ROBERT MATKINS

VOLUME 1 ☐ NUMBER 6 JUNE 1964

IEEE spectrum

features

departments

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE spectrum

the cover

The illustration on the front cover shows a section of an experimental wafer on which several hundred TTL gates are interconnected to perform the function of a multibit shift register. This wafer is representative of the current developmental work in the field of integrated circuits. It evolved as part of a Westinghouse research project calling for experiments with large arrays to determine how complex a logic function may be fabricated on a single slice of silicon. A comprehensive report on integrated circuits begins on page 62 of this issue. It is an edited version of the Modular Magic Symposium held during the 1964 IEEE International Convention. The two leading types of circuits that show promise for integrated-circuit use—monolithic silicon circuits and thin-film circuits—are treated in detail in separate articles in this issue beginning on page 83 and on page 102, respectively.

IEEE SPECTRUM is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: Box A, Lenox Hill Station, New York, N.Y. 10021. Cable address: ITRIPLEE. Telephone: 212—PL 2-6800. Published at 20th and Northampton Sts., Easton, Pa. **Change of address** requires thirty days notice to IEEE Headquarters. **Annual subscription:** IEEE members, first subscription included in dues, one additional subscription $13.50. College and public libraries, $13.50 in U.S.A. and Canada, $14.50 elsewhere. Others, $18.00 in U.S.A. and Canada, $19.00 elsewhere. Single copies: members, one additional copy $1.00, college and public libraries, $1.50 in U.S.A. and Canada, $1.75 elsewhere. Others, $2.00 in U.S.A. and Canada, $2.25 elsewhere. **Editorial correspondence** should be addressed to IEEE SPECTRUM at IEEE Headquarters. **Advertising correspondence** should be addressed to IEEE Advertising Department. 72 W. 45 St., New York, N.Y. 10036. Telephone: 212—MU 2-6606.

The Alaskan earthquake has focused the attention of the world on seismic phenomena. The electronics engineer naturally asks, "Can we use the techniques of radar and seismic sounding to investigate the earth's interior?" Burt Rudman, consultant in AIL's Research and Systems Engineering Division, has been thinking about this problem for some time and we present his thoughts in this advertisement.

Seismic Sounding of the Earth's Interior

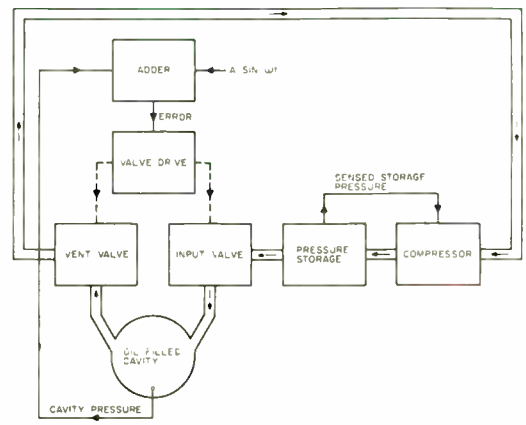
Man's knowledge of the interior structure of the earth comes from astronomical observations and studies of the effects of this interior structure on the paths of seismic energy. In most instances the seismic wave has been generated by an earthquake. Because he has no control over the location, timing, or magnitude of this energy source, the seismologist must accept amplitude measurements at distances where seismographic stations happen to be located—taking into account differences in energy, depth, and radiation patterns and paths.

Although seismological studies using large-scale explosions (about 20 kilotons) have overcome some of these problems, they do not present an ideal solution. This article explores the possibility of using injected coherent energy for seismic sounding of the earth's interior. Because this input signal is known and coherent, it would permit long-term data processing of the detected output signal (a method not feasible in examining data from earthquakes or explosions). This energy could be detected at almost any point on the earth's surface, including a location diametrically opposite the transmitting source. The received signal-to-noise ratio could be increased by optimizing the location of the detectors or by data processing.

To generate a seismic signal by controlled and coherent applications of pressure, consider a spherical cavity excavated at an overburden depth great enough to withstand the required overpressures. The depth will set the upper limit of allowable pressure, and the cavity size will influence the lower limit of required pressure. If this cavity is filled with hydraulic oil, which is subjected to sine-variant pressure, the pressure function will be transferred to the cavity walls and hence to a propagating seismic wave.

A pressure source maintained by a compressor is connected to the cavity through a continuous-control valve-vent system. A pressure sensor in the cavity provides loop control for the valve-vent drive. A sine-wave signal injected into the loop acts as an input source. The pressure in such a system should probably be limited to 200 atmospheres (paralleling current hydraulic-control limitations). At this pressure, overburden requirements dictate a depth of 940 meters or more.

The frequency at which a cavity is resonant is equal to the ratio of the velocity of the compressional wave in the surrounding medium to the radius of the cavity. A 1-cps wave (the frequency best propagated by earthquakes over the path of present interest) would require a cavity of 6-km radius for resonance.



Seismic Wave Generator

A pulse length of 10 seconds at a 1-cps carrier frequency represents a reasonable compromise for the demands of average power, signal-to-noise ratio, and system space resolution.

The bandwidth of the system is 0.075 cps. Since the signal-to-noise criterion of the system was determined from the microseismic noise in a 1-cps band, this bandwidth represents an improvement in the signal-to-noise energy ratio of 40:3; the system's peak-power requirement (869,000 kw for global range) is reduced by the same ratio to 65,000 kw. If the interpulse interval is 1 hour (to avoid intermode ambiguity), the transmitting duty cycle is 1/360. Considerations of bandwidth and duty cycle reduce the requirements to 362 kw of average power at an assumed efficiency of 50 percent.

The system illustrated, with its small duty cycle, does not present any problem for the average power. However, the peak power at this point is 130,000 kw (65,000 kw into the earth), requiring a pressure of 1450 atmospheres for a 1-meter cavity radius. A cavity radius of 7.25 meters would reduce pressure requirements to the hydraulic system limit of 200 atmospheres.

For a receiver at near-surface depths, 50 hours of transmission (10 seconds during each hour) would be required to keep the size of the transmitting cavity at 1 meter and its depth at 1 km. (The actual processing time could, of course, be short—working from tape.) At a receiver depth of about 1900 meters in sandstone, an amplitude signal-to-noise ratio of 7.25 would be achieved, thereby lessening the need for data processing.

We have idealized the problem in omitting the losses due to coupling at the transmitter. For a 7.25-meter cavity, the coupling loss would be on the order of 10^4 but since this loss is an inverse function of cavity size to the fourth power, a 72.5-meter cavity would compensate for it.

A decrease in the range requirement and/or an increase in processing time would further reduce the power required. These requirements could be satisfied by many combinations of receiver/transmitter depth, transmitting cavity size, and processing time. Many assumptions have been made and much remains to be proved—but the system of injecting coherent seismic energy into the earth seems feasible.

A complete bound set of our eighth series of articles is available on request. Write to Harold Hechtman at AIL for your set.

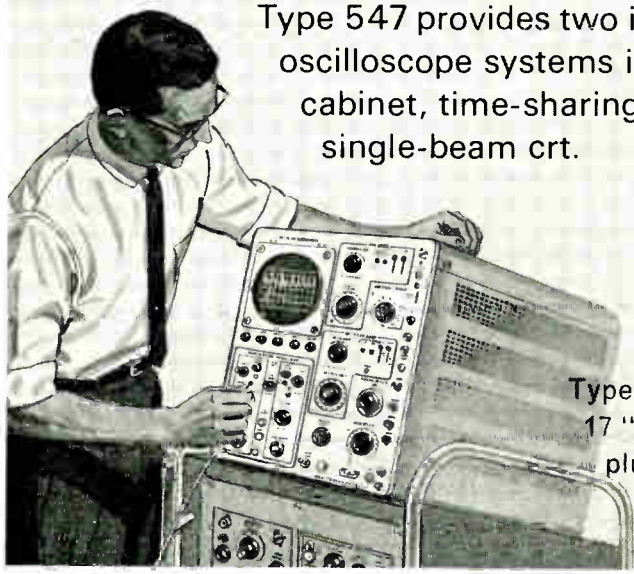
Tektronix oscilloscope displays both time-bases separately or alternately

NEW TYPE 547 and 1A1 UNIT

DUAL TRACE — **DC-to-50 MC 50 MV/CM**
DC-TO-28 MC, 5 MV/CM

SINGLE TRACE — **2 CPS-to-15 MC 500 μV/CM**
(CHANNELS 1 AND 2 CASCADED)

With automatic display switching, the Type 547 provides two independent oscilloscope systems in one cabinet, time-sharing a single-beam crt.



Type 547 also uses 17 "letter-series" plug-in units

Some Type 547 1A1 Unit Features

New CRT (with internal graticule and controllable illumination) provides bright "no-parallax" displays of small spot size and uniform focus over the full 8-cm by 10-cm viewing area.

Calibrated Sweep Delay extends continuously from 0.1 microsecond to 50 seconds.

2 Independent Sweep Systems provide 24 calibrated time-base rates from 5 sec/cm to 0.1 μsec/cm. Three magnified positions of 2X, 5X, and 10X are common to both sweeps—with the 10X magnifier increasing the maximum calibrated sweep rates to 10 nsec/cm.

Single Sweep Operation enables one-shot displays for photography of *either* normal or delayed sweeps, including alternate presentations.

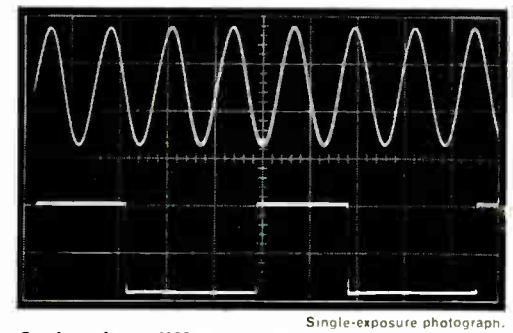
2 Independent Triggering Systems simplify set-up procedures, provide stable displays over the full passband and to beyond 50 MC, and include brightline automatic modes for convenience.

Type 547 Oscilloscope $1875
(without plug-in unit)

Type 1A1 Dual-Trace Unit . . . . . $ 600

Rack-Mount Model Type RM547 . . . $1975

U.S. Sales Prices f.o.b. Beaverton, Oregon



Single-exposure photograph.

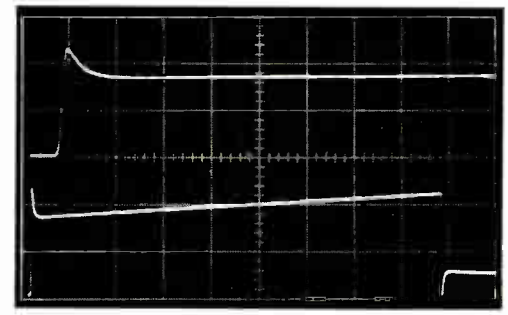
2 signals — different sweeps

Upper trace is Channel 1/A sweep, 1 μsec/cm.
Lower trace is Channel 2/B sweep, 10 μsec/cm.

Using same or different sweep rates (and sensitivities) to alternately display different signals provides equivalent dual-scope operation, in many instances.

Triggering internally (normal) permits viewing stable displays of waveforms unrelated in frequency.

Triggering internally (plug-in, Channel 1) permits viewing frequency or phase differences with respect to Channel 1.

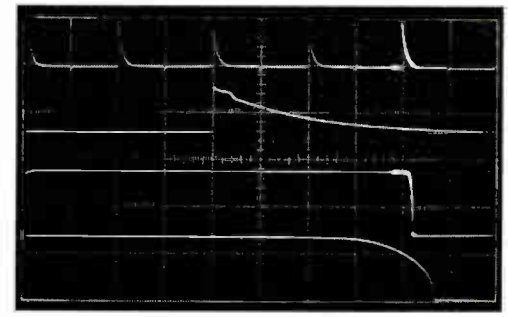


Single-exposure photograph.

same signal — different sweeps

Upper trace is Channel 1/A sweep, 0.1 μsec/cm.
Lower trace is Channel 1/B sweep, 1 μsec/cm.

Using different sweep rates to alternately display the same signal permits close analysis of waveform aberrations in different time domains.



Single-exposure photograph.

2 signals — portions of each magnified

Trace 1 is Channel 2/B sweep, 10 μsec/cm.
Trace 2 (brightened portion of Trace 1) is Channel 2/A sweep, 0.5 μsec/cm.
Trace 3 is Channel 1/B sweep, 10 μsec/cm.
Trace 4 (brightened portion of Trace 3) is Channel 1/A sweep, 0.5 μsec/cm.

Using sweep delay technique—plus automatic alternate switching of the time bases—permits displaying both signals with a selected brightened portion and the brightened portions expanded to a full 10 centimeters.

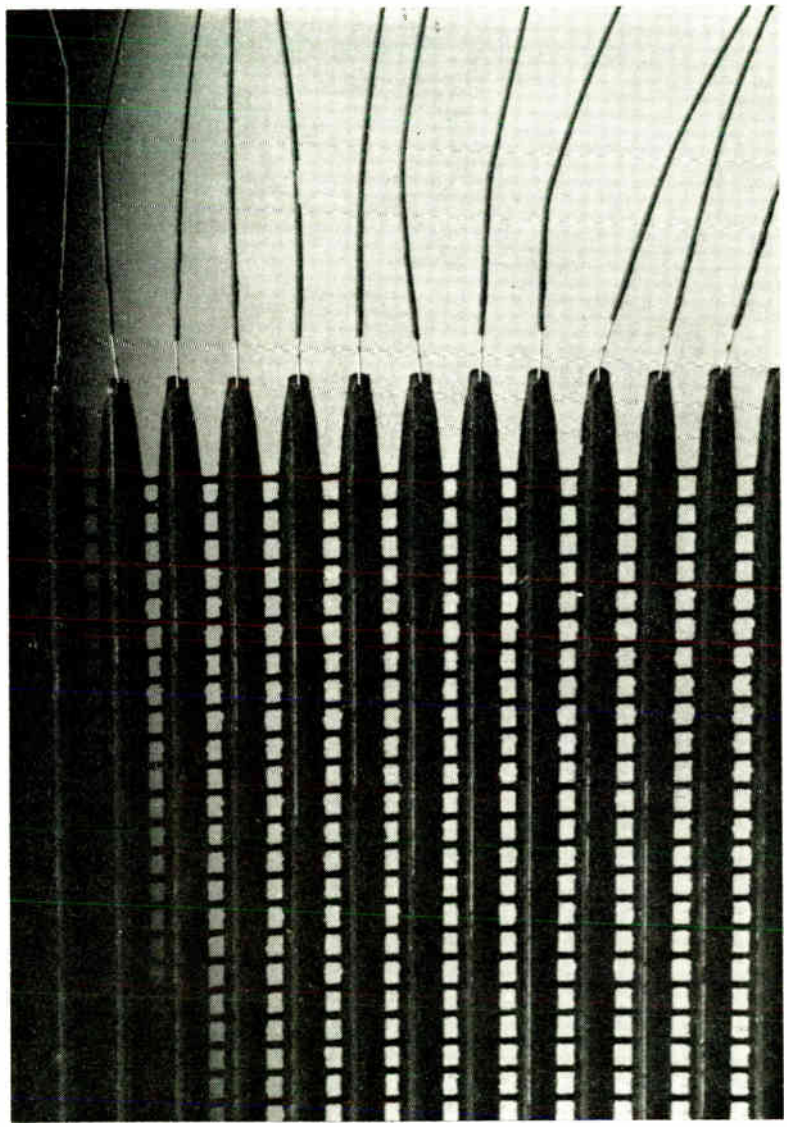
B sweep triggering internally from Channel 1 (plug-in) assures a stable time-related display without using external trigger probe.

For a demonstration, call your Tektronix Field Engineer

Tektronix, Inc. / *P O BOX 500 · BEAVERTON, OREGON 97005 · Phone (Area Code 503) Mitchell 4-0161 · Telex 036-691*
TWX 503-291-6805 · Cable TEKTRONIX · OVERSEAS DISTRIBUTORS IN 25 COUNTRIES
TEKTRONIX FIELD OFFICES in principal cities in United States. Consult Telephone Directory

Tektronix Australia Pty., Ltd., Melbourne; Sydney • Tektronix Canada Ltd., Montreal; Toronto
Tektronix International A.G., Zug, Switzerland • Tektronix Ltd., Guernsey, C. I. • Tektronix U. K. Ltd., Harpenden, Herts

The ART of engineering



Close-up view of an experimental Flute memory array developed by scientists at IBM's Thomas J. Watson Research Center. The highly compact device is made by molding tiny magnetic ferrite "tubes" over a mesh of fine wires. The name of the array is derived from the flute-like appearance of the individual magnetic tubes. It can store 5000 bits of information. Storage is accomplished by changing the magnetization direction of the ferrite material at the intersection of any two wires. More detailed information appears in the Focal Points department in this issue on page 186.

Reflections

75 years ago

New York City Rapid Transit. "It is particularly noticeable that all who have, so far, considered the question of rapid transit for the city of New York, have assumed that it is so deficient in these facilities that the question was of the most immediate and vital importance, and, while I do not for a moment wish to underestimate the urgency of the demand to meet the rapid growth which has already almost overtaxed the elevated system, we should not lose sight of the fact that New York has even today, all things considered, a better rapid transit system than any other city in the world. In evidence of this it is only necessary to cite the facts that one may enter a well-appointed car at the Battery, and, traveling through the best atmosphere of the city, be carried with remarkable smoothness to Fifty-ninth street in 26 minutes, or at the average speed of 12 miles per hour including stops, and in about this manner, the elevated system contrives to carry one-half million passengers per diem for a lower rate of fare than is done in any other city for similar accommodations, and with a smaller proportion of accidents than may be found on any steam road in the world making an equal number of stops per one hundred miles run. On a recent holiday, namely, April 30th last, 835,721 passengers were carried on this much abused system without noticeable detention or accident—but New York wants more, and in view of the rapidly growing traffic which has already almost exceeded the ability of the system, it must be admitted that the demand is a pressing one, as this one half million will grow to one million in a short time, and it is exceeding improbable that any steam motor can be devised to permit of this increase, at least on the present structures, without subjecting them to quickly destructive strains. With steam, therefore, the only avenue open for increasing the passenger capacity, namely the use of longer trains, is obviously out of the question, since it is a notorious fact that the present motors are taxed to their utmost.

"To the mind of the mechanical engineer, having in view the ordinary coefficients of tractive ability, there is no remedy for this, but I think I shall be able to produce at least some little evidence that, owing to certain effects which have not yet been satisfactorily explained, an electric motor may be made capable of solving the problem at least so far as the ultimate strength of the present elevated structure will permit, inasmuch as the effects alluded to certainly do increase the tractive ability beyond that obtained by the use of any coefficients which may be found in the works of that eminent authority, D. K. Clark, or still later in the excellent treatise on the 'Economic Theory of Railroad Location,' by A. M. Wellington. - - -"

"- - - it is now possible to run, and compete with steam on the Ninth Avenue railroad; and, of course, any of the others with a certain concrete economy based upon a series of tests made under circumstances to the last degree prejudicial for an economical showing and without taking into account the great gains which would inevitably result from the use of a large central station, with dynamos of the highest efficiency, and with the obvious advantage resulting from the operation of a number of motive units under varying loads from a central station or dwelling upon such abstract economies as freedom from cinder, dropping water, smoke and noise." (Leo Daft, "Some Recent Electrical Work on the Elevated Railroads and Its Bearing on the Rapid Transit Problem," *Trans. AIEE*, vol. VI, Nov. 1888–Nov. 1889, pp. 359–387.)

50 years ago

All At Sea. "I have repeatedly argued before you that steamships are not so well equipped with radio apparatus as they should be, that the apparatus in use is not so powerful, efficient or reliable as it should be from the standpoint of commercial and distress requirements taken in conjunction with the present knowledge and available apparatus.

"I have particularly tried to point out that it is absurd to give a steamship a comparatively strong voice for ordinary conditions and a hoarse stuttering whisper for use under conditions of extreme distress, when a good strong voice for both ordinary and distress conditions is easily obtainable. Vessels are said to sail as long as five days in some parts of the ocean without being able to carry on communication because of the great average distance of ships and small radio range.

"I have endeavored to bring these points to the attention of all concerned, and have continually described the kind of apparatus I believe should be used. For a time I tried to explain verbally what I believed could be reasonably obtained at the present time.

"A little more than a year ago, I put my ideas of what should be used as radio equipment on board vessels into the form of a set of specifications, and in June, 1913, a number of these specifications were struck off. And it is these revised specifications, with some slight modifications and explanations, that I am presenting tonight for your discussion. In making these specifications, I tried to put myself in the position of steamship owners who desired to secure as satisfactory equipment as could be obtained at a reasonable price at the present point in radio development; and to make the specifications such that when sent out to all manufacturers of radio apparatus, apparatus conforming to the present status of radio development would be offered. This practice of sending out specifications has been adopted by the United States Navy to a considerable extent; and it is probably due to this, in a general way, that we have quite high class radio apparatus manufactured in the United States. Should

KIN TEL

world's largest DC amplifier manufacturer

has an amplifier to fit your

requirements

Solid state, non inverting amplifier

Model 121A/A. Response from DC to beyond 200 kc. Input impedance >10 megohms; output capability ±15 volts, 100 ma; output impedance <0.2 ohm in series with 50 microhenries. Price: $995.

Floating, differential amplifier

Model 114C. High common mode rejection, low effective input capacity. DC to 85 cps. Input impedance >40 megohms; output capability ±10 volts, 10 ma; output impedance <0.25 ohm to 500 cps; <1 ohm to 2 kc. Price: $995.

Floating, differential amplifier

Model 114A. Effective input capacity less than 1$\mu\mu$fd, high common mode rejection. 100 cycle bandwidth. Input impedance >5 megohms. Output capability ±10 volts, 10 ma; output impedance <2 ohms, DC to 1 kc. Price: $895.

Wideband amplifier

Model 112A. Four different plug-in attenuator units. DC to 40 kc. Input impedance 100 K ohms; output capability ±35 volts, 40 ma; output impedance <1 ohm in series with 25 microhenries. Prices: $575 to $680.

Visit our Booth 1017 at WESCON, Sports Arena, Los Angeles, August 25–28.

All KIN TEL DC data amplifiers: have drift less than 2μv • accuracy of 0.01% for any single gain step • fit six to a standard 19-inch rack • are available in portable cabinets • have an integral power supply • are chopper-stabilized • have exceptional linearity and low output impedance.

All prices FOB, San Diego, California. (Additional export charge.)
Write for detailed information. Representatives in all major cities.

5725 Kearny Villa Road
San Diego, California 92112
Phone (714) 277-6700

COHU
ELECTRONICS INC
KIN TEL DIVISION

this practice be followed by the steamship companies, it would probably result in still further development. It certainly would result in the steamship companies obtaining far better apparatus than they have at present." (R. H. Marriott, "Specifications for Steamship Radio Equipment," *Proc. IRE*, June 1914, pp. 165–166.)

The Electric Gyroscope. "Until a few years ago the gyroscope was little known or heard of except as a child's toy, although as far back as 1852 its action was investigated by the French scientist Foucault, who published the results of his experiments in which he laid down very clearly the laws of its action. - - -"

"For nearly half a century after Foucault had demonstrated its action, the gyroscope was put to no practical use. The late American author and scientist Hopkins, whose writings were published for many years in the Scientific American, was the first to drive a gyroscope electrically. He was able to enlarge on Foucault's experiments and obtain much more persistent results with his continuously driven wheel.

"The first serious application of the gyroscope in engineering work was by Obrey, an Austrian naval officer, who made use of the gyroscope to steer torpedoes. About 1903 Dr. Schlick, a German engineer, proposed the use of the gyroscope for stabilizing ships against rolling and built a number of experimental plants which were tried out with some little success. At about the same time Lewis Brennan brought general public attention to the gyroscope by his announcement of a railroad car that would maintain its equilibrium on a single rail. A car of this type was exhibited in this country in 1910 and the subject of monorail traction was much discussed in the press at the time in connection with the possibilities for high speeds which were claimed for this invention.

"Practically all of the serious applications of the gyroscope, and by far the most important of its uses at the present time, have developed in marine work. These are the steering of torpedoes, the use as a compass and the use for stabilizing and controlling the motions of ships, - - -"

"The very high development of the gyro compass during the past few years has been largely due to the efforts of a pioneer in the electrical art, Mr. Elmer A. Sperry, whose long and interesting experience with gyroscopic phenomena and belief in their engineering possibilities led him to devote practically all of his time to its development. . ."

"A number of workers abroad have also been laboring with great persistence and some degree of success to solve the many physical problems involved in translating the idea of Foucault into concrete terms of steel and bronze which will function under the extremely severe conditions on shipboard.

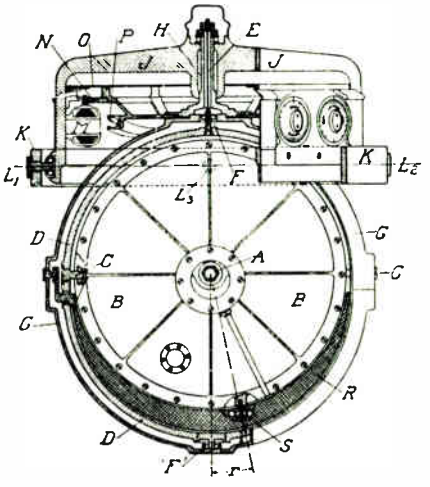
"The nature of the problem is indicated when one realizes that the gyro compass is so sensitive that it responds instantly and with the highest degree of accuracy to the very slowly impressed angular motion in space about the earth's axis four thousand miles distant, while at the same time it remains indifferent to the angular movements of the ship several hundred times greater in intensity. - - -"

"The general arrangement of parts constituting the Sperry gyro compass is shown in Fig. 9, which for greater clearness in illustrating the various functions has been simplified by the omission of electrical circuits and minor details.

"The gyroscope wheel is mounted to rotate on a horizontal shaft A within the casing B, which is pivoted on the horizontal axis C through its center of gravity and carried by the frame or vertical ring D.

"The ring D is suspended by a torsionless strand E and guided by bearings F and F′ to allow a free oscillation of limited amount about its vertical axis within the frame or 'phantom' G, so called because of its characteristic action as a 'shadow' in following up each motion of the ring D. The phantom G has a hollow stem H to which the strand E is attached at its upper end, and the stem forms a journal for rotation in azimuth with respect to the supporting base frame J.

Fig. 9 Method of mounting the directive element.



"The frame J is mounted to swing in the binnacle by the usual Cardan suspension, consisting of the ring K and suitable bearings L1, L2, L3 and L4, the last not shown.

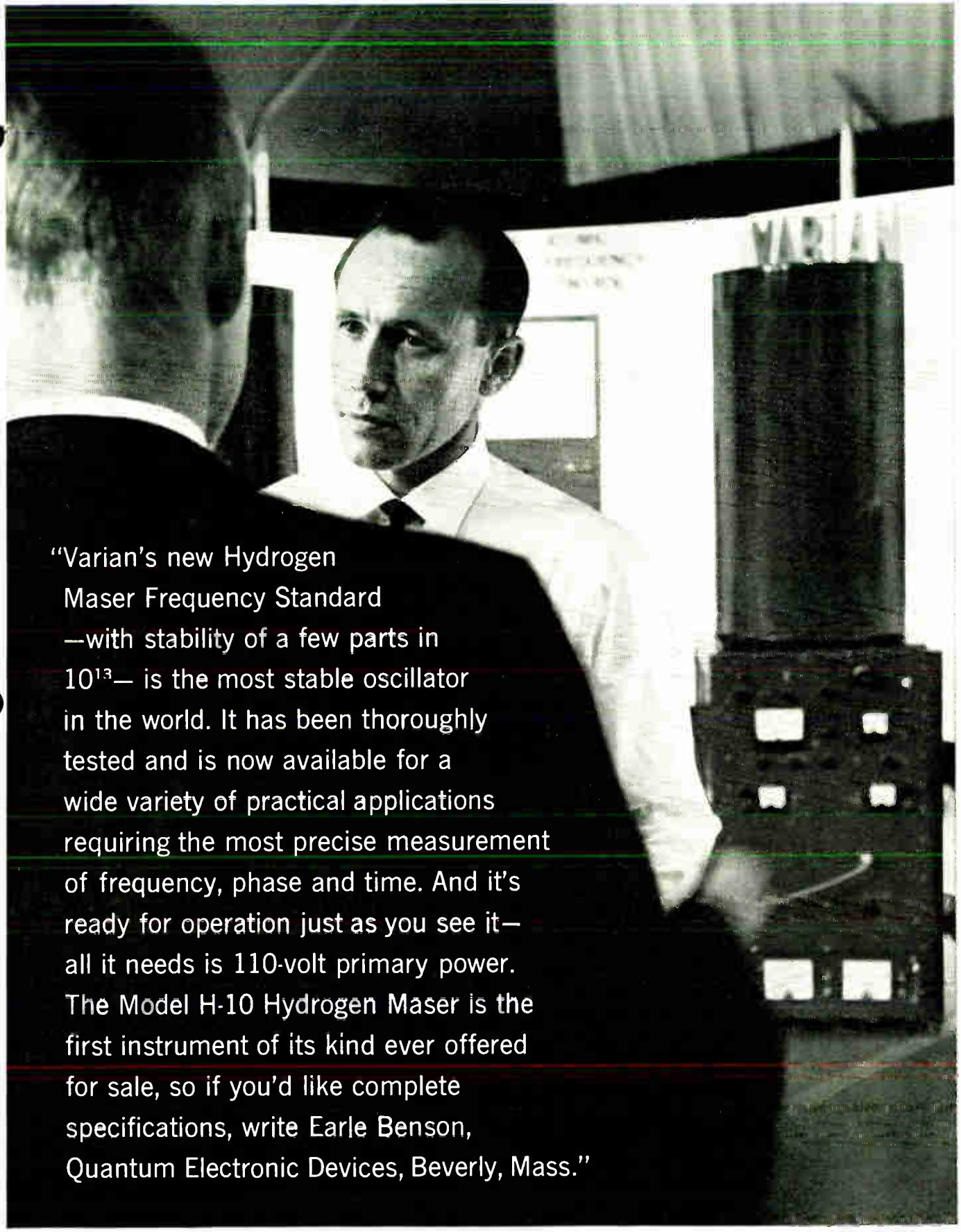
"The frame J carries a 'follow-up' motor M, driving a master gear N, forming a rigid part of the phantom G, whereby the latter is positively driven in azimuth to respond to any movement of ring D.

"The ring D, which with the casing is termed the 'element,' carries a pair of electrical contacts in the form of trolley wheels, which cooperate with stationary contacts mounted upon the phantom G for the purpose of controlling the follow-up motor M, the electrical connections not being shown in the figure.

"The power-driven phantom G is thus automatically controlled so as to respond instantly to all movements of the sensitive element, and in fact to all relative movement in azimuth between the ship and the gyro.

"This characteristic of the phantom enables it to be used for various important functions, namely to carry the scale or card O and a cam P forming part of the automatic correcting device. The power drive is also utilized in operating the transmitter for the repeater instruments, and in overcoming friction of the slip rings needed to carry current to the gyro.

"Since the phantom G, though at all times rigidly anchored to the binnacle, stands in practically constant relation to the sensitive element as regards motion about the vertical axis, but is nevertheless entirely separate thereform, it serves as an anchorage for producing stresses to restrain and correct the movements of the sensitive element. This is accomplished by means of a yielding connection between the phantom G and the gyro case B, in the form of a pendulum or bail R, supported by and forming a part of the phantom. The point of attachment S to the gyro casing is located in a certain position eccentric to both the vertical and horizontal axes of the gyro. By this means several most important results are obtained, including positive orientation and the prevention of oscillation." (H. C. Ford, "The Electrically Driven Gyroscope in Marine Work," *Trans. AIEE*, vol. XXXIII, Part I, Jan.–June 1914, pp. 857–872.)



"Varian's new Hydrogen Maser Frequency Standard —with stability of a few parts in 10^{13}— is the most stable oscillator in the world. It has been thoroughly tested and is now available for a wide variety of practical applications requiring the most precise measurement of frequency, phase and time. And it's ready for operation just as you see it— all it needs is 110-volt primary power. The Model H-10 Hydrogen Maser is the first instrument of its kind ever offered for sale, so if you'd like complete specifications, write Earle Benson, Quantum Electronic Devices, Beverly, Mass."

STABILITY 1×10^{-13} rms for 10-second averaging interval.
5×10^{-13} rms for 1-month averaging interval.
RESETTABILITY 1×10^{-12}
FREQUENCY 1,420,405,751.73±0.03 cps.

VARIAN ASSOCIATES INSTRUMENT GROUP
Beverly, Massachusetts ☐ Palo Alto, California
In Europe contact: Varian A. G., Zug, Switzerland

25 years ago

The Machine Age. "As pointed out by van der Pol in 1934, the analytical investigation of systems producing relaxation oscillations under the influence of an impressed periodic force had at that time hardly begun. The intractable nature of the nonlinear differential equations governing such systems affords a ready explanation for the slow progress encountered. Within recent years, however, with the increased development and use of mechanical means for the solution of mathematical problems, the mechanical solution of these equations is made possible. This paper presents differential-analyzer solutions of the equation of oscillation in certain nonlinear driven systems. These solutions, which have been obtained with the Moore School differential analyzer, illustrate the phenomena of automatic synchronization and frequency demultiplication. The nonlinear theory of oscillating systems, as summarized by van der Pol several years ago, has served as the basis of the work described herein.

"It is observed that in going from the drifting to the synchronized state, synchronization is suddenly effected on the nearest odd subharmonic. Other analyzer solutions, not included in this paper, confirm this. Further, for particular values of K and E, which finally produce synchronization, the resulting steady-state wave form and phase relation with the impressed sinusoid appear to be independent of the initial phase of the sinusoid and of other initial conditions. The only apparent effect of varying the initial phase angle is to hasten or retard by a few cycles the appearance of the steady synchronized state. Inspection of (10a) shows that all such synchronized subharmonics are symmetrical oscillations.

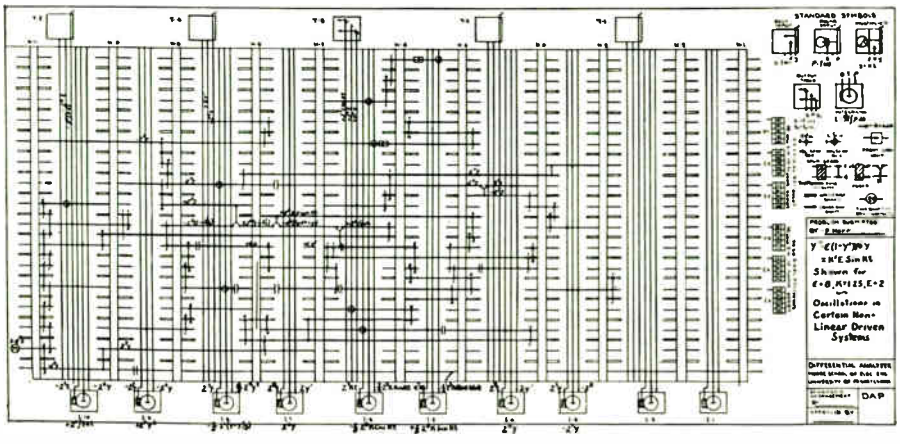
"All of the above phenomena are in agreement with the experimentally observed properties of relaxation oscillations. Especially do they confirm the phenomenon of automatic synchronization over a frequency range as great as an octave (K-2), as mentioned by van der Pol. They also confirm the physical fact that the phenomenon of resonance is practically absent in the case of relaxation oscillations, and that the application of an external electromotive force has as its main effect the variation of time period of the resulting oscillation. This effect is seen to give rise to the phenomenon of subharmonics over a wide range of frequency demultiplication.

"In the course of these studies, employing (10a), only odd-order subharmonics have appeared. The greater the relative frequency of the driving force the higher the order of odd subharmonics it is possible to produce. Increasing the amplitude of the driving force results in successive "driving through" to a lower order of odd subharmonic. It must be stressed that the more general equation (4) is probably the one governing the production of both even- and odd-order subharmonics. Further work in this direction should be undertaken to establish the greater generality of (4) or of a somewhat similar differential equation.

"These results should be of value in the further development of the nonlinear theory of electric oscillations and in application to the design of relaxation oscillators with an impressed sinusoidal electromotive force.

"Fig. B. gives the schematic diagram of the differential-analyzer arrangement used in the course of these studies." (D. L. Herr, "Oscillations in Certain Nonlinear Driven Systems," *Proc. IRE*, June 1939, pp. 396-402.)

Fig. B—Differential analyzer arrangement used in the studies.



How the West Was Won. The first convention of national scope of the Institute to be held on the Pacific Coast is scheduled for San Francisco during the last four days of June. Headquarters will be at the Mark Hopkins Hotel. In the two previous years, meetings were held in Spokane and Portland and gave unmistakable evidence of a substantial interest in a meeting of national character. The San Francisco World's Fair makes that city particularly attractive.

"To one who knows San Francisco, a renewal of acquaintanceship is all that needs be suggested to start making plans and recalling pleasant memories. To one who has never been there, comes indistinct visions of its many charms. Chinatown, the largest Chinese city outside of the Orient, Fisherman's Wharf where the fishing fleet, back from its morning venture, lies at rest, Market Street with its famed clock-towered ferry building at one end and Twin Peaks at the other, are but a few of its world-famed attractions.

"To these one must add Treasure Island and its 400 man-made acres devoted to recreation and education. Sponsored by eleven western states, fifty million dollars have been invested with one thought—to make the visitor want to return for more. An island that may be reached by automobile by way of the new San Francisco-Oakland Bay Bridge, it harbors the giant clipper ships which bring Asian cities within days rather than weeks of America and recall to mind the earlier clippers which once passed through the Golden Gate.

"Our Convention will be held at the same time as the combined Pacific Coast and Summer Convention of the American Institute of Electrical Engineers. Twenty-seven papers will be presented at six technical sessions. Eight are contributions from Pacific coast authors, sixteen have been prepared by authors from the eastern part of this country, two are from Europe and one from Puerto Rico. Thus, the program which follows is of broad geographical and topical interest.

Tuesday, June 27
10:00 A.M.–12:00 Noon

Opening address by R. A. Heising, President of the Institute.

"Communications Engineering in Geo-

physical Exploration," by Herbert Hoover, Jr., United Geophysical Corporation, Pasadena, Calif.

"Federal Communications Commission Engineering Regulations and Standards of Good Engineering Practice for Broadcast Stations," by S. L. Bailey, Jansky and Bailey, Washington, D.C.

"Columbia's West Coast Operations," by L. H. Bowman, Columbia Broadcasting System, Hollywood, Calif.

2:00 P.M.–4:30 P.M.

"Recent Developments in Aerial Navigation," by H. H. Willis, Sperry Gyroscope Company, Brooklyn, N.Y.

"Aircraft Instrument Landing Research at the Massachusetts Institute of Technology," by E. L. Bowles, Massachusetts Institute of Technology, Cambridge, Mass.

"Study of the Effects of Mountains in Radiogoniometry and of the Combined Use of Radio Beacons and Radio Compasses for Aerial Navigation," by Andre Busignies, Le Materiel Telephonique, Paris, France.

"Acoustic Models of Radio Antennas," by E. C. Jordan and W. L. Everitt, Ohio State University, Columbus, Ohio.

"Recent Advances in Receiving Equipment for Transoceanic Telephony," by F. A. Polkinghorn, Bell Telephone Laboratories, New York, N. Y.

Wednesday, June 28

9:00 A.M.—11:00 A.M.

"Electron Optics in Television," by V. K. Zworykin, RCA Manufacturing Company, Camden, N.J.

"Current Division in Plane-Electrode Triodes," by Karl Spangenberg, Stanford University, Stanford, Calif.

"Functions of Electron Bombardment in Television," by I. G. Maloff, RCA Manufacturing Company, Camden, N.J.

"Surface-Controlled Mercury-Pool Rectifier," by T. M. Libby, Pacific Telephone and Telegraph Company, Seattle, Wash.

2:00 P.M.—4:30 P.M.

"Direct-Current and Audio-Frequency Amplifier," by L. J. Black and H. J. Scott, University of California, Berkeley, Calif.

"Golden Gate International Exposition Radio and Sound Distributing Systems," by C. A. Lahar and L. Hewitt, RCA Manufacturing Company, Camden, N.J.

"Radio-Frequency Spark-Over in Air," by P. A. Ekstrand, Heintz and Kauf-

man, South San Francisco, Calif.

"Solar Cycle and the F_2 Region of the Ionosphere," by W. M. Goodall, Bell Telephone Laboratories, Deal, N.J.

"Atmospheric and Radio Transmission Phenomena in Puerto Rico," by G. W. Kenrick and P. T. Sammon, University of Puerto Rico, Rio Piedras, P. R.

"Transmission on 41 Megacycles," by S. S. MacKeown, B. M. Oliver, and A. C. Tregidga, California Institute of Technology, Pasadena, Calif.

Thursday, June 29
9:00 A.M.—11:30 P.M.

(*Joint Session with American Institute of Electrical Engineers*)

"The Klystron as a Generator of Very Short Waves," by W. W. Hansen, R. H. Varian, S. F. Varian, D. L. Webster, and J. R. Woodyard, Stanford University, Stanford, Calif.

"Instruments and Methods of Measuring Radio Noise," by C. V. Aggers, Westinghouse Electric and Manufacturing Company, East Pittsburgh, Penna.; D. E. Foster, RCA License Laboratory, New York, N.Y.; and C. S. Young, Pennsylvania Power and Light Company, Allentown, Penna.

"Methods of Controlling Radio Interference," by C. V. Aggers, Westinghouse Electric and Manufacturing Company, East Pittsburgh, Penna.

"Technical Framework of our Television," by E. W. Engstrom, RCA Manufacturing Company, Camden, N.J.

"A New Standard Volume Indicator and Reference Level," by H. A. Chinn, Columbia Broadcasting System; D. K. Gannett, Bell Telephone Laboratories; and R. M. Morris, National Broadcasting Company; all of New York, N.Y.

2:00 P.M.—4:30 P.M.

"Electronic-Wave Theory of Velocity-Modulation Tubes," by Simon Ramo, General Electric Company, Schenectady, N.Y.

"Recent Ultra-High-Frequency Developments," by B. J. Thompson, RCA Manufacturing Company, Harrison, N.J.

"Simple Television Antennas," by P. S. Carter, RCA Communications, Rocky Point, L.I., N.Y.

"Continuous-Wave Interference with Television Reception," by C. N. Smyth, Kolster-Brandes, Sidcup, Kent, England.

(*Proc. IRE*, June 1939, pp. 405–407.)

Spectral lines

Problems of the Printed Word. The Editorial Board in its discussions last year rather quickly became convinced that publications were the major means by which our Institute would carry out its responsibility for individual member advancement, and that the same means were necessary to advancement of the field itself. However, in discussion of policy, the Board did not immediately assume that publications need follow present formats.

It studied carefully one proposal, received from several sources, which was based on the premise that we all receive more printed pages than we can assimilate each month. The suggestion then followed that the technical journal of the future should consist only of titles of papers, each with a one-page abstract, and some form of reader-service card by which complete papers might be ordered. The member would then order such complete papers as were of interest to him and would save the space required for filing articles outside his field of interest. Perhaps a certain number of reprints per year could be supplied, but beyond this number a charge would be made.

This plan has merit if it reduces the amount of paper to be filed; it has merit in being expandable to include more papers than are now printed, or to citations and papers from foreign sources. It fails if it does not serve the general technical publication needs of the membership, and the Editorial Board rejected it on this basis. They felt that an abstract of a paper would not carry the prestige of the paper in full, that the plan might not attract the best authors, and that in this way it would retard advancement of the field; it also might fail to provide for the archival needs of the profession.

As to the needs of the average member, the plan provides no means of overcoming human inertia in ordering a needed paper in full; in fact, a member might not know that he should order a certain paper for reading a number of years hence. The plan as proposed suggested that only the abstracts should be carefully edited, the complete papers ordered being published without rewriting or condensing. In this case, libraries attempting to maintain a file in full might be burdened with much excess paper.

The Editorial Board was advised that the economics associated with the present state of the art could not justify such a plan. If quick service on complete papers is to be given, the paper must be on the shelf at the time the abstract is published. This entails the use of a very clear crystal ball in predicting press runs; otherwise the Institute ties up its money in printed paper and shelf space, or is faced with expensive reruns. Not the least of the cost would lie in the salaries of paper pickers charged with assembling the papers for an order.

When we have developed our own technology so that one punches holes in a card to order, the card going direct to a machine that can speedily search its microfilmed files for the requested papers, can cheaply produce a photo copy for the order, and drop that copy in a computer-addressed envelope. Then perhaps we will have an economic rival for our present machines, which set type, print a fixed and paid-in-advance number of copies, and place them in computer-addressed envelopes. We must face these facts: even if the achievement of objectives were comparable by the two methods, the good old magazine method is still the economical one.

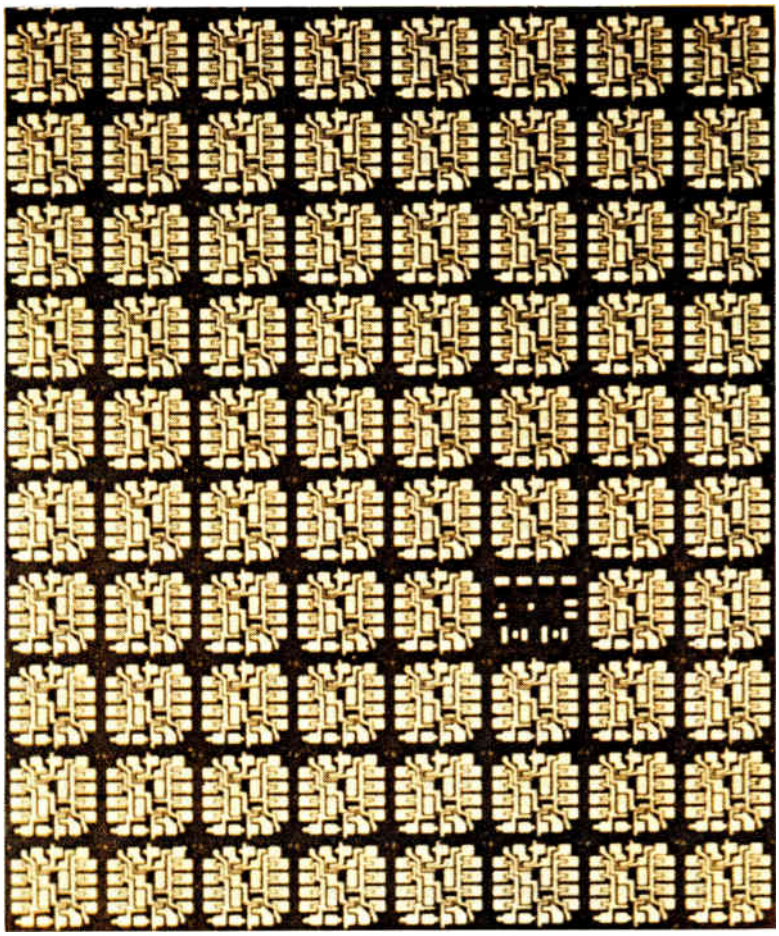
Economics was influential in another Editorial Board decision: that concerned with the supply of annual bound volumes of the former AIEE Bimonthlies. The number of orders for these has always been small, and it is difficult and expensive to predict anticipated demand.

Binding is a hand art, and quantity runs produce no saving. The Editorial Board is convinced that those individual members desiring annual volumes would save money by having their monthly copies bound locally instead of buying extra bound volumes from headquarters. To provide for those desiring the service, however, a decision was made to continue bound volumes at a price ensuring recovery of costs, and print them only to fill advance orders at the beginning of the volume year. Special provision was made to meet the needs of the new Group on Electric Power, since that Group will take over the former Bimonthly POWER APPARATUS AND SYSTEMS.

The Editorial Board also noted the present-day pressures for speedy publication of engineering and research results. The verbal presentation requirements for some of our journals have been delaying publication to ten months or more. On the other hand, IEEE PROCEEDINGS, which does not require that a paper be first presented verbally, is able to publish its papers four months after receipt. However, in competitive areas some journals are exceeding the speed of the PROCEEDINGS and it is noted that authors, understandably, consider the factor of speed in choosing the journal in which to publish. To make our various publications attractive to all authors, and because in an organization of 150 000 members, only a small percentage of these can be given the opportunity of reading their papers at meetings, the Editorial Board urged a reconsideration of policies requiring advance verbal presentation before publication. Members of the Board feel that two different audiences and two different purposes are being served in presentation and publication; they are aware of the delays resulting from the presentation requirement, and they see no logic in coupling the needs of one audience to those of another.

These problems may look like fleas on the editorial beast to some; for more about the beast himself, see page 147 of this issue.

J. D. Ryder



This Westinghouse array of dual DTL gates is fabricated at a density of approximately 1200 gates per square inch. Each gate performs the NAND function and propagates a logic signal at a speed under 17 ns per stage.

Integrated circuits

This edited version of a 1964 IEEE International Convention symposium covers integrated-circuit types; manufacturing principles; the impact of these circuits in government, industrial, and consumer products; and future trends

Patrick E. Haggerty *Texas Instruments Inc.*
C. Lester Hogan *Motorola, Inc.*
Robert N. Noyce *Fairchild Camera and Instrument Corporation*
Leonard C. Maier *General Electric Company*
J. E. Brown *Zenith Radio Corporation*
C. Harry Knowles *Westinghouse Electric Corporation*

Introduction

Patrick E. Haggerty, President, Texas Instruments Inc.

Our subject is integrated circuitry. One generally accepted definition describes it as the physical realization of a number of circuit elements inseparably associated on or within a continuous body to perform the function of a circuit.

This discussion will begin with remarks by Dr. Hogan, who will describe the various types of integrated circuits that are available, including thin films, multiple chip, monolithic silicon, hybrid, and magnetic. He will also compare manufacturing processes, performance, and costs. Next, Dr. Noyce will cover the impact of integrated circuitry in government equipment. Dr. Maier will discuss the impact of integrated circuitry in industrial equipment. Mr. Brown's paper will encompass the effects of integrated circuitry applications in consumer equipment. Mr. Knowles will evaluate research and development in the field of integrated circuitry.

At the conclusion, I shall attempt to summarize the economic impact of integrated circuitry on the electronics industry.

Types of integrated circuits

C. Lester Hogan, Vice President and General Manager, Semiconductor Products Div., Motorola, Inc.

It is difficult to compare the potentials of various technical approaches to a given problem because major breakthroughs often make it possible to achieve particular results cheaper, better, and more reliably by certain methods that may have been discarded as impractical only a few days earlier.

It is impossible to predict breakthroughs in the various present technologies with which we are dealing unless we inject science fiction rather than science.

It is possible, however, to extrapolate efficiencies and productivity within a technology and to set theoretical and practical limits to various technologies that are independent of the breakthroughs. Therefore, I shall make what I feel are proper extrapolations of our current techniques as we become more proficient at the art in which we are engaged.

There are many specific approaches to building integrated circuits today. For example, thin-film integrated circuits can be made by sputtering and anodizing tantalum on a suitable substrate without the use of any other element. They can also be made by evaporation of nichrome, aluminum, gold, and copper alloys, or by the IBM cermet approach. All of these methods have definite advantages and disadvantages. Realistic predictions of the potentials of these individual approaches are difficult to make. Many comments can be made, however, about the general approach of thin-film circuits vs. monolithic silicon blocks.

In the IBM cermet method used for thin films, one first needs either a glass or ceramic substrate. A very high-density alumina ceramic may be used with a very fine surface perfection at a size of approximately one-half

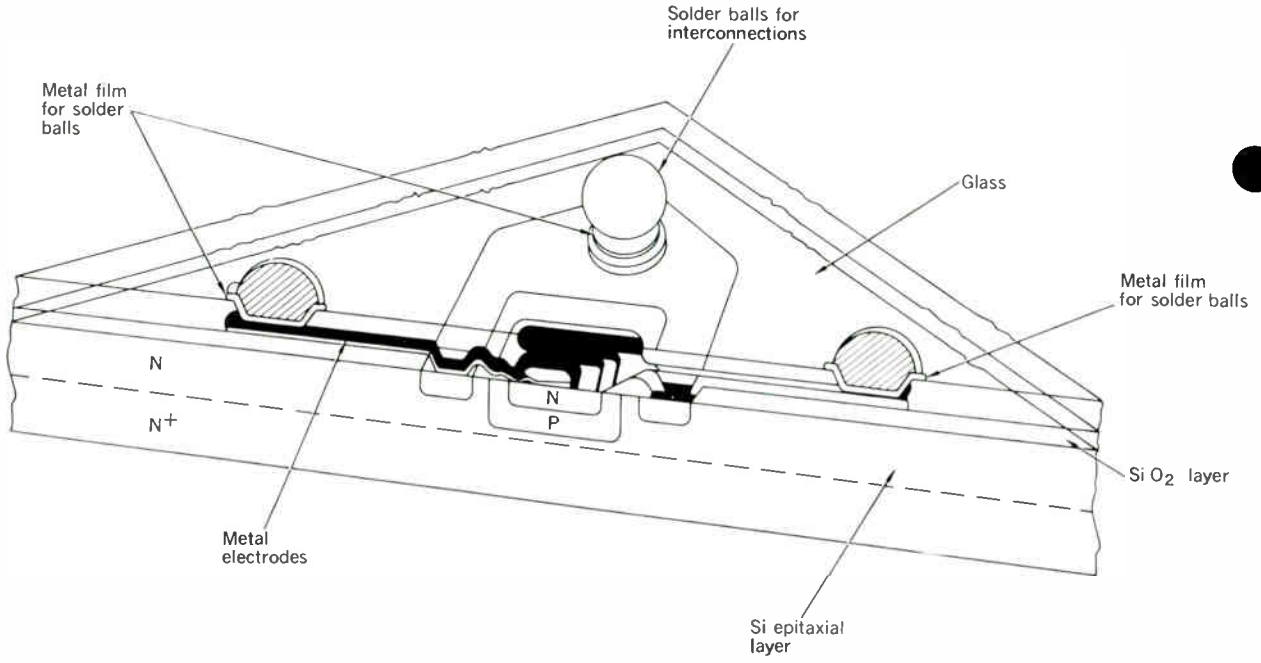
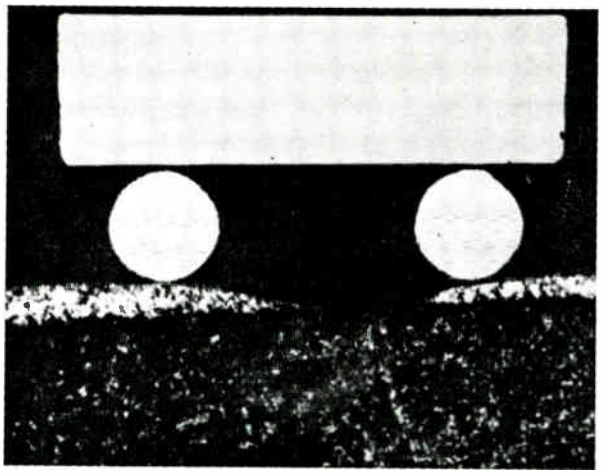


Fig. 1. Cross section of a transistor chip to which three metal balls are attached to make contact with the emitter, base, and collector of the transistor.



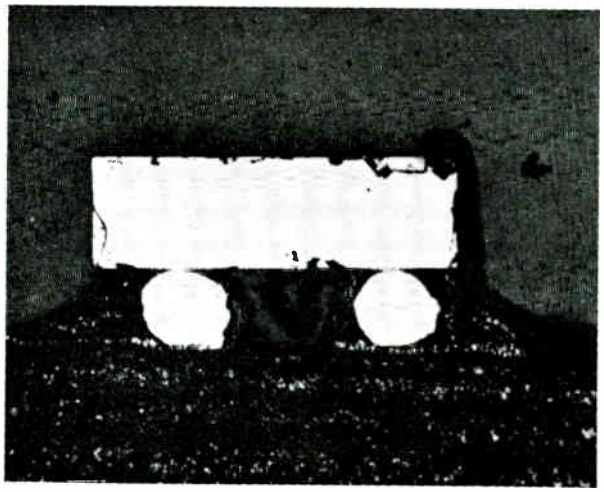


Fig. 2. When the transistor of Fig. 1 is flipped over on the alumina ceramic, the three balls contact the three previously tinned land regions on the substrate.

Fig. 3. After heating, the solder flows around the balls and makes both electrical and mechanical connection to the transistor. Then the unit is potted in plastic.

inch square, metalized to allow interconnections to be made. The interconnections can be screened on the substrate by using inks that contain noble metals such as silver, gold, or platinum. After firing to fix the interconnections, the resistors are printed by the same technique used for the interconnections but the material used is a combination of palladium-silver and glass particles in an organic vehicle. Resistivities ranging from 50 to 50 000 ohms per square can be obtained. The material is fired at about 800 °C; then copper pins are inserted.

Next, the entire substrate is dipped into solder to cover the pins and interconnection pattern. The solder ensures a good electrical connection between the pins and the interconnection land pattern; it lowers the series resistance of the lands, and it provides the bond required for joining the active elements.

The resistors are then tailored to the desired value by abrasive sand blasting. Finally, the transistors are attached by the flip-chip technique in which three metal balls are attached to the transistor chip to make contact with the emitter, base, and collector, as shown in Fig. 1. When the transistor is flipped over on the alumina cer-

Fig. 4. A typical silicon monolithic integrated circuit. This circuit is a current mode gate that performs both the OR and NOR function in a digital logic system.

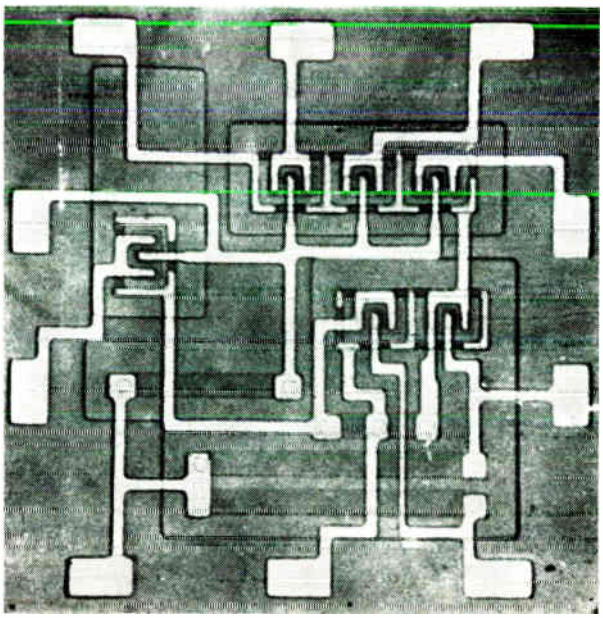


Fig. 5. Another silicon monolithic integrated circuit is this flip-flop with set and reset capabilities.

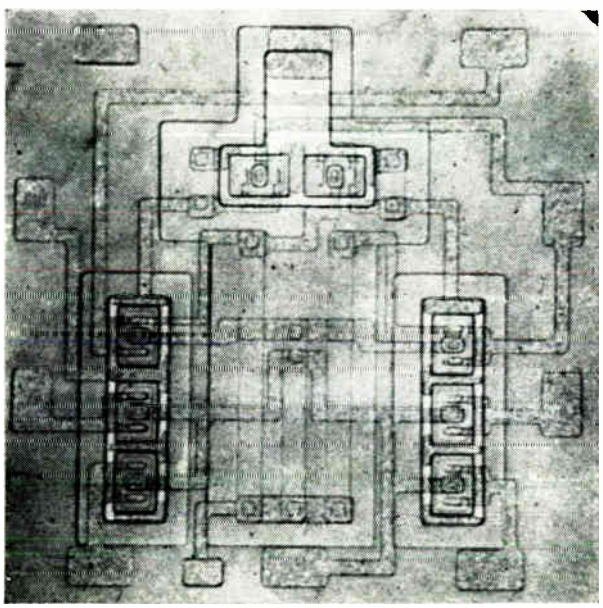
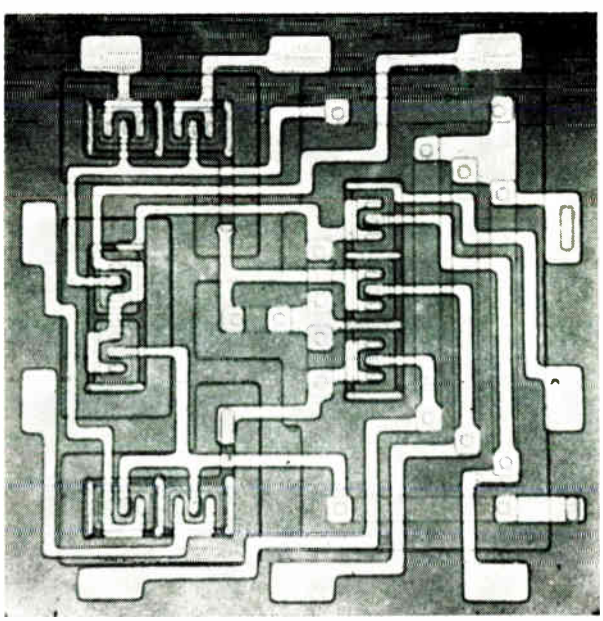


Fig. 6. A third silicon unit is this current mode half-adder.



amic, the three balls contact the three previously tinned land regions on the substrate; see Fig. 2. After heating (Fig. 3), the solder flows around the balls and makes mechanical and electrical connection to the transistor. Finally, the entire substrate is potted in plastic to give a completed thin-film integrated circuit.

