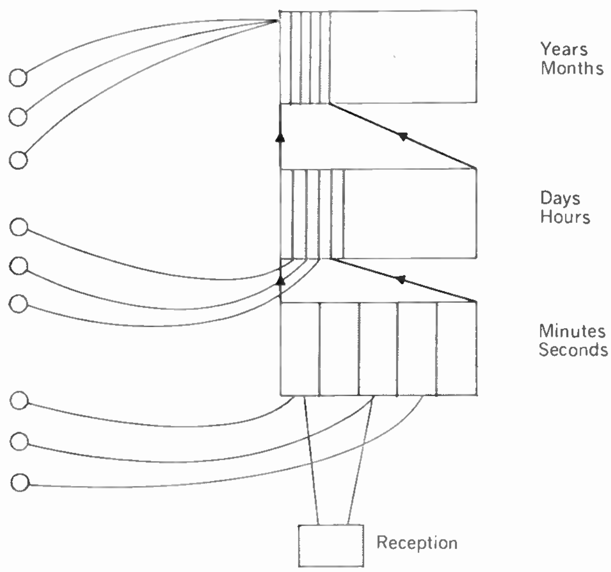


FIGURE 6. Schematic of the "network of time" showing the levels of time associations proposed by the Soviet cybernetician N. M. Amosov.

The evolution of new languages

We noted earlier that change in any part of an evolutionary system brings in its train reverberating changes in the remainder of the system. Thus, we should recognize that much recent research on languages and on methods of representation, and research on new media for communications, will exercise an expansive power in the nature of our dialogue with machines. As D. C. Engelbart points out, because present languages, present media, and present methods of manipulating languages have limitations, it is difficult if not impossible to attack many problems because we have no adequate means of quantifying those problems. And quantify here does not mean merely the traditional sense of measuring and assigning numbers; it also means finding units for measure. New languages could allow us to find new coherent and well-defined shapes or Gestalts in our relation with our environment. For instance, as every scientist and engineer knows, the discovery and invention of new forms of mathematics opened new worlds. So too, computer languages, as they steadily grow more powerful, are opening new worlds.

In the scale of languages, we are weighing not only verbal languages in the classical sense, but languages of gesture, graphical languages of all kinds, and so forth. We shall touch on just a few points of recent relevant research to suggest the potential of new developments.

Graphical representations of two, three, and more dimensions generated through the assistance of computers are currently of keen research interest. One of the important first examples of computer graphics was I. Sutherland's Sketchpad system, and it has been followed by many others since. Sutherland, along with Coons and others, is investigating true three-dimensional display languages. Two-dimensional projections of three-dimensional figures that can be rotated at will with a hand-operated hemispheric control have already been used for some time on the "Kluge" system. R. Stotz and others at Project MAC are now working on low-cost graphic displays for computer time-sharing consoles.[71]

A delightful example of graphical language for writing computer-generated music is that of M. V. Mathews.[72] He has a procedure for drawing scores as graphical functions of time using a light pen on a CRT attached to a small computer. The graphical input is transmitted digitally to a larger computer, which synthesizes the sound and reproduces it immediately with a loudspeaker. Figure 7 shows the arrangement of the man–machine loop. Programs have been developed for producing a variety of sounds using programmed instruments that simulate real-world instruments. Composers such as Varèse and Stockhausen have already experimented with graphic scores. An interesting element of the system is that it can be programmed so that the computer generates parts of the music through random and mathematical algorithms, "suggesting" new possibilities to the composer. Thus, the

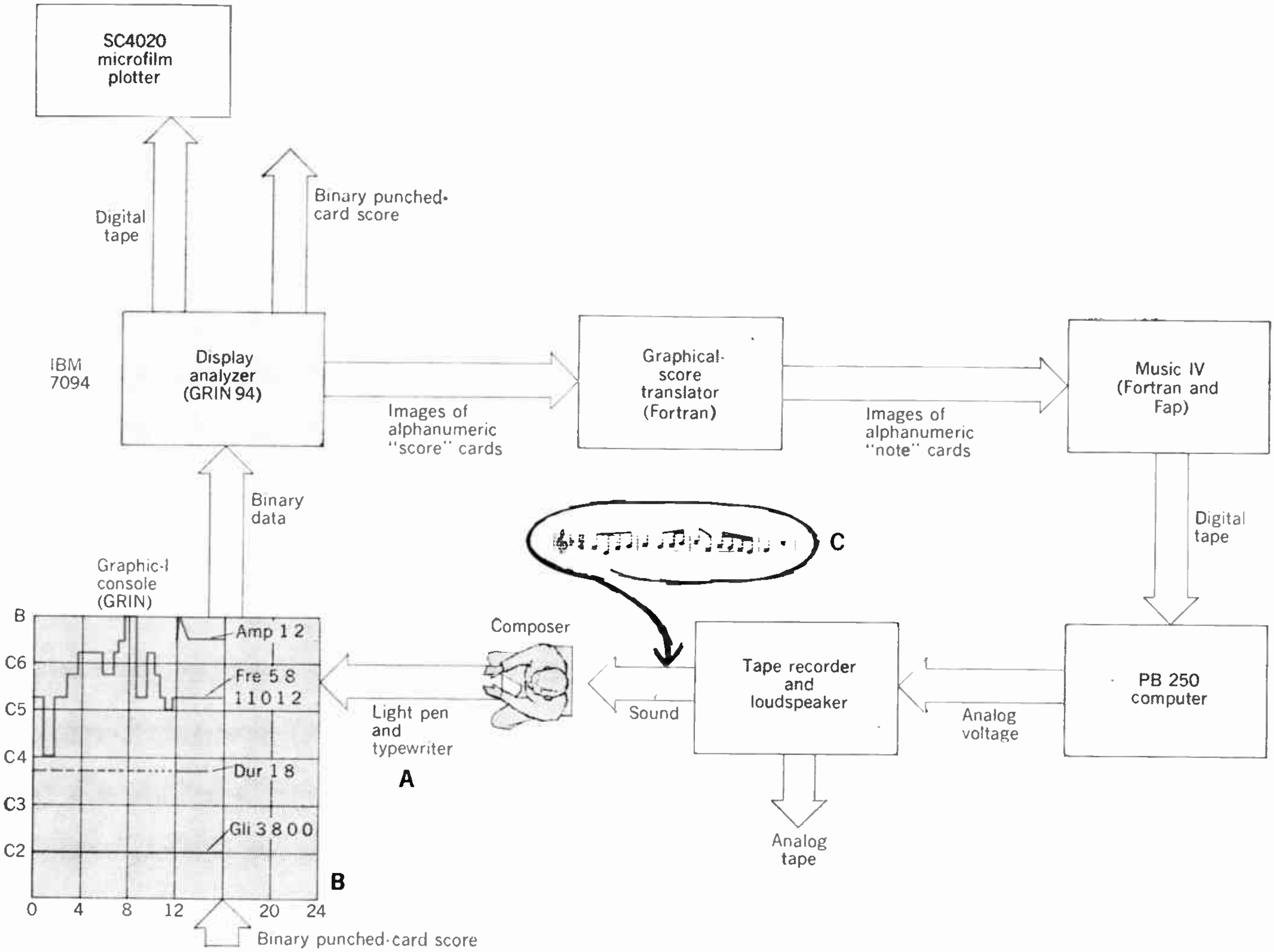


FIGURE 7. Music can be composed graphically with this system (A) developed by M. V. Mathews. One form of the score for the first measures of the march "The British Grenadiers," shown in (B), requires four functions specifying amplitude, frequency, note durations, and any glissando (continuously changing frequency). For comparison, the conventional music score is shown in (C). The system constitutes a powerful new on-line language.

composer plays with musical ideas that might not occur to him on his own. The joyful part of a system like this, aside from its obvious powers, is that it could be used by children, who could learn to compose music directly through making drawings before, for instance, their hands were big enough to play a piano. We might then see an adult interest in children's music comparable to the interest in children's paintings during the past two decades.

The really significant aspect of the Mathews system is that it puts the man directly in the feedback loop with the responsive machine, and although the system is not quite operating in real time yet, it is getting there in orders of magnitude jumps.

Another computer-based graphical system for studying the characteristics of human speech is the system set up by P. Denes. With a light pen and a keyboard, he can vary a dozen or so parameters of speech, (e.g., pitch, amplitude, formant frequencies, etc.), so as to study the interactive roles of such parameters, and hear an immediate synthetic voice output. Thus, he has the power of both visual and audio feedback.[73] The results can be delicious to hear. In interacting with his kind of system, a person can gain new insights into his own speech production. The machine would have a positive effect in the education of children.

Computer-based graphical experiments of an unusual kind have also been carried out by A. M. Noll.[74] He produces three-dimensional images by having the computer calculate separate pictures for the left and right eyes that, viewed stereooptically, fuse and generate the illusion of true 3-D depth. He points out that with currently available systems that use two CRTs mounted on the side of a helmet and with half-silvered mirrors, images can be presented to each eye. In this way, a person can see both the normal surroundings and also the CRT images superimposed on those surroundings.[75] If a computer generates images for this helmet device, and the movement of the head is picked up and fed into the computer, a person can move about and the computer-generated 3-D image will change accordingly, giving the person the sensation that he is *in* that environment. Imagine the excitement of "taking a walk" with such a system in a world governed by the laws of molecular systems, or of actually making those "quantum jumps" one had to study *about* in text-book equations. Would this not be a new world of realization for the scientific student? A computer-driven, 3-D movie that time-drives us and allows us to feel our way into the dynamics of the "underworld" would thus make relationships plainly and simply visible to us that were barely evident before, and that could barely be guessed by the most carefully educated intuition. The value of this kind of system in architectural designing, in which an architect could walk through and get the feel of his building before a physical element ever went up on the site, is incalculable.

Noll also envisions a system in which a mechanical manipulator similar to those used in nuclear installations would be connected to the computer. The user, he says, would put his hands into this manipulator, and could then freely move his hands about with the computer sensing the exact position in space of his hands and fingers. The servomechanism in the manipulator would also be connected to the computer so that as the hands approached a computer-stored object, the servo would lock to give the person the feeling of touching the object. With such a system, the man in this loop would see the object in 3-D, feel it, and even have the sensation of picking it up and moving it about[75] (one could project this scene in to the microcosm or out to the macrocosm).

S. Silverstone and N. Negroponte have been developing graphic languages especially designed for representing aspects of urban growth and large city-size systems.[76,77] S. Boutourline has also been designing systems that employ enriched coupling.[78]

Another exciting venture into new languages is cited by G. Kepes, namely, the language of artificial *light*, a creative medium that has thus far been used mainly by artists. Kepes notes that "in spite of the current vogue of light art, we have only begun to explore the rich potentials of the medium. The works done today . . are no more than the individual words—if not just the letters of the alphabet—of an emerging vocabulary of light art."[79] He also cites the acceptance of randomness—that is, of nonmechanical process—as an important aspect of kinetic design that tended to avoid the deadening effect of mere cyclical events. Anyone who has gone to a pyschedelic nightclub, and simply allowed himself to "ride" the gyrations of strobe lights, knows how he can be led to a decidedly new and provocative perception of the world.

What we have been emphasizing here, largely because of their novelty, are physical applications of new languages growing out of new media. The full speaking of these languages will indeed press far out into territories lying beyond old constraints.

But these new languages, too, will evolve through definite levels of sophistication and abstraction, and each level gained will bring in its wake new powers of evolution of conceptualizing the world. Psycholinguists have also recently discovered that children, as they grow, pass through definite levels of grammatical structure and linguistic capacity that in some deep way must be a reflection of the natural evolution of man.[80]

So too, as S. Papert points out, must computers (or artificial intelligence machines) be trained, so that they will make *their kind* of conceptual leaps.[81]

On theoretical grounds, one must watch also for the work of G. Gunther, W. S. McCulloch, and R. Moreno-Diaz. Their "triadic logic," although not discussed here, will be seen in due course by the artificial-intelligence community as being of crucial importance in future mechanization of logical processes that are like those in living systems, logical processes that depend on "languages of becoming" rather than on "languages of being."[82,83]

A way of grasping how a man's on-line interaction with a computer can bring new meaning is to look back historically to the day that the movie was invented—a process of seeing photographs one after the other, projected at a rate faster than the visual system could accommodate as "stills" (this is, by the way, another example of time driving). Thus, it was discovered that this new medium, the movie, could arouse feelings, sensations, and insights that could not be aroused by single photos. The movie was a new language medium, and so too is the computer. Inasmuch as it is the step to artificial intelligence and to evolutionary applications of machines, we are only beginning to guess what possibilities of communication it holds in store for us and our children.

We owe so much to so many people for their often extensive and generous discussions, criticisms and suggestions during the course of our research that it is impossible to thank them appropriately. Those who have assisted us by joining the dialogue, though not necessarily agreeing with our point of view, include: S. Amarel, M. Arbib, S. Beer, R. Barron, D. Bobrow, S. Boutourline, M. Coler, S. Coons, P. Denes, C. Draper, P. Duran, D. Engelbart, L. Fogel, L. Gilstrap, G. Gunther, H. Hughes, A. Iberall, K. Ingham, A. Johnson, C. Jones, W. Kilmer, E. Konecci, J. Krasner, J. Lilly, W. McCulloch, M. McLuhan, J. Maddux, M. Mathews, M. Minsky, R. Moreno-Diaz, M. Noll, G. Pask, H. Robinson, O. Selfridge, T. Sheridan, S. Silverstone, R. Stotz, I. Sutherland, L. Sutro, R. Swanson, R. Taylor, H. von Foerster, members of the Creative Science Seminar, N.Y.U., and many others.

REFERENCES

1. Darwin, Charles, *On the Origin of Species*. Facs. of 1859 ed., intro. by E. Mayr. Cambridge, Mass.: Harvard Univ. Press, 1964, p. 459.

2. Brodey, W. M., and Lindgren, N., "Human enhancement through evolutionary technology," *IEEE Spectrum*, vol. 4, pp. 87–97, Sept. 1967.

3. Feigenbaum, E. A., and Feldman, J., eds., *Computers and Thought*. New York: McGraw-Hill, 1963.

4. Jones, J. C., "The designing of man-machine systems," Conf. on the Human Operator in Complex Systems, Univ. of Aston, Birmingham, England, July 26–28, 1966.

5. Pask, G., "Comments on men, machines, and communication between them," Vision 67—ICCAS, International Center for the Communication Arts and Sciences, New York Univ., Oct. 19, 1967.

