VOLUME 7 □ NUMBER 2 IEEE spectrum FEBRUARY 1970

IEEE spectrum

features

39 **Spectral lines: The 1970 Convention**

This year's meeting continues the philosophy introduced a few years ago—that IEEE takes care of established fields of specialization by means of 100 symposiums a year, and that the March Convention's function is to supply papers covering other needs

40 **From engineer to entrepreneur: financing the transition**

John F. Jordan

What is the market for your product? The answer to this question is the most important factor in the success of a proposal for a new company as it is very likely to weigh heavily in the considerations of a potential investor

42 **Pulsars may be neutron stars**

Seymour Tilson

Strangely enough, the neutron stars, objects that were theoretical curiosities for over 25 years until the discovery of pulsars tentatively confirmed their existence, may be among the few truly stable astrophysical objects in the universe

56 **The anatomy of integrated-circuit technology**

Harwick Johnson

Although the technology of microelectronics is advanced, it is not basically new. As practiced today, it is largely an extension of silicon transistor technology and thin-film technology developed for passive elements and interconnections

71 **Electrooptic effects in ferroelectric ceramics**

Cecil E. Land, Richard Holland

Thin ferroelectric ceramic plates possess two types of interesting electrooptic effects: electrically controlled scattering and electrically controlled birefringence. Which is more noticeable depends on the average grain size of the ceramic

79 **Data transmission: a direction for future development**

Harry Rudin, Jr.

The strange and fascinating aspect of attempts to improve data communication over the telephone network is that, unfortunately, traffic theory has been neglected

86 **Why space broadcasting?**

R. P. Haviland

In contrast to conventional television techniques, satellites offer full national coverage from the beginning of service, with world coverage possible within ten more years—or less if system parameters are properly chosen

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

SCIENCE/SCOPE

A new launcher for the Phoenix missile now being built by Hughes has a fail-safe device that prevents accidental separation of missile from launcher during aircraft maneuvers. It uses no exotic or critical materials, weighs only two-thirds as much as an earlier model, and can be installed on either side of the fuselage.

A self-cleaning gas system makes it unnecessary to remove the launcher for maintenance after each mission. Hughes is building the launcher for the U.S. Navy's new F-14A fighter under contract to Grumman Aerospace Corp.

An imaging photopolarimeter for the Jupiter probe is being developed for NASA's Ames Research Center by Santa Barbara Research Center, a Hughes subsidiary, for the Pioneer F and G spacecraft to be launched in 1972 and 1973. Instrument will map the density and distribution of "asteroidal debris", measure the gas above Jupiter's cloud layers, and send back two-color spin-scan images of the planet.

Los Angeles has turned to aerospace technology for help in meeting the increasing demands for police, fire, and ambulance service. The city council recently chose Hughes to make a one-year study of the city's over-burdened services and to draw up a plan for a command-&-control system that would provide rapid pinpointing of field forces, computer dispatching, automated status displays, computerized information files, individual communications for hazardous-duty personnel, and automatic transmission and signaling for police vehicles.

NASA's Atmosphere Explorer satellite, now under study at Hughes, will carry a propulsion system that will enable it to climb to an apogee of 2500 miles in its variable orbit around earth. Every two hours it will dip back into the upper atmosphere for 10 to 20 minutes, swooping within 90 miles of earth.

The "yo-yo" satellite's scientific objectives will be to obtain data on the behavioral relationship of the upper and lower atmosphere, solar energy absorption, density of the atmosphere's charged-particle structure, and the diurnal bulge that appears to circle the earth as the sun heats the atmosphere.

Career opportunities for engineers at Hughes include: Signal Processing Systems Analysts, Computer Software Analysts, Radar Systems Engineers, and Circuit Designers. B.S. degree, two years of related experience, and U.S. citizenship are required. Please write: Mr. J. C. Cox, Hughes Aircraft Company, P.O. Box 90515, Los Angeles 90009. Hughes is an equal opportunity employer.

A laser proximity fuse and larger fins are being given to the U.S. Air Force Falcon in a program now underway at Hughes to make the air-to-air missile more effective against maneuvering targets. The proximity fuse's optically focused laser beam, which is reflected off the target, cannot be confused by electronic countermeasures and is virtually impossible to detect.

Because the laser gear is extremely compact, it can be tucked into a collar around the nozzle of the Falcon's rocket motor, leaving space in the missile for a larger and more powerful warhead.

Creating a new world with electronics

HUGHES

HUGHES AIRCRAFT COMPANY

92 **The Esclangon diagram for voltage and current along a transmission line**

Michel Poloujadoff

Although the Esclangon diagram does not replace other methods for solving transmission-line problems, it is the most efficient way of showing the voltage and current variations along an electric line

95 **New product applications**

A staff-written report on some carefully selected new products emphasizing one or more of their potential applications as an aid to engineers who may wish to apply these products to solve their own engineering problems

21 **IEEE International Convention and Exhibition Technical Program**

128 **Japan Electronics Show 1970**

the cover

Of the hundred or so pulsars discovered to date, that with the shortest period (shown here) lies within the Crab Nebula; it emits powerful radio bursts at precisely periodic intervals of 35 ms. During their short history, many theories have arisen concerning these mysterious celestial objects. These are discussed in the article beginning on page 42, which concludes that pulsars actually may be neutron stars.

departments

6 **Forum**

11 **News from Washington**

14 **Focal points**

18 **Calendar**

101 **Scanning the issues**

103 **Advance tables of contents**

Future special issues, 104

110 **Translated journals**

112 **Special publications**

114 **Book reviews**

118 **News of the IEEE**

125 **People**

134 **Index to advertisers**

IEEE SPECTRUM EDITORIAL BOARD J. J. G. McCue, *Editor;* F. E. Borgnis; C. C. Concordia; A. C. Dickieson; Peter Elias, E. G. Fubini; E. W. Herold; D. D. King; B. M Oliver; J. H. Rowen; Shigebumi Saito; J. J. Suran; Charles Süsskind; Michiyuki Uenohara

IEEE SPECTRUM EDITORIAL STAFF Ronald K. Jurgen, *Managing Editor;* Robert E. Whitlock, *Senior Editor;* Seymour Tilson, *Staff Writer;* Evelyn Tucker, Marcelino Eleccion, Paul Hersch, *Associate Editors;* Mary L. Jones, *Assistant Editor;* Stella Grazda, *Editorial Assistant;* Ruth M. Edmiston, *Production Editor;* Herbert Taylor, *Art Director;* Janet Mannheimer, *Assistant to the Art Director;* Morris Khan, *Staff Artist*

IEEE PUBLICATIONS BOARD M. E. Van Valkenburg, *Chairman;* David Slepian, *Vice Chairman;* F. S. Barnes; F. E. Borgnis; P. E. Gray; E. E. Grazda; Y. C. Ho; R. W. Lucky; J. J. G. McCue; Seymour Okwit; R. L. Schoenfeld; N. R. Scott; G. W. Stagg; David Van Meter

PUBLICATIONS OPERATIONS Elwood K. Gannett, *Director, Editorial Services;* Alexander A. McKenzie, *Assistant to the Director;* Patricia Penick, *Administrative Assistant to the Director;* Ralph H. Flynn, *Director, Publishing Services;* William R. Saunders, *Advertising Director for Publications;* Henry Prins, *Manager, Marketing Services and Research;* Carl Maier, *Advertising Production Manager;* Irene Corder, *Assistant Advertising Production Manager*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE SPECTRUM is published monthly by The Institute of Electrical and Electronics Engineers, Inc. Headquarters address: 345 East 47 Street, New York, N.Y. 10017. Cable address: ITRIPLEE. Telephone: 212—752-6800. Published at 20th and Northampton Sts., Easton, Pa. 18042, and additional offices. **Change of address** can be made effective with any particular issue if received by IEEE Headquarters by the 10th of the preceding month. **Annual subscription:** IEEE members, first subscription $3.00 included in dues. Single copies $1.50. Nonmember subscriptions and additional member subscriptions available in either microfiche or printed form. Prices obtainable on request. **Editorial correspondence** should be addressed to IEEE SPECTRUM at IEEE Headquarters. **Advertising correspondence** should be addressed to IEEE Advertising Department, at IEEE Headquarters. Telephone: 212—752-6800.

Responsibility for the contents of papers published rests upon the authors, and not the IEEE or its members. All republication rights, including translations, are reserved by the IEEE. Abstracting is permitted with mention of source.

Second-class postage paid at Easton, Pa., and additional offices. Printed in U.S.A.

OTHER IEEE PUBLICATIONS: IEEE also publishes the PROCEEDINGS OF THE IEEE, the IEEE STUDENT JOURNAL, and more than 30 Transactions for IEEE Groups with specialized interests within the electrical and electronics field. Manuscripts for any IEEE publication should be sent to the editor of that publication whose name and address are shown on page 17 of the January 1970 issue. When in doubt, send the manuscript to E. K. Gannett, Editorial Services, at IEEE Headquarters, for forwarding to the correct party.

One of the exciting projects at AIL is the development of an engineering model of a retrodirective phased array for NASA. In this issue, Communications Systems Department Head Peter Sielman describes its application to a pressing satellite problem.

A Retrodirective Phased Array For The Data Relay Satellites

DATA RELAY

One of the limitations faced by manned and unmanned low earth orbit satellites is that the earth is in the way of a communications link between the satellite and its readout station(s) during much of the orbit. Up to now this problem has been circumvented by providing the satellite with data storage or by setting up a multiplicity of readout stations. Neither of these solutions has been entirely satisfactory. Storage bandwidth is limited by the state of the art, processing is necessary to fit the data into the limited storage and short readout times over ground stations necessitates a fast dump of the data which puts an added burden on the satellite-to-ground data link. The multiplicity of readout stations are not only expensive but also create logistics problems of recorollating the data through a terrestrial communications network.

NASA is keenly aware of these problems and has, therefore, initiated the investigation of an attractive alternative. This alternative is to place two or more Data Relay Satellites in synchronous orbit. The perspective afforded from synchronous orbit is about 160° in latitude. Therefore, a low earth orbit satellite can continuously communicate with a single readout station with the aid of two well placed synchronous Data Relay Satellites.

Clearly, the same Data Relay Satellite could be used to interrogate a variety of earth based sensors and relay the data back to a common readout and analysis center.

SATELLITE ATTRIBUTES

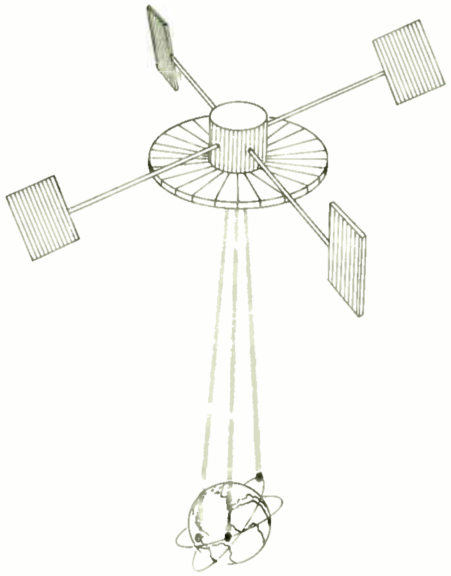
What are the desired attributes of such a Data Relay Satellite? The obvious ones that come to mind are:

1. Multiple simultaneous beams.
2. Sensitivity to the weak signals emitted by small users.
3. High effective radiated power.
4. Isolation between transmission and reception.

All of these requisites appear to point to a large aperture phased array. (For one or two simultaneous users, mechanically steered dishes are cost effective alternates.)

5. Retrodirectivity.

If a Data Relay Satellite is to provide useful service to a multiplicity of users, it would be desirable for it to operate independent of a priori knowledge of the location of the users. A retrodirective antenna achieves this capability by sending signals back in the direction from which they are received. A large aperture retrodirective phased array then would meet all of the desired attributes listed above.



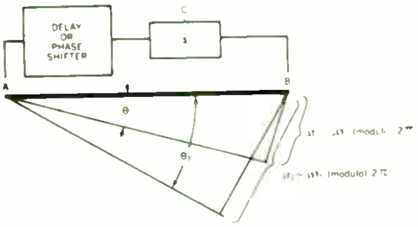


FIG. 1. Phased Array Principles

RETRODIRECTIVE PHASED ARRAY

The basic principles of a phased array are well known. A wavefront striking two elements (A and B) of a phased array at an angle θ_1 will be summed coherently at C if the signal from A is delayed by Δt_1 or equivalently shifted in phase by $\Delta\phi_1$. Likewise, a signal to be transmitted can be launched at an angle θ_2 by delaying the signal transmitted from element A by Δt_2 or the equivalent phase $\Delta\phi_2$.

A retrodirective phased array achieves its capability by isolating a known component (pilot) of the received signal at each element of the array. This pilot is then mixed with the remainder of the received signal (data)

$$\underset{\text{(pilot)}}{A_1 \cos(w_1 t + \phi)} \times \underset{\text{(data)}}{A_2 \cos(w_2 t + \phi)} =$$

$$\frac{A_1 A_2}{2} \cos (w_1 - w_2)\, t + \text{filtered signals}$$

which makes the received signal at each element independent of phase and, hence, easily summed over all elements.

When the pilot (with conjugate phase) is mixed with data to be transmitted

$$\underset{\text{(pilot)}}{A_1 \cos(w_1 t - \phi)} \times A_3 \cos w_3 t =$$

$$\frac{A_1 A_3}{2} \cos \left[(w_1 - w_2)\, t - \phi \right] + \begin{matrix}\text{filtered}\\ \text{signals}\end{matrix}$$

a signal is generated at the element which, when combined with the signals from the other elements, points the beam back in the direction from which the pilot was received.

"SQUINT"

To permit the Data Relay Satellite to receive from and transmit to users at the same time, it is necessary to provide isolation. Frequency offset is the easiest means of achieving this isolation. However, if a signal is received at one frequency and is to be transmitted back in the same direction at another frequency, the received phases are not the right phases to form the transmit beam. This problem is called "squint". In the development work that AIL is currently doing for NASA, this problem has been resolved by placing the elements so that the signal received at element A at one frequency (f_R) is just the right phase at element B to transmit at another frequency (f_T). Each element is still used to receive and transmit; the proper phase is just passed from each element to its appropriately located partner.

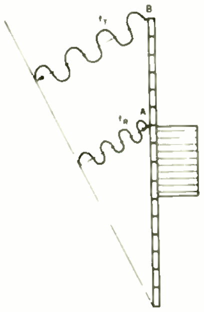


FIG. 2. Squint Problems

MODULES

Since a large aperture phased array has many elements, and since the cost of satellite systems is directly proportional to weight, it is extremely important that the weight of each element and its associated electronics be kept as low as possible. (An added ounce per module in a 400 element array means 25 extra pounds of payload.) This has made the work that Group Leader Larry Schwartz and his group are doing particularly exciting. NASA has asked them to make integrated circuit modules so as to be able to realistically assess what the weight of a Data Relay Satellite will be. So far, the results have been encouraging.

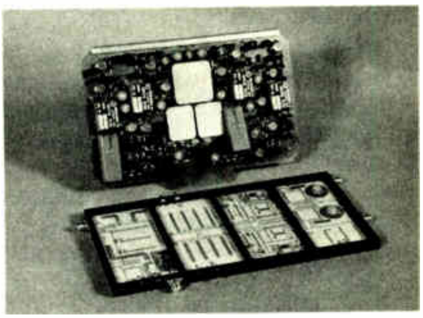


FIG. 3. IF and RF Modules

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

AIL a division of CUTLER-HAMMER
DEER PARK, LONG ISLAND, NEW YORK 11729

Circle No. 6 on Reader Service Card.

Forum

Readers are invited to comment in this department on material previously published in IEEE SPECTRUM; on the policies and operations of the IEEE; and on technical, economic, or social matters of interest to the electrical and electronics engineering profession.

Professional registration

I read with interest W. G. Naef's letter concerning professionalism in the engineering community, and the need for P.E. certification. Although I do not chew, carry a bottle, or use a red kerchief, I feel that I can speak for many of my fellows.

I did not take the E.I.T. or P.E. exams in my home state (Ohio) as I did not feel that the possession of a P.E. sticker could do anything to add to my credentials as a member of the electronics industry.

Although the present P.E. structure may reflect on the abilities of those in the older disciplines, it is not germane to those of us in the electronics business. With the heavy emphasis on concrete, steam, and steel in the E.I.T. exams, is there any wonder that growing numbers of E.E. students are electing to bypass them?

Until the P.E. groups catch up with technology, they will continue to be ignored by large numbers of IEEE members, with justification.

E. G. Machak
IBM
Owego, N.Y.

Professors and research

Charles Anderson, in his article, "Research and the Changing Campus Environment" (*Spectrum*, December 1969), has touched on a variety of topics and issues on which I have strong feelings. Because of time and space limitations, I shall discuss only one basic point: Is classified research an appropriate activity for engineering professors?

Let me begin with the more basic question: Why should an engineering professor participate in a research activity? By definition engineering work is to be "real-world" oriented. Engineering research activity on real-world problems is necessary for professors. They need to maintain a sense of relevancy and their graduate students require a relevant educational and training experience. A professor's research experience should be used in the classroom to point out the applicability of the theory he is teaching. As his research yields solutions to problems, these solutions should be brought into the classroom so that the students are educated in the latest techniques. As technologies develop, new courses should be developed and old courses revised in order that the information is put quickly into the hands and minds of those who will soon have need of it.

How well does classified, defense-oriented, research satisfy the above purposes of university, engineering research? Problems from this area are unquestionably real-world oriented and relevant. Beyond this point, however, classified research cannot satisfy our purposes. Not every graduate student can participate in this area, and whether or not he can is not based on his technical competence. The most damning aspect of classified research is that the professor may not talk about it openly and freely. Thus he is forced to keep valuable new knowledge from his students and colleagues. At that point (in my mind at least), he ceases to be a professor and he becomes a research engineer in the employ of his contractor. If he is that, let him truly be that in a governmental or industrial research laboratory, not in a university.

I appreciate the fact that in some areas, the most advanced work in the area is of a classified nature. If this is the case, then surely the university professor should be allowed to participate in that work. However, the degree of participation is important. I believe that such work, although important, is peripheral to his main function as a professor. Thus classified research should be treated in much the same way as consulting. It should be allowed, but only for a specified limited fraction of his time.

R. M. Anderson
E.E. Dept., Purdue University
Lafayette, Ind.

The changing world

In reading Mr. Anderson's article on the changing environment in the December issue of *Spectrum*, I almost got the feeling that here is a man who is really with it. I say almost because his choice of words suggests a traditional outlook on the environment.

It is certainly true that the changing world will cause many changes in industry and that industry should prepare for this.

I doubt though that a different kind of management is enough. This suggests the traditional setup where a small management group controls an organization by utilizing more and more sophisticated methods of KITA to motivate their subjects. I believe the changes will primarily be in the organization.

There are indeed more and more people who wish to function in an advisory capacity. Of course! That's why they were sent to school in the first place.

The name of the game in industry, as it was at the universities, will be participation in the controlling process. The young people who are coming out of the universities now will not be satisfied with letting the managers of industry use technology to achieve their (the managers') goals. The young people will want to have a say. The under-30 generation has no sentimental attachment to technology or to the great importance of technology in bringing most of the western world out of poverty and misery. As far as they are concerned technology is just one part of life, and in many ways not a very pleasant part.

We should be grateful. Increased participation can only make life richer. The younger generation is by and large motivated by a desire to make the world a better place to live, not only for themselves as individuals, but for everybody.

Leif Lundquist
Bell Telephone Laboratories, Inc.
Holmdel, N.J.

If education is the process by which a community preserves its cultural pattern, then it is time education changed our cultural pattern. So the U.S. is the richest nation on earth. We have a mass televised culture whose intellectual content is almost nil. We have a standard ideal of beauty to which all manner of people try to adapt themselves. We have all manner of mass-produced art objects by which people stifle their own creativity. But most of all, we have the almighty dollar, which, as we know, can fulfill all our needs.

Can it buy arete? And who does not respond to the great books of the past? The writings of the Greek philosophers concern man in any society, in any age. If we are to have a luxurious future in which most things are done for us, we would do well to toy with the idea of arete-excellence.

How will we amuse ourselves in the womb-like future? The Greeks of the Hellenic period believed that each man must develop his abilities as much as possible. And we believe in the fast buck.

Since science has altered man's view of himself it is imperative that the arts continue to question the morality of science. We must always ask ourselves whether science's advances should be employed. They must be used justly (and, of course, they probably won't be). Although man has developed intellectually, he has not changed much emotionally and spiritually. Science has done little to control his anger, violence, and greed. Rather, it was the discovery of the wheel, iron, and use of the horse that put him on the road to power. About 2000 B.C.

I must inform the editor that the Educated Man is alive and well and living in the liberal arts faculty. Many of us want to teach. We will teach the foolish ideals found in arts; the ideals that say man can be creative, humane, moral . . . .

Kathryn Elliott
Sir George Williams University
Montreal, Que., Canada

It will be a better thing for us all if someone starts to control engineering action, and probably the best ones to do so are those who are out of sympathy with industrial enterprise.

Also, their ignorance of engineering is fast disappearing. See, for example,

1. *IEEE Spectrum*, December 1969—color television sets that burn homes and emit radiation.

2. *Time*, December 19, 1969, p. 51—electric power plants polluting the environment.

The reason the "others" will start to control engineering activities is that engineers have steadfastly refused to exert control themselves. For instance, when may I expect to hear that malpractice charges have been brought by the IEEE against the engineers who were (or should have been) *responsible* for the design of those color television sets, those electric power plants?

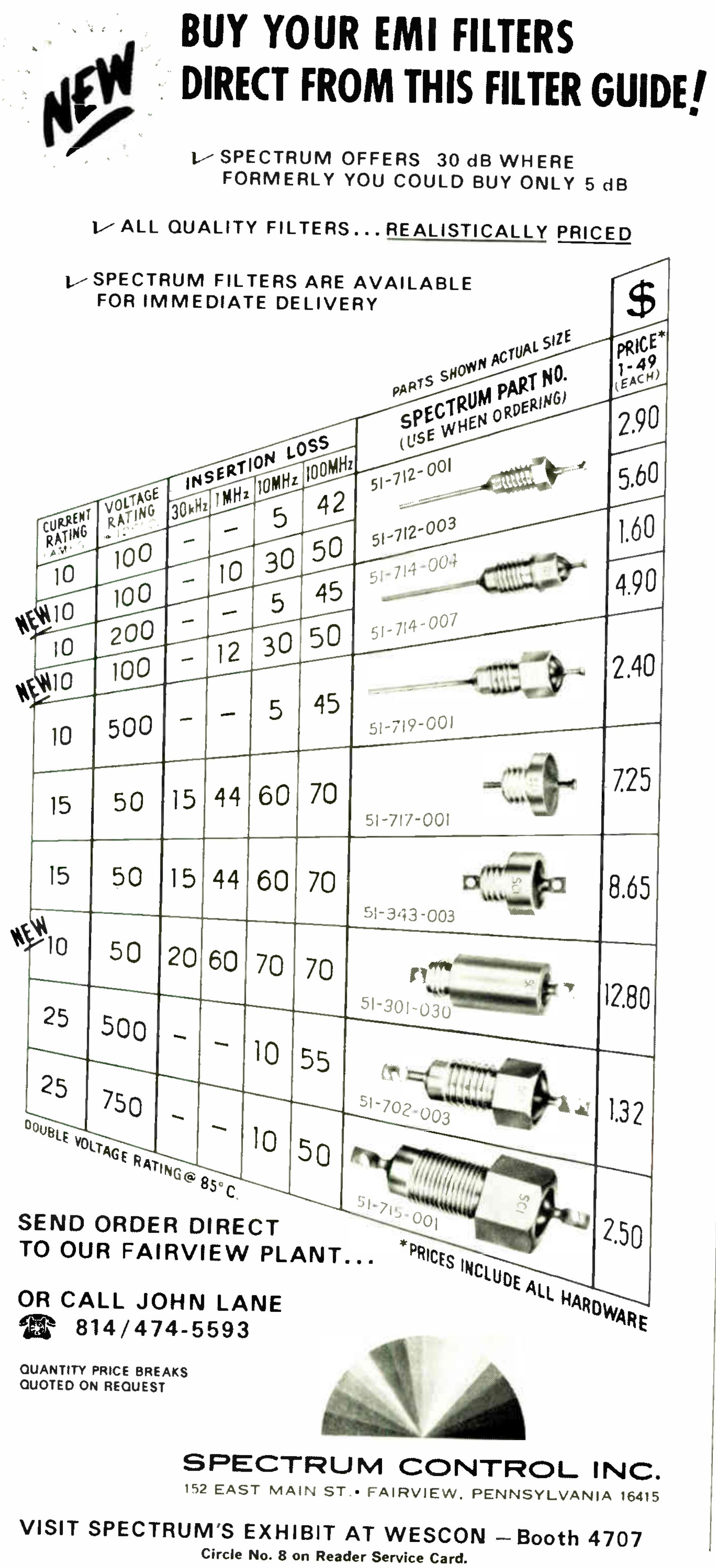
G. H. Friedman
Torrance, Calif.

Engineers and society

"Spectral lines" in the December 1969 issue of *Spectrum* is important to the world of the engineer as it is today. A failure of education to produce people who can maintain and expand our engineering works can, indeed, doom humanity even before we can breed ourselves out of existence. Our Editor is correct in remarking on the college "publish or perish" movement. It may mean publish for the faculty and perish for a civilization. In the past dozen years required credits for the B.S. degree in engineering have been reduced by over 10 percent. However, an increase in forced study of the social sciences and humanities has resulted in a decrease of about 30 percent in engineering subject matter. In the November 1969 issue of the magazine *Electric Light and Power*, Editor A. J. Stegman points with alarm to new, base load power plant equipment gingerly used as undependable peaking capacity. Is this the result of poor education? Stegman suspects the crafts, but testing is the function of the engineer.

In electrical engineering it is basic theory of circuits and electric machines that have suffered most. These are the things that make it possible for us to mine raw materials, manufacture hard goods, process and package food as well as refrigerate it; that operate our heating plants, activate our water supplies, dispose of our sewage, and, in fact, make possible the support of our considerable population.

Our Editor, in his last paragraph, supposes a "good bet" that much engineering in the future will be controlled by social scientists. I wish to disagree. The prejudicial sciences cannot guide man to required production. Production, if it exists, will be guided by the natural and physical scientists, leavened by the art of engineering. The engineer is the truly socially-conscious among us. He ignores defamation and strives greatly to keep up with the needs, and even demands, of our populace; and this in the face of



an increasingly overpopulated and unresponsive civilization.

G. C. Barnes, Jr., P.E.
Blacksburg, Va.

Congratulations to *Spectrum* for taking that first, faltering step from the narrow confines of the laboratory into that sometimes sunny, frequently stormy outside world.

I have observed, with some sadness, how little the engineering fraternity as a whole participates in the great social and economic issues of our time. On an even sadder note is the observation of the withdrawal of the younger engineers and engineering students from the arena of current thinking and social action. Although young in years, they are already middle-aged in outlook. How in the world can we expect them to compete successfully for future positions of leadership with the young people who are firmly in the mainstream of a changing America and are shaping our concepts for a better country?

It is ironic that engineers who have been trained to question, to distinguish between fact and fancy, to innovate in the technical domain, should be so laggard in applying these sharply honed faculties to the more significant problems of living. I think we all know that other forces have counteracted these tendencies; forces such as the exclusion of humanities studies from many engineering school curriculums, the fear of clearance withdrawal in the military-industrial segment, the parochialism of technical societies, and others.

By continuing in this newly chosen path, *Spectrum* may be providing its most important service yet to young, middle-aged, and, perhaps, even to older electrical engineers.

Edward S. White
Santa Cruz, Calif.

I like what the IEEE is doing in widening the forum of debate. Cannot this be extended? Why not *public* debate with leading engineers and members of other disciplines? In the October *Spectrum* for example, Dr. Trauboth argues with Professor Toynbee. How much better if they could both talk together on television. Toynbee would probably win since he is a better communicator to the layman—but this sort of thing could, in time, demonstrate that the engineer is on a par with the historian, doctor, lawyer, or whatever. And perhaps such exposure would remind the engineer of his serious lack of foresight in noting the consequences of his actions. In short, we need educating to complement our training.

D. W. Soughan
Ripley, Derbyshire, England

Become a better speaker

"Engineers and scientific personnel in general are extremely boring speakers, as can be attested to by anyone who ever attended a symposium or technical session at a convention." So stated G. M. Cinque of Stamford, Conn., in his letter in the October *Spectrum*.

"Amen" must have been the reaction of readers who have suffered through such sessions!

However, this need not be so. In every area of this country can be found a toastmasters' club, wherein members are able to practice prepared and extemporaneous speaking—and receive instant feedback (called evaluation). This mutual evaluation of all aspects of the toastmasters' program—also including the functions of chairman, toastmaster, and evaluator—is its forte.

Again and again, during my four years of membership, I have observed hesitant and unimpressive novices develop into confident and effective speakers. Furthermore, this was accomplished in a congenial atmosphere of semimonthly dinner meetings.

Any engineer who has any sensitivity regarding inadequacies in his public presentations should do himself and his peers a favor and inquire of Toastmasters International at Santa Ana, Calif. 92711. He will receive a local contact within two weeks.

Bernard F. Dwyer
Toastmasters International
Watervliet, N.Y.