A silicon-substrate monolithic integrated circuit is made by a series of photoresistant and diffusion steps (see "Monolithic integrated circuits," by A. B. Phillips, this issue, pp. 83–101). The main advantage of this process is that more than 400 individual circuits can be processed on one wafer of silicon approximately one inch in diameter. Hence, both the labor and material costs can be kept lower than any other known process provided they are produced in large quantity at reasonable yields. Usually, several hundred wafers can be diffused at one time. Since there can be from 20 to 1000 circuits per wafer, it is possible to process 100 000 circuits at a time.

Typical circuits that are produced by this technique are a current mode gate that performs both the OR and NOR function (Fig. 4); a flip-flop with set and reset (Fig. 5); and a current mode half adder (Fig. 6). All of these circuits as shown are on silicon chips of 50 by 50 mils and approximately 400 of them are produced at one time on a single wafer.

The performance of all silicon-substrate monolithic circuits made to date is limited by the capacitance of the back-biased junctions that occur between the silicon substrate and the isolated n-type islands in which the elements of the circuit are constructed. This parasitic capacitance has been such a limiting factor that there are many exponents of thin-film circuits who have felt that the demand for higher and higher speeds would eventually force the silicon monolithic circuit out of use since thin-film circuits on insulating substrates do not suffer from this effect.

I. Speed comparison data for various NOR gates

	Standard Monolithic, ns	EPIC Monolithic, ns	Hybrid, ns
Delay propagation time, t_D	6	5	4.6
Rise time, t_R	8	4.7	4.5
Fall time, t_F	9	5.2	5.2

Fanout = 1; fan-in = 3; C_S (test jig) = 13 pF.

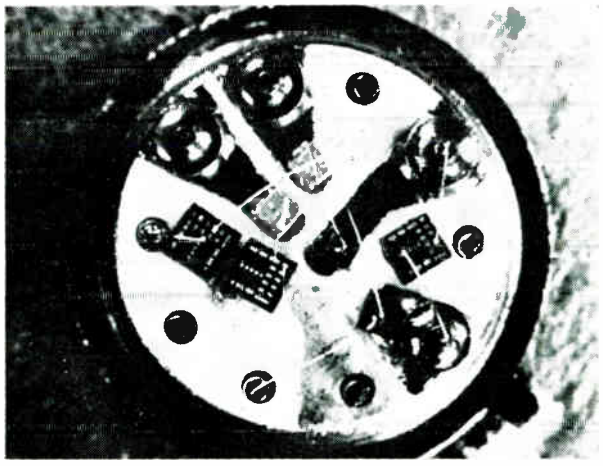


Fig. 7. Hybrid integrated circuits combine one or more integrated circuits with one or more discrete devices.

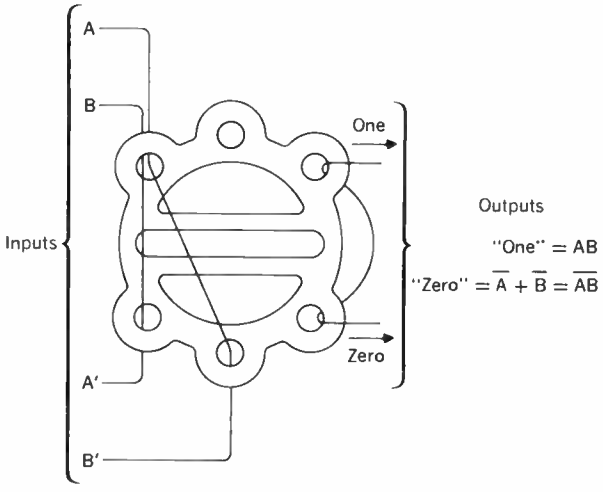
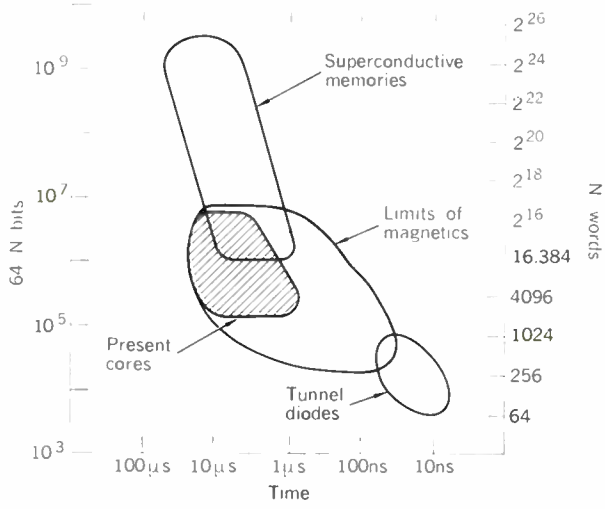


Fig. 8. Multiaperture ferrite cores are used for magnetic elements that can perform most needed logic functions. This particular unit performs as an AND gate.

Fig. 9. Limits of speed and storage capacity of magnetic, superconductive, and tunnel-diode memories.



It is, therefore, quite significant that this problem has been solved and the fabrication process for new circuits that do not suffer from this limitation has advanced from the research laboratory to pilot-line production. The new EPIC circuits (see this issue, p. 84) are made by exactly the same processes as before, but now the cross-section of the wafer shows that an insulating layer separates each isolation island from the silicon substrate that supports it. The extra processing to obtain this insulating layer will add about 0.5 cent to the cost of the finished integrated circuit when the process is in full-scale production. This insulating layer can be made so thick that the effective capacitance is reduced to zero.

Measurements have been made on hundreds of circuits produced in this new way. They are as fast as their hybrid circuit counterparts which do not suffer from this parasitic problem. Table I shows typical data taken on two monolithic MECL OR–NOR gates. The first set of data was taken on today's standard product; the second was taken on the same gate made by the new EPIC process. Even at this early stage of development, the rise and fall times have been almost cut in half and the propagation delay has been reduced by about 20 per cent.

Hybrid integrated circuits can be made by the combination of one or more integrated circuits in combination with one or more discrete devices as shown in Fig. 7. In spite of its superior performance, the only application for this type of circuitry in the long run is for special purposes where complexity or performance are so demanding that the technology does not permit reduction to monolithic form or when only a few circuits are required and one cannot afford the mask charges associated with the manufacture of monolithic circuits.

There probably will always be circuits than can be realized by the hybrid technique and that cannot be realized by other approaches. But when the quantity of circuits required is great enough—1000 or more of the identical circuit—the hybrid approach cannot achieve the low cost of the monolithic silicon approach.

Magnetic elements have long been used to realize certain circuit functions. Most logic functions can be obtained by the proper winding of wires on multiaperture ferrite cores. Figure 8 shows such a core that can perform the function of an AND gate.

Magnetic circuits can be made at very low cost and probably have a reliability exceeding that of all other integrated circuits. However, the magnetic circuits are extremely large and extremely slow compared to silicon circuits. In addition, transistors are required to drive the magnetic circuits and, if you need a transistor to operate your circuit, you might just as well build the entire function on the silicon substrate. This is true because, in one sense, you can get the other elements free if they do not occupy any more area than the transistors.

Finally, magnetic thin films and superconducting thin films have received much attention for more than ten years. While they still offer some promise for performing logical functions, their present state of development limits them to use in computer memories. They certainly have very great promise in this area of our technology. Thin magnetic films promise storage capacity as great as ten million bits with cycle times of the order of 100 ns, as shown in Fig. 9 which is taken from a recent paper by J. A. Rajchman of RCA.

Superconducting thin films promise memories with

II. Two-phase J K flip-flop circuit table

Year	Width, mils	Tolerance, mils	Resistors, ohms per square	Total Resistor Area, mils²	Total Transistor Area, mils²	Surface Efficiency, per cent	Projected Die Size, mils
1964	0.5	0.5	200 (diffused)	830	892	40	70 × 70
1965	0.3	0.3	1000 (film)	45	451	45	40 × 40
1966	0.2	0.2	3000 (film)	14.8	257	50	30 × 30
1967	0.1	0.2	10 000 (film)	4.4	200	55	25 × 25

capacity greater than one billion bits with cycle times as fast as one microsecond. In this particular application these films appear to have no peers but for general circuit use they do not seem to have the promise of achieving the performance, cost, or size of either the silicon monolithic circuit or the general thin-film circuit in which passive circuit elements are deposited on an insulating substrate.

After surveying the field, one is left with two types of circuits that promise to have general-purpose use for integrated circuits of the future. These are thin-film circuits and monolithic silicon circuits. Each has its special field of application but it appears that monolithic silicon will have the edge in digital circuit applications through cost alone.

Very complex silicon circuits can now be constructed on a chip approximately 50 by 50 mils. The cost to carry a wafer through all steps of diffusion and metalization, provided it is run at a moderately high volume and at a 100 per cent yield, is about $10 per wafer including normal overhead. Assuming that a one-inch-square wafer can be processed at a 100 per cent yield before packaging, each wafer will contain 400 individual circuits which would then cost 2½ cents apiece. Even if the ultimate yield is as low as 50 per cent, the cost of the finished silicon monolithic integrated circuit, before packaging, will be less than 5 cents.

Over the next three years, the resolution with which silicon circuits can be built will steadily improve. In the research and development laboratories, one bit of a two-phase-shift register has been placed successfully on a 70-mil-square die. Again assuming a 50 per cent yield, this entire circuit consisting of 33 transistors, 27 resistors, and 2 diodes, could be built for less than 10 cents.

A reasonable projection of what may be achieved shortly is illustrated by Fig. 10. This circuit—a J K flip-flop—contains 14 transistors, 10 resistors, and 2 capacitors. It has already been placed on a 70-mil square die; see Table II. The possibility of placing it on a 25-mil-square die by 1967 may be questioned but there is little doubt that the circuit will be placed on a 30-mil-square die by 1966. At a 50 per cent yield, with a reasonably high volume before packaging, this circuit will cost less than 2 cents to process, including factory overhead.

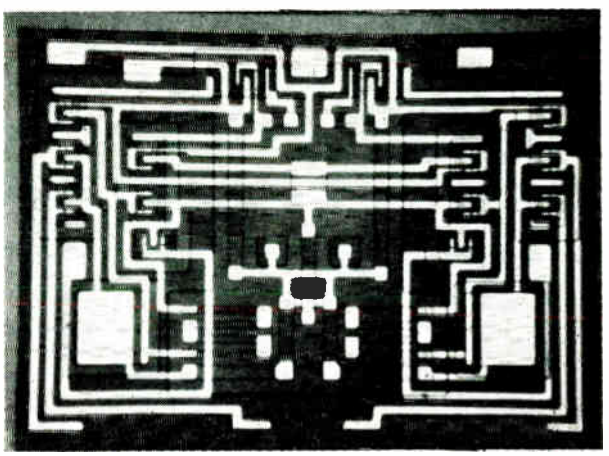
Basically, the cost of producing a silicon monolithic integrated circuit or a silicon transistor is related both to die size and yield. One cannot talk about die sizes much less than 20 mils square because the difficulty in handling and alignment during packaging and attachment of leads mitigates against smaller die sizes. Thus, assuming we are attempting to build transistors on a one-inch wafer rather than integrated circuits and assuming an ultimate die size of 20 mils square, one is led to individual transistor costs in excess of one cent apiece as compared to a complete J K flip-flop for two cents. All these costs, of course, refer to a device before lead attachment and packaging.

Any method that requires the construction and attachment of individual transistors to a circuit can never achieve the low cost of the monolithic silicon approach to integrated circuits, provided the yield percentages are comparable.

Whenever one builds any device through the successive diffusion of impurities into silicon, the resulting devices show a distribution of parameters. In testing transistors, one can divide the distribution into a reasonable set of limits and hopefully find a market for all the devices which come off the production line.

In building integrated circuits, it does not seem feasible to segregate the devices into different categories (as is done with transistors) to accommodate the wide variations that occur from device to device. Also, the yield

Fig. 10. This J K flip-flop silicon integrated circuit contains 14 transistors, 10 resistors, and 2 capacitors.



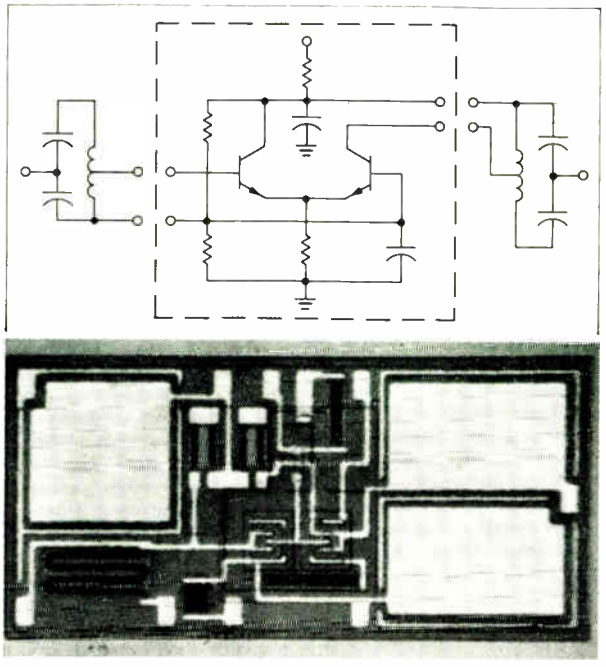


Fig. 11. Circuit diagram and silicon monolithic integrated circuit for a 120 Mc/s RF amplifier. Specifications for the circuit are: B^+ voltage, +12 V; drain, 2.3 mA; power gain, 18 dB at 120 Mc/s; noise figure (1000-ohm source), 6.0 dB at 120 Mc/s; input resistance, 1200 ohms; input capacitance, 4.5 pF; output resistance, 2500 ohms; and output capacitance, 5 pF.

of any device, whether it be a circuit or component, is related exponentially to the area it covers on a silicon chip. Integrated circuits must by nature cover more area than an individual transistor contained within the circuit and, therefore, one should expect for some time to come that more good transistors than integrated circuits will be obtained per wafer.

Present transistor production, however, must strike a balance between a variety of extremely tight specification limits, with the result that realistic yields are often much lower than most of us care to admit. A transistor is usually specified in extreme detail. We don't make the selection from just one distribution curve and its resultant yield loss; we make successive selections from 20 to 40 different distribution curves. Conflicting requirements such as high breakdown voltage, low saturation resistance, high current gain, and low storage time, are all specified. If the distribution curves are so tight that 90 per cent of the transistors fall within the specification limits for 20 specified parameters, the resulting yield to all specifications will be about 12 per cent.

This situation need not occur with integrated circuits since the circuit designer is concerned with black-box specifications concerning circuit performance and not with detailed characteristics of the individual components within the circuit. If the integrated circuit functions at all, it usually meets all the specifications placed on it, once one has learned how to make it using the proper mask design, the proper processing steps, and the proper resistivity of the starting wafers.

To illustrate what has been achieved, Fig. 11 shows both the circuit diagram of a 120 Mc/s RF amplifier and its appearance as a silicon monolithic integrated circuit. Several thousand of these circuits have been built with an actual yield of 90 per cent where yield is measured in terms of the direct-current operating levels and gain vs. frequency of the amplifier. The 90 per cent yield is measured relative to the specifications shown in the caption of Fig. 11. These simple specifications are the kind that might become realistic when one deals with circuits instead of individual devices.

We have accumulated vast data on the yields of our standard integrated circuit and transistor lines. The yield on the former is about the same as that for the very critical transistors now being produced. Hence, from these data, it seems reasonable to assume that monolithic silicon integrated circuit yields and transistor yields are comparable for similar size dies.

These factors may give the impression that monolithic silicon integrated circuits will eventually eliminate thin-film elements, but this is not the case. Diffused resistors have such a large temperature coefficient and the maintenance of microscopic mechanical tolerances is so difficult that one cannot build such resistors with tolerances less than 20 per cent with any reasonable yield. Critical circuits, which require closer resistance control, therefore must be made by thin-film or other techniques. The characteristics of various resistors for use in integrated circuits is detailed in Table III.

Capacitors made with diffused p-n junctions have very low Q, are voltage sensitive, are polar, and can be subjected only to relatively low voltage. For these reasons, thermally grown silicon dioxide or other glass is usually used as a dielectric rather than the junction capacitance. Properties of such capacitors are detailed in Table IV.

The major disadvantage of silicon substrate capacitors, however, is their cost. If thermally grown silicon dioxide is used as the dielectric, Table IV indicates that capacitance values of about 0.25 pF/mil^2 can be obtained. Again considering our $10 one-inch wafer, we see that a capacitor built on a silicon substrate would cost about 0.008 mill per picofarad at 50 per cent yield. Therefore, capacitors placed on silicon substrates must be limited to rather small values.

For comparison, we can calculate the cost of a thin-film capacitor that uses pyrolytically deposited boroaluminosilicate glass as the dielectric:

½ by ½-inch substrate	$0.12
Aluminum interconnect	0.04
Boroaluminosilicate	0.02
Aluminum counterelectrode	0.04
	$0.22

Assuming a 90 per cent yield on this process, the actual cost is 24 cents. Again, if the capacitor covered the entire area, the capacitance value would be 0.1 μF, or a cost of 2.4×10^{-4} cents/pF. Thus, it costs about 30 times more to make a capacitor on a silicon substrate than on an insulating substrate using thin-film techniques.

Completely passive circuits (those requiring no transistors or diodes) with large capacitance and resistance values will be cheaper to construct on insulating substrates. However, this restriction applies principally to *large* capacitors and *large* resistors. If the values required are small enough so that the capacitors and resistors

can fit on a 50-mil silicon chip, it may be cheaper to build on silicon because it is difficult to obtain very small ceramic substrate.

The problem of determining whether to put passive components on silicon or to place them on an insulating substrate and attach the silicon active elements to them is a complex problem involving the actual size of the finished circuit and the cost of the attachment procedure.

If the passive components require an area that is trivial compared with the area needed for the transistors and diodes in the circuit, the use of the silicon substrate will provide an excellent economy. But if the area required for the passive components is greater than that needed for the active components, packaging and interconnecting techniques will determine whether it is cheaper to use the silicon substrate or an insulating substrate. And, if the circuit's passive components, when placed on the silicon substrate, occupy essentially the entire area and there are relatively few active components, the thin-film technique for passive components with active elements attached would be the best approach.

A basic guide is to determine whether the passive components will occupy more than the minimum die size that can be used otherwise. There is a gray area between the two extremes where interconnection, packaging techniques, and actual substrate costs will determine the approach. To illustrate, assume that the split in this gray area occurs when the passive com-

III. Integrated circuit resistor characteristics

Parameter	Monolithic Silicon Components		Thin Film Components	
	Diffused p	Diffused n	Nichrome	Cermet
Sheet resistance, ohms/square	100–300	2.5	40–400	100–1000
Resistance per substrate area*, ohms/mil²	50–150	1.25	20–200	50–500
Temperature coefficient, ppm/°C	+2800 to 1500	+100	±100 adjustable to within ±10%	−55
Power dissipation per active resistor area†, mW/mil²	3	3	2	
Maximum voltage, volts	20	6	—	—
Tolerance for high yield,‡ per cent	±20	—	±8	±8
Distributed capacitance§, pF/mil²	0.2	0.6	—	—

* Assumes 1-mil-wide stripe, and 1-mil-wide spacing.
† Depends upon package and heat sinking.
‡ Tighter tolerances available at higher cost.
§ Negligible using recent EPIC techniques.

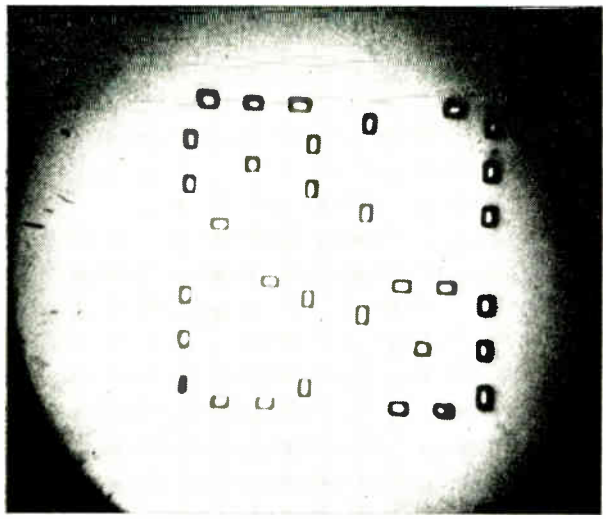
IV. Integrated circuit capacitor characteristics

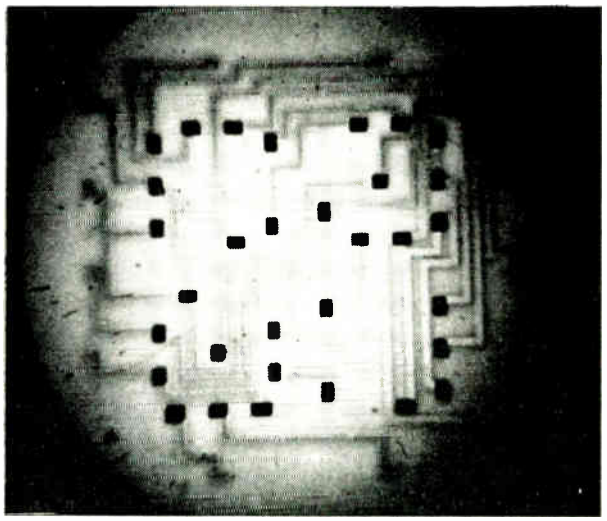
Parameter	Monolithic Silicon Components			Thin-Film Components		
	Single Diffused p-n Junction	Double Diffused p-n Junction	Thermally Grown SiO_2	SiO	Boroalumino-Silicate Glass	Tantalum Oxide
pF/mil²	0.1 at $V_{bias} = 0$	1.0 at $V_{bias} = 0$	~0.25	0.01	0.4	2.5
V_{max}, volts	30	6	50	50	50	20
Dissipation factor, %						
1 kc/s	10	100		2.5	0.2	0.8
1 Mc/s			0.7	0.7	0.2	0.3
10 Mc/s			2.0		1.0	
Capacitor temperature coefficient, ppm/°C	Low	Low	Low	±200 ±50	+115	+400
Capacitor voltage sensitivity	$\sim V^{-1/2}$	$\sim V^{-1/2}$	0	0	0	0
Polar	Yes	Yes	No	No	No	No
Shunt capacitance, per cent component value	25	25	18	≈0	≈0	≈0
Leakage current at 5 volts, A/pF	10^{-9}	10^{-9}	10^{-9}	10^{-14}	10^{-15}	10^{-15}

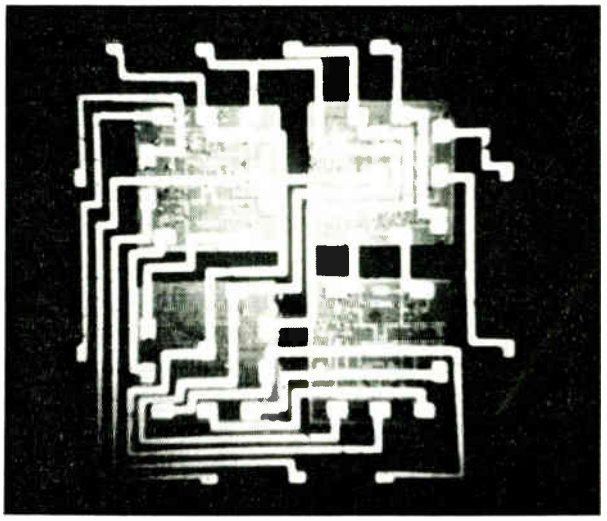
Fig. 12. One approach to the combination of monolithic silicon with thin films. Aluminum bonding islands are placed on a glass substrate by photo-resist methods.

Fig. 13. Thin-film components and interconnections are next deposited on the glass substrate shown in Fig. 12.

Fig. 14. The final step is to flip the monolithic silicon circuits over onto the bonding islands where an aluminum-aluminum bond is made. The result, in this case, is a two-phase shift register made up of two half-adders and two flip-flops bonded to the thin-film substrate.







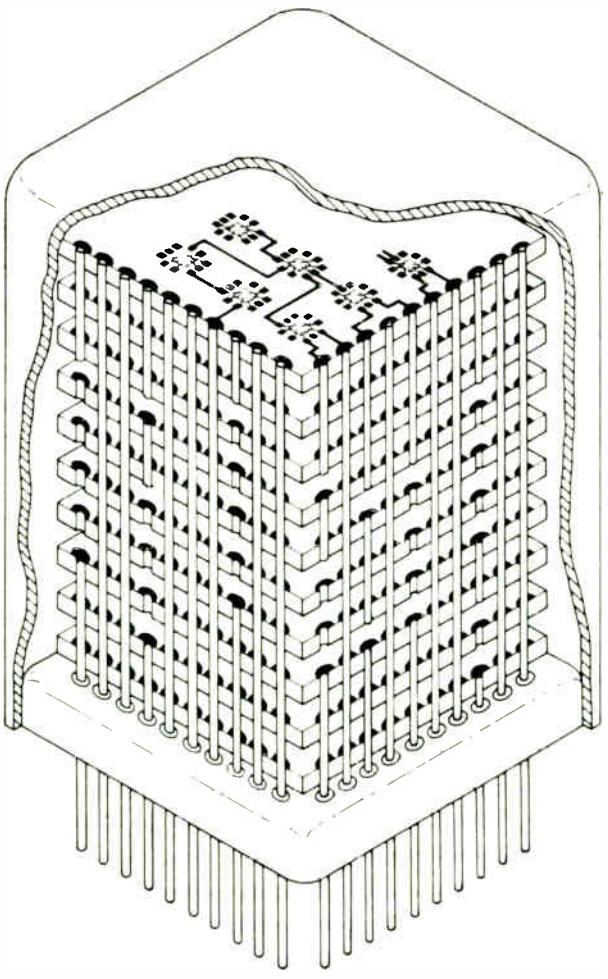
ponents require the same area as the active components. With this guideline and use of Tables III and IV one can arrive at reasonable limits for the resistance and capacitance values that one should attempt to place on a silicon substrate.

From the preceding arguments, one can reach the following conclusion. If the area consigned to passive elements is equal to that allocated to active elements, it may be reasonable to assume that it will be cheaper to put the elements on the silicon substrate than to pay for the extra interconnection problems that will arise through the use of thin-film techniques.

A 50-mil-square die, assuming a space-utilization efficiency of 40 per cent has approximately 1000 square mils available for circuit components. If one half of this area is given to resistance, then we can accommodate about 75 000 ohms. If this same area is used for boroaluminosilicate capacitors, it can handle about 200 pF. If the circuit calls for total resistor and capacitor values greater than this, one should certainly investigate the thin-film approach.

The techniques to be used for forming integrated circuits and the ultimate cost will depend largely on the type of final packaging and interconnection. For example, when attaching discrete transistors, or partially

Fig. 15. Modern packaging techniques under development make maximum use of integrated-circuit sizes.



integrated circuits, to a substrate that contains interconnections and passive components, the break-even point depends on the complexity and the cost of the attachment process. Also, when it is possible to build so much circuit complexity into a silicon chip only 50 mils square, it seems a waste of space to package these circuits hermetically and separately, since the individual packages take up so much space. Research is being done on packaging techniques. Within a year or two there will be integrated circuit packages that will not only contain flip-flops, logic gates, or half adders, but also will include collections of these circuits for the production of functional items.

At first, the complexity of these future packages will be kept low so that individual packages will contain elements such as shift registers or binary or decade counters.

In one technique, Fig. 12, that may combine the best features of monolithic silicon with those of thin-film technology, a group of aluminum bonding islands are placed on a glass substrate by photoresistant techniques. These islands are made about one mil thick so that the silicon chips attached to them will be held above and clear of the substrate. Next, as shown in Fig. 13, thin-film components and interconnections are deposited on the glass substrate. Finally (Fig. 14), monolithic silicon circuits are flipped over onto the bonding islands where an aluminum–aluminum bond is made. In this case, two half adders and two flip-flops are bonded to the thin-film substrate to form one bit of a two-phase shift register. Many functional elements have been made by this promising technique since it offers a compact method for combining the best features of both monolithic silicon and thin-film circuitry.

An example of other packaging techniques being developed is shown in Fig. 15. This assembly contains several six-layer printed-circuit boards that will ultimately hold 16 silicon chips each. Each board has two signal planes and four voltage supply planes. After the boards are assembled with their individual silicon chips, they are stacked very much like the old "tinkertoy" project using vertical pins to make the interconnections between boards as illustrated. Using this technique, we have recently assembled a full parallel–parallel adder with look-ahead logic in a one-inch cube.

Certainly some packaging technique similar to the one just described is needed in order to take full advantage of the size of present-day integrated circuits.

Integrated circuits in military equipment

Robert N. Noyce, Vice President and General Manager,
Semiconductor Div., Fairchild Camera and Instrument Corporation

The military effect on the progress of integrated circuits has been twofold. First, new technology has developed, some through the direct subsidy of military research and development, and much more through the company-sponsored research stimulated by this support. About $100 million of R & D expenditures escalated from an initial Government expenditure of $2 million. The second effect has been the military agencies' interest not only in using integrated circuits but also in providing the market and the motivation for suppliers to complete the development and establish the production capability to supply this waiting market. Military and space applications accounted for essentially the entire integrated circuits market last year, and will use over 95 per cent of the integrated circuits produced this year. Even in 1970, these applications may well be using as high a proportion as 55 per cent of the circuits produced.

The "Dick Tracy wrist radio" characteristics of integrated circuits are widely known, and there are tremendous size and weight reductions in electronic equipment using these techniques. In many applications, particularly those in which a ton of fuel is required for every pound of payload, these reductions are very important. This is not the only attribute, however, that motivated military agencies to use integrated circuits. Reliability is the most important single factor. We have data on two operating medium-sized computers that use integrated circuits. The first is the Apollo guidance computer, designed by MIT and built by Raytheon. It has accumulated 19 million operating hours on its integrated circuits, in which time two failures have occurred—an initial failure, and the other a failure, external to the package, that was caused by moving the computer. The second system, the MAGIC 1, an airborne computer built by the AC Spark Plug Computer Division, has accumulated $15\frac{1}{4}$ million hours with two failures. Fairchild's in-house life-test program, with 33 million total operating hours, has had a total of eight failures; of these, five accumulated during the first $6\frac{2}{3}$ million hours and only three occurred on more recent units during the last 26 334 982 hours.

These data are not extrapolated from accelerated tests, but are actual, observed operational failure rates, and include early production units in some cases. Considering the complexity of the function performed by these circuits, the integrated circuit equipment today is ten times more reliable than its discrete component counterpart. As new failure modes are identified and eliminated, we

may see substantial improvements in the reliability figures.

Today's integrated circuit, with minor exceptions, is just as sensitive to nuclear radiation environments as were yesterday's transistor equivalents. In some military and space applications, this will place a serious limitation on integrated circuits that use conventional transistors for the active elements.

The most liberal way to measure integrated circuit cost is to neglect development expenditures and to consider the total mission—which includes initial cost, maintenance and repair, spare parts logistics, and delivery. For satellite applications, with their premium on weight, integrated circuits are cheaper to use than conventional circuits.

Prices of individual transistors supplied to military contractors range from $3 to $5 in small quantity. In quantities of 50 000 or more, unit prices vary from 75 cents to $2, depending upon transistor type. Tight screening and burn-in for higher reliability will increase these prices.

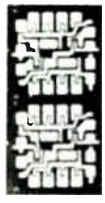
By comparison, if we consider only the transistors in an integrated circuit, typical prices are about $4 per transistor in small quantities; and in quantitites of over 50 000 prices of $1.50 to $1.75 are average. The reason for this lower cost is that the silicon chip size of a typical 12-transistor circuit can be smaller than that for the 2N1613 transistor.

Performance is another factor, and there are large areas of electronic equipment that cannot be equipped with integrated circuits. In general, the same limitations apply to integrated circuits and transistors. For example, we cannot replace the magnetron in the radar set, and it is difficult to make accurately tuned circuits in integrated form. However, many of the integrated circuit limitations are being overcome rapidly.

In developing any new technology, schedule slippages are expected. As an industry, we have a bad reputation in this area. There are many cases where component manufacturers have committed themselves to a delivery schedule for integrated circuits and have not met the deadline.

But as the range of circuits available as off-the-shelf items is expanded, the designers and manufacturers for the military market will find standard components much more compatible with requirements. And as the components industry gains experience with integrated circuits of special design, manufacturing and delivery schedules will be met on time.

Integrated circuits now satisfy many of the military and space requirements and there will be an increasing use of integrated circuits in military systems. Today, the advanced Minuteman, Apollo, Phoenix, and all new military digital computers use integrated circuits for the major part of their electronics systems. With higher reliability, lower cost, and better performance, many missions once considered too imaginative have become or are becoming both feasible and practical.



Integrated circuits in industrial equipment

Leonard C. Maier, General Manager,
Semiconductor Products Department, General Electric Company

The commercial computer industry accounts for more than half of industrial electronics. Although none of the 1963 production equipment used modular circuits, all manufacturers are exploring their use for high-volume logic applications. The choice of modular circuits or discrete components in new computer design is difficult, and the decision cannot be made by comparing one component approach with another because the alternative circuits seldom have the same characteristics. Therefore, the choice must be made from a total systems standpoint.

Some computer engineers feel that propagation delay times of 5 ns are essential before integrated circuits can be considered at any price level, while others feel that there is a potential price at which today's performance level becomes attractive. A third undercurrent, particularly from some designers of large high-speed systems, pushes hard for propagation delay times of one nanosecond and under.

Forecasts indicate that the price level for semiconductor integrated circuits in the next three years will shift the balance of the performance vs. cost equation. Late 1964 price quotations, for quantities approaching 100 000, are in the unit range of $3 for a three-input AND gate. On a system cost basis, such prices will be lower than for comparable circuits assembled from discrete components. Most people in the industry believe that commercial machines using semiconductor integrated circuits should be available by the end of 1965.

There is even less unanimity of feeling in the computer industry regarding thin-film circuits. Some manufacturers, believing that the long-term cost of thin-film circuits is greater, and their performance less impressive, emphasize semiconductor circuits in their development programs. Others have large internal thin-film development programs. Thin-film circuits will most certainly be applied where lower volume and closer performance tolerances make their use more economical.

Considerable work has been done in very large thin-film magnetic memories in the order of 10^7 bits. Perhaps such assemblies will achieve the "penny-per-bit" cost objective for large-size memories.

The computer data-processing industry consumed over one billion signal-level active and passive components in 1963. Fig. 16(A) expresses that usage in terms of 35

to 40 million circuit equivalents, and shows a forecast of 50 to 60 million potential circuits in 1969 and 70 to 80 million by 1974. As modular or integrated circuit performance improves, at prices competitive with assemblies of discrete components, most machines that will be introduced in 1966 to 1967 will use silicon semiconductor integrated circuits in the high-volume logic applications. Also, thin films will be in limited use. Therefore, by 1969, we can predict an integrated circuit penetration of over 18 per cent of the potential 55 million circuit market. Nevertheless, discrete semiconductor assemblies will still be employed in many high performance or low volume applications.

Beyond 1969, industry predictions are mixed regarding the direction modular circuits will take. Some designers feel that the answer will come from putting large numbers of logic circuits in a single package. Others feel that the generation of machines beyond 1969 will require totally new logic techniques to push machine economy and capability beyond the limits attainable today. By 1974 one could predict that 85 per cent of the active circuitry of commercial computers should be using modular circuits in one form or another. Considering the 8 to 10 per cent annual growth of the computer business itself, this indicates the use by 1969 of 10 to 12 million circuits per year and by 1974 of 60 to 70 million circuits per year.

Communications is the second largest part of the industrial electronics industry, and it includes diverse applications in which the use of modular circuits varies widely.

One large segment of communications includes mobile radio, microwave, and broadcast equipment. Here, linear circuitry predominates, and the potential is limited for digital applications. Present integrated circuit activity is restricted to developmental appraisals of thin-film modular circuits, although some exploration of semiconductor circuits is under way. No modular circuit equipment is yet in production, and there is little prototype development.

Since cost reduction would be the principal reason for use of modular circuits in these fields, serious design activity will begin only when performance and reliability that are equal to present circuits have been demonstrated.

Detailed cost comparisons between modular circuits and discrete assemblies are premature, as serious technological obstacles to using modular circuits remain unsolved. Hybrid thin-film circuits appear at this time to offer the best opportunity for their solution.

Commercial products making broad use of either thin-film or semiconductor circuits are not foreseen within the next five years; see Fig. 16(B). By 1974, linear semiconductor circuitry and closely controlled thin-film circuit costs should have fallen, and their use substantially increased. These communication businesses may use about 7 to 8 million circuits per year by 1974, with about the same percentage of semiconductor circuits as integrated thin-film semiconductor hybrids.

Present modular circuit activity in the aircraft communications and navigation equipment is intense. For example, Sperry Gyroscope has in production a digitalized Loran C receiver, originally designed for military aircraft, that is being sold for commercial airline use. Over 90 per cent of its active circuit elements use semiconductor integrated circuits, and hybrid circuits are used in most of the remainder. Such applications benefit from close relationship to prior military designs. The speed of adoption of modular circuits, however, is limited primarily by initial equipment cost and maintenance cost.

Manufacturers are shifting to the use of digital techniques to accelerate introduction of modular circuit equipment. Most digital circuits will change to semiconductor monolithic devices, and most linear circuits are slated for conversion to thin-film or hybrid circuits. Use of semiconductor circuits in linear areas will be limited by the high cost of low volume runs, and the inherent superiority of thin films where controlled parameters are necessary.

Semiconductor digital circuits will soon become competitive, and volume use will be heavy by 1969 as shown in Fig. 16(C). Linear thin-film circuitry is being adopted rapidly, and its use should also be extensive by that date. Within the next 10 years most aircraft communication and navigation circuitry will be converted completely to modular circuits.

The current use of modular circuits by the process and program control segment of the industrial electronics market is relatively modest but is increasing as semiconductor control circuitry displaces pneumatic, hydraulic, mechanical, and electric circuitry. Control equipment life is traditionally very long—upward of 25 years—and long-range maintenance requirements are very important.

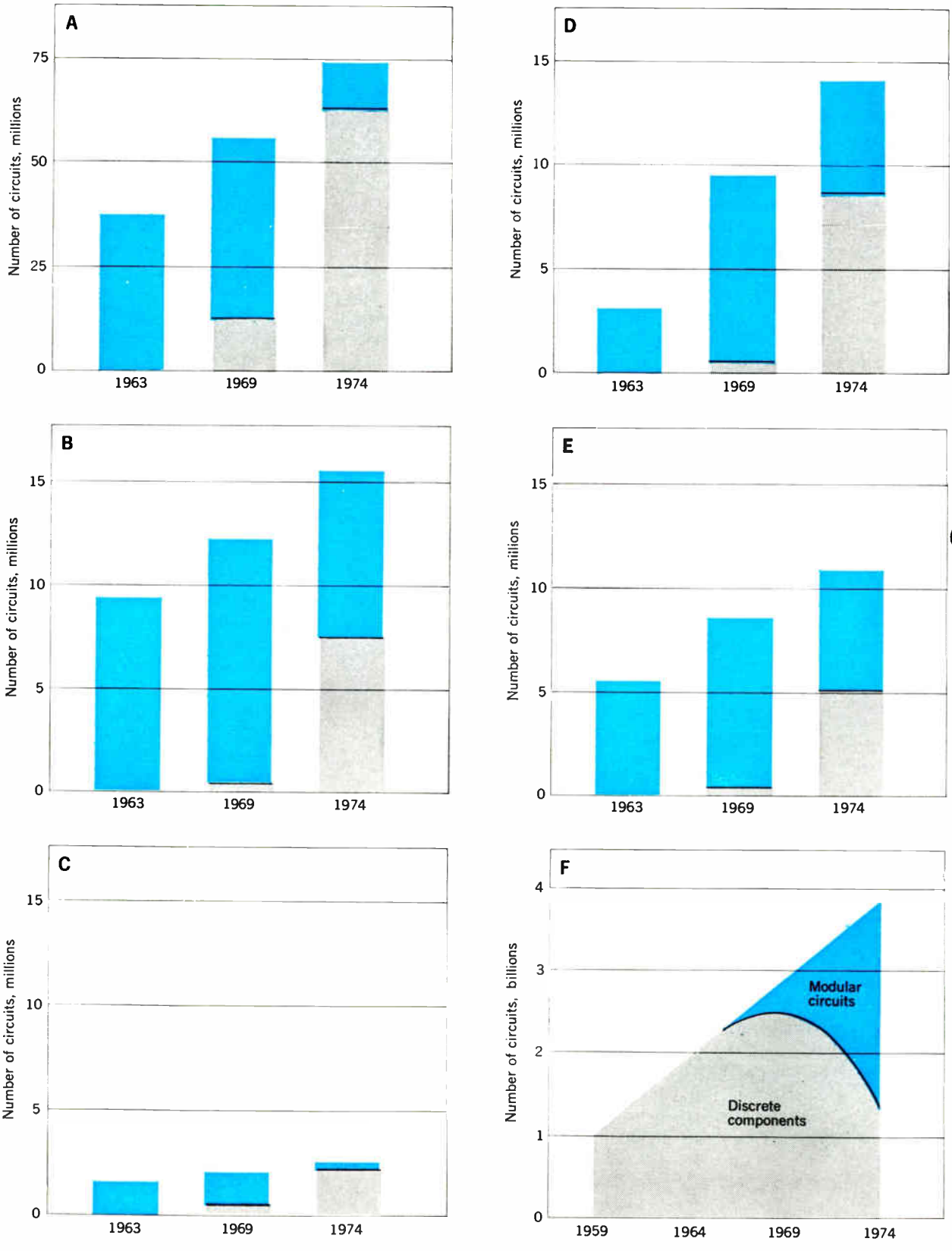
Electronic circuitry in process control has been primarily analog, whereas program control circuits are at present divided between digital and analog. In both fields there is a strong design trend toward digital circuitry. The adoption of modular circuits will depend on system reliability, ease of maintenance, and initial cost. In program control applications, where digital circuits predominate, an elementary flip-flop in the price range of $3 for a double NOR gate or a NAND gate for $2, would make semiconductor monolithic circuits less expensive than discrete assemblies. Prices at these levels are predicted for high volume during the next two years. This should trigger a substantial development activity but limited production use through the 1969 period; see Fig. 16(D). In the analog field, today's discrete component assemblies seem more practical for the low-volume requirements even though evaluation work continues on thin-film and hybrid circuits. It is doubtful whether cost will justify change to these types of circuits in the next five years.

Despite the rapid growth of the use of electronic circuits, the process and program control fields present a very limited opportunity for modular circuitry through 1969. In the subsequent five-year period, however, significant volume should appear for both semiconductor and thin-film circuits.

Both linear and digital circuits are used extensively in the test and measuring instruments field, but there is a continuing shift from linear to digital circuits where possible. No modular circuits are used in these businesses today. In addition, no significant modular circuits equipment design is under way, although some companies are beginning to make advanced development appraisals of modular circuit techniques.

As costs are critical, and production runs are limited, the shift to digital circuits may permit some standardization of components. In higher volume areas, where

Fig. 16. Industrial use of modular circuits. Potential use is shown in color; actual use in gray. Segments of the industrial-electronics market shown are: commercial computers (A), radio communications (B), aircraft communications and navigation (C), process and program control (D), and instruments (E). Industrial-electronics component requirements are shown in (F).



standardization is feasible, semiconductor circuits may equal the cost of conventional discrete assemblies within the next few years. For linear circuits, and digital circuits wherein tight specifications are required for passive elements, there is less probability for modular-circuit use.

The instrument business will move to modular circuits only when costs indicate a definite opportunity to save money. Higher volume digital applications will move to semiconductor circuits late in the first five-year period, Fig. 16(E), and more heavily in the second five-year period. Only minor penetration of lower volume digital circuits and linear circuits by modular devices is expected by 1969. Total use of about 3 to 4 million circuits is expected by 1974.

Having separated industrial electronics into its constituent parts, one now is faced with the temptation of adding together the apples, oranges, and pears into an industrial electronics fruit salad.

The overall impact of modular circuits on industrial equipment is determined by modular circuit costs vs. discrete assembly costs, and the equipment and design turnover cycles in each industry. For the many applications in the industrial electronics field, semiconductor integrated circuits will be very cost competitive during the next five years, and should have a clear competitive edge by 1974 in all except the smallest volume applications. Thin-film usage will be more limited both in time and quantity but, in volume and linear applications requiring close parameter control, a significant market for thin-film circuits should emerge.

Figure 16(F) shows the predicted use in the total industrial electronics market of both passive and active components with the scale now expressed in terms of equivalent discrete components. The overall collection of businesses which make up industrial electronics has been growing at about 12 per cent per year. The ever increasing demand for industrial data processing and automation will continue this overall growth rate. The cost potential of modular circuits and the demonstration of their reliability and performance capability will bring about very impressive inroads for modular circuits during the last five years of the next decade.

Integrated circuits in consumer products

J. E. Brown, Vice President—Engineering, Zenith Radio Corporation

Modular microcircuitry will move into every section of the consumer electronic goods field and perhaps into some areas that are not considered primarily electronic.

The foremost practical usage of the new electronic concepts is in the hearing aid. Small size, light weight, and a degree of reliability and stability much higher than for conventional electronic gear are required for this instrument. Integrated circuitry more nearly meets these specifications than any other system. The problem of making transducers that are more compatible with microcircuits remains to be solved.

In radio or television applications, we must either connect elements of microcircuitry into arrangements with tuned circuits, or we must discover alternatives to our conventional methods. For example, the phonon-amplifier approach may be effective in the intermediate-amplifier stage of a superheterodyne receiver.

One can envisage extremely small radio receivers limited in size only by the problems of the audio-reproducing device and the tuning mechanism. We have solid-state systems right now that perform the function of the conventional inductance and variable capacitor.

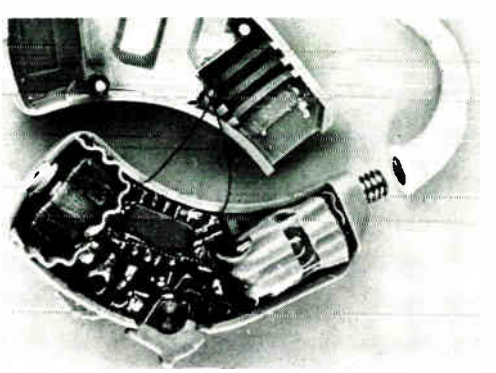
We are studying the problems of building power-handling capabilities into microcircuits, and the accompanying problem of disposing of the heat generated. An example of this would be in audio-output circuitry.

A difficult challenge is presented to the circuit and device designer when he tries to use nothing but microcircuits in a television receiver. Much circuitry is involved in the handling of high-voltage waveforms. Clipping functions are performed as well as various wave-shaping operations. Therefore, some of these methods must be reversed so that current waveforms can be used rather than voltage waveforms.

Perhaps many of the circuits in a television receiver, including the deflection system, can now be built into the base of the picture tube.

Beyond the use of microcircuits in the radio and television fields, there are many other applications in prospect. The problem at this time is somewhat more economic than scientific: reduction of cost is the key to the opening of vast new fields of application.

The hearing aid is the best and perhaps most obvious outlet for integrated circuits in consumer products.



Research and development in integrated circuits

C. Harry Knowles, Manager,
Molecular Electronics Division, Westinghouse Electric Corporation

The integrated-circuits man is pre-empted in research and development by either device technologists or circuit engineers. Conversely, research and development in integrated circuitry must cover the entire field of research and development in electronics to keep up with competition's marketing.

The electronics field has been organized into four groupings: materials, devices, circuits, and systems as shown in Fig. 17. The entire field may be conceived as a stratified stream, moving with time, with the four groupings moving generally together. There is, however, a two-year lag between a development in one field and its subsequent usage in the next. Therefore, the entire cycle is such that from the time a new material or technique is introduced, a system using that material will follow approximately six years later. What has been one man's R & D has become another man's production.

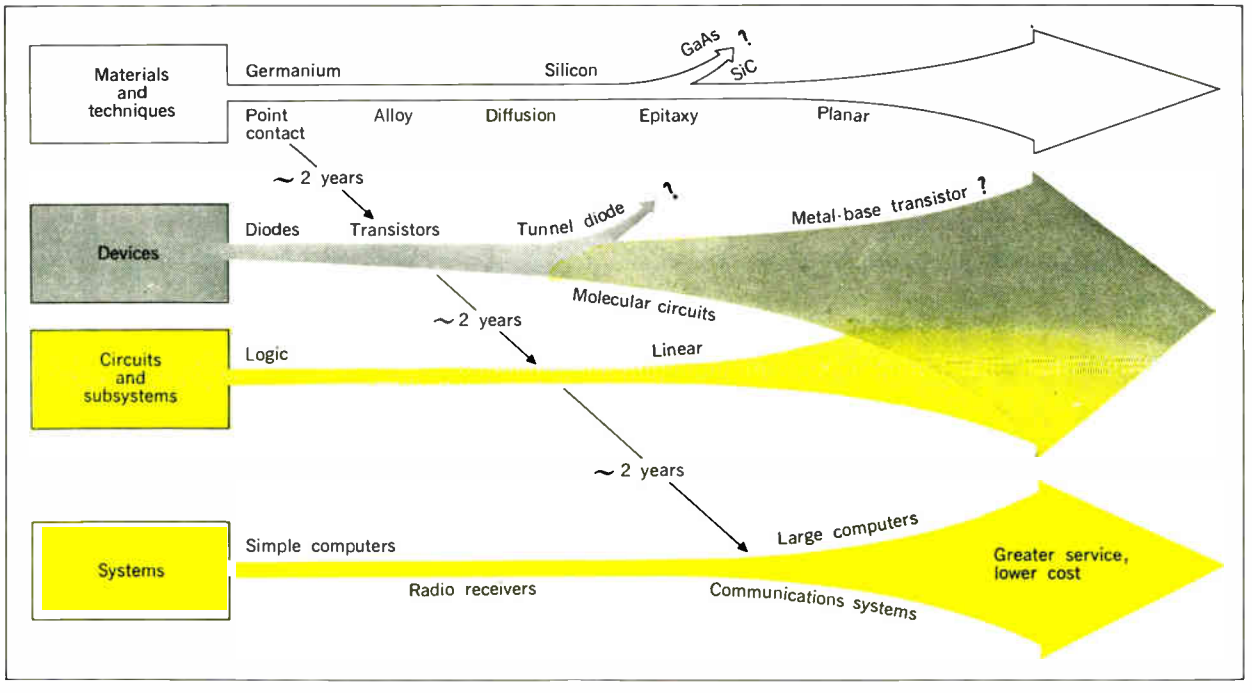
Another factor, already mentioned, is the fusing together of the devices and circuits and subsystems fields. The circuit designer is now in the position of being a device designer and vice versa. The six-year lag between material and system, therefore, will be shortened by at least two years.

Because of the extreme sensitivity of production costs to volume, it is important for the systems manufacturer and the device manufacturer to jointly develop and use circuit types that are mutually compatible to each other. Both groups must make certain that they follow the mainstream of developments in materials through systems. In the mainstream of devices have been alloy transistors, diffused and mesa transistors, and then silicon-diffused epitaxial planar transistors. The rewards to the systems manufacturer who settled on a popular device in the mainstream have been very large. Conversely, diversions away from the mainstream have consumed enormous economical resources and technical talents. These diversions must be selected and monitored carefully. Some examples of traps away from the mainstream have been silicon carbide, gallium arsenide, and the tunnel diode.

The primary course of integrated circuitry over the next several years will be in circuits made from silicon monolithic substrates with extensive use of p-n junction transistors and diodes. A combination of diffused and deposited inactive elements over the passivating silicon dioxide layers will be used and their development and production will be in flat packages of about a dozen leads. Considerable R & D will go into more complex circuit blocks.

The complexity problem facing both designers and users of integrated circuits is that cost increases as each component becomes more complex, as the number of components on a block increases, and as the compatibility mix of components on a block increases its processing difficulty. In each case, as the technology improves, the cost decreases. For example, when better ways of putting p-n-p and n-p-n transistors on a given

Fig. 17. The entire electronics field may be conceived as a stratified stream.



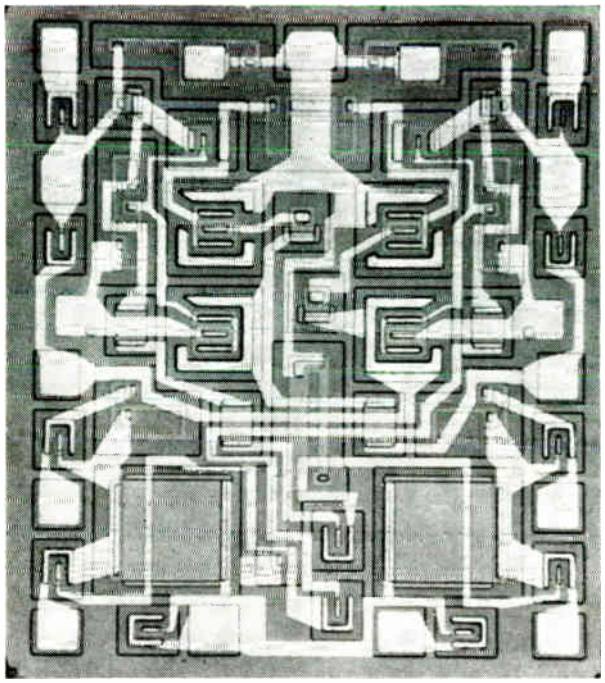
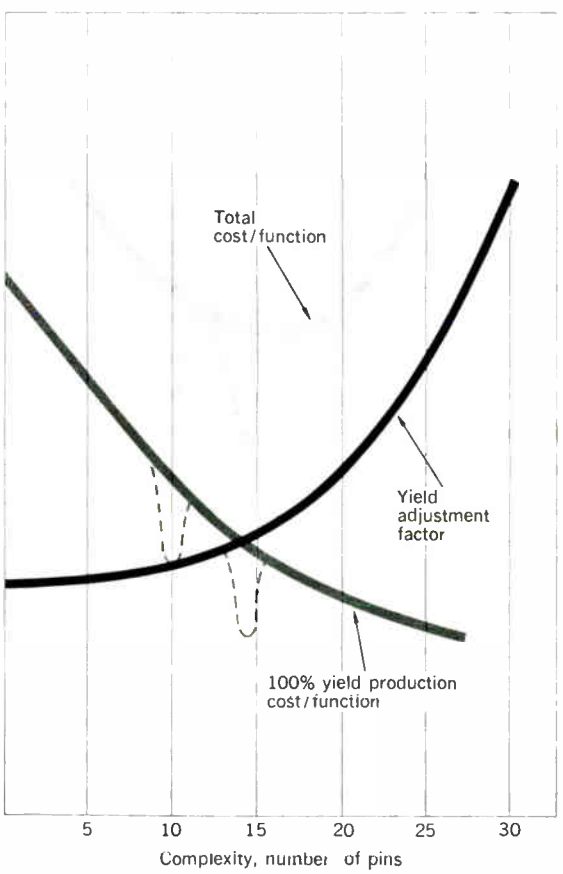


Fig. 18. The current level of practical complexity is illustrated by this gated flip-flop or pulse binary. It contains 45 components with about 12 to 14 leads. The resistors and capacitors are diffused; the transistors are n-p-n units.

Fig. 19. Cost plotted as a function of complexity as evidenced by the number of pins in a package. Total cost per function is a minimum at a complexity of 10 to 14 pins.



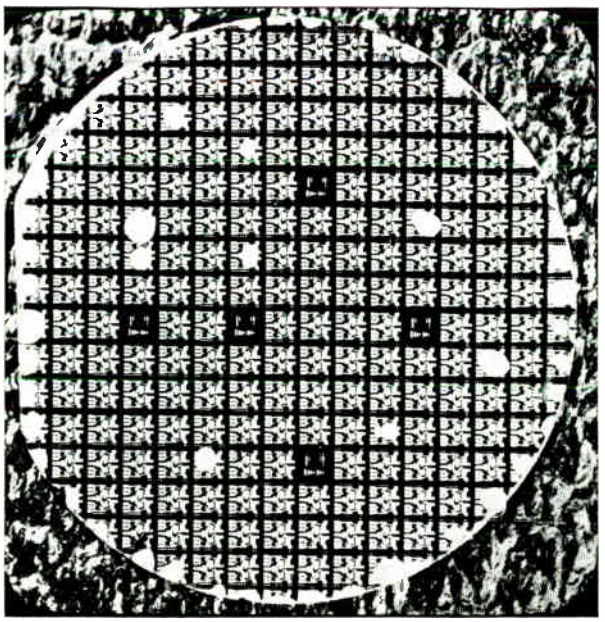


Fig. 20. This one-inch wafer contains 200 double gates.

block are found, yield improves and cost drops. Also, as yields improve, more components can be put on a given block.

Figure 18 shows the current level of practical complexity. The component illustrated is a gated flip-flop or pulse binary. It has about 12 to 14 leads and 45 components. The resistors and capacitors are diffused and the transistors are n-p-n units.

Mechanical configuration is a major problem for the combination device-circuit-system field. Figure 19 depicts cost vs. complexity as based on the number of pins in a package. One curve shows what the cost per function would be if yield were always 100 per cent. This curve is fairly independent of time, but is dependent upon mechanization savings. Where there are localized mechanization savings, there is a downward perturbation in the cost curve indicated by the dashed line. The yield adjustment factor curve is dependent on the processing technology that produces the devices and has a knee that moves outward as a function of time. The total cost per function is the sum of these two curves, as shown, and has a minimum complexity of 10 to 14 pins. Note that dips occur as a result of savings through mechanization.

Figure 20 shows what is felt to be the state of the art in complexity in wafer form. This slice is about one inch in diameter and consists of about 200 double gates. Each double gate has about six transistors, about 12 diodes, and a handful of resistors. In the wafer form, each circuit is completely tested by probes, which touch the aluminum bonding pads. If a circuit is bad, a spot of white ink is marked upon the defective circuit. It can be seen from Fig. 20 that there are only a couple of dozen defective circuits, most of them around the edges of the wafer, as would be expected. There are in the middle, however, several randomly placed defective units.

Since the principal integrated circuit interest is in digital computers, let us look at logic speed. Figure 21 shows the average propagation delay per stage as a function of years for both saturated logic and nonsaturated

circuitry. During 1964, a saturated logic time in the 10-nanosecond range will probably be achieved—and perhaps even a lower value. Speed has doubled every year over the past seven years on the average. Greater speeds depend upon mechanical tolerances. If transistors, diodes, resistors, etc., can be made smaller, the use of smaller parasitic capacitances implied thereby means that the speed of the circuits will be increased.

Where we stand now with respect so size can be illustrated by the fact that a commercially available block now in production has resistors 8 microns wide with 7 micron tolerances on sets of adjacent maskings in the transistors. This block operated in the 20-ns propagation delay range. Experimental blocks using the same tolerances and components operate in the 12- to 15-ns range.

Electron-beam technology as a machining tool in integrated circuitry offers several advantages and may reduce the size of the devices by another order of magnitude. Mechanical tolerances, which are currently controlled by masks, will soon be controlled by sensing prior operations with the electron beam and, after sampling the device characteristics, the pattern will be inscribed.