6. Minsky, M., "Artificial intelligence," *Sci. Am.*, vol. 215, no. 3, pp. 246–260, Sept. 1966.

7. Minsky, M., "Steps toward artificial intelligence," *Proc. IRE*, vol. 49, pp. 8–30, Jan. 1961.

8. Solomonoff, R. J., "Some recent work in artificial intelligence," *Proc. IEEE*, vol. 54, pp. 1687–1697, Dec. 1966.

9. Sklansky, J., "Learning systems for automatic control," *IEEE Trans. Automatic Control*, vol. AC-11, pp. 6–19, 1966.

10. Barron, R. L., "Self-organizing and learning control systems," 1966 Bionics Symp., May 2–5, Dayton, Ohio.

11. Brooks, S. H., "A comparison of maximum-seeking methods," *Operations Res.*, vol. 7, pp. 430–457, 1959.

12. Rastrigin, L. A., "The convergence of the random search method in the extremal control of a many-parameter system," trans. from *Avtomatika i Telemekhan.*, vol. 24, pp. 1467–1473, 1963.

13. Matyas, J., "Random optimization," trans. from *Avtomatika i Telemekhan.*, vol. 26, pp. 246–253, 1965.

14. Gilstrap, L. O., Jr., *et al*, "Study of large neuromime networks," Tech. Report AFAL-TR-67-316, Dec. 1967.

15. Barron, R. L., "Inference of vehicle and atmosphere parameters from free flight motions," Paper no. 67-600, AIAA Guidance, Control and Flight Dynamics Conf., Huntsville, Ala., Aug. 14–16, 1967.

16. Barron, R. L., *et al.*, "Synthesis of a spacecraft probability state variable adaptive control system," Final Project Rept., June 30, 1967.

17. McKechnie, R. M., and Barron, R. L., "Design principles for self-organizing control system flight hardware," NAECON, Dayton, Ohio, May 15–17, 1967.

18. Von Foerster, H., "Quantum mechanical theory of memory," *Trans. Sixth Conf. on Cybernetics*, Josiah Macy, Jr., Foundation, New York, N.Y., pp. 112–145, 1950.

19. Ashby, W. R., *Design for a Brain*. New York: Wiley, 1960.

20. Barron, R. L., Private communication, Oct. 31, 1967.

21. Barron, R. L., "The future for self-organizing control," *Control Eng.*, to be published.

22. Fogel, L. J., *et al.*, "Intelligent decision-making through a simulation of evolution," *IEEE Trans. Human Factors in Electronics*, vol. HFE-6, pp. 13–23, Sept. 1965.

23. Fogel, L. J., *et al.*, *Artificial Intelligence Through Simulated Evolution*. New York: Wiley, 1966.

24. Fogel, L. J., "Inanimate intellect through evolution," *Naval Res. Rev.*, to be published.

25. Fogel, L. J., Private communication.

26. MacKay, D. M., "Ways of looking at perception," in *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, ed. Cambridge, Mass.: The M.I.T. Press, 1967.

27. Engelbart, D. C., "Augmenting human intellect: a conceptual framework," SRI AD 289565, Oct. 1962.

28. Engelbart, D. C., "Augmenting human intellect: experiments, concepts, and possibilities," Summary Rept., SRI Project 3578, March 1965.

29. English, W. K., *et al.*, "Computer-aided display control," Final Rept., SRI Project 5061, July 1965.

30. Engelbart, D. C., *et al.*, "Study for the development of human intellect augmentation techniques," Interim Progress Rept., SRI Project 5890, March 1967.

31. Hoffer, Eric, "Man, play and creativity," *Think*, vol. 33, no. 5, Sept.–Oct. 1967.

32. Engelbart, D. C., "Automated psychomotor skill training," unpublished proposal.

33. Engelbart, D. C., and Sorensen, P. H., "Explorations in the automation of sensorimotor skill training," Tech. Rept. NAVTRADEVCEN 1517-1, Jan. 1965.

34. Pask, G., "Machines that teach," *New Scientist*, vol. 10, June 1961.

35. Pask, G., "Thresholds of learning and control," *Data and Control*, Feb. 1964.

36. Lewis, B. N., and Pask, G., "The development of communication skills under adaptively controlled conditions," *Programmed Learning*, July 1964.

37. *IEEE Trans. Human Factors in Electronics*, vol. HFE-8, June 1967.

38. Weizenbaum, J., "Contextual understanding by computers," *Commun. ACM*, vol. 10, pp. 474–480, Aug. 1967.

39. Rosenblith, W. A., ed., *Sensory Communication*. Cambridge, Mass.: The M.I.T. Press.; also New York: Wiley, 1961. See "Editor's Comment," p. 824.

40. Walter, W. G., "Specific and non-specific cerebral responses and autonomic mechanisms in human subjects during conditioning," in *Progress in Brain Research—Vol. 1*. Amsterdam: Elsevier, 1963.

41. Farley, B. G., "Some similarities between the behavior of a neutral network model and electrophysiological experiments," in *Self-Organization Systems*, M. Yovits, ed. New York: Pergamon, 1962.

42. Farley, B. G., and Clark, Jr., W. A., "Activity in networks of neuron-like elements," *Proc. 4th London Conf. on Information Theory*, C. Cherry, ed. London: Butterworth, 1961.

43. Dewan, E. M., "Communication by voluntary control of the electroencephalogram," AFCRL-66-736, Oct. 1966.

44. Hecker, M. H. L., "The effects of task-induced stress on speech," AFCRL-67-0499, Aug. 25, 1967.

45. Ray, Charles, "The coming internal instrumentation of physiology and medicine," to be published.

46. McCulloch, W. S., *Embodiments of Mind*. Cambridge, Mass.: The M.I.T. Press, 1965.

47. Sutro, L. L., Kilmer, W. L., Moreno-Diaz, R., McCulloch, W. S., *et al.*, "Development of visual, contact and decision subsystems for a Mars rover," M.I.T. Instrumentation Lab. Rept. R-565 (July 1966–Jan. 1967), July 1967.

48. Kilmer, W. L., McCulloch, W. S., and Blum, J., "An embodiment of some vertebrate command and control principles," to be published.

49. Kilmer, W. L., "A reticular formation model decision computer for a robot," unpublished report.

50. Kilmer, W. L., Private communication, Dec. 15, 1967.

51. Arbib, M. A., *Brains, Machines, and Mathematics*. New York: McGraw-Hill, 1964.

52. Papert, S., paper presented at Conf. on Intelligence and Intelligent Systems, Athens, Ga., Jan. 15–18, 1967.

53. Amarel, S., "Introductory comments: on representations and modeling in problem solving," Conf. on Intelligence and Intelligent Systems, Athens, Ga., Jan. 15–18, 1968; publ. in RCA report.

54. Amarel, S., "On the automatic formation of a computer program which represents a theory," in *Self Organizing Systems—62*, M. Yovits *et al.*, eds. New York: Spartan, 1962.

55. Amarel, S., "On machine representations of problems of reasoning about actions—the missionaries and cannibals problem," in *Machine Intelligence 3*, D. Michie, ed., Edinburgh: Univ. of Edinburgh Press, 1968.

56. Amarel, S., "An approach to heuristic problem solving and theorem proving in the propositional calculus," in *System and Computer Science*. Toronto: Univ. of Toronto Press, 1967.

57. Amarel, S., "On the mechanization of creative processes," *IEEE Spectrum*, vol. 3, pp. 112–114, April 1966.

58. Gibson, J. J., *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin, 1966.

59. Johnson, A. R., "A structural, preconscious Piaget: heed without habit," *Proc. Nat. Electron. Conf.*, vol. XXIII, 1967.

60. Johnson, A. R., "A cybernetic trend," *Am. Soc. for Cybernetics Newsletter*, to be published.

61. Hermann, H. T., and Kotelly, J. C., "An approach to formal psychiatry," *Perspectives Biol. Med.*, Winter 1968.

62. Amosov, N. M., *Modeling of Thinking and the Mind*. New York: Spartan, 1967.

63. Ford, J. J., "The development pattern of the Soviet automation program," Am. Soc. for Cybernetics, unpublished.

64. Iberall, A., and McCulloch, W., "1967 behavioral model of man—his chains revealed," NASA Rept. CR-858, July 1967.

65. Brodey, W. M., "Information exchange modeled in the time domain," presented at Am. Psychiatric Assoc. meeting, Atlantic City, N.J., May 1966.

66. MacKay, D. M., "Self-organization in the time domain," in *Self Organizing Systems—1962*, M. Yovits *et al.* eds., New York: Spartan, 1962.

67. Gibson, J. J., "The problem of temporal order in stimulation and perception," *Artorga*, Aug./Sept. 1967.

68. Von Foerster, H., "Time and memory," from panel discussion, Interdisciplinary Perspectives of Time, *Ann. New York Acad. Sci.*, vol. 138, pp. 866–873, Feb. 6, 1967.

69. Brodey, W. M., "The clock manifesto," *Ann. New York Acad. Sci.*, vol. 138, pp. 895–899, 1967.

70. Smith, K. U., *Delayed Sensory Feedback and Behavior*. Philadelphia: Saunders, 1962.

71. Stotz, R. H., and Cheek T. B., "A low-cost graphic display for a computer time-sharing console," M.I.T. ESL-TM-316 Tech. Memo., July 1967.

72. Mathews, M. V., and Rosler, L., "Graphical language for the scores of computer-generated sounds," to be published.

73. Denes, P. B., "The elusive process of speech," *Bell Lab. Record*, pp. 254–259, Sept. 1966.

74. Noll, A. M., "The digital computer as a creative medium," *IEEE Spectrum*, vol. 4, pp. 89–95, Oct. 1967.

75. Noll, A. M., "Description of graphic manipulator," Private communication, Aug. 25, 1967.

76. Silverstone, S. M., and Rusch, C. W., "Communication in the design process," or "The medium is not the solution," *AIA J.*, vol. 48, Dec. 1967.

77. Negroponte, N., and Groisser, L., "Urban 5: An on-line urban design partner," IBM Sci. Center Rept. 36. Y15, June 1967.

78. Boutourline, S., "The concept of environmental management," *Dot Zero IV*, Sept. 1967.

79. Kepes, G., "Kinetic light as a creative medium," *Technol. Review*, M.I.T., pp. 25–35, Dec. 1967.

80. Smith, F., and Miller, G. A., *The Genesis of Language*. Cambridge, Mass.: The M.I.T. Press, 1966.

81. Papert, S., "Why machines can't think," First Ann. Symp. of Am. Soc. for Cybernetics, Gaithersburg, Md., Oct. 26, 1967, to be published.

82. McCulloch, W. S., and Moreno-Diaz, R., "On a calculus for triadas," M.I.T., R. L. E. Progress Rept., no. 84, pp. 335–346, Jan. 15, 1967.

83. Moreno-Diaz, R., "Description of the dialogue implicit in man–environment interaction—a triadic formal language," Private communication, June 1967.

Education for

Engineering involves such a wide range of human endeavor that it demands diversity and innovation in the preparation of its practitioners. Engineering effort can be divided into three major categories: engineering technology, engineering practice, and engineering science. The qualifications of individuals whose activities would fall into each of these mutually interdependent categories are examined, as are the problems involved in training these individuals.

The role of the engineer and his preparation for this role have major significance in a technologically oriented world. It is therefore not remarkable that the problem of providing education for those seeking careers in the engineering profession have occupied the attention of educators, industrialists, and officials in government. Education for engineering is a complex subject on which strong views are held, as reflected in many studies.

The title of this article was chosen deliberately, even though it has a more cosmic ring than these remarks may warrant. From my experience on the faculty and in the administration of an American university and as President of the IEEE, I have had the opportunity to observe the patterns of engineering needs in a number of countries in various stages of technological development. I am persuaded that no one pattern of education can be considered optimum for all countries; moreover, even within one country, uniformity of pattern for all institutions would be stultifying. I am convinced that engineering involves such a wide range of human activities that it must accommodate—and indeed demand—diversity and innovation in the preparation of its practitioners.

Engineering education is now, and has been in recent years, the object of much discussion in the United States, with many within and without the educational establishment expressing their views on the matter. There appears to be a fair degree of agreement that the present patterns of education need modification to meet both present and future needs; there is much less agreement on what pattern or patterns will satisfy these needs. Vigorous arguments have been put forth for technical breadth versus specialization, for more training in the economic aspects, for the humanization of the engineer, etc.

The structure of engineering education

The concern for the structure of engineering education is not unique to our time, although earlier studies in response to such concerns moved at a more leisurely pace. In the United States the Society for the Promotion of Engineering Education was established as a voluntary association of engineering teachers in 1893. Some 14 years after its founding it invited the professional societies to join in a general committee for the study of the activities and objectives of engineering schools. Under the sponsorship of the Carnegie Foundation for the Advancement of Teaching, this effort resulted in the publication of a comprehensive report, which was described as "a single and clear treatment of the complicated problem involved—and the connection between the curriculum and the changing demands of industrial activities and growth." The words have a familiar ring. Such periodic major studies have been a regular feature of the scene in U.S. engineering education: witness the Wickenden Report of 1926–27, the Hammond Reports of 1940–44, the Grintner Report of 1955, and the as-yet-incomplete Study of the Goals of Engineering Education (under the auspices of the American Society for Engineering Education).[1] This most recent study has been a penetrating analysis of the present state of the engineering profession and the projections for its future. The recommendations presented in its preliminary report aroused a great deal of controversy, which highlights, I believe, a lack of clarity in our perception of engineering.

The scope of engineering and the levels of sophistication necessary for its activities have very much widened in the past quarter of a century. Our difficulties in reaching agreement on the issues of engineering education to meet the needs of the present and foreseeable future are hampered, in my view, by historical development and by problems that are semantic in nature. Unfortunately, the term "engineer" has been applied indiscriminately to those who operate engines and to those who create them, so that it is no wonder that the general public has difficulty in recognizing the engineer's professional role. But within the profession of engineering we engage in arguments as to whether the individual pressing at the

Revised text of a talk presented before the Institution of Electrical Engineers in London on October 17, 1967. This article is also scheduled to appear in a forthcoming issue of *Electronics and Power*, published by the IEE.

the engineering mission

Perhaps the most sensible approach to the complex problem of engineering education is to focus our attention first on the nature of the "engineering mission," and then direct our efforts to educating individuals toward its fulfillment

William G. Shepherd *University of Minnesota*

frontiers of science or the man engaged in implementing long-understood principles is the "real engineer." The education best suited to make each most effective is quite different. It should be apparent that if we accept both activities as engineering and at the same time recognize that different levels of ability and preparation are required, one cannot speak of the education of an engineer as a single track along which all must follow.

I believe that if we turn away from the semantic description of an engineer and focus our attention on the engineering mission and the preparation of individuals required for the fulfillment of that mission we may ultimately achieve a greater consensus. For if we continue to push the limits by talking about what *an* engineer *ought* to be, the composite picture of the engineer that emerges as optimally fitting the needs of industry and society is clearly recognizable as a fiction. He would be a man broadly understanding of basic physical science and mathematics, inventive, capable of the design and management of complex systems; a scholar, knowledgeable of man's goals and aspirations, versed in the social sciences so that he can properly assess the impact of his activities on society; a personnel manager who deals effectively with his subordinates, at once willing to put society's good above profit and to provide his employer's shareholders with an adequate return on their investment; someone with a Ph.D. and 15 years' experience, but not over 25 years old—a genius indeed.