Correction

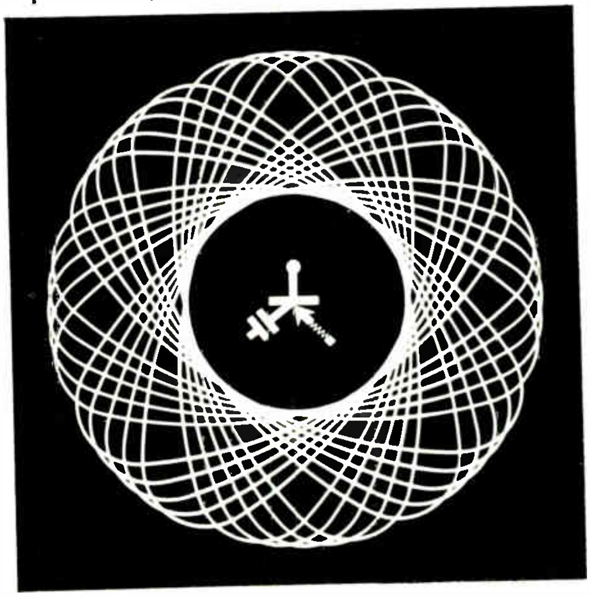
The article by James Tow, "A Step-by-Step Active-Filter Design," which appears on pages 64–68 of the December 1969 issue, contains several errors in the artwork.

On page 67, Fig. 8, upper left, the resistor should be labeled

$$R_1 = \frac{1}{aC_1}$$

In Fig. 9, center right, the resistor should be labeled

$$R_8 = \frac{k_2}{k_1} \frac{ma}{|mb-d|} \sqrt{bR_{10}}$$



VOLUME 7 □ NUMBER 2 IEEE spectrum FEBRUARY 1970

Spectral lines

The 1970 Convention. Again the IEEE March Convention is nearly upon us, posing the questions of whether to attend and which attractions to select.

This year's meeting continues, with some enlargement, the philosophy introduced a few years ago by Dr. Herold. Its basis is that IEEE takes care of the established fields of specialization by means of a hundred symposiums per year, and that the function of the March Convention is to supply papers covering other needs, while serving also as the locus for a maximally useful exhibit of engineered electronic hardware. These two themes interlock. Only for a very large attendance is it feasible to mount such a large exhibit, and for a large gathering of engineers, only a broad-interest program can supply cohesion.

It is unfortunate that the available facilities dictate separation of the exhibits from most of the papers. The separation suggests disunity, and that is an illusion. The Convention is about engineering. Engineering does have its negative side: it sometimes says "Don't do it"; some technical or economical unsoundness may stand in the way. But the usual end result of engineering can serve as the basis for further engineering and, in a large fraction of the cases, can be put on exhibit. Therefore it would be better if the engineering dreams of today and the consequences of the engineering dreams of yesterday could all be set forth under one roof. The Institute does the best it can with the most suitable real estate that is available for the purpose in one of the world's largest cities.

Engineering societies have a difficulty not felt in societies concerned with pure science. The output of engineers tends to be proprietary. A wise tradition of engineering societies dictates that the merits of proprietary developments are not debated in society meetings or publications. A society tries to guarantee the technical quality of the work that it publicizes, but shuns involvement in the relative merits of proprietary schemes or devices.

Nevertheless, decision about whether to adopt and use a device is an important engineering activity. A valuable function of an engineering society is the bringing together of suppliers and users of the things that engineering has produced, so that judgment about adoption can be made, case by case. Much of this bringing together is done by the advertising in the journals. The smartest advertising in technical journals is truly educational. It is the same at the Convention. Exhibitors soon learn that if they merely send salesmen, they do not get the most out of the Convention. What is valuable about the exhibits is the opportunity for users and potential users to talk to engineers who have an intimate and detailed knowledge of the product. These talks are two-way affairs; what the exhibitor's engineer learns from them is embodied in later models of the product.

The purpose of the exhibits and the papers is the same: Continuing Education. A printed article is no substitute for a well-prepared oral presentation; similarly, a descriptive brochure is no substitute for personal contact with the people who designed the device that you are wondering whether you should use.

It is the theme of Continuing Education that holds the Convention together. Noteworthy on the 1970 program of papers are two broad-gauge evening sessions, one on the outlook for technology in the coming decade, and one on adaptation to change by organizations. Following recent precedent are three early-morning tutorial sessions: Computers and Patient Care; Programming for Industrial Process Computers; and Digital Filters: Design and Applications. There will be a two-day course on Monolithic Integrated Circuits, and one on Using Quantum Electronics Today. Details about the program can be found in a special section of this issue. Many members will welcome the expanded roster of Technical Applications sessions at the Coliseum.

As compared with ten years ago, the improvements in recent Conventions have been purchased at a cost. Those of us who are attached to a specific discipline used to be able to find many of our intimate professional acquaintances in the room devoted to papers by a particular IEEE Group. A convivial and informative lunch or dinner often followed. The interdisciplinary emphasis of recent Conventions has broken up that pattern. A few Groups have recognized that perhaps we can have it both ways. The Convention tries to do all that can be done in a single week in a single place, and leaves the rest to the symposiums. But some enterprising Groups have noticed that Thursday is not actually the end of the week; they have organized meetings in New York on the Friday, and some have even trespassed on Thursday afternoon. Possibly emerging is a pattern that will retain the advantages of the present Convention philosophy, but will recapture the values that have been submerged as the Convention has broadened.

J. J. G. McCue

From engineer to entrepreneur: financing the transition

In view of today's inflated economy, the engineer planning to go into business for himself would do well to look to the older, well-established companies for advice and financial assistance

John F. Jordan D. H. Baldwin Company

A well-established company that wishes to extend its activities in technical fields will sometimes find it advantageous to assist and finance a new technology-based company. In this article, proposals that are likely to be considered by such a company are discussed and their preparation and evaluation described.

For the electronics engineer the past few decades have provided unusual opportunities to make the transition to business entrepreneur. He has found that his technical competence and his inventions can be made the basis for a new business. And, because the United States has been in a nearly continuous inflationary cycle, until recently many investors were encouraged to finance newly formed companies, such as those started as a result of expanding electronic technology. These new companies gave promise of swift growth even though they may have been highly speculative, and the investors looked to a rapid increase in stock price in order that their capital investment might increase faster than inflation could deplete it.

As we know, a number of such companies have had phenomenal growth rates, and the stockholders profited handsomely when the stocks were sold to the public at high price-to-earnings ratios. These profits were subject to capital gains taxes with a maximum rate of 25 percent. Thus, high tax rates on ordinary income, together with the ready availability of money at low interest rates, made the financing of new companies relatively easy.

Lately, however, the pace of inflation has increased so rapidly as to cause considerable alarm in fiscal circles, and the Federal Reserve has found it necessary to slow the growth rate of the money supply and drastically increase interest rates.

The anticipation of a slowing economy and resulting lower corporate earnings caused a substantial lowering of stock prices during 1969. Since new stock issues normally require a buoyant stock market, we can now expect a reduction in the flow of speculative funds for the formation of new companies. On the other hand, there are established companies that have always taken the long-term view of business with its inevitable ups and downs. They are interested in increasing company earnings on a long-term basis rather than in stock manipulation.

Corporate growth in today's world hinges upon technological progress with its attendant new products, processes, and materials. However, only the giant billion-dollar corporations can afford the wide spectrum of research and development that will produce a continuous flow of these new products. Smaller concerns, those whose sales total $100 million or less, need to be more alert to opportunities presented by outsiders.

It should be pointed out, however, that the personnel and financial policies of many of these organizations prevent them from financing new companies, and it is assumed, of course, that one would use judgment and approach only those likely to be interested. In the medium-size concern, usually some executive has the responsibility for extending its activities to insure corporate growth, and this person should be contacted to determine whether the formation of a new company is permissible under company policy. If such is the case, then it is usually necessary to make a formal proposal in writing.

The proposal

The written proposal for a new company may be divided into three general sections:

The idea. The idea is generally a product, a process, a material, or a service. It should be described completely and in great technical detail. A proposal submitted to a responsible company will be subjected to a thorough analysis, by the engineers of the company or by consultants, and so you must make certain that it can withstand technical investigation.

The need. What is the need (market) for your proposed product? The answer to this question will be the most important factor in the success of your proposal. Except in those rare instances where your idea involves a major technical breakthrough (for example, the invention of the transistor) in which the advantages of your product are self-evident, it will be necessary to supply maximum information. Your proposal should include your best estimate of the annual market for your product. Be as specific as possible. In the case of an industrial product, if you can name the companies that would comprise a market, the amount of their probable annual purchases, and the names of individuals within these companies to whom you intend to sell, it will be evidence of a carefully thought out plan. In any case, this is the type of question you are likely to be asked. Information as to the extent of the total existing market and names of competing companies, if such already exist, is also valuable.

The importance of this part of the proposal cannot be overstressed because it is likely to be given great weight by a potential investor. Most companies are sales limited and this information is most likely to interest them.

Resources required. The resources that will be required may be divided into two categories:

Human resources. You will require a group or organization to engage in any worthwhile activity. The organization you propose should be completely described, together with the names and qualifications of those whom you propose to employ. The people needed fall into four groups: technical, manufacturing, financial, and sales.

Usually the person or persons supplying the idea are technical people themselves and know other qualified technical people who can work together harmoniously.

The importance of selecting qualified manufacturing personnel is often overlooked by the technical man. The efficiency of his manufacturing operation will determine the cost of the product and the ability to compete. Here an investing company can often be of help.

It is in the selection of financial people that the technical man is usually least qualified. Generally, he has had little contact with these types and little knowledge to enable him to judge their abilities. And yet it is here that many early difficulties arise. Poor financial ability often results in underestimation of financial needs and the time required to become profitable, poor utilization of funds available, and failure to project cash needs properly. When such cash needs develop, lack of close acquaintance with the financial community can result in an inability to arrange further necessary financing. In this area, an investing company can prove most valuable to a new establishment since such services are then readily available. In addition, if further financing is required, it is easier to obtain, particularly in times like the present.

The selection of sales personnel is most critical also. The difference in performance between individual salesmen is probably greater than that of any other skill. Here an existing company, assuming that it has some expertise in the field, can be of help.

Financial resources. Your proposal should include a cash-flow projection, which is a list of all expenditures by category, together with all items of income vs. time. The time should extend at least to breakeven (the point at which income equals outgo). This gives the investing company an estimate of the size of investment required and is, of course, very important. Most organizers of new companies underestimate the time required to become profitable. A few new companies have accomplished it in as little as six months, but two to three years is more likely. Here again, if you need additional funds, dealing with an investing company is easier than dealing with individuals or groups.

An estimate of the cost of producing and selling your product together with its market price and sales volume should be included. This permits the investing company to estimate the probable return on its investment—a very important consideration.

Assuming that your proposal meets with approval, the next step is negotiation of a preincorporation agreement between the idea or founding group and the investing company.

Preincorporation agreement

The preincorporation agreement will cover the arrangements as to salaries, titles, and ultimate participation in stock ownership. There are several methods by which members of the founders group can obtain stocks. One of the simplest is the qualified stock option, which permits employees of the new company to buy stock over a five-year period at an initial low figure and, by observing certain rules, sell it later at a higher price. Under present tax law (October 1969), the profit is then subject to a capital gains tax of 25 percent maximum.

Of course, a higher price is contingent upon the success of the company and thus the stock option serves as a great incentive to work for such success.

Those concerns interested in long-term earnings generally do not wish to sell the stock to the public when the company becomes profitable, although this is not entirely precluded. It is necessary, however, to provide a market for the stock of option holders in order that they may realize a return on their investment in time and effort. Therefore, it can be arranged for the investing company to buy the stock from the option holders after a reasonable lapse of time. Satisfactory formulas such as a suitable ratio of price to earnings after taxes, or an exchange for stock in the investing company at a mutually agreed upon ratio, can be written into the preincorporation agreement.

Rules for the qualified stock option

The U.S. Internal Revenue Service rules under which a stock option is "qualified" can be summarized as follows:

1. The option must be granted pursuant to a written plan that is approved by the shareholders of the granting corporation within 12 months before or after the plan is adopted.

2. The written plan must state the aggregate number of shares that may be issued under the options granted under the plan and must specifically state the employees or class of employees eligible to receive options.

3. The option must be granted within ten years from the date the plan is adopted or the date the plan is approved by the shareholders, whichever is earlier.

4. The option, by its terms, may not be exercised after the expiration of five years from the date of the grant.

5. The option price may not be less than the fair market value of the stock at the time of the grant.

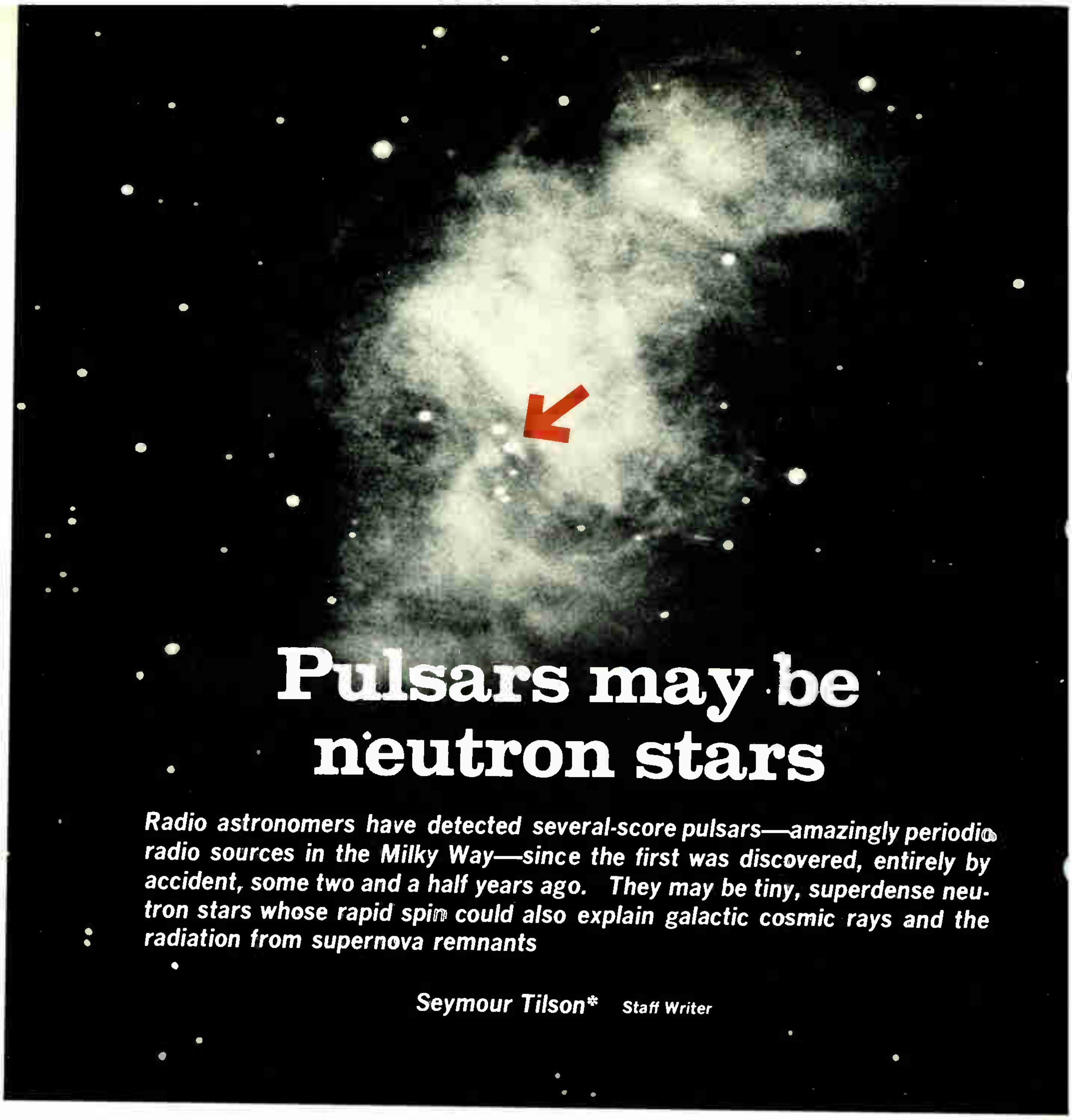
6. The option, by its terms, may not be exercised while there is outstanding any restricted or qualified stock option previously granted to the employee.

7. The option, by its terms, must not be transferable by the employee, other than by will or the laws of descent and distribution, and must be exercisable during the employee's lifetime only by him.

8. The employee receiving the option immediately after the option is granted may not own stock having more than 5 percent of the total combining voting power or value of all classes of stock of the employer corporation or its parent or subsidiary corporation.

Revision of a paper presented at WESCON, San Francisco, Calif., August 19–22, 1969.

John F. Jordan (F) received the E.E. degree from the University of Cincinnati in 1932. He joined the D. H. Baldwin Company in 1935 where he was one of the early workers in the field of electronic music, a field in which he holds a number of patents. He has served at that company successively as research engineer, director of engineering and research, and, since 1956, as vice president. He is also president of Baldwin Electronics, Inc. In 1953 he completed the Advanced Management training program at Harvard Graduate School of Business Administration and he has been active in the formation of several technology-based companies. Mr. Jordan has served as Chairman of the Cincinnati Section. In 1960 he was selected as Engineer of the Year by the Technical and Scientific Societies Council of Cincinnati.



Pulsars may be neutron stars

Radio astronomers have detected several-score pulsars—amazingly periodic radio sources in the Milky Way—since the first was discovered, entirely by accident, some two and a half years ago. They may be tiny, superdense neutron stars whose rapid spin could also explain galactic cosmic rays and the radiation from supernova remnants

*Seymour Tilson** *Staff Writer*

Pulsars are celestial objects that emit brief, intense bursts of energy, principally, but not exclusively, at radio frequencies. Several pulsars "tick" at time intervals whose precision exceeds that of quartz crystal clocks and approaches that of atomic time standards. The announcement of the discovery of the first four pulsars, early in 1968, triggered an unparalleled competitive frenzy among radio astronomers to find additional pulsars. By last fall, some 100 of these precisely periodic radio sources had been discovered, theories about what they might be had proliferated and been laid to rest almost as rapidly, and many important new facts about several of them had been elucidated.

Where careful measurements have been made, pulsar periods have been found to be precise to nine or ten significant figures, though the periods of several have also been found to increase in a systematic way that is important in attempts to explain the emission mechanism. The periods of known pulsars range from 33 ms to 3.75 seconds, but the most frequently observed period is about 1 second. The pulsar with the shortest (33-ms) period is located within the Crab Nebula (see the introductory illustration).

The brevity of the pulses from these strange objects is as striking as their periodicity. Pulses last from 1 to 2 ms in the fastest pulsar to about 150 ms in the slowest. Fifty milliseconds is perhaps the typical pulse length. The pulse length is thus typically about 1/30 of the pulse period.

* Seymour Tilson died suddenly on December 20. An obituary may be found in this issue on page 122.

The Crab Nebula in the constellation Taurus is centered on the spot in the sky at which Oriental astronomers saw a star explode in 1054 A.D. Its interior harbors the shortest-period pulsar of the several score now known. The Crab pulsar emits powerful radio bursts at precisely periodic intervals of 33 ms. The source of these pulses may be the star marked with an arrow on this Naval Research Laboratory photograph. It pulses visibly (see Fig. 6) at the same rate. This star also may be the source of identically periodic X-ray pulses (see Fig. 8) that have been detected from the direction of the Crab Nebula. The pulsar story, now 2½ years old, has obviously just begun.

I. The box score on pulsars (January 1969)

Observatory	Designation	Right Ascension			Declination		Period, seconds	Pulse Width, ms
		hours	minutes	seconds	degrees	minutes		
CP	CP 0328	03	28	52	54	23	0.715	8
CP	CP 0808	08	08	58	74	38	1.292	30
CP	CP 0834	08	34	22	6	20	1.274	35
CP	CP 0950	09	50	29	8	11	0.253	10
CP	CP 1133	11	33	28	16	7	1.188	12
CP	CP 1919	19	19	37	21	47	1.337	16
HP	HP 1507	15	07	40	55	41	0.739	15
PSR	PSR 0833-45	08	33	39	—45	24	0.089	2
PSR	PSR 1749-28	17	49	49	—28	5	0.563	7
PSR	PSR 1929	19	29	52	10	53	0.227	10
PSR	PSR 2045-16	20	45	48	—16	37	1.962	—
PSR	PSR 2218	22	18	—	47	30	0.538	25
AP	AP 2015	20	15	45	28	31	0.558	—
AP	AP 0823	08	23	52	26	48	0.530	—
JP	JP 1933	19	33	—	16	6	0.359	—
NP	NP 0524	05	24	52	21	51	3.746	160
NP	NP 0532	05	32	—	22	—	0.033	3
MP	MP 0731-40	07	31	51	—40	—	0.375	40
MP	MP 0959-56	09	59	51	—56	—	1.438	50
MP	MP 0940-56	09	40	40	—56	—	0.662	30
MP	MP 0943	09	43	—	8	—	1.09	—
MP	MP 0835-40	08	35	34	—40	—	0.765	20
MP	MP 1426-66	14	26	35	—66	—	0.788	10
MP	MP 1451-68	14	51	33	—68	—	0.248	—
MP	MP 1727-50	17	27	50	—50	—	0.835	30
PP	PP 0943	09	43	—	8	—	1.09	—

CP = Cambridge pulsar (England); HP = Harvard pulsar; PSR and MP = Molonglo-Sydney (Australia); AP = Arecibo pulsar (Puerto Rico); JP = Jodrell Bank pulsar (England); NP = National Radio Astronomy Observatory (U.S.); PP = Pulkovo Observatory pulsar (U.S.S.R.)

Table I lists the designations, locations, periods, and pulse widths of the two dozen pulsars that had been discovered by January 1969.

Though brief, the individual pulses are often very intense. Indeed, at their peak, the strongest pulses are among the most powerful signals in the radio sky. In transmitter power they occasionally exceed the output of all the electric generators on earth by perhaps ten billion times. Observations of radio-pulse intensities—when corrected on the basis of estimates of the distances to the objects—indicate that some pulsars, at radio wavelengths alone, are as luminous, on the average, as the sun at all wavelengths. While radiating, however, their surfaces may become 10^6 to 10^{10} times brighter than the surface of the sun at all wavelengths. The equivalent brightness temperature deduced from this peak emission is about 10^{25}°K, suggesting that whatever the pulsar emission process may be, it is not merely thermal.

Since even the most powerful pulses are "on" only about 3 percent of the time, however, and the average pulse is only about one tenth as bright as the strongest pulses, the time-averaged emission coming from a pulsar, though powerful, is not especially noteworthy. Only when the time constant of the radiometer approaches the period of the pulsar is the remarkable intensity of pulses revealed. It is not surprising, therefore, that the discovery came as late—and as accidentally—as it did.

How to make great discoveries

Radio astronomers, led by Anthony Hewish at the Mullard Radio Astronomy Observatory of Cambridge University, in England, first observed signals from one of these objects in the summer of 1967. Not believing their instruments, they didn't announce their startling around the common FM broadcast band centered at about 100 MHz, though sometimes they also peak, not quite as strongly, at around 400 MHz. In all cases, power drops sharply at frequencies above 1000 MHz. The precise spectral distribution of the energy is difficult to determine with accuracy, because large intensity variations, which do not correlate at different frequencies, are always present. The shapes of individual pulses are also quite variable. Shapes range from simple, single-peaked pulses to double- and triple-peaked pulses in some pulsars. One pulsar even has a five-peaked pulse. In any given pulsar there also is a large variation in pulse shape from one pulse to the next. However, the mean pulse shape—obtained by averaging a large number of pulses—changes only very slightly over time periods of the order of months. Figures 2–5 illustrate many of these characteristic features of pulsar records.

The most rapid variations in pulsar records are the violent, unpredictable changes in pulse amplitude from one pulse to the next. These intensity variations must represent genuine variations in the intensity of emission at the sources, because they occur simultaneously over the whole frequency spectrum of the radiated signal. Superposed on these rapid-intensity fluctuations, however, there are slower fluctuations—of the order of minutes at 100 MHz to hours at 500 MHz. These slower-intensity fluctuations can be correlated with each other only over a limited range of frequencies. In addition, when any given pulsar is observed simultaneously at several frequencies, the pulse arrival time is also found to depend on the frequency. Accurate measurements on several pulsars have shown that the frequency-dependent differences in pulse arrival times can be accounted for exactly by the delay imposed by trace amounts of ionized gases (chiefly hydrogen) in interstellar space; in such a plasma a high-frequency pulse travels more rapidly than a lower-frequency pulse. The longer-period fluctuations in pulse amplitude also appear to depend on the quantity of interstellar ionized gas traversed by the signal. Thus, at the source, the pulse probably is emitted over the whole frequency range at the same instant.

Propagation or path effects may also be largely, if not exclusively, responsible for the observed polarization of pulsar signals. Pulsar radiation is typically highly polarized, of the order of 90 percent, and usually in a plane. The Faraday rotation of the plane of polarization as a function of frequency has been measured for several pulsars. These measurements have led to refined estimates of the strength (0.5 microgauss) of the galactic magnetic field, which presumably causes most of the Faraday rotation.

Not every pulsar's radiation is plane-polarized, however; pulsar CP 0328, for example (see Table I), exhibits circular polarization. The circularly polarized emission from this pulsar has a spectrum that is approximately flat in the 1-MHz band centered at 113.6 MHz. The plane-polarized emission from this pulsar, on the other hand, varies, both with frequency and from pulse to pulse. Assuming that the source has a constant Faraday rotation measure—i.e., that a single value of the rotation measure applies to the whole source—Goldstein and Meisel[1] were able to fit some of the plane-polarized spectrums observed for individual pulses to this simple model. But between successive adjacent pulses, changes in this rotation measure of as much as 30 radians per square meter were required to achieve fit to the model. Since such large and rapid changes in rotation measure can only occur in or near the source, Goldstein and Meisel believe the use of pulsar signals to determine the large-scale galactic magnetic field is premature; it must await fuller evaluation of that part of the variable Faraday rotation that may arise within the source itself. They suggest that the Faraday rotation may vary with time during a single pulse; or alternatively, if two or more Faraday rotations occur near the source, that at least one may vary with time.

The most characteristic parameter of some pulsars, at any rate, does vary with time. Months of measurements on several pulsars have revealed that their periods are lengthening with time. Secular increases in period have been observed in six pulsars. In the fastest pulsar yet discovered, the 33-ms pulsar (NP-0532), which lies at the center of the Crab Nebula, the increase in period is also the greatest yet observed, some 40 ns (nanoseconds) per day, or one part in 2400 per year. The basic period of pulsar PSR 0833-45—which, like the Crab pulsar, is also associated with a supernova-like wisp of nebulosity (known as the Vela remnant)—is 89 ms. It is the second fastest pulsar known, and its period is increasing at the second fastest rate yet observed, 10 ns per day. Pulsar CP 1133, on the other hand—which, like just about every other pulsar but the Crab and Vela objects, is not yet associated with any starlike object or extended nebular cloud—has a relatively lethargic basic period of 1.19 seconds. Its period is increasing by a correspondingly lethargic amount—0.3 ns per day.

The rule seems to be: the faster a pulsar beats, the faster its beat slows down. At least that seemed to be the rule until early last spring, when the Vela pulsar temporarily accelerated. For just one week, between February 24 and March 3, this pulsar reversed the general rule; its period decreased by 134 ns, then it resumed its steady increase.

So much for the essential radio facts about pulsars. Now, what do they mean?

The race to determine what pulsars might be

Although the periodicity of pulsars was their most striking characteristic, most people attempting to explain this new phenomenon were equally intrigued by the short pulse durations—typically 50 ms. These indicated that the sources were small, since no source of radiation, no known physical mechanism, can turn on and off in less time than it takes light to cross the source. A 50-ms pulse suggested, for example, that the source must be no more than about 15 000 km in size. And many pulsars pulsed far more briefly than this. No ordinary star could be this small. Radio flares coming from sunspots on ordinary stars could perhaps account for the signals, but no radio bursts emitted by our sun were ever so powerful, or so precisely periodic. And in any case, a sunlike star with disturbed regions of these indicated dimensions should have been visible optically, but until recently no visible object has been associated with a pulsar.

Planets, white-dwarf stars, and neutron stars were the only objects that seemed to fit the dimensional criterions. The race to solve the pulsar puzzle thus became a contest to eliminate two of these three possibilities.

Planets and white-dwarf stars were well-known objects, and both were known to be capable of emitting radio signals. Just a few years ago, in fact, the giant planet Jupiter was discovered to be the source of strong beamlike radio signals that are emitted at reasonably periodic intervals.

Neutron stars, at this stage of the game, were the dark horse in the three-way race. They were tiny, superdense theoretical objects, made of matter in highly condensed forms that only a few physicists had ever thought about. These strange stars envisioned by the theoreticians were made almost entirely of neutrons, the heavy, electrically neutral particles that—together with almost equally heavy but positively charged protons—comprise the nuclei of all atoms but those of ordinary hydrogen, whose nucleus is a lone proton.

A neutron star would come into existence, according to theory, when the interior of an old massive star collapsed, packing a mass equivalent to that of the sun into a small sphere 10 km in diameter. As the old star collapsed, its central density would become 10^{14} times the density of water. To achieve that enormous density, electrons must be driven into protons to form neutrons and the neutrons in turn must be packed so closely that the outer half of the star would be very unstarlike; instead of a fiery gas it would resemble a pseudocrystalline substance—a solid, with a strength many billion times that of steel. (Recent theoretical work suggests, however, that the central part of a neutron star may be fluid.)

Such an extraordinary star could embody a magnetic field a trillion (10^{12}) or more times stronger than the earth's 1-gauss field in a body a thousand times smaller. (Such a field would be 100 000 times stronger than the strongest field yet produced in the laboratory.)

Narrowing the field of choice

Although the small sizes deduced for the emitting objects were consistent with the sizes of planets, it appeared very unlikely that the prevailing increases in pulsar periods were caused by the revolution of a planet circling a distant sun. The observed increases in period were so small, and so obviously related somehow to the pulsars' basic periods, that nearly everyone agreed they must be intrinsic to either the rotation or the mechanical pulsation of the emitting object itself. Although this reasoning did not immediately eliminate planets as a possibility, it did focus most attention on the two remaining choices.