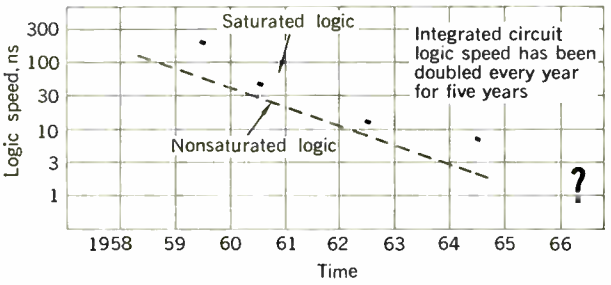
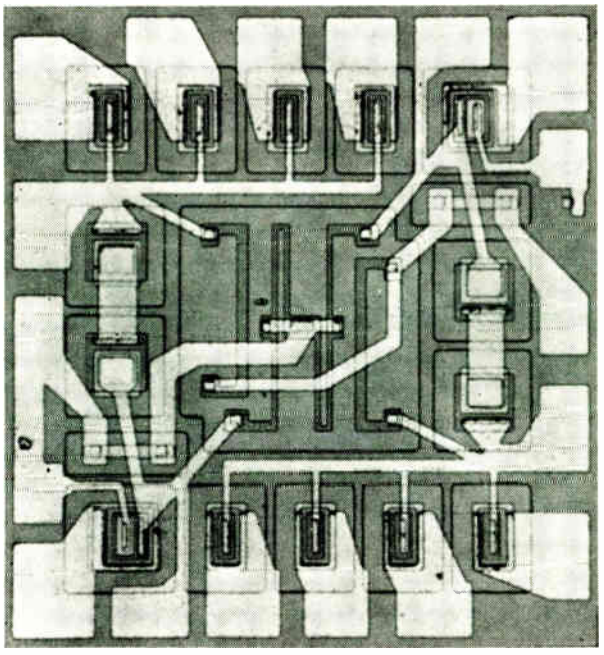


Fig. 21. Average propagation delay per stage as a function of years for both saturated logic and nonsaturated circuitry.

Fig. 22. This one-inch silicon wafer contains about 800 gates per square inch using 8-micron tolerances.



Masking tolerances are now in the 7- to 8-micron range, with a 15-ns saturated logic. We can soon expect 3- to 4-micron tolerances and 5- to 7-ns saturated logic. Electron-beam sensing and machining will permit 0.1-micron tolerances and subnanosecond logic speeds.

Parenthetically, it is interesting to extrapolate on what a one-inch silicon slice would produce if the size of individual components were decreased by a factor of 15. Figure 22 shows an integrated circuit which may be fabricated on a silicon wafer at a density of about 800 gates per square inch, using 8-micron tolerances. If these tolerances were reduced to 0.5 micron, the area of the circuits would be reduced by slightly over 200. The number of circuits per square inch would then be about 16 000. In addition, the gates would operate 15 times faster or in the 1-ns delay range. Such a wafer would have a logic power about 100 times greater than today's largest computer!

In general, integrated linear circuits are limited to frequencies of several dozen megacycles or less and to circuits, which in numbers of components, are generally smaller than their digital brethren. A fruitful field for integrated linear devices is the differential-amplifier area and the closely corresponding dc amplifier field. Here the device designer can take advantage of carefully matched transistors and resistors which, in absolute value, may not be reproducible from one block to another but where the tight balance between sets can be very tight.

Fig. 23. An integrated differential-amplifier circuit typical of several available commercially to systems users.

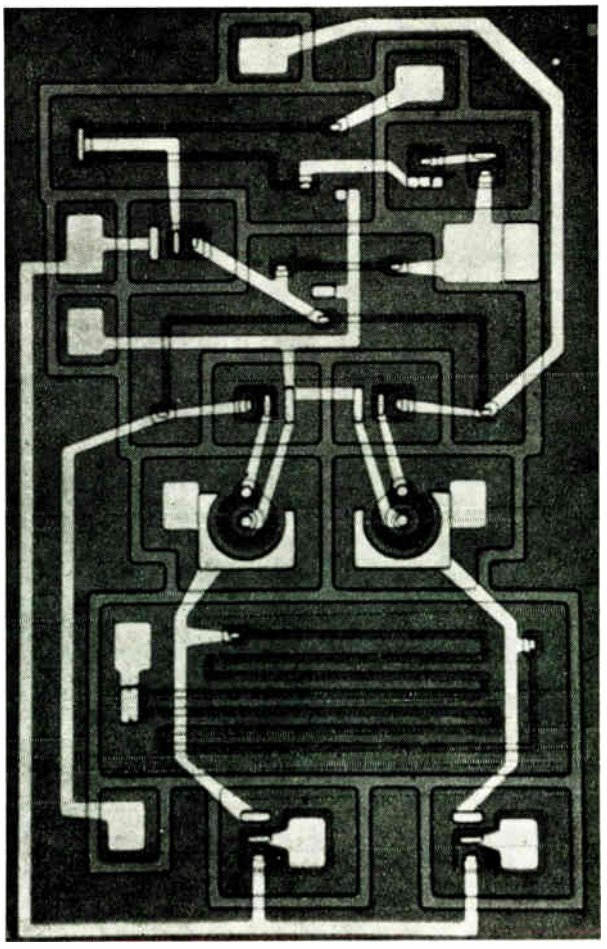


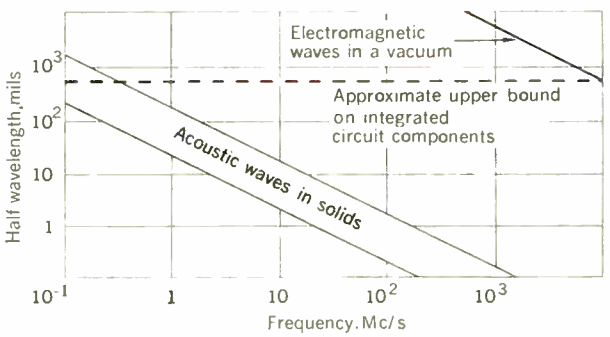
Figure 23 shows a differential amplifier typical of several available to systems users today.

Tuning is a popular topic in integrated circuits and many ways of achieving frequency selectivity have been examined recently. A graphic curve, Fig. 24, shows that if we plot resonant frequency vs. half-wave length in mils, frequency selectivity for sizes compatible with integrated circuits is limited to 100 kc/s or more. Thus, tight mechanical tolerances for high-resolution frequency selectivity will become critically important.

With respect to high-frequency performance, the metal-based transistor appears to be the most promising new device, and recent analyses indicate that it may improve speed performance with equivalent mechanical tolerances by a factor of ten. However, metal-based transistors may have to be aimed at the 100 Gc/s range or better to compete favorably with p-n junction transistors that can probably be designed up to about 50 Gc/s.

Another new device for low-frequency operation may be the metal–oxide semiconductor transistor with almost infinite input resistance. Very low power, slow speed, logic elements, and amplifying elements will probably be developed that will use these devices.

Fig. 24. Frequency selectivity for sizes compatible with integrated circuits is limited to 100 kc/s or more.



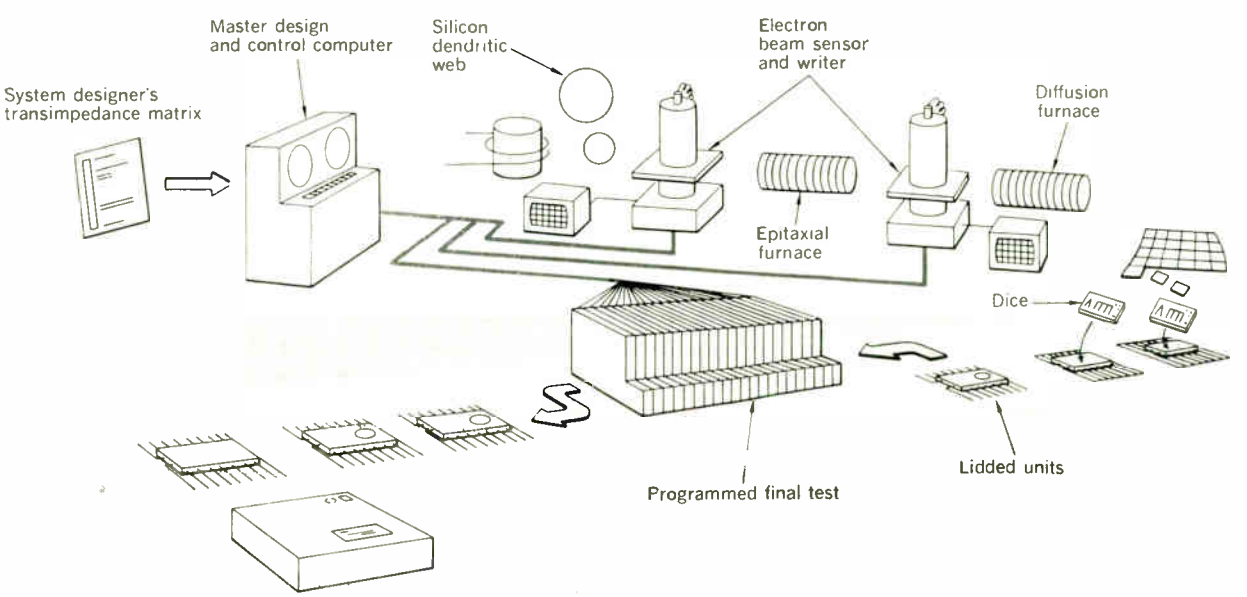
Finally, some promising devices are being developed that use light coupling between molecular elements. Such elements offer virtually complete isolation between one circuit and another—as in a demodulator chopper.

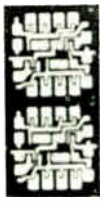
An important consideration in choosing an approach for advanced development is a look ahead at what the consequences may be. In this case, let us examine the possible design and manufacturing cycle of a typical circuit circa 1970, as shown in Fig. 25.

The customer presumably has determined his n-terminal transimpedance matrix which is programmed into a master design and control computer. The computer lays out the functional block design according to modular packaging. The design is converted into a sequence of writing patterns on electron beams. The electron beams perform functions similar to the photographic masks of today. Dendritic silicon web is pulled from a puddle of continuously replenished silicon. The dendritic web passes through a sequence of oxidation furnaces, electron-beam inscriptions, epitaxial growth and diffusions, plus metalizations and the depositions of glass, providing dice at the end of the line, which are completely passivated and tested. These dice are then automatically inserted into similar automatically generated packages, transported under a modular test arrangement which is programmed by the computer. The units are then marked automatically by a marking machine controlled by the computer. The computer has, of course, provided packages, shipping labels, and billing tickets for the customer, as well as performing the usual production control and inventory functions automatically.

Thus, the implication of a fantastically low cost per function appears to be accurate. The shift of problems from materials and devices into systems engineering is emphasized. Our problems as electronics engineers will be to generate requirements for systems at a rate to match the supply of circuitry that the semiconductor industry is able to supply over the next several years.

Fig. 25. Possible design and manufacturing cycle of a typical circuit circa 1970.





The economic impact of integrated circuitry

Patrick E. Haggerty, President,
Texas Instruments Inc.

Let us look at what the integrated circuit market probably was in 1963, as shown in Table V. As applied to semiconductor networks, these are EIA numbers—for all other circuits, out best estimate. Research and development contracts are not included. Remember that we are defining integrated circuitry as "The physical realization of a number of circuit elements, inseparably associated on or within a continuous body to perform the function of a circuit." Under the terms of this definition we exclude RC networks and similar assemblages of completely passive equivalent components.

The total 1963 integrated circuit market reached approximately 950 000 units with a value of $21.4 million. These data, however, give no indication of the rate of change, particularly in the semiconductor network area. Nearly one half of the 515 000 semiconductor networks were shipped during the fourth quarter of the year, and this quantity was four times greater than that shipped in the first quarter—with dollar values trebling over the same period. But even this does not give the total growth picture in this area. At the end of 1962, about 30 types of semiconductor networks were produced, while at the end of 1963, approximately 132 types were made. Looking at it another way, the number of equivalent components in integrated circuits ranged from just three or four per integrated circuit up to a maximum of 69. The 915 000 integrated circuits are probably equivalent to something in the neighborhood of 7.5 million individual discrete components.

Table VI shows an estimate of what the U.S. market for electronic equipment probably will be during 1973. Our best estimate, excluding the replacement parts market, is that the electronic market in 1973 will be about $20.1 billion.

We analyzed the circuit cost of this electronic equipment by studying many typical equipments in the areas of Government, industry, and consumer electronics. Including corresponding engineering costs, and immediately associated overheads, we arrived at total circuit costs as a percentage of the total cost of the equipment in each area. These costs suggest that the aggregate cost of circuits utilized in electronic equipment during 1973 will be about $5.8 billion, or 30 per cent of the total market.

A similar analysis for 1963 indicates that total circuit costs approximated 25 per cent of equipment sales, or a total of $3.6 billion. For the moment, let's accept the statement that we had $3.6 billion in total circuit costs in 1963 and that this figure will go to about $5.8 billion by 1973.

Next we broke down the total circuit costs around active element groups (an AEG consists of an active element plus its associated circuitry, including diodes, resistors, capacitors, relays, inductors, coils, connectors, printed circuit boards, etc.). We then determined the number of active elements on the basis of the total number of transistors; plus the total number of silicon controlled rectifiers; plus three eights of all other rectifiers, reflecting our judgment of their use in bridges; and 170 per cent of the number of receiving tubes going into new equipment, which reflected our best judgment of the equivalent usage of multielement tubes.

If we assume that 1973's electronic equipment market is being generated entirely by today's conventional electronic circuitry, the number of AEGs in the total 1973 market of $20.1 billion is estimated at 1.73 billion, as shown in Table VII. From the examination of typical circuits, we estimated that 80 per cent of such circuits in the Government electronics area could be handled in 1973 by integrated circuits; with 75 per cent in the industrial area, and 70 per cent in the consumer areas—for an overall average of about 75 per cent. Multiplying these percentages by the total number of AEGs for the elec-

V. 1963 integrated circuit market

	Units, thousands	Dollars, thousands
Semiconductor networks	515	15826
All other circuits	400	5600
Total integrated circuits	915	21426

VI. Comparative circuit costs

	Government	Industrial	Consumer	Total
		1973		
Electronic equipment sales, billions of dollars	10.7	5.7	3.7	20.1
Circuit costs, per cent of total equipment	30—	30+	25	30—
Cost of circuits utilized in electronic equipment, billions of dollars	3.1	1.8	0.9	5.8
		1963		
Cost of circuits utilized in electronic equipment, billions of dollars	2.2	0.8—	0.6	3.6—

tronic equipment volume estimated, gave a total of 1.28 billion AEGs technically replaceable by integrated circuitry in 1973.

Similar estimates for 1963 indicate that there were 110 million AEGs—or 15 per cent of the total—which were technically replaceable by integrated circuits during 1963.

The 1973 total circuit values, Table VIII, will be $2.44 billion in the Government category, $1.33 billion in the industrial category, and $650 million in the consumer category—for a total of $4.42 billion.

In terms of these same typical circuits and the AEGs making them up, we attempted to identify the proportion of such AEGs which could not be handled just as economically with integrated circuits as with conventional components. In other words, we assumed that for them to be technically replaceable was not enough. If they were to be considered, the cost of integrated circuits would have to break even with that of conventional components. This sort of dreaming gets double-barrelled because you must project probable manufacturing costs of integrated circuits ten years ahead, as you think they will then exist, and simultaneously assume the capture of varying percentages of the market. If you're wrong about the percentage of the market captured, then the average costs go up and the percentage goes down still further. However, our analysis suggested that a very large percentage of all technically replaceable AEGs would also be economically replaceable. That is, the integrated circuits would cost no more than conventional circuits.

As Table VIII indicates, we estimated that 95 per cent of the AEGs technically replaceable in the Government area will be replaceable in 1973 at no higher a price than that of conventional circuitry. Our percentages in the industrial area are 90 per cent, and in the consumer area 70 per cent—for an overall average of just below 90 per cent. Applying these percentages to the dollar values at at the top of the Table VIII, one arrives at a potential value of conventional circuits, both technically and economically replaceable by integrated circuits, assuming an even break in cost. These numbers are $2.3 billion in the Government area; $1.2 billion in the industrial area; and $0.5 billion in the consumer area—for a total of $4.0 billion.

Even though conventional circuitry can be replaced technically, and at no additional cost, by integrated circuitry, there is no assurance that it will be completely replaced. Market and tooling considerations, obsolescence, differences of opinion, human judgment, etc., produce the "inertia lead-time factor." To determine this, we assumed that integrated circuitry would be a part of new equipment designs; then we estimated the percentage of the equipment being manufactured in 1973, with these new designs. This gave us the inertia lead-time factor, expressed as a percentile; see Table IX. On the basis that 80 per cent of the circuits (both technically and economically replaceable by integrated circuitry in the Government area in 1973) are now generating the potential pressure for conversion to integrated circuitry, 55 per cent of the industrial area circuits, and 90 per cent of the consumer area circuits are exerting the same pressures, for an overall average of 75 per cent. Thus, a potential value of about $2.9 billion in conventional circuits exists as the "pressure" for conversion to integrated circuits.

Analyses suggest that integrated circuitry will actually cost from one third to two thirds as much as conventional circuitry. If we assume that integrated circuitry will cost

VII. Comparison for active element groups (AEG)

	Government	Industrial	Consumer	Total
		1973		
Number of active element groups, millions of units	385	680	665	1730
Potential technical penetration of conventional circuits by integrated circuitry, per cent	80	75	70	75—
Potential number of AEGs technically replaceable by integrated circuits, millions of units	308	510	466	1284
		1963		
Potential number of AEGs technically replaceable by integrated circuits, millions of units	33	44	33	110

VIII. Predicted potential value of integrated circuits for 1973

	Government	Industrial	Consumer	Total
Potential value of conventional circuits technically replaceable by integrated circuits, millions of dollars	2440	1330	650	4420
Portion of conventional circuits economically replaceable by integrated circuits at break-even point, per cent	95	90	70	90
Potential value of conventional circuits both technically and economically replaceable by integrated circuits at break-even point, billions of dollars	2.3	1.2	0.5	4.0

IX. Predicted potential savings with integrated circuits for 1973

	Government	Industrial	Consumer	Total
Potential value of conventional circuits both technically and economically replaceable by integrated circuits at break-even point, billions of dollars	2.3	1.2	0.5	4.0
Value adjusted for probable 'inertia lead-time' factor, per cent	80	55	90	75
billions of dollars	1.8+	0.7—	0.4+	2.9
Potential value of integrated circuits, per cent	50	50	50	50
billions of dollars	0.9	0.4	0.2	1.5
Potential savings resulting from use of integrated circuits at circuit level billions of dollars	0.9	0.3	0.2	1.4
Potential savings resulting from use of integrated circuits at both circuit and chassis fabrication levels, billions of dollars	1.4—	0.4	0.2	2.0

one half as much as the conventional circuitry it replaces in 1973, the actual potential value of the integrated circuitry is about $900 million in the Government area, $400 million in the industrial area and $200 million in the consumer area, for a total of $1.5 billion.

This suggests that $2.9 billion worth of conventional circuitry in the 1973 electronic equipment market of $20.1 billion can be replaced with $1.5 billion of integrated circuitry, with a potential saving of $900 million in the Government area, $300 million in the industrial area, and $200 million in the consumer area, for a total of $1.4 billion.

Still further savings in hardware, chassis fabrication, and wiring may be achieved. These were estimated to be about $500 million in the Government area, bringing the total savings to $1.4 billion; about $100 million in the industrial area, increasing the savings to $400 million; in the consumer area our estimates were below the limit of accuracy, so we left the total of $200 million. Thus, total potential savings to the electronics industry through the use of integrated circuitry in 1973 is estimated at about $2.0 billion.

Does this mean that the market actually generated for integrated circuitry in 1973 will be $1.5 billion? Probably not. However, the $2.0 billion saving must be considered as the pressure that is forcing conversion of electronic equipment to integrated circuitry. If one analyzes the industry today and attempts to project facility and technological growth and the effectiveness with which it will seize the opportunity presented, the conclusion is reached that the actual market to be developed for integrated circuitry by 1973 may be between $500 million and $1 billion.

I suggest that $750 million in 1973 is as good a number as any and this is exactly one half the $1.5 billion potential. It does not seem at all unrealistic to me that one half of such a potential will indeed be reached. This line of reasoning also suggests that about half of the 2.0 billion potential savings to the industry may well be achieved by 1973. Should this occur, the $20.1 billion electronic equipment market with which we started will have become a $19 billion market. It may be, however, that the improved reliability; decreased size, weight, and power consumption of integrated circuitry; and the marked economic gains expected; will expand the use of electronics sufficiently to generate a larger total market. There is little question that this will be the case over a still longer span of time. If a $1 billion saving is indeed developed, about 70 per cent of it will unquestionably accrue to the Government area, which would represent a healthy 7 per cent decrease in the total cost of Government electronics in 1973. Development of integrated circuitry in electronics seems strongly parallel to the development of the jet engine in the aircraft industry, where for the first time improved performance has been accompanied by sharply lower costs.

A $1.5 billion potential for integrated circuitry in 1973, generating an actual market of $750 million, may sound astronomically out of reach. Can the industry in a short ten years generate anything approximating that amount? From a look at the past, I think it may indeed be possible to do so.

The potential can be viewed in still another way. Compare the original transistorized computer for the Minuteman missile and the new integrated circuit computer for the Improved Minuteman. Compared to the Minuteman I computer, the new unit is 33½ pounds lighter (36.5 vs. 70 pounds), occupies about 40 per cent less space, uses less than half as many individual parts (5510 vs. 14 711) and operates on little more than half as much power (195 vs. 350 watts). And functionally, the new computer is more than twice as capable as the original one.

As we conclude, we should refocus on the fact that the potential for gains is still very large and that the potential, in itself, will ensure continued rapid development of integrated circuitry.

In the Improved Minuteman computer, 95 per cent of the electronic functions are performed by the 1895 integrated circuits it contains. The 36.5-pound computer is a miracle of compactness. Yet, the 1895 silicon wafers that perform 95 per cent of its functions weigh only 2.5 grams. Essentially all of the total weight is made up of the few remaining conventional components, the packaging, the interconnections, the plugs, the sheet metal, and the case.

Yes—miracle though it is, at 36.5 pounds we still have a long way to go, because 95 per cent of the functions are being performed by these 2.5 grams. The long-term implications are, I believe, obvious.

Monolithic integrated circuits

A single-crystal chip of silicon contains both active and passive elements. Parasitic capacitance introduced during monolithic construction and circuit performance can now be analyzed

A. B. Phillips *Motorola, Inc.*

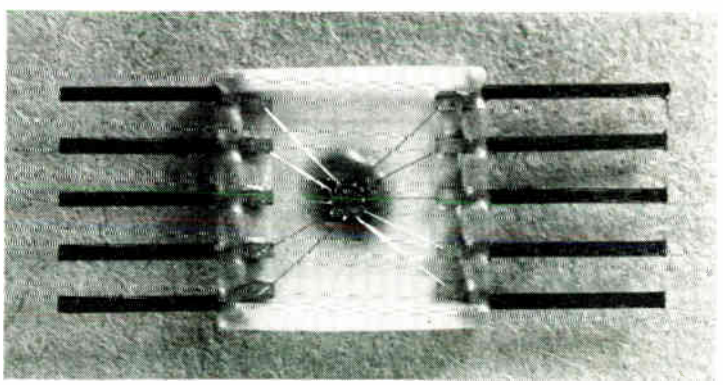


Fig. 1. Example of a monolithic DTL circuit in a flat package (protective covering removed).

Monolithic integrated circuits derive their name from the fact that the entire integrated circuit is formed in a single-crystal chip of silicon semiconductor. The single-crystal chip technique is a main factor distinguishing between monolithic integrated circuits and thin-film circuits.

The advent of monolithic circuits has created a need for knowledge on the part of engineers who design circuits from this modern approach and also other engineers who wish to keep abreast of the technology. Even though there is no completely universal set of guidelines at this time, certain basic design rules and technology limitations with which the integrated circuit maker has to live can be outlined.

To demonstrate the theory and design approach used for these circuits, a typical monolithic digital integrated circuit, such as the completed one shown in Fig. 1, will be used as an example. The steps involved in translating this typical circuit design into a final monolithic integrated circuit will be described. In this process, the highlights of integrated circuit theory, design relationships, processes, and parameters will be developed, covering almost all contemporary applications of monolithic circuits to digital computers.

The DTL digital circuit

The schematic of a high-speed NAND gate is shown in Fig. 2. This is a widely used circuit form in computer systems. The reason for the popularity of diode-transistor logic (DTL) is that, in discrete form, diodes are consider-

Parasiticless integrated circuits

Just prior to the IEEE International Convention in March, Motorola revealed its EPIC process for manufacturing monolithic silicon integrated circuits. This new approach is said to eliminate for all practical purposes the parasitic effects that are now limiting the frequency response and circuit speed of integrated circuits.

Other advantages ascribed to the new technique are (1) simultaneous diffusion of both high- and low-frequency transistors is permitted within the same substrate through selective gold diffusion; (2) fabrication of both n-p-n and p-n-p transistors is readily permitted within the same semiconductor substrate; and (3) fabrication of transistors with much higher gain–bandwidth products is allowed.

Dr. C. L. Hogan, General Manager of the Motorola Semiconductor Products Division, in making the announcement said, "Within a matter of months, devices utilizing this process should be coming off the production lines."

Cross section of a typical EPIC structure shows that circuit elements are completely isolated from each other, and from the substrate, by an insulating layer of material that is said to provide both electrical insulation and protection against peripheral impurity penetration.

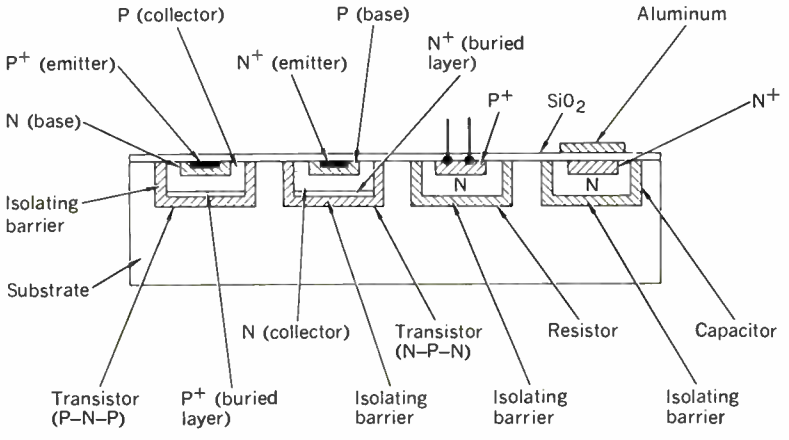
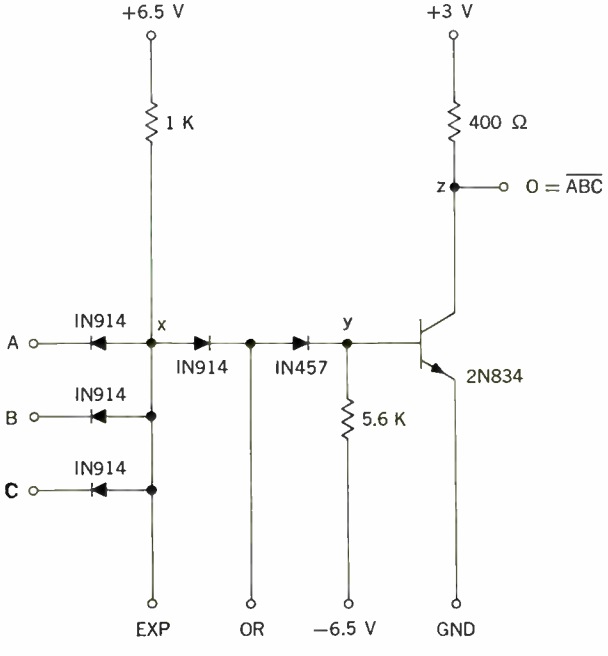


Fig. 2. A typical DTL NAND gate. Average propagation delay is less than 10 ns. When any of the inputs is returned to ground, the gate is off. The gate is turned on when all three inputs go up to +3 volts.



ably less expensive than transistors. This particular circuit consists of an AND (more than one input terminal) diode gate in series with a NOT (an inhibitory circuit) transistor inverter, the combination of which performs the NAND (an AND gating circuit which inverts the pulse phase) function. Thus, an output occurs only when all of the inputs are turned ON.

Figure 2 represents a high-speed circuit that has an average propagation delay of less than 10 ns. Using a simplified breakpoint voltage of 0.7 volt for the various diodes in this circuit, we can then ascertain the significant voltage potentials at points x, y, and z. When any of the inputs is returned to the ground, the gate is OFF, and the voltage at x is clamped to the ground through one diode drop—or 0.7 volt. Thus, there is no base current drive for the transistor, and the output voltage at z is approximately 3 volts. Voltage point y would then be at −0.7 volt, or *two* diode drops below x. This conduction keeps the emitter–base junction of the n-p-n planar silicon-epitaxial transistor in a reverse-biased condition. In the OFF condition, the current through the input diodes is 5.8 mA.

The gate is turned ON when all three inputs go up to +3 volts, at which time both points x and y rise in potential and provide base current to saturate the transistor. Consequently, the base of the transistor goes up to 0.7

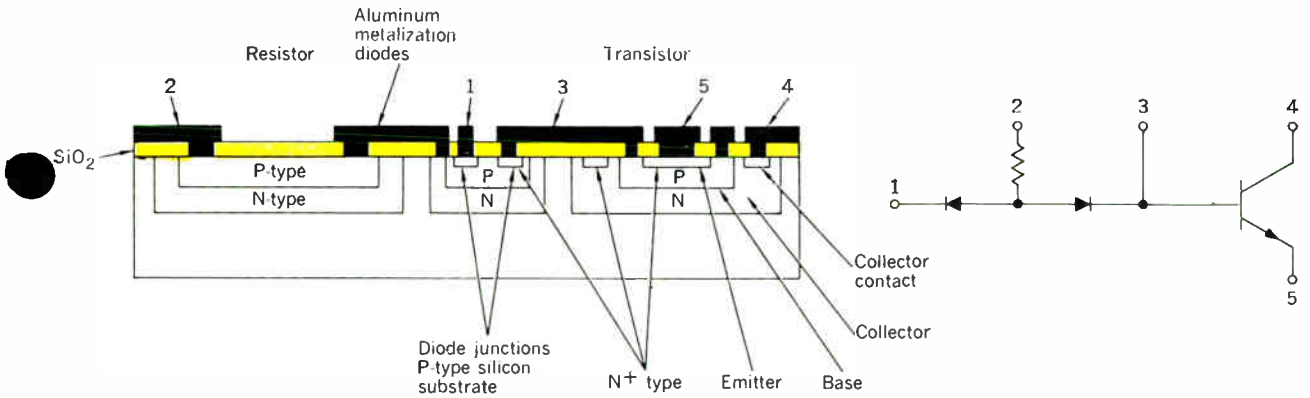


Fig. 3. Cross-sectional view of a typical integrated circuit.

volt in conduction, and therefore point x will rise also by two diode drops, or 2.1 volts. Thus, the input drive at the transistor swings by two diode drops, or 1.4 volts magnitude. The output, of course, goes into saturation to the extreme of approximately 0.24 volt.

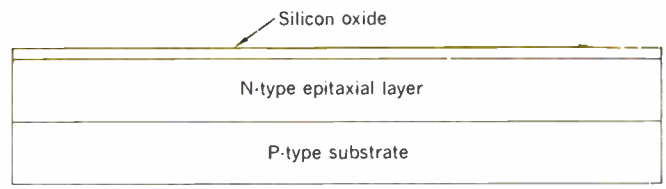
This basic switching circuit will be used to illustrate what factors must be considered in the attainment of a monolithic version of the circuit.

Circuit structure and processes

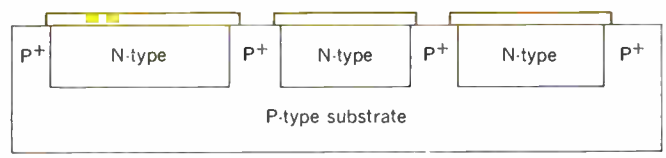
The term "monolithic" means that the entire integrated circuit is formed from a single-crystal chip of silicon semiconductor. The chip itself, shown in Fig. 1, contains the various active and passive circuit elements such as the resistor, diode, transistor—and in some cases—the capacitor. Figure 3 is a sectional view of these elements in a typical monolithic chip. It illustrates the four-layer structure, how each element is electrically connected by metalization, and how each element is electrically "insulated" with isolated junctions. In this case, all of the components are formed in an n-type epitaxial region that was grown on a p-type substrate. The resistor comprises the resistance of a p-type diffused region; the diodes consist of junctions formed by diffusing two n-type regions into a common-p background. Finally, the transistor is a typical n-p-n structure, except that the connections to the collector region are made at the top surface.

The general process sequence for achieving this structure is shown in Fig. 4. The total process sequence consists of a number of photoresist and diffusion steps that utilize the fundamental property of silicon oxide for prevention of impurity diffusion. Thus, the process starts with the deposition of an n-type epitaxial layer on a p-type substrate that has a resistivity of 10 ohm·cm. The n-type epitaxial region can be programmed to have resistivities anywhere between 0.1 and 0.5 ohm·cm, as desired. This n-type region is isolated into individual islands by diffusing a heavy concentration of p-type impurity through the epitaxial layer until it connects with the p-type substrate. This is shown in (B). Following reoxidation and photoresist, the base diffusion is achieved by diffusing a

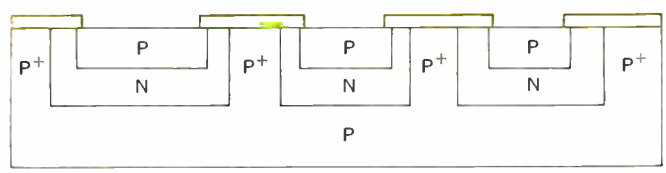
Fig. 4. Sequence of photoresist and diffusion processes.



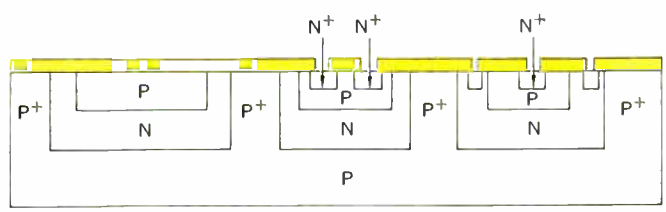
A Epitaxial growth



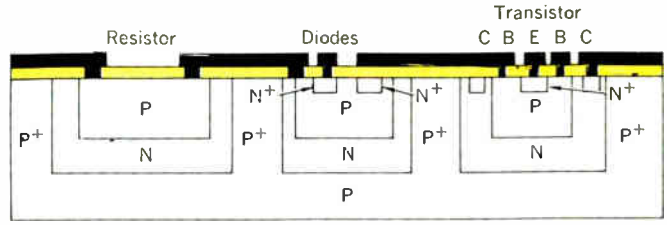
B Isolation diffusion



C Base diffusion



D Emitter diffusion



E Aluminum metalization

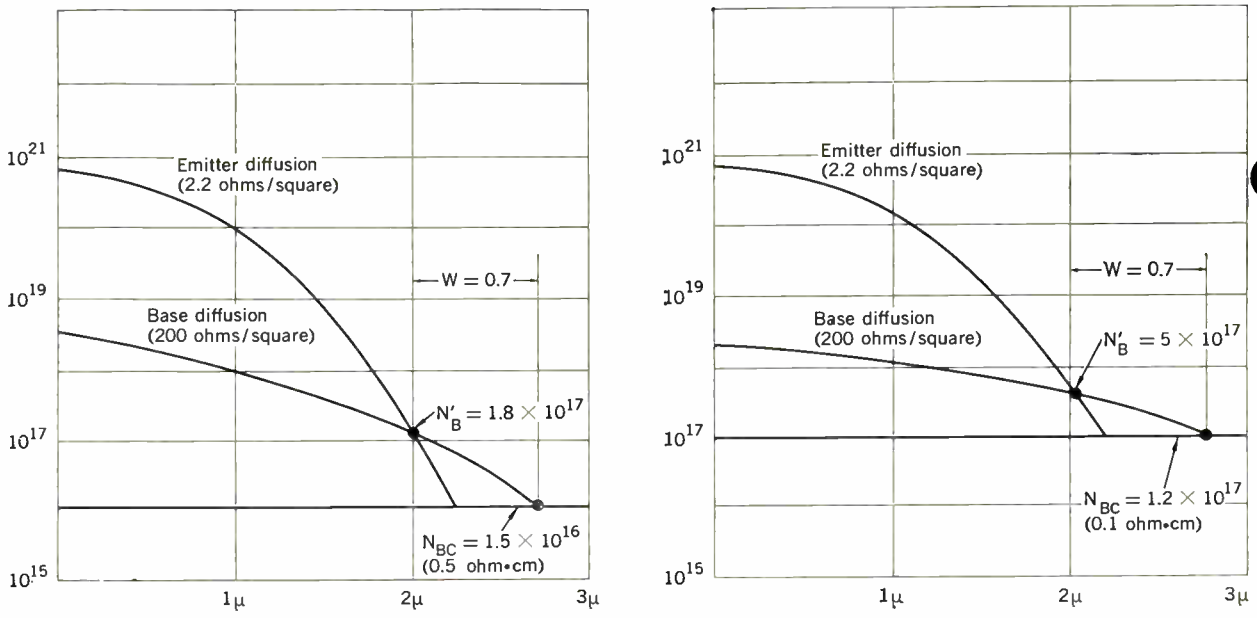


Fig. 5. Typical impurity profiles for monolithic integrated circuit transistors.

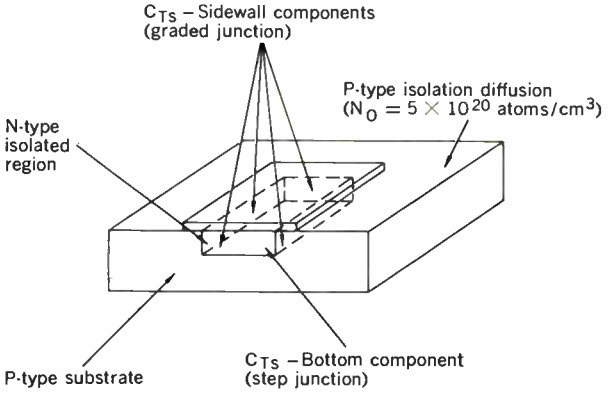


Fig. 6. Details of an isolation region illustrating the diffusion of impurities under oxide.

p-type impurity as shown in (C). Note here that this region forms both the base of the transistor and the resistors. In (D), the oxide is shown removed to permit the shallower n-type emitter diffusion to occur. This step simultaneously forms the transistor emitter, the n-region of the diodes, and the ohmic contacts to the collector region of the transistor. After reoxidation, the oxide is removed at those points where the interconnections are to make electrical contact. The cross section (E) shows the aluminum metalization interconnections after evaporation and photoresistance. This completes the process sequence in which resistors, diodes, and transistors are formed simultaneously in a silicon wafer.

Two typical impurity profiles that are widely used in monolithic structures are shown in Fig. 5. The left profile represents the diffusion profile in an n-type epitaxial background of 0.5 ohm·cm; the right profile is similar, but has an impurity background of a 0.1 ohm·cm. In both cases, the diffusion cycle for the base diffusion is adjusted to produce a sheet resistance of 200 ohms per square at a junction depth of 2.7 microns. The n-type emitter diffusion is doped much more, and produces a sheet resistance of 2.2 ohms per square at a junction depth of 2 microns. In each instance, this produces a net physical base width equal to 0.7 micron.

When these processes are completed, the wafer is scribed into individual circuit chips that are ready for assembly in either conventional multilead TO-5 packages, or the new, flat packages. After hermetic sealing, the integrated circuit is tested for both its dc parameters and ac switching speed.

Capacitance of isolation junctions

In almost all integrated circuits, a reverse-biased silicon p-n junction serves as an excellent means for isolating one voltage point from any other within the monolithic structure. This is because the leakage current under such conditions is measured in nanoamperes and yielding resistances are in the order of hundreds of megohms. As mentioned previously, these junctions are formed by the diffusion of p-type impurities through the epitaxial layer until they link with the substrate. This establishes an isolated n-type region into which either a resistor, capacitor, diode, or transistor can be formed. The details of such a region are shown in Fig. 6. The first important point is that diffusion is a three-dimensional process and, therefore, the isolation impurity will diffuse under the oxide to a depth equal to the thickness of the epitaxial layer. Thus, the effective junction area is not determined by the original oxide pattern but by the actual position of the junction, as shown by the dashed lines in Fig. 6. For example, assuming an epitaxial layer of one-mil thickness, or 25 microns, and an oxide pattern equal to 8 by 10 mils, the effective junction area is reduced by one mil at each side to 6 by 8 mils. The same argument applies to the perimeter. As will be seen later, this is a fundamental

I. Parasitic capacitance values for isolation regions*

Collector-Substrate Capacitance	0.1 Ohm·cm Epitaxial Collector, pF/mil²	0.5 Ohm·cm Epitaxial Collector, pF/mil²	Voltage Variation
Bottom component (10 ohm·cm substrate)	0.078	0.078	$C(V) = \frac{C_{Ts}}{V^{1/2}}$ (step junction)
Sidewall component ($N_0 = 5 \times 10^{20}$)	0.17 ($X_j = 25\mu$)	0.09 ($X_j = 12.5\mu$)	$C(V) = \frac{C_{Ts}}{V^{1/3}}$ (graded junction)

* All values for V = 1 volt reverse.

point in integrated circuit design for the determination of the magnitudes of parasitic capacitances.

We now have to consider that these isolation junctions have a significant capacitance per unit area that cannot be neglected. In Fig. 6, note that the isolation island will have a total capacitance, with respect to the substrate, equal to the sum of both the bottom components and the sidewall components. This total capacitance, known as C_{Ts}, is called the *parasitic capacitance* of the isolation region. In an epitaxial monolithic circuit it is reasonable to assume a step junction for the bottom component, and that the capacitance is determined primarily by the resistivity of the p-type substrate. The step junction is not valid for the sidewall component because of the graded diffusion profile. Here, the sidewall capacitance component is determined partially by the resistivity of the n-type collector region, and must be regarded as a graded-junction capacitance. For the sidewall component, the calculated area is equal to the perimeter of the isolated region times the thickness of the epitaxial layer.

With reference to Fig. 6, Table I summarizes the calculated parasitic capacitance values for isolation regions. It should be noted that the bottom component is independent of collector resistivity, since the substrate is the same in both cases. All values given are for a voltage equal to a one-volt reverse bias. Capacitance values for any other reverse voltage are obtained merely by dividing by the square root of voltage for the step-junction bottom component, and by the cube root of voltage for graded-junction sidewall components.

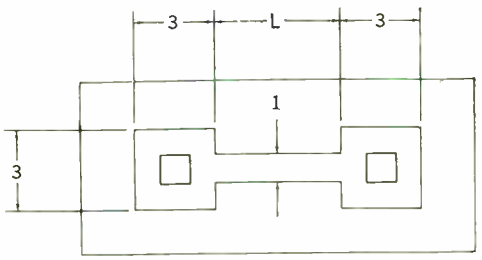
As an example, to determine the value of the parasitic capacitance of a 6- by 8-mil isolation region that has an epitaxial thickness of one mil and a resistivity of 0.1 ohm·cm, with 5 volts reverse from collector to substrate

$$C_{Ts} = C_{Ts(\text{bottom})} + C_{Ts(\text{sidewall})} \tag{1}$$

$$C_{Ts} = \frac{0.078 \text{ pF/mil}^2 (6 \text{ mil} \times 8 \text{ mil})}{\sqrt{5 \text{ volts}}} + \frac{0.17 \text{ pF/mil}^2 (28 \text{ mil} \times 1 \text{ mil})}{\sqrt[3]{5 \text{ volts}}} \tag{2}$$

$$C_{Ts} = 1.68 \text{ pF} + 2.78 \text{ pF} = 4.46 \text{ pF} \tag{3}$$

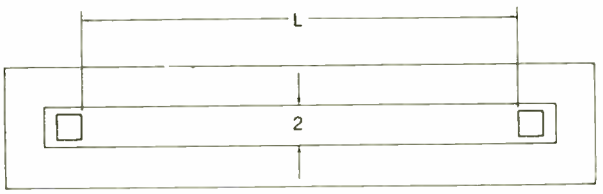
1-mil-width diffused resistor (20% tolerance)



$R = R_{BB}[L + 2(0.65)] = 2000$ ohms

$L = 8.70$ mils $R_{BB} = 200$ ohms/square

2-mil-width diffused resistor (10% tolerance)



$R = \frac{R_{BB}}{2}[L + 2(0.14)] = 2000$ ohms

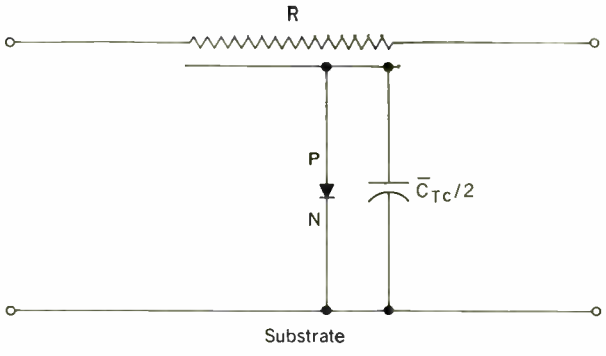
$L = 19.72$ mils $R_{BB} = 200$ ohms/square

Fig. 7. Resistance formulas for two representative types, including correction factors for end contacts.

Diffused resistors

A resistor element in a monolithic integrated circuit is actually the resistance of the p-type base region that is diffused into the isolated n-type collector region. Here, the total resistance is that which is measured between the ohmic contacts of the aluminum metalization to the p-type diffused region. Except for the ohmic contacts, the resistor is completely covered with a silicon oxide. Thus, the design of the diffused resistor is related not only to the sheet resistance of the p-type diffusion but also to the geometry as seen from the top view of the resistor in Fig. 7, where two examples are presented to illustrate the principles of resistor design. In this case, both geometries yield a diffused resistor equal to 2000 ohms. Design formulas will vary, however, because of the correction factors required to include the end effects of the ohmic connection. From these formulas, it is apparent that the resistance of a diffused resistor is equal to the base sheet resistance times the length of the resistor, plus the correction factor for the end contacts.

For the resistor that is one mil wide, the straight length is calculated to be 8.7 mils at 200 ohms per square. In the second case, where the width of the resistor is increased to 2 mils, the effective sheet resistance is reduced accordingly to 100 ohms per linear mil. Thus, the straight length



$C_{Tc} = 0.22$ pF/mil^2 (0.1 ohm•cm $V = 1$ volt, reverse)

$C_{Tc} = 0.11$ pF/mil^2 (0.5 ohm•cm $V = 1$ volt, reverse)

Fig. 8. Equivalent circuit of diffused resistor. Capacitance is lumped equivalent to distributed values.

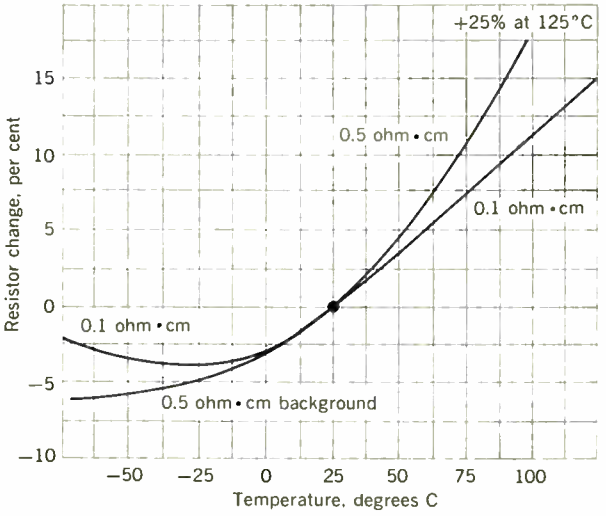


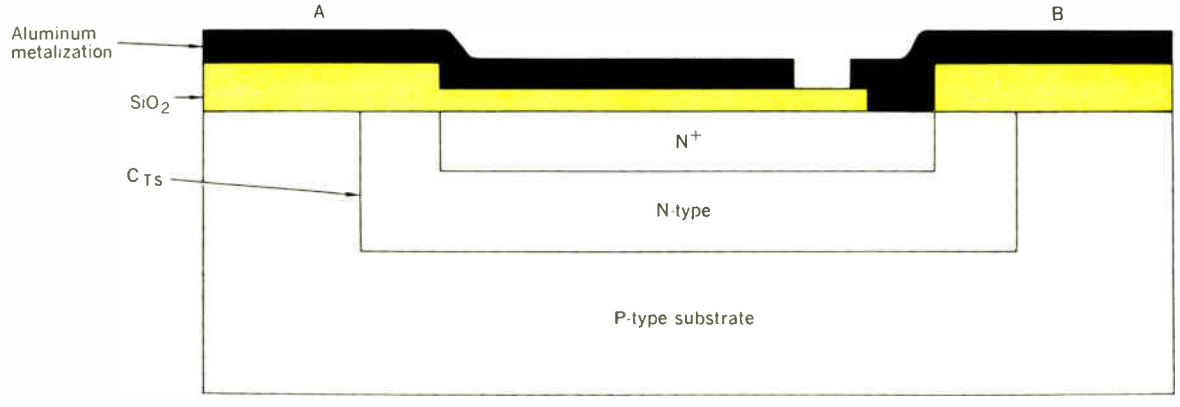
Fig. 9. Resistance changes plotted as a function of temperature for a sheet resistance of 200 ohms per square.

becomes 19.7 mils. Obviously, the wider the resistor, the longer it must become to achieve a given absolute value. This point reveals the first significant limitation of integrated circuit technology that must be considered; that is, the narrower a resistor is made, the more difficult it is to maintain a given tolerance. This is attributed to the fact that the physical width of the resistor is not constant, and could vary as much as 0.1 mil because of the chemical undercutting effects of the oxide after the photoresist process step. From a percentage viewpoint, undercutting variations are more severe for a one-mil-wide resistor than for one of two-mil width. Sheet resistance control is another important process variation; however, R_{BB} can be maintained well within 5 per cent. Because of these results, it is reasonable to expect a 20 per cent tolerance for a diffused resistor of one-mil width, and a 10 per cent tolerance for a 2-mil-wide resistor. It is always preferable to design for the maximum width where tolerances are critical, but unfortunately this requires longer resistor geometries that produce larger monolithic-chip designs. The other disadvantage is that a wider resistor has a larger parasitic capacitance. Therefore, it is essential, from a circuit design consideration, that the widest possible tolerances be required. One way of overcoming the tolerance problem is to design circuits based on resistance ratios. These ratios can be maintained well within 4 per cent, because all resistor errors track very closely together within the same chip.

The equivalent circuit of a diffused resistor is shown in Fig. 8, where the diode is the junction between the diffused resistor and the isolated collector region. In the operation of the circuit, this diode must be reverse biased to prevent forward conduction when positive potentials are applied to the resistor terminals. This is achieved by connecting the resistor isolation region to the most positive potential of the circuit. Figure 8 also shows the parasitic capacitance of the same junction distributed across the entire length of the resistor. This is a graded-junction capacitance that has values of one volt for each resistivity shown in Fig. 8. Later, it will be shown how the distributed capacitance may be equated to a single lumped value equal to $\bar{C}_{Tc}/2$, where $\bar{C}_{Tc}$ is the average parasitic capacitance of the resistor.

The magnitude of diffused resistors is not constant with temperature, as shown by the Fig. 9 curves. These percentage variations, with respect to room temperature, are

Fig. 10. Cross section of an oxide-type capacitor.



based on a sheet resistance of 200 ohms per square. It is seen that for a 0.1-ohm·cm background, the percentage increase in resistance is 15 per cent at 125° C. At −55° C, the resistance decreases to about −5 per cent. The resistance changes for a 0.5-ohm·cm background are more severe because the total impurity concentration is lower for the same sheet resistance level. In any case, these resistance variations are attributed primarily to the change of hole mobility with temperature. The heavier the doping level of the base diffusion, the smaller is the resistance variation.

Oxide capacitors

Although this specific component is not required in the monolithic version of the vehicle DTL circuit of Fig. 2, it is appropriate at this point to discuss the design details of the oxide type of capacitors used in monolithic integrated circuits. The cross section of such a capacitor is shown in Fig. 10. It is completely analogous to the simple parallel-plate capacitor, except that the dielectric used here is silicon oxide. One plate consists of the heavily doped n^+ region that is formed during the emitter diffusion. The other plate is the aluminum metalization that is evaporated over the thin silicon oxide. For an oxide thickness of 500 Å, the capacitance per unit area is 0.35 pF/mil². This type of capacitor is very stable and independent of voltage. The equivalent circuit is shown in Fig. 11, wherein the diode is the collector to the substrate junction. The significant parasitic for this element is C_{TS}, the parasitic capacitance of the collector–substrate junction. In certain circuits, it is absolutely essential that the ratio C/C_{T_s} be as large as possible. This is partially accomplished by applying the maximum reverse bias to the substrate junction to minimize the value of C_{T_s}. The small 2-ohm resistor represents the series resistance of the diffused n^+ region.

Monolithic diodes

The cross-sectional view of the possible diode configurations that are employed in monolithic circuits is shown in Fig. 12. The two basic diode types are the emitter–base, and the collector–base. For comparison, the junction area in each case has been made equal. In actual practice, this area can be made as small as one mil by one mil for

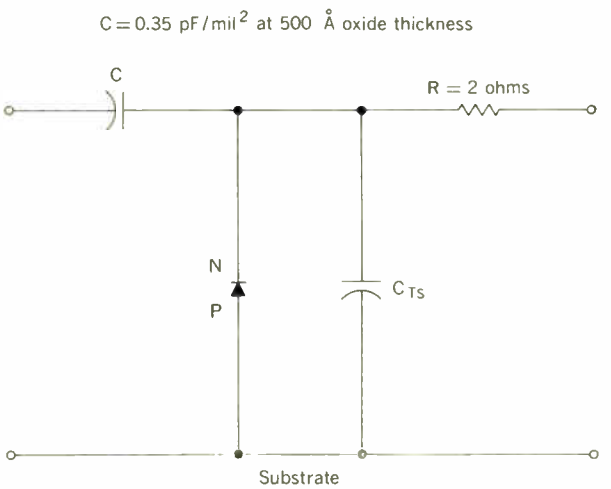


Fig. 11. Equivalent circuit of an oxide capacitor.

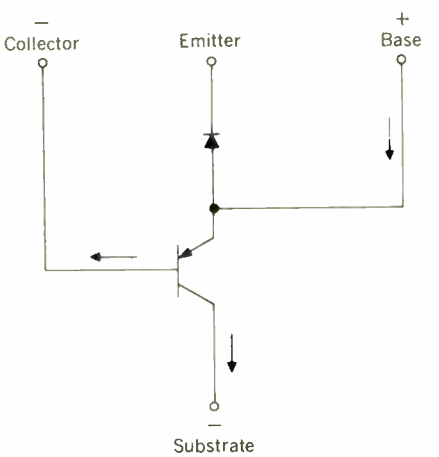
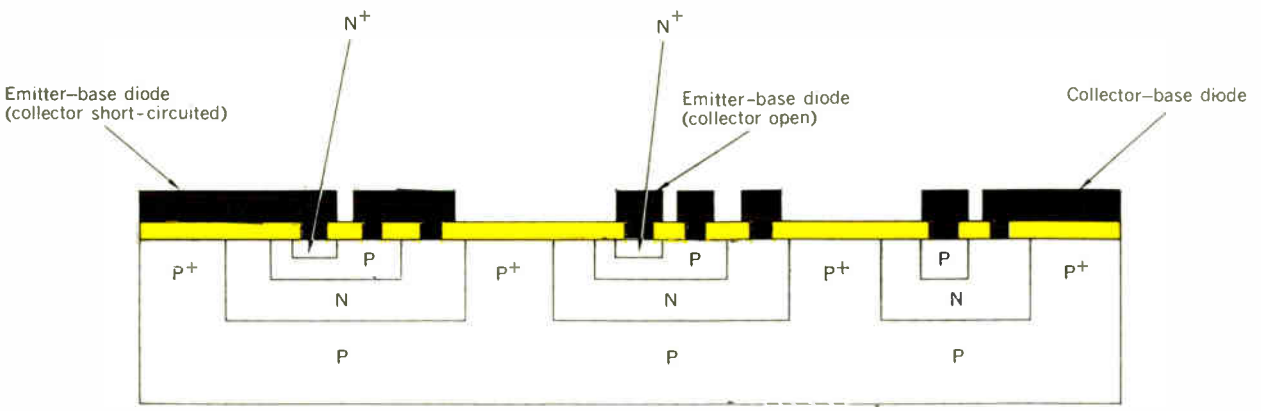


Fig. 13. Equivalent circuit of a four-layer transistor that illustrates a typical p-n-p parasitic transistor action.

Fig. 12. Cross section of various diode structures.



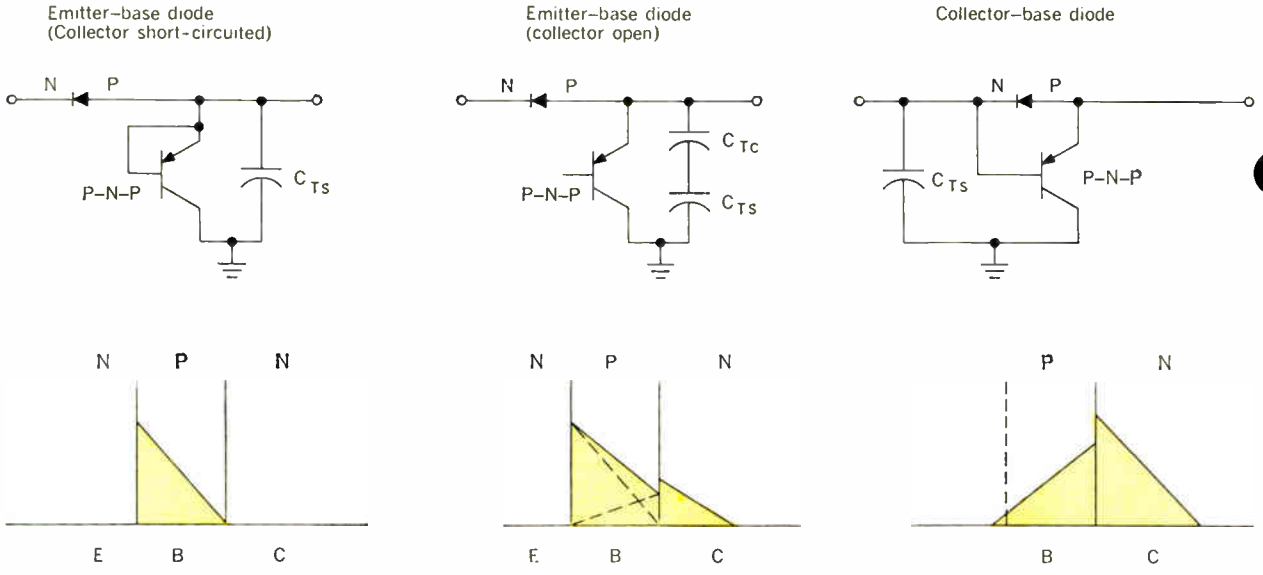


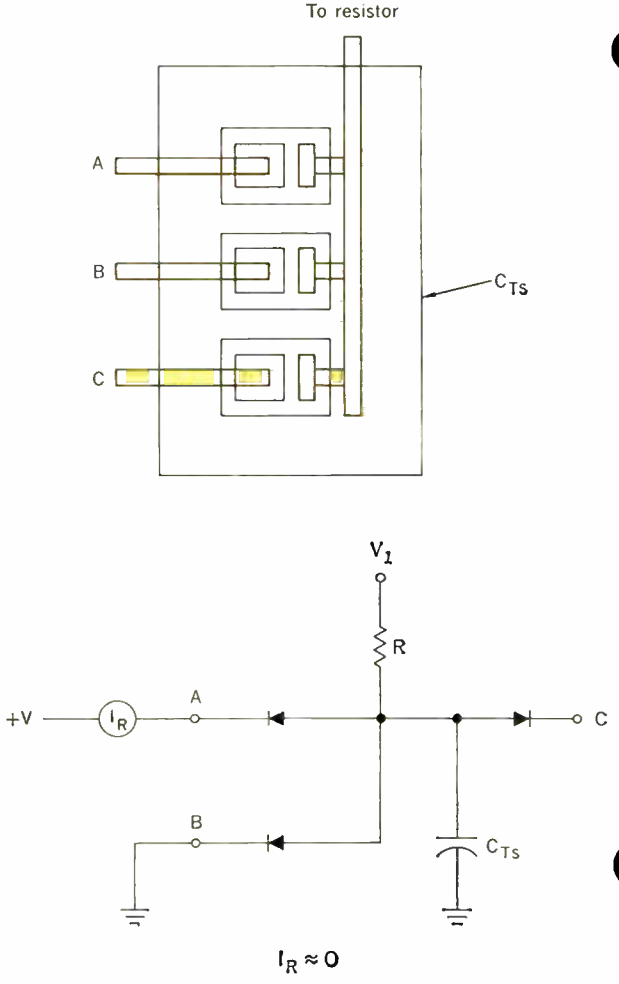
Fig. 14. Equivalent circuit and minority-carrier storage diagrams for various diode configurations.