The absurdity of these composite qualifications is patent, but when one examines the expected goals of some educational programs and the requests of employers it is not too far off the mark. It is not altogether surprising that such elaborate and impossible criteria develop; they are actually an expression of the growing complexity of the technology on which the world increasingly depends and of the engineer's expanding role in the structuring and management of our civilization.

The concept of the engineering mission

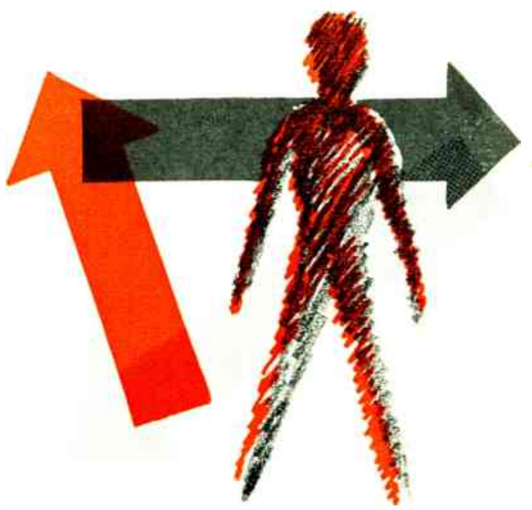
I suggest that we turn then to a look at the engineering mission, assuming that we can come to agreement on a definition of this mission. One of the most often quoted definitions of the mission of the engineer is taken from the charter of the earliest of the British civilian engineering societies—the Institution of Civil Engineers, incorporated in 1828.

The charter refers to civil engineering as the "art of directing the great sources of power in nature for the use and convenience of man . . . as the means of production and traffic in states, both for external and internal trade, as applied in the construction of roads, bridges, aqueducts, canals, river navigation and docks for internal intercourse and exchange, and in the construction of ports, harbours, moles, breakwaters and lighthouses, and in the art of navigation by artificial power for the purposes of commerce, and in the construction and adaptation of machinery, and in the drainage of cities and towns."

The opening of this statement, if we interpret "sources of power in nature" broadly, can serve to define the mission of all engineering and at all levels of sophistication. What follows this opening stresses the role of the engineer in the physical realization of structures, which is not unnatural since, ultimately, this is the way society is served. But the sometimes acrimonious arguments within the profession, centering on whether the title "engineer" should be reserved for those who are engaged in this culminating activity, lead to endless, fruitless debate about the patterns of engineering education and the nature of publications and technical meetings of professional societies.

Our problem, as Harvey Brooks, Dean of the Division of Engineering and Applied Physics, Harvard University, has so clearly stated,[2] arises from the fact that in engineering, as in medicine, science has overtaken art. He notes that the "learned" professions, in which he includes engineering, rest on a base of theoretical and empirical knowledge. He goes on to point out that the professional is much more than a man with knowledge; he is the middleman between a body of knowledge and society and has the responsibility to use both existing and new knowledge to provide the services that society wants and needs. The dilemma of the professional arises, Brooks says, from the fact that both the body of knowledge he must use and the expectations of society are changing and expanding so rapidly. The structure of a profession developed to fulfill a simpler mission must be adapted to meet this more complex situation.

The basis for Brooks' conclusions can be seen in the developments of the past quarter century, which have made engineering a more significant force in terms of the increasing complexity of its contributions to society and in the proportion of the numbers of engineers in the total labor force. All the present signs point to a continuation of these trends.



An increasingly complex technology

The diversity of physical phenomena with which the engineer must now deal is difficult to quantify but conservatively it is of an order of magnitude more complex than it was 25 years ago. Many of the developments that have become a part of our technology involve the engineer with fields, such as quantum and nuclear physics, that were strictly in the domain of the research physicist a quarter of a century ago. Indeed, some of them exploit phenomena and techniques undiscovered at that time. The rate of change of the base in the physical sciences is still increasing; this fact, coupled with the increasing interaction of engineering with the social and biological sciences, suggests that the diversity of the scientific base with which the engineer will be concerned 25 years hence will have expanded by an even greater factor.

The widening contribution of engineering to the development of modern society is reflected in the proportion of the U.S. labor force engaged in engineering. At the turn of the century, the proportion was one engineer to each 673 workers. By 1920 the ratio was one in 311; in 1940, one in 165; and in 1960, one in 76. A continuation at this pace would result in a ratio of one engineer for every 20 workers near the end of this century. Neither the interests nor the qualifications of the general population support an expectation of a continuation of this trend.

The growing complexity of engineering technology is both reflected in and a consequence of the rise in the number of doctoral degrees awarded in engineering in the United States. As is pointed out in the ASEE Interim Report of the Committee on Goals of Engineering Education,[1] doctoral work in engineering is not a recent innovation in the United States but is as old as doctoral education generally. The first earned doctorate in any field in the United States was awarded in 1861 by Yale University. Two years later the first engineering doctorate was awarded by the same institution, interestingly enough to Joshua Willard Gibbs of thermodynamics fame for a dissertation on gear trains.

By the turn of the century the production of doctorates had reached one or two per year, at which time the ratio of production of bachelors–masters–doctorates was approximately 1000:100:1. The rate of production of bachelor's degrees has been increasing for over a half century along a trend line of 5 percent per year, the production of master's degrees along a trend line of 11 percent per year, and the rate of production of doctorates along a trend line of 12 percent per year (twice the rate of annual increase of doctorates in all fields). Even with the rapid increase in production, the supply of doctorates continually is short of the demand. By 1976 it is estimated that about one in 12 bachelors graduates will go on to a doctorate. This leads immediately to the question: Is there a sufficient pool of talent to provide the basis for an expansion of this most highly trained group? A study by Dael Wolfle, entitled "America's Resources of Specialized Talent," is encouraging in this regard; it shows that only one percent of those who have the average intellectual level of the doctorate population now go on to that degree, and that even at the genius level only one in five does so. Many, of course, lack other attributes, or are uninterested in engineering. We must further recognize that the social and life sciences, as they become more mathematically and analytically oriented, will increasingly draw on the same pool of unusual talent.

What can we conclude from these trends and what is their significance for education for the engineering mission?

1. There is every reason to believe that society's demands on the engineering profession for the satisfaction of its needs will continue to expand. Quite apart from the increasing sophistication of society's technological expectations, we need only look at some of the less sophisticated but extraordinarily difficult problems created by our increasing urbanization, such as transportation, energy supply, and air and water pollution.

2. Although the proportion of engineers in the total labor force may be expected to increase, it seems likely that limitations of aptitude and interest will cause a leveling off in the next quarter century.

3. The expansion of our knowledge of the basic sciences, the growing interaction with the social and physical sciences, and the expectations of society of an early translation of science into service will sustain a rising demand for engineers educated to the doctorate level.

4. The limitations of aptitude or interest and the increasing competition for highly trained individuals in

other fields suggests that engineers with doctoral training will be in short supply.

Thus in approaching the preparation of men and women for the engineering mission, we must recognize that we are dealing, on a relative scale, with an increasingly scarce resource. The engineering talents we develop and utilize will become more and more precious, and inefficient usage or inadequate preparation will become less and less tolerable.

A medical parallel

I should like to digress for a moment to point out that our problem in engineering has a parallel in medicine. Recently my attention was directed to an address presented by Dean J. F. McCreary, University of British Columbia, at the 15th International Hospital Congress held in Chicago in August 1967. I should like to quote a few paragraphs from his address because they provide some additional insight into our own problem.

"When I graduated from medical school over 30 years ago it was possible for my instructors to expose me to all that was known about medicine. I could emerge into practice with the comfortable feeling that with ordinary intelligence I should be able to practice medicine as well as anyone in the world. How very different it is today. It has been stated that 80 percent of all that is known about biological science has been discovered during the past 40 years. It is stated that knowledge in medicine doubles every 10 years and the World Health Organization tells us that a new medical article is being published every 23 seconds in the world today. This explosion of medical knowledge has made the tasks of medical educators very much more difficult. There are many evidences of the unrest which this has produced in medical schools. . . ."

"[One] aspect of medical education which has been carefully studied is the length of time required to prepare a physician for his tasks. Rather intense studies have taken place in a number of universities. As you know on this continent the usual pattern is that the individual graduates from medical school eight years after he has graduated from high school. He takes four years for a college degree and then four years of medicine. One indication of the frank uncertainty about the correctness of our present standards is indicated by the fact that among four universities that have studied this matter most carefully, three have reduced the length of time for a segment of their medical students to six years instead of eight whereas in the fourth another year has been added increasing the time to nine years. It is clear that we really don't know how much time is required to educate a physician. . . ."

"It seems likely that if careful studies could be made of the type of health services required by a community it would be apparent that many of the duties now performed by overworked doctors could be undertaken successfully by individuals with significantly lesser training. It is clear that studies of the exact nature of the health needs of a community must be made. We must learn precisely how the public health nurse, the physiotherapist, the social worker and the doctor are contributing and how they could contribute. With the cooperation of all professional groups we must frankly experiment with the use of the various health professions and try to find the most effective and economic methods of meeting the needs of our patients."

If we examine these statements, changing medicine to engineering, we can see our own situation in what he says. And some of his suggestions for the future organization of the health professions to meet society's needs suggest ways in which we may find it both desirable and perhaps necessary to educate and deploy engineering manpower.

It is interesting to compare the first of the foregoing three excerpts with the following statement in December 1966 by Elmer W. Engstrom of RCA[3]:

"As an example, my own career of forty-three years has run parallel in time and experience to the growth of electronics and mass communications from the early days of radio broadcasting. In 1923, when I joined General Electric as a young engineering graduate, an individual could comprehend the bulk of available knowledge about electronic theory and application. The outer limits of the technology were clearly visible, and the most advanced apparatus held no mysteries for anyone who had completed a basic course in electrical engineering. None of us, in fact, would have recognized that we were engaged in electronics, because the term itself had not been coined. We were in radio, and that offered glamour enough."

Categories of engineering effort

The foregoing has been, perhaps, an overly long preamble to my main topic but necessary, I believe, for an understanding of the problem of education for the engineering mission. I propose to distinguish three major categories of effort that must relate to and support each other in order for the engineering mission to be fulfilled. These are necessarily arbitrary and thus boundaries between them are blurred.

The categorizations that I find useful are those defined in an unpublished study dated September 16, 1966, entitled "Statement on Goals of Engineering Education," prepared by Harvey Brooks, F. Karl Willenbrock, and F. H. Abernathy, of Harvard University's Division of Applied Physics:

"*Engineering technology* refers primarily to the application of well-established technology in production and service as well as in some of the supporting aspects of research and development. . . ."

"*Engineering practice* refers to the creative application of existing knowledge to the solution of specific engineering problems. It is not concerned primarily with the development of new knowledge or of generic solutions extending beyond the particular problem attacked. . . ."

"By *engineering science* is meant those fields of science which are of interest primarily from the standpoint of applications. . . ."

It may be noted that the major objective in engineering science is a reduction in the degree to which it is necessary to rely on empiricism in engineering. The marked reduction of empiricism in engineering practice in the past years has accompanied the sharp increase in the numbers of engineers trained to the highest level. Thus, the engineering scientist seeks an extension of the understanding of basic phenomena or the development of generic solutions so that their practical implications can be fully exploited.

I am sure that many of you will find that the classification into three groups is too restrictive and that the groupings will not readily accommodate some classes of activity that clearly involve engineering. This depends on how flexibly they are interpreted and, although they represent an oversimplification, in the main they are adequate. Let us now examine the qualifications of individuals whose activities would fall in the three categories. These are also discussed in the Harvard "goals" statement.

Engineering technology. Individuals engaged in engineering technology are expected to provide technical support for the engineering practitioner and scientist. They should be

1. Well versed in the current state of the art of a particular technology, capable of utilizing handbooks and other forms of codified information with skill and discrimination.

2. Sufficiently versed in mathematics and the sciences related to the particular technology to distinguish sound procedures from unsound ones and to keep up with the current innovations in their special fields as these innovations occur.

Prior to World War II most engineers were engaged in engineering practice and generally served as their own engineering technologists. The increasing demands for more scientifically based engineering curricula has forced most engineering schools to abandon most "state of the art" instruction. The gap thus created has resulted in a growing deficiency of competence in this vital sector of the engineering mission. Recognition of the unfilled need has led to the establishment of institutions and educational programs to train engineering technicians or technologists. In the United States there are now a number of two-year institutional programs. In 1964 these programs produced approximately 14 500 graduates and it is probable that many more were trained in company-sponsored programs. Some estimate that by 1970 (when, perhaps optimistically, there will be 1.5 million engineers) the need will exist for two to four technicians for each engineer and scientist.* Most U.S. two-year technician programs provide for only minimal requirements; as pressures develop for greater sophistication in the qualifications of engineering practitioners, there will probably also be pressures to upgrade the training of the engineering technologists so that the duration of the present two-year programs will have to be lengthened. I am sure that they will eventually reach university degree level. Presumably in response to these trends, there are presently between 70 and 80 institutions in the United States offering four-year programs for engineering technologists. (Most of these four-year programs have been created since 1950.)

Engineering practice. The major center of controversy about the preliminary reports of the ASEE Committee on Goals of Engineering Education has revolved about the question of whether the first *professional* engineering degree should be awarded after a four-year bachelor's program or after a five-year program involving one year of study beyond the bachelor's and leading to a master's degree. However, one discovers that most four-year programs in the United States currently cannot be completed by the average student in less than four and one-half years, so the question of time scale is perhaps moot. The debate about the duration and content of curricula at this level is not insignificant, since it affects the plans of the majority of those embarking on careers in engineering practice or science. But for the present discussion, I should like to set this question aside with confidence that some sensible solution will be reached.

The expectations are that the engineering practitioner will be characterized by

1. An ability to handle mathematics and science related to a general area and to handle problems not in handbooks.

2. A greater concern with finding a needed solution to a specified problem than with an understanding of all aspects of the science or mathematics involved.

3. An ability to synthesize practical designs that satisfy a number of requirements, several of which may be in conflict.

4. A sensitivity to economic factors and an ability to effect trade-offs between partially conflicting objectives.

5. An ability to utilize formal technical background and practical experience to solve problems which are new in detail, but not new in concept.

6. An ability to direct large-scale technical operations by coordinating and supervising the efforts of appropriate specialists.