The apparent boundary for choosing between a neutron star and a white-dwarf star is a period of about one second. White dwarfs, such as the well-known companion to the bright star Sirius, can peak as much mass as there is in the sun into a diameter no greater than several hundred kilometers. And, from considerations of their temperature, angular momentum, and stability against disruptive effects, they therefore can, except under quite unusual circumstances, expand and contract or rotate as often as once a second, but no faster. The much smaller (perhaps 100 times smaller) size and much higher (perhaps 100 000 times higher) density of a neutron star, however, would allow it to expand and contract from 10 to 10 000 times a second. It could also rotate much more rapidly than a white-dwarf star, or any planetary body. Held together by its enormous gravitational force, a neutron star could rotate as rapidly as 10 000 times a second, if required to by observations, before centrifugal forces would cause it to fly apart.

If the only periodic events detected in pulsars always had periods of a second or longer, pulsars would almost certainly be white-dwarf stars; but since periods much shorter than this, especially periods in the range of a few thousandths of a second, were common, it was clear that some pulsars, at least, might be neutron stars. Still, one could not be absolutely sure that this was so on this basis alone.

The next set of observations narrowed the choice, but still not conclusively. When they recorded individual pulses with special electronic equipment that could detect changes in amplitude in times as short as 0.1 ms, radio astronomers at Cal Tech and Arecibo observed that a remarkable transformation occurred in each pulse. A single radio pulse—thought to last 50 ms, for example—turned out to be comprised of an apparently random ensemble of many, much shorter bursts. In some cases, these individual bursts lasted no more than 0.1 ms.

Since an object cannot change its emission in less time than it takes light to cross it, these superbrief bursts mean that the radiating regions were no more than 30 km across. This, plus the power requirements and the observed increases in period, effectively eliminated the planet possibility. It also is smaller than a white-dwarf star's diameter, but the possibility that these regions were bright radio flares on a spinning white dwarf could not yet be discounted.

Clinching the case for neutron stars

Subsequent studies of consecutive individual pulse shapes seem to have eliminated the white-dwarf possibility, however, and perhaps clinched the case for neutron stars. Carried out by radio astronomers at Arecibo, these important studies have shown that, in at least two pulsars, the complex observed pulse structure results largely from the combination and superimposition of two very fast pulsation events with markedly different periods. One set consists of the striking, brief pulsations that were first discerned in these pulsars; these pulses occur at roughly one-second intervals and are further characterized by the stable pulse shape or envelope one obtains by taking the mean of many pulses. The second set of pulses is much more rapid; periods of the order of 10 ms characterize this set. The faster pulsations show phase variations with respect to the slower pulsations of as much as one-half cycle in times of the order of 25 seconds. The march of the faster pulses through the slower pulses may account for the apparently chaotic progression of pulse shapes observed in the typical short-time-constant pulsar records.

The chaos may be only apparent, however. Actually, the faster pulses may be marching through the slower pulses in a remarkably regular way. If so, the steady march of the faster subpulses through the main pulses may mean that the timing mechanism of the subpulses is as precisely controlled as that of the main pulses. It also may mean that the mechanism responsible for the shorter-period emission is completely different from the mechanism responsible for the main pulses. Although the mechanism responsible for the main (1-second-average) pulse could be tied to the mechanical rotation or pulsation of either a neutron star or a white-dwarf star, the 1–15-ms marching speed of the subpulses through the main pulses—in the two pulsars for which it has been measured—can be related only to the mechanical

The first visible pulsar . . . Now you see it, now you don't

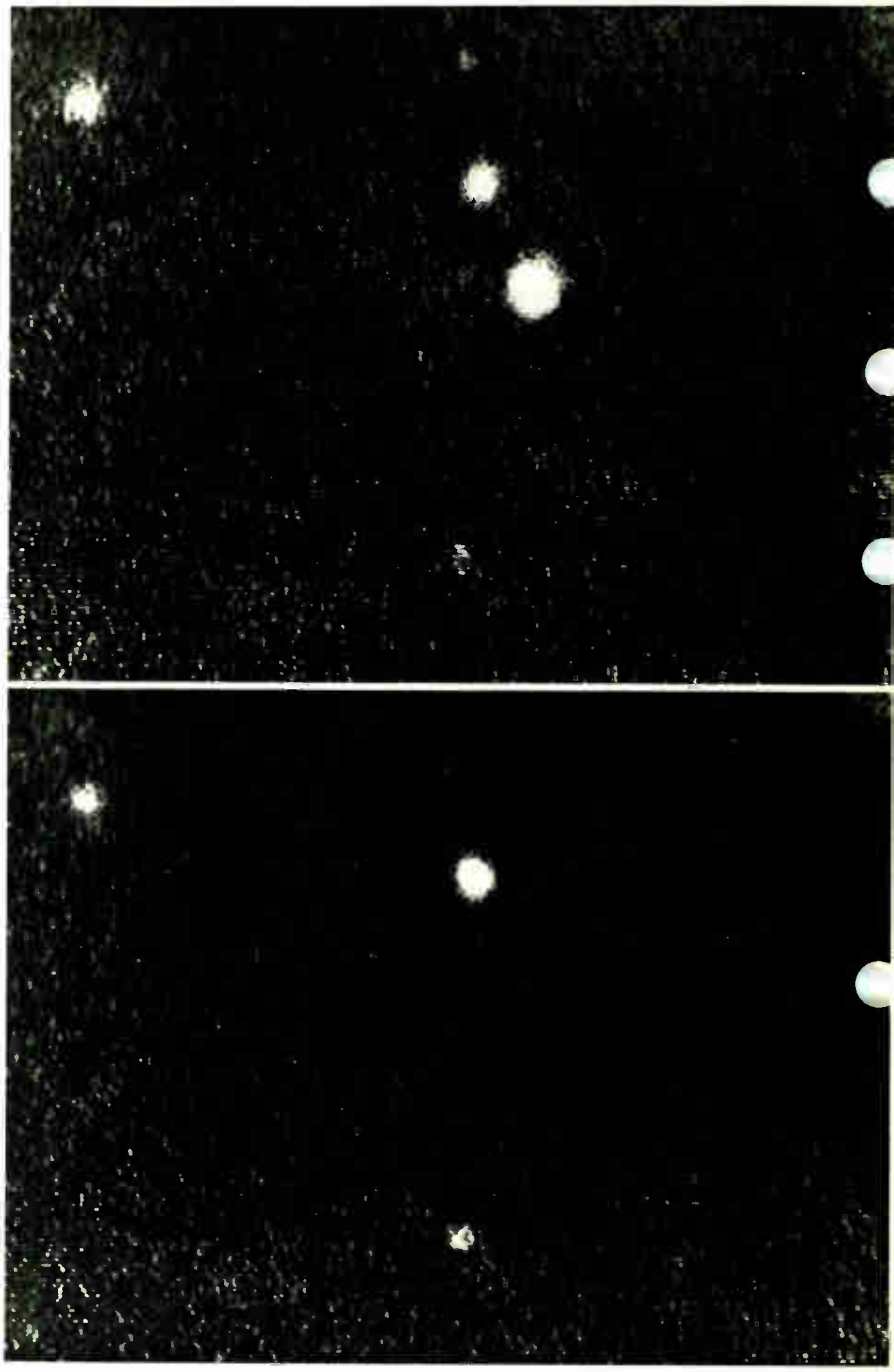


FIGURE 6. These pictures of a star that pulsates visibly at the same rate at which it emits radio signals are surely among the most extraordinary astronomical photographs ever made. E. J. Wampler and J. S. Miller made these by focusing the Lick Observatory's 120-inch telescope on the lower of the two prominent central stars in the Crab Nebula (the star marked with an arrow in the introductory photograph). Its peculiar optical spectrum, as well as its location, suggested it might be the source of 33-ms radio pulses coming from Crab. By interrupting the apparently steady light coming from this star at the same rate—30 times a second—but not quite in phase with the radio pulses, Wampler and Miller caused their stroboscopically generated light flashes to fall out of phase with the actual light flashes coming from the star—i.e., the star was obscured for part of each observing cycle. Viewed on a television screen, the star appears to be pulsing considerably more slowly than it actually does.

pulsation or rotation of one astronomical object—a neutron star.

If these two, and perhaps most other, pulsars are in fact neutron stars, are they rotating or pulsating? There is a simple test of these two possibilities. As a pulsating star gives up energy, it should become cooler and more dense. It therefore should pulse more rapidly. If, on the other hand, the star is rotating, as it loses energy it will spin more slowly. A lengthening period, as already mentioned, was the general rule with pulsars. The pulses therefore appear to be best explained by associating the emission mechanism, whatever it may be, with the axial rotation of tiny neutron stars.

But what about the week-long decrease in period manifested last spring by the contrary pulsar in the Vela supernova remnant?

With only this one fact, the experts can but guess at what may have happened. Some think that a planet may have slammed into the pulsar; others think the Vela pulsar may have suffered something like an earthquake. On the theoretical assumption that at least the outer, cooler, layers of neutron stars may resemble gigantic pseudocrystals made of neutrons, as these surface layers cool, it is conceivable that they can shrink, perhaps suddenly. When a neutron star shrinks it will spin faster in order to conserve angular momentum and its period will decrease. To explain the Vela decrease, a sudden shrinkage of 1 cm would suffice.

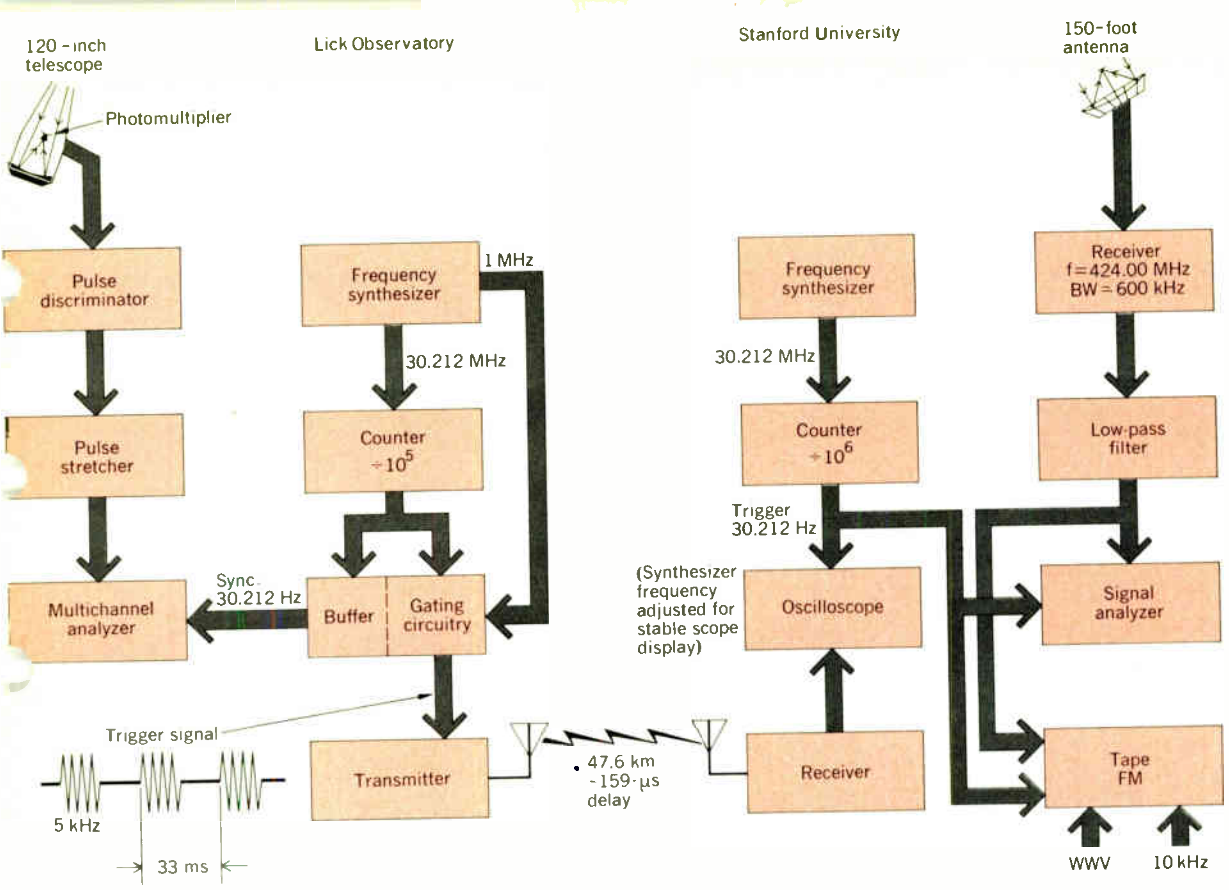


FIGURE 7. Diagram shows key elements of system used by Lick-Stanford team simultaneously to monitor visible and radio pulses from the 33-ms Crab pulsar. Purpose was to determine whether two signals that have identical periods (see Fig. 6) are also exactly in phase. They are not. Radio pulses lag light pulses by about 1.5 ms. Lag may be explained by greater delay imposed on radio signals by interstellar electrons, but source effects cannot be entirely ruled out. (C. Taubman described this experiment in an article in the "Hewlett-Packard Journal," June 1969, from which this diagram was derived.)

One pulsar, so far, can be seen as well as heard

Neutron stars are so small that no one ever expected to hear one, much less see one, unless it was improbably close. The observational story took on an exciting new dimension just a year ago, however, when astronomer John Cocke and his colleagues at the University of Arizona's Steward Observatory discovered that one pulsar, which had previously been tentatively identified with the lower of the two ordinary-looking central stars in the Crab Nebula, was also pulsing at visible-light frequencies.

Figure 6, surely one of the most wonderful astronomical photographs ever taken (this one was obtained by E. J. Wampler and J. S. Miller at the Lick Observatory), is proof that one pulsar can be seen as well as heard.

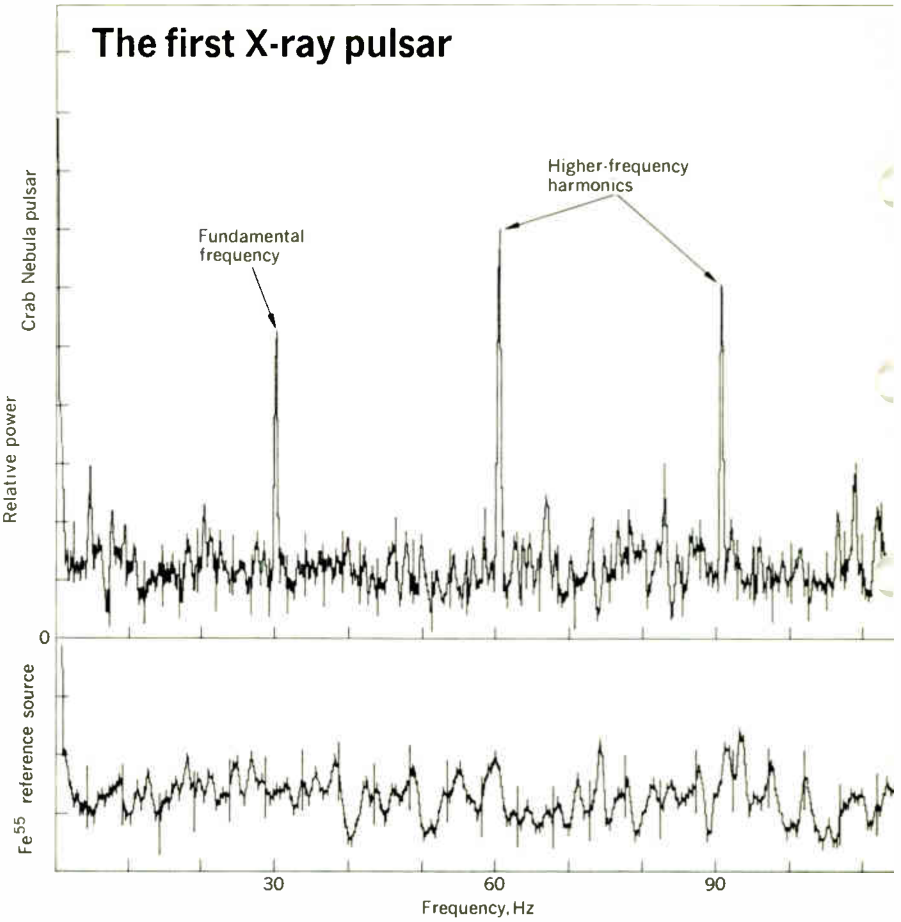
The Crab Nebula has received great attention from astronomers, not only because of its distinctive appearance and relative proximity to us—it is only about 6600 light-years away—but because it is now thought to be the optically visible wreckage of a supernova that Oriental court astronomers saw occur in 1054 A.D. In the 900-some-odd years since this star exploded, its gaseous debris has expanded to fill a volume of some 10^{40} cubic miles. Already about 10 light-years in diameter, this colorful gaseous envelope is still expanding at the extraordinary speed of 1300 km/s and radiating energy at all frequencies—optical, radio, infrared, and X ray—at the enormous rate of some 10^{31} watts. It would take the radiation from 100 000 stars like the sun to match this power output.

Pulsar *aficionados* have also been greatly intrigued by the Crab Nebula because its interior was soon discovered to be the abode of the fastest known pulsar. The 33-ms Crab pulsar "turns on" (to use the currently fashionable phrase) about 30 times a second. Now the Crab pulsar has been associated with an optically visible object—and it blinks at precisely the same interval.

The optical flashes closely mimic the radio flashes. In each 33-ms cycle there is a major pulse followed about 14 ms later by a minor pulse. Moreover, nearly all the light emitted by the pulsar seems to be confined to the flashes; if there is a steady-state component, it does not account for more than 6 percent of the total light emission. This strongly suggests that the source is not the entire surface of an object but more nearly some sort of rotating beacon.

Subsequent work on the Crab pulsar, by scientists at the University of California's Lick Observatory and Stanford University, also seems to support a rotating-beacon theory of pulsars. In this work, early last spring,

FIGURE 8. **Herbert Friedman and colleagues at the U.S. Naval Research Laboratory derived these revealing power spectra from rocket-borne observations of 3- to 15-Å X rays. Radiation coming from the Crab Nebula (upper curve) was calibrated against radiation coming from an onboard Fe^{55} reference source. Relative power (abscissa) is proportional to the square of the signal amplitude in one frequency interval. Fundamental X-ray pulse frequency from the Crab is 30.2 Hz, same as the radio and optical Crab pulsar frequencies. X-ray flux, however, is about 10^2 times more intense than the optical flux and about 10^4 times more intense than the radio flux. Relatively weaker radio signals may be generated by proton–synchrotron radiation, whereas electron–synchrotron radiation may be responsible for the more highly energetic visible and X-ray pulses.**



radio and optical signals from the Crab pulsar were monitored simultaneously for the first time, using the elaborate experimental setup shown in Fig. 7. The purpose was to measure the difference in arrival time between the optical and radio signals, if any, to see whether these identically periodic signals were being emitted at the same time. Allowing for the delay imposed by the interstellar electron plasma, the radio pulses arrived approximately 1.5 ms later than the light pulses. Although this phase difference can be completely accounted for by assuming a higher value for the interstellar electron density, thus imposing a greater delay on the radio waves, it could, in part, also be a source effect.

That's pretty much where the observational part of the pulsar story stood until last March. On March 13, on behalf of Herbert Friedman and his rocket astronomy and aeronomy group at the U.S. Naval Research Laboratory, NASA launched an Aerobee sounding rocket. It carried X-ray detectors to an altitude above the obscuring effects of the earth's atmosphere. An attitude control system then pointed the 3–15-Å X-ray detectors toward the Crab Nebula for 40 precious seconds. Guess what they found!

The first X-ray pulsar

The NRL group observed X-ray pulsations coming from the Crab Nebula.[2] The X-ray pulsation period closely matched the 33-ms beat of the radio and optical pulsations from the Crab pulsar. Figure 8 shows the power spectrum of X-ray pulses from the Crab. The lowest of the three prominent peaks, the one on the left, represents the fundamental 30.2-Hz beat of the X-ray pulsar. The second and third, slightly higher, peaks on the right are the next higher harmonics of this fundamental frequency.

Just as the Crab pulsar's optical flashes closely mimic its radio flashes, so the X-ray pulses mimic both, although the X-ray interpulse structure is slightly different. Although the pulsed X-ray flux amounts to only about 5 percent of the integrated steady-state flux of X rays (3 to 15 Å) coming from the entire Crab Nebula, the X-ray pulses are nevertheless about 200 times more powerful than the optical pulses and about 20 000 times more powerful than the radio pulses. In the time interval of a single pulse—about 1/30 second—the Crab pulsar pours out as much energy in X rays alone as our sun emits at all wavelengths over a period of 10 seconds.

Shortly after this discovery of the first—and, at the moment, still the only—X-ray pulsar, a team of astronomers from M.I.T., McDonald Observatory, and Mt. Palomar Observatory carried the Crab drama a step further. The M.I.T. team confirmed the NRL group's observations, and showed that the pulses in all three spectral ranges—radio, visible, X ray—coincided not only in period but in time of origin.

What sort of emission mechanism can account for this broadband radiation from the Crab pulsar, and at the same time account for the radio signals that alone testify to the existence of the 100 or so other pulsars discovered to date?

'he emission is too powerful to be thermal

One important clue to the emission mechanism came rom the observed equivalent brightness temperature of ulsars—10^{25}°K, while radiating, as you may recall. 'hat's a pretty high temperature. If in fact it were a hermal effect, most of the radiation coming from a pulsar ould be in the form of gamma rays, not radio waves, nd all pulsars—not just the Crab pulsar—would emit trong light flashes and X rays.

The very intense, extremely short-lived radio brightness f the pulsars—from which this high equivalent brightess temperature was calculated—indicates, on the conary, that in pulsars astronomers are observing the emison of energy from a process that involves only relavely modest amounts of heat. Heat-producing processes ield disorderly, chaotic, inefficient, and therefore, in ie end, less intensely luminous results, whereas the ulsar energy-production process, as indicated by the igh equivalent brightness temperature, is so powerful could only be the product of an orderly, extremely ficient process.

How could so small an object as a neutron star proide powerful pulses of highly polarized radio waves, s well—in the case of the Crab pulsar, at least—as still iore powerful pulses of visible light and X rays? This ould seem to require a Goliath among the stars, yet eutron stars are tiny; they are more appropriately decribed, metaphorically, as "little Davids." The answer this apparent paradox may lie in the fact that the acuracy of the latter image may extend even to the slingiot.

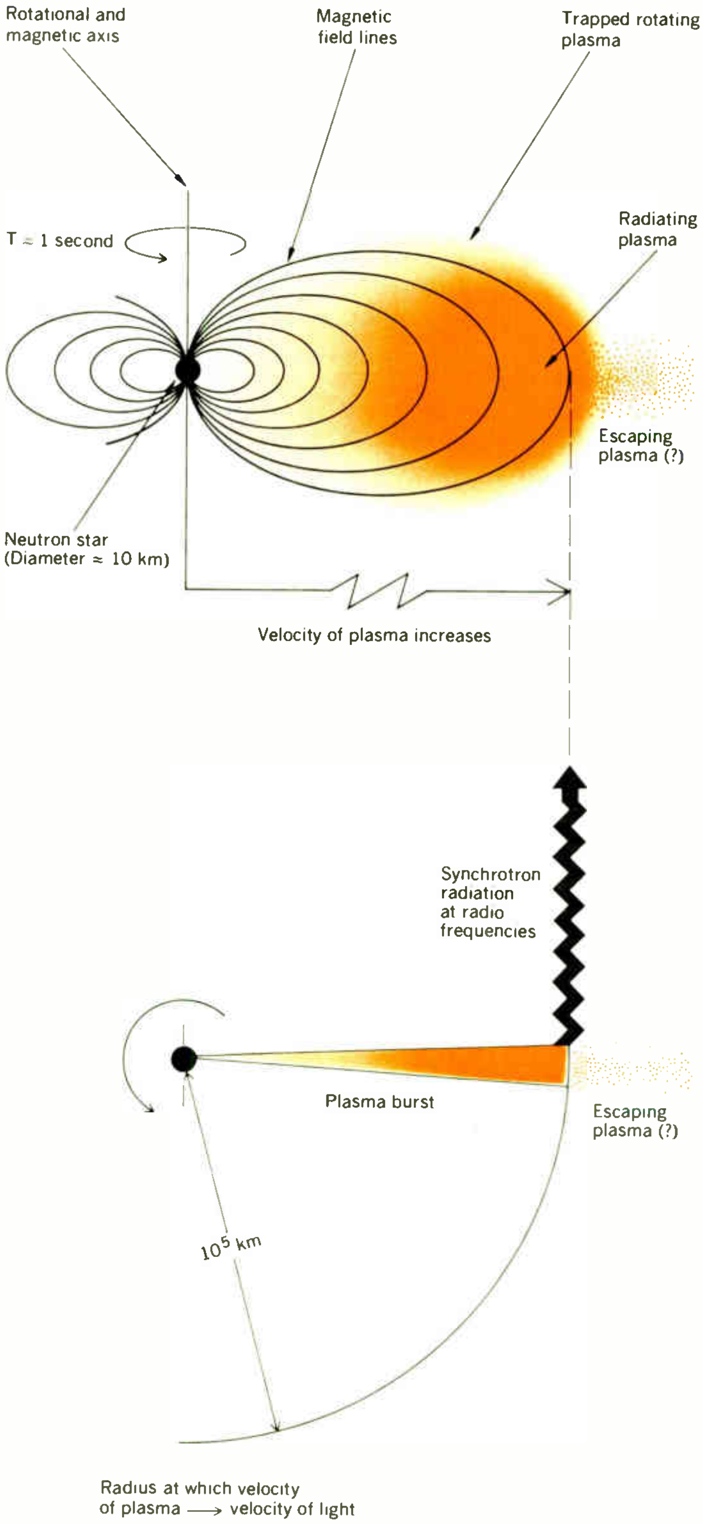
ittle David was small, but oh my!

The "sling" used by these energetic little stars to overme the barriers to our perception of them may consist the lines of force of the intense (10^{12} gauss) magnetic eld that theory suggests they should have. A powerful agnetic field is an integral part of every theory that eks to explain pulsars. The magnetic field may grow strong as the middling-strength magnetic field of an dinary star, whose death throes spawn a neutron star, tensifies as "frozen-in" magnetic field lines converge er more closely during the collapse of the ordinary star om normal dimensions to the dimensions of a neutron ar.

The "shot" whirled about by this magnetic field, to locities approaching the velocity of light, may consist some of the ordinary, charged, subatomic particles—otons and electrons—that remain in a neutron star d comprise perhaps one percent of its mass. Leaking it of some kind of "hot" spot on the surface of the star, it constrained by their electric charges to spiral outird from the surface only along magnetic field lines, ese particles rotating with the rapidly rotating star ay, at sufficient distances from the star's surface, at- ı velocities high enough to cause them to emit elecomagnetic radiation. The emission process may be akin that observed in the particle accelerators used in highergy physics experiments. In these experiments, the -called synchrotron radiation that is emitted results om the acceleration of charged particles to extreme 'ocities by, and in, an intense magnetic field.

f various sectors of a neutron star's magnetosphere e fed different quantities of charged particles, or if ferent sectors are made to radiate in slightly different

FIGURE 9. Diagram portrays essential elements of theory advanced by Thomas Gold of Cornell University to explain pulsars. Plasma particles, leaking from "hot" spot on a rapidly rotating neutron star, spiral outward along magnetic field lines, which rotate with star. At sufficiently great distance from the star's rotational axis, velocities of orbiting particles approach the velocity of light, at which point radio-frequency pulses may be generated by the synchrotron process. Plasma escaping from the neutron star's powerful (10^{12} to 10^{13} gauss) magnetic field may feed and energize an expanding gaseous nebula such as the Crab Nebula shown in the introductory photograph, while the neutron star itself may be born in the same star explosion that creates the nebula. The leading alternative theory is shown in Fig. 10.


Rotational and magnetic axis
Magnetic field lines
Trapped rotating plasma
T ≈ 1 second
Radiating plasma
Escaping plasma (?)
Neutron star (Diameter ≈ 10 km)
Velocity of plasma increases
Synchrotron radiation at radio frequencies
Plasma burst
Escaping plasma (?)
10^5 km
Radius at which velocity of plasma ⟶ velocity of light

directions, the short, intense radio bursts typical of pulsars will occur.

At the outer fringes of the neutron star's magnetosphere, where the field intensity has perhaps diminished sufficiently, some of the proton–electron plasma may even be able to escape, to fill and energize an expanding nebular cloud such as the Crab and to contribute to the galactic cosmic-ray flux. The latter suggestion depends, in part, on the assumption that a neutron star may be created each time a supernova occurs in the Milky Way—perhaps once every hundred years.

These, at any rate, are the essential elements of one leading theory—the "slingshot" or "snap-the-whip" theory illustrated in Fig. 9—that Thomas Gold of Cornell University advanced within a few months after their discovery to explain the strange radio pulsations from the objects that their discoverers had named pulsars.

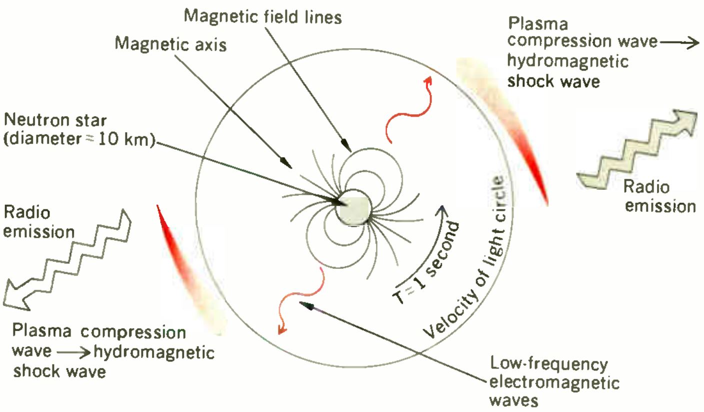
The Gunn–Ostriker theory

At this writing, one other theory also appears to have survived the flood of new pulsar observations, a theory advanced by James Gunn and Jeremiah Ostriker of Princeton University.