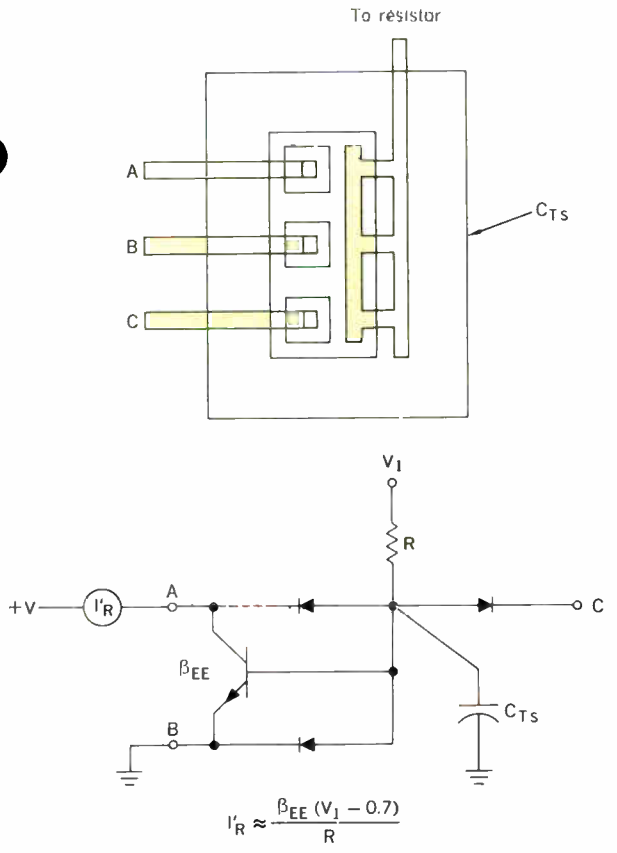
maximum, high-frequency performance. The emitter–base diode has two variations, the collector short-circuited, or the collector open. As shown in Fig. 12, this is accomplished by the use of aluminum metalization.

In each case, when the diode is forward biased, the p-type base region can be forward biased with respect to the n-type collector region. Furthermore, the substrate is always reverse biased to prevent forward conduction. Consequently, in each diode we see the appearance of a p-n-p parasitic transistor having properly applied bias potentials that cause it to shunt current from the diode to the substrate. The equivalent circuit of this parasitic p-n-p transistor is shown in Fig. 13, wherein the base of the diode becomes the emitter of the p-n-p transistor. Fortunately, in digital integrated circuits, the lifetime of the monolithic chip is usually shortened by gold diffusion that reduces the storage time in the diodes and transistors. As a result of this very short lifetime, and the extreme thickness of the epitaxial collector region (which is the base of the parasitic p-n-p transistor), the beta (current-transfer ratio) of the latter is essentially equal to zero. For nongold-diffused structures, however, the beta can range as high as from 2 to 4.

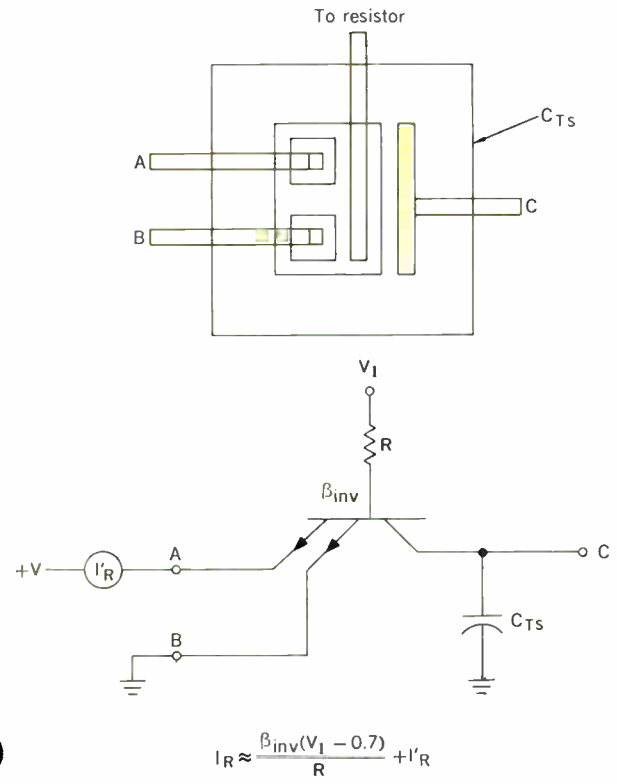
A better understanding of the properties of the three-diode configurations is obtained by reference to the equivalent circuits of Fig. 14. In the first case, we have the emitter–base diode with the collector short-circuited. In this instance, the emitter–base junction of the p-n-p is short-circuited, and the parasitic capacitance of the collector–substrate junction is tied directly to the base of the diode. For this structure, the minority-carrier diagram shows charge stored only in the base region. If the collector is open, as shown in the second case, the p-n-p parasitic transistor appears off the base node in a BV_{CEO} condition, wherein the base–collector junction becomes slightly forward biased, and the collector–substrate junction is reverse biased. This manifests itself in a forward-bias capacitance C_{Tc}, which appears in series with the substrate capacitance, C_{Ts}. Under the conditions of an

Fig. 15. Equivalent circuit of triple emitter–base diode configuration that uses separate base regions.





A Collector short-circuited



B Collector open (TTL)

open collector, the minority carrier diagram would show additional stored charge in both the base and collector regions. Thus, the diode-recovery time would be longer than in the short-circuited case. The third arrangement is where there is a collector–base diode. In this situation, the emitter–base junction of the p-n-p parasitic transistor is forward biased, but the current shunted to the ground is still almost negligible because of the extremely low beta. The interesting point to note, however, is that here the substrate capacitance appears from the n-type side of the diode. Diode recovery of this structure may be higher, since charges are stored as shown, in both the base and collector regions. In reference to forward drop, the emitter–base diode, with the collector short-circuited, is usually the lowest of the three—by as much as 0.1 volt. Also, the breakdown voltage rating of an emitter–base diode is lower than that of a collector–base diode. In the former case, it is equal in magnitude to the BV_{EBO} of a transistor, that is, from 6 to 8 volts. For the base–collector diode, the voltage breakdown is approximately BV_{CBO}—or in the range of 20 to 45 volts, depending upon the resistivity of the epitaxial collector.

It is important here to examine the design considerations to determine which diode configuration should be selected for a particular monolithic circuit. Let us assume the input diode configuration of a DTL circuit, wherein we have two diode inputs meeting at a common node, and a similar diode, serving as an output, coming from the same node. The first possible arrangement is shown in Fig. 15, with three emitter–base diodes (collector short-circuited) using separate base regions. Inputs *A* and *B* enter the individual emitter-diffused n-regions, and the output circuit leaves as shown in Fig. 15. The total substrate capacitance at the common node is related to the total area of the isolated island. If input *A* is returned to a positive voltage, and input *B* is grounded, the reverse current I_R that flows is essentially zero, since it consists only of the nanoampere leakage current of two diodes connected back to back.

As each base region is at the same electric potential, there is no reason why space cannot be conserved by the establishment of a common base region. The new configuration is shown in diagram (A) of Fig. 16; in this case the diode junctions still have the same area respectively, but the total isolation island dimensions have been reduced accordingly, thereby producing a lower substrate capacitance. The leakage current observed between inputs, however, is now slightly larger because of the presence of a parasitic n-p-n transistor as shown in the equivalent circuit. This parasitic transistor is formed between the n-type inputs and the common p-type base region. Because of the wide spacing between inputs and gold diffusion, the emitter-to-emitter beta β_{EE} is usually about 0.001 or less. Although small in value, this does increase the reverse current to a higher level in the microampere range.

By disconnecting the collector–base short circuit we have available an additional junction that can be used as an output diode. This eliminates the need for the third

Fig. 16. Equivalent circuit of triple emitter–base diode configuration that uses common base regions.

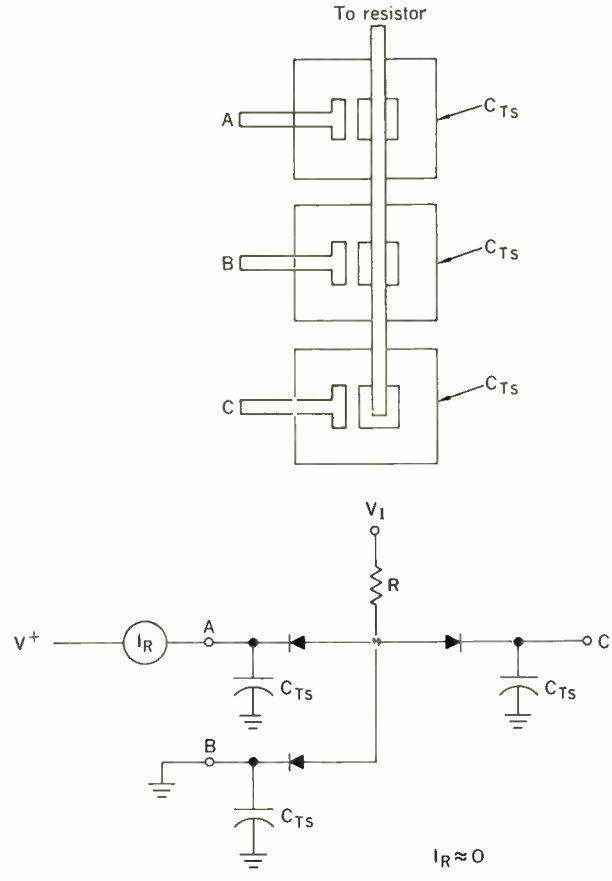
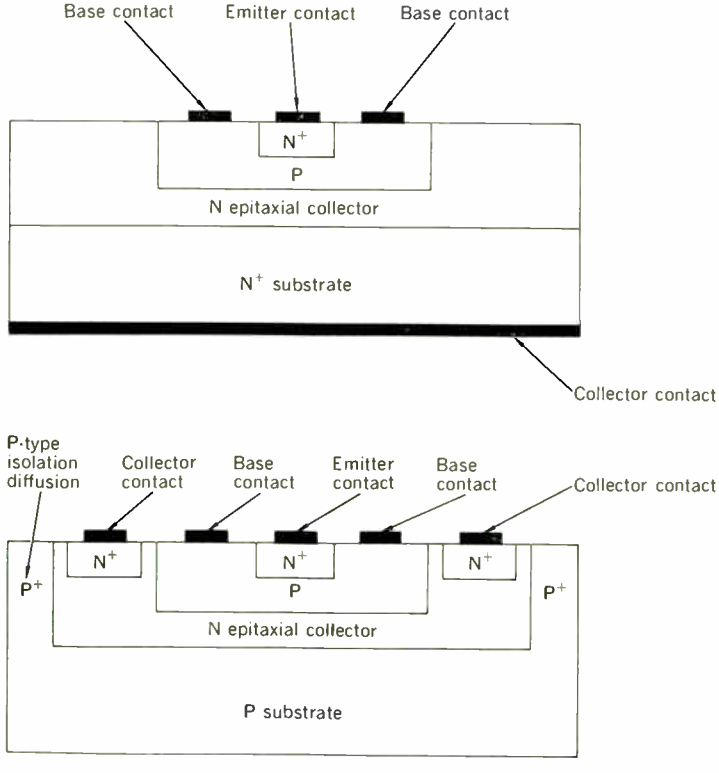


Fig. 17. Equivalent circuit of triple collector–base diode configuration that uses isolated collectors.

Fig. 18. Comparison of cross sections of integrated-circuit transistor with discrete planar-epitaxial transistor.



diode, so the configuration is reduced to diagram (B) of Fig. 16. This is redrawn schematically as a transistor that has two emitter inputs. It is commonly known as transistor–transistor logic (TTL), where the total isolation area is reduced considerably and produces a lower collector-to-substrate capacitance. The reverse current is even higher in this case, because, in addition to the I_R' component from β_{EE}, there is another component from the inverse beta (β_{inv}) of the input transistor. In these configurations, the total reverse current that flows is calculated by the formulas given in Fig. 16. The salient advantage of the TTL configuration, with reduced parasitic capacitance (for higher speed operation), is offset by the disadvantage that inverse beta can impose on the worst-case logic design.

The fourth, and final, arrangement, using collector–base diodes, is shown in Fig. 17. Here, each input is an isolated collector region. In this case, the leakage current I_R is very low, since there are no inverse beta mechanisms. The main advantage is that each parasitic capacitance is now hanging from the inputs rather than from the common node. This considerably improves the speed of the circuit as the input capacitance is usually driven by a lower resistance value.

Since the advantages and disadvantages of each diode configuration have been demonstrated, the designer must now determine which form will be most compatible with his circuit requirements. The design compromises usually involved are: circuit switching speeds, as effected by the magnitude of the parasitic capacitance, vs. reverse-input currents, as effected by β_{inv} problems.

Monolithic transistors

The cross section of a monolithic transistor, shown in Fig. 18, is compared to its discrete n-p-n counterpart that was produced by the epitaxial process. In both cases we have an epitaxial collector—except that the discrete device is self-isolated, and the monolithic transistor is isolated by the collector-to-substrate junction. Furthermore, in the discrete device, the collector current flows down through the collector to the substrate, and thence to the bottom collector contact; whereas, in the monolithic structure, the current flows along the collector in a narrow sheet to the top collector contacts. For this reason, the collector series resistance of the monolithic transistor is considerably higher than the epitaxial transistor at the same resistivity level. Thus, the lower epitaxial resistivities are required for monolithic transistors to provide sufficiently low enough $V_{CE(sat)}$. Therefore, we have both 0.1-ohm·cm, and 0.5-ohm·cm, n-type collectors as design parameters. For each of these backgrounds, the impurity profile is exactly as shown in Fig. 5.

A typical geometry of the monolithic structure is indicated in Fig. 19. This geometry uses an emitter one mil wide, and all other contacts and spacings are equal to 0.5 mils. These dimensions, and their tolerances, are compatible with the process limitations of contemporary integrated-circuit art. In Fig. 19, the emitter length is 1.5 mils, and is confined by two base contacts. As indicated, the collector contact completely surrounds the base region. A summary of the pertinent parameters of this 1- by 1.5-mil double-base stripe transistor is given in Table II for both the 0.5-ohm·cm, and the 0.1-ohm·cm epitaxial collector. With respect to breakdown voltages

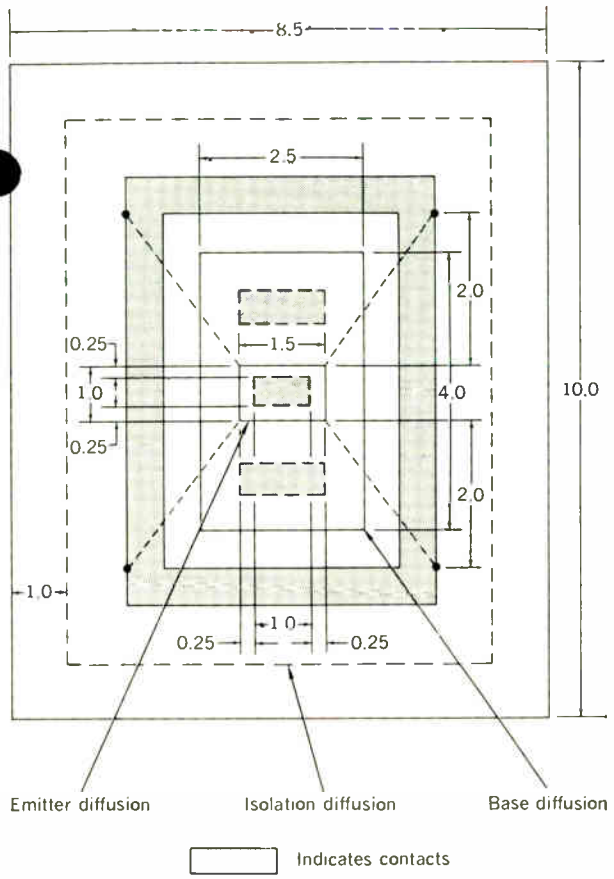


Fig. 19. A typical double-base stripe geometry of integrated circuit transistors.

and capacitances, the 0.5-ohm·cm device is superior to the 0.1-ohm·cm device. However, r_{sc} for the 0.5 ohm·cm collector is 75 ohms when compared to the 0.1-ohm·cm case, wherein it is reduced to 15 ohms. Therefore, to attain the higher impurity doping level of the collector, one device is penalized by lower breakdown voltages and higher capacitances.

The calculations of the series collector resistance are best demonstrated by reference to Fig. 19. Assuming that the current is emitted only from the emitter edges, then, within the collector region, the current will fan out within the dashed lines to the collector contact. Thus, the series resistance of the structure is determined by calculating the resistance of the trapezoid. For a double-base geometry, where there are two trapezoidal regions in parallel, the following expression is used for collector series resistance

$$r_{sc} \approx \frac{\rho_c \cdot d}{W_c(2l_e + 3)} \qquad (4)$$

where r_{sc} is the resistivity of the collector in ohm·cm, W_c is the thickness of the collector region in centimeters, l_e is the length of the emitter in mils, and d is the distance between the inside edges of the emitter and the collector contact. The ratio of ρ_c/W_c is actually the sheet resistance of the collector region.

When the value of r_{sc} is determined, the magnitude of $V_{CE(sat)}$, at any particular collector current (for maximum fanout), can be determined by using the approximate relationship

$$V_{CE(sat)} \approx 0.150 + I_c r_{sc} \text{ volts} \qquad (5)$$

The 150-millivolt term in (5) represents the sum of the transistor junction potential in saturation and the grounded-emitter offset voltage. The former is calculated from the familiar Ebers-Moll equation for a typical beta of 50, an I_c/I_B ratio of 10, and an inverse alpha of 0.4. The offset voltage is attributed to the effects of gold diffusion, which is needed to decrease lifetime for purposes of switching speed. Using (5), we see in Table II, that for the 0.5-ohm·cm collector, $V_{CE(sat)}$ at 5 mA, we have 0.525 volt; for 0.1 ohm·cm, $V_{CE(sat)}$ is 0.225 volt. The use of (4) and (5) may be illustrated by considering $V_{CE(sat)}$ at 35 mA for a 1 by 6 monolithic transistor, that has a 0.1-ohm·cm collector one mil thick. From (4) we can calculate that r_{sc} equals 5.3 ohms; from (5) we know that $V_{CE(sat)}$ equals 0.335 volt.

II. Characteristics for 1- by 1.5-mil double-base stripe monolithic transistors

Transistor Parameter	0.5 ohm·cm	0.1 ohm·cm
BV_{CBO}, volts	45	20
CV_{EBO}, volts	8	6.5
BV_{CEO}, volts	19	11
C_{Te} (forward bias), pF	6	10
C_{Te} at 1 V, pF	1.5	2.5
C_{Tc} at 5 V, pF	0.65	1.3
h_{FE} at 10 mA	50	50
r_{sc}, ohms	75	15
$V_{CE(sat)}$ at 5 mA, volts	0.525	0.225
V_{BE} at 10 mA, volts	0.8	0.8
f_T at 5 V, 5 mA, Mc/s	563	678

The frequency response of the monolithic transistor is inherently the same as its discrete counterpart insofar as emitter and base time constants are concerned. The measured f_T, under short-circuit output conditions, is considerably lower because the collector capacitance and the parasitic capacitance are charged through a higher collector series resistance that gives the considerably lower f_T shown in Table II. Actually, the 0.1 ohm·cm column has a higher f_T because the decrease in r_{sc} more than adequately compensates for the increase in capacitance.

The key design consideration in the selection of a monolithic transistor is the determination of the emitter length, since the current rating is proportional to the emitter periphery. The typical design value is 4 mA for each mil of emitter periphery. Thus, the 1- by 1.5-mil geometry of Fig. 19 is capable of handling a maximum current of approximately 12 mA. If a higher current rating is required, the stripe must be lengthened accordingly. For example, a 1- by 6-mil emitter would be capable of handling 48 mA. This current capability is demonstrated in Fig. 20, where beta, as a function of collector current, is shown for different monolithic transistor geometries. The only variable is the length of the emitter; all other dimensional spacings remain exactly the same. Reference to (4) also indicates that by scaling up the length of the

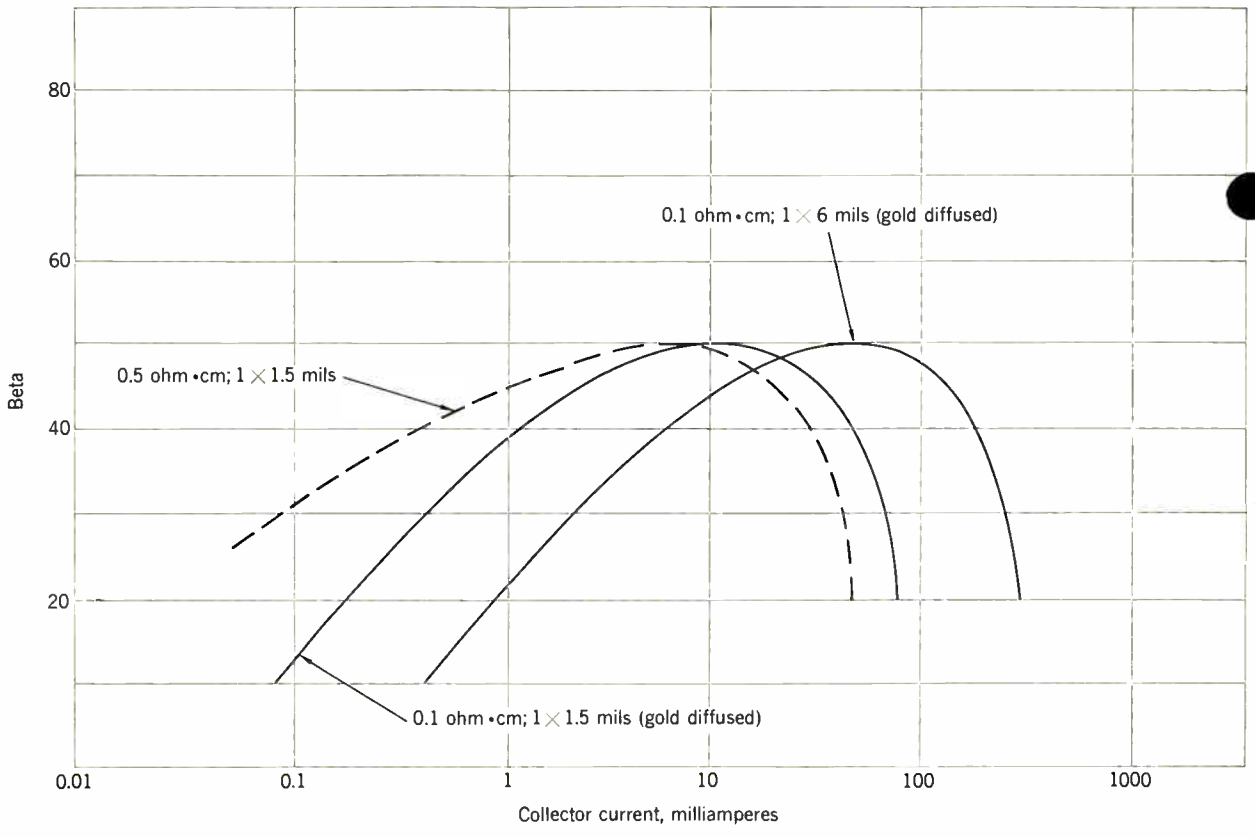
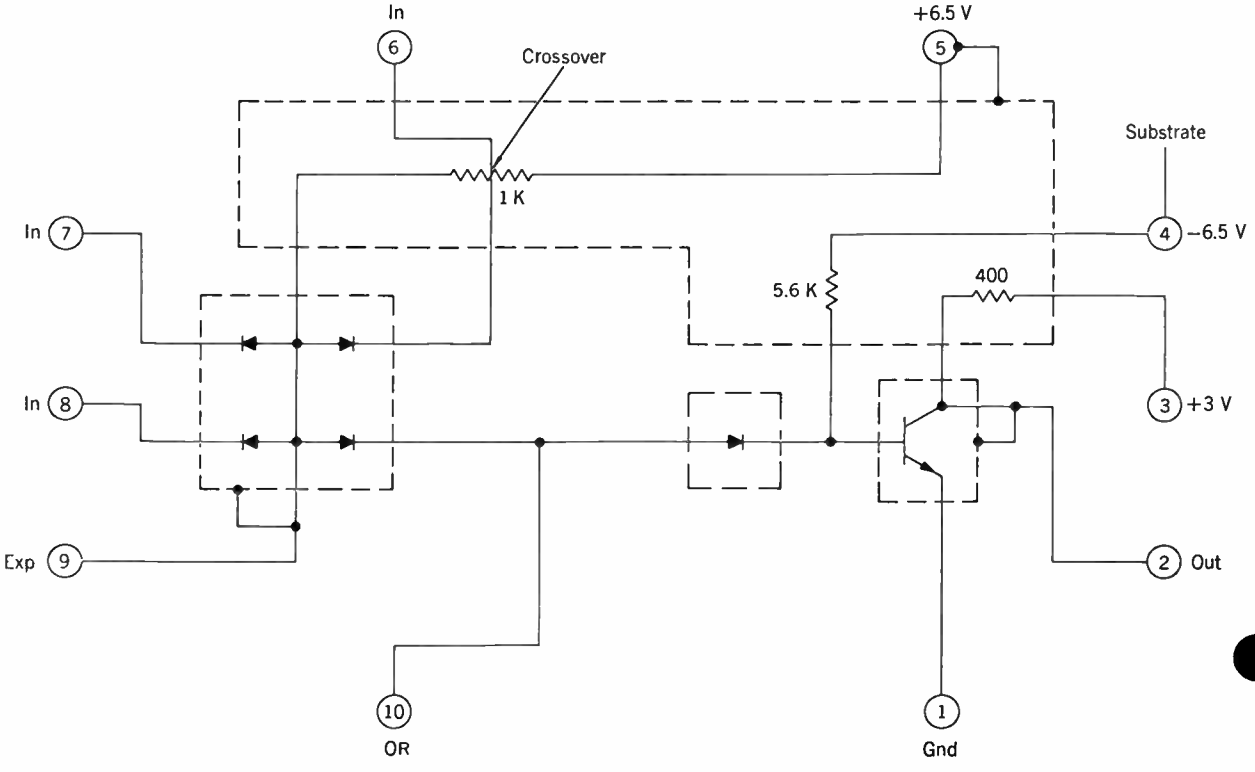


Fig. 20. Representative beta values plotted against current characteristics for various monolithic transistor structures.

Fig. 21. Redrawn schematic diagram of a DTL NAND gate to satisfy pin connections and to determine the number of isolation regions.



emitter, the r_{SC} will be reduced, since I_e appears in the denominator.

In summary, for a digital application, it is apparent that a 0.1-ohm·cm collector is necessary if one is to achieve a lower $V_{CE(sat)}$ for the monolithic transistor. The next design consideration, then, is the current level for which $V_{CE(sat)}$ is to be determined. This in turn determines the maximum magnitude of r_{SC} that will yield the desired minimum length of the one-mil-wide emitter. Thus, we have the concept of a basic monolithic transistor that has a 1-mil by X-mil emitter. With reference to the dimensions in Fig. 19, the establishment of the emitter length automatically fixes the dimensions of the transistor isolation island.

Layout and design rules

In the previous sections the general design details of resistors, capacitors, diodes, and transistors have been presented as they appear in a monolithic integrated circuit. With this as a fundamental background, the discrete circuit shown in Fig. 2 can be transformed into a monolithic integrated circuit.

The first step in the sequence of monolithic design is to redraw the schematic to satisfy the pin connection requirements that may be dictated by the particular system to which the integrated circuit is to be applied. In Fig. 21 such a redrawn schematic is shown. The 10-pin array was drawn first and then, through trial and error, the schematic was arranged until the designated pin connections were supplied without producing undesired crossovers. It is apparent from Fig. 21 that the diode input coming in on pin 6 is shown crossing over the one-kilohm resistor. A legitimate crossover is accomplished because it implies that the aluminum metalizing connection can cross over that portion of a diffused resistor that is insulated with silicon oxide. This is a most important and useful design concept; all connecting wires can cross over resistors in order to conform to any particular set of pin connections.

The next step is to determine the number of isolation regions that are required. They are dependent upon how the potentials of the various n regions for each element vary within the circuit. Therefore, at this point, one must also determine the particular type of configuration to be used for the various elements of the circuit.

The transistor shown in Fig. 21 must be isolated, since its collector swings independent of any other part of the circuit of Fig. 2. The dashed-line rectangle drawn around the transistor indicates the first isolation region. The collector also represents the circuit output point; therefore the transistor isolation island is shown with an electric tie to the output lead pin 2.

The next isolation consideration is that of the series offset diode. To minimize the amount of parasitic capacitance at this node and also to gain the additional stored charge effect, one would select the configuration of the emitter–base diode with the collector open as shown in Fig. 14. Since the collector is floating or open-circuited, this diode must be isolated by itself as shown in Fig. 21.

Because of the requirements of low forward drop and high fanout levels, it was decided to make the input diodes of the emitter–base type with the collector short-circuited. Since all the bases of these four diodes are tied together electrically to a collector that swings from approximately 0.7 to 2.1 volts, this input cluster must be isolated as shown in Fig. 21. Because of the short-circuited collector, the isolation region is shown tied to the expander point of pin 9.

The next isolation region is one that contains all of the resistors of the circuit, since all the resistors are returned to fixed power supply potentials. In terms of parasitic substrate capacitance, this resistor isolation island is not critical because the substrate capacitance appears as bypass or decoupling capacitance at the power supply terminals. It is essential, however, that the resistor isolation region be connected to the most positive potential of the circuit, such as the 6.5 volts at pin 5. If this were not the case, we would find that the p-type diffused resistors would conduct into their n-type isolation islands.

The last and most important consideration is that the substrate must be tied to the most negative potential in the circuit. In this case, the potential is −6.5 volts at pin 4. In any configuration where there are no negative power supply potentials, the most negative potential, of course, is ground. The reason for this tying step is to maintain the maximum reversed bias potential across the substrate-to-collector junctions in order to assure true isolation among the various regions.

With all this done, one can prepare to draft the monolithic layout. This, too, is a trial-and-error process. The goal is to achieve simultaneously the analyzed requirements and maintain the smallest possible circuit chip size. Chip size is primarily an economic consideration in that the integrated circuit cost is related to the area of the final chip design.

In Fig. 22 the dashed lines represent the aluminum metalizing connections drawn to scale at a one-mil width. The transistor is chosen to be a double-base stripe geometry with a 1- by 6-mil emitter. The series diode is drawn as an emitter–base diode having a 1- by 1.5-mil emitter base. The four emitter–base diodes at the input were laid out with 1- by 1.5-mil emitters and separate base regions. The separate base regions in this case were selected in order to minimize any reverse leakage due to β_{EE}. A square diode array was chosen in order to minimize the area of the total isolation island. (A rectangular array with the inputs side-by-side would have yielded a larger area.) One should also note how the bases of these diodes are short-circuited to their collector isolation region by the metalizing. Finally, the various resistors are drawn within one isolation island, which is electrically tied to the 6.5-volt supply with the metalizing run that comes from pin 5. The 400-ohm and 1000-ohm resistors were designed with a width of 2 mils. To conserve space, the 1000-ohm value was achieved by the use of two 500-ohm resistors tied together. In Fig. 22, the input from pin 6 can be seen crossing over the 1000-ohm resistor. The −6.5 volts at pin 4 is connected to the substrate at the contact pad by making an ohmic connection to the p-type isolation diffusion that goes down to the substrate. Thus, Fig. 22 presents the final layout of the entire monolithic DTL circuit of Fig. 2 in a 40-mil-square silicon chip.

The critical design procedure for monolithic layout may be summarized as follows:

1. Redraw the schematic to satisfy the required pin connections with the minimum number of crossovers. Any resistor may be used as a legitimate crossover point

for the aluminum metalizing.

2. Determine the number of isolation islands from collector potential considerations, and minimize the areas as much as possible.

3. Place all resistors having fixed potentials at one end in the same isolation island and return that isolation island to the most positive potential in the circuit.

4. Connect the substrate to the most negative potential of the circuit.

5. In layout allow an isolation border equal to twice the epitaxial thickness to allow for underdiffusion.

6. Use one-mil widths for diffused emitter regions and ½-mil widths for base contacts and spacings, and collector contacts and spacings.

7. For resistors, use widest possible designs consistent with chip size limitations.

8. Always optimize the layout arrangement to maintain the smallest possible chip size and, if possible, compromise pin connections to achieve this.

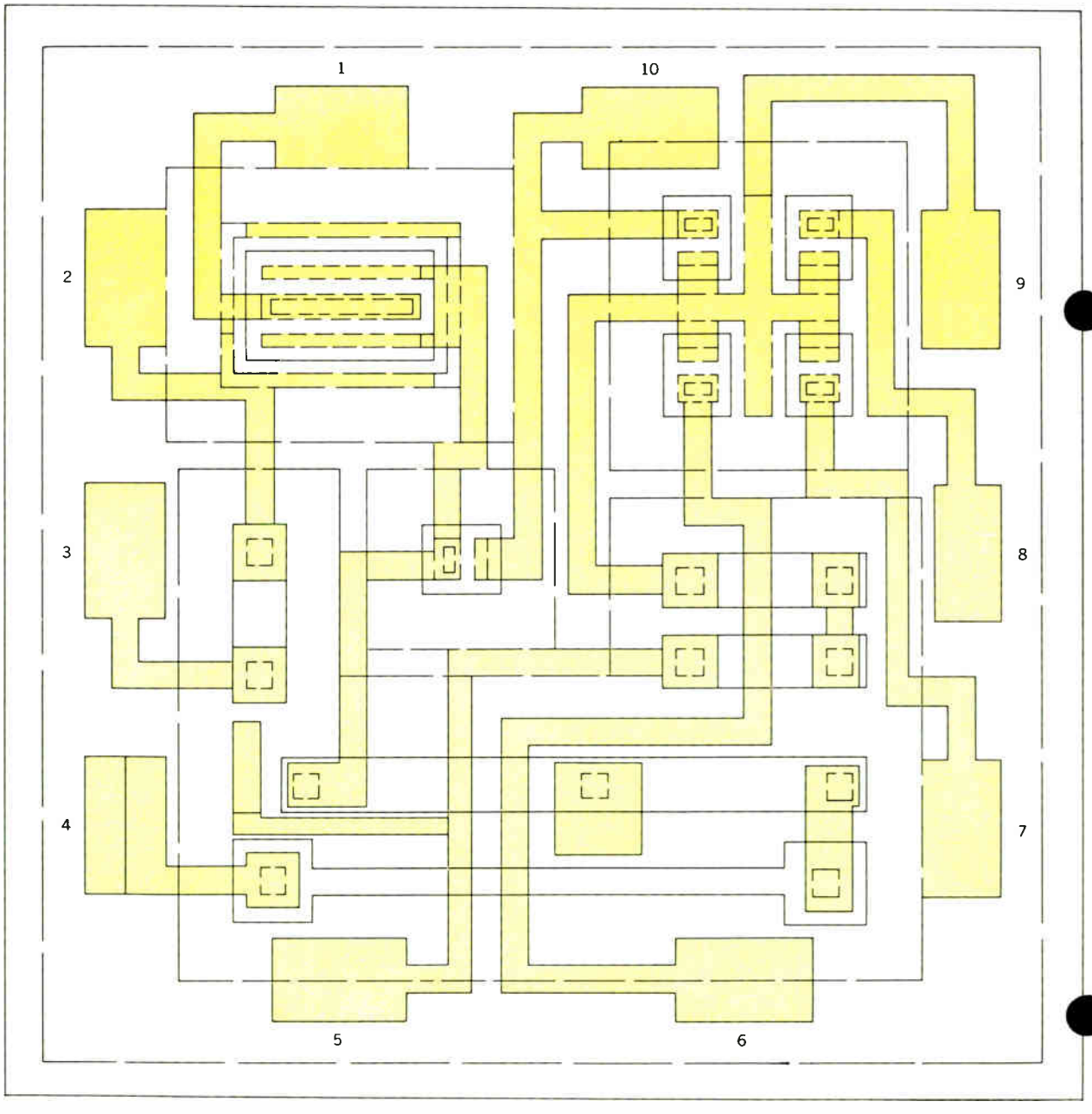
9. Determine component geometries from the performance requirements of the circuit.

10. Keep all metalizing runs as short and as wide as possible, particularly at the emitter and collector output connections of the saturating transistor.

Parasitic capacitances

From the final layout drawing shown in Fig. 22, we can calculate the various parasitic capacitance values that must be added to the discrete circuit to anticipate the final switching performance of the monolithic version. The various formulas for calculating these values are summarized in the following equations.

Fig. 22. Diagram of a DTL monolithic design layout.



Isolation region (bottom component):

$$\bar{C} \approx \frac{2.0 C_{Is}}{V_2 - V_1} (V_2^{1/2} - V_1^{1/2}) \quad (6)$$

Isolation region (sidewall component):

$$\bar{C} \approx \frac{1.5\, C_{Ts}}{V_2 - V_1} (V_2^{2/3} - V_1^{2/3}) \quad (7)$$

Diffused resistor:

$$C \approx \frac{1.9\, C_{Ic}}{V_2 - V_1} [(V_0 + V_2)^{2/3} - (V_0 + V_1)^{2/3}] \quad (8)$$

These equations are expressions for the average capacitance of a reverse-biased p-n junction as it swings between two voltage potentials, V_1 and V_2. In the case of Eq. (8), V_0 is the fixed absolute potential at one end of the resistor. C_{Ts} and C_{Tc} are the capacitance values given in Table I and Fig. 8, respectively. When these formulas are applied to the circuit in Fig. 22 with regard to the various dimensions and voltages, the parasitic capacitance values shown in Fig. 23 result.

Fabrication processes

Assuming that the circuit designer is satisfied with the analysis of the effect of parasitic capacitances on the circuit performance, we can now examine the method used to process the vehicle DTL circuit of Fig. 2. The sequence of events is shown in Figs. 24 and 25. Figure 24 shows the basic steps for preparation of the photomasks from the initial art work. The initial layout, which is

Fig. 23. DTL schematic diagram that includes lumped values of parasitic capacitances.

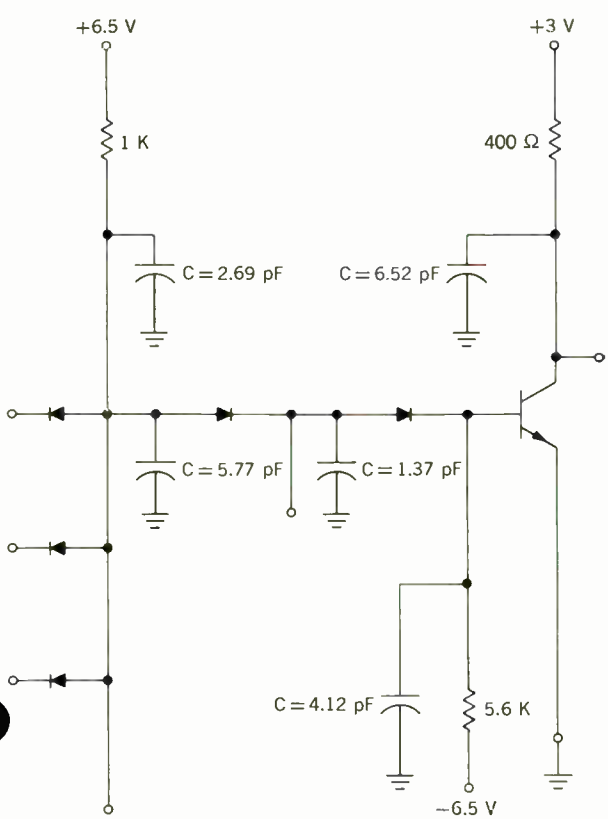
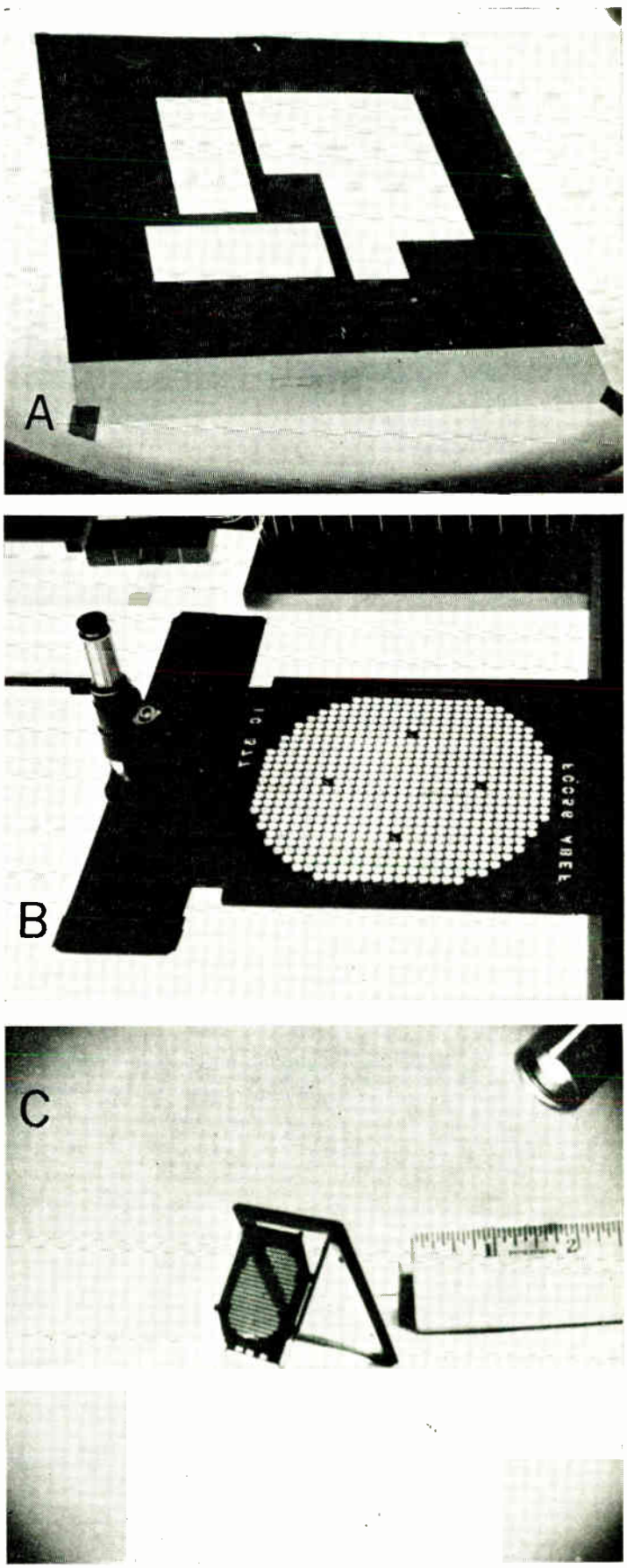


Fig. 24. The basic steps for preparation of photomasks.
A. Layout of the initial art work.
B. The intermediate reduction stage.
C. Final mask size of the completed circuit.



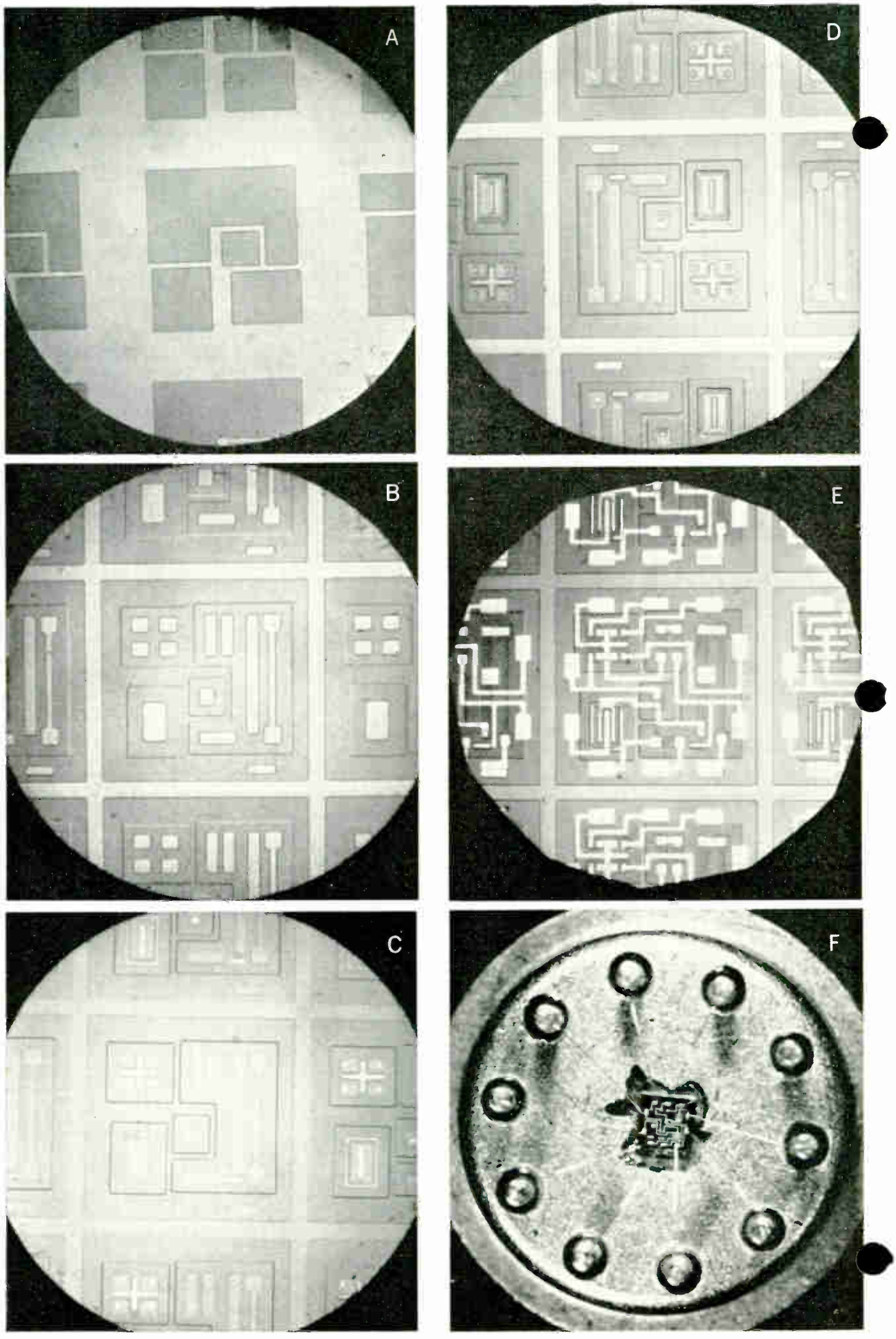
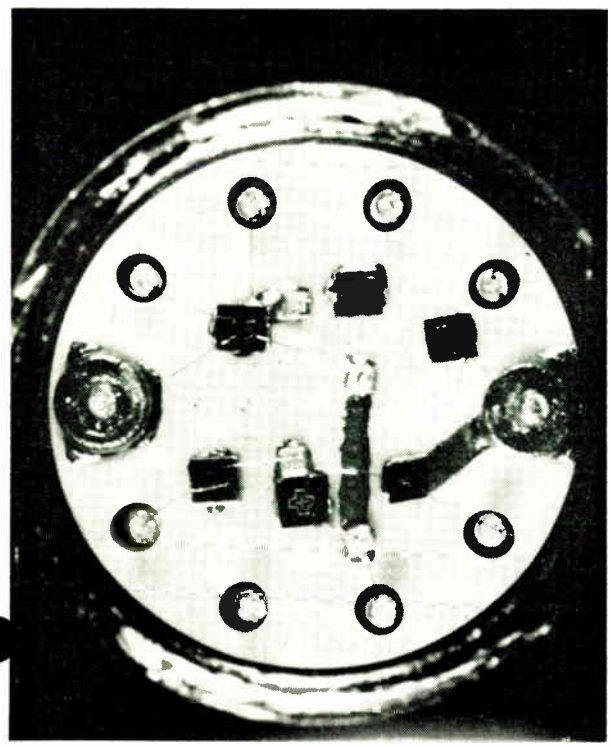


Fig. 25. The six steps in the preparation of a monolithic DTL integrated circuit.
A. Isolation diffusion in which islands are formed.
B. Base diffusion of diodes, transistors, and resistors.
C. N-type emitter-diffusion connections.
D. Preohmic etch for oxide removal.
E. Completed DTL circuit with aluminum metalization.
F. The assembled circuit in a 10-pin, TO-5 header.

accurately drawn at a magnification of 400 times actual size is subjected to an intermediate reduction in which the same pattern is stepped and repeated throughout the photographic plate. Then it is reduced again to final size in which the complete circuit is contained within a 40-mil^2 area. This process is repeated over 100 times across the final mask. An individual photomask is required for each of the five steps in Fig. 4, and these are shown in Figs. 25(A) through (E). In Fig. 25(A) we see the first step in which the various isolation islands have been formed. The base regions of the diodes, transistors, and resistors are shown in Fig. 25(B). In Fig. 25(C), the various n-type emitter-diffusion and collector-ohmic connections are indicated. After the wafer is completely oxided, it receives a preohmic etch [Fig. 25(D)] which removes the oxide where the aluminum-metalized connections are to be made. The final DTL circuit, with aluminum metalization, is depicted in Fig. 25(E). After scribing and assembly, the completed monolithic DTL integrated circuit is shown assembled in a 10-pin, TO-5 header in Fig. 25(F).

To compare the monolithic DTL circuit performance to a discrete circuit, a hybrid or multichip circuit (Fig. 26) is extremely useful. This hybrid circuit is the exact equivalent of Fig. 2; however, individual chips are used for the transistor, diodes, and resistors, thereby eliminating all the isolation junctions and their inherent parasitic capacitance. Note that the input diodes in the Fig. 26 hybrid circuit consist of one common-p chip of silicon into which four separate n-type regions are diffused. The resistor chips are conventional diffused resistors that have available tapped points to achieve any desired resistor value. Electric connections between chips in the hybrid circuit must be made by wire bonding instead of the evaporated and photoresisted metalization that is used in the monolithic version of a circuit.

Fig. 26. Multichip (hybrid) circuit in TO-5 assembly.



Monolithic vs. discrete design

Now we can make a true comparison of the hybrid or discrete circuit performance with the monolithic circuit design that has evolved from the treatments of the previous sections. The latter circuit is specifically related to the case of the 0.1-ohm·cm epitaxial collector that is one mil, or 25 microns thick. The essential "black box" parameters of the detailed circuit are summarized in Table III. From the standpoint of parasitic capacitance, there is an addition of 13.95 pF at the input node, and a total of 6.52 pF at the output node. There is no change in the reverse currents at the inputs in the transition from discrete to monolithic. The forward current is reduced insignificantly because of the slightly greater forward drop of monolithic diodes. In terms of the dc output characteristics, there is a severe degradation of $V_{CE(sat)}$ at 35 mA, where a typical value increases from 0.24 volt to 0.34 volt. The OFF voltage requirements are fundamentally the same, since the reverse leakage current is essentially negligible.

The primary considerations of switching speed reveal the essential effects of the parasitic capacitance introduced

III. DTL circuit performance comparison table

Integrated Circuit Parameter	Hybrid or Discrete Circuit	Monolithic Circuit (0.1 ohm·cm-25 μ epi.)	Buried-Layer Monolithic Circuit (0.5 ohm·cm-12 μ epi.)
Total parasitic capacitance:			
Input node, pF	—	13.95	6.70
Output node, pF	—	6.52	3.12
Input characteristics:			
Forward current, mA	5.8	5.7	5.8
Reverse current, μA	0.01	0.01	0.01
Output characteristics:			
ON voltage at 35 mA	0.24	0.34	0.25
OFF voltage	3.0	3.0	3.0
Switching speed (C_L = 20 pF):			
Turn-on time (R_L = 100 ohms), ns	16	26	20
Turn-off time, ns	26	38	34
Propagation delay time, ns	10	16	13

by the monolithic construction. The switching data shown in Table III are based on a circuit in which 20-pF capacitance is maintained as an output load. The turn-on time is the sum of the delay plus rise times, when the output load is shunted with a 100-ohm resistor. The turn-off time is the sum of the storage plus fall time. In discrete form, the DTL circuit has $t_{ON} = 16$ ns, and $t_{OFF} = 26$ ns, corresponding to a propagation delay of 10 ns.

In the 0.1-ohm·cm monolithic version, we would expect the t_{ON} time to be longer, principally because of (1) an increase in the delay time caused by charging the parasitic capacitance at the input node through the 1 K resistor and (2) an increase in the rise time by charging the capacitance at the output node through the effective load resistance of 80 ohms. The expression for the approximate additional time delay for this circuit is

$$\Delta t_{ON} \approx RC_{in} \ln\left(\frac{V - 0.7}{V - 2.1}\right) + 10\ R_L C_{out} \qquad (9)$$

The first term is the additional turn-on delay as the input node changes from 0.7 to 2.1 volts, when supplied from 6.5 volts. The second term is the additional rise time, and the factor of 10 is the magnitude of the Ebers-Moll function for a beta of 50 and a circuit drive of 10. Inserting the appropriate values into Eq. (7), we obtain

$$\Delta t_{ON} \approx 3.9 + 5.2 = 9.1 \text{ ns} \qquad (10)$$

Allowing about one nanosecond for miscellaneous stray capacitance, we find that the total turn-on time for the monolithic circuit is 26 ns.

The turn-off time for the monolithic DTL circuit is increased according to the relationship

$$\Delta t_{OFF} \approx \Delta t_s + 2R_L C_{out} \qquad (11)$$

Here, the first term is the additional storage time that may result from insufficient storage effects in the series diode. The second term is the increase in fall time that results from the charging of the parasitic capacitance at the output node through the 400-ohm resistor. The factor of 2 is the magnitude of the fall-time Ebers-Moll circuit-drive function. With the appropriate values in (11), the additional turn-off time becomes

$$\Delta t_{OFF} \approx 6.0 + 5.2 = 11.2 \text{ ns} \qquad (12)$$

This produces a total turn-off time of 38 ns, as shown in Table III. Thus, the average propagation delay increases to 16 ns as a result of integrated circuit design. If the series diode in monolithic form could have a recovery time as slow as that obtainable in discrete form, the turn-off time for the monolithic circuit would be as low as 32 ns, and would produce a propagation delay of 14.5 ns. This is difficult to achieve with present technology, and in some cases a circuit must be redesigned by decreasing the value of the 5.6 K resistor to increase the turn-off current during switching.

Buried layers

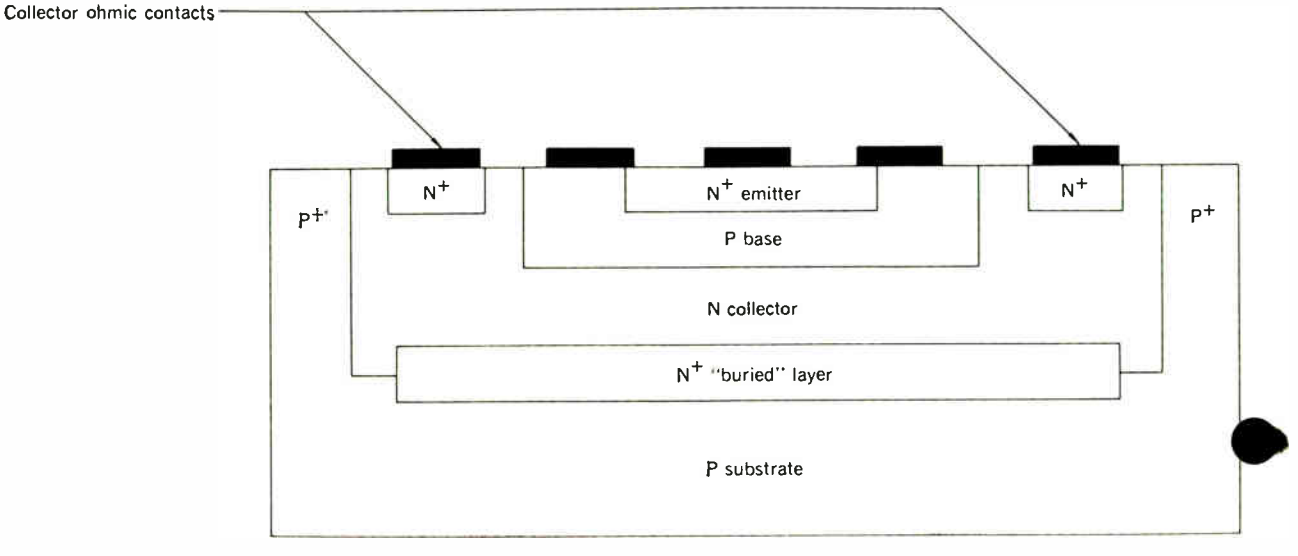
Primarily because high capacitance per unit area is associated with the 0.1-ohm·cm, n-type epitaxial collector, the monolithic circuit introduces larger capacitances.

As previously shown, the resistivity and thickness were selected as a nominal design point to yield $V_{CE(sat)}$ levels that were suitable for digital applications. Considerably lower capacitances can be achieved by increasing the resistivity to 0.5 ohm·cm, as evidenced by the Table I values. This, however, makes $V_{CE(sat)}$ extremely high and untenable.

The introduction of the buried layer step in the process sequence permits the achievement of the 0.5-ohm·cm resistivity level for low capacitances, while still retaining an extremely low $V_{CE(sat)}$. This process modification is best described by reference to the cross section shown in Fig. 27. Prior to the growth of the epitaxial layer, a separate masking step is introduced to permit the diffusion of a heavily doped n-type impurity into the p-type substrate. This n^+ layer is then literally "buried" into the monolithic block since, in the next step, the 0.5-ohm·cm epitaxial collector is grown directly above. From here, the remaining process steps are identical to those described for the 0.1-ohm·cm design.

The striking feature of this approach is that the high

Fig. 27. Utilization of "buried" n^+ layer to reduce collector series resistance.



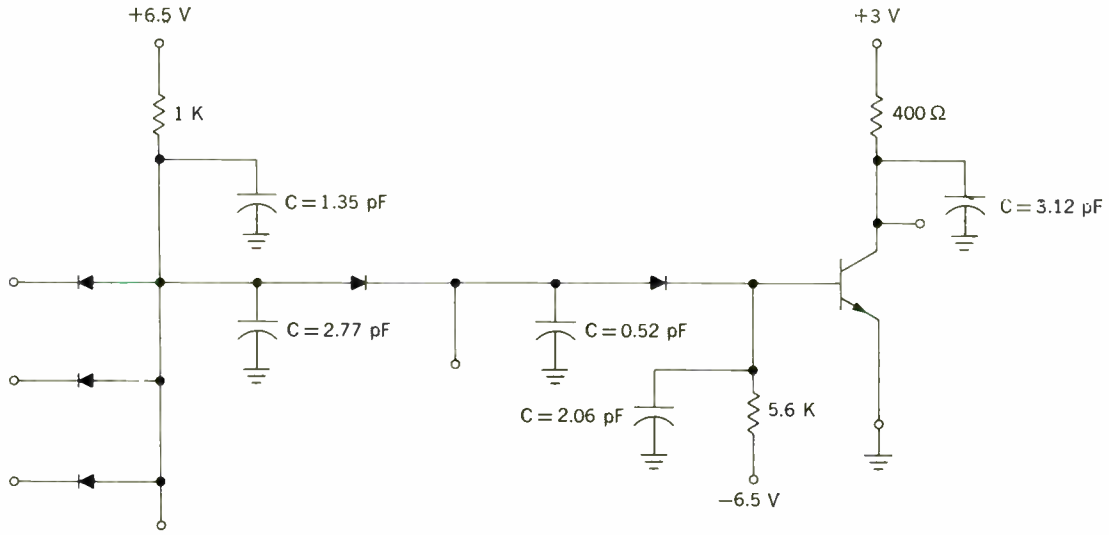


Fig. 28. Reduction of parasitic capacitance values through use of buried layers and 0.5-ohm·cm epitaxial collector.

sheet resistance of the collector is virtually shunted by the low sheet resistance of the diffused buried layer, thereby producing an extremely low collector series resistance. Because of the low sheet resistance shunt, the epitaxial layer thickness can be reduced to approximately 0.5 mil, or 12.5 microns. The design theory shows that the $V_{CE(sat)}$ of the 1 by 6 double-base stripe transistor reduces to 0.25 volt at 35 mA, which is essentially equal to the discrete value. Thus, the use of the buried-layer process, although it requires an extra diffusion and masking step, achieves monolithic transistors that have characteristics equal to their discrete epitaxial counterparts.