Engineering science. The activities and methods employed by the engineering scientist generally differ little from those used by people who have sought careers in the basic sciences. It should not be surprising that there is a fair degree of mobility of individuals across the boundaries between pure and applied scientific activities. Many of our most able scientists have recognized the practical implications of their basic studies and, conversely, some of the most important contributions to basic knowledge have resulted from the pragmatically motivated investigations of engineering scientists. But difference of motivation between those who work on the two sides of the ill-defined boundary should be stressed, since it results in a basic difference in their end actions. For semantic purposes I will distinguish between the engineering scientist and the scientist. The latter has as his basic concern the expansion of the knowledge of fundamental laws of nature without necessary regard for the relevance of his findings for purposes other than the intellectual satisfaction of society. The engineering scientist may be equally concerned with expanding the understanding of the fundamental laws of nature but he does so with the clear intent of asking another question: What are the implications of this new knowledge for the use and convenience of man?

Important attributes of the engineering scientist are inventivity and creativity. A significant aspect of his training is the encouragement to see these as representing the highest fulfillment of his efforts. Both the engineering scientist and the scientist are working at the frontiers of the known and both will normally have been educated to the doctorate level.

I have proposed, then, that we consider the need for educational preparation of three different and reasonably distinct groups of students, all of whom will eventually engage in engineering, albeit in different ways. It is through their cooperative efforts that the engineering mission can be most effectively met. I should like to stress the mutual interdependence of their activities. The efforts of the engineering scientist would be sterile without the activities of the engineering practitioner and the pace of implementation would be slowed without the support of the engineering technologist.

* This statement has been challenged as an overestimate on the grounds that the expanded uses of the computer as a technical tool will eliminate many of the tasks now performed by technologists.

If the whole is the sum of its parts, efforts must be made to insure that each segment is strong and designed to coordinate with the other segments. Thus faculties concerned with the preparation of engineering practitioners need to collaborate with those whose primary efforts are directed toward engineering technology or engineering science. This will insure, on the one hand, that new knowledge will flow usefully into educational programs and, on the other, that each segment of the engineering profession will develop an understanding of the expectations of the other segments.

Training problems—technical and nontechnical

The remainder of this discussion will be devoted to some of the problems we face in the preparation of engineering technologists, practitioners, and scientists. A major concern has to do with the *nontechnical* aspects of the engineer's education, a concern that has been most evident at the practitioner and the scientist levels. If those involved in the engineering mission are to assume an increasingly prominent role in decision-making processes affecting both the economic and esthetic well-being of our communities, they must have an educational background that will provide them with both an understanding of social expectations and the skills to deal with them. Engineering curricula of the past unfortunately have been deficient in these aspects. Engineering graduates who have not had the opportunity to broaden their intellectual background in this respect are sharply critical of this curricular deficiency and urge reform.

Similarly, a number of our present-day engineering practitioners, and some engineering scientists, feel handicapped by their lack of background in management. Management cuts across all levels of sophistication in the engineering mission, and the problems of management are not only those internal to engineering activities per se, but encompass the consequences of engineering decisions on the welfare and quality of our society. In addition, internal to engineering one may distinguish skills in technical management and general management.

Particularly in the case of the engineering practitioner, we face major training problems having to do with specialization. The breadth of the scientific base of present-day engineering, its rate of expansion, and its complexity impose an increasing premium on a concentration on the fundamentals of basic and engineering science that undergird the broad engineering specialties at the bachelor's level. Failure to concentrate on the fundamentals will limit the effective engineering lifetime of the graduate, although an exclusive concentration on fundamentals produces an engineer who is sometimes described as one who "doesn't know how to do anything." And, as I mentioned earlier, there is the further pressure on precious curricular time: the desires of engineers for the opportunity to become "educated men," through wider acquaintance with the humanities and social sciences, which cannot be ignored either.

The demands of a curriculum that will establish the fundamental base needed to meet the broad expectations laid down earlier and provide an opportunity for a liberal education will leave little time in a bachelor's program for any intensive technical specialization. Thus, unless there is a greater willingness to extend undergraduate engineering programs beyond the nominal four years traditional for the American bachelor's degree in engineering than has thus far been evident, specialization must be left for some form of postbaccalaureate study. This training might be accomplished either in graduate programs or through continuing education. The trends toward general programs in the undergraduate engineering curricula have led many to suggest that there should be pre-engineering programs paralleling those now expected prior to admission to the study of law or medicine in the United States. Engineering would then become "professional" or "graduate" in the same sense as medicine or law. Engineering does not yet seem ripe for this development, however; a compromise involving specialization through graduate or continuation study seems a more likely pattern for some time to come.

There are countervailing pressures against general studies in favor of programs that will steer students into an early concentration in subspecialties. Although these programs can provide for immediate competence in dealing with present problems, they will narrow the employment options of the graduate and, because his training makes him less adaptable, are likely to subject him to the risk of rapid professional obsolescence. Thus, in my opinion, early specialization at the expense of a basic foundation will ultimately underutilize a precious resource. Such programs seem inappropriate for the engineering practitioner, as earlier defined, but they may be suitable for the preparation of the more able individuals who seek careers in the area of engineering technology.

Doctorate programs

Let me next comment on the educational programs for the engineering scientist. Present programs leading to the doctorate in engineering fields have generally developed along lines parallel to those of the sciences. The doctorate is a research degree and since the great majority of those who earn these degrees enter research or teaching, at least initially, this approach is not inappropriate.

However, one criticism that has been leveled at engineering doctorate programs is that their very similarity to the science doctorates engenders an emphasis on a unique contribution to knowledge, resulting in a lack of the concern for the applicability of the results. In essence, this charge is that the engineering scientist apes the scientist, who has a different mission. Since this criticism has some validity, it is suggested that industrial laboratories and academic engineering institutions could profitably join in stimulating, by a variety of means, an interest in the relevance of doctoral research findings to application. An understanding of what is frequently called the "real world," either through past experience or consulting, is a valuable asset for faculty engaged in the supervision of graduate research in engineering. The use of competent industrial personnel as adjunct professors joining with regular faculty in supervising graduate thesis research in the university milieu, or conversely in an industrial laboratory, would encourage an interest in the inventive consequence of research.

I should like to suggest a role for the engineering scientist that departs from the traditional concentration on research. The solution of increasingly complex engineering problems demands a breadth and depth of disciplinary competence that can rarely, if ever, be found in an individual, and thus reliance must be placed on the composite competence of a team of specialists. The level of sophistication required in making decisions about the synthesis of the contributions of highly trained specialists demands of the team leader an intellectual and critical capacity equal to and probably greater than that of a research investigator. Such individuals must be generalists who have sufficient depth to appreciate the contributions of specialists and to understand how these can be marshaled to accomplish a complex goal. Individuals assigned such responsibilities presently emerge from the ranks of engineering practitioners or engineering scientists.

That these individuals are in short supply is evidenced by the widespread urging for universities to train men for the systems role. Whether this type of training can be systematized remains to be determined, but the whole problem is relevant to engineering education. The experience in training should be real, just as the research experience of the doctoral candidate should be real. Within universities, however, the opportunities for experience with real rather than synthetic or simulated systems is rare. It may be that meaningful programs can be mounted only through the cooperative efforts of universities with major industries or government agencies. An innovation of this kind would require a willingness on the part of graduate faculties to depart from tradition and on the part of government agencies and industry to provide opportunities for a significant intern experience for the students.

The benefits of cooperation between practicing engineers and faculty in the educational endeavor are not confined to the graduate level. Particularly in some of the areas of engineering specialization, such as design, practicing engineers have much to contribute, and continuing efforts should be made in this direction.

Continuing education

In the earlier discussion of education for the engineering mission only a passing reference was made to continuing education. However, it is implicit in the expectations defined for those engaged at all levels of the engineering mission that continued study will be necessary if an individual is to maintain his competence and keep pace with the expanding base of his field. All institutions at all levels that provide formal educational programs have an obligation to instill in students the concept that continuity in education is an ongoing obligation they assume if they are to meet fully the changing requirements of their profession. Employers, in their own self-interest in conserving their limited resources of engineering talent, should also foster this attitude and aid in providing opportunities for continued education. The need for physical refreshment during annual holiday leaves is commonly accepted; leaves for intellectual refreshment would return at least as great dividends to employer and employee alike. A number of industries have recognized that the device of sabbatical leaves is useful for their most valuable personnel. This practice could provide opportunities for exchange of personnel between universities and industries that could benefit both the individuals and their institutions.

Continuing education in engineering cannot be the sole responsibility of educational institutions; rather, it should be shared by universities, professional societies, and industries. However, the divided responsibility for organizing continued education has resulted in opportunities deficient in scope, quality, and availability. An intensive effort to find a solution is warranted.

Conclusions

In the foregoing discussion I have presented a picture of the complex professional structure in engineering that has been emerging from a simple historical base. The complexity reflects a parallel emergence of complexity in society itself. In one sense the problems and patterns emerging for engineering education are a case in point of the problems faced by all branches of modern education. That the problem is difficult should not amaze us; the nature of our world does not permit the easy certainties or formulary answers possible in an earlier age. That the answers to problems are tentative and often perplexed by operational difficulties should not surprise us. In a world of change, each new response must be accompanied by a readiness for further change.

Yet there is an enduring perspective on the mission of engineering that should influence each of our particular decisions on the design and conduct of engineering education. Within the mission is a complex of tasks requiring the contributions of individuals of varying interests and intellectual capacities. The effective accomplishment of the mission requires that these people work together with a mutual respect for the relevance and interdependence of their roles. We need to delineate roles and devise educational patterns that will fully develop the talents of those whose aptitudes and interests best suit them to a given role. This complex mission makes increasing demands on talent that must be shared in the fulfillment of the requirements of other human endeavors. Once we have developed talent at all levels to its fullest potential we must deploy it so that each individual contributes, in accordance with his ability, to the accomplishment of the engineering mission of serving man in harmony with nature. It is a problem that demands our utmost effort.

REFERENCES

1. American Society for Engineering Education, "Goals of engineering education," Preliminary Rept., Oct. 1965; Interim Rept., Apr. 1967; Office of Dean of Engineering, Purdue University, Lafayette, Ind.

2. Brooks, Harvey, "Dilemmas of engineering education," *IEEE Spectrum*, vol. 4, pp. 89–91, Feb. 1967.

3. Engstrom, E. W., "Science, technology, and statesmanship," *Am. Scientist*, vol. 55, pp. 72–79, Mar. 1967.

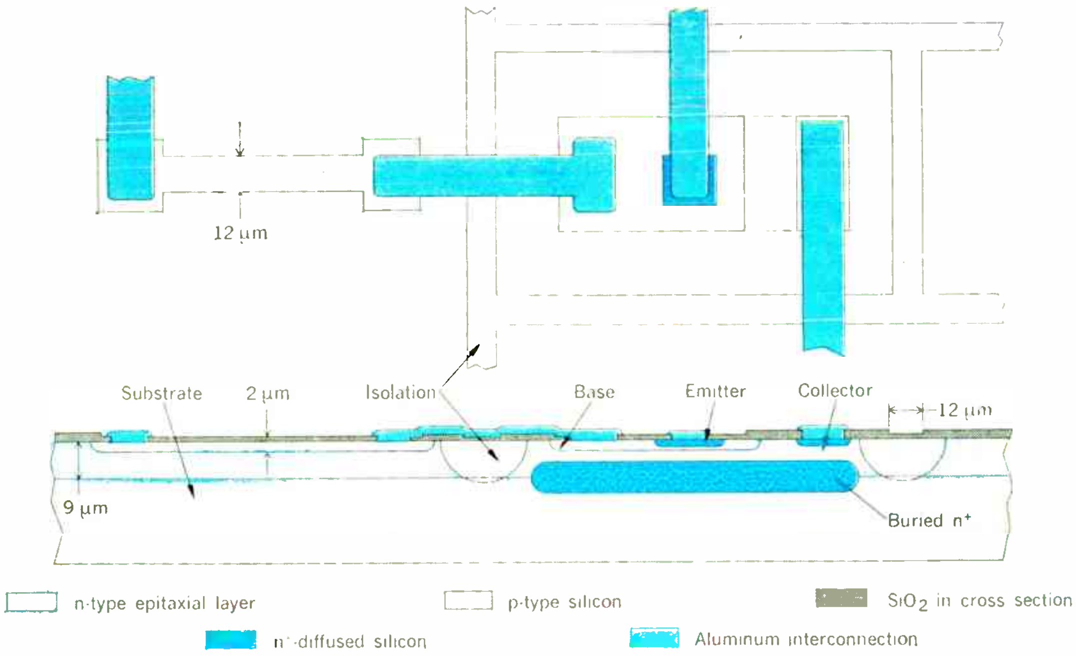


FIGURE 1. Cross section of bipolar SIC.

Silicon device technology

Bipolar and MOS silicon integrated circuits are having a dramatic effect on the electronics industry. On the basis of processes developed in the past few years in Great Britain, recommendations are made for future developments in this field

D. H. Roberts *The Plessey Company Limited*

Despite the recent gains in silicon integrated circuit technology, further developments are necessary to provide lower costs, increased control capability, improvements in performance, and better overall understanding of the problems and potentials of this burgeoning field. The concept of optimum-scale integration is examined in detail.

In view of the many detailed descriptions of the technology used in the fabrication of silicon integrated circuits (SICs) that have appeared in the literature,[1-3] there is no intention of covering the same ground in this article. It will be noticed that "silicon technology" has already been qualified to read "SIC technology." This change is considered to be adequately justified by the dramatic impact that SICs—both bipolar and metal-oxide semiconductor (MOS)—are beginning to have on the electronics industry and subsequently will have on civilization itself.

The purpose of this article is to use two existing SIC production processes as a basis for a critical examination of the following questions: (1) Is further development of SIC technology necessary or desirable? (2) If so, in what ways should this development be steered in order to maximize its usefulness?

Finally, brief mention will be made of areas in which SIC technology may not penetrate, but in which similar techniques, when applied to other semiconductor materials, certainly will.