The Gunn–Ostriker theory is a more fully developed version of an idea, shown in Fig. 10, that was originally proposed by F. Pacini, now of Cornell University. In this theory, the pulsar is also postulated to have a magnetosphere with a very strong magnetic field, which rotates with the star. However, the axis of the magnetic field does not coincide with the star's rotational axis. It is tilted. This magnetic dipole thus rotates at the rotation frequency of the star, and as a result electromagnetic waves are generated, in the classical fashion, at the same frequency. These electromagnetic waves are low-frequency waves. They cannot be expected to escape from any reasonably dense cloud of charged particles that may surround the star. More exactly put, since it is almost certain that the frequency of these waves generated by the rotating magnetic dipole is below the plasma frequency of any atmosphere surrounding the star, the cannot propagate. But the radiation pressure they can exert may be considerable, and this pressure may give rise to periodic compressions of the atmosphere. Compression waves traveling outward through the atmosphere may develop into hydromagnetic shock waves at sufficiently great distances from the star, and these plasma disturbances may be the source of the wide-band emissions at the higher, radio frequencies that characterize pulsars.

In effect, the original low-frequency electromagnetic waves emitted by the rotating magnetic dipole may catch up the charged particles surrounding the star and propel them outward through the star's magnetosphere, like surfboards on an ocean wave. And, since even low-frequency electromagnetic waves travel at the speed of light, they may propel successive groups of outwardly moving charged particles to nearly the speed of light, at which point, as in Gold's theory, successive groups of particles may generate powerful bursts of radio-frequency radiation, pulse after pulse after pulse.

FIGURE 10. Pulsar model developed by James Gunn and Jeremiah Ostriker of Princeton University is a more detailed version of ideas shown diagrammatically here, which were originally suggested by F. Pacini, now of Cornell University. Magnetic and rotational axes of neutron star do not coincide in this theory, as they do in the leading alternative theory (Fig. 9). Magnetic dipole thus rotates at rotation frequency of star, generating low-frequency electromagnetic waves. These cannot propagate through the plasma surrounding star, but may exert sufficient radiation pressure to compress plasma periodically, at about star's rotation rate. Outward-traveling compression waves may in turn develop into hydromagnetic shock waves where plasma approaches velocity of light, and these shock waves may yield the wide-band emission at MHz-range radio frequencies characteristic of pulsars.



Two kinds of synchrotron emission—maybe

These surviving theories, and many others that now appear to have fallen by the wayside, were advanced originally to explain only the amazingly periodic radio pulses. Neither is vitiated by last spring's spectacular series of observations on the Crab Nebula. Friedman and his colleagues believe, however, that the radio emission mechanism—in the Crab pulsar, at least—differs from the mechanism producing the more powerful optical and X-ray pulses. Although synchrotron radiation can explain all three kinds of radiation, they think that two kinds of synchrotron radiation may be involved. A proton-accelerating synchrotron process—operating at distances from the stellar surface where the magnetic field is still of the order of several hundred gauss—may explain the radio emission. But in order to explain the considerably more powerful optical and X-ray pulses, they call upon the higher-energy radiation that can be generated by electron acceleration. In a magnetic field of any given strength, electrons of course can be more highly accelerated than protons traveling at the same velocity, and thereby emit higher-energy radiation by the synchrotron mechanism, because they are only about 1/1800 as heavy as protons.

The key thing that almost everyone seems to agree on at this stage of the pulsar game is that rotating, highly magnetic, neutron stars lie at the heart of the matter.

Strangely enough, these objects that were mere theoretical curiosities for over 25 years, until the discovery of pulsars tentatively confirmed their existence, may be one of the few truly stable astrophysical objects in the universe. Along with such similarly diminutive objects as the planet earth and white-dwarf stars, they may be the only objects forever capable of resisting their own gravitational forces. Relativity theory predicts that all other objects must sooner or later succumb to the crushing forces set up by their own mass, and collapse catasophically to a singularity, the point where their volume ecomes zero and their density infinite, and where the nysical space around them may be "pinched off," perups banishing them from our view.

Neutron stars are thus the perhaps no-longer-missing nk connecting a wealth of observational and theoretical cts. Neutron stars, certain pulsing visible and X-ray ars, radio stars; exploding stars and the nebular clouds at succeed them; cosmic rays; and long-held theories oout the gravitational collapse of matter and the ultiate implications of general relativity—all of these n now begin to be woven into a fabric whose intricate ttern reveals some of the elements of creation.

he extraordinary death of an ordinary star

To begin weaving this fabric we must go back to the edside of a conventional "dying" star, a star very much ce our sun. But we cannot pick just any sunlike star. he prescriptions of the theoretical astrophysicists limit to stars not too much heavier than perhaps 1.5 or 2 olar masses. Stars less massive than this simply will not o, not if we wish to end up with a neutron star. Smaller ars are more likely to end up as white-dwarf stars, not eutron stars, while very much more massive stars—range as it may seem—are doomed to crush themselves ut of ordinary existence.

Stars of perhaps 1.5 or 2 solar masses, having finally khausted the nuclear fuels in their interiors in the maniold fusion-reaction chains that powered them for bilons of years, do in fact begin to darken and die. The toms comprising the outer portions of these dying stars re no longer sustained in frantic motion by the intense eat coming from nuclear reactions in the star's interior. hey begin to contract inward, toward the star's center mass, as gravitational forces continue to exert their nexorable pull. Then strange things start to happen, and new physics of matter and energy, foreign to everyhing we are familiar with on earth, begins to unfold najestically from the theoretician's esoteric equations.

As the dying star contracts, the material in its core is queezed into an ever-smaller volume. As a direct conequence, the gravitationally induced pressure on the core naterial increases correspondingly. This causes the star o contract still further, of course, and the pressure again ncreases correspondingly, and so it goes, on and on and n . . . but not quite ad infinitum. The sheer squashing ogether of atoms in the star's tortured interior finally trips the orbital electrons in the atoms free of their nt nuclei. The outward pressure exerted by this nergetic cloud of so-called Fermi electrons may slow he star's contraction, but contraction eventually proeeds, to a critical point at which the pressure on the napless star becomes so great that these free electrons egin to combine with positively charged protons in the tar core's atomic nuclei. Greater and greater numbers eutrons are produced by this process, which physiists call "inverse beta decay" to distinguish it from the ar more familiar process of beta decay in which neurons split into electrons and protons.

At this point, the star no longer merely contracts; it egins to collapse catastrophically. It doesn't fall apart, xplode; it falls together, or implodes. In one sense, astrophysicist's, what happens next is all over in a ew seconds. In other senses—the radio astronomer's, he X-ray astronomer's, and the optical astronomer's, as we have seen—it is just the beginning of the story.

Catastrophic gravitational collapse is initiated by inverse beta decay when the star, which has been contracting relatively slowly until then, reaches a central density of perhaps 10^9 to 10^{11} g/cm^3. At such densities masses as great as the sun wouldn't take up much more space than earthlike planets or white-dwarf stars. But even this highly condensed matter is frothy stuff compared with the dense material that remains when the catastrophic implosion has run its course.

Within a fraction of a second after collapse begins, nearly all the electrons and protons inside the star's core have been transformed into neutrons, and the core itself is in free fall, toward a fantastic configuration in which the space between these neutrons, in only several additional seconds, will be reduced another thousand to ten thousandfold, to yield densities of the order of 10^{14} g/cm^3. At this density all the human beings on earth would fit into the volume of a single raindrop.

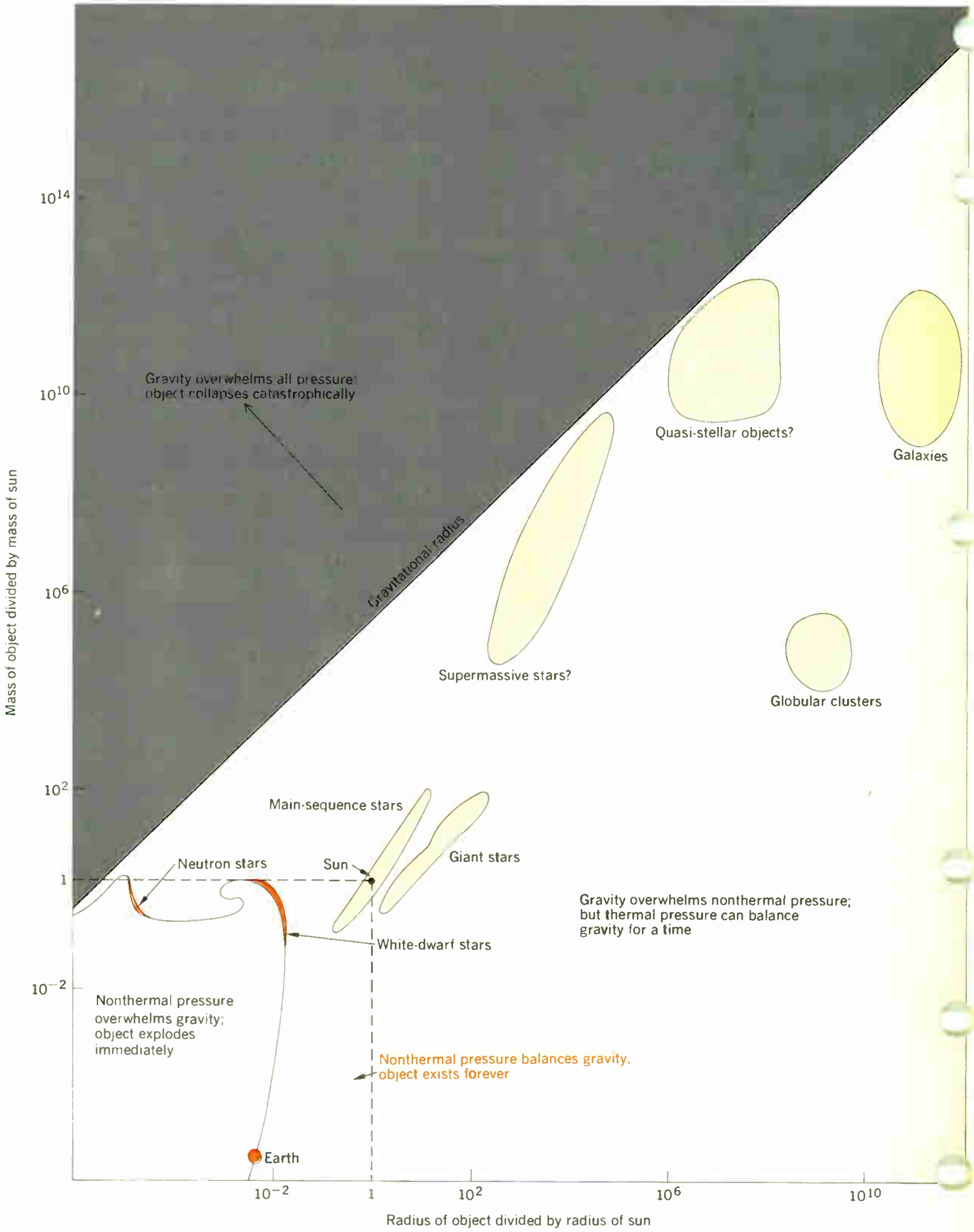
The resulting confining pressure is so intense that the neutrons (and the very few remaining protons and other nuclear particles) begin to cozy up to each other, in spite of the enormous repulsive forces tending to drive them apart. This formidable pressure wave brings the collapse of the core to a halt. At the same time, it sends a shock wave of equally formidable proportions propagating outward through the core. In the core shock front, the huge kinetic energy of collapse is converted instantaneously into heat. Temperatures of over 10 billion degrees are reached. At such high temperatures and densities, however, the elementary particle transformations so dear to the hearts of physicists proceed at a terrific rate, and the heat behind the core shock front, in the interior of the star, is converted with equal rapidity into high-energy neutrinos.

Neutrinos really heat things up

Neutrinos should not be confused with neutrons. Although neutrinos are like neutrons in that they also are electrically uncharged, neutrinos are much smaller and lighter. They are so small and so light, in fact, that they behave as though they have no rest mass. Since they always travel at the velocity of light, however, they do exhibit mass, but their interactions with the more conventionally massive kinds of matter are so feeble they are hard to detect. Interactions between neutrinos and other forms of matter are among the weakest in the universe. Because of this, neutrinos, which fly about the universe in huge numbers, pass pretty much unhindered and unobserved through nearly everything. Their mean free paths are measured on the scale of light-years.

But under the conditions of extreme density in the core of a collapsing star, even the mean free path of these agile neutrinos is scaled down. It is reduced to such an extraordinary degree that the reduction may help you to comprehend in a still more tactile way how extreme these core densities are. Instead of zipping freely away, unhindered as usual, for distances of thousands of light-years, the neutrinos trying to escape from the core of the star move less than 100 meters. Thus constrained, they deposit the energy they contain—all the energy released by the core's collapse—into the outer envelope of the star. This of course raises the temperature of the by-now relatively cool envelope; it may raise the envelope temperature to as much as 200 billion degrees.

FIGURE 11. This mass–radius diagram (after Kip Thorne in "Scientific American") helps put the astrophysical objects that are prime pulsar candidates—neutron stars and white-dwarf stars—into fundamental physical perspective. These objects, like all others known to exist, lie in the area to the right of the limiting mass–radius ratio known as the gravitational radius. An object's destiny is determined by the outcome of the confict between gravitational forces and outwardly directed pressures that resist them. Thermal pressure (of atoms in motion within stars, or of stars in motion within galaxies) can resist gravity for a while, but ultimately thermal energy is dissipated and gravity forces objects to contract. If an object is of the right mass to begin with, its contraction may terminate in the formation of such uniquely stable objects as neutron stars, white dwarfs, and even earthlike planets. In such objects, nonthermal pressure (of electrons in atoms, or of particles in atomic nuclei) may balance gravity forever. If contraction led an object into region at lower left, however, an explosion would destroy it immediately; hence no objects lying this region have been observed. To upper left, beyond the gravitational radius, lies oblivion, but not nonexistence. An object carried into this region by contraction would collapse catastrophically, but probably unobservably, to zero volume and infinite density.



At such an enormous temperature a new, brief phase f explosive nuclear burning is initiated in the star's uter envelopes. Nuclear reactions that yield heat and ght, and that synthesize new elements and new isotopes f preexisting elements, begin to occur once more. The eat released by these nuclear reactions, when added to e already huge thermal energies generated by neutrino eposition, finally creates an outwardly directed counterorce to the gravitational force that has been calling he tune—clearly a rather lively dirge—until now. The outer envelopes of the star then suddenly become gravitationally unbound from the core of the star.

star dies, a star is born

A thermal shock wave forms, and this explosion blows he star's envelope away from its core at speeds approaching the speed of light. The huge thermal energy of the expanding envelope is thereby converted into radiation so intense that the visible light coming from the exploding star is almost as bright as that which comes from an entire galaxy of 100 billion ordinary stars. This brilliant burst that brings the visibly radiant life of the star to an end with a bang is, of course, a supernova.

It's the biggest bang definitely known to occur in the universe. Still bigger bangs, like the one with which some theorists believe the universe itself began, or the ones with which the gravitationally connected star systems called galaxies themselves may collapse en masse (as in quasars, perhaps), are subjects of speculation and dispute. A supernova is an observational fact.

Viewed from a distance, from the earth for example, if the explosion occurs in our galaxy, this metamorphosing star, heretofore perhaps an utterly insignificant speck in the sky, suddenly shines both night and day, for a period a few weeks. Its brilliance exceeds that of the brightest starlike objects in the nighttime sky, the planets Venus and Jupiter. If the mass of material in the outer layers of an exploding star is about equal to the mass of our sun, as it would be if the original star were about twice the solar mass, the energy released, per second, in the explosion is comparable to the energy output of our sun over a billion years. Small wonder, then, that the supernova shines day and night for two or three weeks with the brilliance of billions of suns. Then its brilliance rapidly wanes, to be replaced, eventually, by the more diffuse but longer-lived luminosity of that expanding cloud of gas known as a nebula, the ghostly remnant—the telescopically visible wreckage— of the supernova.

ut what of the star core that is left behind? It has n crunched by the original gravitational collapse of the star; it has been buffeted by the pressure-induced shock wave propagating outward through it from its central region; and it has been counterbuffeted by the blast of the supernova explosion that has blown its enveloping layers into space at nearly the speed of light. npletely transformed by the cataclysm that has blown away perhaps half of its original substance, the star that gives birth to a nebula nevertheless lives on.

This star core, if its mass is no greater than twice that of the sun, and if other rather strict conditions are also met, survives as a neutron star.

ting neutron stars into cosmic perspective

The mass-radius relationship of neutron stars to other celestial objects such as the earth, the sun, galaxies, quasars, and white-dwarf stars, is shown in Fig. 11.

The neutrons are packed together so closely that they may actually touch each other, in a pseudocrystalline array. A neutron star can thus be said to resemble a single, gigantic atomic nucleus—of somewhat unusual composition, to be sure. But this crude analog must not be taken too seriously: neutron stars are held together not by ordinary nuclear binding forces but by their own immense gravitational force.

Strange as they are, neutron stars still obey the fundamental laws of physics. One law they obey is the law of conservation of angular momentum ($\omega = mvr$). Since a neutron star is so small, it must spin much faster than the star from which it came in order to conserve angular momentum, even allowing for the angular momentum that may be carried off in the supernova. Evidence derived principally from the Crab Nebula pulsar suggests that when they are first born neutron stars may spin as fast as 50 times per second.

Thomas Gold deduced this initial rotation speed by starting with the observed period of the Crab pulsar and working backward in time to the time of its origin in the supernova of 1054 A.D. To do this he simply reversed its observed increase in period, which amounts to about 36.51 billionths of a second per day, or about one part in 2400 per year.

Assuming that this neutron star now spinning 30 times a second, but then spinning 50 times a second, has a mass equal to that of the sun and a moment of inertia appropriate to its mass and small size, Gold's calculations also led to the two important, though still tentative, conclusions I mentioned briefly earlier.

First, the rapidly rotating Crab pulsar is dissipating energy at the prodigious rate of 10^{31} watts. This is far more than enough to supply the energy being dissipated by this pulsar's radio bursts, some 10^{24} watts. It is equal to, and may account for, the radiant energy being emitted by the vast Crab Nebula itself, at all frequencies.

Second, this energy contribution may be enough, when the energy and plasma contributions of other relatively young and rapidly spinning neutron stars are added to it, to account for the cosmic rays that are bombarding us at this very moment, thus clearing up a decades-old mystery.

The writer is indebted to Frank Drake (Cornell University), Kip Thorne (California Institute of Technology), J. J. G. McCue (M.I.T. Lincoln Laboratory), and Gerald Feinberg (Columbia University) for personal communications that contributed greatly to improving an early draft.

REFERENCES

1. Goldstein, Meisel, *Science*, vol. 163, pp. 810–811, Feb. 21, 1969.
2. *Science*, vol. 164, pp. 709–711, May 9, 1969.

RECOMMENDED READING

Drake, F. "Pulsars," Paper 1E2, *1969 IEEE Internat'l Conv. Dig.*

Drake, F., "Pulsars," in *Science Year 1969*. Chicago: Field Enterprises Educational Corp., in press.

Hewish, A., "Pulsars," *Endeavour*, vol. 28, pp. 55–59, May 1969.

Smith, F. G., and Hewish, A., *Pulsating Stars* (a collection, with additional material, of articles on pulsars that appeared in *Nature* over the past 2½ years). New York: Plenum Press, 1969.

Thorne, K., "Gravitational collapse," *Sci. Am.*, vol. 217, pp. 88–98, Nov. 1967.

Thorne, K., "Gravitational collapse and the death of a star," *Science*, vol. 150, pp. 1671–1679, Dec. 24, 1965.

The anatomy of integrated-circuit technology

Although it is predicted that monolithic silicon arrays will reach a level of circuit integration that is one or two orders of magnitude greater than exists today, future needs will most certainly increase the role of film, hybrid, and other technologies in microelectronics

Harwick Johnson RCA Laboratories

One definition of integrated electronics can be given 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 definition gives only an indirect clue to the reasons for the tremendous impact of integrated electronics upon the electronics field in particular, and upon technological aids to society in general. This article attempts to explore these reasons and to discuss some of the consequences of the growth of this advanced technology. It complements J. J. Suran's January article, "A Perspective on Integrated Electronics," which explores the problems in design theory that have resulted from the extreme complexity of IC fabrication.

Much has been written concerning integrated electronics and it is no longer necessary or even feasible to detail the technological processes and achievements in a short review. We shall attempt to show the nature of the field through emphasis of the major circumstances affecting its development, and hope to provide some answers to such questions as: Why are certain techniques dominant? Why are certain functions easier to accomplish than others? What circumstances give significance to supplementary techniques? What are the inadequacies of the present system?

To understand the significance of integrated electronics, it is necessary to adopt a systems viewpoint. As such, integrated electronics is an efficient organization of the contributions of the diverse disciplines of materials, active and passive devices, circuit design, circuit fabrication, and packaging into a single product entity. Today this entity is usually a subsystem of the overall electronic system. Tomorrow it may well encompass that portion of the system between the input and output

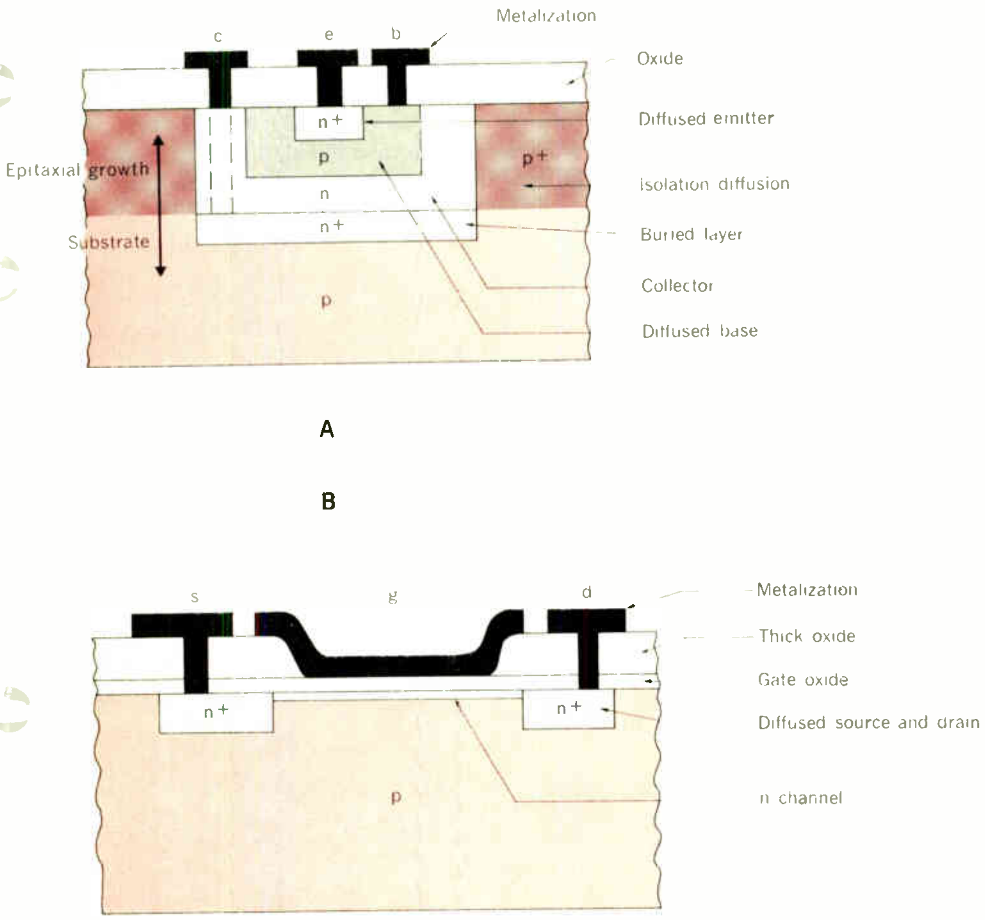


FIGURE 1. Bipolar and MOS transistor structures.

translators necessary to contact the real world.

The key words here are efficient organization and single entity, the key to these possibilities being microelectronics. Literally, microelectronics means very small electronics. But it is not only this attribute of size that is revolutionizing electronics. Small size simply means more electronics for a given volume or weight. Though essential in space applications or in very fast computers, it does not account for the pervasiveness* with which integrated circuits are penetrating all electronics. Rather, it is the small size *in combination with* the attributes of microelectronics that lead to greater reliability, better performance, lower cost, and/or greater circuit sophistication.

Although the technology of microelectronics is advanced, it is not basically new. As practiced today, it is largely an extension of silicon transistor technology and thin-film technology developed for passive elements and interconnections. The most important contributors to the success of integrated electronics are the features of batch or simultaneous processing and the adaptability to automation. These features also played an important role in the success of discrete transistors. It is their extension to interconnections and circuitry in the form of microelectronics that has given impetus to the development of integrated electronics. The importance of these features to the economics of integrated electronics can scarcely be overemphasized.

But there are other advantages. Because transistors are made simultaneously, they are better matched and because the circuit is contained on a common substrate, temperature differentials are minimized. These circumstances, for example, make it relatively easy to produce differential

* This word is so appropriately descriptive that its first use in the present context by P. E. Haggerty should be acknowledged.

amplifiers, in comparison with conventional methods utilizing discrete components.

Although the basic circuit principles remain true to Maxwell, the design constraints are different. The optimum integrated-circuit design will not be recognized as a one-to-one translation of a circuit for the same purpose designed with discrete components. Moreover, because additional elements are available at a minimal cost on a chip basis, the design will tend toward greater sophistication and to superior performance.

Thus, a variety of interlocking "reasons" account for the movement toward integrated electronics. The significance of these factors will vary from one electronic system to another. However, the greater the number of these factors that can be brought into play, the more pervasive will be the resulting electronics.

Monolithic silicon technology

The heading of this section is somewhat a misnomer, but is used loosely, as it usually is, to apply to the system of integrated-circuit fabrication from design to the finished product. Two transistor structures are dominant in integrated-circuit practice. These are the bipolar and MOS structures shown in Fig. 1. The structures and their principles of operation are entirely different. Nevertheless, the fabrication processes are generally the same and differ only in technological detail. The elements of the overall system are outlined in Fig. 2. It should be noted that silicon technology, per se, is included in the step of wafer processing.

Characteristic of almost all the steps in wafer processing is that they are physical or chemical processes carried out under controlled conditions in a controlled environment. This control contributes much to the reliability of the product. It is also characteristic of most of these processes that it is possible for wafers to be batch-processed.

An insight into the efficacy of the batch process and the economic importance of the many detailed steps involved in transistor fabrication may be gained by considering a typical single diffusion step. For the *diffusion process alone*, a labor expenditure of one man-day will typically process 1300 wafers. For today's packing densities, this corresponds to 2×10^7 circuit elements (transistors, diodes, resistors).

Actually, the diffusion process itself is but a minor part of a diffusion step. Indeed, the wafer cleaning, etching, photoresist, and other processes associated with a single diffusion require a much greater labor expenditure than does the diffusion itself. For the *complete single diffusion step*, 10^6 circuit elements are processed per man-day.

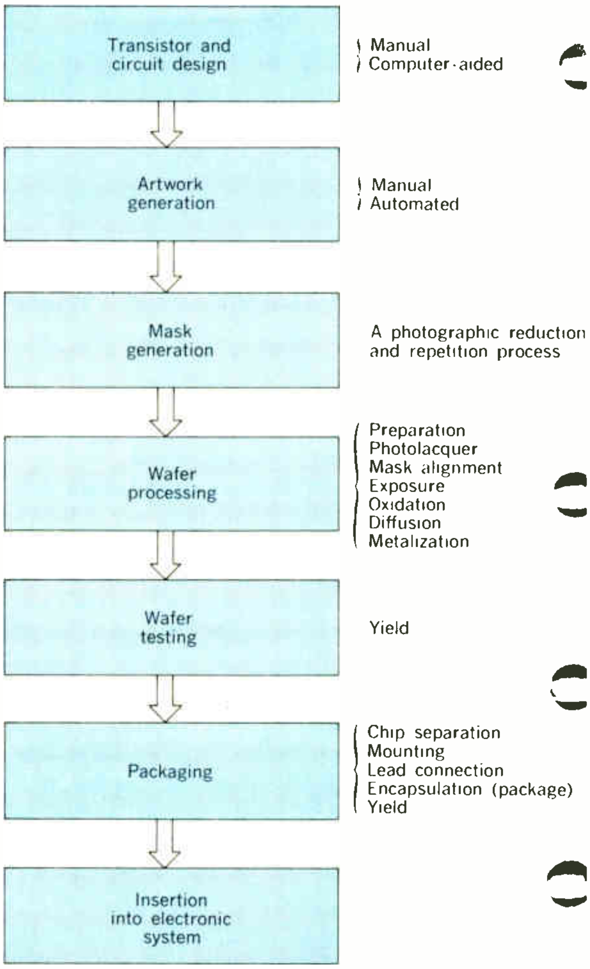
With efficient operations such as this, vast numbers of circuits can be processed at an extremely low unit cost. Herein lies the basis of silicon technology in providing large quantities of electronics at low cost. Even so, these numbers do not represent a limit and it is technologically feasible to increase them. Wafer processing is but one element of the overall system, and it can already be foreseen that its very efficacy is throwing it out of balance with the other elements of the system; that is, further lowering of the costs of wafer processing has but a minor effect on the cost of the product and, hence, does little to further the penetration of the product into electronic systems. It is equally important to refine the other elements of the system.