The higher resistivity and reduced thickness of the epitaxial collector region, have a compounding effect on the magnitude of the parasitic capacitance, primarily because the isolation diffusion depth is reduced to 0.5 mil; in addition, all isolation islands in the layout can be drawn with a one-mil border instead of a 2-mil border, thus reducing the area of the bottom component, and decreasing the perimeter. The area of the sidewall components is also reduced by the decrease in thickness. Finally, the use of the 0.5-ohm·cm collector permits a lower capacitance per unit area for both the base-to-collector junctions and the sidewall collector–substrate junctions.

When applying these buried-layer considerations to the monolithic design of the DTL circuit in Fig. 2, we must prepare a new reduced area layout, and recalculate all the lower parasitic capacitance values. And, we see that, for the 1 K resistor, the capacitance reduces to 2.69 pF. For the input diodes, there is a reduction in bottom component area from 90 mils^2 to 72 mils^2, and there is also a sidewall component area reduction from 38 mils^2 to 17 mils^2. Coupled with the lower capacitance of the 0.5-ohm·cm sidewall component, a total parasitic capacitance of 2.77 pF is calculated. The same considerations, when applied to the other components, show a similar reduction in capacitance, and these are all summarized in the new, equivalent circuit shown in Fig. 28. Thus, by going from 0.1-ohm·cm monolithic to a 0.5-ohm·cm buried-layer monolithic, the total capacitance at the input node is reduced from 13.95 pF to 6.7 pF. Similarly, at the output node, the parasitics are reduced from 6.52 pF to 3.12 pF, as summarized in the third column of Table III.

This substantial reduction manifests itself in a considerable improvement in switching speed when recalculated on the basis of Eqs. (9) and (11). The results in Table III indicate that the turn-on time decreases from 26 ns to 20 ns. This compares favorably with the discrete value of 16 ns. The turn-off time is 34 ns, as compared to the discrete value of 26 ns. If additional stored charge effects are neglected, this would correspond to about 28 ns for the buried layer device. Nevertheless, we see that the propagation delay decreases to 13 ns.

Other variations

The entire technical treatment presented to date has been limited to the particular DTL circuit of Fig. 2. This is an excellent vehicle to illustrate the application of the basic design principles for digital monolithic circuits. All of the major principles have been introduced in the example and, with proper understanding and use, these principles can be applied to any digital, integrated circuit problem. Although DTL logic was used as the particular example, the design theory is applicable to transistor–transistor logic (TTL), direct-coupled transistor logic (DCTL), resistor-coupled transistor logic (RCTL), and current-mode logic (MECL). In each case, the design procedure is an empirical process that starts from the performance requirements of the integrated circuit. From this requirement a designer evolves a monolithic layout drawing from which he can calculate all of the parasitic elements, and introduce these into his breadboard discrete circuit to anticipate the final characteristics of the monolithic version. Thus, the effect integrated circuits have in blending semiconductor technology with circuit design is apparent. Inevitably, the contemporary semiconductor engineer and the electronic circuit engineer will become the future integrated circuit engineers.

Thin-film circuit technology

Part III—Active thin-film devices

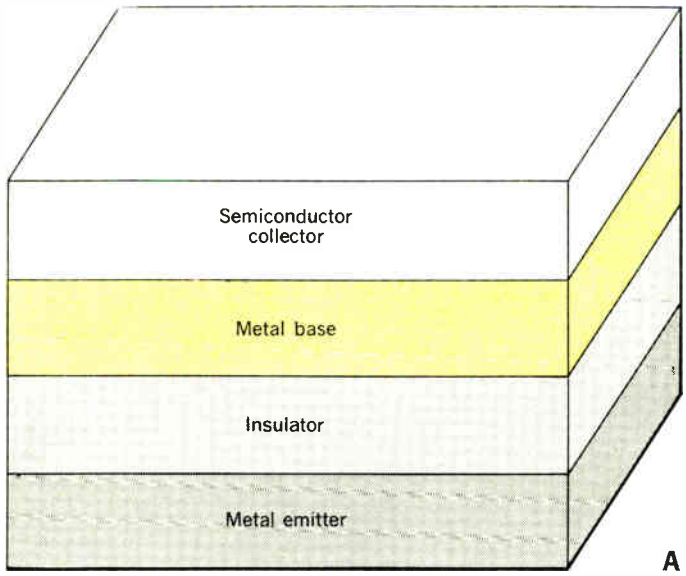
The insulated-gate thin-film transistor is the active device under the most intensive investigation today. Solutions to its problems should simplify the future development of other types of active thin-film devices

Alan B. Fowler *International Business Machines Corporation*

The first article in this series[1] discussed thin-film circuits and the considerable success that has been achieved in the fabrication of inactive circuit elements by evaporative techniques. However, the thin-film technology will not be complete until active devices can be made compatibly with the inactive ones. At present the active elements—usually transistors—are soldered on the circuit boards. The usefulness of evaporated circuitry would be immeasurably increased if the active devices themselves were also evaporated.

Thin-film techniques have been applied to several types of active devices, including cryogenic devices, which were discussed in the second article in this series.[2] These units are a somewhat special case and are not compatible with the usual evaporated circuits because of their low impedance levels and temperature requirements. Of active devices, the most nearly compatible with evaporated circuits are semiconductors. The one of most interest and under most intense development at present is the insulated-gate thin-film transistor (TFT). Evaporative technology can be applied to other active devices, but many of the problems are common to all; therefore, a study of the TFT will demonstrate the types of technological problems to be faced. In this article the emphasis will be on the TFT; but first a brief survey will be made of some other possibilities.

Conventional bipolar transistors have been fabricated by use of a technology compatible with other thin-film techniques. The single crystals desirable in their manufacture must be large enough that a planar transistor can be made within one crystallite. Rasmanis[3] attained some success by recrystallizing silicon evaporated onto molten



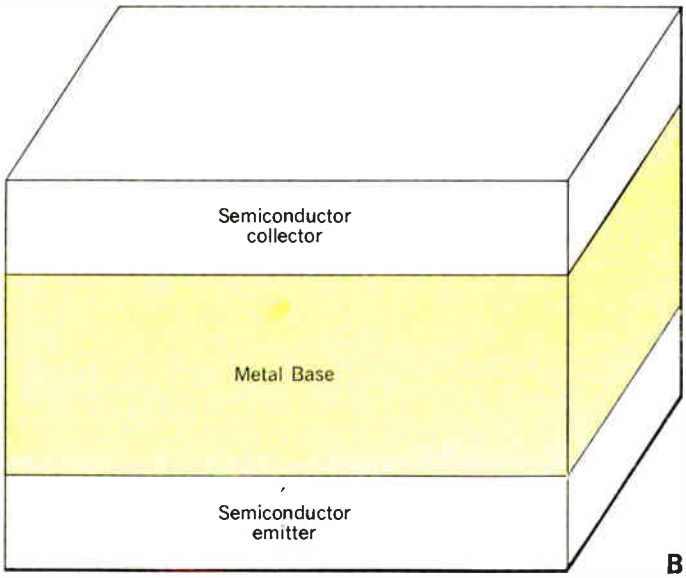


Fig. 1. Two types of hot-electron metal-base transistors. A—Tunnel emitter triode. B—Semiconductor–metal emitter triode.

glass layers. Crystallites up to several microns in size have been made by this technique, and transistors have been fabricated with common base current gains of 0.9 and frequency cutoffs of 50 Mc/s.

This or other techniques for recrystallization or for epitaxial growth may allow some of the forms of metal-base transistors shown in Fig. 1 to be evaporated. These devices, which are similar in operation to conventional transistors except that the base is a thin metallic film, were recently reviewed by Atalla and Soshea.[4] The problems of their fabrication, starting from a single-crystal substrate, are still at the research stage; and thus the realization of completely evaporated, compatible devices of these types is still in the future.

Many parametric amplifiers of the varactor type made by use of thin-film techniques have been suggested and some have been studied. Perhaps the best understood is the silicon–insulator–metal capacitor studied by Lindner.[5] Such devices have large variations in capacitance over a restricted voltage range and can be used for parametric amplification. Other thin-film varactors have been made by means of metal–semiconductor–metal structures, but their principles are not so well understood.

Origin and development of field-effect devices

The insulated-gate thin-film transistor is one of the more recent of many attempts to realize a useful majority-carrier amplifying device by modulating the conductivity of a semiconductor or semi-insulator by induction of charge with an electric field. This family of devices derives from disclosures of Lilienfeld[6] in 1930 and Heil[7] in 1935.

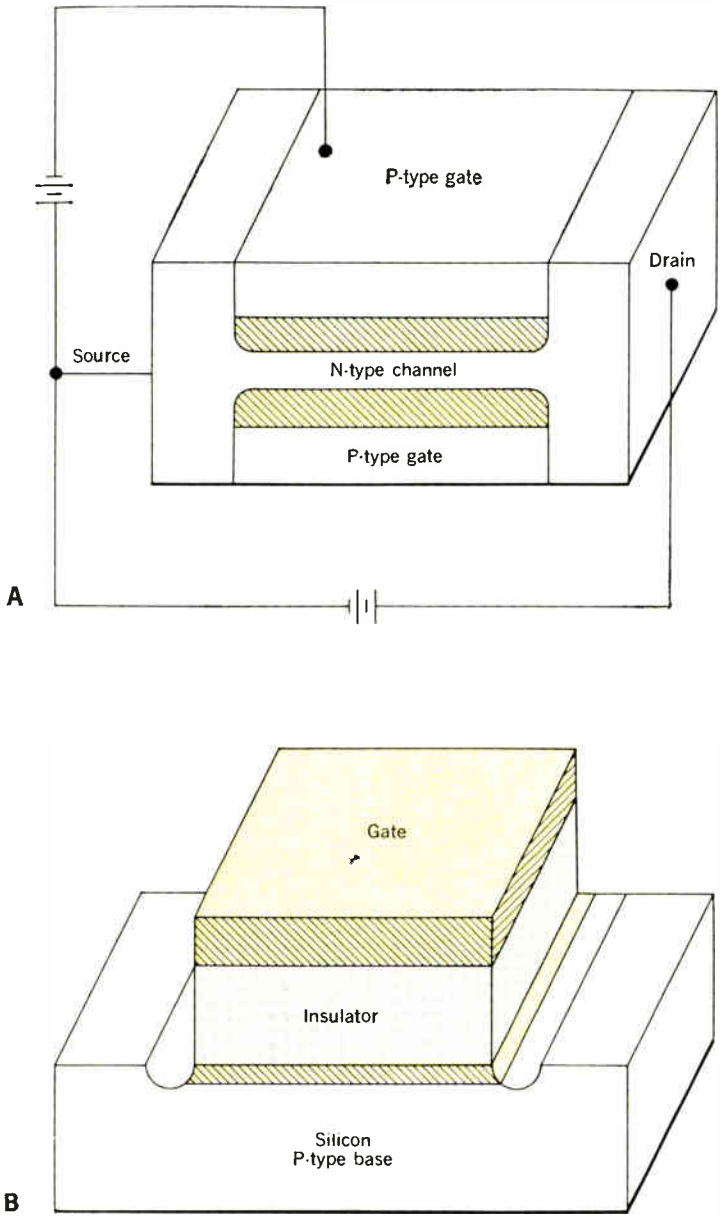


Fig. 2. Field-effect transistors. A—Unipolar transistor. Shaded portion is a space-charge region, which varies in thickness as the gate voltage is varied. B—Insulated-gate field-effect transistor. Dark regions are the source and drain, made by diffusion of n-type dopant into the p-type base. Cross-hatched region under the insulator is an n region induced by application of a positive voltage to the gate.

In the late 1940s attempts by Shockley and Pearson[8] to modulate the conductivity of thin germanium films were successful in the sense that current through the device could be modulated. However, much less modulation was possible than might have been expected. The reason is that most of the induced charge went into localized surface states and was immobilized; therefore, no useful amplification was achieved. These structures were useful in the study of the surface states, and much of our present knowledge of semiconductor surfaces has been derived from investigations of these field-effect structures.

Shockley's[9] unipolar transistor is a field-effect device that operates by the application of a field across a p-n junction and reduction of a channel by the creation of a space-charge region free of carriers; see Fig. 2(A). The problem of surface states is thereby eliminated, but some excess current is introduced from the control electrode. These devices were successfully made and studied by Dacey and Ross,[10] whose analysis has been essentially followed in describing the newer field-effect devices. These units were made by diffusion into single crystals, but no report has been made of an attempt to make them by the use of thin-film techniques.

A device of much current interest is the MOS (metal oxide semiconductor) or silicon insulated-gate field-effect transistor.[11,12] In this device the current flows between p- or n-type electrodes diffused into n- or p-type single-crystal wafers as shown in Fig. 2(B). An inverted surface channel of p or n type may be induced between the electrodes by the application of a voltage to the gate electrode that covers the silicon between the two diffused regions and is separated from them by a silicon oxide layer. These devices are very similar to TFTs and offer many advantages, especially since they are based on a well-developed silicon technology. However, because they are made on silicon wafers, they are not strictly thin-film devices.

Recent interest in insulated-gate thin-film transistors (Fig. 3) fabricated through use of evaporative techniques has resulted from the investigations of Weimer,[13] who showed that useful field-effect devices can be made by a method that employs evaporated films of cadmium sulfide. More recently other semiconductors have been used successfully. The remainder of this article will be devoted to a discussion of these devices.

Electronic characteristics of TFTs

In a circuit sense, TFTs are similar to triodes, although their characteristics more nearly correspond to those of pentodes. The gate corresponds to the grid of a tube, the source to the cathode, and the drain to the plate. This terminology is common to all field-effect devices. The current flows between the source and drain, with the sense depending on whether the current carriers are electrons or holes. The holes or electrons are in equilibrium; no injection is involved. In the discussion that follows, devices with electron currents will be described because they have been studied the most, but the general considerations apply to hole-current devices as well. A voltage V_g between the gate and source modulates the current I_{sd} between the source and drain by inducing a greater or lesser number of electrons and varying the resistance. In electron-current devices the electrons originate from the source, and therefore the drain is biased positive with respect to it.

In general, TFTs are of two types that are said to operate in the depletion or enhancement modes. In the first case electrons are present with the gate short-circuited to the source; thus, current normally flows to the drain. A negative voltage applied to the gate reduces the number of electrons and may be used to reduce the current or, in the limiting case, pinch it off completely. This mode is especially effective when induced holes are incapable of carrying current because of low mobility, as in cadmium sulfide. The enhancement mode is complementary in that the device is normally off and may be

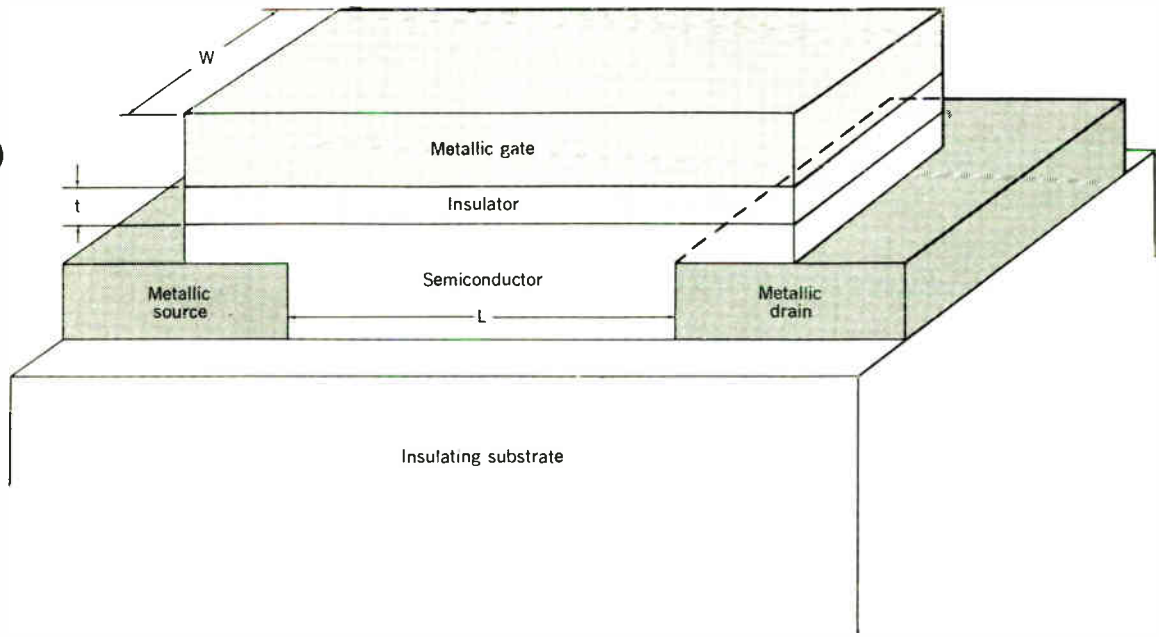
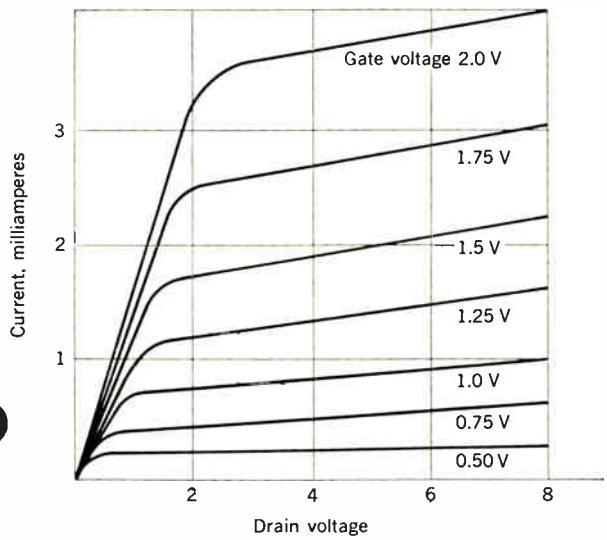


Fig. 3. Schematic diagram of a thin-film transistor. Typical dimensions are $t \approx 10^{-5}$ cm, $w = 2 \times 10^{-1}$ cm, and $L = 5 \times 10^{-4}$ cm.

turned on by applying positive bias to the gate. The possibility of these two modes of operation coupled with the possibility of both hole and electron devices allows great flexibility in circuit design. However, all of these possibilities have not yet been realized.

When the voltage between the drain and source is greater than that between the gate and source, no negative charge will be induced in the region of the semiconductor near the drain. Ideally, when the drain voltage exceeds the gate voltage the current does not increase as the drain voltage is further increased; see Fig. 4. Complete saturation of the current is never observed. Failure to saturate may be caused by space-charge-limited current,[14] by space-charge effects of the sort discussed by Wright,[15] by breakdown currents in the high-field region near the drain, or by leakage currents due to contaminants. The saturation current is one of the least understood characteristics of these devices.

Fig. 4. Typical current-voltage characteristics of an enhancement-mode TFT, for which $V_0 = 0$.



Cadmium sulfide TFTs with extremely useful characteristics have been fabricated. Transconductances of 10 000 μmhos, dynamic output resistances of 8000 ohms, and voltage amplifications of 80 have been reported[16] in units operated with drain and gate voltages in the range of 0 to 10 volts. The input resistance may be of the order of 10^9 ohms with an input capacitance of 30 pF. Useful amplification has been observed as high as 50 Mc/s. Switching times of TFTs have been as low as 30 ns. While these characteristics are not satisfactory for high-speed computer operation, for instance, they are interesting for many applications.

Principles of thin-film transistors

When a charge is induced on the surface of a semiconductor, it may occur as an exhaustion region from which the balancing mobile charge has been driven, leaving an uncompensated ionic charge—either as excess mobile charge of electrons or holes, or as fixed charge in localized states associated with the surface. When the bulk of the semiconductor is of the p type, the surface may be converted to n type and then is said to be inverted. Electrons trapped in the surface states or in the insulator are lost to the conduction process, and the gain is thereby reduced. To some extent this effect may be overcome by use of the high fields (up to 10^6 volts/cm) possible with thin insulating layers (500–2000 Å). Such high fields tend to saturate the surface states which are present in the order of 10^{11} to 10^{12}/cm^2 on germanium[17] surfaces, 10^{11}/cm^2 on

silicon surfaces,[18] and less on cadmium sulfide surfaces.

Because cadmium sulfide has high intrinsic resistivity and low hole mobility, current in the "off" state may be extremely low. Low "off" currents may also be achieved by using very thin semiconductor films. In the case of cadmium sulfide, these films are 1000–2000 Å thick. For materials with lower intrinsic resistivity, such as lead sulfide, it is necessary to use films approximately 200 Å thick.

The operation of TFTs may be understood in terms of a theory presented by Borkan and Weimer.[16] They show that the current in a TFT is given by

$$I_0 = \frac{\mu C_g}{L^2}\left[(V_g - V_0)V_d - \frac{V_d^2}{2}\right]$$

where μ is the mobility of the carriers, V_0 is the gate voltage necessary for onset of drain current (which may be negative for depletion-mode devices), V_g and V_d are the gate and drain voltages relative to the source, and C_g is the gate to semiconductor capacitance ($= \kappa wL/t$, where κ is the dielectric constant of the insulator and w, t, and L are dimensions corresponding to those shown in Fig. 3). In the saturation region ($V_d > V_g - V_0$) the current is

$$I_0 = \frac{\mu C_g}{2L^2}(V_g - V_0)^2$$

Therefore, below the saturation knee the transconductance is given by

$$g_m = \frac{\mu C_g V_d}{L^2}$$

whereas in saturation it is

$$g_m = \frac{\mu C_g (V_g - V_0)}{L^2}$$

The gain–bandwidth product, in the absence of any parasitic capacitance, is

$$\text{GBW} = \frac{\mu V_d}{2\pi L^2} = \frac{g_m}{2\pi C_g}$$

$$\text{GBW} = \frac{\mu(V_g - V_0)}{2\pi L^2}$$

below and above the knee, respectively.

These equations are derived on the assumption that the mobility is a constant independent of the local carrier density. Furthermore, when an effective or field effect mobility is extracted by fitting the equations to TFT curves, mobility values of 20 to 100 cm²/V·s result. Measurements of the Hall mobility are in disagreement with both the assumption of constant mobility and with the high values of effective mobility.[19] The Hall mobility is approximately 10 to 20 cm²/V·s and increases with field. The increase with field is consistent with observations of increase of mobility in polycrystalline CdSe films,[20] where the electron density is increased by light, although similar effects have not been seen in PbS films.[21] There has been some indication[19] that the mobility goes through a maximum as might have been expected if surface scattering[22] played a role at high fields.

Although the foregoing equations are not consistent with the Hall data, they do fit the TFT characteristics surprisingly well and serve as a basis for TFT design. It may be seen that desirable properties are high mobility and small source-drain gaps. The gaps are limited by present technology to 5 to 10 microns. The mobilities in cadmium sulfide devices are 20 to 100 cm²/V·s, but it is reasonable to expect that much higher mobilities might be attained in epitaxially grown films of such semiconductors as silicon.

Fabrication

In principle, the fabrication of thin-film transistors is straightforward. The devices are made by successive evaporations through metal masks. A typical TFT might be made by evaporation of the semiconductor, followed by evaporation of the metallic source and drain electrodes. The insulator is evaporated, and then the metallic gate is deposited. The order of the evaporations can be altered to achieve various permutations of this geometry. The evaporations are usually made in vacua of 10^{-6} to 10^{-7} torr. Most of the materials used may be evaporated from conventionally heated sources. Some refractory metals and insulators have been evaporated by means of an electron beam for heating the sources. This technique has the additional advantage of reducing contamination.

Cadmium sulfide[12, 16, 23–26] and selenide,[27] germanium,[28] and lead sulfide[29] have been employed as the semiconductors. The compound semiconductors must be evaporated in such a way that stoichiometric balance is preserved. This problem is particularly severe in a compound such as gallium arsenide because the vapor pressure of arsenic is much higher than that of gallium. In general, semiconductor should have low intrinsic conductivity so that conductance in the absence of induced charge will be small. Doping may be most accurately achieved by equilibration of the semiconductor in a vapor of the dopant. The stoichiometry may also be equilibrated as in the case of cadmium sulfide by doping to the desired resistivity by heating in sulfur vapor. This technique introduces cadmium vacancies, which are acceptors.

The insulators most used—silicon monoxide and dioxide and calcium fluoride—may be evaporated from conventional sources. Insulators generally present the most serious problems in the fabrication of reliable TFTs. Pinholes, low breakdown strength (the fields in these devices may exceed 10^6 V/cm), deterioration of the insulation in humid environments, and slow relaxation effects are problems that have plagued attempts to make these devices, and evaporated capacitors as well. Refractory insulators, such as tantalum and aluminum oxides, may offer distinct advantages but are more difficult to evaporate. Additional problems may arise if the insulator reacts with the semiconductor; for instance, silicon monoxide reacts with cadmium sulfide by removing sulfur from the surface, and a strongly n-type surface results. This kind of reaction may be useful if a depletion-mode device is being made, but it is undesirable in the case of enhancement-mode devices. It can be avoided by evaporating a sulfur-rich film or by evaporating the cadmium sulfide and silicon oxide in an overpressure of oxygen.[25]

The metals commonly used are aluminum and gold.

Refractory metals are more durable, but they also are more difficult to evaporate. An additional requirement for the source and drain electrodes is that they make ohmic contacts with the semiconductor. The order of evaporation may play a role in making good contacts, especially if the metal easily forms an oxide.

Because the source–drain separation is of the order of 5 to 10 microns, precise alignment of the masks is essential so that the gate does not overlap the source or drain. Overlap increases the capacitance without increasing the transconductance and thus reduces the gain–bandwidth product. Alignment for an individual device may be achieved optically, but the simultaneous alignment of many devices in a large integrated circuit is a more difficult problem.

Fabrication problems common to all thin-film technology apply to the active devices.[30] The various materials must be compatible in that they must not react chemically, must have thermal expansion coefficients sufficiently similar to prevent strains and cracking, and must adhere to each other reasonably well. Problems of this nature multiply as the circuits become more complex.

Summary and conclusions

The insulated-gate thin-film transistor is the active thin-film device farthest advanced for application to integrated circuits. This is not to say that it represents a fully developed, trouble-free circuit element or even that all of the basic problems associated with it have been solved. Nor can it be said that it is necessarily the best possible thin-film device. However, it can be reasonably stated that the TFT shows much promise.

There are several major problems in existing devices. Low-frequency relaxation effects[25] are, in the worst cases, evidenced by hysteresis of the characteristics and slow drifts of the operating points. Many of these effects, which are probably associated with the insulators, or possibly with the grain boundaries, can be eliminated by the use of improved insulators. In fact many individual devices do not exhibit them at all. Another problem is the unpredictable lifetime of the TFTs. Some last indefinitely but others made similarly may have lifetimes of only a few days. Careful control of the evaporation processes is necessary if this type of failure is to be avoided.

Even if these problems are eliminated, the present devices based on polycrystalline cadmium sulfide will be limited in usefulness because of the limited gain–bandwidth product. Until films with higher mobility—such as those that might be attained with epitaxial silicon—can be made, the TFTs will probably be limited to low- to medium-frequency uses in applications demanding the potential low costs of fully integrated thin-film technology. In addition to limiting the frequency response, low mobilities limit the packing and increase the heat dissipation. Therefore, even for low-frequency applications, high mobilities are desirable.

The main problems involved in reducing the TFT to a practical device appear to lie in areas of improving evaporated insulation and in increasing the mobility of the carriers in the semiconductors. More understanding of the interfaces between surfaces of different materials is needed. Since many of these problems are common to all thin-film devices, their solution should improve the chances of making devices other than insulated-gate thin-film transistors.

REFERENCES

1. Lessor, A. E., L. I. Maissel, and R. E. Thun, "Thin-Film Circuit Technology. Part I—Thin-Film R-C Networks," *IEEE Spectrum*, Apr. 1964, pp. 72–80.
2. Caswell, H. L., "Thin-Film Circuit Technology. Part II—Thin-Film Superconducting Devices," *Ibid.*, May 1964, pp. 84–89.
3. Rasmanis, E., "Silicon Thin-Film Diodes and Transistors," *1962 NEREM Record*, p. 160.
4. Atalla, M. M., and R. W. Soshea, "Hot-Carrier Triodes with Thin-Film Metal Base," *Solid-State Electron.*, vol. 6, no. 3, 1963, pp. 245–250.
5. Lindner, R., "Semiconductor Surface Varactor," *Bell System Tech. J.*, vol. 41, May 1962, pp. 803–831.
6. Lilienfeld, J., U. S. Patent No. 1 900 018, 1933.
7. Heil, O., British Patent No. 437 457, 1935.
8. Shockley, W., and G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," *Phys. Rev.*, vol. 74, 1958, p. 232.
9. Shockley, W., "A Unipolar Field-Effect Transistor," *Proc. IRE*, vol. 40, Nov. 1952, pp. 1365–1376.
10. Dacey, G. C., and I. M. Ross, "Field Effect Transistor," *Bell System Tech. J.*, vol. 34, Nov. 1955, pp. 1149–1189.
11. Atalla, M. M., U.S. Patent No. 3 056 888.
12. Hofstein, S. R., and F. P. Heiman, "The Silicon Insulated-Gate Field-Effect Transistor," *Proc. IEEE*, vol. 51, Sept. 1963, pp. 1190–1202.
13. Weimer, P. K., "The TFT—A New Thin-Film Transistor," *Proc. IRE*, vol. 50, June 1962, pp. 1462–1469.
14. Dumke, W. P., "Current Saturation in Field-Effect Devices," to be published.
15. Wright, G. T., "Space-Charge-Limited Solid-State Devices," *Proc. IEEE*, vol. 51, Nov. 1963, pp. 1642–1652.
16. Borkan, H., and P. K. Weimer, "Analysis of Characteristics of Insulated-Gate Thin-Film Transistors," *RCA Rev.*, vol. 24, June 1963, pp. 153–165.
17. Harnik, E., Y. Goldstein, N. B. Grover, and A. Many, "Fast Surface States in Germanium at Low Temperatures," in *Semiconducting Surfaces*, J. Zemel, ed. New York: Pergamon Press, Inc., 1960.
18. Statz, H., G. A. deMars, L. Davis, Jr., and A. Adams, Jr., "Measurements of Inversion Layers on Silicon and Germanium and Their Interpretation," in *Semiconductor Surface Physics*, R. H. Kingston, ed. Philadelphia: University of Pennsylvania Press, 1957, pp. 139–167.
19. Waxman, A., V. E. Henrich, F. V. Shallcross, H. Borkan, and P. K. Weimer, "Electron Mobility Studies in Surface Space-Charge Layers in Evaporated CdS Films," *Bull. Am. Phys. Soc.*, vol. 8, 1963, p. 1539.
20. Fowler, A. B., "Photo-Hall Effect in CdSe Sintered Photoconductors," *J. Phys. Chem. Solids*, vol. 22, Dec. 1961, pp. 181–188.
21. Woods, J. F., "Investigation of the Photoconductive Effect in Lead Sulfide Films Using Hall and Resistivity Measurements," *Phys. Rev.*, vol. 106, 1957, pp. 235–240.
22. Schreiffer, J. R., "Effective Carrier Mobility in Surface Space-Charge Layers," *Ibid.*, vol. 97, 1955, pp. 641–646.
23. Miksic, M. C., E. S. Schlig, and R. R. Haering, "Behavior of Thin Film Transistors," *Solid-State Electron.*, vol. 7, 1964, p. 39.
24. Tantraporn, W., "How Good Are Thin-Film Triodes?" *Electron.*, vol. 35, Dec. 28, 1962, pp. 29–32.
25. Zuleeg, R., "Electrical Evaluation of Thin-Film Cadmium Sulfide Diodes and Transistors," *Solid-State Electron.*, vol. 6, 1963, p. 193.
26. Zuleeg, R., and R. S. Miller, "Space-Charge-Limited Currents and Schottky Emission in Thin-Film CdS Diodes," *Ibid.*, to be published.
27. Shallcross, F. V., "Cadmium Selenide Thin-Film Transistors," *Proc. IEEE* (Correspondence), vol. 51, May 1963, p. 851.
28. Wilson, H. L., W. T. Layton, W. A. Gutierrez, and C. Feldman, "A Germanium Thin-Film Device," *Electrochem. Soc. Extended Abstr.*, 1963, p. 109.
29. Pennebaker, W. B., "PbS Thin-Film Transistors," to be published.
30. Thun, R. E., W. N. Carroll, C. J. Kraus, J. Riseman, and E. S. Wayda, "Thin-Film Circuits," in *Microelectronics*, E. Keonjian, ed. New York: McGraw-Hill Book Co., Inc., 1963, p. 173.

W. L. Everitt

L. M. K. Boelter

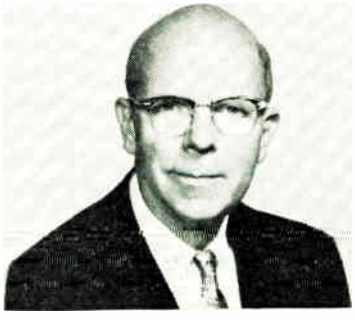
A symposium on continuing education

The unrelenting growth of technology, fed by an ever expanding science, underlines the need for post-baccalaureate education during the engineer's working life. Illustrated herein are some programs under way at the university and industry level

W. L. Everitt *University of Illinois*
L. M. K. Boelter, R. R. O'Neill *University of California*
George E. Moore *Polytechnic Institute of Brooklyn*
Edgar C. Gentle, Jr. *American Telephone and Telegraph Company*
D. G. Ebeling *General Electric Company*
Harold K. Work *Engineering Foundation*
IEEE Committee on Education

Introduction

W. L. Everitt
University of Illinois



Engineering has long been considered not merely a learned profession but also a learning profession—one whose practitioners must first become and then remain students throughout their active years. But concurrent with the explosion in technology characteristic of the past two decades has been an increasing realization that organized programs of a wide variety of types must be made available to support the lifetime of learning in which engineers must engage.

It has been realized for many years that the education of the engineer is not complete upon graduation from the university—whether he stops at the bachelor's level or continues on to a master's or doctor's degree. Because of the limited time available, it is apparent that the universities must restrict their teaching to those subjects that they are best equipped to cover with their faculties and facilities and leave to industry the responsibility of instruction in the fields where it is most capable because of its specialized facilities and personnel. For this reason engineering educators have put an increasing emphasis in academic engineering curricula upon basic fundamentals of science and mathematics—what has been referred to as "engineering science"—and a more adequate introduction to the social sciences and humanities. There are those who have been calling for a reversal of this trend, but the increasing impact of new discoveries in science upon practical engineering and the increasing participation of engineers in the world of affairs make such a reversal improbable if not impossible.

R. R. O'Neill

George E. Moore

Edgar C. Gentle, Jr.

D. G. Ebeling

Harold K. Work

Until comparatively recently, however, the university education and the formal programs of industry have been considered sequential in character—something to be completed at a relatively early age, after which the engineer who is ambitious can depend upon technical literature, professional society meetings, short conferences, and self-study under his own direction for his continuing education.

However, the increasing sophistication of the new and constantly expanding scientific and mathematical ideas, and their continuing influence upon practical engineering, are making it increasingly difficult for the older methods of education alone to satisfy modern needs. The impact of modern computers upon design; of quantum theory upon such areas as nuclear reactors and modern electronics; of plasma and atomic physics upon new concepts of energy conversion and satellite utilization; and of new knowledge in biophysics and biochemistry is determining much of the engineering of the future. Even graduates with advanced degrees find that, in order to be creative leaders, they need the availability of organized programs as part of their own career management.

The Quaker philosopher, Trueblood, has said

"*The terrible danger of our time consists in the fact that ours is a* cut-flower *civilization. Beautiful as cut flowers may be, and much as we may use our ingenuity to keep them looking fresh for a while, they will eventually die, and they die because they are severed from their sustaining roots.*"

Trueblood was commenting upon the lack of religious and moral roots in our modern civilization, and his view deserves the consideration of thinking men in its original context. But I am quoting his metaphor because is is also important that we do not have "cut-flower engineers" in our industrial organizations, in government, or in our universities. Unless these engineers maintain tap roots which can supply intellectual nourishment from the constant flow of knowledge welling up through research and development, they too will eventually sicken or die as effective members of our profession.

This clearly indicates that engineering career planning in the future must provide for periods of formal intellectual interaction with informed associates or teachers, periods devoted primarily to the learning process and distributed fairly regularly throughout the lifetime of a professionally active engineer.

Many articles have been written on the need for programs in continuing education and the combined responsibility of universities and industry in meeting this need. The editors of IEEE SPECTRUM suggested that a group of short articles be assembled to illustrate some of the active programs now being conducted. They felt such a paper symposium would be more appropriate and capable of wider dissemination than ideas garnered from a round table discussion at a convention session. The papers by no means cover the wide range of current activities, but they do represent a sample of programs in being. The more usual advanced graduate programs conducted by many universities, both in residence or by extension, have purposely been omitted as falling into the older pattern of university graduate programs, which are intended primarily to supply education in the early years of an engineering career.

It is to be hoped that this symposium may stimulate increasing activity of a similar nature on the part of other universities, industries, and government agencies. Broad recognition of the need for continuing education in engineering is the necessary first step that has been taken. Now further creative thinking—of the type represented here—is required to fill this need.

The role of the university

L. M. K. Boelter
University of California

R. R. O'Neill
University of California

There are many outstanding programs of continuing education for engineers offered by universities. But rather than use these as a basis for discussion, we shall consider the role of the university in continuing education and the principles upon which programs should be based. Industry and the engineering professional societies have important roles in continuing education, but the university must accept the primary responsibility.

Education is a lifelong activity, not an activity confined to the brief span of years when the person is preparing for a professional career. The institution that has the resources for the student's education during the preprofessional years, the institution that sets standards of performance and has preserved, over the centuries, its spiritual quality should not and cannot terminate its responsibility to the student after he has earned his highest degree. Learning does not end with the receipt of the bachelor's, master's, or doctor's degree. The university must always be at the intellectual frontier of knowledge, an instrument of dissemination. Only in this way can it be of optimal value to society, which it serves in a most crucial way. In the face of an exploding technology, engineering education becomes an important factor in the survival of our society.

For purposes of this discussion, the spectrum of professional education will be divided into three phases. Phase I consists of the undergraduate and graduate education, which provides the base for a man's professional life. The extent of this education determines the point or level at which he embarks upon his career. Phase II provides the opportunity for him to return to the center of learning to study in greater depth. One of the pertinent factors in this phase of post-baccalaureate work is that the engineer can bring experience and maturity to his studies. Phase III keeps the individual stimulated so that he continues to be productive.

The first phase has been the subject of much study. UCLA's Department of Engineering's Educational Development Program and ASEE's (American Society for Engineering Education) Goals of Engineering Education Committee are two recent examples of attempts to improve undergraduate curricula and graduate programs. Phase I is concerned with preparation for the profession. Phases II and III involve continuing education. Viewed in this manner, credit and noncredit courses; degree and nondegree programs; full-time or part-time; and work–study and intern arrangements are all considered as part of continuing education.

The emphasis in Phase II is on the acquisition of knowledge—old or new, or both, but in depth. Examples include the master's and doctoral programs undertaken by the student who returns to the university full time or part time on his own or with financial assistance from industry or the Government. The Engineering Executive Program (EEP) at UCLA is such a program. A sequence of eight graduate courses has been developed to prepare

engineers, already in middle management, for top executive positions in industry. These courses are designed for mature individuals who are able to carry a rigorous graduate program concurrently with an active professional commitment. The average age of the EEP student is 35. The Master of Engineering degree is awarded to those who successfully complete the courses. Other examples are certificate programs consisting of carefully planned sequences of 6 to 16 courses; short courses usually lasting from three days to two weeks; and lecture series on current topics organized by discipline (e.g., modern physics, modern mathematics, modern chemistry) or, by application, (e.g., cryogenic technology).

Phase III is concerned primarily with stimulation. These are the educational experiences designed to keep the mind alert and willing to accept new ideas and, even more important, to act on the new ideas. Fresh avenues of interest are opened and explored through reading, discussion groups, lectures, and workshops. There are some cases where a particular program may belong in Phase II or Phase III, depending upon the individual. As an example, the "cryogenic technology" lecture series could provide the necessary new knowledge for someone engaged in the frontiers of technology or it could be the stimulus for an engineer working in a related field who seeks the stimulation found in becoming aware of new ideas.

These programs involve education, not training, and are the responsibility of the modern American university —the multiversity. However, continuing education programs, especially the successful ones described, are the result of close cooperation between the university and industry. In some cases even the leadership for specific programs comes from industry or government. Although the university has a long history of disseminating, storing, and searching for information, it no longer holds a monopoly of these activities in the technical fields. Industry often participates not only in the planning and organization but in teaching as well.

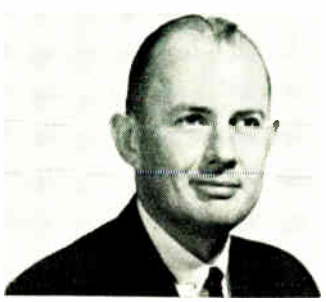
There are still several important challenges which must be considered by the university. Meeting these challenges is fast becoming an urgent matter. Over the years the universities have evolved valid criteria for admission to the educational programs in Phase I. The criteria for programs in Phases II and III are less developed and less valid. It is true and accepted that not every student who earned just an average record of scholarship as an undergraduate is capable of completing the work for an advanced degree—but many are. These are known as "late bloomers." In general, the university has no way of determining their potentials. But if our human resources are to be exploited, the problems of admission must be faced realistically, giving some recognition for the achievements that the individual has made in the profession. Patents, publications, and professional performance should be given the same serious consideration as the promise of grades received years earlier. Admitting the right students to the right programs is basic to the success of the educational activity.

Second, the university must provide the proper recognition for achievement, but here it must look to industry and government for assistance. Degrees are coveted. Industry and government put a premium on them because universities have maintained standards of quality over the years. A degree is negotiable; an individual takes it with him wherever he goes. The university is also capable of maintaining standards of quality in nondegree programs. UCLA's programs called Engineering and Management, and Modern Engineering for Engineering Executives, and certificate programs in propulsion and power conversion systems, and in nuclear technology are examples of quality work. To exploit noncredit work with its tremendous potential, industry must find a way of recognizing this kind of achievement by its employees.

Finally, the university has the responsibility for educational programs which look far into the future. Care must be taken to make certain that the treatment for each topic opens doors and makes it possible for new facts and concepts to be securely bonded to the old so that they become a part of generalized knowledge. Because this is the essence of continuing education, the university must be central. For centuries the university has had few if any peers in the manner in which it has accomplished the building of knowledge on knowledge so that each generation is able to think creatively and to break through with new insight and wisdom.

Education for survival

George E. Moore
Polytechnic Institute of Brooklyn



Technological revolution, information explosion, systems automation, and data retrieval are exciting words to the recent, or soon-to-be, graduate engineer. These are more than words. They are challenging, changing, accelerating, and exhilarating idea-worlds in which a young engineer may contribute and profit. With a foundation of modern mathematics, materials science, systems, and communications—and trained in the modern engineering method—the graduate is well prepared to launch his career.

These words, however, may generate apprehensive

feelings in the not-so-recent graduate engineer of 10, 15, or more years. The idea-worlds they conjure up may lie outside his ken of experience, rendering his engineering knowledge inappropriate to present-day needs. With a foundation of classical mathematics, preatomic physics, and ac and dc circuits, and trained in empirical methods, the older graduate is not well prepared to modify or change his career.

Pre-1953 engineering graduates—and according to the March 1963 issue of *Chemical Week*, 300 000 B.S. graduates of the 1942–1953 era are in this category—are at a distinct disadvantage. The majority of these men have not been working at advanced research and development levels where the changing demands of the work both force and help a man stay abreast of new developments. On the contrary, many have, by corporate necessity, developed engineering specialties of a close-to-routine nature in product areas that are now threatened with or have already succumbed to obsolescence. Such engineers are themselves being typed as obsolescent—a matter that is receiving growing attention and publicity (perhaps to the detriment of new engineering enrollments) at a national level.

In general there are but two courses of action available to the experienced engineer. The first is to attend the many new lecture series being sponsored by professional groups. Lectures on plasmas, control, solid state, and lasers attract engineers in increasing numbers. All too often, however, these lectures are not instructional, not at the anticipated level, and in the final analysis more frustrating than fulfilling; they are basically informational, providing benefits primarily to those who are currently close to or involved in the new emerging technical areas. Second, full-time or part-time graduate study also has inherent limitations in satisfying the educational requirements of the experienced engineer. Modern graduate work covers a domain of knowledge based on today's undergraduate fundamentals—not those of 10 or 15 years ago. Each graduate course builds on the knowledge and fundamentals covered in preceding courses and, in turn, provides a base for more advanced study.

While for some time practicing engineers have encountered this educational "void," only recently have some engineering institutions (and some industries) seen the need for developing new programs and approaches to fill the technological gap and modernize the engineering base of pre-1953 graduates.

At the Polytechnic Institute of Brooklyn, through its new Office of Continuing Professional Studies, a broad and comprehensive program is beginning to evolve.

The program for executive technical development

In early November of 1963 the Institute announced this Modern Engineering Program to industry. Designed to update engineering managers and directors, the course will focus on those areas of engineering and science which are continuing to accelerate and change our technology. The major areas will include modern mathematics, materials science, systems engineering, communications, electromagnetics and plasmas, and developments in applied mechanics.

A majority of the lecturers will be senior professors of the Institute, supported by eminent educators from other universities. This program was conducted on a full-time basis during the month of April 1964.

The modern engineering series

In April of this year, the first of a series of three one-year courses got under way. The first course, Modern Mathematics and Engineering Applications, attracted 37 participants sponsored by these companies located in Long Island and surrounding Metropolitan area (60 companies had been invited): Airborne Instruments Laboratory, General Telephone & Electronics Laboratories, Grumman Aircraft Engineering Corporation, Instruments for Industry, Inc., Long Island Lighting Company, New York Telephone Company, Republic Aviation Corporation, Sperry Gyroscope Company, U.S. Merchant Marine Academy, U.S. Naval Applied Science Laboratory, U.S. Naval Training Device Center, and Wheeler Laboratories, Inc.

The course, held at the Graduate Center in Farmingdale, N.Y., every Monday morning from 8:30 a.m. to 12:30 p.m., has been specially designed to develop both a conceptual and computational understanding of linear systems analysis, feedback theory, control systems, analog and digital simulation, numerical analysis, probability, statistics, modern communication theory, and reliability. The participants, to our initial surprise, were representative not of practicing engineers but of the higher levels of engineering management. The two principal reasons for attending advanced by this group were (1) to update and strengthen their technical backgrounds, and (2) to evaluate the program as a means of strengthening the technical proficiency of practicing engineers who would be released for attendance in 1964. As of this date, the class has held six meetings with the overwhelming judgment being that the course is fulfilling its objectives. A few statistics on the participants show that:

1. The average participant received his B.S. degree in 1946.
2. The 35 per cent with M.S. degrees completed their degree requirements, on the average, in 1948.
3. Thirteen different degree fields are represented, with 50 per cent having degrees in electrical engineering, 10 per cent in aeronautical engineering, 10 per cent in mechanical engineering, and the remaining 30 per cent in chemistry, physics, mathematics, civil engineering, etc.
4. Of the total group, three are vice presidents, 11 are engineering directors or department heads, 14 are project or section heads, and 8 are supervisory or senior engineers.
5. Alumni of more than 30 universities are represented in the course. Such universities include Iowa State, Cornell, M.I.T., and Arkansas. Eleven participants are alumni of the Polytechnic Institute.

Plans have been made to offer the modern mathematics program in 1964 at both the Graduate Center and at the main campus in Brooklyn. Also, two new courses of the series are in the development stage: "Modern Physics and Engineering Applications," and "Engineering Science and Engineering Applications." The first-named will be scheduled at the Graduate Center in the fall of 1964.

SCORE: Short Courses on Research and Engineering

The Polytechnic Institute has, during the past year, increased the number and diversity of short courses

available for research and development engineers and scientists in industry. Courses in X-ray diffraction and polymer chemistry have attracted national and international attention. To these have been added courses in microwave field and network techniques, optimal control theory, space communications, optimal stochastic control theory, modern metallurgy, and others still in the planning stage.

These concentrated courses are proving to be a needed and an effective means of presenting significant theoretical developments and new important applications to those whose work is in the same field or closely related fields. These short courses literally tend to minimize the informational time lag between significant research endeavors and their eventual publication.

Looking to the future, the domain of continuing professional studies offers much in the way of challenge and opportunity to the Polytechnic Institute. The need and the "market" for continuing professional studies are limited only by the constraints of available, qualified teaching staff and facilities. We are even now discussing ways and means of developing (1) an effective approach to updating or strengthening engineering faculty, (2) a program for members of the House and Senate who are not technically trained but whose decisions are more and more shaping the technological future of the country, and (3) industry's broad support of the idea that programs of technical development must be factored into the engineer's career on a planned and continuing basis in a manner that has come to characterize industry's programs in management development.

To close the loop, the 1963 graduates of whom I spoke glowingly are prospective students of the Office of Continuing Professional Studies in 1973 or sooner!

A challenge to industry

Edgar C. Gentle, Jr.
American Telephone and Telegraph Company

The overall situation in the education of engineers by industry in general and the communication industry in particular is influenced by two trends:

1. Design. The four-year college of today does not cover the application of engineering principles in the detail and manner required by industry.

2. Speed of development of science and technology. For example, a technologically sophisticated communications network has been created wherein innovation and application are separated by a steadily narrowing time interval. The total system grows increasingly complex. To maintain his effectiveness, the engineer of today must adapt to the continuing need for education throughout his professional life.

These trends present a heavy challenge to industry. Let us look at the situation at the Bell System, and the solutions now under way to meet it. The development of the Bell System into a fully mechanized network entails increased capital expenditures and engineering activity of more than usual flexibility. This situation means that the Bell System engineer must not only keep technically up to date but must broaden his concepts to include the economic impact on company business of his activity. Among the educational and training requirements are re-educating the older engineer in new theoretical knowledge, challenging the graduate to practical problems, and presenting to all engineering personnel the study of common management programs.

Two categories of programs

A good part of the education and training is obtained on the job and through self-development. The two categories of formal study programs are (1) academic programs conducted on campus, and (2) programs more directly oriented to our specific needs, which are conducted off campus by the Bell System.

Regional communications. The program of regional communications engineering falls in the first-category and is conducted on several college campuses under this name. It is a basic academic program for qualified engineers in the operating companies of the Bell System.

Conducted entirely on college campuses—one Canadian and five American—and taught by professional educators, this program combines electronics, telephone transmission, switching theory, and engineering economics at the senior and graduate levels. College credit may be granted. The program operates for 20 to 24 weeks in sessions of 4 weeks each during a 12-month period and is considered to be education rather than training. All students are graduates of regular four-year technical courses, of which one third are in electrical engineering. Many students, in addition, have taken nonengineering majors. Of the 280 students in the 1963-1964 program, none has been out of school less than three years nor more than ten. The validity of the regional program as a valuable service is under continual critical examination.

Operating engineers. The training program for operating engineers is a technological program of the second

type and is conducted at the Bell Telephone Laboratories. Sessions are of 18 months' duration.

The principles and concepts underlying new devices, circuits, and total systems under current development for the Bell System are covered. Five work assignments enable the trainee to test, consolidate, and apply his academics by participating in development projects. These are the equipments and systems of the future that are to be integrated into an existing network. Trainees have been out of college an average of seven years.

At the Laboratories, students spend one third of their time on graduate level academics, covering principles and concepts. The program is flexible and it changes with the technology. Studies bridge the gap between college and the current state of engineering and science. The remaining two thirds of the time is spent on rotational work where the trainees learn by doing. The plan is to provide the trainees with a general view plus an intensive appreciation of particular developments from either a systems engineering, a systems development, or an apparatus development standpoint. The graduates and the new systems and apparatus are "old friends" when they meet again in their home companies. In addition to this thorough grounding in the new arts, our men receive considerable classroom work in the field of engineering economics.

Data communications training program. Conducted at Cooperstown, N.Y., the program in data communications training is of the second type and directly related to specific applications of current problems. The emphasis is on the total communications system and provides the graduate with a deep understanding of the Bell network and how it meets the customers' complete systems requirements. Practical instructions include a wide variety of Bell System test and communication equipment as well as an understanding of representative business machines and computers. The program's aim is to make the graduate immediately effective in the data communications field.

The course material is carefully selected to avoid duplication and overlap with the programs previously discussed. This is a 12-week live-in program. Attending engineers generally have had their bachelor degrees in electrical engineering an average of 12 years prior to taking the course.

Course study involves 300 contact hours of classroom and laboratory work, regular homework, and examinations, making the work load equivalent to that of a university. Engineers have been trained at the rate of 150 per year; they come from the United States and Canada.

Another program just getting started in this category is an engineering course relating to the introduction of electronic switching into the communication network. This course considers message switching as well as line switching concepts and is a live-in, one-month program covering principles, features, limitations, and economies as well as problems involved in integrating these newer types of systems into an existing switching network, predominantly electromechanical.

Maintaining engineering proficiency

D. G. Ebeling
General Electric Company

In recent years, General Electric has conducted a new educational program for its senior technical managers. This six-week course is conducted for a class of 25 engineering managers at an off-the-job facility. Approximately 100 survey lectures on fundamental scientific theories and new methods of engineering analysis are presented by 25 visiting professors and company experts.

The basic objective of the course is to give the participant an integrated, overall view of the major advances in science and engineering in recent years.

Course content and organization

Since the physical world cannot be described accurately without recourse to mathematical models and operations, this course consists of four complementary parts: (1) mathematics as a "descriptive language"; (2) science, or the "behavior of natural systems"; (3) engineering, or the "design and evaluation of synthetic systems"; and (4) environment or "social, marketing and economic trends."

Mathematics. Many problems of modern engineering involve random processes and decision making on the basis of inadequate or uncertain data. To deal with these probabilistic situations, one of three groups of mathematics lectures is: Boolean algebra, set theory, probability, statistical inference, decision theory, reliability, and simulation techniques.

The second group of mathematical techniques was originally developed for the analysis of electric circuits but these methods can now be applied to mechanical, hydraulic, and thermal systems. All these problems can be handled as dynamic systems represented by networks of lumped parameter components. For these system problems, the topological ideas of networks, branches, loops, and transfer coefficients, and the techniques of signal flow diagrams, simultaneous differential equations

and Laplace transforms are shown to be powerful tools.

The third group of mathematical ideas are those that are useful in the analysis of potential field or distributed parameter problems, such as may arise in electrical, magnetic, gravitational, and stress fields, and in heat conduction, fluid flow, and mass transport. For these purposes, complex variables, vectors, matrices, and the concepts of vector calculus are particularly useful.

Science. The science lectures describe the natural physical world: various forms of matter and energy, and their various interactions. The opening physics lectures on the structure of the atom, the nucleus, and the elementary particles of matter serve to introduce the quantized nature of matter and energy. The basic concepts of relativity theory are then discussed and the intimate connection between electricity, magnetism, and the invariant speed of light is demonstrated. Using these concepts, the relativistic transformations of mass, distance, and time, and the all-important mass–energy equivalence are deduced. The concepts of quantum and statistical mechanics are used to develop the basic models of solid-state physics and to describe the structures and properties of metals, semiconductors, ceramics, and polymers. The important structural features of materials are described, such as crystal lattices, molecular geometries, grain textures, dislocations and vacancies in metals, and the polymerization and cross linking of polymer chains.

Engineering. A unified approach in our survey of engineering analysis methods is attempted by classifying engineering problems into three general groups, corresponding to the mathematical concepts used in their solution: decision problems involving random variables, systems problems involving lumped parameter components, and field problems involving distributed parameters. In recent years, mathematical analysis has made great strides in the sense of being successfully applied to much more complicated situations. While the initial application of these new methods requires a rigorous mathematical development, often this logic can be converted into adaptable, numerical-analysis computer programs. By using adjustable parameters to specify the boundary conditions and restraints, it is possible for one "generalized computer program" to handle an almost unlimited number of problem variations. Furthermore, rapid strides are being made whereby the engineer can express his problem in familiar engineering terms, and these new computer programs can process this input without any elaborate, intermediate programming efforts.

Environment. The engineer must be sensitive to society's ability and desire to pay for his designs. He must recognize the economic limitations of his organization to invest capital in his designs, as well as the vital role played by marketing. Finally, his designs must anticipate the features and the styling that will motivate the ultimate customer. For these questions, a number of our company experts discuss industrial design, economic forecasting, and marketing considerations.

Conclusion

The course, which began with basic concepts of mathematics, quantum mechanics and relativity, concludes with consideration of the vital impact upon each other of an explosive technology and a rapidly changing social–economic environment. This course illustrates the variety of sophisticated knowledge required for a professional approach to today's engineering problems.

However, a problem still exists. Modern engineering is so voluminous and complex, that an unprogrammed, casual effort at self-education does not seem to be a realistic program to ward off the phenomenon of technological obsolescence. The solution may be that we must never permit a new graduate to get out of the learning habit. This could be done by creating the proper job situations, incentives, and work schedules to ensure that a definite portion of each engineer's time is regularly devoted to the process of formal education.

An American Institute for Continuing Education of Engineers

Harold K. Work
Engineering Foundation

Currently, engineers are confronted with a huge mass of new scientific information. They must learn how to assimilate the new information and continue to assimilate it as it is produced during their working careers. This problem is not new but its importance has greatly increased as a result of the massive support of research in this country in recent years. The broad solution to the problem is designated as continuing education; the term applies only to nondegree work.

Ten or fifteen years ago a large part of the continuing education of engineers came from the professional societies. More recently, industry and the universities have

become deeply involved. The present situation is a confused mass of activity, and there are those who feel that the time has come to inject some degree of organization into the problem of continuing education.

One suggestion offered recently[1] involves the formation of an organization similar to one initiated by the medical practitioners in 1947. This suggestion has aroused considerable interest and a possible way to implement it will be briefly described. The suggested organization will be referred to as the American Institute of Continuing Education. The effectiveness of such a plan will, however, depend upon the full cooperation of the engineering societies, the universities, industry, and the engineers themselves.

The needs that justify such an organization are (1) a need for some centralizing agency for evaluation and dissemination of information, (2) a need for status or recognition of accomplishment for the ones who put effort into keeping up to date, and (3) a need for guidance for those wishing to set up programs.

If these needs are accepted as valid, a plan at the national level to improve and expand continuing education would seem desirable to keep the engineer abreast of the rapid advances of scientific and technological knowledge. Such plans already exist in medicine, metallurgy, and management. They are represented by the American Academy of General Practice, The Metallurgical Engineering Institute, and the American Management Association.

To meet the needs of engineers, it has been suggested that the aforementioned American Institute for Continuing Education of Engineers might well be established as a cooperative body of the national engineering societies. This national organization, it is presumed, would be developed by and affiliated with the Engineers Council for Professional Development (ECPD), the Engineers Joint Council (EJC), and the American Society for Engineering Education (ASEE), with the primary responsibility assigned to one of them.

To guide the activities of the Institute, an advisory committee would undoubtedly be established, composed of leading engineers and educators with experience and interest in engineering education.

The objectives of the organization might be outlined as follows:

1. To assist the engineer throughout his career to maintain a high degree of technical competence.
2. To provide professional education and training designed specifically to fit the needs of experienced practicing engineers.
3. To involve industry, government, academic institutions and the professional societies so that the full force of the engineering profession can be brought to bear on this problem.
4. To make engineering a lifelong, exciting, and rewarding profession and attract more of our qualified young people to technical careers.

It might be desirable for the institute to include the following activities:

1. A national register of industrial programs for continuing education.
2. A national register of engineering school programs.
3. A national register of society programs.
4. Development of new methods of continuing education, such as teaching machines, information retrieval techniques, radio, television, homestudy, etc.
5. Development of engineering and science-oriented monographs, aimed specifically at practicing engineers.
6. A consulting service to industry, engineering schools, and professional societies.
7. The development of certification criteria and the practical implementation thereof.

Chapters or division of the engineering societies throughout the country should help implement these activities on a nation-wide basis.

Continuing education of scientists and engineers has assumed major proportions so that coordination of such activities through a national organization such as the American Institute for Continuing Education of Engineers is needed. Such an organization would affiliate itself with the existing professional societies.

This plan has been presented only as a basis for discussion.

REFERENCE

1. Friberg, Bengt, "The Challenge of Continuing Technical Excellence," *J. Profess. Practice* (*Proc. ASCE*), Jan. 1963.

Post-baccalaureate education in industry: A report of the IEEE Committee on Education

This report brings up to date the work of the old AIEE Committee on Education. Its one-year survey of post-baccalaureate education in industry, completed in January 1963, forms the basis of the report.