Typical SIC processes and component characteristics

A cross section of a silicon slice processed to form bipolar SICs is shown in Fig. 1. As will be immediately apparent this particular process uses epitaxial isolation with a subepitaxial diffusion ("buried n+"). Some details of the process parameters and the subsequent device characteristics are given in Tables I and II respectively. This process, known as "Process I," with its subsequent component characteristics, was developed

I. Bipolar process parameters

Epitaxy:

Layer thickness = 10.5 ± 1 μm
Resistivity ≈ 0.68 ohm-cm (n-type)
(Defined such that subsequently 12 volts < BV_{CEO} < 24 volts)

Arsenic diffusion:

Sheet resistance = 7 ± 2 ohms per square

Isolation diffusion:

Sheet resistance = 25 ± 2 ohms per square

Base diffusion:

Junction depth = 2.15 ± 0.1 μm
Sheet resistance = 98 ± 12 ohms per square

Emitter diffusion:

Sheet resistance = 3 ohms per square
(Depth control by h_{FF} measurement)

Aluminum interconnects:

Thickness ≈ 1 μm
Sheet resistance ≈ 0.025 ohm per square

Oxide thickness:

Before isolation ≈ 1 μm
Regrowth over base = 0.4–0.5 μm

II. Bipolar device characteristics for circuit designers

Breakdown voltages:

$BV_{CEO} > 12$ $BV_{CIO} > 25$
$BV_{CBO} > 20$ $BV_{EBO} = 5.05–5.5$

Capacitances (for a typical geometry), pF:

$C_{OB} = 2 \pm 0.15$ $C_{IB} = 1.8 \pm 0.25$
$C_{CI} = 5 + 0.7, - 0.3$
C (decoupling) ≈ 2.4 pF/mil² ≈ 3700 pF/mm²

Transistor current gain:

$h_{FE} = 40–200$

Frequency f_T:

Dependent on geometry, but typically 750 ± 100 MHz

Resistor tolerance:

2-mil (50-μm) width = ±14%
1-mil (25-μm) width = ±17%
0.5-mil (13-μm) width = ±20% (selection to ±15% for RTL circuits)

Resistor ratios:

2-mil (50-μm) width = ±3.5%
1-mil (25-μm) width = ±4%
0.5-mil (13-μm) width = ±5%

Transistor matching:

h_{FE} ± 10% (90th percentile)
V_{BE} ± 0.65 millivolt (50th percentile)
± 3.5 millivolts (90th percentile)

III. MOS process parameters

Substrate material:

Resistivity = 10 ± 3 ohm-cm (n-type)

Oxide thickness:

Thick oxide = 1.6 ± 0.1 μm
Gate oxide = 0.15 ± 0.02 μm

Source drain diffusion:

Sheet resistance = 65 ± 10 ohms per square
Junction depth = 3.1 ± 0.3 μm

IV. MOS characteristics for circuit designers

Breakdown voltages:

$BV_{DSS} = 24–80$ $BV_{GSUB} = 24–80$ for gate protection diode

Threshold voltages:

$V_T = 4–6.5$ $V_T > 24$ (> 30 by selection)
for thick oxide "spurious" MOS transistor

Bottomed "on" resistance, ohms:

$r_{ON} < 2800$ for ~6-μm source-drain spacking and 100-μm channel width

Capacitances, pF/μm²:

Aluminum substrate
Thick oxide = (2.4 ± 0.5) × 10^{-5} (zero bias)
Gate oxide = (1.9 ± 0.02) × 10^{-4} (zero bias)

in the summer of 1964 at the Allen Clark Research Centre of the Plessey Company Limited. The various comments to follow are based on three years' experience with this standard process.

The corresponding information regarding a p-channel enhancement mode technology is given in Fig. 2 and Tables III and IV. This process (referred to subsequently as MOS I) was developed about 2½ years ago and has been used for the last two years for the fabrication of MOS array integrated circuits, or MOSAICs.[10]

Residual limitations of SIC technology

In spite of the use of the word "limitations," the fact is that the technology discussed briefly in the preceding section and the similar processes used for the production of SICs in many parts of the world, would in itself revolutionize the electronics industry even if all technology R&D ceased as of now. Luckily, however, such an occurrence seems unlikely and it is profitable, therefore, to consider the ways in which the technology needs to be further improved. These ways can be considered under four major headings: (1) cost reduction, (2) control improvements, (3) performance improvements, and (4) better understanding.

Before we examine the various steps in SIC technology against these criteria, it is necessary for the writer to set out the general philosophy of SIC design and usage within which his other judgments will be made:

1. SIC technology is not just *a* way of making cheap components, but rather *the* way to make electronic equipment.

2. On this basis, designers of circuits and equipment should be encouraged to apply SIC technology in the manner that leads to overall minimum "cost of ownership."

3. This approach leads directly to the concept of optimum-scale integration (OSI), in which the *optimum* complexity of a silicon chip depends on the type of equipment, the market for which it is intended, the available time for development, the production volume and time scale, etc. This is an essentially "evolutionary" attitude to chip complexity. It is opposed to the artificially "revolutionary" attitude of simple general-purpose building blocks (such as multiple gates, counters, etc.) being directly superseded by custom arrays of standard gates (with or without discretionary wiring), usually termed large-scale integration (LSI).

4. The achievement of a flexible approach to SIC design and usage in an economical manner demands, par-

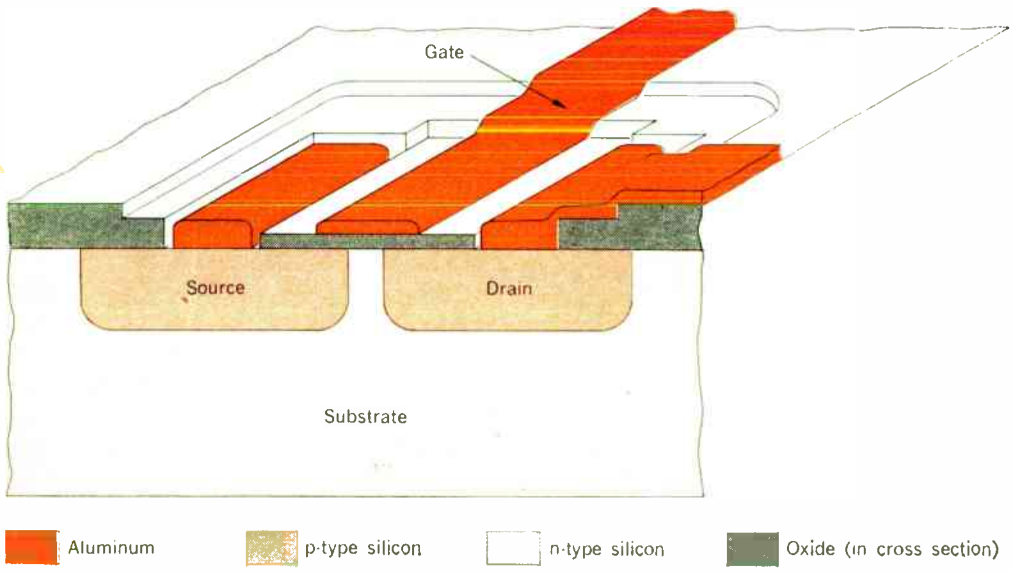


FIGURE 2. Cross section of MOSAIC.

adoxically, a high degree of inflexibility in the actual silicon technology, on the principle that it is much cheaper to invent a new circuit solution to a problem than to develop, control, characterize, and establish the reliability of a modified SIC process (for example, to increase the epitaxial layer resistivity to make "high-voltage" devices). More detailed accounts of this process standardization and the approach to circuit design have been given respectively by Hester[4] and Foss.[5] In Fig. 3 examples are given of the range of circuit requirements that have been satisfied by the single "Process I," referred to previously.

We can now return to the ways in which SIC technology should, and will, be improved.

Cost reduction

Although it is recognized that the most significant area for production cost reduction at the present time lies with the development of new assembly and packaging techniques (for example, flipped SICs on ceramic substrates), it is considered that such activity lies outside the terms of reference of this article.

When considering cost reduction in the direct context of SIC technology it is necessary to separate the problems into two phases: design and production.

Design cost. In the case of design cost, the need is to increase the use of computers, not only to decrease costs directly by reducing the number of "engineer-hours" at the various design stages, but particularly to decrease the possibility of design errors attributable to human fallibility. This need for computer-aided design is particularly apparent when

1. Checking logic design for complex logic blocks.
2. Checking circuit design (e.g., carrying out tolerance calculations of circuit response). This step is becoming vital, since orthodox "breadboard" experiments become meaningless because of their inferior performance in terms of active device matching and parasitics.
3. Checking layout design artwork. Tape-controlled coordinate tables are rapidly replacing operators at the artwork stage. The full benefits, however, depend on having the facility to check the tape for errors, since this capability gives the most significant advantage over the use of a human operator.

Production cost. The production cost of an SIC chip depends on the cost of processing a silicon slice (regardless of quality) and the subsequent yield of "good" chips. The slice processing cost depends primarily on such factors as volume of throughput, fixed overheads, and production efficiency. For the purpose of this article these topics can be dismissed as "management problems."

Yield, on the other hand, is determined by the detailed problems that arise during slice processing (unless inadequate circuit design is asking more of a process than it reasonably should) and some of these problems will now be considered. The assumption is made that the mask quality, in terms of registration and localized defects, is sufficiently good for this not to be a first-order yield limitation.

Silicon crystal quality. The use of scanning X-ray topography (see, for example, Ref. 6) permits the nondestructive observation of certain crystallographic defects in silicon slices—defects either existing prior to processing or introduced during the various stages of oxidation, diffusion, and epitaxy.

A comparison between the pattern of defects in a silicon slice and the pattern of device yield indicates that in some circumstances poor device characteristics can be directly correlated with bulk defects in the silicon substrate material.[7] In applications requiring the use of the entire silicon slice (e.g., MOSAIC stores, silicon "vidicon," etc.), greater attention must be paid to the quality of crystal pulling and the specification of an acceptable quality of starting material.

Uniformity of resistivity in silicon substrate slices is adequate in bipolar technology, where the active devices are in an epitaxially grown layer. However, for MOSAICs, where the devices generally are formed in the "as-pulled" silicon, one would prefer tighter control on the resistivity and on the degree of impurity compensation. For this reason, among others, it is highly likely

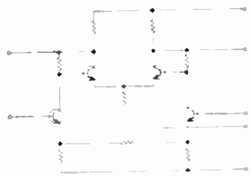
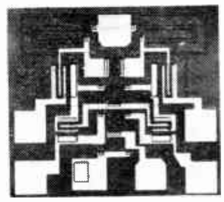
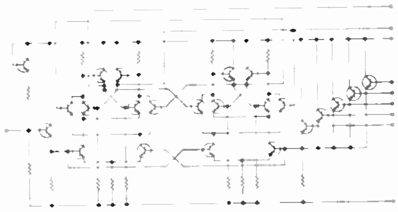
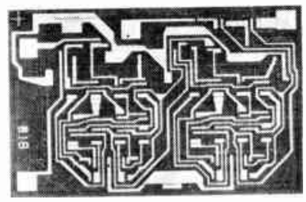
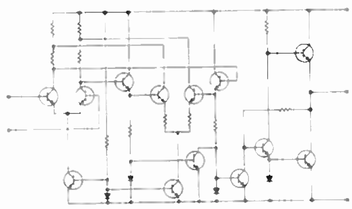
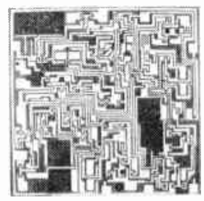
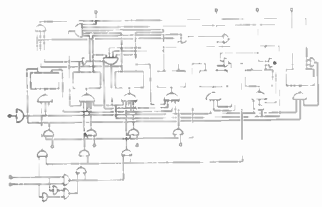
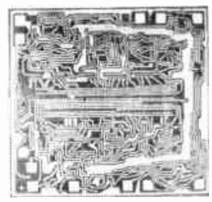
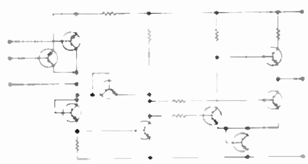
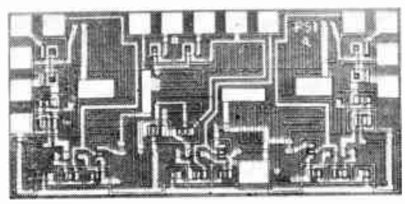
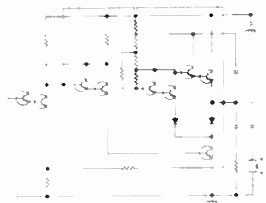
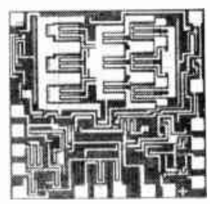
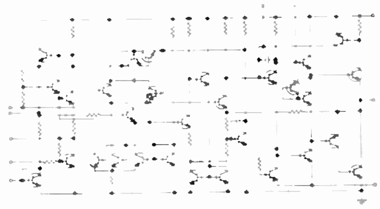
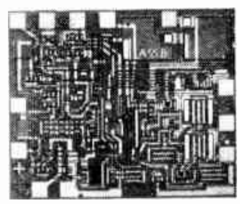
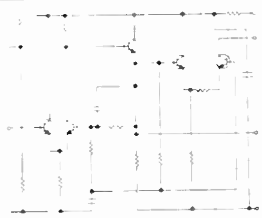
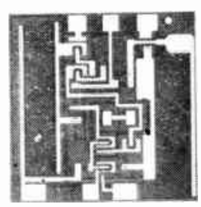
Example of	Circuit Diagram	Chip Photograph
Bandwidth-rise time (Rise time 1.5 ns. Bandwidth > 350 MHz.)		
High-speed digital circuit (Divide-by-four circuit plus output buffer. Consumes 80 mW at 200-MHz input frequency.)		
Linear complexity (Triple operational amplifier, each with Nyquist network included.)		
Digital complexity (Variable decade divider. 80 gates. 40-MHz max. binary count rate. 2 mW per gate.)		
Output voltage (28 V peak to peak with 50-ns edge times.)		
Output power and current (1.5 A, 3 W)		
Component matching (Operation of this AF amplifier depends on matching and tracking of 26 components.)		
Novel linear function (Log. amp. bandwidth of 170 MHz, with 1200 pF on the chip; limiting function with built-in low-threshold detector.)		

FIGURE 3. Examples of circuits made by bipolar Process I.

that future MOSAICs will be based on epitaxial material where better radial uniformity can be achieved and the control lies wholly with the SIC manufacturer.

Silicon surfaces. Ignoring localized contamination in the form of dust, which needs to be minimized at all process stages, the residual contamination on the silicon surface following such operations as photolithography, oxide window etching, or slice "cleaning" is a cause for much concern from the point of view of the long-term stability of surface-controlled device characteristics (for example, leakage currents, low-current h_{FE}, V_T, and V_T'). Apart from such obvious steps as simplifying cleaning schedules, choosing etchants and solvents intelligently, and using ultrapure reagents, there are two significant aspects that should be studied further. The first involves the measurement of surface cleanliness. One approach being pursued is the use of ellipsometry as an on-line monitor of silicon surfaces immediately after cleaning. The second is concerned with the avoidance of photoresist operations. In the long term one would like to see a technology whereby silicon slices could be selectively oxidized, diffused (several times), epitaxed, and provided with multilayer metalizing—all without ever being removed from a single clean chamber. Processes that might permit such a pipe dream to be achieved are ion-beam implantation and photolysis.

Oxidation. To date, SIC technology has been based almost entirely on the use of thermally grown silicon dioxide films both for diffusion masking and junction passivation. These films are far from ideal in terms of their fundamental properties (e.g., relative permeability to contaminants, such as sodium) and the effect that their growth has on the underlying silicon (e.g., boron outdiffusion and vacancy formation). In the near future we are likely to see much greater use of deposited layers such as silicon dioxide (by pyrolysis), silicon nitride (by RF discharge-induced chemical deposition), and alumina. Such techniques will give the device designer and the device technologist new degrees of freedom, exemplified by the use of a polycrystalline silicon gate electrode on a "nitride/oxide" sandwich to form an MOS transistor with reduced gate overlap.[8]

Multilayer metalization. There is a real need for a high-yield multilayer technology to be made available as a standard production process. This development will frequently assist designers in overcoming topological problems and will reduce costs by enabling smaller silicon chips to be used.