The steps of the system prior to wafer processing—those involved in setting up a design—are not repeatec more than a few times until the design is finalized. Foi transistors and simple integrated circuits, this cost prorated over a large production, is reasonable. But thi may not be true for complex circuits that require mc design and artwork labor, as well as readjustment o tentative designs. Nor may it be so if the design is specific to an application with a limited market—e.g., a custom design.

Under these given circumstances, the design set-up costs could be dominant. The question of keeping t flexibility of the system in adapting to new designs h. not yet been resolved. To alleviate the problem, artwork generation has been automated, and computer aids in th form of circuit simulation and topological formulas ar used. Nevertheless, wafer processing is so prolific tha the economics of instituting a design must receive mor attention in the future.

One possibility for improving the overall performanc is to reduce the size of the individual components—tha is, to put more "micro" in the electronics. This require an increased resolution in the mask and the photographi processes by which it is produced. Presently, line width down to 5 μm with tolerances of 0.3 μm are commonl

FIGURE 2. Outline of monolithic silicon technology.



employed using radiation of 0.5-μm wavelength. Since practical limits are ordinarily encountered before fundamental limits are attained, some change in these techniques will be necessary to effect order-of-magnitude improvements.

Two characteristics of the photographic process should be pointed out. One is the economic advantage of exposing the whole wafer (e.g., 400 integrated circuits) at one time. On the other hand, mask alignment is the one remaining step that relies on operator skill. Proposals have been made to dispense with the mask (and even the photoresist) and expose the wafer with a computer-controlled scanning electron beam of high resolution. [illegible] beam resolution of better than 1 μm, an inordinately long and uneconomic exposure time is required. Some composite system may be feasible, however.

The elements of the system following wafer processing are concerned with preparing the integrated circuit for insertion into the overall electronic system. Here again, because of the efficacy of silicon technology, packaging techniques of the past must be modified if their cost is not to limit the applicability of integrated circuits. The individual and often manual operations of mounting and lead attachment can no longer be tolerated. Even for discrete transistors, the packaging operation costs almost as much as the transistor chip.

For integrated circuits of high complexity, which require increasing numbers of pins, packaging costs will increase. It becomes an important consideration in system organization and integrated circuit design to minimize the number of external connections required. The most promising solution to the overall packaging operation appears to be the beam-lead approach.[1] A beam-lead circuit before chip separation is shown in Fig. 3. The beam leads extending out over the silicon that will be etched away to effect chip separation are readily identified as the relatively large, rough, rectangular depositions extending from the integrated circuit proper. After removal of the underlying silicon, the outer extremities of these leads will be free, the leads being cantilevered from the silicon chip. The circuit measures 1.35 mm from beam tip to beam tip.

Aside from the technical sophistication of beam-lead technology in the use of materials and the elimination of a separate package structure, its promise in the present context arises from the use of batch processing and elimination of the individual manual operations of conventional procedures. Chip separation is now a chemical-etch process, lead formation is an electroplating process, and encapsulation is a chemical-deposition process. These functions can be carried out under controlled conditions to maintain the reliability of the product and, with batch processing, at low unit cost.

The diagram of Fig. 2 includes a final element indicated as "insertion into electronic system." This block was added to emphasize the importance of keeping the overall system always in mind. Here the beam-lead approach is also beneficial because insertion into the electronic system may be done with simultaneous lead attachment.

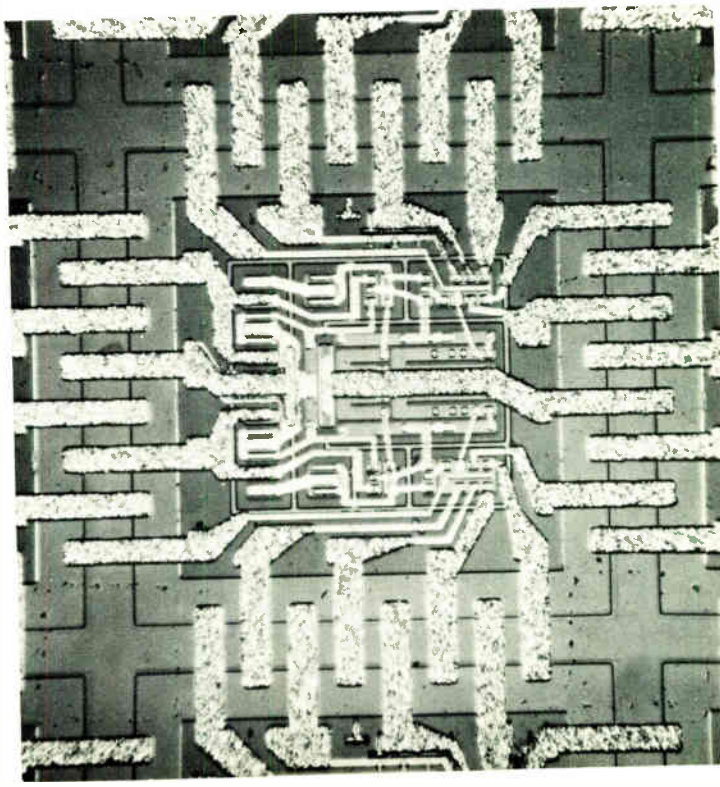
In summary, we have tried to point out in this section that the system of integrated-circuit fabrication, through reliance on automation, materials sophistication, and batch or simultaneous processing, can produce vast quantities of low-unit-cost, reliable, and complex integrated circuits.

However, it is not to be expected that all electronic systems can be adapted to this method of fabrication; some alternatives are discussed later. Moreover, when the adaptation can be made, a serious challenge arises in what to do with these vast quantities of electronics. The potential now apparent is greater than the rosy predictions of a decade ago. Yet, there is an irony. With these efficient processes, the manufacturer can supply vast quantities of electronics, yet his dollar volume may be disappointing because of the low unit costs. So not only must the advantages of integrated electronics motivate the penetration into all of electronics, but the market for electronics must be enlarged if the overall system is to flourish.

The past and future roles of silicon electronics in enhancing the use of computers is an example of such changing considerations. The transistor is credited with making the large computer practical. Yet the computer is but a tool, and the extent of its use will depend on the cost of making a computation or calculation. Only 2.5 percent of the cost is said to be attributed to the silicon electronics.[2] Simple reduction in the cost of silicon electronics is approaching a marginal contribution to the wider use of large-scale computers. Attention to the other costs of computation could be more effective. Even as the electronics industry is forced by the low cost of silicon technology to revamp design and packaging procedures, so the computer industry must give attention to those other costs. Hopefully, the greater use of electronics may aid in achieving these ends.

Although cost reduction, per se, may no longer be a paramount objective in the development of silicon electronics for computers, the important objectives of

FIGURE 3. A beam-lead circuit before chip separation. (Courtesy Bell Telephone Laboratories)



greater capability and flexibility remain. The future role of silicon electronics will be to aid in achieving these objectives. One approach is through the development of large-scale integration, an important step forward in the sophistication of silicon electronics.

We turn now to examine some of the design fields to which integrated electronics has been applied. In addition to simply pointing out what has been accomplished, we have tried to make clear the significance of the microelectronic aspects and to indicate the consequent technological and economical challenges that have arisen.

Digital circuits*

Of the various fields to which it has been applied, integrated electronics has been most impressive in digital circuits and, in particular, in logic circuits of computers. In this application, there is a fortuitous coincidence between the application requirements and the simpler capabilities of silicon technology in integrated electronics.

The basic requirement of an elementary logic circuit is quite simple: It must provide a means of signal quantization. Usually only two levels are required. The quantization function is readily provided by a p-n junction or a transistor, and various logic elements can be devised employing only transistors, solid-state diodes, and resistors. Transistors are advantageous in providing the gain necessary for the fanout of element assemblies.

Logic-element design today goes far beyond this basic consideration to provide operation at high speed, low-power consumption, etc.—that is, the various attributes of improved performance. These important details will not be discussed here. We wish only to emphasize that logic circuits demand of integrated electronics the ability to produce only combinations of transistors, diodes, and resistors. This is almost the simplest device capability that could be demanded. Yet further simplicity results from the substitution of transistors or diodes for [illegible] resistor elements so that the elementary circuits are composed entirely of semiconductor devices. Since resistors were traditionally less expensive and more reliable than discrete transistors, such substitution may appear surprising.

This is just a small but good example of the reorientation in values that comes from integrated electron[illegible] Here, production expense is related to the chip area occupied by the element. Since a resistor occupies more area than a transistor, in integrated form it may well be more expensive than a transistor or diode.

It is characteristic of the central processing unit of computers that complex logic functions are realized [illegible] combinations of large numbers of simple and simi[illegible] digital circuits. To this problem of expeditiously combining large numbers of elementary circuits, the microelectronic aspect of silicon technology provides many favorable answers. The small size is not only convenient but, in high-speed computers, is essential in minimizing signal propagation delays. Batch processing offers low cost for both the device elements *and* their interconnections, which also benefit from the reliability of silicon technology.

Altogether, a complex logic circuit is a very favorable vehicle for displaying the capabilities of integrated electronics and much progress has been made. The silicon chip displayed in Fig. 4 is typical of a relatively simple product that is produced today in large numbers. T[illegible] chip is an example of transistor–transistor logic (T[illegible]) and contains 28 device elements with interconnections to comprise four logic gates in an area of 14.9×10^{-3} cm^2. This is about half the area contained on the head of a common pin.

It is customary, as we have done, to cite the number of

FIGURE 4. A typical high-volume digital circuit chip. An example of a transistor–transistor logic gate.

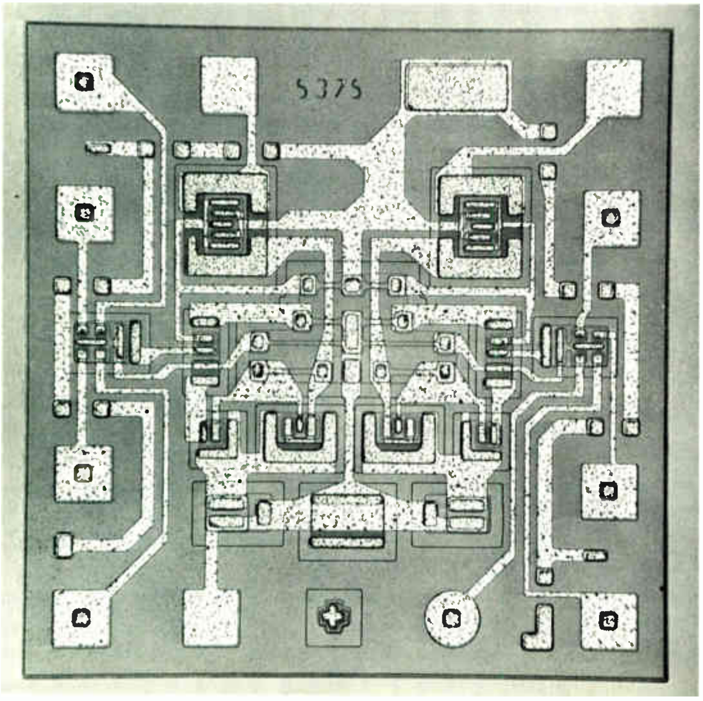
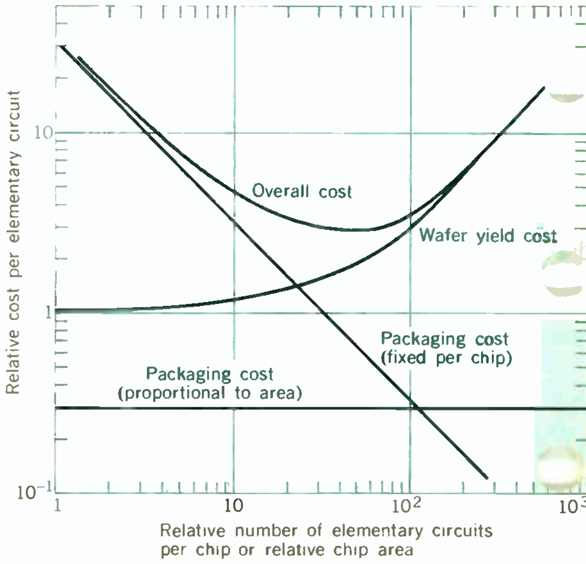


FIGURE 5. Trends in manufacturing costs with increasing levels of integration.



* For details of the modern technology of digital integrated circuits, see Ref. 3.

device elements contained in the chip. But it is of equal economic significance that the chip also contains 60 interconnections, all made in a most reliable manner by the semiconductor manufacturer.

If this modest level of integration is good, as already has been shown in the marketplace, more integration would be better. But how much more? This question leads to the challenge of large-scale integration (LSI), the philosophy of incorporating a large subsystem or even a complete system on a chip. The potential and the limits of LSI are topics of much concern today. Industry needs answers to such questions as: Are the limits to be set by silicon technology, by conditions on system organization, or simply by the economic problem of generating enough volume for a product so that LSI is economically competitive? Certainly all of these are factors influencing today's designs. But this begs the basic question. In the absence of limiting relations such as those derived for the performance of charge-controlled electron devices, recourse is made to extrapolating current experience and to defining the roadblocks of present practice.

From the viewpoint of the semiconductor manufacturer, a possible guideline to a suitable degree of circuit integration is one that minimizes his manufacturing costs per elementary circuit. In this way, the manufacturing operation offers the customer the most value for his dollar.

Manufacturing costs include the costs of processing and packaging the silicon chip. The chip-processing cost per elementary circuit may be written as

$$C_u = \frac{pa}{y(n)} \tag{1}$$

where p is the processing cost per unit area, a is the area of an elementary circuit, and y is the chip yield. The effect of increasing levels of circuit integration (the number n of elementary circuits contained on a chip) is contained in the yield factor. The yield decreases rapidly at high levels of circuit integration.

Packaging costs are in a state of flux for reasons mentioned earlier, so that there is little consensus on how they will vary with circuit complexity. However, the details affect our conclusions only in the matter of degree. We will argue that packaging operations of the future will be done by batch processing and simultaneous lead attachment. With simultaneous lead attachment, the attachment cost per wafer may be taken, for illustrative purposes, as fixed and independent of the number of leads to be attached. The packaging cost per elementary circuit will take the following form:

$$P_u = \alpha a + \frac{\beta}{n} \tag{2}$$

where the first term accounts for batch processes, and the second for lead attachment and any other fixed cost per wafer. The inverse dependence on n for the second term results from β representing the fixed cost per wafer containing n elementary circuits. Since packaging yield is a multiplication factor in both C_u and P_u, it will not be included in this discussion.

Since C_u increases and P_u decreases with level of circuit integration, a minimum in the manufacturing cost $(C_u + P_u)$ per elementary circuit will exist at some level of circuit integration. Roughly (but not exactly), it occurs when the chip costs approximate the packaging costs. The trends are shown in Fig. 5. In general, the minimum is very broad, and not too great a significance should be attached to the minimum point.

The trend of the chip cost curve depends on the yield function, which, in turn, depends on the defect distribution. We may conceive that the defects are located by chance and that, in the limit of large numbers, it is possible to define an average defect density, D_o. This is the simple Poisson description of "randomness." These concepts lead to the probability of not finding a defect in an area a (the area of an elementary circuit) of

$$P = e^{-aD_o}$$

which is defined as the yield function y_e for an elementary circuit.

For an integrated circuit comprising n elementary circuits, the pertinent area is na and the yield would go as y_e^n, giving rise to the so-called "y^nth" problem. However, it has been pointed out[4] that the defect distribution is not random in a simple Poisson sense. Experience has shown that this simple two-parameter description is not adequate, and that the yield does not fall off as rapidly at low yields as does the yield for a simple Poisson distribution of defects.

This phenomenon serves to extend the upper end of the cost curve and shifts the minimum to higher levels of integration. A somewhat startling consequence is that, under some circumstances (relatively high packaging costs), the yield factor at the minimum can be in the neighborhood of 15 percent. This result is a consequence of attempting to put enough value into the chip to roughly match the packaging costs. In this example, the lowest manufacturing cost per elementary circuit is obtained by designing up to a level of integration that results in six out of seven units being discarded.

Broad as it is, it is at least semantically comforting to find a minimum in the unit cost in order to define a possible objective. However, with current trends in "packaging" procedures, this comfort is likely to be lost. As packaging costs are reduced, the minimum shifts to levels of integration that may be lower than that desirable from other systems considerations. In the limit, if the lead attachment term in Eq. (2) becomes negligible or if the manufacturer supplies batch-processed chips, then a minimum based on these arguments no longer exists. The manufacturing cost is a monotonically increasing function of the level of integration and depends primarily on the yield factor. Aside from the steep rise in cost at low yields, a desirable level of integration must be found by considering factors in the overall system aside from the manufacturing costs. One factor, for example, could be the cost of insertion into the electronic system, as noted in the last block of Fig. 2.

Several particular analyses and projections[4-6] into the future have been carried out that are similar to these arguments but differ in detail—particularly in the formulation of packaging costs. Considering arrays of bipolar gates, Seeds and Noyce predict that the manufacturing costs per gate can go from 9 cents per gate in 1968 to 0.7 cent per gate in 1974—the latter gates being produced in arrays of 200. The important conclusion for our purpose is that projections of the capabilities of silicon technology predict the feasibility of arrays having a level of circuit integration one or two orders of magnitude greater than that in common practice today, with a substantial reduction in the manufacturing cost per elementary circuit. Or, more simply, the technological limits appear to be far beyond today's practice.

Another possible design limit is the so-called pin-limited condition. The pins or connection leads are placed along the perimeter of a chip. As the number n of elementary circuits on a chip increases, the area increases in direct proportion to n, but the perimeter or linear dimension available for lead placement increases only by $\sqrt{n}$. If, with increasing level of circuit integration, the system interconnections between chips increase faster than $\sqrt{n}$, more pins will soon be required than can be placed on the periphery of the chip. This limitation would determine the level of integration that can be incorporated on one chip. This is a problem in system design and layout rather than a general limitation. It is evident that subsystems should be sought that requir minimum number of interconnections between chips.

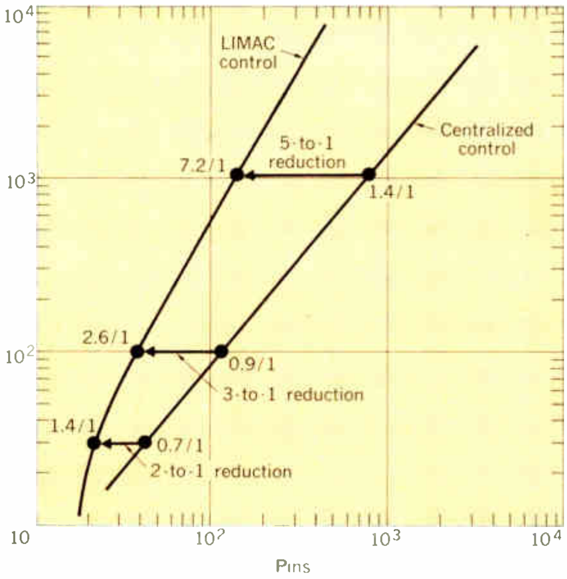
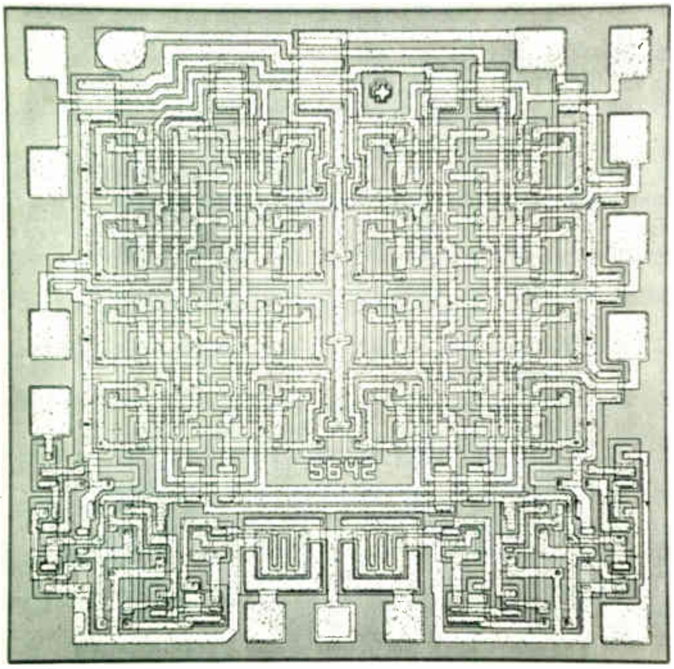


FIGURE 6. Pin requirement reduction by system reorganization (S. Y. Levy *et al.*[7]).

FIGURE 7. A 16-bit T²L memory chip.



This problem emphasizes the point so often made that close cooperation is necessary between the system designer and the semiconductor manufacturer. Alternatively, systems may be reorganized so that the number of pins per chip increases no faster than $\sqrt{n}$. An example has been given by S. Y. Levy *et al.*[7] The authors point out that, in many computer designs, memory and data-processing elements are ordered in a fairly regular structure and by this token are amenable to large-scale integration.

On the other hand, however, the centralized control logic structure, which may comprise about one half of the machine, is highly irregular because of the diverse control signals it must generate. This formlessness of the control structure leads to many interconnections and to pin-limited conditions for large-scale integration. Restructuring the system using functional partitioning for both the data and control paths, the authors have succeeded through improved ordering of the component structures in reducing the pin requirement of an array by a factor of two or more, as shown in Fig. 6.

Another associated pin-number problem that may tend at any time to limit the degree of integration is that of testing. Testing costs are an important part of production costs and it is observed that the number of tests for comprehensive evaluation tends to increase exponentially with the number of pins. Evidently, the advent of large-scale integration will call for fast low-cost testing techniques, and, most likely, a new philosphy of testing. Perhaps a greater reliance will be placed on functional rather than parametric testing.

Another promising application for large-scale integration in computers is semiconductor memories. This application is in its infancy, the first commercial product having appeared in 1968. The basic circuit element is the flip-flop, again a circuit comprised only of transistors and resistors. These elements are assembled to perform the functions of shift registers and decoders. Memory patterns are very regular and have a high gate-to-pin ratio.

Because volatility is a problem in semiconductor memories, initial development will not attempt to compete with the popularity of magnetic cores in large-capacity random-access memories. Rather, those memory functions that can exploit the advantages of semiconductors will be developed. Among these functions is the read-only memory, in which the memory information is wired in and memory volatility is not a problem. The use of semiconductor material is efficient, requiring one transistor per bit.

Because of the efficiency that is added to the computation process, fast semiconductor scratch-pad memories are also being developed. Figure 7 shows a chip containing a 16-bit T²L memory. Compared with Fig. 4 the chip contains about four times as many device elements in twice the area.

The capacity of present semiconductor memories is limited by the chip size that will permit a reasonable yield. For a typical chip of 6.45 × 10^{-2} cm^2, 64 bits with bipolar construction or 256 bits with MOS techniques appear currently practical. The size advantage of the MOS technique is readily apparent for large-capacity memories. Bipolar techniques, however, do enjoy an advantage of about 20 times in speed, and technique refinements are being explored to mitigate the size disadvantage.[8,9] On the other hand, refined MOS techniques through internal capacitance reduction will increase the speed of operation within a chip; however, the problem of connection to the high capacitance of the outside world remains to be solved.

As techniques are improved, it is only reasonable to expect semiconductor memories to have capacities approaching 5000 bits in the near future.

To provide nonvolatile alterable semiconductor memories, various ideas have been explored to introduce the necessary long-term hysteretic effect into the gate insulator in insulated-gate field-effect transistors. Current interest centers on the injection of charge by tunneling into silicon-nitride insulators.[10]

Analog circuits*

The transition from discrete transistors to integrated circuits proceeds more slowly for analog or linear circuitry than it did for digital circuitry. To account for this lag, there are almost innumerable reasons, both economical and technological.

We have seen that a computer system utilizes large numbers of similar circuits, as in logic or memory subsystems. On the other hand, diversity is more characteristic of circuits in analog systems, in which the attractiveness of the repetitive pattern of an elementary circuit is lost and the number of circuits for the system is decreased.

We have noted that, at least in the beginning, logic circuits, for example, placed relatively simple requirements on the transistor devices. On the other hand, most analog circuits are high-performance circuits with each discrete transistor designed specifically for the role it is to play. Until improved integrable transistor forms were devised, only the most modest of analog circuitry could be considered for translation into integrated form. Important contributions to improved integrated transistors were made with the advent of improved isolation techniques and the "buried layer" technique (see Fig. 1).

Low-pass or baseband operation is typical of digital circuits; it requires, for the most part, only transistor and resistor elements. Many analog circuits are characterized by bandpass operation with some requiring the bandpass to be shifted on demand. For these functions, high-Q inductors and mechanically variable capacitors are conventionally used. These are not part of the repertoire of integrated circuit technology. It is evident that much rethinking and innovation is required to accomplish similar functions most expeditiously with the use of integrated circuits.

Solutions to these problems will be radical departures from conventional practice. Gain functions will probably be provided in broadband form. Initially, bandpass determining elements in conventional discrete form will be lumped at the functional interfaces of the system. To provide inductive reactances more suitable for integration, miniature mechanical resonators, electronically simulated reactances,[12] and electroacoustic effects in solids are being investigated, with some use being made of piezoelectric resonators. The problem of variable-capacitive reactance in integrated form appears less formidable with the availability of the voltage-controlled capacitance of a p-n junction (varactor).

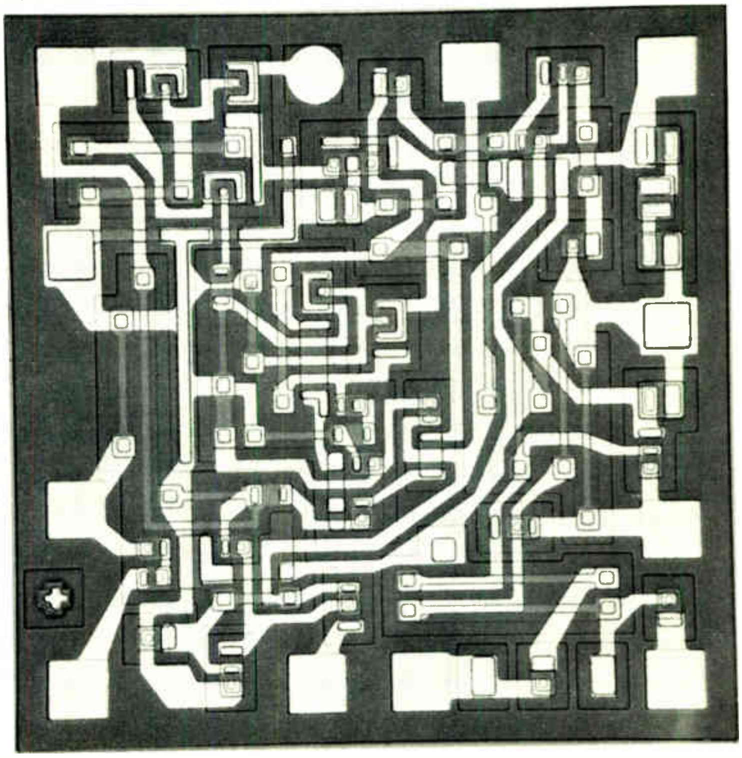
Finally, a large potential market for integrated analog circuitry exists in the consumer entertainment and appliance markets. In these markets, cost considerations are dominant; little, if any, premium is available for greater reliability. Consequently, acceptance in this market depends almost entirely on cost, although the other attributes of microelectronics and integrated circuits are recognized as desirable and beneficial.

Notwithstanding this unfavorable comparison with digital circuits, much progress has been made, leaving no doubt that integrated circuits will be as pervasive in the analog field as they are in the digital field. This growth will be accompanied by many changes in design philosophy to adapt and to exploit integrated-circuit technology. Some of these accommodations are already becoming clear; others remain vague.

Perhaps one of the major adjustments to be made is to determine how analog circuits are to be partitioned and "standardized" to produce an economic volume for the semiconductor fabricator without essentially eliminating the possibilities for initiative on the part of the circuit designer.

Although there undoubtedly will be "standard"-type products developed that will service widespread needs, a

FIGURE 8. An analog subsystem chip combining functions of IF amplification, limiting, detection, and audio amplification.