For purposes of the study, "post-baccalaureate" is defined as a broad field of education that includes, in addition to formal graduate study, programs of continuing education for the professional development of engineers after completion of their baccalaureate programs. No attempt to draw conclusions was made, outside of a few obvious generalizations.

For practical reasons, the survey was confined to the electrical engineer in the electrical industry, disregarding the other fields of engineering specialization. It is reasonable to assume that whatever is being done for the electrical engineer in a particular industry is also being

done for the mechanical, chemical, and other engineering types in the same industry.

A questionnaire was mailed late in 1962 to 246 industrial organizations selected for diversity in size and geographical distribution and six major departments of the Federal Government, employing large numbers of engineers. Three types of companies were included: manufacturers, electric utilities, and communications utilities. Of the 128 companies replying, approximately 88 per cent reported some type of post-baccalaureate educational program. Almost 145 000 engineers are employed by the respondent companies.

Procedure for survey

The questionnaire delineated three distinct types of educational programs, as follows:

1. Graduate study assistance programs. Programs of assistance, financial or otherwise, to employees studying for advanced degrees at recognized institutions of higher education outside of the company. Such programs may involve part-time or full-time study.

2. Formal company training programs. Prearranged structured programs, perhaps involving rotating work assignments, a planned study program, or a combination of work and study (excluding graduate-degree-producing programs).

3. Company-sponsored courses. Generally individual courses, developed to meet specific needs of groups of present employees, taught by either company or outside instructors, or conducted by other companies.

Graduate study assistance programs

Graduate study assistance programs were the most frequently used type of post-baccalaureate educational program covered in the study. Since the large majority are tuition refund programs of some type, the emphasis in this portion of the report will be on the details of some of these tuition refund plans.

Tuition refund programs. Of the 128 companies responding, 103 report graduate study assistance and 84 of these are tuition refund in character.

Tuition refund programs can be found in very small as well as very large companies; in fact, some of the more liberal arrangements were reported by small companies who, perhaps, are able to give more supervision to their student employees. The percentage of tuition refundable by the company varies widely, with the two common fixed amounts being 50 per cent and 100 per cent.

The great majority of the manufacturing and electric utilities have set no maximum amount that the company will pay. Among communications utilities, however, all those participating in the survey report maximums of $165 to $350 annually. The few respondent manufacturing and electric utilities report maximums from $80 to $1000 annually, with most refunds falling between $200 and $500.

Virtually all companies surveyed required, for tuition refund consideration, that a course be job-related and "satisfactorily completed," which is generally defined as a passing grade.

Loans and time off. Only eight companies, four manufacturing and four electric utilities, reported company-sponsored loan programs.

The practice of allowing time off from work to attend classes under tuition refund seems to be more common among manufacturing firms than among the other two categories. Forty-two of the manufacturing firms have some provision for allowing such time off, whereas only eight of the electric utilities and none of the communications utilities do so. Most companies have established limits on a per course, per week, or per year basis.

Only 13 of the manufacturing and two of the electric utilities reduce salary when time off is granted. Interestingly, most of those who do are among the larger companies in each category.

Leave of absence. Sixty companies reported a willingness to grant leaves of absence for educational purposes.

The granting of a stipend or living allowance to employees on leave of absence is limited and is directly related to company size. Fifteen manufacturing companies and two public utilities will consider such grants; in no case do any of these firms employ less than 500 engineers.

Other programs. In addition to or in place of tuition refund programs, several companies report programs where all or part of the tuition for advance degrees is paid directly by the company. These are generally much more formalized than normal tuition refund programs with students pursuing a specified degree within a set time period.

Participants in such programs are usually carefully selected, with the company exercising more control over the courses taken, general progress, etc. Such programs often involve a certain amount of time off from work, sometimes at full pay, to pursue courses. A few of the more liberal programs allow full time off with pay during class time, although these are the exception. Such graduate study programs may be found in companies of all sizes—some quite small.

Perhaps the most unusual program is in the research laboratory of a major company where participation in the master's producing program is mandatory for all new employees entering at the bachelor's level. Another unusual program, again in a large company, is a full-time, full-pay arrangement for the three years of doctoral study. One hundred selected students are now benefiting from this arrangement.

Participation. Many companies did not attempt to give the percentage of their engineers participating in graduate assistance programs. Where this figure was given, it was generally rather low.

The question of legal or moral obligation of program participants to remain with the company drew an interesting response. Only the two service organizations (Army and Air Force) reported a legal obligation. More surprising, however, was the fact that only 17 companies felt that participants had even a moral obligation to remain with the company. None of the respondents could see anything but an expanding future for such graduate assistance programs.

Formal company training programs

The majority of the formalized programs are for the orientation of new employees. These programs are therefore usually found in the medium to large companies where the annual influx of new engineers is sufficient to justify such a program on a continuing basis.

The number of participants during the course of a year varied according to company size or by the

number of engineers employed, ranging from 3 to 1650.

Many of the formal training programs involve a combination of work and study. Some of these utilize alternate periods of work and classroom study over the time period, while others intermix both work and study throughout the entire period. Selection of participants, in many cases, is automatic upon employment, especially in the case of initial orientation programs.

In addition to the orientation programs, some companies report a course content of highly specialized advanced technology. Illustrative of these are an instrument manufacturer with an initial training program that includes courses in electrical and temperature measurements, heat treating, automatic control, etc.; and a communications utility with an updating program given after three to five years' employment with courses in advanced communications theory, information theory, engineering economics, etc.

Another example of the in-company advanced technical program is that of a company that offers carefully selected young engineers three one-year courses in varying depth. The purpose is to help participants apply more effectively the fundamental principles of engineering as learned in college to the solution of actual engineering problems. Classes meet 4 hours each week, on company time with about 25 hours per week of outside study required. Three colleges are currently considering this program, with slight modifications, for graduate credit.

Company-sponsored courses

The term "company-sponsored courses" describes educational courses, other than formal programs, to meet the specific needs of employees or to correct certain deficiencies in their education. Although the variety of such courses is greater than in either the graduate assistance or formal training programs, few companies had developed a continuous program of uncovering present or future employee needs and the training program to meet them. In this area, the majority of companies rely largely on intuition.

Company-sponsored courses differ considerably in content and duration, and are conducted both during and after hours. Courses taught during working hours are usually conducted on company premises and staffed by employee instructors. Selection of students is made by management, and attendance is usually mandatory. The after-hours programs are frequently taught by an employee who is compensated for this extra work. Some of the larger companies use faculty members from nearby educational institutions to teach after-hour courses, either on the campus, if close by, or on company premises.

Of the wide variety of courses existing in this area, it may be helpful to describe a few examples. One large manufacturer uses the instructional staff of several educational institutions and the facilities of a local school system to conduct noncredit courses after hours. Enrollment is voluntary and students are charged no tuition. Courses are approved and sponsored by departments of the company, based on the needs of their employees. This is the maximum use of outside instructional assistance and facilities revealed by the survey.

Another large manufacturer has a highly developed program of individual courses, mostly at graduate level, which are taught at several plant locations. Courses are organized into central fields of study, such as optical systems, systems engineering, programming, etc., with a core curriculum and recommended electives. Planned study programs are arranged for individual engineers. Courses are developed by a permanent central education staff and are taught primarily by company employees, both during and after hours.

Of the 51 companies employing less than 200 engineers, only 11 conduct any kind of company-sponsored course. On the other hand. 26 of these same companies had graduate assistance programs. While it may be difficult for the small company to organize and maintain an internal program, there is evidence of this being done. Four companies with 75 or less engineers reported one or more courses being conducted. At the other end of the spectrum, one company, employing 4200 engineers, reported no internal training program, apparently having elected to depend entirely on established courses included under a graduate assistance program.

Twenty-six companies report courses specifically directed to the "updating" of engineers. In this connection, a number of companies mention problems associated with the education of older employees. These include home problems, lack of motivation, lack of advanced mathematics background, outside activities, and poor study habits.

Summary

No effort has been made to relate, in any quantitative way, the data obtained in this survey to the entire electrical industry. However, considering the total number of engineers (145 000) employed by the 128 reporting companies, the diversity in size and geographical location of these companies, and the varying types of businesses represented, this limited sample should be representative of the kinds of educational programs being conducted and of the relative extent of this activity.

The general tone of the replies indicates an increasing interest on the part of industry in this field of education. Many of the programs described are relatively new, others are being broadened in scope, and still others are reported as being only in the planning stages. All of this suggests that the tempo of post-baccalaureate education in industry is increasing. Furthermore, the interest shown in the results of the survey by respondents, educators, and industrial educational people in general suggests that there is a growing interest in a cooperative approach by industry and educators to the increasingly important problem of continuing education for today's engineer.

Additional reading

The December 1963 issue of the *IEEE Transactions on Education* contains several papers on the subject of continuing education. They are recommended for additional reading and consideration. — *Editor*

Matrix functions and applications

Part IV—Matrix functions and constituent matrices

This fourth part of a five-part series starts with the expansion of an analytic function of an arbitrary square matrix, involving real or complex numbers, in terms of its constituent matrices. The conjoint matrix and characteristic polynominal are computed by a convenient algorithm and are then used for calculating constituent matrices. Finally there is a discussion of the two-sided matrix equation $AY + B = YC$, and of the reduction of an arbitrary matrix to a similar Jordan matrix.

J. S. Frame *Michigan State University*

4.1 Analytic functions of a general n × n matrix

Solutions of certain systems of linear differential equations with constant coefficients, such as those derived in Part III for the state model of a system, were found in Part II to be expressible by analytic functions, such as e^{At} and $\cos A^{1/2}t$, of an $n \times n$ constant matrix A. Analytic functions of a *diagonable* matrix A could be expressed in the form

$$f(A) = \sum_{j=1}^{s} f(\lambda_j)A_j \quad \text{for } A \text{ diagonable} \quad (4.1.1)$$

where λ_j are the distinct eigenvalues and A_j the corresponding constituent idempotent matrices of A. Unfortunately, (4.1.1) and (2.4.18) of Part II for the constituent idempotent matrices A_j are *not valid* if A is *not diagonable*, and it is not easy by examining a matrix A superficially to tell whether it is diagonable or not. Furthermore, when $n > 3$, the methods of Part II are not the most efficient methods for computing the constituent idempotents of even a diagonable $n \times n$ matrix A.

For the general $n \times n$ constant matrix A, the expansion of (4.1.1) is to be replaced and generalized by the following theorem.

Theorem 4.1.1 If f is an analytic function in a simply connected domain of the complex λ plane that contains all the eigenvalues λ_j of a given $n \times n$ matrix A with real or complex entries, then any analytic matrix function $f(A)$ may be expressed by the n term sum

$$f(A) = \sum_{j=1}^{s} \sum_{k=1}^{n_j} \frac{f^{k-1}(\lambda_j)}{(k-1)!} Z_{jk} \quad (4.1.2)$$

where n_j is the "characteristic" multiplicity of λ_j (as a root of $|\lambda U - A| = 0$), where the matrices Z_{j1} are polynomials in A known as the constituent idempotents A_j for A, and where Z_{jk} are matrices defined in terms of the idempotents A_j by the equations

$$Z_{jk} = (A - \lambda_j U)^{k-1} A_j \qquad k = 1, 2, \ldots n_j \quad (4.1.3)$$

(Note: The matrices called Z_{jk} in this article are denoted $(k-1)!Z_{jk}$ by Gantmacher.[1])

Definition 4.1.1 If $Z_{jm_j} \neq 0$, but $Z_{jk} = 0$ for $k > m_j$, then m_j is called the *minimal multiplicity* of λ_j, $m = \Sigma m_j$ is called the *minimal degree* of A, the m nonvanishing matrices Z_{jk} are called the *constituent matrices* for A, and their coefficients $f^{k-1}(\lambda_j)/(k-1)!$ in (4.1.2) are called the *values of f on the spectrum of A*.

The matrix A satisfies a minimal polynominal equation of degree m, but no equation of lower degree.

The proof of Theorem 4.1.1 depends on the following properties of the constituent idempotent matrices A_j, which we shall assume now and prove later.

$$\sum_{j=1}^{s} A_j = U \tag{4.1.4}$$

$$(A - \lambda_j U)^{n_j} A_j = 0 \tag{4.1.5}$$

$$A_j A_k = 0 \quad \text{for } j \neq k \quad \text{(orthogonality)} \tag{4.1.6}$$

$$A_j^2 = A_j \quad \text{(idempotency)} \tag{4.1.7}$$

Proof of the expansion Theorem 4.1.1: Let $\alpha_j(\lambda)$ be a polynomial in λ such that $\alpha_j(A) = A_j$; see (4.3.4). Assuming (4.1.4), we have

$$f(A) = \sum_{j=1}^{s} f(A)A_j = \sum_{j=1}^{s} f[\lambda_j U + (A - \lambda_j U)]A_j \tag{4.1.8}$$

To the scalar function $f(\lambda)\alpha_j(\lambda)$ with convergent series expansion

$$f(\lambda)\alpha_j(\lambda) = \left[f(\lambda_j) + f'(\lambda_j)(\lambda - \lambda_j) + \frac{f''(\lambda_j)}{2!}(\lambda - \lambda_j)^2 + \dots\right]\alpha_j(\lambda) \tag{4.1.9}$$

corresponds the matrix function whose corresponding expansion is

$$f(A)A_j = \left[f(\lambda_j)U + f'(\lambda_j)(A - \lambda_j U) + \frac{f''(\lambda_j)}{2!}(A - \lambda_j I)^2 + \dots\right]A_j \tag{4.1.10}$$

$$= f(\lambda_j)Z_{j1} + f'(\lambda_j)Z_{j2} + \frac{f''(\lambda_j)}{2!}Z_{j3} + \dots + \frac{f^{(k-1)}(\lambda_j)}{(k-1)!}Z_{jk} + \dots$$

All but the first $m_j \leq n_j$ terms of this sum vanish, by (4.1.5) and (4.1.3), so the series on the right is convergent and defines the function on the left. Substitution of (4.1.10) in (4.1.8) proves Theorem 4.1.1.

Calculation of the constituent matrices Z_{jk} for A will require

1. Calculation of the *eigenvalues* λ_j of A.
2. Calculation of either the *powers* $U, A, A^2, \dots A^{n-1}$ of A, or some alternative set of *polynomials in A*.
3. Determination of the Z_{jk} as linear combinations of these powers or polynomials. The coefficients in these linear combinations will not depend directly on the entries of A, but only on the coefficients d_k and the roots λ_j of its characteristic polynomial $D(\lambda)$.

Specifically these coefficients may be expressed in terms of the (generalized) Vandermonde matrix V and the symmetric coefficient matrix W, which when $D(\lambda) = |\lambda U - A| = (\lambda - \lambda_1)^3(\lambda - \lambda_2)^2(\lambda - \lambda_3)$ take the form

$$V = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \lambda_1^3 & \lambda_1^4 & \lambda_1^5 \\ 0 & 1 & 2\lambda & 3\lambda^2 & 4\lambda_1^3 & 5\lambda_1^4 \\ 0 & 0 & 1 & 3\lambda_1 & 6\lambda_1^2 & 10\lambda_1^3 \\ 1 & \lambda_2 & \lambda_2^2 & \lambda_2^3 & \lambda_2^4 & \lambda_2^5 \\ 0 & 1 & 2\lambda_2 & 3\lambda_2^2 & 4\lambda_2^3 & 5\lambda_2^4 \\ 1 & \lambda_3 & \lambda_3^2 & \lambda_3^3 & \lambda_3^4 & \lambda_3^5 \end{bmatrix} \quad \text{and}$$

$$W = \begin{bmatrix} d_5 & d_4 & d_3 & d_2 & d_1 & 1 \\ d_4 & d_3 & d_2 & d_1 & 1 & 0 \\ d_3 & d_2 & d_1 & 1 & 0 & 0 \\ d_2 & d_1 & 1 & 0 & 0 & 0 \\ d_1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.1.11}$$

In the general case, the entries v_{ir} of column r of the Vandermonde matrix $V = (v_{ir})$ are the values on the spectrum of A of the function $f(\lambda) = \lambda^{r-1}$. The entry corresponding to the ith matrix Z_{jk} is

$$\binom{r-1}{k-1}\lambda_j^{r-k}$$

The ij entry of W is $d_{n+1-i-j}$ if $d_k = 0$ for $k < 0$. Setting $f(A) = A^{r-1}$ in (4.1.2) gives

$$A^{r-1} = \sum_{i=1}^{n} v_{ir}Z_i = \sum_{j=1}^{s}\sum_{k=1}^{n_j}\binom{r-1}{k-1}\lambda_j^{r-k}Z_{jk} \tag{4.1.12}$$

where Z_i denotes the ith matrix $Z_{jk}(j = 1, \dots s, k = 1, \dots, n_j)$, of which some may be 0 if $m_j < n_j$.

Since V is nonsingular, these equations can be solved to express each Z_i as a polynomial in A with coefficients obtained from row i of $(V^T)^{-1}$.

An easily inverted triangular coefficient matrix T will replace V^T as coefficient matrix in (4.1.12) if the powers of A on the left are replaced by the polynomials

$$U, A - \lambda_1 U, (A - \lambda_1 U)(A - \lambda_2 U), \dots, \prod_{j=1}^{n-1}(A - \lambda_j U)$$

where, in this instance, consecutive eigenvalues λ_j, λ_{j+1} may be equal. Even simpler is the coefficient matrix related to certain polynomials C_{jk} derived below from the adjoint of $\lambda U - A$.

Example 1 Find the constituent matrices Z_{jk} of matrix A if

$$A = \begin{bmatrix} -13 & -21 & -11 \\ 5 & 8 & 5 \\ 6 & 11 & 4 \end{bmatrix}$$

$$A^2 = \begin{bmatrix} -2 & -16 & -6 \\ 5 & 14 & 5 \\ 1 & 6 & 5 \end{bmatrix} \tag{4.1.13}$$

Solution: The characteristic polynomial is

$$|\lambda U - A| = \begin{vmatrix} \lambda + 13 & 21 & 11 \\ -5 & \lambda - 8 & -5 \\ -6 & -11 & \lambda - 4 \end{vmatrix} = \lambda^3 + \lambda^2 - 8\lambda - 12 = (\lambda - 3)(\lambda + 2)^2 \tag{4.1.14}$$

Eigenvalues are $\lambda_1 = 3$ $(n_1 = 1)$, $\lambda_2 = -2$ $(n_2 = 2)$. Then, using (4.1.2), we have

$$\begin{aligned} U &= Z_{11} + Z_{21} \\ A &= 3Z_{11} - 2Z_{21} + Z_{22} \\ A^2 &= 9Z_{11} + 4Z_{21} - 4Z_{22} \end{aligned} \qquad V^T = \begin{bmatrix} 1 & 1 & 0 \\ 3 & -2 & 1 \\ 9 & 4 & -4 \end{bmatrix} \tag{4.1.15}$$

The inverse of V^T is

$$(V^T)^{-1} = \frac{1}{25}\begin{bmatrix} 4 & 4 & 1 \\ 21 & -4 & -1 \\ 30 & 5 & -5 \end{bmatrix} \qquad (4.1.16)$$

Hence, solving for Z_{jk} we have

$$Z_{11} = \frac{1}{25}(4U + 4A + A^2)$$

$$Z_{21} = \frac{1}{25}(21U - 4A - A^2) \qquad (4.1.17)$$

$$Z_{22} = \frac{1}{25}(30U + 5A - 5A^2)$$

The constituent matrices Z_{jk} for A are found to be

$$Z_{11} = \begin{bmatrix} -2 & -4 & -2 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix} \quad Z_{21} = \begin{bmatrix} 3 & 4 & 2 \\ -1 & -1 & -1 \\ -1 & -2 & 0 \end{bmatrix}$$

$$Z_{22} = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The alternative computation suggested above, using eigenvalues 3, −2, −2, is carried through as follows:

Functions		Values on spectrum		
	$f(\lambda)$	$f(3)$	$f(-2)$	$f'(-2)$
$\begin{bmatrix} 1 \\ \lambda - 3 \\ (\lambda - 3)(\lambda + 2) \end{bmatrix}$	$F(\lambda)$	$\begin{bmatrix} 1 & 1 & 0 \\ 0 & -5 & 1 \\ 0 & 0 & -5 \end{bmatrix}$		T

Equations

$$\begin{aligned} U &= Z_{11} + Z_{21} \\ A - 3U &= -5Z_{21} + Z_{22} \\ (A - 3U)(A + 2U) &= -5Z_{22} \\ F(A) &= TZ \end{aligned} \qquad (4.1.19)$$

We build successively the matrices U, $A - 3U$, and

$$\begin{aligned} &(A - 3U)(A + 2U) \\ &= \begin{bmatrix} -16 & -21 & -11 \\ 5 & 5 & 5 \\ 6 & 11 & 1 \end{bmatrix}\begin{bmatrix} -11 & -21 & -11 \\ 5 & 10 & 5 \\ 6 & 11 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 5 & 5 & 5 \\ 0 & 0 & 0 \\ -5 & -5 & -5 \end{bmatrix} \end{aligned} \qquad (4.1.20)$$

Then either solve the equations of (4.1.19) successively for Z_{22}, Z_{21}, Z_{11}, or invert the matrix T in (4.1.19) and express the Z_{jk} in terms of the chosen polynomials in A.

$$T^{-1} = \begin{bmatrix} 1 & 1/5 & 1/25 \\ 0 & -1/5 & -1/25 \\ 0 & 0 & -1/5 \end{bmatrix} \qquad (4.1.21)$$

$$\begin{aligned} Z_{11} &= U + 1/5(A - 3U) + 1/25(A - 3U)(A + 2U) \\ Z_{21} &= -1/5(A - 3U) - 1/25(A - 3U)(A + 2U) \\ Z_{22} &= -1/5(A - 3U)(A + 2U) \end{aligned}$$

4.2 The conjoint matrix

For an arbitrary $n \times n$ matrix A, the characteristic polynomial has the form

$$\begin{aligned} D(\lambda) &= |\lambda U - A| \\ &= \lambda^n + d_1\lambda^{n-1} + \ldots + d_k\lambda^{n-k} + \ldots + d_n \\ &= \prod_{j=1}^{s} (\lambda - \lambda_j)^{n_j} \end{aligned} \qquad (4.2.1)$$

The adjoint of $\lambda U - A$, which we denote by $B(\lambda)$ and call the *conjoint* of A, is the transposed cofactor matrix of $\lambda U - A$. It may be written as a matrix $\beta(\lambda)$ with polynomial entries, or as a polynomial with matrix coefficients B_k, of the form

$$\begin{aligned} B(\lambda) &= \text{adj}\,(\lambda U - A) \\ &= B_0\lambda^{n-1} + B_1\lambda^{n-2} + \ldots + B_{n-2}\lambda + B_{n-1} \end{aligned}$$

$$B_0 = U \qquad B_n = 0 \qquad (4.2.2)$$

Since the product of any matrix by its adjoint is a diagonal matrix with the determinant of the given matrix in each diagonal entry, we have

$$(\lambda U - A)B(\lambda) = D(\lambda)U \qquad (4.2.3)$$

For the matrix A of (4.1.13) this product would be as follows:

$$\begin{aligned} &(\lambda U - A) \\ &\begin{bmatrix} \lambda + 13 & 21 & 11 \\ -5 & \lambda - 8 & -5 \\ -6 & -11 & \lambda - 4 \end{bmatrix} \times \\ &\qquad B(\lambda) \\ &\begin{bmatrix} \lambda^2 - 12\lambda - 23 & -21\lambda - 37 & -11\lambda - 17 \\ 5\lambda + 10 & \lambda^2 + 9\lambda + 14 & 5\lambda + 10 \\ 6\lambda + 7 & 11\lambda + 17 & \lambda^2 + 5\lambda + 1 \end{bmatrix} \\ &= \begin{bmatrix} \lambda^3 + \lambda^2 - 8\lambda - 12 & 0 & 0 \\ 0 & \lambda^3 + \lambda^2 - 8\lambda - 12 & 0 \\ 0 & 0 & \lambda^3 + \lambda^2 - 8\lambda - 12 \end{bmatrix} \\ &= D(\lambda)U \end{aligned} \qquad (4.2.4)$$

An important theoretical and computational role is played by this conjoint matrix $B(\lambda)$ with matrix coefficients B_k. Computations of n^2 polynomial cofactors of $\lambda U - A$, quite laborious for $n > 3$, and subsequent computation of $D(\lambda)$ by (4.2.3) are avoided by an algorithm discovered independently at about the same time by several mathematicians, including Souriau, Faddeev, Fettis, and Frame.

After the scalar coefficients d_k in $D(\lambda)$ and the matrix coefficients B_k in $B(\lambda)$ have been simultaneously computed, they can be used to find the eigenvectors of A, its modal matrix S (such that $S^{-1}AS$ is a Jordan matrix), and its constituent matrices Z_{jk}, including the important idempotent matrices $A_j = Z_{j1}$. The theoretical results include as a by-product the proof of the important Hamilton-Calyey theorem.

Theorem 4.2.1 The trace of the conjoint matrix $B(\lambda)$ equals the derivative of the characteristic polynomial $D(\lambda)$.

$$\text{tr}\, B(\lambda) = D'(\lambda) \qquad (4.2.5)$$

$$\text{tr}\, B_k = (n - k)d_k \qquad (4.2.6)$$

Proof: For any square matrix $M(\lambda)$ with polynomial entries $m_{ij}(\lambda)$ and adjoint $\tilde{M}(\lambda)$, the derivative of the determinant has the expansion

$$\frac{d}{d\lambda}|M(\lambda)| = \sum_{ij}\left[\frac{\partial}{\partial m_{ij}}|M(\lambda)|\right]\frac{dm_{ij}}{d\lambda}$$
$$= \sum_{ij}[\text{cofactor of } m_{ij}]\frac{dm_{ij}}{d\lambda}$$
$$= \sum_{j}\left(\tilde{M}\frac{dM}{d\lambda}\right)_{jj} = \text{tr}\left(\tilde{M}\frac{dM}{d\lambda}\right) \quad (4.2.7)$$

When $M = \lambda U - A$, we have $|M(\lambda)| = D(\lambda)$, $\tilde{M} = B(\lambda)$, $dM/d\lambda = U$, and (4.2.7) reduces to (4.2.5). Comparison of coefficients of λ^{n-k-1} in (4.2.5) gives (4.2.6).

The reader may check in (4.2.4) that tr $B(\lambda) = 3\lambda^2 + 2\lambda - 8 = D'(\lambda)$.

Theorem 4.2.2 (Frame, Faddeev, and others) The scalar coefficients d_k in the characteristic polynomial $D(\lambda)$ and matrix coefficients B_k in $B(\lambda) = \text{adj}\,(\lambda U - A)$ are given recursively by

$$d_k = -\frac{1}{k}\,\text{tr}\,AB_{k-1} \quad k = 1, 2, \ldots, n \quad d_0 = 1 \quad (4.2.8)$$

$$B_k = AB_{k-1} + d_kU \quad k = 1, 2, \ldots, n \quad B_0 = U \quad B_n = 0 \quad (4.2.9)$$

Proof: Equating coefficients of λ^{n-k} in both members of

$$(\lambda U - A)B(\lambda) = D(\lambda)U \quad (4.2.3)$$

and defining $B_n = 0$ to be the constant term of $\lambda B(\lambda)$, we have

$$B_k - AB_{k-1} = d_kU \quad \text{for } k = 1, 2, \ldots, n \quad (4.2.10)$$

Taking traces, and applying Theorem 4.2.1, we have

$$(n - k)d_k - \text{tr}\,AB_{k-1} = nd_k \quad \text{for } k = 1, 2, \ldots n \quad (4.2.11)$$

Equations (4.2.9) and (4.2.8) follow immediately from (4.2.10) and (4.2.11).

Corollary 4.2.3 The inverse of A is

$$A^{-1} = \frac{-B_{n-1}}{d_n} \text{ if } d_n = |-A| \neq 0 \quad (4.2.12)$$

This follows by setting $k = n$ in (4.2.9), since $B_n = 0$.

Theorem 4.2.4 (Hamilton-Cayley theorem) Every square matrix satisfies its characteristic equation.

Proof: From (4.2.9), we find successively for $k = 1, 2, \ldots n$

$$B_k = A^k + d_1A^{k-1} + d_2A^{k-2} + \ldots + d_{k-1}A + d_kU \quad (4.2.13)$$

For $k = n$, we have $B_n = 0$ and hence

$$A^n + d_1A^{n-1} + d_2A^{n-2} + \ldots + d_{n-1}A + d_nU = D(A) = 0 \quad \text{Q.E.D.} \quad (4.2.14)$$

A systematic computation of the d_k and B_k may be arranged as follows, for the matrix A of (4.1.13)

	A		diag A
	−13	−21	−11	−13
	5	8	5	8
	6	11	4	4
tr A =				−1
d_1 =				1

	$B_1 = AB_0 + d_1I$			diag AB_1	$B_2 = AB_1 + d_2I$			diag AB_2	$B_3 = AB_2 + d_3I$		
	−12	−21	−11	−15	−23	−37	−17	12	0	0	0
	5	9	5	22	10	14	10	12	0	0	0
	6	11	5	9	7	17	1	12	0	0	0
tr AB_1 =				16				36			check
d_2 =				−8							
d_3 =								−12			

(4.2.15)

$$D(\lambda) = \lambda^3 + \lambda^2 - 8\lambda - 12$$

$$D'(\lambda) = 3\lambda^2 + 2\lambda - 8 = (\text{tr}\,B_0)\lambda^2 + (\text{tr}\,B_1)\lambda + \text{tr}\,B_2 \quad \text{(check)}$$

There are n steps, $k = 1, 2, \ldots, n$, of which step k computes d_k and B_k, and step n serves to check the computation.

To compute d_k, we first write in a column the diagonal entries of AB_{k-1}, sum them, and divide by $-k$. To compute B_k, add this value of d_k to the diagonal entries just computed and place them in the diagonal of B_k. Then fill in the other entries of AB_{k-1}. Note that the equal diagonal entries of AB_{n-1} are $-d_n = (-1)^{n-1}|A|$, and that the matrix B_n should come out 0 if there is no round-off error. In machine computation, the relative deviation of AB_{n-1} from the scalar matrix $-d_nI$ gives a measure of the round-off error involved.

4.3 Constituent matrices and related matrix polynomials

We shall now define and derive the constituent idempotents A_j for an $n \times n$ matrix A whose eigenvalue λ_j has multiplicity n_j in its characteristic polynomial

$$D(\lambda) = |\lambda U - A| = \sum_{k=1}^{n} d_k\lambda^{n-k} = \prod_{j=1}^{s}(\lambda - \lambda_j)^{n_j} \quad (4.3.1)$$

First we expand $1/D(\lambda)$ in partial fractions of the form

$$\frac{1}{D(\lambda)} = \sum_{j=1}^{s}\frac{\beta_j(\lambda)}{(\lambda - \lambda_j)^{n_j}}$$

$$\beta_j(\lambda) = \sum_{k=1}^{n_j}\gamma_{jk}(\lambda - \lambda_j)^{k-1} \quad (4.3.2)$$

Then we define the scalar polynomials $D_j(\lambda)$ and $\alpha_j(\lambda)$ by

$$D_j(\lambda) = \prod_{k\neq j}(\lambda - \lambda)_k^{n_k} = \frac{D(\lambda)}{(\lambda - \lambda_j)^{n_j}} \quad (4.3.3)$$

$$\alpha_j(\lambda) = \beta_j(\lambda)D_j(\lambda) = \sum_{k=1}^{n_j}\gamma_{jk}(\lambda - \lambda_j)^{k-1}D_j(\lambda) \quad (4.3.4)$$

and define the corresponding polynomials D_j and A_j in the matrix A by

$$D_j = D_j(A) = \prod_{k\neq j}(A - \lambda_kU)^{n_k} \quad (4.3.5)$$

$$A_j = \alpha_j(A) = \beta_j(A)D_j = \sum_{k=1}^{n_j} \gamma_{jk}(A - \lambda_j U)^{-1}D_j \tag{4.3.6}$$

Theorem 4.3.1 The matrices A_j defined by (4.3.6) are constituent idempotent matrices for A that satisfy the fundamental identities

$$\Sigma A_j = U \tag{4.3.7}$$

$$(A - \lambda_j U)^{n_j} A_j = 0 \tag{4.3.8}$$

$$A_i A_j = 0 \quad \text{for } i \neq j \tag{4.3.9}$$

$$A_j^2 = A_j \tag{4.3.10}$$

Proof: If we substitute $D_j(\lambda)$ from (4.3.3) in (4.3.4) and apply (4.3.2), we have

$$\sum_j \alpha_j(\lambda) = \sum_j \frac{D(\lambda)\beta_j(\lambda)}{(\lambda - \lambda_j)^{n_j}} = \frac{D(\lambda)}{D(\lambda)} = 1 \tag{4.3.11}$$

Also

$$(\lambda - \lambda_j)^{n_j}\alpha_j(\lambda) = \beta_j(\lambda)D(\lambda) \tag{4.3.12}$$

These polynomial identities in λ remain valid when λ is replaced by the matrix A, and λ^0 by U. Thus (4.3.7) follows from (4.3.11) and (4.3.8) follows from (4.3.12) if we note that $D(A)$ must be 0 by the Hamilton-Cayley theorem, 4.2.4.

Since for $i \neq j$ both D_i in (4.3.5) and hence A_i in (4.3.6) have the factor $(A - \lambda_j U^{n_j})$, equation (4.3.9) follows from (4.3.8). Finally the validity of (4.3.10) follows from (4.3.9) and (4.3.7) since

$$A_j^2 = \sum_{i=1}^{s} A_i A_j = UA_j = A_j \tag{4.3.13}$$

Definition 4.3.1 A square matrix is called *nilpotent* if any of its powers is the zero matrix.

Theorem 4.3.2 The constituent matrices

$$Z_{jk} = (A - \lambda_j U)^{k-1}A_j \tag{4.3.14}$$

are nilpotent for $k > 1$.

Proof: Since polynomials in A are commutative, we have

$$(Z_{j2})^{k-1} = (A - \lambda_j U)^{k-1}A_j^{k-1} = (A - \lambda_j U)^{k-1}A_j = Z_{jk} \tag{4.3.15}$$

$$(Z_{j2})^{n_j} = (A - \lambda_j U)^{n_j}A_j = 0 \tag{4.3.16}$$

Theorem 4.3.3 The conjoint matrix $B(\lambda)$ for A has the expansion

$$B(\lambda) = \sum_{j=1}^{s} D_j(\lambda) \sum_{k=1}^{n_j} (\lambda - \lambda_j)^{n_j - k} Z_{jk} \tag{4.3.17}$$

in terms of the constituent matrices Z_{jk}.

Proof: The scalar function $f(\lambda) = (\mu - \lambda)^{-1}$ corresponds to the matrix function $f(A) = (\mu U - A)^{-1}$, called the *resolvent* of A. It is an analytic function of μ except at the eigenvalues of A. The kth derivative of $f(\lambda)$ with respect to λ is $k!\,(\mu - \lambda)^{-1-k}$. Thus for fixed μ the resolvent can be expanded by Theorem 4.1.1 in the form

$$(\mu U - A)^{-1} = \sum_{j=1}^{s} \sum_{k=1}^{n_j} (\mu - \lambda_j)^{-k} Z_{jk} \tag{4.3.18}$$

Replacing μ by λ in (4.3.18) and multiplying by

$$(\lambda U - A)B(\lambda) = D(\lambda)U = D_j(\lambda)(\lambda - \lambda_j)^{n_j}U$$

we obtain (4.3.17) for each value of λ that is not an eigenvalue of A. Continuity establishes the result for the eigenvalues as well.

Assume next that the polynomial $D(\lambda)$ has the finite series expansion

$$D(\lambda) = \sum_{k=1}^{n-n_j+1} g_{jk}(\lambda - \lambda_j)^{n_j+k-1} \tag{4.3.19}$$

Then the scalars g_{jk} may be defined by

$$g_{jk} = \begin{cases} D_j^{(k-1)}(\lambda_j)/(k-1)! \\ \quad = D^{(n_j+k-1)}(\lambda_j)/(n_j + k - 1)! \quad k > 0 \\ 0 \text{ if } k \leq 0 \text{ or } k > (n - n_j + 1) \end{cases} \tag{4.3.20}$$

Then by Theorem 4.1.1 the expansion of the matrix D_j in constituent matrices is

$$D_j = D_j(A) = \sum_{k=1}^{n_j} g_{jk}Z_{jk} \tag{4.3.21}$$

since $D_j(\lambda)$ has vanishing derivatives of all required orders at other roots λ_i.

To facilitate the computation of constituent matrices, we now introduce a special set of n polynomials in A, which we denote C_{jk}, and define in terms of $B(\lambda)$ and its derivatives, as follows:

$$C_{jk} = \frac{B^{(k-1)}(\lambda_j)}{(k-1)!} = \sum_{r=k}^{n} \binom{r-1}{k-1} \lambda_j^{r-k} B_{n-r}$$
$$k = 1 \ldots n_j \quad j = 1 \ldots s \tag{4.3.22}$$

We then assemble the powers A^{r-1}, the conjoint coefficients B_{n-r}, the conjoint derivatives C_{jk}, and the constituent matrices Z_{jk} into $n \times 1$ vectors whose entries are $n \times n$ matrices

$$A^{\natural} = \begin{bmatrix} U \\ A \\ A^2 \\ \cdot \\ \cdot \\ \cdot \\ A^{n-1} \end{bmatrix} \quad B^{\natural} = \begin{bmatrix} B_{n-1} \\ B_{n-2} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ B_0 \end{bmatrix} \quad C^{\natural} = \begin{bmatrix} C_{11} \\ C_{12} \\ \cdot \\ \cdot \\ C_{21} \\ \cdot \\ \cdot \\ C_{sn_s} \end{bmatrix} \quad Z^{\natural} = \begin{bmatrix} Z_{11} \\ Z_{12} \\ \cdot \\ \cdot \\ Z_{21} \\ \cdot \\ \cdot \\ Z_{sn_s} \end{bmatrix} \tag{4.3.23}$$

Any $n \times n$ matrix $V = (v_{ij})$ with scalar entries v_{ij}, such as the Vandermonde matrix V or coefficient matrix W in (4.1.11) may be enlarged to an $n^2 \times n^2$ matrix $V^{\natural}$ or $W^{\natural}$ by replacing the scalar entry v_{ij} by the submatrix $v_{ij}U$ to obtain the $n^2 \times n^2$ partitioned matrix $V^{\natural} = V \times U$, technically known as the *direct* product, or Kronecker product, of V and U. We also construct the $n^2 \times n^2$ matrix $G^{\natural}$ and its inverse $\Gamma^{\natural}$ by enlarging the matrix G and its inverse Γ, defined below.

Theorem 4.3.4 The four vectors (4.3.23) whose entries are polynomials in A are related by the matrix equations

$$A^{\natural} = (V^T)^{\natural}Z^{\natural} \tag{4.3.24}$$

$$B^{\natural} = W^{\natural}A^{\natural} = (WV^T)^{\natural}Z^{\natural} \tag{4.3.25}$$

$$C^{\natural} = V^{\natural}B^{\natural} \tag{4.3.26}$$

$$Z^r = \Gamma^r C^r = (G^r)^{-1} V^r B^r \tag{4.3.27}$$

where

$$G = \begin{bmatrix} G_1 & 0 & \cdots & 0 \\ 0 & G_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & G_s \end{bmatrix}$$

$$G_j = \begin{bmatrix} 0 & 0 & \cdots & g_{j1} \\ 0 & 0 & \cdots & g_{j2} \\ \vdots & \vdots & & \vdots \\ 0 & g_{j1} & \cdots & g_{j,n_j-1} \\ g_{j1} & g_{j2} & \cdots & g_{j,n_j} \end{bmatrix} \tag{4.3.28}$$

$$\Gamma = \begin{bmatrix} \Gamma_1 & 0 & \cdots & 0 \\ 0 & \Gamma_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Gamma_s \end{bmatrix} = G^{-1}$$

$$\Gamma_j = \begin{bmatrix} \gamma_{j,n_j} & \cdots & \gamma_{j2} & \gamma_{j1} \\ \gamma_{j,n_j-1} & \cdots & \gamma_{j1} & 0 \\ \vdots & & \vdots & \vdots \\ \gamma_{j2} & \gamma_{j1} & \cdots & 0 \\ \gamma_{j1} & 0 & \cdots & 0 \end{bmatrix} = G_j^{-1} \tag{4.3.29}$$

and where V and W are the Vandermonde matrix and symmetric coefficient matrix illustrated in (4.1.11). The enlarged matrices V^r, W^r, G^r, and Γ^r in equations (4.3.24) to (4.3.27) are obtained by replacing each scalar entry c_{ij} in V, W, G, or Γ respectively by the scalar matrix entry $c_{ij}U$, where U is the $n \times n$ unit matrix.

Proof: Equation (4.3.24) is equivalent to (4.1.12), which was obtained by expanding the matrix functions A^{k-1} by Theorem 4.1.1. Equation (4.3.25) restates the expression for B_k in (4.2.13) used in the proof of the Hamilton-Cayley theorem, and expresses B^r in terms of Z^r using (4.3.24). Equation (4.3.26) restates the expression for C_{jk} in the defining equations (4.3.22). It remains to prove (4.3.27).

Differentiating $B(\lambda)$ in (4.3.17) repeatedly at λ_j and noting that all required derivatives of $D_i(\lambda)$ vanish at λ_j if $i \neq j$, we apply (4.3.20) and (4.3.21) to obtain

$$C_{jk} = \sum_{r=1}^{n_j} \frac{1}{(k-1)!} \frac{d^{k-1}}{d\lambda^{k-1}} \left[D_j(\lambda)(\lambda - \lambda_j)^{n_j - r} \right]_{\lambda = \lambda_j} Z_{jr} \tag{4.3.30}$$

$$C_{jk} = \sum_r g_{jr} Z_{r,\,r+n_j-k} = (A - \lambda_j U)^{n_j - k} D_j \tag{4.3.31}$$

From (4.3.6) we now obtain

$$Z_{jr} = (A - \lambda_j U)^{r-1} A_j = \sum_k \gamma_{jk} (A - \lambda_j U)^{r+k-2} D_j$$

$$= \sum_k \gamma_{j,n_j-r-k+2} C_{jk} \tag{4.3.32}$$

Equations (4.3.31) and (4.3.32) are equivalent to $C^r = GZ^r$ and $Z^r = \Gamma G^r$. Thus equation (4.3.27) is proved with $\Gamma = G^{-1}$.

Theorem 4.3.5 The determinant and inverse of the Vandermonde matrix V of (4.1.11) are given by

$$|V| = \prod_{i>j} (\lambda_i - \lambda_j)^{n_i n_j} \tag{4.3.33}$$

$$(V^T)^{-1} = G^{-1}(VW) = \Gamma VW \tag{4.3.34}$$

Proof: By combining equations (4.3.24) through (4.3.27) we obtain

$$Z^r = \Gamma^r C^r = \Gamma^r V^r W^r (V^T)^r Z^r \tag{4.3.35}$$

Given the characteristic polynomial $D(\lambda)$, there is a corresponding $n \times n$ matrix A having $D(\lambda) = 0$ as its minimal equation, for which there are n independent constituent matrices Z^r. Hence the coefficient of Z on the right of (4.3.35) must be the unit matrix. Thus

$$\Gamma VWV^T = U \tag{4.3.36}$$

$$VWV^T = \Gamma^{-1} = G \tag{4.3.37}$$

Equation (4.3.34) follows immediately from (4.3.37). Taking determinants we find

$$|V|^2 |W| = |G| = \prod_j |G_j| \tag{4.3.38}$$

$$|V|^2 (-1)^{n(n-1)/2} = \prod_j (g_{j1})^n j (-1)^{n_j(n_j-1)/2}$$

$$|V|^2 = (-1)^{(n^2 - \Sigma n_j^2)/2} \prod_j \prod_{i \neq j} (\lambda_j - \lambda_i)^{n_i n_j}$$

$$= \prod_{i>j} (\lambda_i - \lambda_j)^{2 n_i n_j}$$

Extracting the appropriate square roots of both sides, we obtain (4.3.33). The sign can easily be verified from the sign of the term that is the product of the diagonal entries of V.

Entries of both the matrices VW and G can be calculated by use of the synthetic division algorithm for the polynomial $D(\lambda)$. Writing the coefficients 1, d_1, d_2, ..., d_n in row one, we copy the 1 in row three, multiply by λ_j and place λ_j under d_1 in row two, add to get $\lambda_j + d_1$ in row three. Then successively multiply by λ_j and add the next coefficient, until $D(\lambda_j)$ appears in row three under d_n in row one. Row three contains the coefficients in the quotient

$$\frac{D(\lambda) - D(\lambda_j)}{\lambda - \lambda_j}$$

followed by the remainder $D(\lambda_j)$. If λ_j has multiplicity n_j, we repeat the process n_j times, dropping the final column each time to obtain in alternate rows the columns of VW in reversed order. Then we continue the process n_j more times in order to compute the values of the coefficients g_{jk} needed for G. Take $g_{jk} = 0$ when $(k + n_j) > (n + 1)$.

The format of synthetic division appears, in part, as follows:

$$\begin{array}{cccccc} 1 & d_1 & d_2 & \cdots & d_{n-1} & a_n \\ & \lambda_j & \lambda_j(\lambda_j + d_1) & & & \lambda_j[\ldots] \\ 1 & \lambda_j + d_1 & \lambda_j^2 + d_1\lambda_j + d_2 & \cdots & & 0 = D(\lambda_j) \\ & \lambda_j & \lambda_j(2\lambda_j + d_1) & & \lambda_j[\ldots] & \\ 1 & 2\lambda_j + d_1 & 3\lambda_j^2 + 2d_1\lambda_j + d_2 & \cdots & D'(\lambda_j) & \end{array} \tag{4.3.39}$$

Example 2 Compute VW and G if

$$\begin{aligned} D(\lambda) &= (\lambda - 2)^3(\lambda + 3)^2(\lambda - 4) \\ &= \lambda^6 - 4\lambda^5 - 15\lambda^4 + 70\lambda^3 + \\ &\qquad 20\lambda^2 - 312\lambda + 288 \end{aligned} \tag{4.3.40}$$

Repeated synthetic divisions for the roots 2, −3, 4 have the format

$$\begin{array}{cccccccl} 1 & -4 & -15 & 70 & 20 & -312 & +288 & \mid 2 \\ & 2 & -4 & -38 & 64 & 168 & -288 & \\ 1 & -2 & -19 & 32 & 84 & -144 & 0 = D(2) & \\ & 2 & 0 & -38 & -12 & 144 & & \\ 1 & 0 & -19 & -6 & 72 & 0 = D'(2)/1! & & \\ & 2 & 4 & -30 & -72 & & & \\ 1 & 2 & -15 & -36 & 0 = D''(2)/2! & & & \\ & 2 & 8 & -14 & & & & \\ 1 & 4 & -7 & -50 = g_{11} & & & & \\ & 2 & 12 & & & & & \\ 1 & 6 & 5 = g_{12} & & & & & \\ & 2 & & & & & & \\ 1 & 8 = g_{13} & & & & & & \end{array}$$

$$\begin{array}{cccccccl} 1 & -4 & -15 & 70 & 20 & -312 & 288 & \mid -3 \\ & -3 & 21 & -18 & -156 & 408 & -288 & \\ 1 & -7 & 6 & 52 & -136 & 96 & 0 = D(-3) & \\ & -3 & 30 & -108 & 168 & -96 & & \\ 1 & -10 & 36 & -56 & 32 & 0 = D'(-3)/1! & & \\ & -3 & 39 & -225 & 843 & & & \\ 1 & -13 & 75 & -281 & 875 = g_{21} & & & \\ & -3 & 48 & -369 & & & & \\ 1 & -16 & 123 & -650 = g_{22} & & & & \end{array}$$

$$\begin{array}{cccccccl} 1 & -4 & -15 & 70 & 20 & -312 & 288 & \mid 4 \\ & 4 & 0 & -60 & 40 & 240 & -288 & \\ 1 & 0 & -15 & 10 & 60 & -72 & 0 = D(4) & \\ & 4 & 16 & 4 & 56 & 464 & & \\ 1 & 4 & 1 & 14 & 116 & 392 = g_{33} & & \end{array} \tag{4.3.41}$$

In this example, the matrices G and VW may be read off as follows:

$$(G, VW) = \left[\begin{array}{ccc|cc|ccc} 0 & 0 & -50 & & & -144 & 84 & 32 & -19 & -2 & 1 \\ 0 & -50 & 5 & & & 72 & -6 & -19 & 0 & 1 & 0 \\ -50 & 5 & 8 & & & -36 & -15 & 2 & 1 & 0 & 0 \\ & & & 0 & 875 & 96 & -136 & 52 & 6 & -7 & 1 \\ & & & 875 & -650 & 32 & -56 & 36 & -10 & 1 & 0 \\ & & & & & 392 & -72 & 60 & 10 & -15 & 0 & 1 \end{array}\right] \tag{4.3.42}$$

Reduction of this matrix to the echelon form $U, (V^T)^{-1}$ by row operations evaluates the inverse of V^T by (4.3.34).

Note as a check that

$$\begin{aligned} (\lambda_1 - \lambda_2)^{m_2}(\lambda_1 - \lambda_3)^{m_3} &= (2 + 3)^2(2 - 4) = -50 \\ (\lambda_2 - \lambda_1)^{m_1}(\lambda_2 - \lambda_3)^{m_3} &= (-3 - 2)^3(-3 - 4) = 875 \\ (\lambda_3 - \lambda_1)^{m_1}(\lambda_3 - \lambda_2)^{m_2} &= (4 - 2)^3(4 + 3)^2 = 392 \end{aligned} \tag{4.3.43}$$

Complex eigenvalues. In case a real matrix A has a pair of conjugate complex multiple roots, the corresponding echelon reduction of the complex system

$$\begin{bmatrix} G_1 + iG_2 & 0 \\ 0 & G_1 - iG_2 \end{bmatrix} \begin{bmatrix} M_1 + iM_2 \\ M_1 - iM_2 \end{bmatrix} = \begin{bmatrix} (V_1 + iV_2)W \\ (V_1 - iV_2)W \end{bmatrix} \tag{4.3.44}$$

can be reduced in the real domain by solving instead for M_1, M_2 the equivalent real system

$$\begin{bmatrix} G_1 & -G_2 \\ G_2 & G_1 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} = \begin{bmatrix} V_1W \\ V_2W \end{bmatrix} \tag{4.3.45}$$

4.4 Similarity reduction of quasi-triangular matrices to Jacobi and Jordan forms

A matrix M is called quasi-upper triangular if it can be partitioned into submatrices M_{ij} so that each diagonal submatrix (block) M_{ii} is square and each of the M_{ij} below the diagonal ($i > j$) is 0.

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1r} \\ 0 & M_{22} & & M_{2r} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \cdots & M_{rr} \end{bmatrix} \quad \text{quasi-upper triangular} \tag{4.4.1}$$

Such a matrix M is sometimes called a *subdirect sum* of the diagonal blocks M_{ii}. It is a direct sum if and only if $M_{ij} = 0$ for all $i \neq j$. If the submatrices M_{ij} in (4.4.1) are all one-dimensional then M is upper triangular.

It is often desirable to transform a subdirect sum M by a similarity transformation into a direct sum matrix, and this can be done if no two of the diagonal blocks share a common eigenvalue. Consider, for example, the system of nonhomogeneous linear differential equations

$$\frac{dZ}{dt} = AZ + BF \tag{4.4.2}$$

in which A and B are constant $n \times n$ and $n \times m$ coefficient matrices, but the $m \times 1$ variable driving vector $F(t)$

satisfies a differential equation with $m \times m$ constant matrix C

$$\frac{dF}{dt} = CF \tag{4.4.3}$$

We may combine (4.4.2) and (4.4.3) into the single homogeneous matrix equation

$$\frac{d}{dt}\begin{bmatrix} Z \\ F \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}\begin{bmatrix} Z \\ F \end{bmatrix} = M\begin{bmatrix} Z \\ F \end{bmatrix} \tag{4.4.4}$$

where M is quasi-upper triangular. The system will split if M is similar to the direct sum of A and C. We seek a rectangular $n \times m$ matrix Y such that

$$\begin{bmatrix} A & B \\ 0 & C \end{bmatrix}\begin{bmatrix} U & Y \\ 0 & U \end{bmatrix} = \begin{bmatrix} U & Y \\ 0 & U \end{bmatrix}\begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix} \tag{4.4.5}$$

This condition (4.4.5) is satisfied if and only if Y satisfies the two-sided linear matrix equation

$$AY + B = YC \tag{4.4.6}$$

If (4.4.6) has a solution Y, the nonhomogeneous system (4.4.2) can be replaced by the homogeneous system

$$\begin{aligned} \frac{d}{dt}(Z - YF) &= (AZ + BF) - YCF \\ &= A(Z - YF) \end{aligned} \tag{4.4.7}$$

An explicit solution of the two-sided matrix equation (4.4.6) can be obtained as follows whenever A and C have no common eigenvalues.[2]

Theorem 4.4.1 Let A, B, C be $n \times n$, $n \times m$, and $m \times m$ constant matrices, respectively, such that none of the s distinct eigenvalues λ_j of A are eigenvalues of C. Then the matrix equation $AY + B = YC$ in (4.4.6) has the unique solution

$$Y = \sum_{j=1}^{s}\sum_{k=1}^{m_j} Z_{jk}B(C - \lambda_j U)^{-k} \tag{4.4.8}$$

where $Z_{jk} = (A - \lambda_j U)^{k-1}A_j$ are the constituent matrices of A.

Proof: The constituent idempotents A_j of A are polynomials in A with sum U, and they commute with A. Thus equation (4.4.6) holds if and only if

$$AA_jY + A_jB = A_jYC \quad \text{for } j = 1, 2, \ldots, s \tag{4.4.9}$$

Subtracting $\lambda_j A_j Y$ from both sides, we write

$$(A - \lambda_j U)A_jY + A_jB = A_jY(C - \lambda_j U) \tag{4.4.10}$$

Since λ_j is not an eigenvalue of C, the matrix $(C - \lambda_j U)$ has an inverse and we obtain

$$\begin{aligned} A_jB(C - \lambda_j U)^{-1} &= A_jY - (A - \lambda_j U)A_jY(C - \lambda_j U)^{-1} \\ &= Z_{j1}Y - Z_{j2}Y(C - \lambda_j U)^{-1} \end{aligned} \tag{4.4.11}$$

Multiplying on the left side by Z_{jk} and on the right by $(C - \lambda_j U)^{1-k}$, and summing over k, we obtain a telescoping sum in which all terms cancel but the first.

$$\begin{aligned} &\sum_{k=1}^{m_j} Z_{jk}B(C - \lambda_j U)^{-k} \\ &\quad = \sum_{k=1}^{m_j}[Z_{jk}Y(C - \lambda_j U)^{1-k} - Z_{j,k+1}Y(C - \lambda_j U)^{-k}] \\ &\quad = Z_{j1}Y = A_jY \end{aligned} \tag{4.4.12}$$

Summing equations (4.4.12) over j gives (4.4.8), since $\Sigma A_j = U$. Q.E.D.

Theorem 4.4.2 If no two of the diagonal blocks M_{ii} of the quasi-upper-triangular matrix M in (4.4.1) share a common eigenvalue, then M is similar to the direct sum of the M_{ii}.

Proof: First let C in (4.4.5) be M_{rr} in (4.4.1) and apply Theorem 4.4.1 to find a Y in (4.4.5) that eliminates the nonzero entries above $M_{rr} = C$ by a similarity transformation. Continue the procedure to reduce the remaining submatrix A to a quasi-diagonal matrix.

Every $n \times n$ matrix A over the complex numbers was shown in Theorem 2.3.2 of Part II to be unitarily similar to an $n \times n$ upper triangular matrix T_s in which equal eigenvalues are next to each other along the diagonal. This T_s can be partitioned so that each diagonal block J_j is upper triangular with equal eigenvalues λ_j on its diagonal, but no two diagonal blocks J_j and J_k belong to the same eigenvalue. The diagonal blocks J_j are called *simple Jacobi* matrices and have the form $J_j = \lambda_j U + N_j$

$$J_j = \begin{bmatrix} \lambda_j & * & * & \cdots & * \\ 0 & \lambda_j & * & \cdots & * \\ \cdot & \cdot & \cdot & & * \\ 0 & 0 & 0 & \cdots & \lambda_j \end{bmatrix}$$

Simple Jacobi matrix

$$N_j = \begin{bmatrix} 0 & * & * & \cdots & * \\ 0 & 0 & * & \cdots & * \\ \cdot & \cdot & \cdot & & * \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = J_j - \lambda_j U \tag{4.4.13}$$

Nilpotent Jacobi matrix

Their direct sum J is a Jacobi matrix similar to T_s and hence to A.

The matrices $N_j = J_j - \lambda_j U$ are nilpotent strictly upper triangular matrices having 0's as entries on and below the diagonal. Any matrix similar to the simple Jacobi matrix J_j is the sum of a scalar matrix $\lambda_j U$ and a nilpotent matrix similar to the nilpotent Jacobi matrix N_j.

Jordan matrices. An $m \times m$ simple Jordan matrix $\Lambda_j = \lambda_j U_m + H_m$ has all its entries 0 except for m equal diagonal entries λ_j and $m - 1$ entries 1 in positions $(i, i + 1)$ to the right of the diagonal. A Jordan matrix Λ is a direct sum of such simple Jordan blocks Λ_j.

$$\Lambda = \begin{bmatrix} \Lambda_1 & & & 0 \\ & \Lambda_2 & & \\ & & \ddots & \\ 0 & & & \Lambda_q \end{bmatrix}$$

$$\Lambda_j = \begin{bmatrix} \lambda_j & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_j & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & \lambda_j & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_j \end{bmatrix} = \lambda_j U_m + H_m \tag{4.4.14}$$

An $n_j \times n_j$ nilpotent Jordan matrix $H = H_{\sigma_j}$ is one whose diagonal entries λ_j are all 0. It is characterized by a partition $\sigma_j = \{\sigma_{j1}, \sigma_{j2}, \ldots\}$ of $n_j = \Sigma\sigma_{jk}$ into the sum of positive integers σ_{jk}, which are the dimensions of its simple Jordan blocks.

We construct a *rank diagram* for H_{σ_j} by arranging the blocks in nonincreasing order of size, $\sigma_{jk} \geq \sigma_{j,k+1}$, and

placing σ_{jh} equally spaced dots in column h of an array in which the ρ_{j1} top dots are equally spaced in a horizontal line, and ρ_{jk} dots are in line k; see (4.4.16). The key fact is that this number ρ_{jk} is the number of nonzero diagonal blocks in $H_{\sigma_j}{}^{k-1}$ and, hence, is the difference between its rank $r_{j,k-1}$ and the rank r_{jk} of $H_{\sigma_j}{}^{k}$

$$\rho_{jk} = r_{j,k-1} - r_{j,k} \qquad r_{j0} = n_j \tag{4.4.15}$$

The rank differences ρ_{jk} form a standard partition ρ_j of n_j, which is said to be *associated* with the partition σ_j.

Example 3 Construct the rank diagram for a nilpotent Jordan matrix with $n_j = r_{j0} = 8$, $r_{j1} = 4$, $r_{j2} = 1$, $r_{j3} = 0$.

Solution: The rank differences ρ_{jk} are 4, 3, 1; the diagram is given by

$$\begin{array}{c|cccc} & 3 & 2 & 2 & 1 \\ 4 & \bullet & \bullet & \bullet & \bullet \\ 3 & \bullet & \bullet & \bullet & \\ 1 & \bullet & & & \end{array} \qquad \text{Rank diagram} \tag{4.4.16}$$

$$\rho_j = \{4, 3, 1\} \qquad \sigma_j = \{3, 2, 2, 1\}$$

and the powers of the nilpotent Jordan matrix are

$$H_{\sigma_j} = \begin{bmatrix} 0 & 1 & 0 & & & & & \\ 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & & & & & \\ & & & 0 & 1 & & & \\ & & & 0 & 0 & & & \\ & & & & & 0 & 1 & \\ & & & & & 0 & 0 & \\ & & & & & & & 0 \end{bmatrix}$$

Rank: $r_{j0} = 8$, $r_{j1} = 4$

$$H_{\sigma_j}{}^2 = \begin{bmatrix} 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & & & & & \\ 0 & 0 & 0 & & & & & \\ & & & 0 & 0 & & & \\ & & & 0 & 0 & & & \\ & & & & & 0 & 0 & \\ & & & & & 0 & 0 & \\ & & & & & & & 0 \end{bmatrix} \qquad H_{\sigma_j}{}^3 = 0 \tag{4.4.17}$$

Rank: $r_{j2} = 1$ $\qquad$ $r_{j3} = 0$

Theorem 4.4.3 A nilpotent Jacobi matrix N_j in (4.4.13) is similar to a unique nilpotent Jordan matrix H_{σ_j} that has the same rank diagram, and also has its blocks arranged in nonincreasing order of size.