Control improvements

Insofar as control is concerned, the requirements are relatively simple: Improve the control of temperature, gas flow, gas mixture, etc., in order that epitaxial layer thickness and resistivity can be controlled to, say, ± 1 percent and similarly that the sheet resistance of diffused layers (particularly those used to form transistor base regions and resistors) shall also be within a ± 1 percent range. Such a development would be of considerable benefit to circuit designers and equipment manufacturers.

Performance improvements

Bipolar circuits. There are in principle two ways in which bipolar circuit performance can be improved by technology development. Increased process control will result for example in operational amplifiers with lower offsets. On the other hand, one more usually thinks of improvements in speed, frequency response, or voltage capability (current vs. power being a more straightforward economic trade-off of how much silicon area you can afford). Before developing a new SIC process to give such improvements, it may be profitable to note in Fig. 3 the performance of circuits already made by the use of Process I. It is worthwhile noting that at a time when it was being widely maintained that dielectric isolation was essential to achieve bandwidths greater than 30 MHz,[9,10] Process I was being used to manufacture and sell a 100-MHz amplifier (SL501); moreover, that same process now yields amplifiers with bandwidths of up to 350 MHz. The difference between this result and those of Refs. 9 and 10 would appear to lie in the field of circuit design and its close integration with technology in order to extract the maximum benefit from it. It is interesting to note in passing that the SL501 also included 1350 pF of junction capacitance on board the chip, using a technique discussed in Ref. 11, a fact that seems to have escaped the attention of many workers in the field who still suggest that such values are impossible.[12] The point of this digression is that it is all too easy for inadequate understanding to lead to the conclusion that expensive process changes are essential when in fact the relatively cheap process of thought would provide a superior solution.

Summarizing the available information on Process I, it would appear that a different process is required to be developed only for:

1. Linear circuits above 200–400 MHz.
2. Digital circuits above 100–200 MHz.
3. Output voltages greater than 15–30 volts.

Combining the first two of these needs leads to a requirement of transistors of higher f_T and lower C_{OB} without any increase in r_{bb}'. There is an incidental need to reduce isolation capacitance, C_{CI}, but this requirement follows automatically from the process changes necessary to achieve the better transistor characteristics (e.g., thinner epitaxial layer, shallower diffusions, smaller-area devices, tighter photoengraving clearances). The keen interest in size reduction is reflected in the activities of several laboratories in applying electron-beam technology to the fabrication of submicrometer-dimensioned devices. The use of dielectric isolation would certainly reduce C_{CI} but this in itself would give only a marginal advantage (say 30 percent higher speed or bandwidth). Such a process, for higher speed and frequency operation than Process I, is currently in its final stages of development. Known as Process II, it will use transistors with the following typical values: f_T ± 2.5 GHz, C_{OB} ± 0.4 pF, and C_{CI} ± 1.5 pF. It is intended primarily for linear and digital circuits in the range of 200 MHz to 1 GHz. However, there is another significant advantage of Process II. Once the teething troubles of dealing, for example, with shallower diffusions are sorted out, it will be cheaper than Process I in the context of OSI simply because the smaller sizes of devices will enable more circuit complexity to be achieved within a given area—and it is silicon area that costs money.

It is interesting to note that the development time scales of Processes I and II show a significant departure from transistor process development. The reason is that the frequent introduction of small technology improvements was avoided, since it was assumed that the loss of circuit design information would more than outweigh the apparent gain in device performance.

Conversely, the problem that causes most concern when contemplating the development of a higher-voltage process is that it will be more expensive, for the following reasons:

1. The use of higher-resistivity material makes it more difficult to achieve the same precision.

2. Surface-controlled device characteristics are more difficult to control with stability on higher-resistivity material. The techniques used with discrete transistors (e.g., diffused guard ring) is wasteful of silicon area and hence expensive as part of an SIC technology.

3. Higher voltages obviously mean wider depletion regions, thicker epitaxial layers, etc. This results in significant wastage of silicon area relative to Processes I or II.

As a result it is stressed that although higher-voltage processes are technically quite feasible, from the point of view of optimizing system and equipment cost the question should be asked "Is the high-voltage requirement absolutely essential?" In many cases the answer that will come from the intelligent juxtaposition of system, circuit, device, and process engineers will be to come up with an alternative, cheaper solution. The usual argument in favor of high-voltage devices is that they are needed at the periphery of equipment. However it is worth remembering that we can often trade expensive volts for less expensive amperes, and that as more and more "peripheral" devices go solid state (for example, electroluminescent diodes for data presentation) the present arbitrary interest in high voltage will diminish.

MOS devices. As is the case with bipolar devices, there are certain aspects of MOS circuit performance that can be improved by increased control over the SIC technology. Typical examples are reductions in the spread of V_T and V_T' and in the drift of V_T. These values, though now satisfactory for most logic circuits, are probably too high for uncorrected dc amplifiers and comparators.

The ways in which one would like to improve the performance of MOSAICs, and for which a new, improved technology is needed, are:

1. Provide bipolar transistors on the same chip in order to improve the economics of providing current to drive external capacitance. An MOS transistor is less efficient than a bipolar transistor, in terms of current per unit area (and hence milliamperes per dollar).

2. Improve the speed of MOSAICs. The incentive is obvious: the cost-per-function advantage of MOS over bipolar is presently limited to digital speeds in the region of 1 MHz with a present limit of approximately 5 MHz. A factor-of-10 improvement (which seems quite feasible) would lead to a significant effect on the role of saturating bipolar logic families for medium-speed logic. It seems reasonable to expect the bulk of the digital business over the next 5 to 10 years to become channeled into MOS and high-speed bipolar current steering, with saturating bipolar logic squeezed out. It is interesting to note that there are already some similarities of circuit concept in current steering and MOS logic, since both are primarily speed-limited by capacitance charging. In both, therefore, the advantage of logic performed with currents at a "virtual-ground" input can be exploited.

The objectives have been set; now the question is: "How should the technology react in order to achieve them?" First of all, with regard to the problem of compatible MOS/bipolar technology, it must be stressed that the process must be kept simple or it will destroy one of the main advantages of MOS technology. There is a great advantage, therefore, in avoiding the need for isolation technology (one of the disadvantages of complementary MOS circuits) where possible. Failing that, we should be able to use the same doping level in the MOS substrate and the n-p-n collector layer, in order to avoid the need for selective area epitaxy (i.e., etch and refill). This will most probably lead to a reduction in the resistivity of the MOS substrate, which will incidentally ease the problems of device parameter control and stability.

Speed improvement is dependent on such factors as

1. Improved photoengraving precision (or some other trick such as the silicon/nitride/oxide/silicon, or SNOS, device) to reduce the gate overlap contribution to the input capacitance.

2. Channel mobility more closely approaching bulk values, possibly by improved substrate surface preparation and/or annealing treatments.

3. Choice of optimum value of V_T. It may seem advantageous at first sight to reduce V_T in order to increase the speed. However, if one reduces $V_S - V_T$ (where V_S is the signal swing, usually 2 to 3 times V_T), the switched-on device will have a lower β for a given area. Hence it is apparent that some integrated-circuit device technology thinking is needed to determine the optimum value of V_T for economical higher-speed performance.

4. The simplest dielectric isolation technology. "Silicon on sapphire"[13,14] is promising as a vehicle for MOS technology because of the way in which vertical junctions can give significant reductions in source- and drain-substrate capacitances. Also, the capacitances from interconnections to substrate are reduced by SOS technology.

5. Circuit and system innovation. Exploitation of the MOS transistor as a circuit element is still really at the exploratory stage. Early ideas on dc and two-phase clocked logic are beginning to give way to various forms of four-phase clocked logic with no clearly defined power supplies, and it is fairly clear that we are still only at the beginning of a period of rapid advance in achieving optimum application of MOSAIC technology to real equipment needs.

The complexity of the technology required to fabricate p- and n-channel devices on the same chip, together with the loss of useful area that arises from the isolation technology, leads me to the conclusion that the complementary MOS transistor is not of great immediate significance. Its virtues of low standing power need to be compared with the performance currently being achieved with four-phase clocked p-channel MOSAICs.

Better understanding

A great deal of progress at the level of device fabrication and circuit design has been achieved by the semiconductor industry in spite of the fact that large areas of activity have been based on incomplete understanding of the phenomena, processes, and devices being used. In the opinion of the writer, the full benefits of SIC technology will be achieved only on the basis of improved understanding, at many levels, as discussed in the following paragraphs.

Silicon technology understanding. It is fairly obvious that more fundamental research is required in order to throw more light on such topics as

1. The role of bulk crystallographic defects in determining device behavior.

2. The nature and origin of "interface states" at the silicon–silicon dioxide interface and within silicon dioxide and other dielectrics.

3. The relationship between oxidation kinetics and "interface states."

4. The reaction kinetics of diffusion from sources such as BBr_3 and $POCl_3$ in the presence of silicon dioxide and/or silicon nitride masking layers.

5. The metallurgical and physical/electrical nature of ohmic contacts between, for example, aluminum and shallow diffused p^+ and n^+ layers of silicon.

Device design understanding. As device performance improves it becomes even more important to be able to predict the characteristics of a new design, rather than simply to make a quantity of items and accept what you get. In particular, it is increasingly necessary to have an adequate equivalent circuit for the active elements. It should be remembered that improvements in the technology can easily invalidate an accepted equivalent circuit and hence make it difficult to optimize the design of the device itself or of a circuit embodying it. One example of this situation is that as the gate-overlap capacitance of an MOS transistor is reduced (relative to the active gate–channel capacitance) it becomes more complex to take account of the actual input capacitance when performing transient analysis on a switching circuit.

Circuit design understanding. It is accepted that the prediction of circuit behavior prior to SIC fabrication must be based primarily on a paper study rather than a physical breadboard, simply because in many instances the breadboard performance is far inferior to that of the final SIC, by virtue of the large, poorly defined parasitics and inferior active-device matching. Incidentally, that is one reason why my colleagues and I have never been able to accept the enthusiasm of some organizations for multichip circuits as a means of making initial samples of new high-performance circuits. As a rule, the technique simply does not work.

It is also accepted that such paper studies can consume a great deal of engineering time and that the use of a computer to perform frequency-response and circuit-tolerance calculations, for example, is invaluable. However, it must be stressed that there is no point in feeding an open-ended question to a computer; it might have hysterics. Optimum use of computer-aided design can follow only from a full appreciation by the circuit designer of how his circuit works and just what the relevant technology-controlled component characteristics are, and how they are likely to vary with time, temperature, voltage, etc., and from batch to batch.

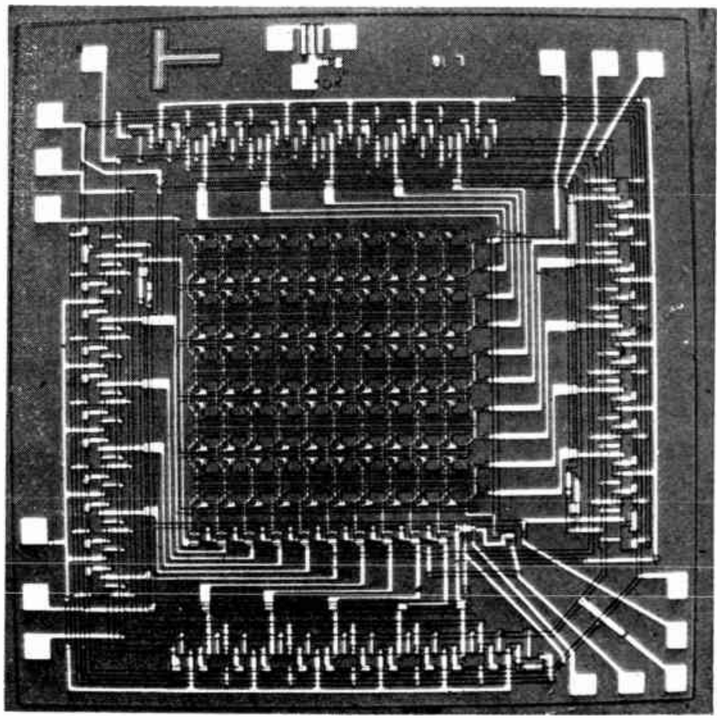
Total understanding. The need for further understanding of SIC technology at all levels—from materials chemistry and metallurgy through device physics to circuit and system engineering—cannot be overemphasized. It is the *totality* of understanding that needs most emphasis: research on silicon diffusion technology (for example) in isolation would most likely be a waste of time. Gold diffusion is a good example of this problem. Its improved control and understanding is a vital aspect of saturating bipolar logic circuits. The choice between improving that process or concentrating on current steering and faster MOSAICs is difficult and calls for interdisciplinary judgment.

Framework for the implementation of SIC technology development

The keynote of SIC development is *integration.* Development of silicon technology in a separate compartment from circuit design would not have led to the establishment of Process I with its emphasis on meeting the competent circuit designer's needs for reliable data on components made reproducibly as opposed to the less competent (in our view) approach whereby the technology is changed to optimize a particular device characteristic. The evidence in Fig. 3 gives ample justification for the "standard process, nonstandard circuit geometry" concept.

It follows that when the next stages of R & D on silicon technology are considered, the integration needs to go all the way to system design. As an example, let us consider the application of silicon technology to the modernization of sonar Doppler signal processing. Current nonmicroelectronic equipment uses an analog system with large numbers of *LC* filters in the receiver. Immediate advantages in size, weight, performance, and cost can be obtained by replacing *LC* filters by active *RC* filters based on the use of SIC operational amplifiers.[15] If one now considers the ways in which silicon technology

FIGURE 4. Integrated sensor array.