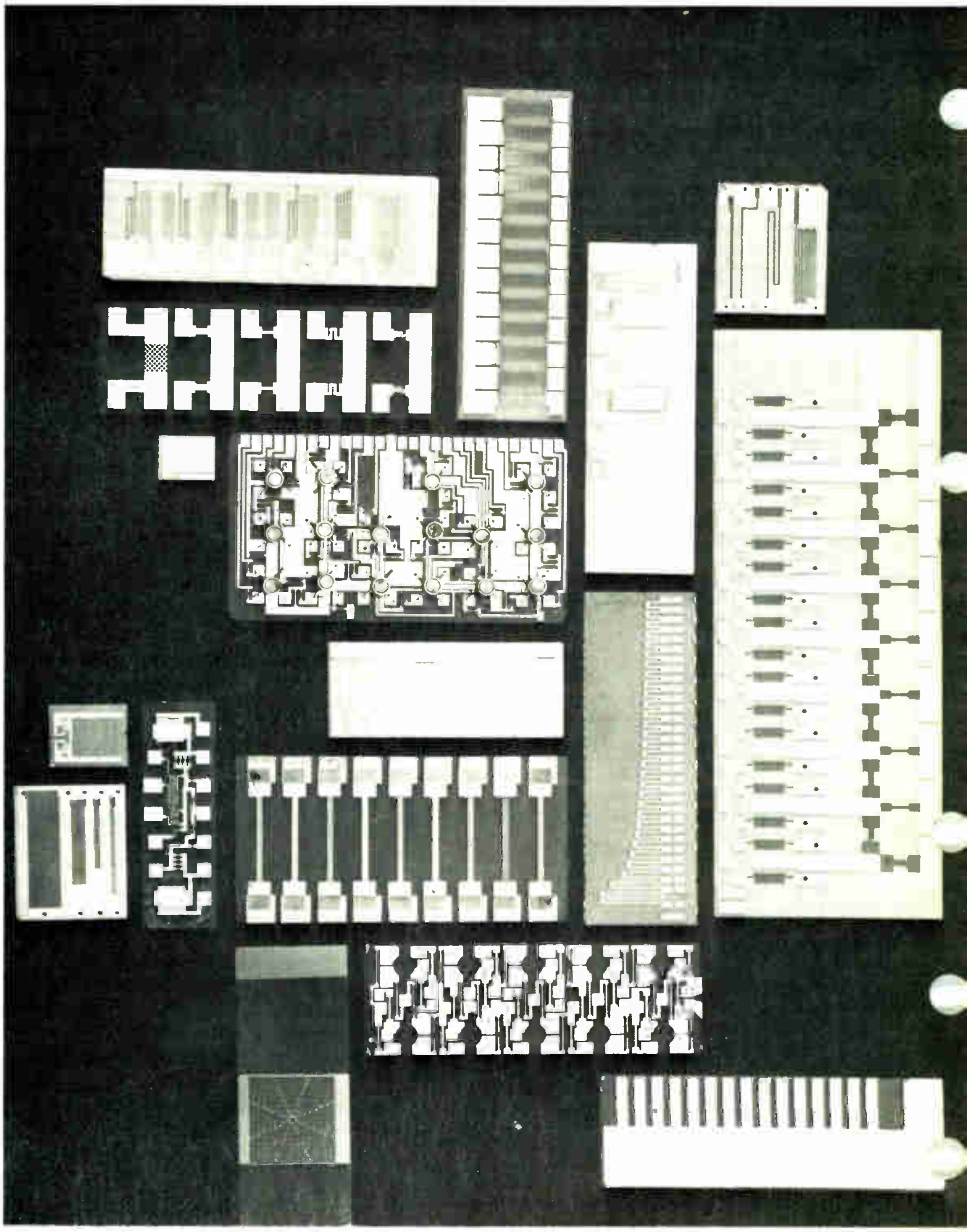
* For details of the modern technology of linear integrated circuits see Ref. 11.

reorientation in thinking is revealing a suitable compromise to accept innovations in circuit design as well. Instead of thinking only in terms of a completely "standardized" product, the semiconductor fabricator will think in terms of a substantially "standard" process and provide for circuit innovation within the bounds of that process. The circuit designer will accept these constraints and similarly display his ingenuity within the bounds of the "standard" process. Thus, it is the process rather than the circuit that will be "standardized."

This argument does not imply that the process is to be a static one. On the contrary, it will continue to develop a new and more efficient fabrication techniques are found Nor does the argument imply a single process; for example, a manufacturer may have one "standard" proces for low-frequency and another for high-frequency circuits.

However, it certainly means that the semiconducto manufacturer will not explicitly tailor each transistor t serve its function on the chip most efficiently. Fo

FIGURE 9. A collection of tantalum thin-film circuits. (Courtesy Bell Telephone Laboratories



example, area can be readily adjusted but special doping profiles are unlikely. Some transistor devices will be better than they need to be; in other instances, more than one transistor device may be necessary to accomplish what might have been done with one discrete transistor.

At present it is difficult to generalize about the trends in partitioning. An elementary choice, to partition by function, may be an initial choice for commercial or industrial products, where cost is not all-important. However, the partitioning is likely to be more subtle than this and take into account the number of external leads and other factors. Partitioning by elementary function is almost certainly not satisfactory for consumer products where cost *is* all-important. In this area, it is likely that every effort will be made to place as much as possible on one chip.

The 1.25-mm silicon chip shown in Fig. 8 is an analog subsystem combining several functions. This chip is illustrative of the functional diversity of analog circuits and of the desire to place as much on one chip as the state of the art will allow. Designed for application in FM receivers, the chip provides IF amplification, limiting, detection, and audio preamplification, while providing an AFC signal and incorporating regulation of the voltages supplied to the various sections. In that, the bandpass-determining elements and a discriminator transformer must be additionally supplied. The chip is also illustrative of the challenge to devise low-cost electronically simulated reactances, or to circumvent the need by means of innovation.

Thin-film, thick-film, and hybrid technologies

Although monolithic silicon technology goes a long way toward being "all things to all people," the diversity of electronic systems is too great for one technology to attain that utopia. Microelectronics should not be a synonym for monolithic silicon electronics, for its horizons are not limited to those of a silicon technology. It is probably true that, if an electronic function can be expressed in terms of a monolithic silicon structure, then this, more often than not, will offer the preferable solution. However, there will be instances where it is not economically preferable, and instances where it is not technically preferable or even feasible. An individual judgment must be made between alternative solutions.

The utilization of film technology is presently the popular alternative to a monolithic silicon structure. Film technology may be used alone or in conjunction with integrated or discrete silicon structures. In the latter case, the technology is referred to as a hybrid technology. Film technology is further divided into two general classes—thin-film and thick-film technologies.

Thin-film technologies are diverse and may utilize evaporation, sublimation, sputtering, pyrolysis, electrolysis, or chemical reaction. Film thickness is less than one micrometer and may range down to 100 Å. Pattern definition, when defined photolithographically, can approach that of silicon technology. Through suitable selection of process and materials, fabrication of transistors, diodes, resistors, conductors, and capacitors may be facilitated.

Thick films are derived from a mixture of metallic powders, vitreous binders, and a carrying vehicle. The mixture is applied as a paste to a ceramic substrate in the desired pattern by screening and is then fired. Film thicknesses range upwards from a few micrometers, and pattern definition is correspondingly coarser. Composition of the mixture is varied to produce resistors, conductors, and capacitors.

The details of these alternative or supplementary technologies are not described in depth in this article. Using the restraints of monolithic silicon technology as a point of departure, we wish only to illustrate that the range and promise of microelectronics is greatly enhanced in various ways through the use of thin-film, thick-film, and hybrid technologies.

One of the necessities of a solid silicon structure is that the various elements of the circuit be isolated from the underlying silicon and from each other. This requirement is ordinarily accomplished with p-n junctions. Aside from other undesirable characteristics of p-n junctions, this step results in considerable unwanted circuit capacitance, which limits the response of high-speed circuits.

Among the many varied solutions that have been proposed, we cite, as an example, the possibilities of thin-film silicon[13,14] deposited on an insulator—the "silicon-on-sapphire" technology. Isolation is attained simply by removing the unwanted silicon. Other potential advantages include a higher packing density, ease of fabricating complementary structures, and a low-loss dielectric substrate for passive elements.

Monolithic silicon technology imposes restraints on the magnitude and the precision of capacitances and resistances that may be employed in an integrated-circuit design. In one sense, this is a challenge to the ingenuity of the designer to work within these limits. But the obligation remains to explore alternative means that could prove more satisfactory in meeting tue designer's objectives.

One of the possible alternatives is a combination of silicon chips (which may or may not be in integrated form) and thin- or thick-film passive circuitry; that is, a hybrid configuration. In general, the thin- or thick-film portion would contain those critical elements not amenable to fabrication on the silicon chip and thus alleviate the restraint imposed by silicon technology. In practice, the division would depend as well on other factors such as the number of interconnections required.

Thin-film components have been made of many different materials. Perhaps the most sophisticated, yet elegantly simple, system from the standpoint of the formation of components, conductor paths, and insulators by modifications of a single material is the tantalum thin-film technology[15] developed by Bell Telephone Laboratories. A collection of tantalum thin-film circuits displayed in Fig. 9 illustrates the flexibility of the technology in producing a wide variety of circuits. Material compatibility assures stability and reliability. Configuration geometries are defined photolithographically. Resistance and capacitance magnitudes can be precisely controlled by anodization, which is a controllable electrochemical process.

In many of its aspects, tantalum technology promises to do for the integration of passive elements what silicon technology does for the integration of active elements.

In the same vein, silicon technology is inadequate to provide the passive structures and some active elements of interest in microwave applications. Again, a hybrid construction greatly enhances the possibilities for integrated microwave circuits.*

As a semiconductor, silicon has certain prescribed characteristics; hence, some electron devices cannot be made in silicon. For example, the band structure of silicon prohibits the fabrication of efficient recombination–radiation devices or transferred-electron devices in this material. Hybridization is presently the easiest way to incorporate such devices within microelectronic circuits.

Alternatively, an integration technology could be developed in another semiconductor. Thus, the foregoing examples could be fitted into a gallium arsenide technology of integration. The economic impetus for the rapid development of an alternative integrated semiconductor technology is not yet very strong.

Since semiconductors, and hence semiconductor devices, are inherently quite temperature-sensitive, certain restraints are imposed on the development of monolithic silicon technology. To the present, thermal design has not been a major consideration because the bulk of the effort on integrated circuits has been concerned with low-power information-processing circuitry. Where temperature sensitivity has appeared, it has been handled by circuit design that depends on resistance ratios rather than on absolute magnitudes or by temperature-compensation circuitry. Monolithic construction has been favorable to these techniques because of the strong thermal coupling between integrated elements. Nevertheless, foreseeable developments of greater packing density, the desire to extend integration to include output elements, and the further development of integrated linear circuits will increase the thermal design problems. Hybrid construction can aid in handling thermal problems through isolation of those elements dissipating large amounts of power, the separation of thermally critical components, and providing, where necessary, an improved thermal circuit.

These illustrations of the enhancement of the ability of microelectronics through the use of supplementary techniques are by no means exhaustive. Indeed, a plethora of possible extensions is being explored in laboratories throughout the industry. As powerful as monolithic silicon electronics may be, microelectronics is not bound by the limitations of that technology.

Concluding comments

This review has emphasized those considerations of integrated-circuit technology that can be derived from or answered by the technical aspects of present-day microelectronics.

From the progress that has been made and any reasonable extrapolation of that progress, the technical horizons extend far beyond what has been accomplished to date. Although this is not to say that no technological problems remain, the interesting questions today are perhaps not what can be accomplished but to what extent logistic and economic factors may operate to limit the development of these possibilities.

Will circuit individuality move from the subsystem level to the system level or will equipment manufacturers demand similar but different chips from the semiconductor manufacturer? As integration is extended, can multiterminal chips be tested without prohibitive cost? In computer development, will it become better strategy to cost-reduce peripherals and software with less effort on the development of semiconductor microelectronics or can the latter lend a hand? A major attraction of microelectronics is low cost, but this is traditionally achieved through large-volume production. This approach sprea engineering and tooling costs and permits the buildup of production yields to a highly efficient level. Will this volume be available? If the volume rate is marginal, must we infer that improved systems requiring new chip designs must come at a slow rate?

These and other similar and tantalizing questio without clear answers today face the microelectron industry in the years ahead. Perhaps some of the future answers will be described as a paraphrase of a popular commercial—it's not how complex you make it, it's how you make it complex.

* Details of the modern technology of microwave integrated circuits are reviewed in Ref. 16.

REFERENCES

1. Lepselter, M. P., "Beam-lead technology," *Bell System Tec. J.*, vol. 45, p. 223, 1966.

2. Rice, R., "Impact of arrays on digital systems," *IEEE J. Solid-State Circuits*, vol. SC-2, p. 148, 1967.

3. Special issue on large-scale integration, *IEEE J. Solid-State Circuits*, vol. SC-2, Dec. 1967.

4. Murphy, B. T., "Cost-size optima of monolithic integrated circuits," *Proc. IEEE*, vol. 52, p. 1537, 1964.

5. Seeds, R. B., "Yield and cost analysis of bipolar LSI," presented at IEEE Fall Devices Meeting, 1967.

6. Noyce, R. N., "Making integrated electronics technology work," *IEEE Spectrum*, vol. 5 p. 63, May 1968.

7. Levy, S. Y., Lindhardt, R. J., Miller, H. S., and Sudnam, R. D., "System utilization of large-scale integration," *IEEE Trans. Electronic Computers*, vol. EC-16, p. 562, 1967.

8. Murphy, B. T., Neville, S. M., and Pederson, R. A., "New, simplified, bipolar technology and its application to systems," *Dig. Tech. Papers 1969 Internat'l Solid-State Circuits Conf.*, p.

9. Hodges, D. A., *et al.*, "Low-power bipolar transistor mem cells," *Dig. Tech. Papers 1969 Internat'l Solid-State Circuits Conf.*, p. 194.

10. Pao, H. C., and O'Connell, M., "Memory behavior of an MNS capacitor," *Appl. Phys. Letters*, vol. 12, p. 260, 1968.

11. Special issue on linear integrated circuits, *IEEE J. Solid-State Circuits*, vol. SC-3, Dec. 1968.

12. Mitra, S. K., "Synthesizing active filters," *IEEE Spectrum*, vol. 6, p. 47, Jan. 1969.

13. Manasevit, H. M., and Simpson, W. I., "Single-crystal silicon on a sapphire structure," *J. Appl. Phys.*, vol. 35, p. 1349, 1964.

14. Mueller, C. W., and Robinson, P. H., "Grown-film silicon transistors on sapphire," *Proc. IEEE*, vol. 52, p. 1487, 1964.

15. McLean, D. A., Schwartz, N., and Tidd, E. D., "Tantalum film technology," *Proc. IEEE*, vol. 52, p. 1450, 1964.

16. Special issue on microwave integrated circuits, *IEEE Trans Electron Devices*, vol. ED-15, July 1968.

Harwick Johnson (F) received the B.S. degree from the Michigan College of Mining and Technology and the Ph.D. degree from the University of Wisconsin. Currently with the David Sarnoff Research Center of RCA, he has been concerned in various capacities with the development of diverse electron devices, from vacuum and gas tube transistors and integrated tronics. Dr. Johnson has authored six papers on electronics development and was coeditor and contributing author of a book on field-effect transistors He holds a patent filed in 195 on what would now be calle integrated-circuit phase-oscillator. He is a member of the American Physical Society and Sigma Xi.

Electrooptic effects in ferroelectric ceramics

Ferroelectric ceramics exhibit some remarkable properties. One consequence is that light passing through these materials can be scattered or polarized in an electrically controllable way

Cecil E. Land Sandia Laboratory

Richard Holland Jet Propulsion Laboratory

Ferroelectric ceramics are piezoelectric and optically birefringent. Moreover, their coefficients of piezoelectricity and birefringence are electrically variable. Consequently, these ceramics are applicable to a variety of devices: electrically tuned oscillators and transformers; miniaturized high-Q, high-frequency filters; FM discriminators; optical memories; electrically controlled light shutters and valves; and electrically activated multicolor displays. This article deals with some possibilities for devices that exploit the electrically controllable optical effects. A later article will consider piezoelectric applications.

Materials known as ferroelectric ceramics possess unusual properties that are very attractive for device technology. Ferroelectric materials derived their name from a response pattern that is analogous to that of ferromagnetic materials. In particular, for ordinary room temperatures, the large-signal relation between electric field E and electric polarization P is hysteretic (Fig. 1).

A ferroelectric ceramic is an aggregate of many tiny single-crystal grains, each randomly oriented with respect to all the others. Typical grain sizes vary from 0.5 μm to 35 μm, depending on the chemical composition and method of fabrication.[1,2] The origin of the ferroelectric ceramic hysteretic effect is best visualized by considering one of these individual grains.

In discussing a ferroelectric (or ferromagnetic) material, it is necessary to specify the temperature range under consideration. At elevated temperatures, polarization is not a markedly hysteretic function of electric field. Rather, in a high-temperature environment, the dielectric relationship tends to be nearly linear in a ferroelectric. It is only when temperature is reduced below some critical threshold value that the dielectric properties become hysteretic. This critical temperature is usually referred to as the Curie point; it varies, of course, from material to material.

Figure 2 shows a single high-temperature unit cell of barium titanate.[3] For this compound, (one of the simpler ferroelectric materials) the Curie point is 120°C. Above the Curie point the cell is cubic with a titanium ion occupying the body-center position. However, upon going below the Curie temperature, the structure deforms and becomes tetragonal. Then the titanium ion is displaced from its central position toward one of the six face centers. Because the titanium ion carries a net charge, the cell acquires a dipole moment. Also, the cell becomes about one percent longer in the dipole direction than along the other two axes.

Adjacent cells tend to deform in the same direction, thus favoring a single domain state within each grain. However—except for material with very small grain size—free energy is lowered by the formation of multiple domains within each grain. Thus the grains of a ferroelectric ceramic that have never been subjected to a large

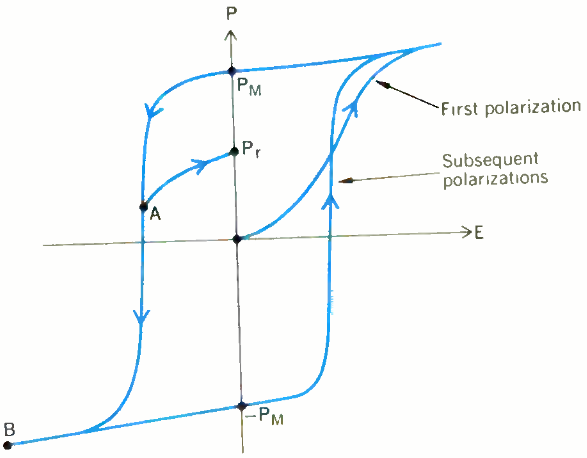
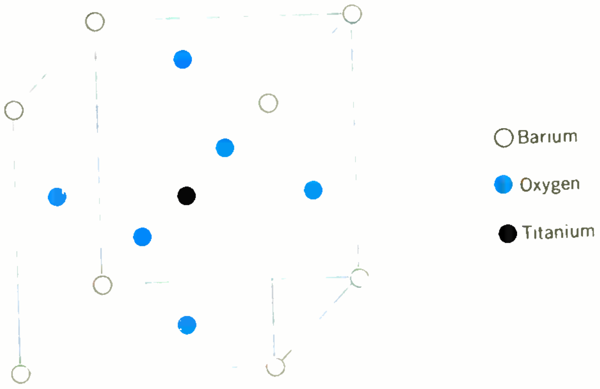


FIGURE 1. Electric field vs. polarization for a ferroelectric ceramic at temperatures below the Curie point.

FIGURE 2. A single unit cell of barium titanate at temperatures above 120°C (the Curie point). In this temperature region, barium titanate has a cubic crystal.



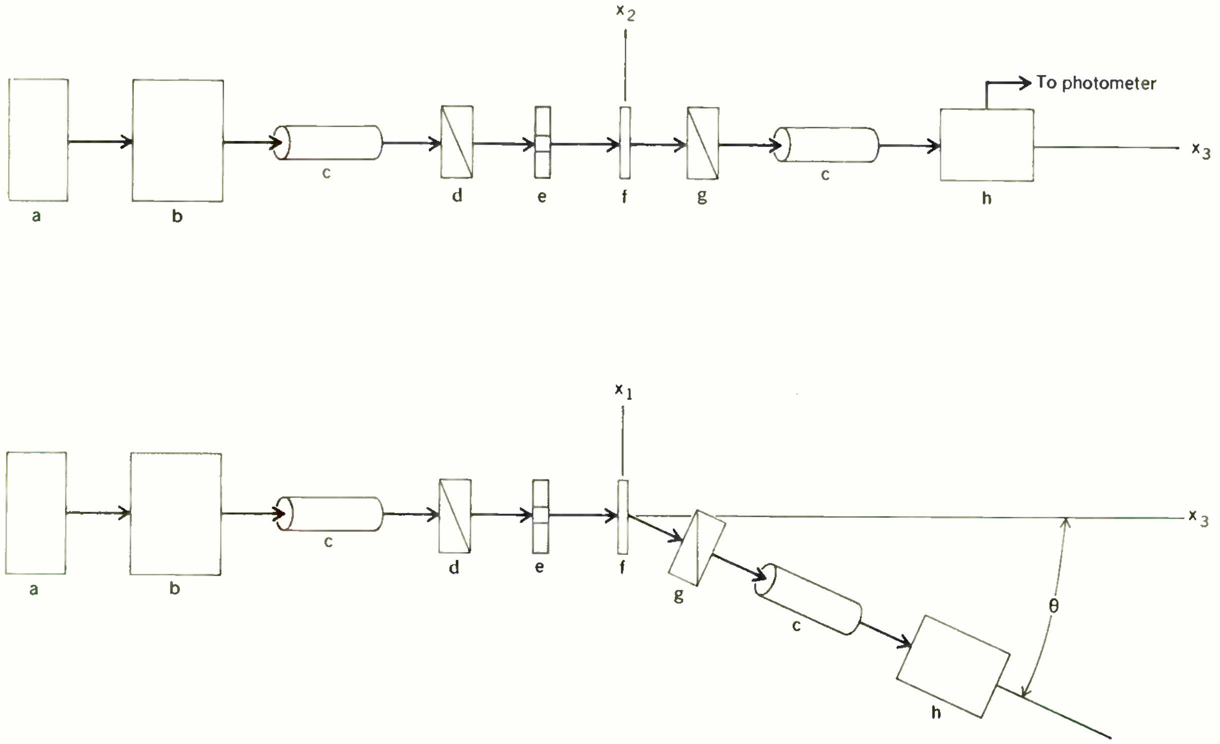
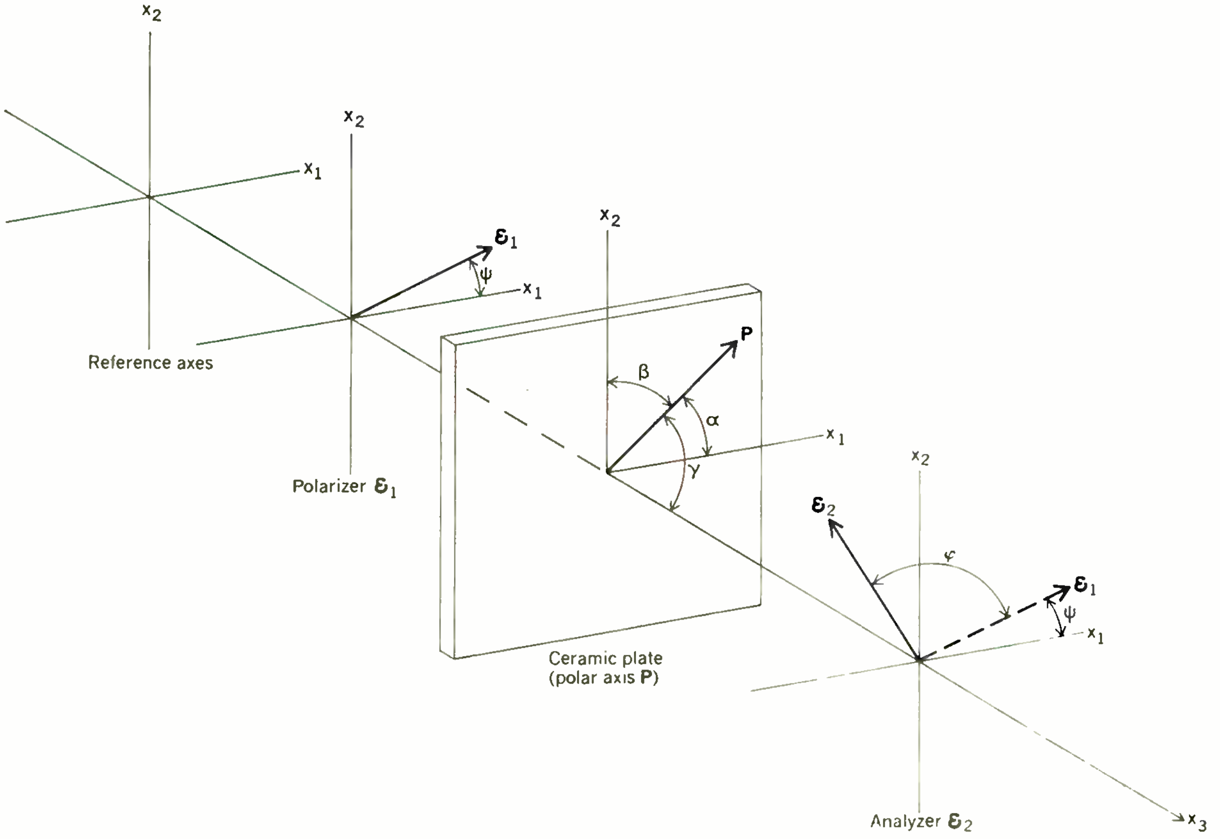


FIGURE 3. Side view (upper illustration) and top view (lower illustration) of the ellipsometer and accessories used for scattering measurements. Components are: (a) light source, (b) monochrometer, (c) collimator, (d) polarizer, (e) aperture, (f) ceramic sample, (g) analyzer, (h) photomultiplier.

FIGURE 4. Orientation of the ceramic polar axis P, and the optical polarization axes $\vec{\mathcal{E}}_1$ and $\vec{\mathcal{E}}_2$ relative to Cartesian coordinates.



electric field are usually broken up into multiple domains.[4] However, when a very large electric field is applied, all favorably oriented domains in each grain grow in size at the expense of the other domains. This is an irreversible process in that the titanium ions do not all return to their original face-center locations when the large electric field is removed. Consequently, the material retains a net remanent polarization (P_M in Fig. 1), a macroscopic elongation, and other anisotropic properties, even when the large polarizing field is no longer applied.

It is possible, of course, to obtain a remanent polarization P_r that is intermediate between the maximal values, $+P_M$ and $-P_M$, in Fig. 1. This state is achieved by supplying the polarizing charge through a very large current-limiting resistance (say several thousand megohms) and then turning off the trickle of current at the instant the polarization reaches point A on the hysteresis in Fig. 1. In this way, all the domains will not complete their switching to point B. The ratio P_r/P_M is called the normalized remanent polarization. This ratio, which can vary between -1 and $+1$, is a fundamental variable characterizing the condition of the ceramic.

General remarks

Thin ferroelectric ceramic plates (< 0.1 mm) possess two types of interesting electrooptic effects: electrically controlled scattering and electrically controlled birefringence. These two effects are said to be electrically controllable because each is dependent either on polarization magnitude and electric field bias [(P, E) of Fig. 1] or on remanent polarization orientation. Whether scattering or birefringence is more noticeable depends on the average grain size of the ceramic under consideration.

One composition that has been thoroughly studied for electrooptic behavior is the hot-pressed, lead-zirconate-titanate ceramic $[Pb_{0.99}Bi_{0.02}(Zr_{0.65}Ti_{0.35})_{0.98}O_3]$.[5-7] (All electrooptical measurements presented here refer to that composition, which we designate PZT 65/35.) In PZT 65/35, with average grain size less than 2 μm, birefringence is the more important effect. For average grain size greater than 2 μm, scattering becomes more significant.

For this composition, grain size depends on manufacturing conditions during hot pressing.[1,2] Temperatures between 800°C and 1300°C, pressures between 7.0 million and 55.0 million newtons per square meter and durations of 1 to 64 hours have been tried. In general, fabrication temperature is the dominant parameter governing grain growth: If the temperature is below 1100°C, small grain size (< 2 μm) results, whereas temperatures above 1100°C lead to large grain size (> 2 μm).

If one is clever enough to generate a switching field of the proper shape, both small- and large-grained materials have the property that localized areas as small as 25 μm by 25 μm can be switched independently of surrounding areas. These locally switched areas are stable with time in that they retain their remanent polarization after the switching field has been removed. They can, however, be restored to their original state by applying a sufficiently large switching field with reverse polarity.[5-7]

Since electrooptic scattering and birefringence depend on polarization, ferroelectric ceramics make it practical to construct a number of interesting electrooptic devices. By judicious choice of material, electrode configuration, and incident illumination spectrum, it is possible to devise, for example, electrically controlled light valves and shutters, spectral filters, optical memories, light modulators, and controlled-persistence black-and-white or multicolor displays.

Electrically controlled light scattering

Consider the experimental arrangement sketched in Fig. 3, which shows a block diagram of an optical ellipsometer. We used this ellipsometer for making scattering measurements on large-grained ferroelectric ceramics. As shown in Fig. 3, collimated monochromatic light passes through a plane polarizer (d) and a 1-mm aperture (e). The light then impinges normally on the major face of a ferroelectric ceramic plate (f). Upon emerging from the plate, the light passes through an analyzer (g)—i.e., a second polarizer—through a second collimator (c), and to a photomultiplier (h). (The angular aperture of our photomultiplier was 0.65°.)

The intensity of light emerging from the ceramic plate was measured as a function of the angle θ shown in Fig. 3. Several other angles are also relevant in this measurement and these are defined in Fig. 4: ψ is the angle between the polarizer axis $\vec{\mathcal{E}}_1$ and the x_1 axis, and φ is the angle between the analyzer axis $\vec{\mathcal{E}}_2$ and the polarizer axis $\vec{\mathcal{E}}_1$*. Additionally, the ceramic polar axis (i.e., the remanent polarization direction of the ceramic) is oriented at angles α, β, and γ, to the x_1, x_2, and x_3 axes, respectively.

Scattering profiles of transmitted light intensity I vs. the angle θ were first measured with the PZT 65/35 ceramic plate polarized normally to its plane—i.e., along x_3. Results for optical polarizer and analyzer parallel ($\psi = 0°$, $\varphi = 0°$) and perpendicular ($\psi = 0°$, $\varphi = 90°$) are shown in Fig. 5.† It is apparent from this graph that plane polarization of light transmitted through this plate is detectable only for scattering angles of less than one degree. There is scattering at larger angles, but hardly any preservation of polarization.

Scattering profiles were also measured with the plate polarized in its own plane, along x_2. Results in this case for the optical polarizer and analyzer parallel ($\psi = 0°$, $\varphi = 0°$) and perpendicular ($\psi = 0°$, $\varphi = 90°$) are shown in Fig. 6. In this case, plane polarization is detectable only for large scattering angles.