Proof: Since like powers of two similar matrices have the same rank, all matrices similar to N_j must have the same rank diagram. Let N_j have the minimal equation $\lambda^m = 0$. Then its constituent matrices are $Z_{j1} = U$, $Z_{jk} = N_j{}^{k-1}$, $k \leq m$. Letting row i be a nonzero row of Z_{jm}, we construct a nonsingular matrix R with inverse S whose first m-rowed submatrix R_1 has in row k the ith row of Z_{jk}. Then $Z_{jk}N = Z_{j,k+1}$, so

$R_1N = HR_1$ where

$$R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} \text{ and } H = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{4.4.18}$$

Here $N = N_j$ is $n \times n$ and $H = H_m$ is $m \times m$. The matrix RNS, similar to N, and its powers, have the form

$$RNS = \begin{bmatrix} H & 0 \\ B & A \end{bmatrix} \qquad RN^pS = \begin{bmatrix} H^p & 0 \\ \sum_k A^{k-1}BH^{p-k} & A^p \end{bmatrix} \tag{4.4.19}$$

To transform RNS into the direct sum of H and A, we seek a rectangular matrix Y such that

$$\begin{bmatrix} U & 0 \\ -Y & U \end{bmatrix} \begin{bmatrix} H & 0 \\ B & A \end{bmatrix} \begin{bmatrix} U & 0 \\ Y & U \end{bmatrix} = \begin{bmatrix} H & 0 \\ 0 & A \end{bmatrix} \tag{4.4.20}$$

This requirement is satisfied if and only if Y satisfies the equation

$$AY + B = YH \tag{4.4.21}$$

which has the same form as (4.4.6). Here Theorem 4.4.1 does not apply, since A and H have all their eigenvalues 0. However, by equating the pth columns in (4.4.21) for $p = 1, 2, \ldots, m$ we obtain

$$AY_{\cdot p} = B_{\cdot p} + YH_{\cdot p} = Y_{\cdot p-1} \tag{4.4.22}$$

provided that $Y_{\cdot 0} = YH_{\cdot 1} = 0$. Choosing $Y_{\cdot m}$ arbitrarily we may then obtain the general solution of (4.4.22) and (4.4.21) by computing the remaining columns of Y in reversed order as follows:

$$Y_{\cdot m-t} = AY_{\cdot m-t+1} + B_{\cdot m-t+1} = \sum_{k=1}^{i} A^{k-1}B_{\cdot m-t+k} \tag{4.4.23}$$

For $i = m$ the additional requirement $Y_{\cdot 0} = 0$ is satisfied, since $Y_{\cdot 0}$ is the last column of $\Sigma A^{k-1}BH^{m-k}$, and this is a submatrix of RN^mS in (4.4.19), which is 0 inasmuch as $N^m = 0$. Hence Y is a solution of (4.4.21), and N is similar to the direct sum of H and A. Similar processes reduce A to Jordan form. Hence every nilpotent Jacobi matrix is similar to a nilpotent Jordan matrix having the same rank diagram. Applying Theorems 2.3.2, 4.4.2, and 4.4.3, we obtain the following important theorem regarding similarity.

Theorem 4.4.4 Every $n \times n$ matrix A over the complex number field is similar to a Jordan matrix Λ, called the spectral matrix of A.

The nilpotent constituent matrices for a Jordan matrix Λ are powers of direct sums of matrices like the matrix H in (4.4.18). The idempotent constituent matrices for Λ are diagonal idempotents.

(The concluding article in this series will appear in July.)

REFERENCES

1. Gantmacher, F. R., *Matrix Theory*. New York: Chelsea Publishing Co., 1960.
2. Frame, J. S., "The Rectangular Matrix Equation $AY + B = YC$" (abstract), *Notices Am. Math. Soc.*, vol. 10, Oct. 1963, p. 566.

BIBLIOGRAPHY

Bellman, R., *Introduction to Matrix Analysis*. New York: McGraw-Hill Book Co., Inc., 1960.

Dwyer, P. S., *Linear Computation*. New York: John Wiley & Sons, Inc., 1951, chap. 15.

Faddeev, D. K., and I. S. Sominskii, *Collection of Problems on Higher Algebra*, 2nd ed. (in Russian). Moscow: Gostekhizdat, 1949.

Faddeeva, V. N., *Computational Methods in Linear Algebra*. New York: Dover Publications, Inc., 1959.

Fettis, H. E., "A Method for Obtaining the Characteristic Equation of a Matrix and Computing the Associated Modal Columns," *Quart. Appl. Math.*, vol. 8, 1950, pp. 206–212.

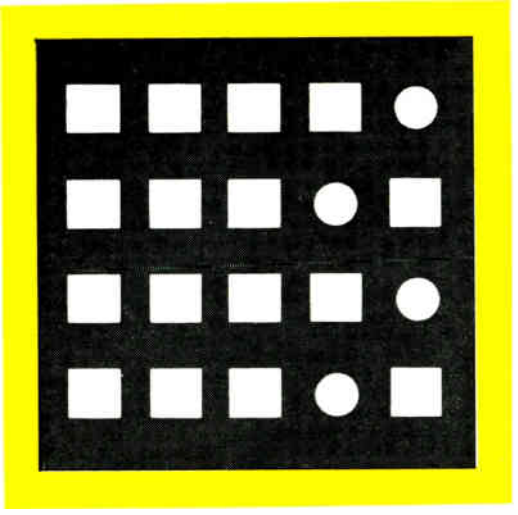
Forsythe, G. R., "The Sourlau-Frame Characteristic Equation Algorithm on a Digital Computer," *J. Math. Phys.*, vol. 34, 1955, pp. 152–156.

Frame, J. S., "A Simple Recursion Formula for Inverting a Matrix" (abstract), *Bull. Am. Math. Soc.*, vol. 55, 1949, p. 1045.

Souriau, Jean-Marie, "Une methode pour la decomposition spectrale a l'inversion des matrices," *Compt. rend.*, vol. 227, 1948, pp. 1010–1011.

Hybrid computation...

What is it? . . .
Who needs it? . . .



Parallel (analog)

Simulation of complex engineering systems is a growing need. The hybrid computer—evolved from the analog and digital computer—offers a means, otherwise impractical, for achieving many such simulations

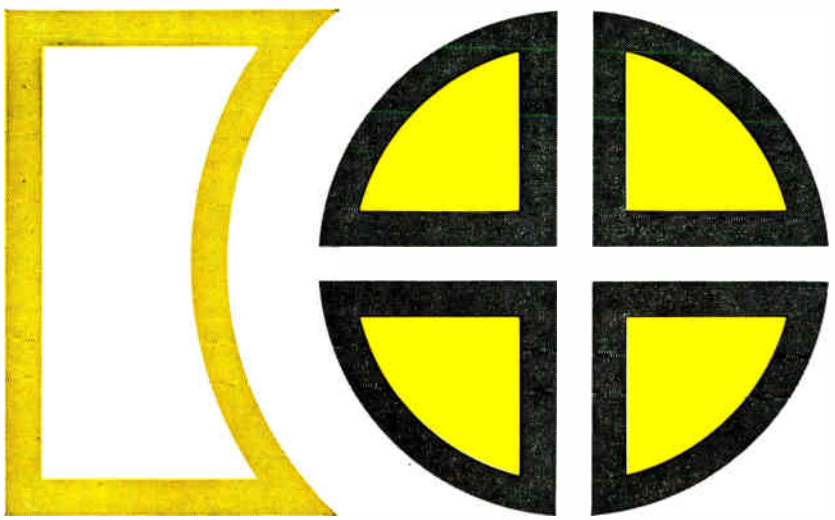
Thomas D. Truitt *Electronic Associates, Inc.*

During the 1950s the capabilities of electronic computers stayed well ahead of the needs of the average user. Such was the case in both the analog computer and scientific digital computer fields. One effect of this situation was the formation of two schools of experts with opposite views on the choice of the "best" general-purpose scientific computer. Differences of training, experience, and semantics led to a serious barrier to interchange of opinion between these two groups. At best, the fashionable middle-of-the-road position was to admit that each computing technique "had its place," which did little to break down the barrier. Only with the appearance of a computational task that could not be accomplished satisfactorily by either type was the barrier cracked and a small opening made.

The earliest attempts to combine the computation of analog and digital computers took place about 1958 at the Convair Astronautics plant in San Diego and at the Space Technology Laboratories in Los Angeles.[1-3] In both cases the job at hand was the complete mission simulation of the trajectory of a long-range missile. The speed of the analog computer was a necessary element in the study to permit a real-time simulation of the rapid motion of the vehicle and of control surfaces. However, the dynamic range required of the simulation was in excess of that of the best analog computers. That is, the ratio of the total range of the trajectory to the required terminal phase resolution (a dynamic range of 10^5 to 10^7) was greater than 10^4, the upper limit of analog computer dynamic range for small programs. Hence the digital computer was used to calculate those variables for which such dynamic range was necessary. The most important of these were the calculation of navigational coordinates: the open integration of velocities to determine the vehicle's position plus the dynamic pressure, a function sensitive to altitude and velocity.

It is fortunate that in such long-range aerospace trajectory simulations the variables with wide dynamic range requiring precise calculation are not, at the same time, rapidly changing. Moreover the "high-speed" variables do not require precise calculation. The early combined computer systems employed the largest and fastest digital computers available at the time—UNIVAC 1103A and IBM 704—together with 300 to 400 amplifiers of general purpose PACE analog equipment. In both cases even these computers were only just fast enough to perform the required repetitive calculations for the slowly changing variables of the simulation in real time.

Since the first combined computers were installed, at least a dozen computer laboratories have employed general-purpose analog and digital computers together to solve simulation problems, and a number of attempts have been made to devise special purpose systems of analog and digital devices. Among the latter are the CADDA of the National Bureau of Standards and the pulsed analog computer of the MIT Electronic Systems Laboratory.[4-9] Hybrid computers of a unique type are the combinations of a general-purpose digital computer and a digital differential analyzer (DDA), illustrated by the Bendix G15 with the DA-1 attachment and the Packard Bell PB250 with the Trice DDA.[10] The former system consists of a small, slow computer with an even smaller serial DDA ($50 000 and $10 000 respectively). In contrast to this the Packard Bell system combines a small, medium-speed computer ($40 000) with a large serial-parallel DDA computer ($500 000). Among the



Sequential (digital)

systems of general-purpose computers, generally large analog computers have been combined with both large (IBM 7090) and small digital computers (PB 250, LGP 30).[11–14]

A brief analysis of the applications to which existing installations of combined systems have been applied leads to these generalizations: For the most part, the analog computers in these systems have been employed in a normal manner to simulate the dynamic behavior of physical systems by solving sets of nonlinear, ordinary differential equations, while the digital computer has performed one or more of the following three functions: complex control logic, storage of arbitrary functions or sampled analog functions, and high-precision arithmetic primarily for numerical integration. Examples are:

1. *Analog computer plus digital control logic.* A system that in itself contains discrete control functions of continuous dynamic variables is appropriately simulated by a hybrid computer. The kinetics of a chemical process are simulated by continuous analog means while its digital control system is represented by a digital program.[15,16] Similarly, an aerospace vehicle with an on-board digital computer, control system or autopilot is simulated by hybrid techniques.

2. *Analog computer plus digital memory.* A common difficulty in the simulation of a chemical or nuclear reactor is to provide an adequate representation of the transport of fluid in pipes from one point to another—from reactor to heat exchanger. The simulation of this transport delay of a dynamic variable, such as the time variation of the fluid temperature, is neatly accomplished by the use of a digital computer for storage of the temperature function for a fixed, or variable, length of time. Digital computer memory has also been used effectively to store multivariable arbitrary functions—an operation which is seriously limited in the analog computer.

3. *Analog computer plus digital arithmetic.* This type of application is the "classic" one where the digital computer is used to perform precise, numerical integration of space vehicle velocities to keep track of the exact position of the vehicle over a very long-range flight.

It should be noted that a significant difference is apparent in the applications of computer systems employing a very small digital computer and those with large, very fast computers. In general, the small machines are limited to execution of control logic programs, one or two channels of transport delay simulation, or limited function generation programs. Since numerical integration and complex function generation by digital programs require considerable time for each calculation and for each discrete step in time, only the fastest digital machines can be used effectively for these tasks.

Evolution of hybrid simulation

The term "computer simulation" appears in so many contexts it is important to emphasize that its use in these pages is limited to the modeling of complex physical systems for the study of their dynamic behavior. These systems are represented by sets of differential equations, algebraic equations, and logic equations. As in most simulation studies, the objectives of hybrid simulation may be experimental design, prediction and control, design evaluation, verification, or optimization. It is not expected that the important applications of hybrid simulation will include data processing system simulation; information handling simulation; business system

simulations; simulation of television coding, character-reading machines, communications coding systems, etc. (These are all digital computer simulation applications.) Similarly, hybrid computers are not warranted for the simulation of circuits, devices, and systems for which the analog computer is quite adequate. It is in the simulation of total systems bringing together components, some of which are suited for digital and some for analog simulation, that the newer hybrid techniques are required. There are probably few, if any, simple applications.

If hybrid computation can be said to be a field of endeavor, it must be considered to be in the formative stage. Developments to date have led to equipment configurations and programming techniques that were dictated by specific problems and limited objectives. The growth rate of the field will be determined by the extent to which a broader view is taken of hybrid computer programming techniques and applications. The greatest advances in computers have been made when the experiences of users and programmers have been brought to bear on design of equipment. For the most part, hybrid computers of today consist of general-purpose analog and digital machines that are *not* designed for hybrid operation but tied together by "linkage equipment" designed only to solve the communication problem. This has been a necessary first step. Newer hybrid systems will be designed for efficient solution of hybrid problems and for convenient programming also.

The next generation of hybrid computers will undoubtedly achieve a greater degree of integration of parallel computing elements with the sequential stored-program principle. Eventually patch-panel programming of parallel elements will be replaced by an automatic system, thus affording a fully automatic method for computer setup and check-out. Even today a secondary activity of the digital part of the hybrid machine is the partial automation of setup and check-out of the analog computer. This feature becomes increasingly important as the computer system grows in size and the programs grow in complexity.

The elements of hybrid computers

Digital computers. Many conflicting factors have influenced the choice of digital computers used in hybrid systems. Computer speed and economics have probably been dominant. Because there are so many computers on the market today that have sufficient speed and that span the complete range of prices, it is more instructive to examine the essential features.

Speed. The speed of execution of arithmetic operations is most important, and this is a function of memory access time and multiplier speed. The access time of drum and delay line memories is too slow. Magnetic core access times of 2 to 5 μs are currently popular. This means the time for addition of two numbers is 4 to 10 μs. Multiplication and division take longer; times of 15 to 40 μs are generally available and quite satisfactory. Overall program speed can be increased by the use of index registers—three registers are desirable; more are useful. Special instructions for subroutine entry, for executing commands out of sequence, and for testing and skipping can help increase computing speeds.

Word structure. The basic requirement is for a fixed-point binary word of at least 24 bits. Because round-off errors affect the last several bits, a smaller word size would result in a dynamic range limitation of less than 10^6. A longer word may be useful in a few applications where fixed-point scaling may be difficult. Floating-point computations may make things easier for the programmer but should not be depended upon at the expense of computational speed. It may be noted that the equivalent of fixed-point scaling is a necessary part of the analog program, and hence floating-point operations may not prove as advantageous as for some all-digital programs. Decimal format and character-oriented machines do not offer any advantages for hybrid computation, and usually they are slower than binary computers of equivalent size.

Input–Output (I/O). High data rates in and out of core memory and any feature that minimizes loss of computing time for input-output operations are highly desirable. In addition, a fast, flexible means for communicating control signals to and from the analog section of the hybrid system is necessary. Three kinds of control signals are usually provided: interrupt, sense lines, and output control signals. It is by means of these controls that the sequential operations of the digital machine are made compatible with the parallel simultaneous operations of the analog machine. Since communications must be made with many points in the analog computer, a number of these control signals are needed. Interrupt signals, from outside the computer stop the current sequence of calculations and force transfer to another sequence. Sense lines simply indicate to the digital program the current state of devices outside the computer, which may be sensed by specific programmed instruction. Other programmed instructions will send control signals outside the computer on the output control lines.

Memory. As already noted, the digital computer main memory should be a high-speed, magnetic core. Most hybrid applications do not require a large memory for either program instructions or data, and four, eight, or twelve thousand words of core memory should suffice. Larger memories may be desired for special digital programs and larger hybrid problems when more experience has been gained in this field; thus expandability of a memory to 16 000 words is a good feature. Newer computers are being introduced with small, very-high-speed, "scratch pad" memories. Such memories may have cycle times of less than one ms and are used to store intermediate arithmetic results. This feature increases the overall computation speed of the computer.

The normal manner of operating an analog computer involves a fair amount of noncomputing time when the computer remains in the HOLD or RESET mode. These intervals may range from seconds to minutes while adjustments are made, pots are set, or recorders changed, or while the programmer analyzes results. It is not possible for the analog computer to operate on other programs at these times; however, with a hybrid system in which such waiting periods are likely to occur also, it is reasonable to consider having the digital computer work on a different program during the intervals, whatever their length. Appropriate "interrupt" and "memory lock-out" features are possible to permit time sharing of the digital machine without affecting the hybrid program and without the danger of one program interfering with the other. The secondary program (involving a strictly digital problem) is simply stored in a "protected" part

of the core memory and it utilizes all the bits of time not required by the hybrid program.

Peripheral equipment. In many digital computer installations, the investment in peripheral equipment rivals that in the central computer. Current hybrid computer applications require only a minimum of digital peripheral equipment. The graphic output equipment associated with the analog computer is sufficient for computational results: punched paper-tape reader, punch, and typewriter may be all that is required for programming. Larger systems in the future will employ punched cards and magnetic tape, primarily for rapid change-over of problems and automatic check-out. Large-capacity off-line data storage does not appear necessary.

In summary, the digital computer must be characterized as a *sequential* machine. For effective use within a hybrid system the machine must (1) have sufficiently high internal speed for it to appear as though a number of calculations were taking place simultaneously; (2) be organized for maximum speed in executing mathematical calculations; and (3) have efficient means for input and output of data during calculation.

Analog computers. In contrast, the analog computer is a parallel machine, with many computing components and I/O devices operating in concert. There are few, if any, features of the modern analog computer that are not appropriate to a hybrid system. However, only the largest analog machines have been used for general-purpose hybrid simulation. The common measure of a large computer is that it has 100 to 200 operational amplifiers; because two or more computers may be "slaved" together, larger systems are possible.

Analog computer features that are important for hybrid systems can be simply listed as

1. Integrators with multiple time scales
2. Amplifiers for tracking and storing voltages
3. Very fast control of the modes of individual amplifiers
4. Automatic remote control of the setting of potentiometers
5. Fast, accurate multipliers and trigonometric resolvers
6. High-speed comparators with logic signal outputs
7. Electronic switches (logic signal gating of analog signals)

In the early days, logic equations or switching functions were programmed with relays and stepping switches, which were connected to the patch board by various means. Present-day technology employs electronic switching of integrator modes and voltage signals at high speeds, and the delays inherent in relay devices can no longer be tolerated for logic operations. The logic building blocks common to the digital computer designer, (flip-flops, gates, inverters, monostable multivibrators, shift registers, and counters) are ideally suited to these operations. Thus with electronic switches replacing relay contacts, logic modules have become an integral part of all new, large, analog systems. These modules are programmed like the other analog components by interconnections at a patch panel. Many signals occur simultaneously but they are logic signals—two values, zero and one, that change as functions of time. Input signals to logic programs come from comparators, manual switches, and external control signals. Logic program outputs go to integrator mode controls, storage amplifier controls, electronic switches (DA switches) to gate analog signals. As will be shown later, it is essential for a hybrid system to have a very significant complement of digital logic components. A few gates and flip-flops are not sufficient. The potentialities for use of logic components in an analog computer for hybrid operation are so great that the EAI HYDAC digital operations system is an entire computer console with its own patching system used entirely for the programming of digital components for parallel computation. The console is really a complete logic computer. It is used together with a conventional analog computer to form what is truly an all-parallel hybrid computer.[17–21]

In summary, the modern analog computer must be characterized as a *parallel* machine. It is not solely a computer for continuous variables. It is a parallel assemblage of building blocks: integrators, multipliers, etc., for continuous variables; and flip-flops, gates, counters, etc., and "digital" circuits for discrete variables. It is organized for convenient representation of an "analogous" physical system by means of a computer model constructed of these building blocks.

Conversion devices. In providing data communication between a sequential computer and a parallel computer, three kinds of devices are commonly used: the multiplexer, the analog-to-digital converter (ADC), and the digital-to-analog converter (DAC). In addition, all early systems have employed a timing and control unit that performs a relatively fixed set of operations, with manual switches to select options such as sampling frequency, and number of channels. Such "linkage systems" thus consisted of a timer unit plus a group of linkage building blocks prewired to perform a specific task. With the integration of digital logic components into the parallel computer, however, greater programming flexibility is possible by use of these logic units for timing control of the data conversions. Furthermore, the converters and multiplexer can act very naturally as additional building blocks in the parallel computer. Thus it is likely that future hybrid systems will simply incorporate the "linkage system" within the parallel computer.

Usually several or many analog signals in a hybrid program will be sampled, converted, and transmitted periodically to the sequential digital program. The numbers, of the several sequences of numbers to be entered into the core memory, can be accepted only one at a time. Because this is so, the conversions from voltage to number form can be performed one at a time—first from one analog variable and then from another. The multiplexer is used to select one from many analog signals, to step through a sequence of signals, and thus to furnish voltage input signals to the ADC.

The output of commonly used ADCs is a binary number of 10 to 14 bits. A 13-bit binary output probably is the best compromise; it represents a resolution of one part in 8000, and resolution of analog voltage signals is at best one part in 10 000. Conversion times ranges from 50 to 300 μs. A typical time of 100 μs would allow the converter to be shared by 16 analog signals each with a frequency spectrum extending to 20 or 30 c/s.

DAC units should have the same binary word size as the ADC, except for special low-accuracy uses. Conversion times are not determined by the converter but rather by the bandwidth of the analog amplifier following

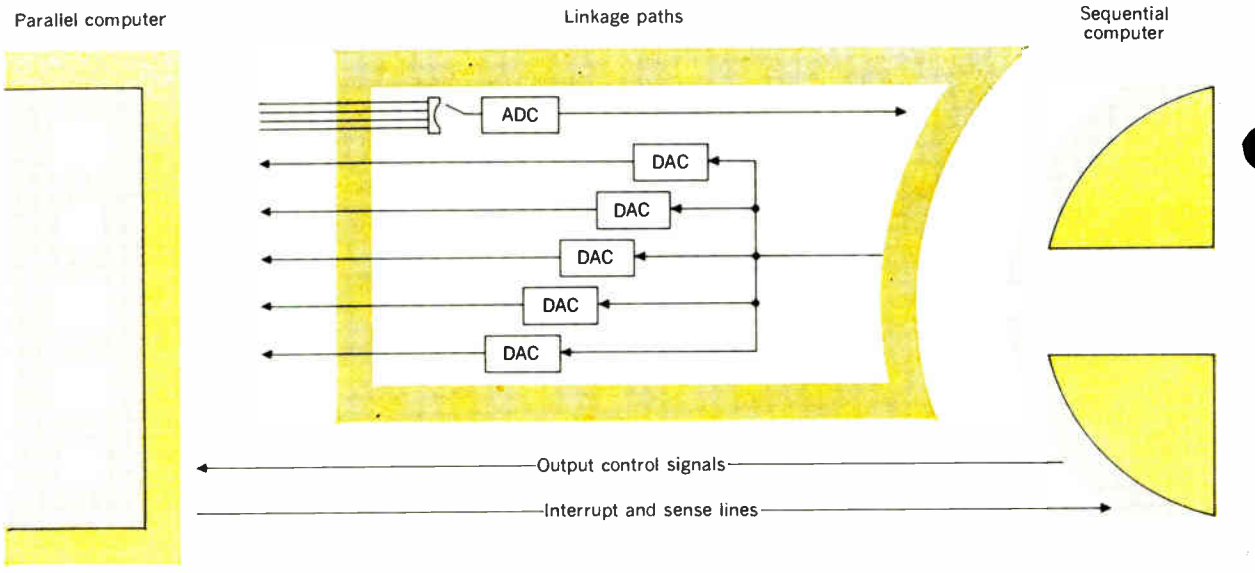


Fig. 1. ADCs and DACs are format converters, changing voltage to numbers and vice versa. As components of the parallel machine, the converters, together with logic components, must act to "match impedances," i.e., resolve the incompatibility between the parallel and sequential computer programs.

the converter. Data from the sequential computer appear only one word at a time, and some means for retaining the latest value, of each sequence of numbers, for each output function is needed. The sequence of numbers coming from the computer may first be converted to voltage values by a single DAC, and then distributed to storage amplifiers for each channel. It is more customary, however, to hold each of the latest words for each channel in a digital register, which is an integral part of the DAC assigned to each channel (Fig. 1).

Special forms. As a passing thought it may be noted that while the primary emphasis here is being placed on the distinction between parallel and sequential operation, the term "hybrid," historically, has been used to imply the combination of continuous and discrete calculations, and that therefore consideration might be given to two special kinds of hybrid computers:

The parallel hybrid computer. Parallel hybrid computer is a proper term for the EAI HYDAC 2000 machine. This system combines an analog computer with a general-purpose system of programmable logic building blocks, multiplexer, ADC, DACs, digital memory units for storage of sampled analog functions, and several digital numerical adders and subtracters. The application of this system encompasses an intermediate range of hybrid problems, such as: (1) transport delay simulation[20]; (2) single and multivariable function generation; (3) logic control systems[19]; (4) automation of the analog computer for parameter searches and optimization studies[16]; and (5) simulation of numerical and sampled data control systems.[17,18]

The sequential hybrid computer. The sequential hybrid computer is exemplified by the experimental pulsed-analog-computer techniques developed at MIT for use in an aircraft flight trainer.[4] This system employs a sequential digital computer which controls a small number of analog functional components—one multiplier, one reciprocal generator, one integrator, and several adding units. These units are interconnected and receive inputs by digital program control. They form, in effect, "analog subroutines" for the sequential computer.

The sequential-parallel hybrid computer

The term "hybrid" is most appropriately used to indicate the combined use of sequential and parallel computing techniques because in part of the machine many operations are taking place simultaneously, and many time-varying problem variables exist in parallel; while elsewhere a number of operations take place, one at a time in a repetitive manner, so as to effect the generation of several problem variables, *as if* they occurred simultaneously. Furthermore, the parallel computation is tied to a real-time base: the very passage of time itself accounts for the changing of the basic independent computer variable. The sequential program is asynchronous—not controlled by a clock. Operations are executed in sequence at whatever rate is possible, and for any reference to be made to the actual elapsed time, external communication is necessary. This basic incompatibility requires that the interface between the two types of operation embody more than the simple format conversions performed by the ADC and DACs. It is necessary for data and control information in the parallel machine to be available to the sequential machine and, conversely, that the latter be able to send data and control signals to many points in the former. Coincidence or simultaneity of events communicated to and controlled from the sequential program are particularly difficult to handle. The logic and data control of the interface equipment must resolve these differences in timing and operation. What might be termed an impedance-matching device is needed between the parallel and sequential program in order to make most efficient use of both machines.

A simple example. An example will illustrate some basic considerations in defining a general-purpose hybrid system. First the operation of the simplest of linkage

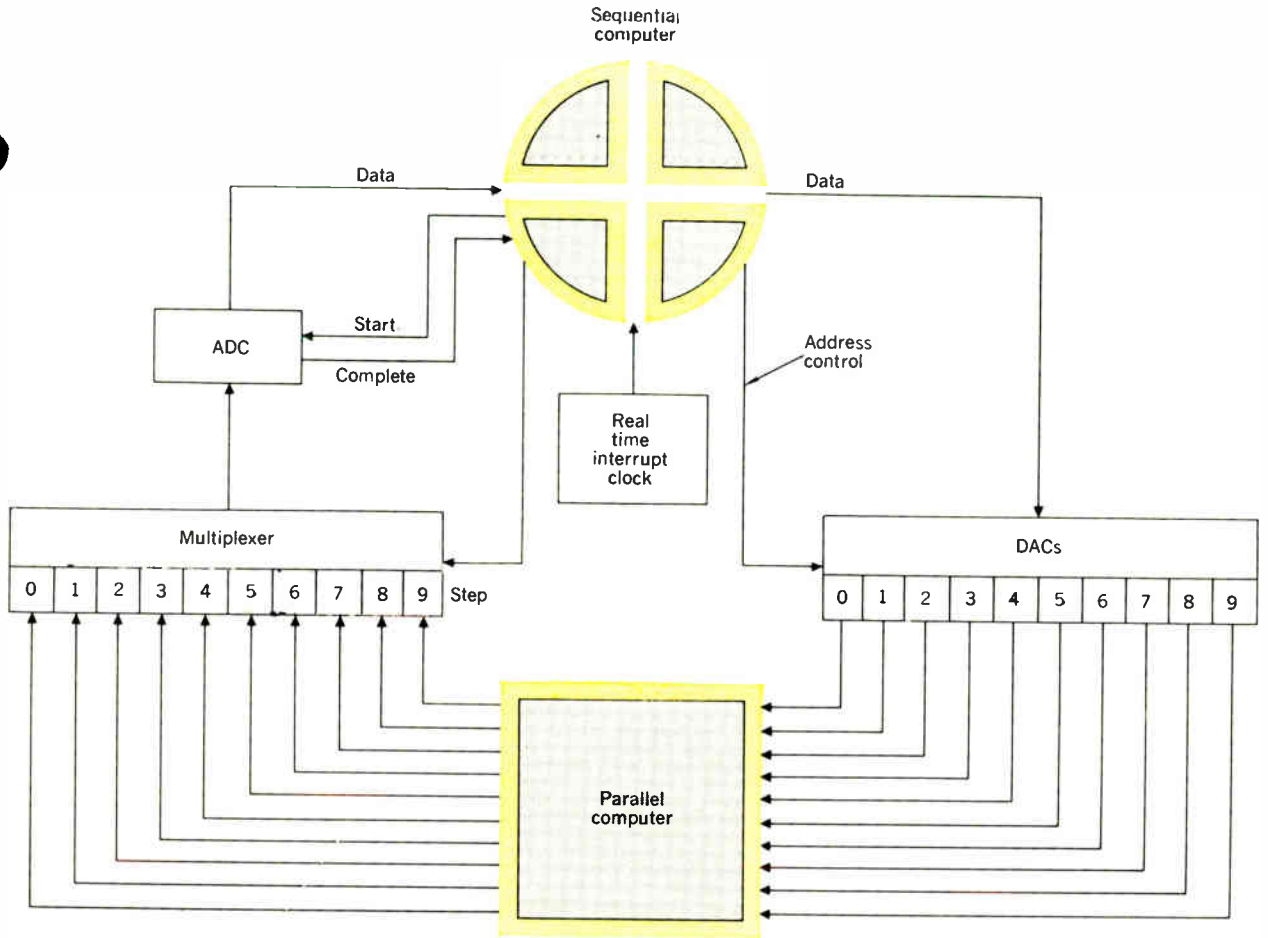


Fig. 2. An early hybrid system configuration. The sequential program controls the timing of the conversion cycle. The cycle is initiated by the interrupt clock. For a typical problem the clock might be set for 100 c/s; with 7 to 9 ms per cycle for calculation.

systems calls for an ADC and multiplexer, a number of DACs, and an interrupt clock. The flexibility of the stored-program digital computer is relied on for control of these units. Assume ten analog signals are to be converted one at a time. These words are placed in memory (average program time: 40 μs per word) and then about 7.5 or 8.0 ms of sequential digital calculations takes place, followed by output from memory of ten words (20 μs per word) to ten DACs. The entire cycle requires 7500 + 10×40 (input) + 10×20 (output) = 8100 ms of digital program time. If it is assumed the conversion of the data (A to D) requires 150 ms per word, then 1 additional ms, or 9 ms are needed if everything proceeds sequentially. Assume further that because of the frequencies of the analog signals, it is necessary to sample at least some—and therefore all—of the channels at 100 samples per second. A real-time interrupt clock is set at 100 c/s. This timer unit is an adjustable oscillator that sends an interrupt pulse to the digital computer. The latter then activates the ADC, waits 150 μs for a completion or READY signal, steps the multiplexer to the next channel, stores the converted word in memory, and then repeats this cycle ten times. With the tenth step the multiplexer resets to the first channel. The program then proceeds with the 7.5 ms of calculation, outputs ten words, one at a time to ten DACs, and then waits for the next interrupt pulse (see Figs. 2, 3, and 4). Manual controls are provided for selecting the interrupt clock frequency and the number of channels in the multiplexer stepping cycle.

This is certainly a simple system and it appears to satisfy the basic requirements for communications. Some of the shortcomings of the system are apparent: sampling and outputting of each channel do not take place simultaneously; 15 per cent of the sequential program time is "waiting" time; and 3 to 5 per cent is used to select and control devices external to the sequential computer. In other problems, these percentages may be higher. Another weakness in this system is that it was not designed to be a general-purpose system. It is restricted in application to a class of problems for which the periodic "input–calculate–output" cycle is useful.

System improvements or variations. By programming the parallel digital components of the parallel computer to perform timing and control functions, these changes to the system are suggested:

1. *Simultaneous sampling:* If the sequential program operates on two or more of the input numbers together to calculate an output, then errors may occur since the input numbers were sampled at different times and correspond to different values of the independent variable. A similar effect may occur at the output since the

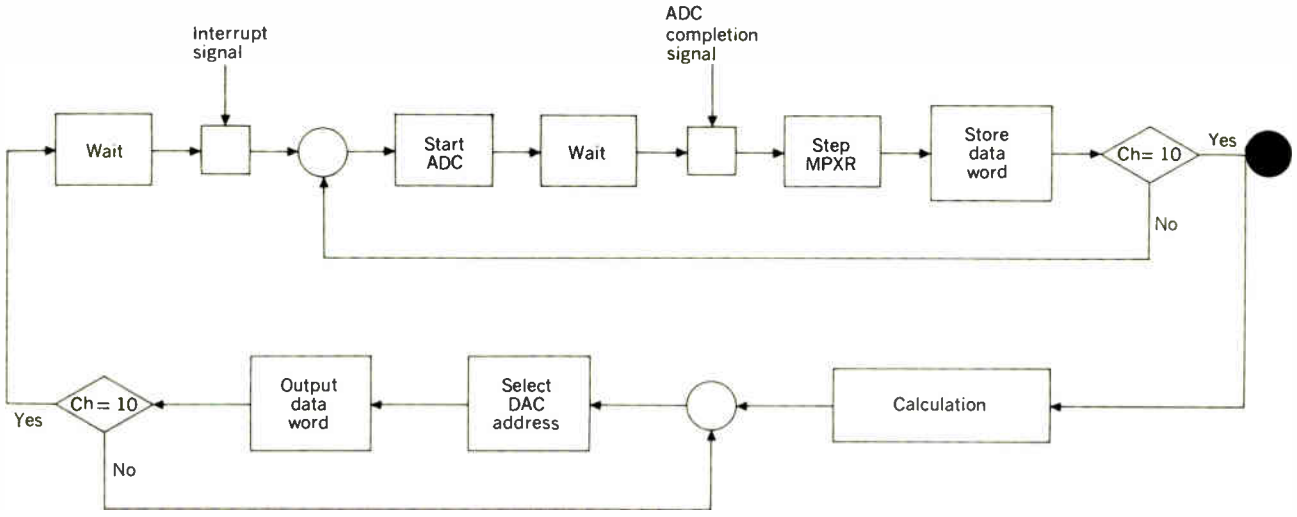


Fig. 3. Sequential program flow diagram, for the system of Fig. 2, shows steps required for control of converters.

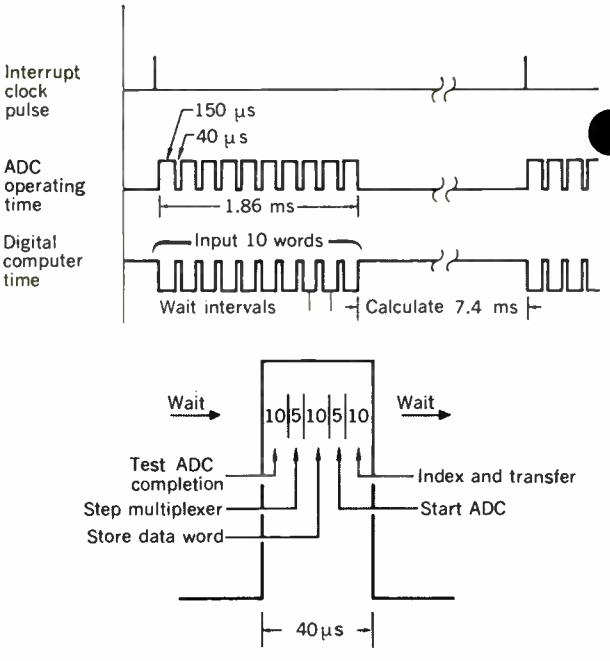
numbers in a group of ten appear at the ten DACs at different times. It is certainly possible by numerical means to compensate for the errors, at the penalty of additional program time.[22,23] The common solution is to add memory to each of the ten input and ten output channels. Ten track–store amplifiers are added before the multiplexer and a control signal causes them all, by storing the voltages, to sample simultaneously. At the output, 13-bit registers are added in front of the DACs. When all ten registers have been loaded, a transfer pulse causes all DAC values to be updated at once.

2. *External timing of ADC and multiplexer:* Sequential program time can be saved by permitting the control of the ADC, multiplexer, track–store cycle to be controlled externally. A simple clock, a counter, a flip-flop, and a group of logic gates will permit the input conversion cycle to run at its own rate—interrupting the sequential program only at the completion of a conversion. Thus the conversion time can overlap the calculation time, eliminating the waiting time. Upon interrupt, only 10 to 20 μs may be required per sample; many control steps are eliminated. Similarly, on output, the addressing and selection of output channels can be done by simple circuits rather than using sequential program time (Fig. 5).

3. *Real-time clock to establish sampling frequency:* If the sequential calculation involves numerical integration over a long term, the accuracy of the sampling interval is just as important as the round-off and truncation error in the numerical calculation. Although numerical means may be resorted to for very accurate integration, in a hybrid program the calculations still need to be referred to a real-time base. This is done by using an accurately calibrated source for setting the sampling interval, or frequency. A good high-frequency, crystal-stabilized oscillator is an important part of the parallel digital subsystem. Sampling frequencies lower than the oscillator frequency are selected by use of preset counters.

4. *Use of sense lines to reduce number of conversions:* In the simple example problem, only whole number data are transmitted to the sequential program. Thus, if the relative magnitude (greater or less than) of two

Fig. 4. Typical times are shown for steps in conversion cycle of the system of Fig. 2. All ten channels are converted and stored in memory as fast as possible before proceeding with calculations—at the expense of "waiting time."



analog signals is needed in the digital calculation, the two numbers must be converted, stored, and then compared. This can be accomplished more simply by use of an analog comparator, the output of which is sent to the digital computer by a sense line—saving time and equipment. The state of any parallel logic component may be monitored conveniently by sense lines. These are tested in one memory cycle (2 to 5 μs). If many such communications are needed, the savings will be significant. Sense lines should also be added to allow the sequential program to determine the modes of the

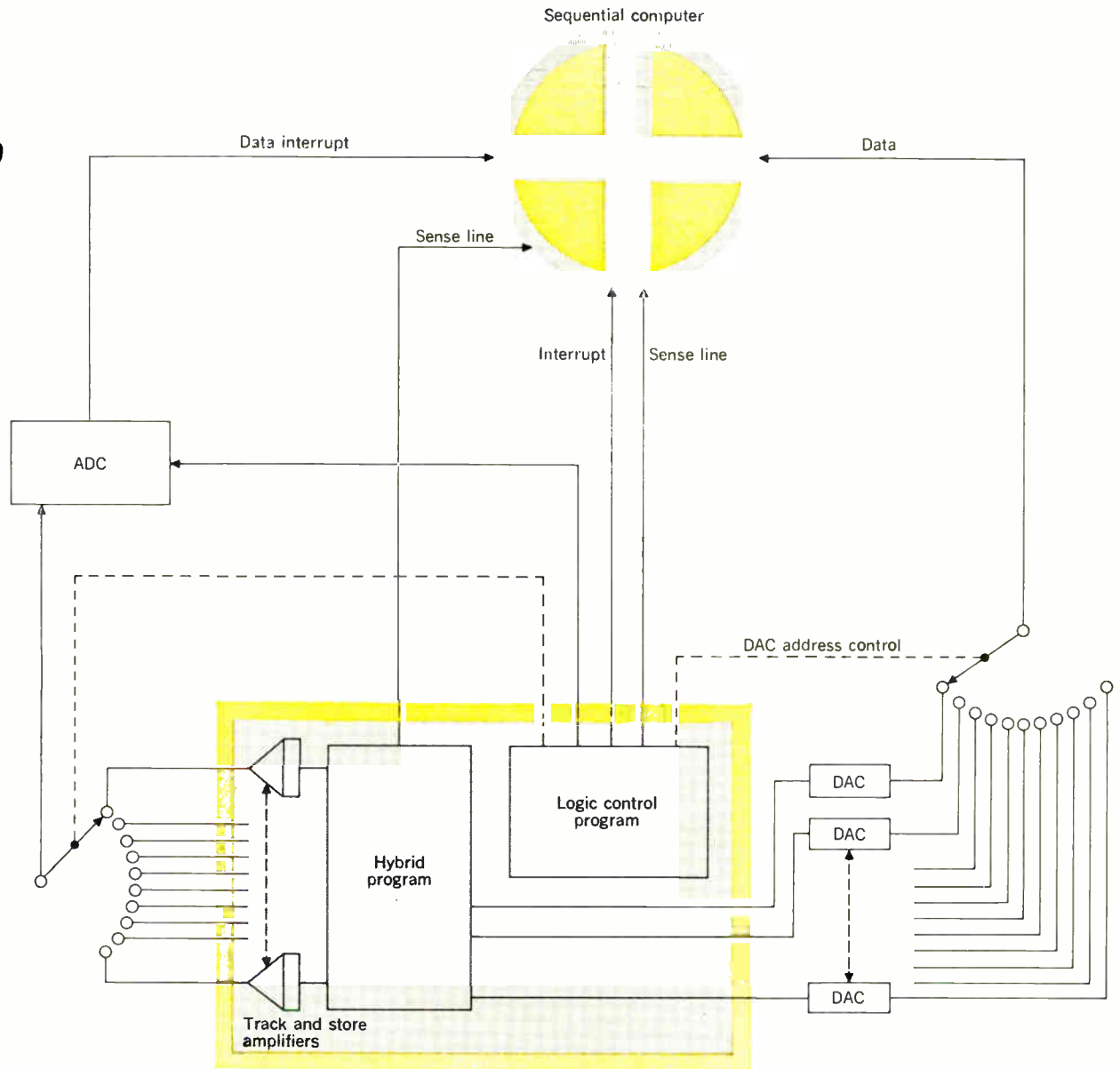


Fig. 5. Several improvements to the interface system of the example of Fig. 2 include: simultaneous sampling, external ADC and multiplexer timing by a parallel logic program, use of sense lines to reduce conversion channels, detection of random events, multiple sampling frequencies or asynchronous sampling, or both.

analog computer, the relative sizes and signs of error quantities, and the states of recording devices.

5. *Detection of random events:* With fixed, periodic sampling, the sequential program cannot tell exactly when events take place in the parallel machine. With comparators and parallel logic, complex functions of analog variables can be monitored. For example, it might be required to determine when the overshoot in x_1 exceeds x_2 after the third cycle; but only when x_3 is negative and x_4 is less than x_5. After determination, the sequential program can be interrupted to perform specific conversions and calculations—asynchronously with respect to the primary conversion cycle. In this manner, the parallel logic avoids the delays in the sequential program and uses the latter only when required. The parallel logic program analyzes the data, interrupts the sequential program, and sets up the proper channels for conversion in and out of the digital computer (Fig. 6).

6. *Multiple sampling frequencies:* In the example problem, all channels are sampled at a frequency determined by the highest frequency of any one channel. Often there may be two or more groups of variables with different ranges of variable frequencies, when it would be appropriate to sample each group at different frequencies. Another approach using different sampling rates is to use several eight-channel multiplexers in cascade so that the output of two of them feed two channels of a third which feeds the ADC. On each cycle of the third unit the first two are stepped, yielding different variables for those two channels on each cycle. Alternatively, each time the third steps to the two special inputs the corresponding multiplexer makes a full cycle.

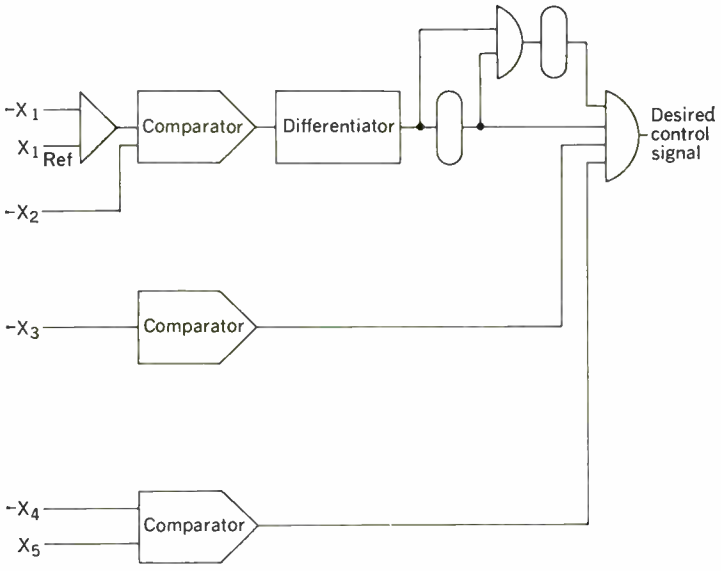


Fig. 6. Parallel logic program to detect random events in the parallel computer. "Desired control signal" interrupts sequential program when overshoot in x_1 exceeds x_2, after the third oscillation, but only when x_3 is negative and x_1 is less than x_5; x_1 through x_5 are analog voltages.

Timing control of these operations is performed by parallel logic components.

7. *Asynchronous sampling:* A completely asynchronous conversion system has been designed by one computer laboratory, in which the sequential program is interrupted only by comparators. Twenty analog problem variables are compared to reference values that are adjusted by the digital computer when necessary. Each comparator calls for conversion of some group of variables (the same variables may be called for by different comparators). When two or more comparator signals occur simultaneously or during a conversion operation, two levels of priorities are set up by logic elements to determine what interrupts are to be made. Though the system appears complex, the operation is simple in the parallel computer, with good use made of the sequential computer time.

A longer list of useful variations in the control and timing of sequential–parallel communications can be compiled. For the most part, however, they should be explained in terms of the particular problem applications.

Operating times for typical mathematical functions. Emphasis on efficient utilization of the sequential program time, high arithmetic speeds, and programming tricks to gain speed seems fully justified when one examines the sequential operating times for several typical mathematical functions, which, on the analog computer, would be executed continuously and in parallel.

a.	The sum: $a + b + c + d$	40 μs
b.	The expression: $ax + by + cz$	160 μs
c.	Sin ωt or cos ωt	215 μs
d.	Square root of $x^2 + y^2$	432 μs
e.	Generate $z = f(x,y)$, where two dimensional interpolation is required between functional values evenly placed in x and y	0.5 to 1.5 ms
f.	Rotate a vector through three coordinate angles	2 to 6 ms
g.	Perform one integration of a single derivative for a single time step	0.1 to 1.3 ms

A program of three first-order differential equations where derivatives are calculated from the functions (a) through (f) would not be a large program; and yet for a single step in time, the calculation time would be about 11.2 ms.

$$\frac{dx}{dt} = -x + y + z + f(x,y)$$

$$\frac{dy}{dt} = ax - by + cz$$

$$\frac{dz}{dt} = x^2 + y^2 + \sin zt$$

Allowing another millisecond or two for control and input–output instructions, one can estimate the real-time speed performance of this program. The speed is best expressed in terms of the useful upper frequency (at full scale) in a problem variable that can be represented by the computer. Although the example equations have no real meaning, the frequency limit for such a program is about 1½ c/s. This does not seem like very fast performance of a few simple equations. On the other hand, it is fast compared to frequencies of some of the variables in an aerospace simulation program for which digital precision is required.

Sampling rates. When parallel and sequential machines are connected in a closed loop, it is assumed that at least part of the task of the sequential program is to accept sequences of sampled values of continuous input, and calculate functions of these inputs, which are then transferred out as sequences of numbers to be smoothed into continuous signals. The digital computer appears to simulate a parallel computer; as in a movie projection, the effectiveness of this simulation is determined by the ratio of the frame or cycle rate to the rates of change of the signals communicated, and by the time resolution of the person or computer receiving the information. Thirty frames per second would not be fast enough to catch the information in the trajectory of a hummingbird. A higher rate is needed for an accurate recording of the flight. The human eye, however, cannot resolve time intervals of less than 1/30 second. Thus an accurate recording can be transmitted to the eye only by a time-scale change to slow motion. Fortunately the parallel computer has a time-resolving power sufficient to detect the shortest practical cycle time on the sequential computer; so the limiting factor in determining useful cycle frequencies is the rates of change of the variables that pass between the two computer sections. It is customary to speak of the bandwidth or spectrum of these signals—or more particularly the highest useful frequency that must range over the full magnitude scale. The sampling rate or cycle rate must be selected in terms of the number of discrete numbers or voltage samples necessary to represent this highest frequency at the desired accuracy.

In sampled-data theory, the "sampling theorem" states that the sampling frequency must exceed two times the highest signal frequency if all the information in the original signal is to be retained.[24] That is, some number greater than two samples per cycle is necessary. Another important point comes from the theory that in sampling

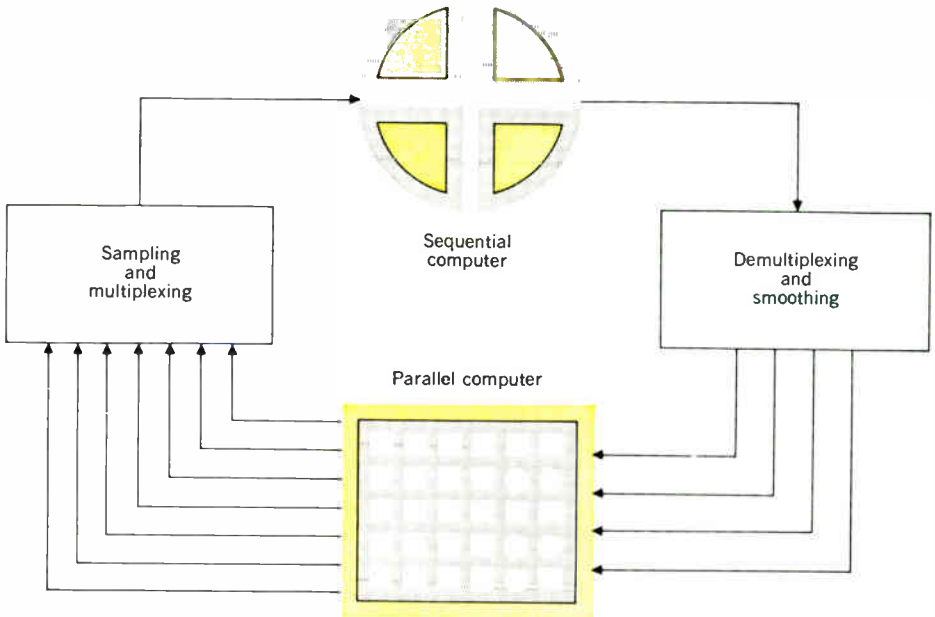


Fig. 7. A sequential-parallel hybrid computer. The noise and delays due to sampling and multiplexing on one side and the magnitude and phase errors due to demultiplexing and smoothing on the other are dominant factors in determining the proper sampling frequency.

voltages at the input to the digital computer the rate must exceed twice the frequency of any signal present. If noise signals are present that are higher in frequency than the desired signal and exceed one half the sampling frequency, it is possible for this noise to be reflected into the signal spectrum, thereby destroying useful information. This can be avoided with appropriate noise filters. If this were the only limitation it would be fortunate. However, too few samples per cycle make it difficult for the sequential program to extrapolate and predict what takes place between samples. It is possible, of course, by numerical means, but at a cost in program time. Furthermore, numerical algorithms for integration are sensitive to the ratio of sampling interval to the rates of change of the variables, and the calculations may become unstable if too few samples are used (Fig. 7).

The most important criterion for determining sampling rates appears in the conversion of the discrete data to continuous analog functions. Two sources of error affect the accuracy of the resultant continuous function. The first is the delay due to conversion and the sequential program itself. The output numbers are functions of input numbers that were sampled at an earlier time. Since the delay is unavoidable, but predictable, numerical means are used to extrapolate the data to the time of actual digital-to-analog conversion.[22,23] The second error source is in the mechanism for conversion from discrete to continuous form. A sequence of discrete values fed to a DAC results in a "staircase" analog function. The

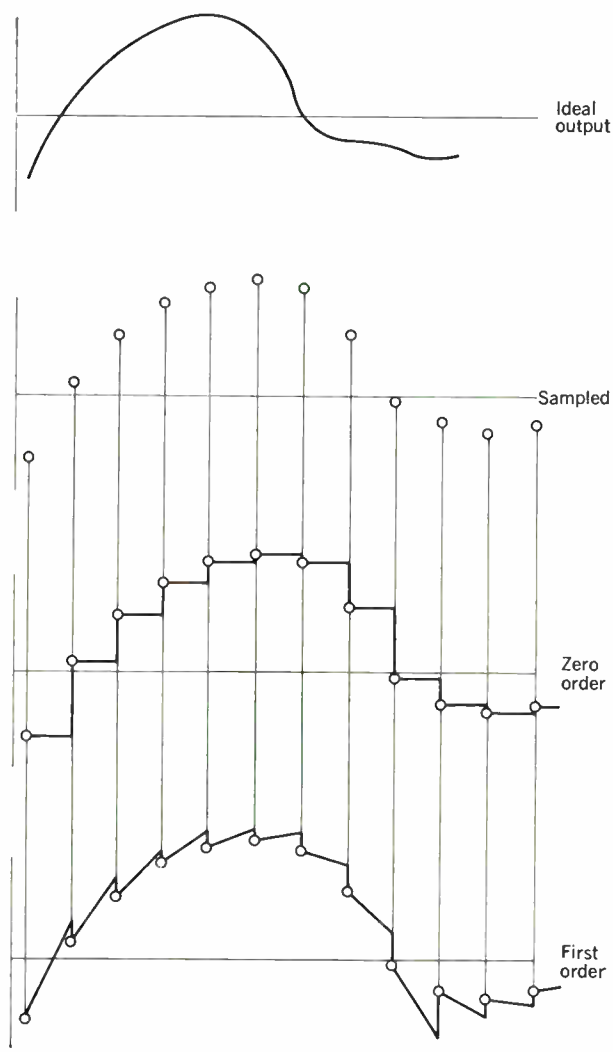


Fig. 8. Discrete to continuous signal conversion requires smoothing, and hence yields only an approximation to the ideal output from sequential computer. Zero order conversion simply holds output voltage at last sampled value until the next arrives. When "steps" are filtered out, the result is shifted in time by a width of one-half step. The first-order scheme uses the past two converted values to predict the value between points, before "resetting."

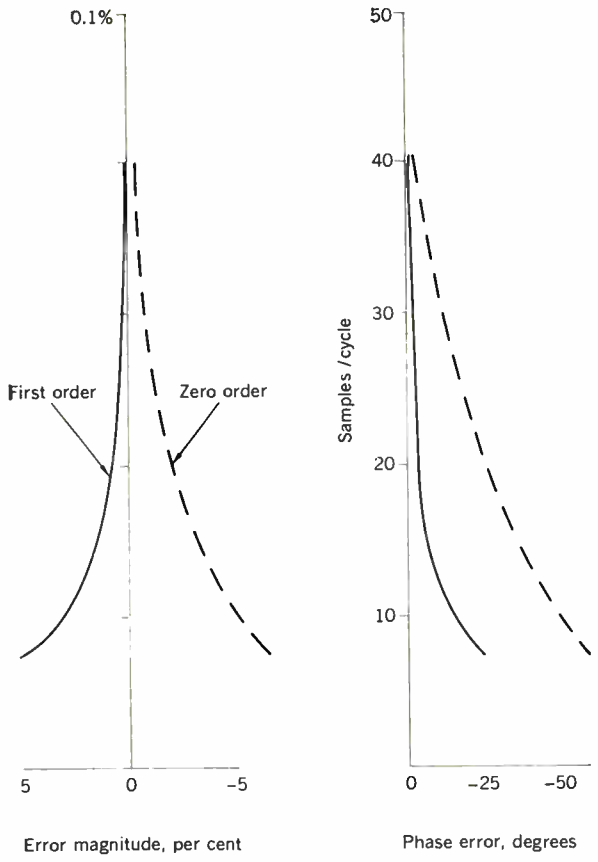


Fig. 9. Zero and first-order conversion methods are compared after filtering. Errors are functions of the number of samples per cycle of full-scale signal frequency. A good rule of thumb for the zero order filter is: for 50 samples per cycle errors are under 0.1 per cent and 3.6 degrees.

desired output is a smooth curve passing through each data point (the left corner of each step). If the staircase is smoothed with an analog circuit, the result is a smooth curve shifted in time ½ Δt, passing through the center of each step. The amplitude of this curve is attenuated from what it should be. This technique of smoothing is called "zero order" filtering (Fig. 8). The size of the errors is a function of the number of samples per cycle. At ten samples per cycle, the magnitude attenuation is about 1.1 per cent and the phase shift is 18 degrees. At 30 samples per cycle, the errors are 0.7 per cent and 6.5 degrees.

A first-order filter may be applied to the DAC output to reduce the errors. For special purposes, extrapolating filter of higher order are feasible; they are programmed from analog components. The first-order filters extrapolate from the last two discrete values to generate intermediate values until the output voltage is reset to the next discrete value from the DAC. The first-order filters have a much improved phase characteristic, but at low sample rates the magnitude is erroneously accentuated. The error characteristics of the two filters are shown in Fig. 9. A good rule of thumb for the simple zero order filter is that for 50 samples per cycle the errors are less than 0.1 per cent and 3.6 degrees. This rule is convenient for estimating the required sample rate and hence the time available for the sequential program. If the variables converted from digital form vary at a maximum frequency of 2 c/s, then 100 samples per second are needed, and a period of 10 ms are available for the sequential program for input–output and calculations.