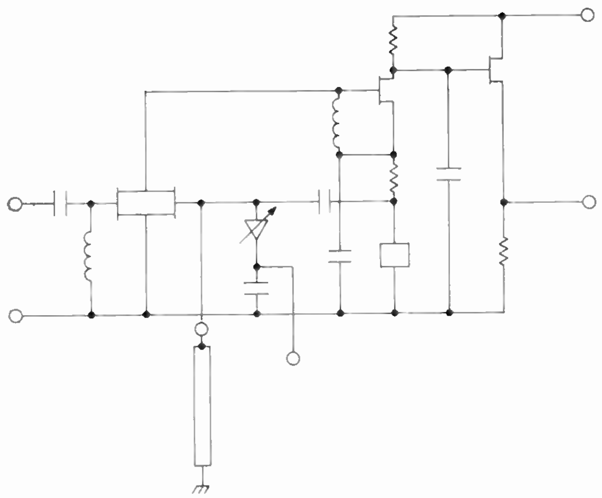


FIGURE 5. Schematic diagram of typical GaAs microwave integrated circuit.

should be influenced by that application, the answer is to put three fully Nyquist-stabilized operational amplifiers on one chip (see Fig. 3) and then to consider Process II to achieve lower cost and higher bandwidth-to-power ratio. An alternative approach is to digitize the signals and use MOSAIC technology to achieve economical digital cross-correlators for the subsequent signal detection. This approach puts demands on yield improvement (and speed improvement) of complex MOSAICs. Yet a third approach is to convert the electric signals into their optical equivalent, and then to make use of optical cross-correlation techniques.[16] The contribution of silicon technology in this case is to provide the integrated sensor arrays[17] required to detect the correlated output; see Fig. 4. This puts the usual demands on MOSAIC technology, together with additional demands on minority-carrier lifetime, leakage-current control, etc.

It is clear, therefore, that to judge independently how sonar system design should take advantage of silicon technology, and how silicon technology should react to the (ill-defined) needs of sonar systems, would be catastrophic. In fact, though, that is what usually happens: a system design is adopted and then the SIC manufacturer in the role of component manufacturer is introduced to the problem when it is often too late to contribute to its solution.

Further examples of this type of problem include

1. *Telephone systems.* The demands on SIC technology for the frequency-division multiplex speech crosspoint are utterly different for the pulse code modulation equivalent.

2. *Color television tubes.* One wonders whether the three-gun "shadow mask" tube, with its open-loop, very high mechanical precision technique, would have have been chosen vis-à-vis the beam index tube with cheap microelectronic feedback around the tube.

3. *Computer systems.* The need to re-evaluate the system breakdown into hardware and software in the context of cheap SIC hardware is generally recognized. Just as important is the need to re-examine the hardware design principles (for example, the replacement of lumped "stores" by a distributed concept, with logic and storage becoming less distinguishable—a particularly valid concept with MOSAICs).

Other material technologies

There seems little doubt that silicon will remain indefinitely the dominant microelectronic material. However, the next five years are going to see rapid extensions of microelectronics philosophy into the microwave region on the one hand, and optoelectronics on the other. It is interesting to speculate on the role of silicon in these areas in the face of competition from other materials.

Microwave integration. We are in the interesting situation wherein the development of monolithic microwave receivers (embodying an RF amplifier, local oscillator, mixer, IF amplifier, and the appropriate lumped *LC* filters) could be undertaken based on either silicon or gallium arsenide technology. However, the latter material possesses a greater potential frequency range. In Fig. 5 is shown the design (by M. J. Gay of the Allen Clark Research Centre) for such a GaAs microwave integrated circuit.

Optoelectronics. For the detection of visible radiation to approximately 10 000 Å, silicon junction photocells offer excellent performance (for example, quantum efficiencies of about 40 percent at 8500 Å) and relative ease of integration into MOSAICs in order to obtain the functions of photocurrent amplification and integration, and image scanning. Such a device is shown in Fig. 4. It comprises 10 × 10 photocells and 526 MOS transistors on a silicon chip 2.5 mm square. The fabrication of a 100 × 100 array on a chip approximately 1 cm square is currently envisaged. In principle, this type of device can be given an extended spectral coverage—for example, by replacing the silicon diffused-junction detector by deposited infrared-sensitive detectors.

Although arrays of gallium phosphide and gallium arsenide phosphide lamps of promising visual efficiency have been described,[18] the full potential of this device is likely to await the successful integration of the electroluminescent diodes with active elements for local storage and drive selection. One possibility is to use an array of silicon devices as the substrate for gallium phosphide epitaxy and subsequent lamp fabrication. Another is to use a gallium arsenide substrate (as is usually the case in present work) and form field-effect transistors on the gallium arsenide; these transistors could be of the metal-insulator-semiconductor variety[19] or junction-gate variety.[20]

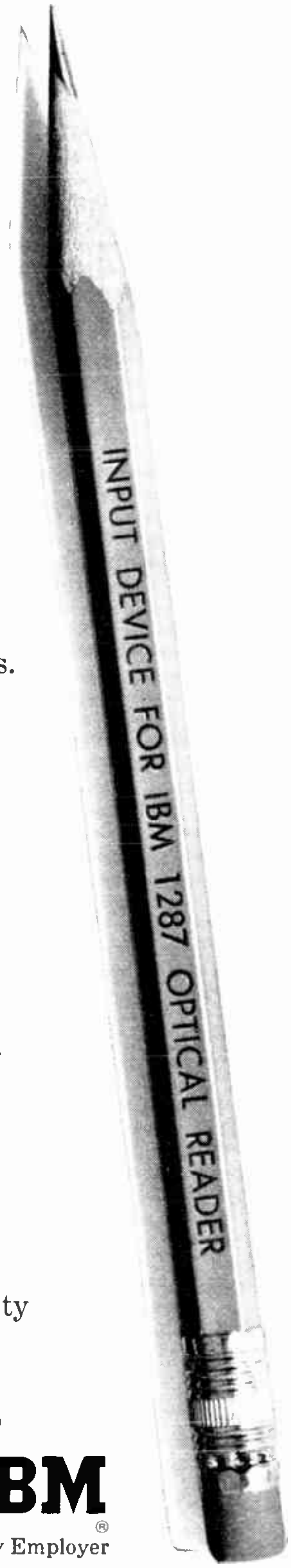
Conclusions

Further development of SIC technology is necessary, to give higher yields for complex OSI subsystems and improved performance in terms of frequency and speed beyond 200–400 MHz. The relevant research and development must be carried out within an overall materials-to-systems context if it is to be adequately motivated and controlled. Failure to do so will result in less than optimum use of this new equipment-constructing technology. In the opinion of the writer, the component–equipment interface is no longer relevant.

Although the views expressed in this article are his own, the author is indebted to his many colleagues at the Allen Clark Research Centre. Thanks are also extended to the directors of the Plessey Company Limited for their kind permission to publish.

Electronic Engineers

Can you design the circuits that make this pencil work?



If you can, IBM would like to hear from you.

We need electronic engineers for development work on sophisticated electromechanical applications. Including circuit design for input devices like our 1287 Optical Reader. This is the first mass-produced machine to read hand-printed numbers (which explains the boastful message on our pencils). It's a whole new way of talking to computers.

And it's only one of many exciting projects you could work on at IBM Rochester (Minnesota).

The outer limits

Join us, and you'll have a job that takes you to the outer limits of today's technology. You may, for example, get involved with sophisticated computer design techniques and new material applications.

And you'll be backed by specialists in mechanical design, metallurgy, plastics, computer-aided design, R&S, acoustics, dynamic analysis, electromagnetics, and all phases of manufacturing.

It's a rare opportunity to make a name for yourself in the development of data-processing devices.

An ideal place to live

These openings are in our development lab in Rochester, Minnesota.

Don't jump to conclusions about this town (pop. 50,000). It has many advantages you'd normally find only in a big city. The Mayo Clinic and a major IBM facility help to give it an unusually high percentage of professional residents.

The school system is excellent (75 percent of Rochester's high school graduates go on to college), and there's an abundance of cultural and recreational facilities. You can enjoy the Minneapolis Symphony in the winter, for example, and three uncrowded golf courses in the summer.

Since Rochester sits in the southern tip of the state, it offers a wide variety of sports all year round. Typical commuting time is 5 to 10 minutes.

Look into it. If you're the kind of engineer we're looking for, you could catalyze your career here. Send a resume or outline of your experience to Robert Lewkowicz, IBM Corporation, Dept. QC7-O, Highway 52 and Northwest 37th Street, Rochester, Minnesota 55901.

IBM®

An Equal Opportunity Employer

Advance abstracts

The IEEE publications listed and abstracted below will be available shortly. Single copies may be ordered from IEEE, 345 East 47 Street, New York, N.Y. 10017. Prices are listed with the abstracts of each publication; libraries and nonmembers outside the United States and Canada should add $0.50. (M—Members; L—Libraries; NM—Nonmembers.)

Copies of individual articles are not available from IEEE but may be purchased from the Engineering Societies Library at the foregoing address.

Proceedings of the IEEE
IEEE Journal of Quantum Electronics
IEEE Journal of Solid-State Circuits
IEEE Transactions on

- Aerospace and Electronic Systems
- Automatic Control
- Communication Technology
- Computers
- Electrical Insulation
- Electromagnetic Compatibility
- Electron Devices
- Information Theory
- Instrumentation and Measurement
- Systems Science and Cybernetics

Proceedings of the IEEE

Vol. 56, no. 2, February 1967
M—$2.50; L—$3.75; NM—$5.00

Terminals for a High-Speed Optical Pulse Code Modulation Communication System: I. 224-Mbit/s Single Channel, *R. T. Denton, T. S. Kinsel*—The design of an experimental single-channel 224-Mbit/s optical pulse code modulation terminal is described and data are presented that have been obtained with such a terminal using a helium-neon laser operating at 6328 Å. The basic element is an optical gate using lithium tantalate whose design and operating characteristics are described.

Terminals for a High-Speed Optical Pulse Code Modulation Communication System: II. Optical Multiplexing and Demultiplexing, *T. S. Kinsel, R. T. Denton*—Techniques are described for optically multiplexing and demultiplexing individual pulse code modulation channels in order to develop the terminal capability for a high-capacity optical communication system. It is shown that, using these techniques, an information capacity in excess of 10^{10} bit/s can be achieved on the output beam of a single laser.

The Spectrum of Microwave Emission from InSb, *B. Ancker-Johnson*—Conflicting descriptions in the literature about the frequency dependency of microwave emission from InSb are shown to be the result of faulty experimental techniques. The amplitude as a function of frequency is apparently not periodic nor is frequency bunching confirmed. Several authors have claimed that the amplitude decreases with increasing frequency and this is confirmed.

An Electron-Beam-Activated Switch and Associated Memory, *N. C. MacDonald, T. E. Everhart*—A new type of electron-beam-activated switch (EBAS) is described that utilizes electron-beam-induced charge storage in the metal-oxide-semiconductor system. The state of the EBAS is determined by monitoring the surface conductance of the semiconductor. The basic charge storage phenomena are discussed; memory arrays that use the electron beam for storing and reading information are described. A matrix array of EBAS's in which information is stored using the electron beam and read by row-column access circuits is discussed in detail. The time to store a bit of information is a function of the current density of the electron beam; an approximate dosage of 10^{-5} C/cm² is required for storage. A memory design using Schlesinger's microspot tube for the electron optics is discussed. It is shown that storage of 1.0×10^7 bits per tube should be possible with presently available electron optical design and semiconductor technology.

Charge Storage Frequency Multipliers, *R. H. Johnston, A. R. Boothroyd*—Frequency multiplication by means of a nonlinear charge storage element is investigated. This element is assumed to have an abrupt transition from an infinite to a zero capacitance. An analysis of a frequency multiplier circuit utilizing such an element is carried out with certain limitations imposed on the mode of operation in order to make possible an algebraic solution, namely, that (1) only a fundamental and one harmonic current be applied to the nonlinear element, (2) only particular conduction angles be permitted, and (3) the fundamental and harmonic be subject to a particular phase relationship. These limitations permit an algebraic solution for circuit performance for any multiplication factor and for a range of conduction angles. The analysis yields directly the input and output resistances of the multiplier, in terms of which the conversion efficiency and power-handling capability are derived.

The Theoretical Analysis of Data Compression Systems, *L. D. Davisson*—The concept of reducing the required transmission rate for a given system through prediction, interpolation, or other such techniques loosely labeled as "data compression" is now well known. The problems in analyzing such systems by theoretical means are formidable in even the simplest situations due to the inherent nonlinear nature of the operations performed. These difficulties are discussed, some approximate and exact solutions presented, and areas suggested where further work is needed.

Proceedings Letters

Because letters are published in PROCEEDINGS as soon as possible after receipt, necessitating a late closing date, we are unable to list here the letters in the February issue. This will appear in the next issue of SPECTRUM. Listed below are the letters from vol. 56, no. 1, January 1968.

Electromagnetics and Plasmas

Effect of Neglecting Hall Term on Dispersion Relation for Hydromagnetic Waves, *M. Abbas*

Comment on "Energy-Transport Velocity in Electromagnetic Waves," *R. F. Adrion, D. V. Geppert*

Antenna Array Excitation for Maximum Gain, *J. K. Butler*

The Wave Nature of Megagauss Field Production, *F. J. Young*

On the Behavior of Electromagnetic Horns, *E. V. Jull*

Electromagnetic Scattering by Two Spheres, *J. H. Bruning, Y. T. Lo*

Two-Dimensional Electromagnetic Field Problems Specified in Terms of a Vector Admittance Function, *I. V. Lindell*

Circuit and System Theory

Maximally Flat Approximation Techniques, *H. J. Orchard, G. C. Temes*

On the Compensation of the Nonideal NIC, *C.-K. Kuo, K. L. Su*

A Discussion of an Auxiliary Matrix, *D. C. Fielder*

Simultaneous Oscillations at Two Frequencies in *RLC* Circuits, *I. Bruyland*

Further Comments on "Star-Delta Transformation: A Second Solution (Also Yielding a Star-Star and Delta-Delta Equivalence)," *M. F. Moad*

Comments on Phase Intercept Distortion, *V. C. Vannicola*

An Extension of the Definition of Power Factor to Bounded Periodic Nonsinusoidal Waves, *L. Unger*

Maximum Flow in a Communication Network, *V. V. B. Rao, K. S. Rao, P. Sankaran, V. G. K. Murti*

Electronic Circuits and Design

Uniform Distributed Amplifier Analysis with Fast and Slow Waves, *W. Jutzi*

Immittance Transformation Chart of a Two-Port, *K. Hirano, S. Kanema*

Intermodulation Products Generated by a p-i-n Diode Switch, *R. L. Sicotte, R. N. Assaly*

A Fast Controlled Monostable Circuit Using Tunnel Diodes, *Z. H. Cho*

Monostable Behavior of an *LR* Loaded Tunnel Diode with Two Stable States, *M. A. Schapper*

Intermodulation Analysis and Design of a Schottky Barrier Diode Mixer, *H. J. Peppiatt, A. V. McDaniel, Jr.*

A DC Reference Voltage with Very High Rejection of Supply Variation, *P. Williams*

Some Observations on Dual Input Null Networks, *M. N. S. Swamy*

Electronic Devices

Rectifying Contacts Under Evaporated CdS, *J. A. Scott-Monck, A. J. Learn*

Calculation of LSA Oscillator Noise, *K. Matsuno*

High-Current Triode with Nonintercepting Control Electrode, *C. Süsskind*

Characteristics of Planar Transistors Under Localized Compressive Stresses, *A. Ohwada*

Turn-On Delay Time of MOS Transistors, *D. W. Peters*

Noise Reduction in Crossed-Field Guns by Cathode Tilt, *M. L. Sisodia, R. P. Wadhwa*

Parametric Amplifier Nonresonant Gain Maximum, *P. J. Khan*

A Method for Measuring Collector Series Resistance of an Integrated Circuit from Module Pins, *W. W. Wu*

Continuous-Wave Planar Avalanche Diode with Restricted Depletion Layer, *H. G. Kock, D. de Nobel, M. T. Vlaardingerbroek, P. J. de Waard*

An Improved Avalanche-Injection Transistor, *G. A. May*

Noise of Gunn-Effect Oscillator, *K. Matsuno* summer. According to a recent announcement, they will include the following.