It is interesting to compare the results in Figs. 5 and 6, as these data form the basis for some attractive applications. First note the difference in the scales of the two curves: The peak transmitted intensity for **P** along x_3 is about seven times higher than the rather broad maximum observed with **P** along x_2. Also, the **P**-along-x_3 maximum occurs much closer to the x_3 axis than the **P**-along-x_2 maximum: $\theta = 2°$ vs. $\theta = 15°$.

Two other parameters to consider in making scattering measurements are plate thickness, τ, and light wavelength,

* Limitations of printing prevent the setting of $\mathcal{E}$ in boldface type to represent a vector. The arrow is used instead.

† In Fig. 5 and all figures following, the plotted intensity $I(\theta)$ is multiplied by $\sin\theta$ to make the data more compatible with photodetector applications: The area under an $I(\theta)\sin\theta$ curve between 0 and θ_0 is proportional to the light a detector of half-angle θ_0 would intercept when centered on the x_3 axis. The intercepted power is

$$\iint\limits_{\substack{\text{aperture}\\\text{area}}} I(\theta)\,dA = 2\pi(x_3 \sec\theta_0)^2 \int_0^{\theta_0} I(\theta)\sin\theta\,d\theta$$

λ. The data for Figs. 5 and 6 are based on $\tau = 80\ \mu m$ and $\lambda = 0.55\ \mu m$. Some indication of the effect of τ and λ on the scattering profiles is given by Figs. 7 and 8. Figure 7 shows $I(\theta) \sin \theta$ vs. θ for various values of τ when $\lambda = 0.55\ \mu m$, whereas Fig. 8 shows $I(\theta) \sin \theta$ vs. θ for various values of λ when $\tau = 60\ \mu m$.

Generally speaking, for large-grained ceramics, the peak of $I(\theta) \sin \theta$ is lower and occurs at a larger scattering angle as plate thickness increases, as wavelength decreases, as angle γ of P orientation increases, and as grain size increases.

Devices utilizing electrically controlled scattering

The most elementary device utilizing electrically controlled scattering is a light valve of the type shown in Fig. 9. The gap between the two electrode regions is the optically active area of the valve that either transmits or blocks light. This valve is considered to be in the ON state when polarized perpendicularly to its plane by electrode voltages of the polarity shown in the upper half of Fig. 9. Such a polarization state permits maximum light transmission through the electrode gap and into the detector that intercepts only light scattered at a small angle. The OFF state, on the other hand, is obtained by polarizing the ceramic in its own plane using the electrode polarities shown in the lower half of Fig. 9. The OFF polarization state minimizes small-scatter-angle light transmission through the gap and into the detector.

Two-valued light valves of the type just described may be combined in two-dimensional arrays on a single plate to form binary memories or visual displays. In order to change the information stored on such a plate, some method of electrically selecting the (x_1, x_2) addresses of the various light valves is required. Figure 10 illustrates one scheme for accomplishing this addressing while storing and erasing information. In this figure, the solidly outlined electrode configuration represents the x_2 address of the various valves. This electrode configuration is on the upper surface of the plate. Similarly, the dashed-outline electrode configuration represents x_1 addresses and is located on the lower surface of the plate.

Prototype memories and displays of this type that can store 256 bits on a plate 5.0 mm by 6.3 mm have been built. One of these prototypes is shown in Fig. 11. If this array is to be used as a visual display, merely illuminate it from

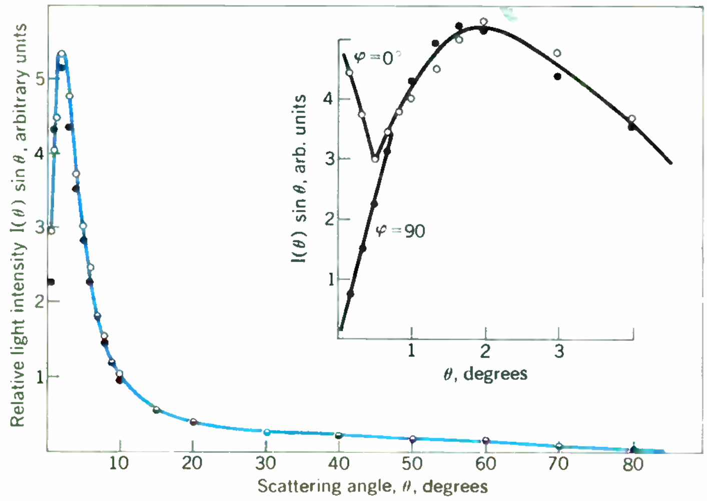


FIGURE 5. Scattering profile of light transmitted by a large grained PZT 65/35 ferroelectric ceramic plate polarized normal to the major surfaces ($\gamma = 0°$, $\psi = 0°$). This figure is based on a sample with thickness $= 80\ \mu m$ and average grain size $= 5\ \mu m$, with light wavelength $= 0.55\ \mu m$.

FIGURE 6. Scattering profile of light transmitted by a large-grained PZT 65/35 ferroelectric ceramic plate poled parallel to its major faces ($\beta = 0°$, $\psi = 0°$). This figure is based on the same sample, wavelength, and intensity units as Fig. 5.

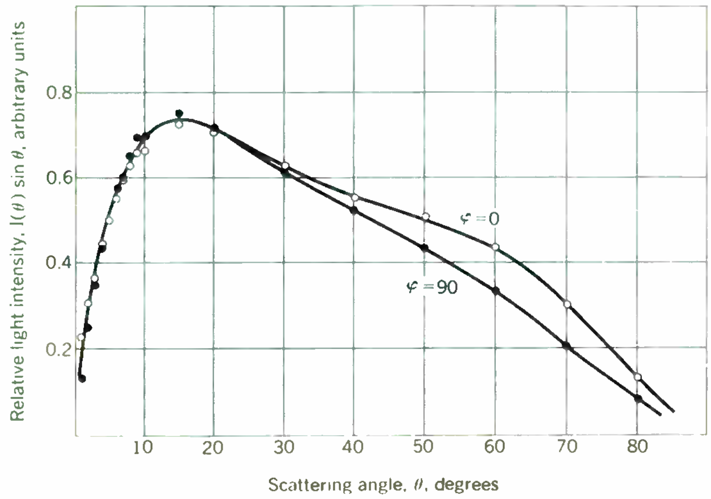
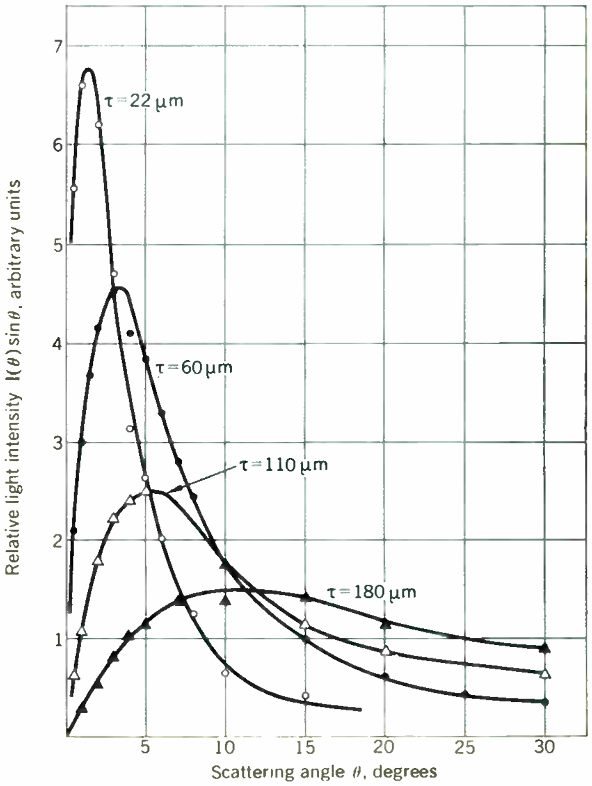


FIGURE 7. Scattering profile variation with plate thickness for large-grained PZT 65/35 ceramic. These curves are based on light wavelength $\lambda = 0.55\ \mu m$ and average material grain size $= 5\ \mu m$, with $\gamma = \varphi = 0$. The vertical scale is not related to that used in Figs. 5 and 6.



behind after the desired information has been written in. Alternatively, if the array is to be used as an optical memory, readout might be achieved, for example, by illuminating the entire array from behind and placing a photodiode in front of each valve.

It is possible to define, as a function of photodetector half-angle aperture θ_0, an ON/OFF ratio for these valves. This ratio is the light received by the photodetector when the valve is ON divided by the light received when the valve is OFF:

$$\chi(\theta_0) = \int_0^{\theta_0} I_{\rm ON}(\theta) \sin \theta \, d\theta \Big/ \int_0^{\theta_0} I_{\rm OFF}(\theta) \sin \theta \, d\theta \quad (1)$$

Values of $\chi(\theta_0)$ as high as 10 are fairly easy to realize experimentally with half-angle apertures θ_0 of the order of 1°, light wavelength 0.55 μm, and plate thickness 60 μm. The dependence of the ON/OFF ratio χ on θ_0 can be determined for all θ_0 by substituting into Eq. (1) the experimental curves for $I(\theta) \sin \theta$ from Figs. 5 and 6.

Electrically controlled birefringence

Birefringent materials are those in which the velocity of light depends on the orientation of the optical electric vector—that is, on the plane of polarization of the light. In other words, the dielectric constant of a birefringent material is anisotropic at optical frequencies.

Small-grained ferroelectric ceramics are optically uniaxial with the optical and remanent polarization axes coincident. This means that light polarized in a plane perpendicular to the ceramic polar axis propagates with a velocity c/n_o, whereas light polarized parallel to the ceramic polar axis propagates with a different velocity, c/n_e. The quantities n_o and n_e are called the ordinary and extraordinary indexes of refraction, respectively. In

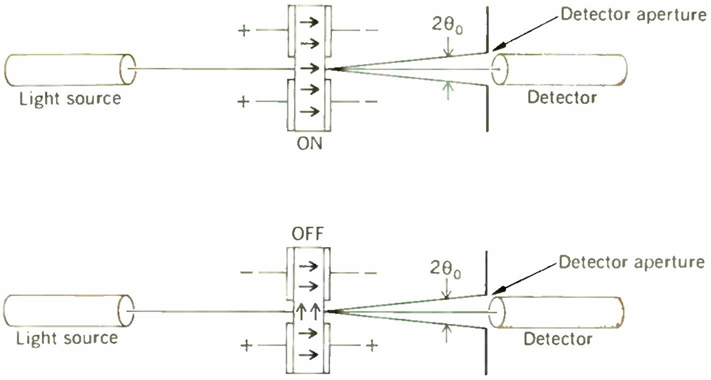


FIGURE 9. Electrode configuration and switching polarities for the two states of the large-grained ceramic light valve. The state illustrated at the top permits transmission of light into the detector, whereas the state illustrated below blocks light.

FIGURE 8. Scattering profile variation with light wavelength for large-grained PZT 65/35 ferroelectric ceramic. Curves are based on a plate of thickness = 60 μm and average grain size of 5 μm, with $\gamma = \varphi = 0°$. The vertical scale has been adjusted so that the $\lambda = 0.55$ μm curve of this figure corresponds to the $\tau = 60$ μm curve of Fig. 7.

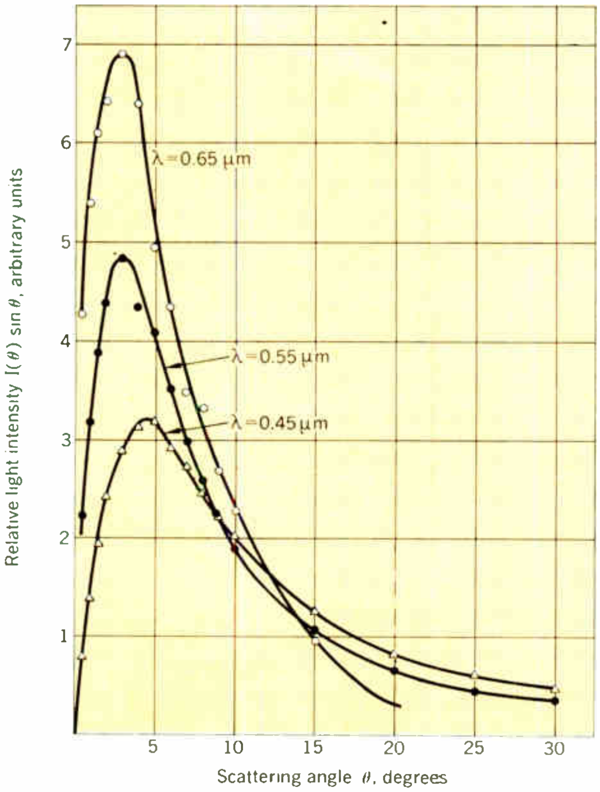
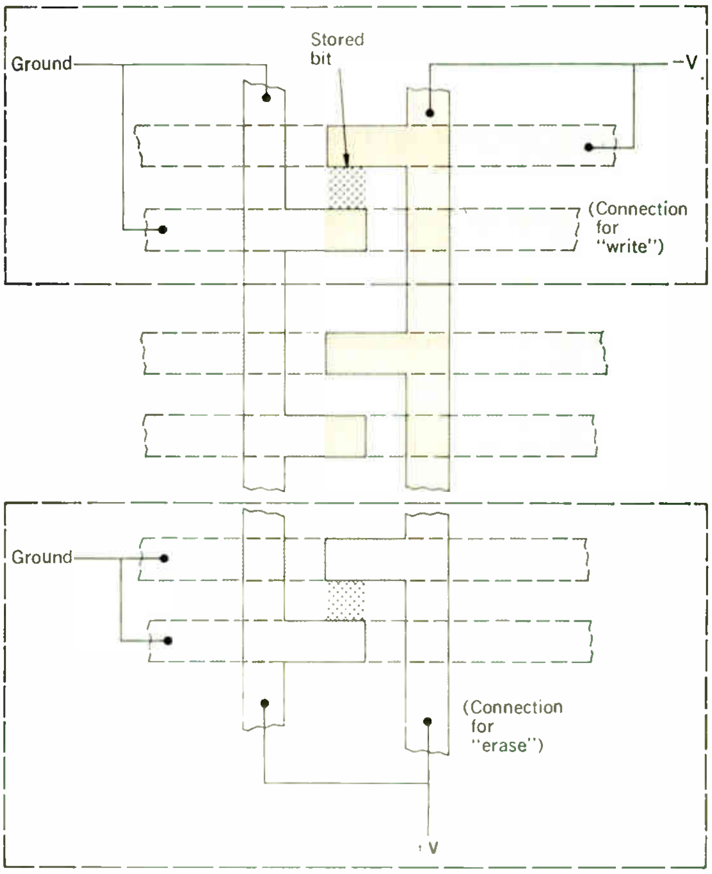


FIGURE 10. Electrode arrangement for selecting the (x_1, x_2) addresses of the bit locations in the memory or display array. The solid lines represent electrodes on the upper surface of the array, and give the x_2 address; the dashed lines represent electrodes on the under surface, and give the x_1 address.



ferroelectric ceramics, n_e is usually less than n_o; such ceramics are said to be negatively birefringent.

Birefringence in small-grained PZT 65/35 ferroelectric ceramics was measured with essentially the same experimental arrangement used for scattering measurements in large-grained ceramics (illustrated in Figs. 3 and 4). The one addition was a compensator between the polarizer (*d*) and aperture (*e*) in Fig. 3.

A measure of the magnitude of birefringence in some particular optically uniaxial plate of thickness τ is the retardation Γ. As the optical polarization angle (ψ in Fig. 4) is rotated, the thickness of a uniaxial birefringent plate (expressed in optical wavelengths) will pass through a maximum and a minimum. The difference between the maximum and the minimum is defined to be the retardation Γ. This Γ may be shown to have a $\sin^2 \gamma$ dependence on γ, the angle between the optical and x_3 axes (Fig. 4):

$$\Gamma = \Gamma_{\max} \sin^2 \gamma \qquad (2)$$

The quantity $\Gamma_{\max}$ is simply the number of optical wavelengths across the plate for **P** in the plane of the plate and $\vec{\mathcal{E}}_1$ parallel to **P** minus the number of optical wavelengths across the plate for **P** in the plane of the plate and $\vec{\mathcal{E}}_1$ perpendicular to **P**:

$$\Gamma_{\max} = \frac{(n_e - n_o)\,\tau}{\lambda_{\text{vacuum}}} = \frac{\Delta n \tau}{\lambda_{\text{vacuum}}} = \frac{\Delta n \omega \tau}{2\pi c} \qquad (3)$$

Here, ω is the angular frequency of light, and quantity Δn is called the effective birefringence. Consequently, the retardation for arbitrary γ is

$$\Gamma = \frac{\Delta n \omega \tau}{2\pi c} \sin^2 \gamma \qquad (4)$$

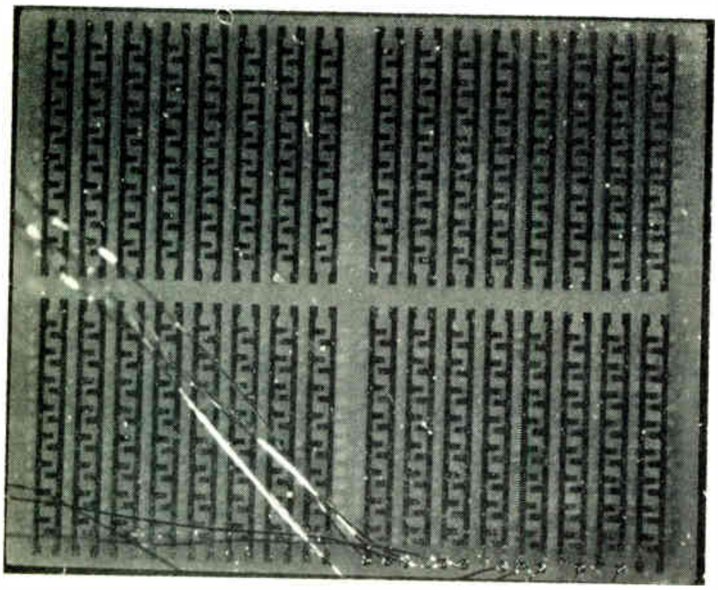


FIGURE 11. Prototype memory unit using a large-grained ceramic plate and the electrode logic system illustrated in Fig. 10. Four sixty-four bit words are stored on a 5.0- by 6.3-mm plate 100 μm thick. Bit size is 50 μm by 50 μm.

FIGURE 12. Effective birefringence of small-grained PZT 65/35 ferroelectric ceramic plates as a function of vacuum wavelength. Data are shown for three different samples. The plates had an average grain size of about 1 μm.

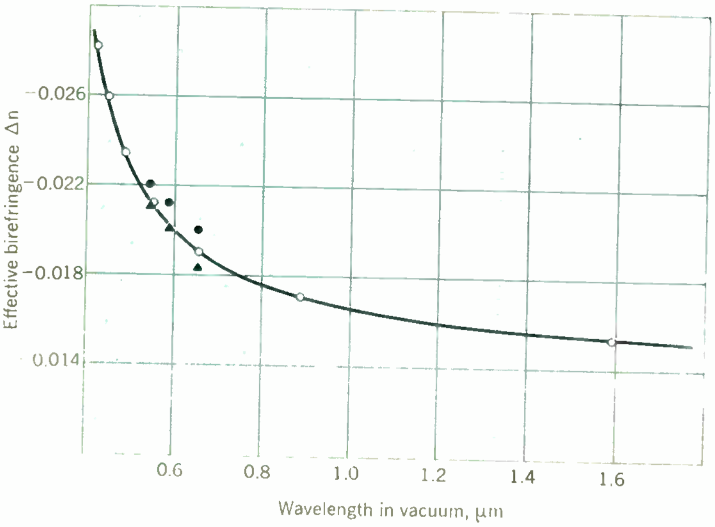
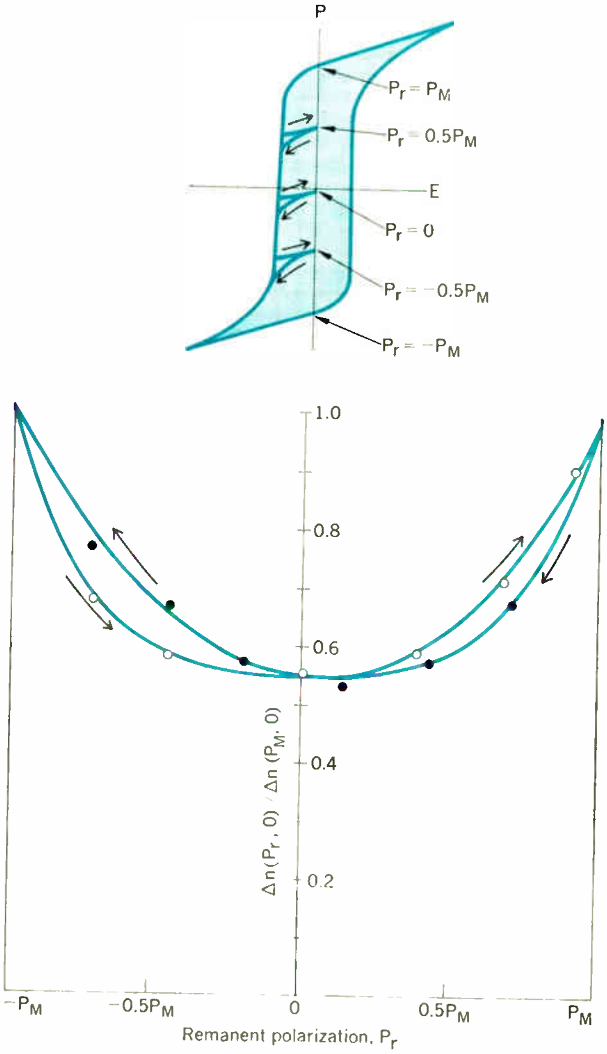


FIGURE 13. Normalized effective birefringence $\Delta n(P_r, 0)/\Delta n(P_M, 0)$ vs. remanent polarization P_r at zero-bias field, $E = 0$. Figure is based on light wavelength of 0.656 μm, average grain size of 1 μm, and $\Delta n(P_M, 0)$ of $-.0214$. Polarization values between $+P_M$ and $-P_M$ are achieved as shown on the hysteresis loop at the top of the figure.



Small-grained, ferroelectric ceramic plates are characterized by an effective birefringence that depends on the vacuum wavelength, on the remanent polarization ($P = P_r$ in Fig. 1), and on the biasing field (E in Fig. 1). The dependence of Δn on λ is illustrated in Fig. 12. Observe that Δn is a decreasing function of λ over the visible and near-infrared spectrum.

A convenient method for presenting data on Δn vs. remanent polarization P_r and bias field E is the normalized form, $\Delta n(P_r, E)/\Delta n(P_M, 0)$. Figure 13 shows the normalized effective birefringence $\Delta n(P_r, 0)/\Delta n(P_M, 0)$ vs. remanent polarization for no bias field. This figure is based on a light wavelength of $\lambda = 0.55\ \mu$m.

The only way to obtain anhysteretic polarization changes with a large electric field is to start with a sample maximally poled (at P_M in Fig. 1) and then apply the large field in the direction of P_M. The dependence of $\Delta n(P_M, E)/\Delta n(P_M, 0)$ on E under these conditions is shown in Fig. 14. The magnitude of this dependence is such that a field along P_M of 13.5 kV/cm will produce a retardation ¼ wavelength greater than a zero field in a plate 63 μm thick. This statement applies for P_M and E in the plane of the plate ($\gamma = 90°$), for an average grain size of 1 μm, and for a light wavelength of 0.65 μm. Figure 14 indicates that the incremental birefringence due to E in this case is approximately proportional to E^2.

Devices utilizing electrically controlled birefringence

Applications for electrically controlled birefringence include light modulators, electrically controlled shutters, optical memories, and spectral filters.[6,7]

A configuration basic for these devices is illustrated in Fig. 15. The ceramic is variably polarized or biased along the x_2 axis ($\beta = 0$), and the light is polarized at $\psi = 45°$. The analyzer is oriented at $\varphi = 90°$.

Consider first what happens to monochromatic light incident on the ceramic. The optical electric vector is parallel to $\vec{\mathcal{E}}_1$, and may be resolved into components along and perpendicular to **P**:

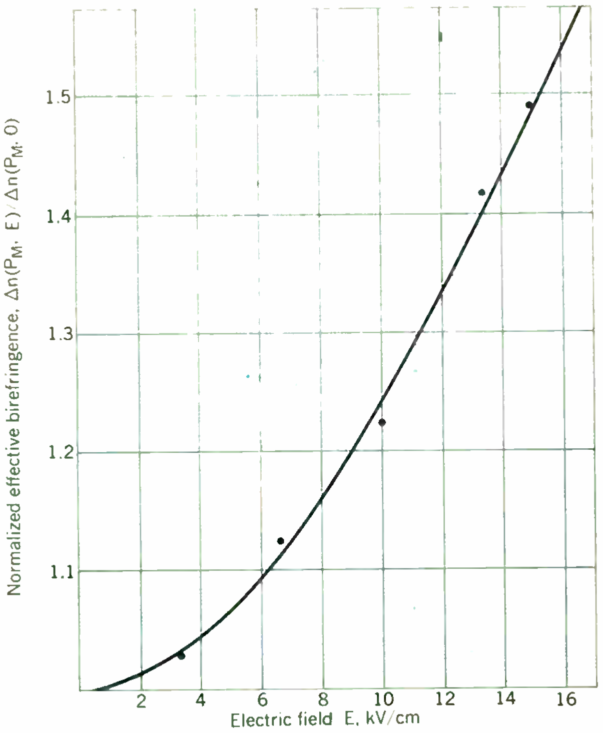


FIGURE 14. Normalized effective birefringence $\Delta n(P_M, E)/\Delta n(P_M, 0)$ vs. bias field $+E$ for E applied with the sample at maximum positive remanent polarization P_M. This graph is based on a wavelength of 0.656 μm, average grain size of 1 μm, and $\Delta n(P_M, 0)$ of -0.0214.

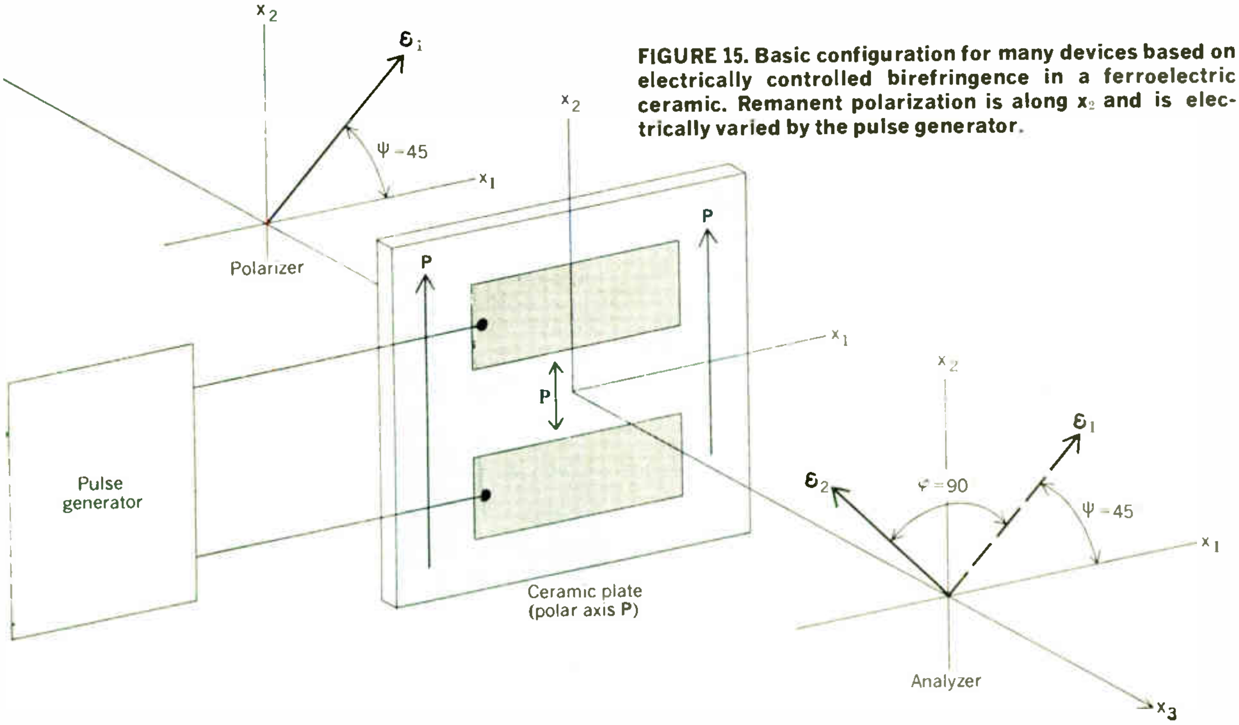


FIGURE 15. Basic configuration for many devices based on electrically controlled birefringence in a ferroelectric ceramic. Remanent polarization is along x_2 and is electrically varied by the pulse generator.

$$\vec{\mathcal{E}} = \frac{(\mathbf{i}_{\parallel} + \mathbf{i}_{\perp})\mathcal{E}}{\sqrt{2}} \tag{5}$$

The component parallel to **P** will propagate with a velocity c/n_e, whereas the orthogonal component propagates with a velocity c/n_o,

$$\vec{\mathcal{E}}(x_3) = \frac{\mathcal{E}}{2}\left\{\mathbf{i}_{\parallel} \exp\left[i\omega\left(t - \frac{n_e x_3}{c}\right)\right]\right.$$
$$\left. + \mathbf{i}_{\perp} \exp\left[i\omega\left(t - \frac{n_o x_3}{c}\right)\right]\right\} \tag{6}$$

Consequently, the light emerging from the plate, referred to the $\vec{\mathcal{E}}_1$ and $\vec{\mathcal{E}}_2$ axes is

$$\vec{\mathcal{E}}(\tau) = \frac{\mathcal{E}}{\sqrt{2}} \exp i\omega t\left\{\mathbf{i}_1\left[\exp\left(\frac{-in_e\omega\tau}{c}\right) + \exp\left(\frac{-in_o\omega\tau}{c}\right)\right]\right.$$
$$\left. + \mathbf{i}_2\left[\exp\left(\frac{-in_e\omega\tau}{c}\right) - \exp\left(\frac{-in_o\omega\tau}{c}\right)\right]\right\} \tag{7}$$

$$= \mathcal{E} \exp i\omega\left[t - \frac{(n_e + n_o)\tau}{2c}\right]$$
$$\cdot\left[\mathbf{i}_1 \cos\frac{\omega(n_e - n_o)\tau}{2c} - i\mathbf{i}_2 \sin\frac{\omega(n_e - n_o)\tau}{2c}\right]$$

The light energy transmitted by the analyzer is proportional to the absolute value of the squared $\vec{\mathcal{E}}_2$ component of $\vec{\mathcal{E}}$:

$$T = \mathcal{E}^2 \sin^2 \omega\frac{\Delta n\tau}{2c} \tag{8}$$

or, from Eq. (3),

$$T = \mathcal{E}^2 \sin^2 \Gamma_{\max}\pi \tag{9}$$

Thus, if the plate is illuminated by monochromatic light, and if the remanent polarization or bias field is adjusted so the retardation is an integer, that is, $\Gamma_{\max} = N$, all light will be extinguished by the system. On the other hand, if $\Gamma_{\max}$ is adjusted to be an integer minus a half, $\Gamma_{\max} = N - \frac{1}{2}$, a maximal amount of light is transmitted by the system.