Dynamic range of dependent and independent variables. The time resolving power of the analog computer has been mentioned in connection with an analogy to the resolution of the human eye. Pursuing this concept further, the time-resolving power of a computer is

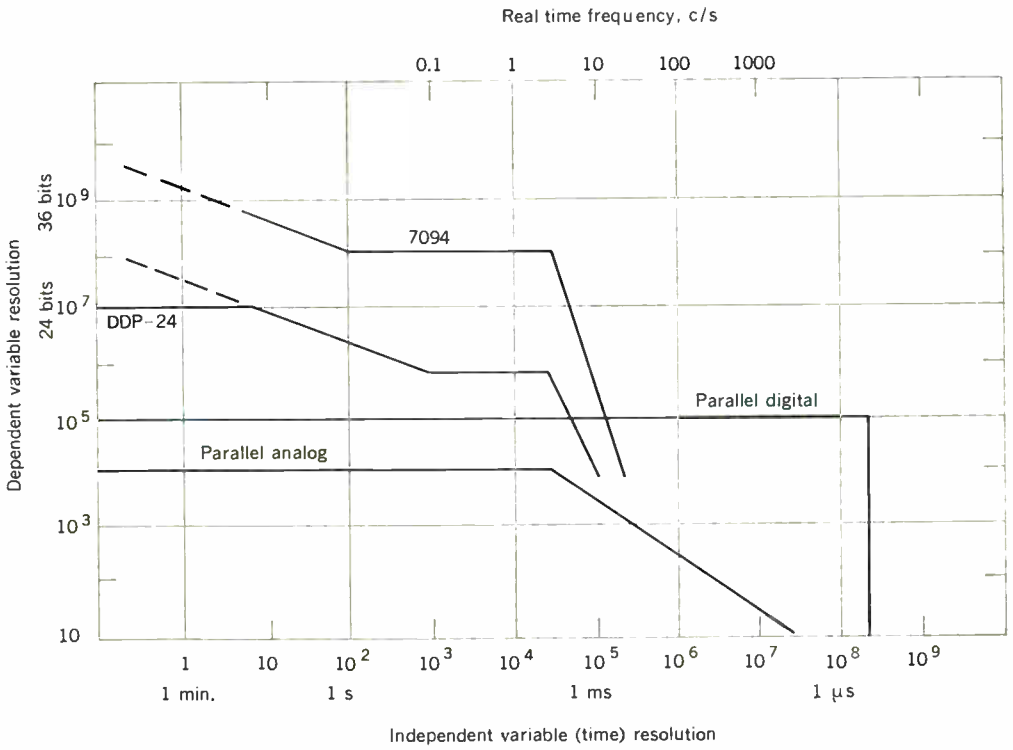


Fig. 10. Certain performance characteristics of computers can be deduced from this comparison of available resolution in computer representation of dependent and independent variables. Notes: (1) limits on the parallel-digital curve are given as those of the EAI Hydac, 0.5 μs in time and 16 bits in magnitude; (2) attenuation on the right-hand end of DDP-24 and 7094 curves corresponds to truncation errors; (3) slope at left of these curves corresponds to reduction of round-off errors at expense of speed, including use of double-precision methods (broken lines); (4) real-time frequency scale refers to signal frequencies passed by sequential program (DDP-24, 7094) assuming a maximum rate of 100 samples per cycle.

measured by the shortest time interval that can be accounted for in a calculation. For all signals in an analog computation, the resolution is directly related to the bandwidth of the components; however, the computer's ability to respond to on–off signals and very short pulses, or to discriminate between two closely spaced events is a closer description of time resolution. In a digital computer, the absolute minimum resolution might be taken as the time to execute three instructions; however, within a hybrid system the resolution of the sequential program is either the sample interval just discussed, or at best, for asynchronous operation, the time for a complete interrupt program to respond to an event. In the parallel computer the time resolution is, of course, much greater because computing elements need not be time-shared. For relay circuits the resolution is about 1 ms; for electronic switching of analog signals, from 10 to 100 μs; and for parallel digital operations, from 0.1 to 10 μs. If these numbers are compared to the total length of a typical computer run, say 1.5 minutes, computer time resolution can be measured by a nondimensional number:

Parallel digital logic operations	$1 : 10^9$
Parallel digital arithmetic operations	$1 : 10^7$
Parallel analog, electronic switching	$1 : 5 \times 10^6$
Sequential program, minimum useful program	$1 : 3 \times 10^5$
Parallel analog, relay switching	$1 : 10^5$
Sequential program, typical sampling	$1 : 5 \times 10^3$

The digital computer is employed in a simulation where the dynamic range of dependent variables requires a wide dynamic range (reciprocal or resolution) in the computation. It is seen that resolution of the independent variable is traded for that of the dependent variables when a particular calculation is moved from analog to digital computer. Fig. 10 shows these functions plotted against each other for different computers. The flat part, or "operating range," of the sequential computer plot is found to be limited on one end by truncation error and the other by round-off error. This is to be interpreted as meaning that, for a given set of mathematics, short cuts and approximations may be used to improve the time resolution up to a point where the truncation error becomes serious. On the other end, special techniques may be used to reduce round-off error, including double precision operations, at the expense of time resolution.

When the digital computer curves are related to the frequency scale, at the top of Fig. 10, a program of a particular size must be considered. For example, at the 2-c/s point, six second-order differential equations for a trajectory simulation could be calculated, because that point corresponds to a 10-ms (bottom scale) sequential program time interval. At the 20-c/s point, very crude integration algorithms and approximations are used for the same problem; or otherwise much smaller calculations are in order. Moving to the left on the curve, more time is available either for more accurate calculation or for computation of more functions. The useful operating ranges for the various techniques are evident in this figure.

Simulation model and programming

Mathematical analysis of the behavior of physical processes and systems is a basic tool for the design engineer, and its use is growing fast. At first, analysis was restricted to the smaller elements in a system, to linearized approximations, or to phenomena that can be isolated from interaction with its environment. For example, there have been many studies of noninteracting servo control loops, heat diffusion in devices of simple geometry, the "small signal behavior" of aircraft and their control systems, and batch and continuous chemical reactors of simplified geometry. Analysis starts from a consideration of the basic laws of physics as applied to the process at hand and proceeds to develop a mathematical model. The solution to the equations of this model for a range of the independent variable(s) constitutes a simulation of the process. The nature of the designer's task and the fact that the analysis has been limited to an element of a more complex system, require that many such solutions be calculated. The simulation is performed many times over to determine the variations in the process behavior with changes in (1) internal design parameters of the process, and (2) environmental conditions. Electronic computers of both types have aided immeasurably in reducing this task to a manageable one. The ease in obtaining simulation results that computers offer the designer and analyst has given a boost to general acceptance of the analytical approach.

Also, the successful correlation of experimental results with analytical predictions has built confidence in these techniques, and led to simulations of more complex systems. A simulation model can be made so large that coping with the variables is difficult. This can happen when poor engineering judgment leads to a model with many more variables than known conditions and assumptions. However, a complex model carefully built up from verified models of subsystems may lead to valuable results attainable by no other means. Thus, as analysis and simulation have yielded understanding of the behavior of small systems, a natural process of escalation has led to simulation of systems of greater and greater complexity. The increased sophistication of simulation models has made the analyst even more dependent upon computers for effective control of the simulation and for interpretation of results.

One might well ask, What are the implications of this escalation of complexity? If simulation models must necessarily grow larger, just how does this affect the procedures of analysis, computer programming, computation, and interpretation of results? How are the hazards, of earlier concern, of becoming overwhelmed with useless data and meaningless computation to be avoided? There is a pernicious theory about programming for a very large digital simulation, that says if two men can do the job in 6 months, four men will take 12, and eight men would never complete it. How can the step from mathematical model to the first computer simulation run be held within bounds—to avoid inordinate investment in programming that may never work or may have to be scrapped for a better approach? How can the analyst or design engineer stay in touch with his model?

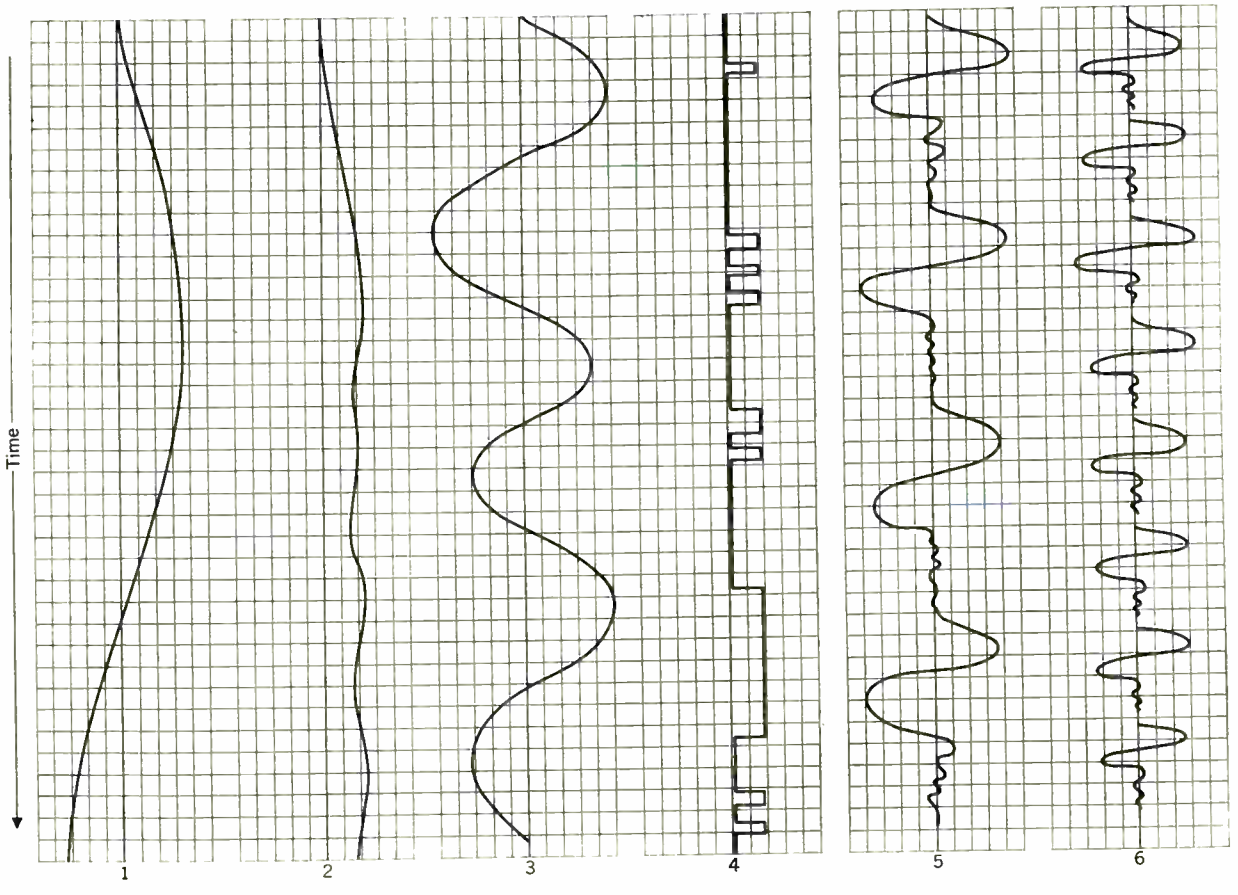
Surely there are no answers that yield guaranteed result. But these are serious questions and some direction is needed in order to evaluate properly the true potential of advanced computer techniques. The implications in the field of hybrid simulation may be divided into three categories: (1) model building in programming, (2) software, and (3) automation.

Model building in programming. The analyst, design engineer, and programmer of a large hybrid simulation must all (if they are more than one person) become involved in all phases of the simulation process. Responsibility cannot be divided, as it often is at the digital "closed-shop" facility, between analyst and computer programmer. The hybrid computer laboratory must be operated, as are many analog laboratories, on an open-shop basis with expert programming support available from the laboratory. The design engineer must have a genuine understanding of the computers to be used, even though he may not do the actual computer programming. Since the computer actually becomes the model of his system, he must know the limitations imposed by the machine as well as by the mathematics, and he must be able to communicate effectively with the computer. Moreover, during the construction of the mathematical model, the analyst must keep in mind the features of the parallel and sequential parts of the hybrid computer in order to achieve a proper partitioning of the mathematical model to suit the computer.

Much attention has been given here to the relative speeds of computation inherent in the different computing techniques. It may be evident at this point that the presence of a very wide range of signal frequencies in a system to be simulated is the one characteristic that most clearly indicates the need for a hybrid computer. As an example, consider the simulation of the long-range flight of a space capsule. In real time the position coordinates probably vary at 0.01 c/s over most of the range and, at most, at 3 c/s during launch and re-entry. At the same time, pitch, roll, and yaw rates and thrust forces may reach 10 c/s or more. Adaptive control functions and control surface may have transient frequencies as high as 50 c/s; and a simulation of reaction jet control forces may require torque pulses as narrow as one millisecond. Because there is little or no damping in an orbital flight, these pulses have a long-term effect, and accuracy in their representation is important. If an on-board predictive computer is included in the simulation, iterative calculations on the analog computer may involve signal frequencies of 100 to 1000 c/s. Thus, this simulation spans a frequency of 10^5 as well as a dynamic range in some dependent variables of 10^5 or 10^6 (Figs. 10,11).

The following observations may be fairly apparent, but in considering division of a problem between computers it is well to note the types of mathematics for which each is best suited. The forté of the digital computer is the solution of algebraic equations. If the equations are explicit, the calculation time is easily determined.

Fig. 11. Wide range of signal frequencies is here suggested, although not to scale (a dynamic range of 10^6 could not be illustrated conveniently. Consider simulation of a long-range flight of space capsule. Curves might represent (left to right): position coordinates varying at 0.01 c/s; deviation from a desired path, 0.1 c/s; pitch, roll, or yaw rate, 1 to 5 c/s; reaction jet control pulses, 1-ms pulses at several hundred pulses per second; thrust forces or control surface transients, 1 to 50 c/s; iterative calculations for trajectories, 100 to 1000 c/s.



Implicit equations often require a variable length of time, and if there are not too many of them they may be readily solved continuously on the analog computer. Numerical integration comes as a by-product of the computer's power in solving algebraic problems. Time is the only penalty. If the high precision is not needed, the integration is better done by the analog computer, since the solution of ordinary differential equations is its strong feature.

Evaluation of arbitrary functions is performed with ease by both computers, as well as by parallel-digital components; however, if speed is not important, and if the data are functions of two or more variables, a digital program is the best choice. Simulation of nonanalytic nonlinear functions, such as limits, backlash, dead zone, striction, and hysteresis again is amenable to both techniques but the analog technique is probably the more economical. Evaluation of trigonometric and hyperbolic functions also can be done both ways and the choice seems to depend on the particular problem. In this case there is a third choice, for there are techniques and equipments for executing these functions by parallel digital components.

Clearly, logic equations that must be evaluated continually with respect to their relation to analog variables must be programmed with parallel logic elements. On the other hand, decision and control functions connected with the occasional evaluation of states of the computer and error signals, and with sequencing various sections of the total system through different modes and states of operation, may require both parallel and sequential operations.

Hybrid simulation requiring solution of partial differential equations (PDE) opens a whole new subject for discussion. Let it just be said that although there is little practical experience in this field, it appears to offer one of the most promising areas of growth for hybrid simulation. The digital computer approach to the solution of PDE is often limited by available computer time—particularly in simulation problems where it is desired to solve the problem many times for various conditions. The analog computer can solve some PDE problems very efficiently but it is often seriously limited by the availability of large amounts of equipment for complex problems. Moreover, only with memory to store complete functions (either in a parallel-digital system or a sequential computer) can certain kinds of boundary value problems be approached. The hybrid computer has the ability to store boundary conditions as well as complete sets of intermediate solutions so that parallel programs can be used for speed, but then be time-shared over again for different parts of the space domain. Challenging problems and promising possibilities face the experimenter in this field.

Let us now return to the implications of the growth of complexity in simulation. The important point in programming is that whoever prepares the computer program must be a model builder and familiar with the system to be simulated. The computer model should be put together from working models of subsystems. Each subsystem, or group thereof, should be verified for correct performance in a linear or simplified mode before being connected to other parts of the model. At each point in the expansion of the model, including the final one, at least one test case of a linear mode of operation should be checked against known or precalculated behavior.

Software. Standard programs and routines for digital computers, of general utility to programmers, known as "software," are in such wide use that the production of software is virtually an industry in its own right. "Automatic programming systems," which make computer programming easy for the noncomputer expert, are responsible for the almost universal acceptance of the digital computer in scientific research and development. A total dependence upon automatic programming, however, has the disadvantage of isolating the problem analyst, and even the programmer himself, from the computer.

In the analog computer field, the reverse situation exists. No comparable "software" usage has grown up. However, here the problem analyst maintains rapport with his computer model during the simulation.

The role that software must play in hybrid simulation of large complex systems is evident. Hybrid software must ease the programmer's burden, and at the same time bring the analyst closer to his model rather than isolating him from it. Hybrid software must include not only coding for the sequential computer but also interconnection diagrams and prewired patch panels for the parallel machine. The following types of software are needed to support growth of hybrid computation to meet the simulation needs of today:

1. *Compilers and assemblers:* These aid in processing data before computation and in processing results for interpretation. Automatic programming systems for hybrid computation may differ from conventional systems in only three ways: (a) running time of the object program is minimized at the expense of compiling time; (b) actual running time for each program statement is precalculated or estimated to aid the programmer with timing of the parallel-sequential interface; and (c) while the programmer's language is "problem-oriented," it must be machine-dependent. The programmer is required to utilize special machine features and to program for control of all interface operations.

2. *Utility library:* In addition to the conventional utility routines for mathematical functions, format conversions, and input-output operations, the hybrid simulation library should expand with routines for specific transfer functions of useful subsystem models that have become standard and are used in larger models, e.g., a typical servo controller. Another type of example is a function generator program for any number of aerodynamic functions. A standard program for a complete aerodynamic vehicle simulation is also feasible.

3. *I/O routines:* Direct on-line control of the computer model by the analyst is needed. In a convenient language, it must be possible to experiment with time scales, parameter values, and even to make substitution of mathematical algorithms (particularly integration algorithms), without any penalty in running time. This means that a complete symbol table and all definitions of parameters must be available to an executive routine that will accept operator instructions to modify a particular problem variable and proceed to calculate changes in all the machine variables that are affected. The executor also permits interrogation of the state (and time history) of any problem variable in engineering units. Upon operator command the model can be modified by changing the linkages between submodels or subroutines.

Automation. One important aspect of hybrid computation, which has been barely mentioned, is the opportunity for automating much of the routine parts of programming and check-out of the analog computer. It is evident that a necessary feature of any hybrid computer system is the mechanization and sequential program control of as many of the manual operations on the analog computer as possible. This includes setting of potentiometers, switches, modes, time scales, recorders, and the selector and readout system. This kind of control is important for some of the software functions previously mentioned. It also makes possible automatic set-up, testing, and diagnosis of machine and program faults. Some interesting diagnostic programming for such a hybrid computer was developed in 1958 to 1960 at General Electric MSVD in Philadelphia.[14,25,26]

A different type of programming automation is offered by the APACHE system developed at Euratom, Ispra, Italy, for the IBM 7090 and PACE analog computers.[27] This is a digital program that translates a mathematical statement into detailed programming instructions for the analog computer. While APACHE is not intended for hybrid computing the appropriateness of such a program should be apparent.

One last important characteristic of hybrid simulation concerns the automation of the model-building process itself. Simulation inherently involves trial and error experimentation. The elements of a model are verified; sensitivity to environment is explored; and variation of performance due to parameter changes are evaluated. When a criterion for optimality can be specified, experiments are made to obtain optimum performance. The sequential computer is perfectly suited to the automation of these procedures. Between calculations the digital computer can evaluate the results, decide upon changes to the model or the data, and implement the changes. At the same time, the analyst can monitor the progress of the simulation and interrupt the automatic process whenever human judgment is required.

Conclusions

The main points developed are:

1. Hybrid computation is built upon the technology of analog and digital computers and is equally dependent upon the programming methods, software, and procedures of problem analysis that have been developed for each.

2. A hybrid computer is a compatible system of parallel computing components, both digital and analog, and a stored-program sequential machine. The hybrid programmer must be constantly aware of the relative timing of computational events in the parallel and sequential parts of the system.

3. There is an ever growing need for simulation of very complex engineering systems. The process of analysis and building of a computer model for evaluation and prediction of behavior is a required step in many large development programs. The hybrid computer offers a means for many such simulations that would be impractical by other means. Hybrid computation is inherently a tool for very complex simulations rather than simple studies.

REFERENCES

1. Greenstein, J. L., "Application of ADDA Verter System in Combined Analog-Digital Computer Operation," AIEE Pacific General Meeting, June 1956.

2. Hartsfield, E., "Timing Considerations in a Combined Simulation System Employing a Serial Digital Computer," *Proc. Symp. Combined Analog Digital Computer Systems*, Philadelphia, Dec. 16–17, 1960.

3. Wilson, A., "Use of Combined Analog-Digital System for Re-entry Vehicle Flight Simulation," *Proc. East. Joint Computer Conf.*, vol. 20, Dec. 1961, pp. 105–113.

4. Connelly, M. E., "Simulation of Aircraft," *Servomechanisms Lab. Rept. 7501-R-1*, M.I.T., Cambridge, Mass., Feb. 15, 1958.

5. Connell, M. E., "Analog-Digital Computers for Real Time Simulation," *Final Rept. ESL-FR-110*, Project DSR 8215, M.I.T., June 1961.

6. Connelly, M. E., "Real-Time Analog-Digital Computation," *IRE Trans. on Electronic Computers*, vol. EC-11, no. 1, Feb. 1962, p. 31.

7. Cox, F. B., and R. C. Lee, "A High-Speed Analog-Digital Computer for Simulation," *Ibid.*, vol. EC-8, no. 2, June 1959, pp. 186–196.

8. Skramstad, H. K., A. A. Ernst, and J. P. Nigro, "An Analog-Digital Simulator for Design and Improvement of Man-Machine Systems," *Proc. East. Joint Computer Conf.*, Dec. 1957, p. 90.

9. Wortzman, D., "Use of a Digital Analog Arithmetic Unit within a Digital Computer," *Ibid.*, Dec. 1960, p. 269.

10. Mitchell, J. M., and S. Ruhman, "The Trice—A High Speed Incremented Computer," 1958 *IRE Nat. Conv. Rec.*, pt. 4, pp. 206–216.

11. Bauer, W. F., and G. P. West, "A System for General Purpose Analog-Digital Computation," *J. Assoc. Computing Mach.*, vol. 4, no. 1, Jan. 1957, p. 12.

12. Baxter, D. C., and J. H. Milsum, "Requirements for a Hybrid Analog-Digital Computer," Mechanical Engineering Rpt. MK-7, National Research Council of Canada, Oct. 1959 Ottawa; and Paper no. 59-A-304, ASME.

13. Burns, A. J., and R. E. Kopp, "A Communication Link between an Analog and a Digital Computer (Data-Link)," *Research Rept. RE-142*, Research Dept., Grumman Aircraft Engineering Corp., Oct. 1960; ASTIA no. AD 244 913.

14. Gelman, H. D., "Evaluation of an Intercept Trajectory Problem Solved on a Combined Analog-Digital System," *Proc. Symp. Combined Analog Digital Computer Systems*, Philadelphia, Dec. 16–17, 1960.

15. Noronha, L., "An Integrated General Purpose Hybrid Computing System," presented at the International Symposium on Analogue and Digital Techniques Applied to Aeronautics, Liége, Belgium, Sept. 1963.

16. Witsenhausen, H., "Hybrid Techniques Applied to Optimization Problems," *Proc. Spring Joint Computer Conf.*, vol. 21, San Francisco, Cal., May 1962.

17. Barnett, R. M., "NASA Ames Hybrid Computer Facilities and their Application to Problems in Aeronautics," Internat'l Symp. on Analogue and Digital Techniques Applied to Aeronautics, *op. cit.*

18. Cameron, W. D., "Determination of Probability Distributions using Hybrid Computer Techniques," Internat'l Symp. on Analogue and Digital Techniques Applied to Aeronautics, *op. cit.*

19. Landauer, J. P., "Simulation of Space Vehicle with Reaction Jet Control System," *EAI Bulletin ALHC 62515*, 1962.

20. Landauer, J. P., "The Simulation of Transport Relay with the Hydac Computing System, *EAI Bulletin ALHC 63011.*

21. Witsenhausen, H., "Hybrid Simulation of a Tubular Reactor," *EAI Bulletin ALHC 6252-25*, 1962.

22. Gelman, R., "Corrected Inputs—A Method for Improving Hybrid Simulation," *Proc. Fall Joint Computer Conf.*, vol. 24, Las Vegas, Nev., Nov. 1963.

23. Miura, T., and J. Iwata, "Effects of Digital Execution Time in a Hybrid Computer," *Ibid.*

24. Susskind, A. K., "*Notes on Analog-Digital Conversion Techniques*," Cambridge, Mass.: Technology Press, M.I.T., 1957.

25. Feucht, K., "Diagnostic Programs for a Combined Analog-Digital System," *Proc. Symp. Combined Analog Digital Computer Systems*, Philadelphia, Dec. 16–17, 1960.

26. Paskman, M., and J. Heid, "Combined Analog-Digital Computer System," *Ibid.*

27. Deeroux, A., C. Green, and H. D'Hoop, "APACHE—A Breakthrough in Analog Computing," *IRE Trans. on Electronic Computers*, vol. EC-11, no. 5, Oct. 1962.

(Based on a paper presented at the Spring Joint Computer Conference, Washington, D. C., Apr. 21–23, 1964. The original paper is copyrighted by SJCC and appears in the *Proceedings* of that conference.)

The genesis of an editorial policy

With the merger of AIEE and IRE in 1963, a reappraisal of existing editorial policy was undertaken which resulted in the birth of a new journal and a serviceable publications policy for the voice of IEEE

J. D. Ryder *Michigan State University*

At the time of the merger of the American Institute of Electrical Engineers and the Institute of Radio Engineers the two Institutes were producing 39 separate publications, which included the *Proceedings of the IRE; Electrical Engineering;* the three AIEE Bimonthlies—*Power Apparatus and Systems, Applications and Industry*, and *Communication and Electronics;* the annual bound volumes of these, entitled the *AIEE Transactions;* the 30 *Professional Group Transactions of the IRE;* the *IRE Convention Record;* and *EE Digest* and the *IRE Student Journal* for the student members of the two organizations. The circulation of these various journals ran from a few hundred to over 100 000, and among them all but the *Proceedings* were subsidized in some way.

The Merger Committee recognized that simply adding together the publication programs of the two Institutes would not produce an ideal economic situation or even

a good technical one, and in October 1962, the Committee appointed an Editorial Policy Committee composed of J. M. Pettit, T. F. Jones, D. B. Sinclair, S. W. Herwald, W. A. Lewis, W. K. MacAdam, C. T. Hatcher, and the late L. F. Hickernell; and the author of this paper as Chairman. Later, Hendley Blackmon and B. M. Oliver were invited to represent the interests of the Technical Operations Committee and the Professional Technical Groups. During the period from October 1962 through March 1963, this group held frequent meetings to discuss the problems of amalgamating the publication programs of the two Institutes.

The Committee's work began with a review of the status of the existing publications, their circulations, their editorial and technical coverage, and the editorial and publication costs insofar as these could be determined from existing records. A review was undertaken of the editorial policies followed and of the advertising potential that the several publications appeared to exhibit. A general declining trend in the advertising incomes as of that time was noted. Early in 1963 it was proposed that the services of an outside consultant be secured to provide advice to the Committee from the field of professional publications. Employment of such a consultant was approved by the Executive Committee of the new IEEE, and over a period of several months Mr. Ralph Flynn brought to bear his experience in the technical publications field with particular attention given to securing opinions of leading advertisers toward past policies of the two Institutes, and toward their publications.

Publications—the voice of IEEE

In the Committee discussions during this period it was pointed out that because the IEEE was neither the IRE nor the AIEE, the publications policy of the new Institute should solidly indicate its identity as a new Institute. It was apparent that the membership interests of the IEEE were oriented more toward management operations and application in the electrical engineering field than had been the membership of the IRE, and also that the interests of the new membership were directed more toward development, research, and the engineering-science interface than had been the interests of the membership of the AIEE.

In terms of technical policy and concerning the editorial content, it was obvious that plans for the ultimate group of publications had to be based on the need for (1) rapid publication of research results and (2) a thoughtful program of educational and tutorial papers that would interpret the results to those applying and using the research information.

The size of the membership, together with its technical and geographical spread, indicated that the Institute's publications would represent its major service to that membership. The Committee emphasized that it is, today, possible to serve only very small percentages of our membership through their participation in meetings and discussions, and that rapid publication of papers, resulting in prompt dissemination of information to the membership has replaced many of the functions once associated with the meetings of technical societies. Also it seemed apparent, in a world moving at its present technological rate, that publications were the only reasonable channel open to the Institute through which to advance the abilities of the members as the field advanced. To carry out educational and tutorial objectives, it was apparent that the technical level of the several publications would have to be above the median level of the abilities of the members. For editorial purposes, it was proposed that this median level should be slightly above the technical level to be expected of the holder of a B.S. degree.

Advertising considerations

The consultant's report on his discussions with leading advertisers pointed to the need for clarifying the distinctions between reader groups. It was emphasized that the consistent advertiser requires clearly defined, identifiable reader groups, and that this essential was found lacking in the two previous core journals *Proceedings* and *Electrical Engineering*.

Proceedings had not developed a clear image of the reader in the minds of the advertisers, chiefly because of its highly technical character. (However, it was noted that *Proceedings* had succeeded as an advertising medium because advertisers recognized its very wide readership and because of its prestige as the most often cited journal in the electronics and radio field.)

Electrical Engineering, had developed a power image; in addition, some concern was expressed that *Electrical Engineering* had not been serving its readership to the best advantage in technical depth.

The ideal situation for advertisers was described as containing a medium with an editorial program that would preselect a clearly defined reader audience with easily visualized fields of professional activity or reader interest. On this ideal was based the definition of purpose for a new publications program. Though formulated on the basis of economic requirements, it was also considered a very desirable one for the membership.

Following several months of study and discussion, six possible courses of action for a new publications program were developed and considered. They are presented here along with the conclusions that led in one way or another to their ultimate elimination:

1. To continue both the *Proceedings* and *Electrical Engineering*. The Committee rejected this proposal since it was thought that the confusion incident to the merger would be perpetuated among both members and advertisers, and would give the impression that we were two units of a merged society, rather than a new society. For the long-range good of the merged society, and of the field of electrical engineering, the need was felt for one common magazine for the entire IEEE membership.

2. To continue *Proceedings* only. Concern was expressed that the research orientation of this journal in the past might limit its service to the full membership, and there was considerable doubt that it would be able to serve without subsidy if its editorial coverage and circulation were expanded to cover the entire field of electrical engineering and the entire membership.

3. To continue *Electrical Engineering* only. This proposal was subject to the same criticism that applied to (2) concerning orientation; in this case the identity of the journal was associated with the power area. Whether it could attract enough advertising was also questionable. To discontinue *Proceedings* (the one unsub-

sidized journal) and to continue *Electrical Engineering* would be to sacrifice sources of income unnecessarily.

4. To publish a "split-run" journal, one portion of which would carry electronics content and advertising, and the other, power-directed material—the individual member to choose which of the two editions he desired to receive. Rejection of this plan was based on cost and the difficulties to be expected in differentiating between the electronics and power industries; and on the probable confusion that such an involved arrangement would create among both members and advertisers.

5. To create a single new publication to serve as a common denominator among the total membership. It was judged that by attempting to serve both the entire technical field and the entire functional field such a publication would not really serve the membership, would be so bulky as to be almost unmanageable, and would not identify the audience to the advertiser.

6. To introduce a new publication and to continue *Electrical Engineering* and *Proceedings*. This proposal was rejected on the basis of cost, lack of available material to maintain three general publications, and difficulty in identifying areas of reader interest for the several publications.

After considerable discussion, a compromise between proposals 5 and 6 was arrived at that received general approval. This program provided for a new periodical to serve as a core or general publication for the Institute, and for the continuation of *Proceedings* as a broad research and development-oriented journal. The advantages of this solution seemed to be that it would (1) ultimately weld the society into one group through the services of the core publication, (2) satisfactorily identify the reader audiences, and (3) set the stage for developing, through the tutorial function, a warm and personal touch. At the same time, the orientation and prestige of *Proceedings* would be retained. Also, this plan seemed to offer easier readership identification in the case of the special role that *Proceedings* plays in research and new development in all electrical fields.

The consultant's opinion that it would be more difficult to sell advertising on the basis of a new concept of an old name than of a new publication with a new name influenced the choice of a name for the new magazine. First, the Editorial Policy Committee proposed "Electrical and Electronics Engineering," which seemed to indicate appropriately the name of the society. Later research showed that there were already in existence 33 publications with quite similar names. As a result, during its consideration of the report of the Editorial Policy Committee in Toronto in June, the Executive Committee recommended the title "*IEEE Spectrum*," as one that would avoid confusion and distinctly identify the new publication.

The designation "spectrum" was deemed appropriate because of the intended coverage of all frequencies of the electromagnetic spectrum, and because of the implications it provided for broad coverage of both technical and professional functions. It is interesting to recall that the name *Spectrum* was originally proposed in 1962 by the Board of Consultants of the *IRE Student Quarterly*, as a potential name for that journal.

Following this comprehensive study of the publications program of the former Institutes, made possible through its own thoroughgoing research and extensive consideration of alternatives, the Editorial Policy Committee proceeded to prepare a report and recommendation which were presented to the Board of Directors at its Toronto meeting on June 19, 1963. Briefly summarized, this report recommended, first, that the IEEE publish one general periodical, "*IEEE Spectrum*," which should (1) perform a tutorial and educational function for the membership at large; and (2) include news of the IEEE and of IEEE people, news of the profession and of education, and letters to the Editor on topics of general interest.

Second, it was recommended that *Proceedings* continue on a subscription basis as a research and new-development publication, carrying the results of important current research and engineering developments; that it continue its practice of publishing special issues, and that it retain its "quick reaction function"—in the Correspondence section, which is devoted to technical topics, and which had so successfully served as a medium for the speedy dissemination of current ideas. The special value of this section is contained in placing material before the readership without delays of reviewing and extensive editing that are necessary in the processing of formal papers.

The Committee also recommended that, as general periodicals, *Proceedings* and *IEEE Spectrum* should not be separated into technical fields of interest, thus definitely to indicate to the engineering profession and the advertising fraternity that electrical engineering was indeed one technical field.

The Board of Directors, after thorough discussion, adopted the report, including the new journal's name, *IEEE Spectrum*. The Editorial Policy Committee, as an *ad hoc* body, was disbanded, and the Board of Directors proceeded to the appointment of the Editorial Board as required by the Institute bylaws. This Editorial Board for 1963 included J. D. Ryder as Editor, Donald Sinclair, T. F. Jones, S. W. Herwald, C. T. Hatcher, W. K. MacAdam, and A. H. Waynick, with E. K. Gannett ex officio as Managing Editor. T. A. Hunter, Editor of the *IEEE Student Journal*, was later appointed as another ex officio member. In the discussion by the Board of Directors it was pointed out that the Merger Agreement had called for a decision on the future of the two major magazines of the former Institutes before the end of 1963. Therefore, because advertising is sold many months in advance of publication, it was necessary that decisions on editorial policy and publication programs be arrived at about the middle of the year. Because of the time element a tactical report by the Editorial Policy Committee was indicated, rather than one covering broad strategy, and the new Editorial Board was instructed to proceed with further development of basic publication recommendations which could be presented to the Board of Directors at a later date.

The Editorial Board then undertook an extensive series of meetings looking to the development of a statement of basic publication policy. In this they were aided by correspondence with Editor Emeritus A. N. Goldsmith, and by a group from the Board of Directors headed by Director F. Karl Willenbrock.

These groups strongly urged the necessity for rapid dissemination of research results and new technological developments. They recommended that we recognize the necessity for publication of research results at high level

so that such results would be available in the archives for later application.

It was agreed that the publication of high-quality papers, oriented toward research and development, and written by leading specialists in the field, was a major responsibility of any leading technical society.

A further statement of publications policy was presented by the Editor to the Board of Directors at a special meeting of that body in Chicago on October 31. After a thorough review and discussion of all facets of the proposal and of Institute publications, the Board of Directors adopted the recommendations of the Editorial Board by a unanimous vote.

Editorial policy statement

For the information of the members of the IEEE the two policy statements adopted by the Board of Directors on June 21 and October 31, 1963, have been amalgamated into one statement, which follows:

In order to advance through publications the theory and practice of electrical engineering and electronics, the allied branches of engineering, and the related arts and sciences, the Editorial Board recommends the following as basic principles of publication policy for the Institute of Electrical and Electronics Engineers:

1. The paramount objective of the IEEE through its publications shall be to encourage the preparation and dissemination of scientific and professional information, thereby to advance the profession and to educate the members thereof, as required by the continuing technical progress of engineering and its associated sciences.

2. The publications of the IEEE shall maintain a standard of technical excellence consistent with the position of the IEEE as a leader in its field, and thereby act as a positive force in developing the technical and professional capabilities of the members.

3. It is recognized that electrical engineering is one broad field and publications of the Institute that are designed to reach all members shall cover that field broadly. Material limited in interest to a portion of the membership only shall appear in publications of the Institute that are voluntarily subscribed to by those members.

4. The technical material of the field shall be presented in a minimum number of separate publications consistent with adequate service to the members and consonant with the value of the material presented for publication.

5. A new format shall be employed in a core publication or magazine for all members: to contain review and tutorial papers, occasional theoretical papers of broad and fundamental import, papers of application and economic significance, news of the profession and of the Institute, announcements of and reports of conferences and conventions, news of education, letters to the Editor on topics of general concern as well as of technical interest, news of scientific and engineering advancement, news of political and social interest to the profession, and abstracts of or references to material in other IEEE publications. The technical level of this core publication shall be a positive force in upgrading the median level of the membership ability. This core publication is to be known as *IEEE Spectrum.*

6. A second publication to be known as the *Proceedings of the IEEE* shall be devoted to research-oriented papers, shall reflect the introduction of new science in the field, and shall be of interest to broad segments of the membership. It shall report on significant research on all aspects of electrical and electronics engineering and related sciences, shall include appropriate letters to the Editor and news of scientific and engineering advancement. It shall provide for broad coverage of important fields as they emerge through the use of "Special Issues" whereby those fields can be provided with adequate reference background and a foundation laid for the development of the abilities of the members of the profession in those fields. This publication shall be available to members and others on a subscription basis which shall insure an adequate circulation to serve the needs of the profession.

7. The Bimonthlies and the *PTG Transactions* and the *Student Journal* shall be left under study and review with the view to reducing the number of separate publications in so far as seems consistent with serving the needs of the membership.

8. Papers resulting from certain conventions and technical conferences may be combined in special publications approved by the Editorial Board.

9. The contents of the several publications, with the possible exception of those resulting from paragraph eight above, shall be subject to review procedures which shall assure appraisal on the basis of technical excellence and value to the technical advancement of the membership of the field. These review procedures shall be reported to the Editorial Board at least once a year.

10. Both *IEEE Spectrum* and *Proceedings* shall carry advertising. Certain of the *Transactions* may be permitted to carry advertising under policies to be developed by the Editorial Board and approval by the Board of Directors. The frequency of publication of *IEEE Spectrum* and *Proceedings* shall be monthly.

To implement the above policies, the Editorial Board shall:

a. Make every effort to enhance the scholarly reputation of the publications of the Institute so that leading research workers will wish to publish their best papers therein.

b. Coordinate all publication activities of the Institute in accordance with IEEE policies in order to provide balance and coherence, to assure adherence to those standards of quality and good practice that will reflect to the credit of the Institute, and to permit logical assignment and utilization of funds budgeted to publication purposes.

c. Be responsible for the establishment of an appropriate review procedure for all material published by the IEEE. These procedures shall be designed to insure that reviewers are chosen from among recognized experts in the appropriate portions of the field.

d. Insure expeditious publication of important research results and technical progress in the field in a form as complete as is consistent with the interests of the members to be reached. In general, prior verbal presentation delays publication of a technical paper, and should be required only after careful consideration of all factors involved.

The Editor acknowledges the aid of Dr. F. Karl Willenbrock in the preparation of this report.

Authors

Patrick E. Haggerty (F), president of Texas Instruments Inc., holds the B.S.E.E. degree from Marquette University and honorary degrees from Marquette, St. Mary's University of San Antonio, and the Polytechnic Institute of Brooklyn. He joined the predecessor company to Texas Instruments in 1945, following employment in Milwaukee with the Badger Carton Company and three years of U.S. Navy assignment to the Bureau of Aeronautics in Washington, D.C., where he served as head of the Electronics Production Branch of the Electronics Products Group, which was responsible for procurement and production of naval airborne electronic equipment. When he joined Texas Instruments he was put in charge of the development of the research, engineering, and manufacturing phases of the company's operations. Since that time the operations of the company have been diversified and tremendously expanded. Mr. Haggerty became executive vice president of Texas Instruments in 1950 and was elected to his present office in 1958.

He served as president of the IRE and as co-chairman of the IRE-AIEE Merger Committee in 1962, and continues to serve on the resultant IEEE Board of Directors. He also is a member of the Defense Science Board.

C. Lester Hogan (F) received the B.S. degree in chemical engineering from Montana State College in 1942. He pursued graduate studies at Lehigh University, from which he received both the M.S. and Ph.D. degrees in physics, with a major in solid-state and electromagnetic theory. In 1954 he received the honorary A.M. degree from Harvard University. As a member of the technical staff at Bell Telephone Laboratories, Dr. Hogan performed experiments that demonstrated nonreciprocity at microwave frequencies, and also carried out a theoretical analysis substantiating the experiments. This work was published in 1952 and has become a classic reference. While at Harvard, he was awarded the Gordon McKay Professorship of Applied Physics. He also led the Symposium on Microwave Properties and Applications of Ferrites. He joined Motorola in 1958 and is now vice president of Motorola, Inc., and general manager of the Semiconductor Products Division.

Dr. Hogan has held important positions in many professional organizations. He serves as a member of the Advisory Council of the Department of Electrical Engineering of Princeton University and is on the Standing Committee on Fundamental Aspects of Material Research of the Materials Advisory Board. He is an active member of several technical societies.

Robert N. Noyce (M) was one of the founders of Fairchild Semiconductor, where he established the research department and directed the initial development of the silicon mesa and planar transistor lines. He was subsequently vice president and general manager of what was then Fairchild Semiconductor Corporation. Since 1962, when the company became a division of Fairchild Camera and Instrument Corporation, he has served as vice president of FCI and general manager of Fairchild Semiconductor Division.

He received the B.A. degree from Grinnell College in 1949, with a double major in physics and mathematics. He received the Ph.D. degree in physical electronics from the Massachusetts Institute of Technology in 1953. As a member of the staff of the Research Division of Philco Corporation, he led a solid-state physics group who worked on the development of germanium and silicon high-frequency transistors. He left Philco in 1956 to join Shockley Semiconductor Laboratory, where he directed the design and development of diffused silicon transistors and investigated the basic recombination process in semiconductors. He remained with Shockley until the formation of Fairchild Semiconductor.

Dr. Noyce holds 12 patents on semiconductor methods, devices, and structures. He is a member of the American Physical Society and Phi Beta Kappa.

Leonard C. Maier (M) received the B.A. degree in physics from Williams College in 1944, and the M.S. and Ph.D degrees, also in physics, from the Massachusetts Institute of Technology in 1948 and 1949, respectively. From 1944 to 1946 he was on active duty with the U.S. Navy, engaged in electronics maintenance in the Pacific Theatre. From 1949 to 1950 he was associated with the Behr Manning Corporation, Troy, N.Y., as research physicist. He joined the General Electric Company in 1950, and for several years held various engineering and supervisory engineering positions in research and development of military and commercial electronic products in the Electronics Laboratory and the Heavy Military Equipment Department. From 1954 to 1959 he was with General Electric's Cathode Ray Tube Department, where he was manager of engineering, and later manager of the industrial and military tube operation. After a period as vice president of defense product marketing with the Crosley Division, Avco Corporation, he joined GE's Semiconductor Products Department, where he is now general manager.

J. E. Brown (F) was born in Greenport, N.Y., on September 11, 1902. He studied electrical engineering at Cornell University. From 1924 through 1936 he was associated with the Radio Division of the United States Department of Commerce, the Federal Radio Commission, and the Federal Communications Commission. He joined the staff of the Zenith Radio Corporation in 1937, where in 1940 he became chief engineer and in 1943 he was named assistant vice president. He was elected to his present position, vice president in charge of engineering, in June 1958.

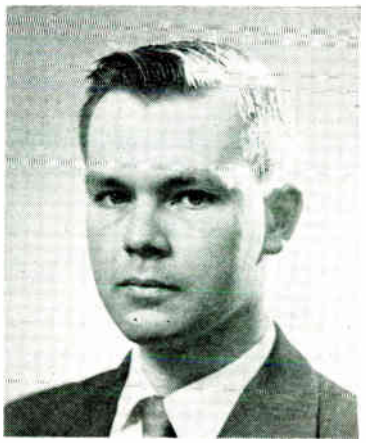
Mr. Brown has served on the IRE Board of Directors and many IRE committees through the years. In 1947 he was IRE representative to the Army Signal Association. He has also served as chairman of panels of the Radio Technical Planning Board and the National Television Systems Committee. He is a member and past president of the Radio Engineers Club of Chicago.

C. Harry Knowles (SM) is general manager of the Molecular Electronics Division of the Westinghouse Electric Corporation, at Elkridge, Md. Prior to joining Westinghouse, he was assistant general manager of research and advanced development at the Motorola Semiconductor Products Division, where he worked on the development and exploratory production of video vhf and uhf transistors. He was with Bell Telephone Laboratories from 1953 through 1957, where he supervised the development of high-frequency transistors. He was also responsible for the Bell Laboratories production engineering in high-frequency transistors at the Western Electric Company in Laureldale, Pa.

Mr. Knowles received the bachelor's degree from Auburn University and the master's degree from Vanderbilt University.

A. B. Phillips (SM), manager of integrated circuits operation at the Motorola Semiconductor Products Division, is responsible for the engineering and production of all hybrid and monolithic integrated circuits for both digital and amplifier applications. Prior to his present position, he was assistant to the vice president and general manager of the division. He formerly worked in engineering development, production, and marketing for both Motorola and the General Electric Company. As manager of mesa engineering and manager of mesa production at Motorola, his duties included the engineering development and production of germanium and silicon mesa, planar, and epitaxial transistors for high-frequency switching and amplifier applications.

He received the B.E.E. degree from the Polytechnic Institute of Brooklyn in 1950, and has taken graduate work toward the M.B.A. degree at Syracuse University. He is the author of a textbook, *Transistor Engineering and Introduction to Integrated Semiconductor Circuits*, the first volume of the "McGraw-Hill Series in Solid-State Engineering." He is also consulting editor and editorial advisor for this series. Mr. Phillips is a member of Tau Beta Pi, Eta Kappa Nu, and Sigma Xi.



Alan B. Fowler was born in Denver, Colo., on October 15, 1928. He received the B.S. and M.S. degrees in physics from Rensselaer Polytechnic Institute in 1951 and 1952, respectively, and the Ph.D. degree in applied physics in 1958 from Harvard University, where he was engaged in research on germanium surfaces. While serving in the U.S. Army, he was a member of the former Thermionics Branch of the Signal Corps Engineering Laboratory and worked on semiconductor devices. From 1953 until 1956, while attending Harvard University, he was employed by the Research Division of Raytheon Manufacturing Company, where he worked on silicon diodes and transistors and on vacuum ultramicrobalance studies of absorption on germanium.

Since joining the IBM Research Division in 1958, Dr. Fowler has been engaged in research on photoconductors, optical properties of heavily doped germanium, and gallium arsenide lasers. He is at present in charge of a group studying thin-film transistors at the Thomas J. Watson Research Center. Dr. Fowler is a member of the American Physical Society and Sigma Xi.

J. S. Frame A biographical sketch of Dr. Frame appears on page 239 of the March issue.

W. L. Everitt (F) holds degrees from Cornell University, the University of Michigan, and the Ohio State University, and has taught at all three of these institutions. During World War II he was director of the Operational Research Staff of the Chief Signal Officer of the U.S. Army. Soon after the war he became head of the Department of Electrical Engineering at the University of Illinois, where he has held the post of dean of the College of Engineering since 1949. He is past president of IRE and the American Society for Engineering Education, and was president of the Engineers' Council for Professional Development from 1958 to 1961.

Dr. Everitt is the recipient of several professional awards, including the IRE Medal of Honor in 1954, the American Society for Engineering Education's Lamme Medal in 1957, the AIEE Medal for Electrical Engineering Education in 1957, and the IEEE Mervin J. Kelly Telecommunications Medal in 1963. He has served on a number of Government boards, committees, and panels, and as a director and consultant for several industrial organizations. He is an eminent member of Eta Kappa Nu, and is currently chairman of the Commission for Engineering Education.

L. M. K. Boelter was appointed dean of the College of Engineering at the University of California, Los Angeles, in 1944, and also serves as professor of engineering at UCLA, as visiting professor of engineering education at Purdue University, and as agricultural engineer at the Experiment Station in Davis, Calif. He received the B.S. degree in mechanics in 1917 and the M.S. degree in electrical engineering, both from the University of California, Berkeley. He became instructor in electrical engineering in 1919, assistant professor of experimental engineering in 1923, associate professor of mechanical engineering in 1927, professor of mechanical engineering in 1934, and associate dean of the College of Engineering in 1943, all at Berkeley. He has published several scientific articles on heat and mass transfer, thermodynamics, production, traffic, transportation, illumination, theory of planning, and engineering education. Among the honors he has received are the ASEE Lamme Medal in 1956 and the ASME Medal in 1957. In 1962, ASME and AIChE presented him with the Max Jakob Award in Heat Transfer. He is a Fellow of ASME and IES, and a member of ASHRAE, ASEE, AIChE, SAE, and the American Chemical Society.

R. R. O'Neill is professor of engineering and associate dean of engineering at the University of California, Los Angeles. He received the B.S. and M.S. degrees in mechanical engineering from the University of California, Berkeley, and the Ph.D. from UCLA. He joined the staff of UCLA in 1946 and seved as assistant head of engineering and then as assistant dean for graduate studies in engineering before being appointed to his present position in 1961. At present he is also the coordinator of the Engineering Executive Program, a two-year program of part-time study leading to the degree of master of engineering.

Prior to 1946 he was associated with the AiResearch Manufacturing Company in Los Angeles, as design engineer; with the Dow Chemical Company in Midland, Mich., as design and development engineer; and with Dowell, Inc., also in Midland, as engineer.

He is a member of several honorary and technical societies, including Sigma Xi, Tau Beta Pi, ASME, the American Society for Engineering Education, and the American Materials Handling Society, and is a licensed mechanical engineer in the State of California.

George E. Moore (SM) has held the position of director of continuing professional studies at the Polytechnic Institute of Brooklyn since 1963. Programs initiated to date include the Modern Engineering Series and the Executive Technical Development Program, as well as the Advanced Management Program planned for the Fall of 1964. He received the B.S.E.E. and M.S.E.E. degrees from the University of Pittsburgh, and was also enrolled in Harvard University's Middle Management Program. He subsequently was engaged in engineering research at the Electronics Research Laboratory in Pittsburgh. After a five-year period as assistant professor of electrical engineering at the University of Pittsburgh, he joined Westinghouse Electric Corporation, where he was manager of university relations from 1952 to 1956 and manager of graduate student training from 1956 to 1963.

Mr. Moore is a member of ASEE and is a registered professional engineer. He is a member and past chairman of the IEEE Professional Technical Group on Education and secretary of the ECPD Committee for the Development of Young Engineers. He is also a member of Sigma Tau and Eta Kappa Nu.

Edgar C. Gentle, Jr. (SM) is data communications planning administrator for the American Telephone & Telegraph Company, New York City. He is responsible for market planning of services, systems, and uses of data communications in the Bell System services. He also conducts the Bell System Business Communications Seminar at Chicago. This program is designed to acquaint business executives with the role of data communications in planning and implementation of business information systems. He received the B.E.E. degree from Auburn University in 1942. Prior to his present assignment he was successively dean of engineering, dean of marketing, and director of the Bell System Communications Training Program at Cooperstown, N.Y. He formerly was with Southern Bell Telephone & Telegraph Company, serving as chief engineer and later general plant manager for the Georgia area. He had previously held the position of transmission engineer for the Southern Bell Company in its nine-state operation. During this time he helped in the planning and implementation of the first of the communication engineering courses in the Bell System at Clemson College.

D. G. Ebeling is consultant, materials engineering, in General Electric Company's Engineering Services. He received the B.S. degree in 1940, the M.S. degree in 1948, and the Ph.D. degree in metallurgy in 1950, all from Rensselaer Polytechnic Institute. He worked one year with the U.S. Steel Corporation, and then served five years in the U.S. Navy as ordnance specialist, attaining the rank of Lieutenant Commander. He joined General Electric in 1946 and held positions in the Chemical Department (later Metallurgical Products Department), Research Laboratory, Large Steam Turbine-Generator Department, and Knolls Atomic Power Department prior to his present assignment. During this period he was concerned with the development of magnetic materials, high-temperature alloys, large forgings and castings, and atomic power plants. In addition to his regular duties, he is actively engaged in the development of the company's Modern Engineering course for engineering managers. He holds patents in the fields of magnetic and heat resisting alloys.

Harold K. Work received the A.B. and professional engineering degrees from Columbia University and the Ph.D. degree from the University of Pittsburgh. He has been director of the Engineering Research Division of New York University since 1949. In addition, he is associate dean of the School of Engineering and Science. Prior to his association with New York University, he was associated with The Mellon Institute as a research fellow, with the Aluminum Company of America as a chemical engineer and as a division head in the Aluminum Research Laboratories, and with Jones and Laughlin Steel Corporation as manager of research and development and as director of research.

At present he is also director of the Engineering Foundation of United Engineering Trustees, Inc. He is a member of the board of directors of the University Corporation for Atmospheric Research and a member of the Conference on Administration of Research. He is active in an EJC-Engineering Foundation committee to establish a National Academy of Engineering.

Thomas D. Truitt (M) received the B.S. degree in mathematics from Texas College of Arts and Industries and the M.S.E. degree from Princeton University. As a graduate student at Princeton and as a faculty member at Texas A & I and the Moore School of Electrical Engineering, University of Pennsylvania, he taught mathematics and electrical engineering courses and gave numerous lectures on various aspects of analog and digital computers. He joined Electronic Associates, Inc., in 1956, and since then has performed and directed extensive studies of digital and analog computer organization and techniques, including automatic programming, parallel-parallel digital computer structure, advanced incremental computing techniques, special purpose digital computer organization, and combined analog-digital computer systems. As head of the Advanced Study Group, he is concerned with the conception and development of advanced computing systems.

J. D. Ryder (F) received the B.E.E. degree in 1928 and the M.S. degree in 1929 from Ohio State University, and the Ph.D. in electrical engineering from Iowa State University in 1944. From 1929 to 1931, he worked with General Electric Company in vacuum and gas tube development. He then joined Bailey Meter Company as supervisor of the electrical and electronics section of its research laboratory.

Dr. Ryder began his teaching career in 1941 as an assistant professor of electrical engineering at Iowa State University. In 1949 he was appointed head of the Department of Electrical Engineering at the University of Illinois, and in 1954 he became dean of the College of Engineering at Michigan State University. He is Editor of IEEE, a Fellow of the American Association for the Advancement of Science, and a member of the Michigan Society of Professional Engineers, the Michigan Engineering Society, and the American Society for Engineering Education. He is the author of four textbooks: *Electronic Engineering Principles; Networks, Lines and Fields; Electronic Fundamentals and Applications*; *and Engineering Electronics with Industrial Applications and Control*.

How to Use Reader Service Card!

IEEE offers you a new service . . .

so that you may easily request further information about products advertised in Spectrum. The postage paid Reader Service Card enclosed in this issue provides space for you to request data from eight firms.

PLEASE FOLLOW THESE STEPS!

A. First enter on the card the specific item in which you are interested, as several items may be described by the advertiser. He needs to know specifically which one interests you.

B. Then name the advertiser so that we may know to whom to send your inquiry. This step is essential.

C. Now, please enter the page number of the advertisement. This is a safeguard to help us avoid mistakes.

TO TIME THIS SERVICE!

The advertiser wants to know how urgently you need this information. The third column of each item line is a block into which you should enter the number 1, 2 or 3 to help the advertiser judge the kind of follow-up to give you, thus:

"1" Indicates that you need the product in question immediately. Advertiser is to rush information even if he cannot send a finished type of literature.

"2" Indicates that you think you will need the product in 3 months. This means prompt follow-up is necessary, but can wait a few days for detailed dictation or finished literature.

"3" Indicates you want the information for your file. Thus, you would prefer to receive very complete data rather than a rushed response.

You will get the best service if you provide very specific information and fill out the card completely.

Don't forget your name and address!

Correspondence

Errata. In my article in the May issue of SPECTRUM, "Protecting Communication Systems from EMP Effects of Nuclear Explosions," there is a significant error in the last line of the third complete paragraph on page 121. The quantity "1 million" should have been "100 000." Also, in Detail *A* of the unnumbered illustration, the conduit or pipe should not continue on to the left since the purpose of the two shield wires is to protect cable that is not enclosed in conduit or pipe. In Fig. 1 curve *d* is for 0.01 in sheet copper with the joints soldered.

J. B. Hays
Bell Telephone Laboratories

A note on problem solving. The "previously considered unsolvable problems," mentioned on page 290 of the March issue, are those involving Poisson's equation. These problems, as a matter of fact, were solved for arbitrary boundary conditions and space charge distribution more than ten years ago by means of a resistance network analog.[1,2] By this means the problems that are solvable with the Poisson cell developed by the University of Michigan engineers can be solved equally well or better.[3,5]

Dr. J. R. Hechtel of the Litton Industries Electron Tube Research Laboratory has been instrumental in developing highly accurate network analogs for the solution of Poisson's equation. Network analogs have been built at Litton during the last few years and offered for sale to industry. Compared to the Poisson cell the precision resistance network analog advantages are:

1. The accuracy of the network analog digital computer combination is one part in 10^4, or at least one to two orders of magnitude better than the published accuracy of 2 to 3 per cent for the Poisson cell.[3-5]

2. With the network analog, simulating or changing boundary conditions requires merely connecting various network nodes together. Current can be injected into or taken from any of the nodes to represent space charge. Upon completion of a problem a new problem can be set up within minutes, whereas with the "Poisson cell" a new problem requires construction of a homogeneous resistance board. With the analog it is not unusual for personnel to solve two or three complicated electron gun problems in one day.[6]

The analog is more versatile and can solve more varied problems than can the Poisson cell. With the analog in combination with a high-speed digital computer, problems involving the following are reduced to a routine solution: electron motion in (1) axially symmetric electric and magnetic fields, (2) superimposed axially symmetric electrostatic and magnetic fields, (3) crossed electric and magnetic fields, (4) time-varying electric fields, and (5) at relativistic velocities; in heat flow problems; in hydrodynamic and aerodynamic problems; and in magnetic-field mapping.

Computer programs are available for each of these problems. In most cases, punched-card data are obtained from the network analog in less than half a day. Electron trajectories are computed and plotted with a high-speed digital computer at an estimated $9 per problem. The final solution of a problem can usually be obtained in two or three iterations.

We believe that the Litton precision network analog provides a less expensive and more adequate solution to problems involving Poisson's equation. In our opinion, its simplicity, versatility, accuracy, and ease of use provide highly superior advantages.

Leonard E. Bernier
Litton Industries
San Carlos, Calif.

1. Liebmann, G., "Precise Solution of Partial Differential Equations by Resistance Networks," *Nature*, vol. 164, July 1949, p. 149.

2. Hechtel, J. R., "Ein Widerstandsnetzwerk sur Losung der Poisson'schen Gleichung," *Die Telefunken-Rohre*, vol. 32, Feb. 1955, p. 30.

3. Hechtel, J. R. and J. A. Seeger, "Accuracy and Limitations of the Resistor Network Used for Solving Laplace's and Poisson's Equations," *Proc. IRE*, vol. 49, May 1961, p. 933.

4. Hechtel, J. R., "Electron Ray Tracing by Means of Resistor Network and Digital Computer," *IRE Trans. on Electron Devices*, vol. ED-9, Jan. 1962, pp. 62–68.

5. Hechtel, J. R., D. F. Brauch, A. Mizuhary, "New Numerical Methods in Electron Dynamics," *Proc. 4th Internat'l Cong. on Microwave Tubes*, Scheveningen, Netherlands, Sept. 1962, pp. 527–532.

6. "Precision Resistance Network Analogue," Litton Electron Tube Corp., San Carlos, Calif., July 28, 1961.

Long transmission telephone delays. The program summary (March issue, pp. 80, 82) of the informative paper by Gleason, *et al.*,[1] states that the paper "also established conclusively that the delay inherent in the 50000 mile transmission path is completely acceptable and tolerable for voice communications." Since the study included a large number of actual telephone conversations, there may have been a tendency to extrapolate from the author's results to a prediction of the acceptability of such delays in normal commercial telephony.

Experiments on delay have been conducted for the past two years in the laboratories of several European telephone administrations and at Bell Telephone Laboratories.[2] Data were presented to CCITT in 1963 and further data are scheduled for consideration by CCITT Study Groups XII and XVI in May and June, in arriving at tentative recommendations regarding the use of international circuits with delay.

In general, the CCITT data suggest that where telephone subscribers have been accustomed to high-quality low-delay connections such as are encountered on the transatlantic cable, some subscribers will consider round-trip delays of half a second as degrading the circuit, and will react adversely. At present, for situations involving lower-quality circuits like high-frequency radio, the greater reliability of satellite communications may well provide a more acceptable alternative, transmission delay notwithstanding.

In the light of available evidence it is not clear why the conversers in the Syncom II experiment appeared to have so little difficulty in communicating. Perhaps this is related to the differences in land-line conditions and psychological environment between this experiment and typical commercial telephone traffic. The discussions at CCITT should provide further insight into the acceptability of transmission delays.

J. E. Karlin
Bell Telephone Laboratories
Murray Hill, N.J.

1. Gleason, T. R., W. T. Tobias, and R. G. Keyes, "Communication Operations with Syncom II," *1964 IEEE International Convention Record*, to be published.

2. Riesz, R. R., and E. T. Klemmer, "Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications," *Bell Sys. Tech. J.*, vol. 62, no. 6, Nov. 1963.