"Integrated Circuits," to cover the theory, design, and fabrication of solid-state devices and integrated circuits, will be held May 27–June 1; fee, $300.

"PCM Telemetry Systems," covering theoretical and engineering state-of-the-art telemetry systems for transmitting analog and digital data over noisy digital channels, will also be held May 27–June 1; fee, $200.

"Modern Automatic Control," for those working in the fields of guidance, estimation, and control, will be given May 27–June 7; fee, $150 for one week or $300 for two weeks.

"Computational Methods for Power Systems Analysis," designed for those concerned with the planning, design, and operation of electric power systems, is being offered June 3–21; fee, $200 for one week, $320 for two weeks, $420 for three weeks.

"Computer Design and Cybernetics," designed to introduce the analysis and design of digital computers, will be held June 24–29; fee, $150.

"Theory and Applications of Information Processing" will present the foundations of the theory for modern information processing on June 24–29; fee, $175.

Further details and registration forms for the courses described may be obtained from the Conference Division, Memorial Center, Room 116, Purdue University, Lafayette, Ind. 47907.

Columbia offers graduate interdepartmental programs

"Solid-state science and engineering" and "plasma physics" are the subjects of two interdepartmental graduate programs recently announced by the School of Engineering and Applied Science at Columbia University.

The program in solid-state science encompasses the study of the electrical, optical, magnetic, and mechanical properties of solids and their applications. At the school's laboratories, research equipment and low-temperature facilities are available for the study of optical absorption, internal friction, ultrasonic attenuation, Mossbauer resonance, dielectric properties, Hall effect, neutron spectroscopy, and field ion microscopy.

Students may enroll in this program for M.S., M.Eng., D.Eng.Sc., and Ph.D. degrees. Fellowships, traineeships, and assistantships are available on a competitive basis. For additional information write to Prof. Arthur Nowick, Chairman, Solid State Interdepartmental Committee, 1146 S.W. Mudd, Columbia University, New York, N.Y. 10027.

The program in plasma physics leads to the M.S., Ph.D., or Eng.Sc.D. degrees; however, the emphasis is on doctoral training. The academic program includes extensive experimental and theoretical research in the Plasma Research Laboratory, where such equipment as a high-energy electromagnetic shock tube and an alkali plasma *Q* machine are available.

Fellowships, traineeships, and assistantships are available, on a competitive basis, for students who wish to pursue doctoral studies in plasma physics. Requests for further information should be addressed to Prof. Robert A. Gross, 236 S.W. Mudd, Columbia University, New York, N.Y. 10027.

Midland, Mich., will be site of privately owned nuclear plant

Plans for the construction of a privately owned dual-purpose nuclear plant have been announced by Consumers Power Company. The facility, to cost $267 million, will be built south of Midland, Mich.

The twin-reactor plant will generate 1.3 million kW of electricity. Additionally, it will deliver four million pounds (approximately 175 000 kilograms) of steam per hour for industrial use by The Dow Chemical Company in Midland. Total energy capacity is equivalent to about 1.5 million kWe.

The installation will use two water-cooled nuclear reactors as its source of heat for generation of steam and power. Plans call for the first reactor to become operational in 1974 and the second unit in 1975. The plant will be totally owned and operated by Consumers Power and its output will be fed into the Michigan Power Pool.

Georgia Tech announces continuing education program

The Georgia Institute of Technology, as part of its program of continuing education, has scheduled the following events, to be held this spring.

"Management for Engineers," a short course designed to improve the manager's ability in planning, organizing, controlling, and innovating, as well as giving him a more comprehensive understanding of his role in the operations of his organization, will be held April 8–12; fee, $150.

"The Chief Industrial Engineer's Seminar," with the objective of enabling the participant to improve the effectiveness of the function he supervises, will take place April 15–19; fee, $300.

"Traffic Engineering," intended for those who are responsible for traffic engineering and who have had limited training in the field, will be given April 29–May 3; fee, $125.

"Creativity in Engineering," a course with the aim of providing the registrant with a more deliberate approach to solving problems found in industry, is scheduled for May 6–10; fee, $150.

"Impending Technology—Its Challenge to Livable Cities" will be the subject of a conference to be held May 7–9; fee, $35.

Additional information and final announcements are available from the Director, Department of Continuing Education, Georgia Institute of Technology, Atlanta, Ga. 30332.

M.I.T. will offer image processing course

An intensive two-week course in image processing will be offered by the Massachusetts Institute of Technology during August 5–16. The course will cover the theory and techniques of optical and computer image processing. Topics from image enhancement and biomedical image analysis will be discussed.

Further information about the course can be obtained from the Director of the Summer Session, Room E19-356, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

Vibration course will be held in St. Louis

An intensive five-day course in vibration and shock testing will be held at the Albert Pick Cavalier Motel in St. Louis, Mo., April 22–26. The principal instructor will be Wayne Tustin, president of the Tustin Institute of Technology, Inc.

The course will cover fundamentals of vibration, measurement and analysis of vibration, sinusoidal and random vibration, and shock testing. The fee is $200. Further information may be obtained from the Tustin Institute, P.O. Box Q, Santa Barbara, Calif. 93102.

Technical correspondence

Basic comments

Your editorial in November's IEEE SPECTRUM makes a very important point—that those who dispense fiscal support for research are sometimes too much obsessed with military goals, and too little concerned with the quality of the civilization they believe they are defending or with less destructive ways of defending it. One sometimes gets the impression that military defense is in itself the primary goal; this is necessarily the case with unrepresentative governments, but it is unworthy of such countries as the United States.

There is, however, one small exception I take to your editorial, and that is in the phrase "impoverished minorities." In fact, the impoverished constitute a majority of the world's population. Being a nonnational organization, the IEEE can be presumed upon to always take a nonnational view, except when singling out a specific area or country by name. But you are right about the importance of making the poor self-sufficient and giving them some hope of attaining the better things of life, which modern communications make them fully aware of. Without hope, they are literally desperate, and there is no telling where their desperation may lead.

Keep up the good work!

Nelson M. Blachman
Sylvania Electronic Systems
Mountain View, Calif.

I agree wholeheartedly with your sentiments.

Note, however, that I used minorities in the plural. The majority of the world's population are members of some minority!

Were the hungry a minority, we would not really have a problem feeding them!

C. C. Cutler
Editor

While I basically agree with your editorial in the November 1967 issue of IEEE SPECTRUM on supporting basic research, I would like to make several comments.

In the second half of your editorial you say "... our political representatives should be looking for means whereby science can be supported as *science*, rather than as a weapon system or as a means to outshine communists ..." These sentiments however, are not really supported by your opening remarks in which you state "How else will we feed the expanding multitudes, replenish depleting resources, clean a polluted environment, and protect our children from a barbaric enemy?"

In other words, you were really implying that while we can say now that we support basic research for its own sake, we really know that in the future it will pay off in some material end. It is not saying that science should be supported for science's sake, even if all the future information that is gained doesn't help us one whit materially. However, it is only with this approach that we can expect basic research not to be hampered by present-day controls.

Second, in the second quoted statement above you use the phrase "...and protect our children from a barbaric enemy?" What does this mean? Isn't this pandering to popular emotions? Don't we have enough of a militaristic, warlike mentality already? (Especially in the electronics field, I might add.) I don't think we should put warlike statements together with positive goals because the end result is not positive. The statement (in the form of a question) only goes along with the all too popular conclusion that spending 60 billion dollars a year on defense (in which the electronics industry gleefully reaps benefits) is socially desirable.

I would be glad to hear any comments you have concerning this letter.

Franklin Cohen
Brooklyn, N.Y.

It might seem inconsistent to say in one breath that science should be supported for its own sake, and in another to point out the material benefits to be derived therefrom. Happily, many scientists do find a purer motivation than the promise of profits or fame. However, it is not an inconsequential fact that the indirect benefits of science are very real and provide additional force to the assertion that science should receive generous support. It is not evil to expect, and to watch for, the benefits of scientific research nor to be motivated by the promise of useful results.

What we deplore is the attempt to define and limit the goals of research; attempts to disguise research projects as weapon systems in order to ruin support, or weapon systems as science in order to win commendation. This is a dishonesty that is fostered by present budgetary and political methods.

Is it bad to define the benefits of science in terms of military protection? As long as international relations are such that force of arms is used, science will be an important factor in meeting or negating it. Let us hope that it will be a means of illuminating the problem.

C. C. Cutler
Editor

Words on engineering and art

C. C. Cutler's editorial projectile in December's IEEE SPECTRUM approaches his target and then veers off without transmitting any information about it. That's the way it usually is with engineers approaching art, as well as with artists approaching engineering.

Agreed, engineers and artists have basically the same elements in their motivations: both strive to "make it work" with every tool at their disposal, but this means different things to the two disciplines. "Economy" too is important to both, but again with different meanings. The same applies to "inventiveness" or "creativity." In addition, talented individuals from both areas often prefer to earn their daily bread by activities subjectively called "less impure" or "less commercial" than those forced on them. This is because their personal objectives may not relate very well to what society will pay for their work, despite the stomach growlings and the expensive educational wants for one's children. The real product of the work of each is "communication," although what is communicated, the receivers' reactions, and the codes of transmission are different. Both are usually alike in often being unclear in their own minds about this being a key aim.

Although these semantic divergences cannot be detailed in a letter, the most important matter is recognizing their existence, rather than using words loosely and increasing the confusion.

Cutler's examples (bridges, planes, ICs, thyratron glow, breadboard junk art) are to the point because they il-

lustrate the engineer's lack of understanding of these distinctions in meanings and objectives. If an artist, painting a nonobjective work, by chance produced a drawing Cutler immediately recognized as looking a lot like a magnified, type 709 op amp, would he call the artist an "engineer"? It is not a bridge's, nor a bridge designer's, function to be dramatic, even though large structures, man-made or natural, produce a dramatic reaction in the beholder.

It is true, however, that, as Cutler says, the artist should always, and indeed must, take advantage of the new tools provided by technology. But it is true of the artist only in the same way and degree that it is true of everyone in a society. Marshall McLuhan (in *Understanding Media* and *The Medium Is the Massage* in particular), who crosses the fields of sociology and art, makes it clear why society must utilize its technological tools—because these tools are its means of cohesion, its tools for communication.

Another kind of blurring in the editorial is not emphasized, but runs just under the surface. Like labor unions, technical societies tend to avoid emphasizing capability differences between members. There is sufficient reason for this, but it should be an aware blurring, not a fumbling one. "Amateur painters, musicians, and sculptors, even without training" are to professional artists what amateur engineers are to professionals. How does Cutler feel about amateur engineers who design bridges, airplanes, cars, etc.?

Finally, although I agree that excellence should be matched with competence and "hope that some can successfully bridge the gap" (producing some thousands or more modern da Vinci's, which, I gather, is the point of the editorial), I think it unlikely that anything significant can be done. Our engineering works *are* less esthetic because they are useful. Economic trade-offs determine esthetic media (which are technological tools); but, in most cases, economic trade-offs do *not* determine esthetic content, and do *not* determine what is communicated. Nobility in art is not its engineering content, but is closer to its opposite, purity from trade-offs.

G. F. Quittner
Cleveland Heights, Ohio

What target?

C. C. Cutler
Editor

A power-full argument

Erdelyi and Barnes in a recent article[1] suggest that electric power utilities support the education of engineers they will eventually hire. Dwon,[2] representing one of the most progressive power companies, American Electric Power, replies, "AEP will not support the concept as stated." In other words, "No."

That electric power companies are reluctant to support education and research is no surprise. In the modern technical world electric power systems are considered by most electrical engineers to be in a sort of technological stone age. Research that must be done before transmission can go underground is not being done. Our green hills are blanketed more and more every day by bigger and uglier outdoor transmission lines. Simple and obvious means for increasing the efficiency of power generation, for decreasing distribution losses, for storing energy in off-peak periods and making use of waste heat, are being inadequately investigated, if at all.

These developments might have far-reaching consequences. For example, they could make heating with heat pumps cheaper than heating with fossil fuels. Using electricity obtained from nuclear power stations, clean heat from heat pumps could substantially reduce smog and curb the alarming build-up of atmospheric carbon dioxide.

Why do power companies spend only a miniscule amount of their income on research, as compared with the 2 percent of gross[3] typical of most American industries ($300 million per year for the electric utilities)? Is their management more greedy or short-sighted? I do not think so. The principal reason, in my opinion, is that they lack incentive. Electric power companies are monopolies. Public commissions set rates. To oversimplify somewhat, these rates are adjusted so that, regardless of circumstances, power companies make a healthy profit. Why should anybody innovate in such a situation? If equipment is inefficient, it is the consumer who pays, not the power company. If, through gifted and daring management, the cost of electricity is reduced, it is the consumer who profits, not the company. Without new and powerful incentives, the status quo is sure to remain.

It is clear that new rate-setting formulas should have been adopted long ago by the rate-setting commissions. In essence, if power companies substantially decrease the cost of power, they should be allowed temporarily high profits; if they do not, their profits should be drastically reduced.

With such an incentive, the power companies would have direct motivation to do what every engineer knows they should do, set up a centralized laboratory similar to Bell Telephone Laboratories. This lab could consolidate the gargantuan and expensive equipment needed for research in high-power apparatus. It could undertake comprehensive research projects aimed at reducing power costs and putting transmission underground. It could support education by providing a place where research-minded Ph.D.-level students could work after completing their schooling. It could underwrite dissertation-type research done by these students prior to their graduation.

Without incentives, such a laboratory will never be formed. Thus, at the root, it is the rate-setting commissions, not the power-system executives, who must take the blame for the apathy and inaction that now distinguishes research and education in electric power.

Richard McFee
Syracuse University
Syracuse, N.Y.

REFERENCES

1. Erdelyi, E. A., and Barnes, F. S., "Co-operation of universities and utilities for the education of power-system engineers," *IEEE Spectrum*, vol. 4, pp. 71–76, Dec. 1967.
2. Dwon, L., *IEEE Spectrum*, vol. 4, pp. 79–81, Dec. 1967.
3. Furnas, C. C., *Research in Industry*. Princeton, N.J.: Van Nostrand, p. 151.