By incrementally varying the remanent polarization along x_2 between the states corresponding to $\Gamma_{\max} = \frac{1}{2}$ and $\Gamma_{\max} = 1$, it is possible to achieve a multistate light valve with gray scale. Alternatively, by picking a plate thickness or light wavelength for which $\Gamma_{\max} = 1$ at $P = P_M$ and $E = 0$, it is possible to obtain a momentarily opened shutter when the ceramic is pulsed with a biasing field. (The biasing field may, for example, effect a transition from the $\Gamma_{\max} = 1$ state to the $\Gamma_{\max} = 3/2$ state.)

If the ceramic is illuminated by white rather than monochromatic light, those wavelengths for which $\Gamma_{\max} = \frac{1}{2}$ will be preferentially transmitted. (This, of course, assumes plate thickness is selected so that $\Gamma_{\max}$ does not significantly exceed one for any visible wavelength.) Consequently, the arrangement shown in Fig. 15 can spectrally filter white light to yield color. Moreover, as the wavelength at which $\Gamma_{\max} = \frac{1}{2}$ is electrically variable, the color that the arrangement transmits can be electrically varied.

These ceramic light valves or spectral filters can be combined in arrays to provide optical memories or display units. For example, monochromatic light the use of to illuminate the array from behind makes it possible to obtain a single-color display. Such a display would essentially be a screen with spatially and temporally variable transmissivity. On the other hand, illumination of the array with white light would permit a multicolor display—such as might conceivably be used as a new foundation for color television.

This work was supported by the United States Atomic Energy Commission and by the National Aeronautics and Space Administration.

REFERENCES

1. Haertling, G. H., and Zimmer, W. J., "An analysis of hot-pressing parameters for lead zirconate-lead titanate ceramics containing two atom percent bismuth," *Am. Ceram. Soc. Bull.*, vol. 45, pp. 1084–1089, Dec. 1966.

2. Haertling, G. H., "Improved ceramics for piezoelectric devices," *WESCON/66 Technical Papers*, vol. 10, pt. 3, paper 3/1.

3. Mason, W. P., "Electrostrictive effect in barium titanate ceramics," *Phys. Rev.*, vol. 74, pp. 1134–1147, Nov. 1948.

4. Känzig, W., "Ferroelectrics and ferroelectronics," *Solid State Physics*, vol. 4, F. Seitz and D. Turnbull, eds. New York: Academic Press, 1947, pp. 97–124.

5. Land, C. E., "Ferroelectric ceramic storage and display devices," 1967 International Electron Devices Meeting, Washington, D.C.; Sandia Laboratories reprint SC-R-67-1219, Oct. 1967.

6. Land, C. E., and Thacher, P. D., "Ferroelectric ceramic electrooptic materials and devices," *Proc. IEEE*, vol. 57, pp. 751–768, May 1969; also Sandia Laboratories research rept. SC-RR-78-866, Dec. 1968.

7. Thacher, P. D., and Land, C. E., "Ferroelectric electrooptic ceramics with reduced scattering," *IEEE Trans. Electron Devices*, vol. ED-16, pp. 515–521, June 1969.

Cecil E. Land (SM) has been on the staff of the Sandia Laboratory, Albuquerque, N. Mex. for the past 14 years, working in ferroelectric ceramics for the latter nine. Previously, he was at the Electronic Division of the Westinghouse Electric Corporation in Baltimore, Md. (1949–1956). After moving to Sandia, he first designed weapons systems and associated instrumentation. Mr. Land received the B.S.E.E. degree from Oklahoma State University (1949) and has since taken graduate courses at the University of Maryland and the University of New Mexico. A registered engineer, Mr. Land is also a member of the American Physical Society and the American Ceramic Society.

Richard Holland (M) began collaborating on this article while with the Sandia Laboratory, Albuquerque, N. Mex. (1964–1969). During this period he designed piezoelectric devices and optical techniques for nanosecond photography of microscopic elastic waves, and worked at radiation-hardening of semiconductor components. He later transferred to Jet Propulsion Laboratory, Pasadena, Calif. As a senior engineer there, he is developing parabolic spacecraft-borne antennas. Mr. Holland received the B.S., S.M., and Ph.D. degrees from M.I.T. in 1962, 1964, and 1966 respectively. His contributions to the technical literature include a book, "Design of Resonant Piezoelectric Devices," (M.I.T. Press, 1969), and some 25 journal articles.

Data transmission: a direction for future development

Because the public telephone network was designed only for voice communication, problems arise when data communication is attempted on this same network. Data waveforms can be converted for ease of handling but it is difficult to match the discontinuous flow of data from a terminal to the continuous flow of information in the communication channel

Harry Rudin, Jr. **IBM Zurich Research Laboratory**

Despite the rapid advances in many regions of data transmission, there is a rapidly growing number of applications for which existing data-transmission techniques are inefficient. A look at the status of data-transmission development indicates that, although a very successful campaign has been waged to map the data waveform into a waveform ideally suited for transmission on the communication channel, very little has been done to match the often discontinuous flow of data from the terminal to the continuous flow of information in the channel. Combining randomly occurring messages from several sources into a more continuous flow is described by the mathematics of traffic theory. Although this theory has been extensively applied to speech traffic, it is rarely applied to data traffic. As a specific example of the gains in channel efficiency that can be had through the application of traffic theory, the multiplexer–concentrator is examined. In the author's opinion, it is in this area of application of traffic theory to data communication that many of the more significant future developments in data transmission will be made. To be sure, some work has been done, but it is relatively little when the gains that can be made are considered.

The public telephone network, naturally enough, was designed only with voice communication in mind. The result of this design is that efficient data communication over the telephone network is difficult to achieve because a number of factors that do not affect the quality of voice transmission seriously impede the flow of data. Delay distortion in the frequency domain is an example of one of these phenomenons.

A tremendous research and development effort has been made to overcome many of the restrictions imposed by the voice-communication network. The bit rate (and the reliability) at which data can be transmitted has been markedly increased within the last few years as the result of such technological advances. But the strange and fascinating aspect of this effort is that it has been carried out almost exclusively in the domain of signal design. That is, research has been almost exclusively concentrated on the development of techniques that convert the data waveform into another waveform that will easily pass through the transmission channel.

However, the public telephone network was designed not only with the voice waveform in mind, but with the *statistics* of voice communication in mind as well. Although it has been well-recognized that the telephone network is not suitable for transmission of many pure data waveforms, the assumption is usually made (albeit tacitly) that the statistics of data communication allow direct transmission of data over the telephone network. This—more often than not—is a bad assumption.

Traffic theory is the study of the statistical flow of messages and this theory is largely responsible for economical voice transmission as it exists today. In fact, the development of the theory was motivated mainly by telephony itself. Unfortunately, traffic theory is only now beginning to be applied to the data-communication problem.

In this article, the present direction of data-communication development and possibility for improvement is assessed in a cursory fashion. Next, the efficiency of conversational and interactive data-communication systems is examined. Finally, an example (slightly rigged to be sure) is considered to illustrate the gains that can be made in certain circumstances. The point of the example is to show that there are cases in which tremendous increases in channel efficiency can be made by taking traffic statistics into account.

Recent developments in data transmission

Recent developments in data transmission can be divided into three areas: modulation techniques, equalization techniques, and error control.

The term "modulation techniques" is used in the most general sense—i.e., the modulation process is that process which maps the data signal (often digital) into a form that is suitable for transmission over whatever transmission facility is at hand. It might include serial-to-parallel conversion, translation from baseband to a higher frequency, and band-limiting methods. There are two techniques very much in vogue.

The first modulation technique is multilevel baseband transmission combined with vestigial sideband modulation.[1,2] Multilevel transmission permits several bits to be sent at the same time so that higher speeds may be achieved. The vestigial sideband modulation is a good, practical approximation to achieve the spectrum utilization efficiency of single-sideband modulation.

A second technique is that of partial response.[3] In this technique a correlation is introduced between what is transmitted at one signaling instant and one or more other signaling instants, with the result that various spectral shapes can be obtained. Practically, the advantage of this family of techniques is that it permits transmission at new combinations of speed and sensitivity to noise.[4]

The developments in modulation techniques have permitted very efficient design of modems that convert the data signal into a form ideally suited to a specific, known communication channel (the average telephone channel, for example). Unfortunately, the difference between a specific channel and the average channel can be quite large, and this difference (in addition to manufacturing tolerances) was the major hindrance to efficient data communication until a few years ago. The development of a variety of automatic equalization techniques has since removed that hindrance.[5-7] Using these techniques, it is possible to compensate automatically for the variations of the individual channel and also the manufacturing tolerances.

One of the more up-to-date devices for this application is the transversal filter. It provides a dynamic and flexible approach to the problem of equalization and is well-suited to adaptive as well as automatic operation, qualifying it for use when distortion is time variable.

A very large percentage of the literature devoted to communications and information theory concerns various coding and decoding techniques for the detection and correction of errors incurred in the process of data transmission. It is possible to design a data-transmission system that operates with a very high speed but with an unacceptably high error rate.[8] However, by using a small percentage of the transmitted information for the insertion of redundancy, a disproportionately large gain can be made in the reduction of the error rate. Unfortunately, the coder and, particularly, the decoder in the vast majority of cases described in the literature are prohibitively expensive. However, these techniques can be both practical and economically reasonable.[9]

The techniques just described have been put into practice in the case of the telephone channel, and the results are truly impressive. For example, it is possible to transmit at the rate of 9600 bits per second (b/s) over carefully prepared leased lines.[10] Also, very extensive

FIGURE 1. Simplified data-communication complex.

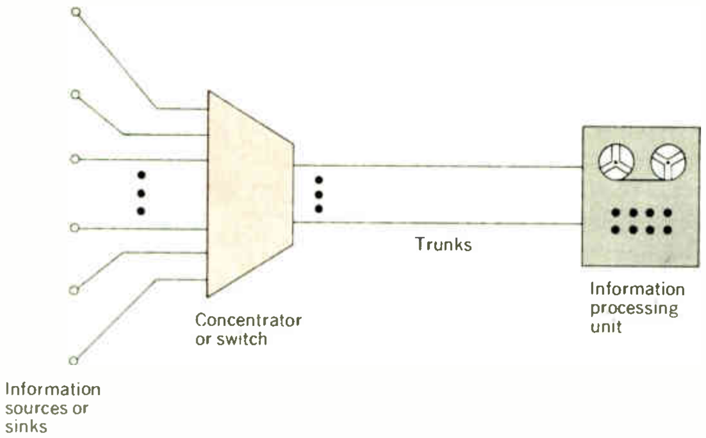
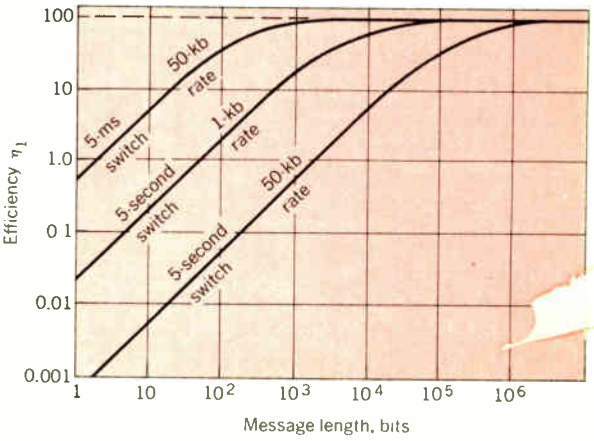


FIGURE 2. Efficiency of some data-communication systems—dial and transmit (computer processing included).



tests have been made over the Direct-Distance Dialed (DDD) network to prove the reliability of 3600-b/s transmission over unprepared facilities.[11] Experimental work indicates the possibility of still higher speeds over the DDD network. Unfortunately, these units are very complex and the complexity seems to increase rapidly with speed.

The question should be asked, "How much room for improvement is there?" Shannon's theory[12] provides the answer for error-free transmission in a white-noise-only environment in terms of the following equation:

$$C = W \log_2 (P/N + 1)$$

where W is the channel bandwidth in hertz and P/N is the signal-to-noise power ratio. Estimating the bandwidth of the voice channel to be 2400 Hz and the signal-to-noise ratio to be 30 dB results in a very rough approximation of the channel capacity: C equals 24 000 b/s. There is, then, room for improvement, but not very much. What progress there is to be made in this direction is apt to come at great cost. This is not to say that the advent of large-scale integration (and with it the promise of economical digital- and active-filter technologies, for example) will not sharply reduce the cost of modems even more complex than the ones produced today. However, there may well be a more profitable direction in which to carry out research.

In summary, much effort has been expended and much success has been had in matching the transmitted signal waveshape to the channel's characteristics. In the voice-communication-channel area, excellent work has been done in making the data waveform look like the speech waveform so that it will easily pass through the channel designed for speech.

Data-transmission system efficiency

As long as the flow of data is continuous, the conventional data-transmission system performs very well indeed. But when the flow of data becomes nonregular or random, the data-communication system can be rather inefficient. The vast majority of research and development efforts in data transmission have been devoted exclusively to the case of continuous or nearly continuous data flow. A look at one of several books devoted to data transmission will verify this.[13,14]

Unfortunately, there is a rapidly growing number of applications in which the flow of data is far from continuous. One example is found in a large inquiry system. Here, several remotely located or satellite terminals make requests to a central computer—perhaps to ascertain a customer's credit, to find out if a seat is available on an aircraft flight, or to check the inventory of a certain item. Both the request for information and the required information are usually very short messages; further, these short messages are very likely separated by long idle-time periods.

Another example of discontinuous data flow is the use of a time-shared computer, perhaps for scientific programming. In this case, the likelihood is that a communication channel will remain idle for long periods of time while the programmer or the computer ponders some aspect of a problem.

One way of characterizing the continuity of data communication is by means of message length. Given the message length, the bit rate of the communication channel, and the time required to set up the connection, it is possible to determine the efficiency of a communication system in terms of the percentage of time actually spent transmitting needed information. The efficiency η_1 is given by

$$\eta_1 = \frac{\text{Time required to transmit the information}}{\text{Total time required to make the connection and transmit the required information}}$$

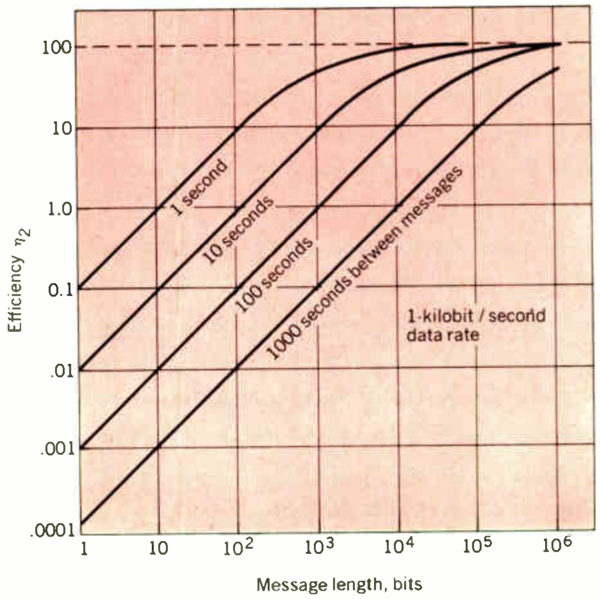
For the very simple, single-switch communication system shown in Fig. 1, η_1 can be plotted as a function of the message length for assumed switching time and rate of transmission as shown in Fig. 2. Two data rates are assumed: The 1-kilobit-per-second (kb/s) rate is taken as a fairly representative number for many data modems capable of operating on the Direct-Distance Dialed telephone network; the 50-kb/s rate is taken as representative of a wide-band data service or of a single pulse-code-modulation voice channel.

Two switching or interconnecting times are assumed: 5 seconds to represent a conventional, mechanical switching system and 5 ms to represent a futuristic solid-state switching system.[15] (It should be noted that these switching times include the processing time of a computer used to control the switch itself.) Combinations of these parameters and a specified message length determine the efficiency.* The results of this simple calculation are shown in Fig. 2. It is clear that in many cases of interest the efficiency is indeed low. Of course, this is not the only criterion of goodness that exists for the evaluation of a communication system.

In the face of such low efficiency, there is another path open to the user. Rather than reestablish the connection for each call (and thereby suffer the penalty of connect time), the user can establish the call only once and ac-

* The combination of 1-kb/s transmission speed and 5-ms switching time does not appear as it is an unreasonable combination. In the 5-ms period there is insufficient time to transfer the information necessary to establish the call, even at the 1-kilobit rate.

FIGURE 3. Efficiency of some data-communication systems—hold and transmit—at a 1-kb/s data rate.



cept, instead, the penalty of holding time between messages when the line is idle. A simple calculation can also be made for this case where the efficiency η_2 is now defined as follows:

$$\eta_2 = \frac{\text{Time required to transmit the information}}{\text{Sum of average holding time between transmissions and time required to transmit information}}$$

The efficiency η_2 is plotted as a function of the average message length for a number of assumed holding times between transmissions in Fig. 3. A data-transmission rate of 1 kb/s is assumed; another set of curves could be drawn for transmission at 50 kb/s with a corresponding degradation in efficiency. Again it is seen that the efficiencies are not high for many cases of interest.

The recognition of such inefficiencies is not new. In a forward-looking paper[16] published in 1961, the authors calculate a "system utilization index" or efficiency for a specific system of some 2 percent. Unfortunately, not a great deal has happened in the intervening period. For example, a very recent (and very timely) paper by Jackson and Stubbs gives a range of 1 to 5 percent for the communication efficiency of several multiaccess computer systems.[17]

So far, only the point of view of the communication user has been considered, but the present state of affairs is also a source of difficulty to the communication supplier. If the customer chooses to operate in the first mode (dial and transmit) and makes many such transactions, the computer that directs the switching in a modern exchange may become overloaded. If the customer chooses to operate in the second mode (hold and transmit) many trunks are blocked for very long periods of time. Thus there is pressure for change from the communication supplier as well.[18]

These inefficiencies are all a result of the nonuniform character of data flow or, in other words, the traffic statistics of data. The job of converting the often discontinuous data waveform into a continuous waveform suited to the transmission channel's capabilities has been well done. The job of converting the often discontinuous traffic pattern of data into the continuous flow of information in the transmission channel has barely been begun.

Some approaches to a solution

In order to discuss and compare several solutions for the problem of data communication with random-message flow, these solutions will be discussed with respect to a specific system—that shown in Fig. 1. The geographical distribution implied in Fig. 1 (a cluster of terminals far removed from a common destination) is chosen to emphasize the gains in efficiency that can be made if circumstances are ideal.

Although the geographical clustering of terminals in Fig. 1 may seem arbitrary, such clustering is commonly observed.[19] Yet another system might be one wherein communication costs are negligible and no extra hardware costs are warranted to reduce already negligible communication costs.

Obviously, what is needed to improve trunk-line efficiency is a technique for combining the communication needs of many terminals (when this is feasible) so that the total data-communication flow becomes more regular and the communication capacity of the trunks better utilized. In fact, this approach has been taken in some nonvoice transmission cases, but infrequently and for relatively long messages—telegraph messages, for example.[20]

The most straightforward approach is the use of a pure multiplexing technique. Multiplexing is here defined as a technique that assigns portions of the channel's capacity to various users on a fixed, a priori basis. A multiplexer has the same total input and output capacities determined on an instantaneous basis.

The channel capacity can be subdivided either in frequency or in time. Compared with transmission on an entire trunk, each terminal "sees" a reduced capacity. When this reduced capacity can be tolerated (for example, when possible resultant delays are reasonable), the restriction on transmission rate itself acts as a filter to smooth the varying traffic flow. Since several terminals can use the same trunk, utilization or efficiency is higher.

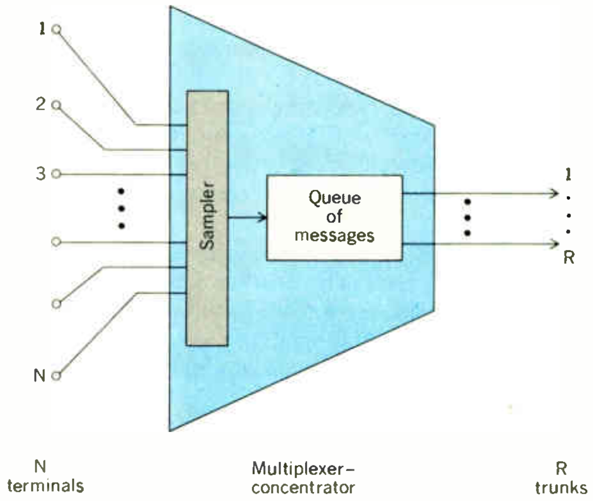
Frequency-division multiplex (FDM) is a well-used technique in the telephone system.[21] In FDM, a segment of the channel's frequency range is allocated to each subchannel. The other multiplex possibility is time-division multiplex (TDM). This is the approach used for subdividing pulse-code modulation channels[22] and in which time slots rather than frequency slots are assigned to subchannels or terminals.

FDMs and TDMs are made by a number of manufacturers for subdividing the voice channel into a number of low-speed data channels.

A more flexible system can be had by using concentrating, as opposed to multiplexing, equipment. "Concentration" is here defined as the assigning of the trunk's capacity on a dynamic or demand basis. A concentrator, however, may have more input that output capacity and so may generate errors.

A concentrator can be simply a relatively fast switch that connects the various terminals, upon request, to the lines or trunks leading to the information-processing unit. The switch can be made to function very rapidly because it has a very limited number of customers and need only perform a limited number of tasks for these

FIGURE 4. Message-multiplexer system.



customers—in contrast to a conventional telephone exchange.

If two messages require transmission at the same time in this approach, one must wait until the other's transmission has been completed. One such system has been described by J. J. Watson.[23] The Bell System has recently announced DATREX,[18] which performs a similar function. Devices such as these are essentially line-switching concentrators.

Although the line-switching concentrator permits significant improvement in the utilization of transmission facilities, it does have its limitations. This is so because it is restricted to setting up a line connection and maintaining that connection until the entire message, including the possibility of some idle time, has been transmitted. Other messages cannot be accepted by the concentrator switch until the first message has completely cleared the system. The signaling techniques (used for specifying the connection path and end of message) are conventional, thus limiting the connection speed.

A system that is capable of higher transmission efficiencies is the multiplexer–concentrator with buffer storage, shown in Fig. 4. This system has three main features:

1. Signaling (consisting only of the address of the calling terminal) is handled by the multiplexer–concentrator and at the trunk's bit rate.

2. The messages are of fixed length but can be very short—perhaps only a single bit or character—and each message is addressed.

3. A queue or waiting line for messages is provided so that every terminal may make the assumption that the system always has room for another message.

As a result of these features it is clear that this system also has several disadvantages.

First, the maximum efficiency of the system is limited—a result of the fact that the address is transmitted with each message. If, for example, five bits are used to represent the message (a single character) and there are $32 = 2^5$ terminals, the maximum attainable efficiency is 50 percent (five information and five address bits). In contrast, the line-switching concentrator may approach 100 percent, but only for very long messages. A glance at either Fig. 2 or 3 reveals, however, that 50 percent or even 17 percent (if the message is a single bit in the example above) is a relatively high efficiency compared with that of many conventional systems.

Second, the system is not a very flexible system because of the heavy dependence of efficiency on the number of addresses.

Third, since the queue must be of limited size in a realizable system, there is a definite probability of error in this system. This is a key parameter and will be discussed further.

The notion of such a multiplexer–concentrator is not new.[16] In the field of digital computer input–output systems, a particularly large effort has been made in this area. In fact, such an approach is mandatory if a computer memory is to have the capability of communicating efficiently with card readers, punches, printers, disks, tapes, and terminals.[24,25] The preoccupation of the designers of these internal multiplexers for large digital computers has been to provide the most efficient use of the computer's memory in an effort to minimize the interference of the communication process with the internal operation of the computer. What would seem desirable in the case of the time-shared computer system is the extension of the multiplexing from inside the computer to a point closer to the remote terminals so that the communications facilities can be efficiently shared as well as the computer's memory is.

Some work has been done in this direction by essentially using a full-size digital computer to achieve multiplexer–concentrator performance as well as to accomplish a great number of other services simultaneously.[26,27] Some computer manufacturers and some manufacturers of peripheral computer equipment do provide remote multiplexer–concentrators that are employed to increase communication efficiency. (Unfortunately, the design of such units is apparently considered highly proprietary and very little information appears in the open literature.) IBM, for example, has models 3967 and 3968 in its product line. These units, as well as the IBM model 2905 (which has recently been described in the literature[28]), can serve as multiplexer–concentrators.

A closer look at a multiplexer–concentrator

A specific form of the multiplexer–concentrator with buffer storage provides an interesting model for closer examination and demonstration of increased communication channel efficiency. The message length is chosen as short as possible, i.e., a single bit or a single character. This author has recently been engaged in a study of the statistical behavior of such a model.[29]

This multiplexer–concentrator is an interesting model because it represents an extreme in simplicity and attainable efficiency. The system examined here can be thought of as perhaps the most primitive message switch[20] because the messages are all of the same and shortest possible length. Also, the "header" (which identifies the message) has the shortest possible length, consisting only of the address. Examination of this system, then, should provide a kind of lower bound that can be improved upon by more sophisticated systems of the future.

Although the multiplexer discussed here provides an interesting model for study, it also has a number of practical advantages. One advantage is that the multiplexer looks like a variable-speed terminal to the user: Whenever the user wants to deliver a bit or a character, he may. This is important, for example, in the case of remote display of computer graphics where the need is for a burst of high-speed data followed by a long quiescent period.[30] Another advantage is that, in addition to shared costs for expensive trunks and modems, a very powerful error-control unit might be shared as well.

A high attainable line utilization or efficiency results from the statistical averaging process of the many randomly operating input terminals. Specifically, it is the inclusion of memory that allows the messages to collect randomly in a queue for subsequent transmission at a regular rate on the trunks.

As stated earlier, it is necessary to limit the size of this queue in any physical device and it is clear that such a restriction introduces the probability of error. A question that must be answered immediately is, "How large must the memory be to achieve a high information-transmission efficiency and at the same time maintain a low probability of error from overflow of the queue?" If the memory must be extremely large, the system is unreasonable from the economic point of view.

In the work mentioned earlier* such an analysis was made assuming that:

1. The messages to be transmitted are of constant length in agreement with what has been stated here.

2. The number of terminals connected is finite (specifically, 15 in the curves shown here).

3. The maximum queue length is specified.

4. The number of messages arriving during the transmission period of one message is binomially distributed; the terminals operate independently.

The results of this analysis are most encouraging and are shown for a single trunk in Fig. 5. The probability of message loss or probability of error is plotted as a function of the maximum allowable queue length. Three curves are drawn for varying levels of channel utilization ρ.

The channel utilization ρ is simply the percentage of the time during which the channel is carrying data. The overall efficiency of the system is thus determined by multiplying the channel utilization by the percentage of useful information in the data. For example, for $\rho = 0.75$ and the five information- and five address-bit message mentioned earlier, the overall efficiency is

$$\eta = (0.75)(0.5) = 37.5\%$$

The memory capacities indicated by Fig. 5 are quite reasonable in the light of what is expected by contemporary data-transmission standards that often have an error rate of one in 10^5. The error rate in the multiplexer–concentrator, for example, could be held to one in 10^8 with a maximum queue length of about 70 messages even for the very high trunk utilization of 90 percent.

In addition to providing the desired smoothing of data flow, the memory must introduce delay. It is, of course, important to make certain that this delay remains within reasonable bounds. Figure 6 shows the average waiting delay (in relation to the unit time required to send a single message) as a function of the maximum allowable queue size. As in Fig. 5, three values of utilization are indicated. It is evident from Fig. 6 that, even in the case of high utilization, the delays are reasonable.

As a summary, consider a system consisting of 15 high-speed (1000-b/s) terminals clustered together and connected by 2000-bit trunks to a distant information processing unit. The terminals can operate at 1000 b/s but do so only 6 percent of the time (producing uniformly distributed, single-character messages) in an independent fashion.

One communication possibility is 15 separate lines with an assumed error rate on the trunk of one in 10^5 messages. Alternatively, two channels can be multiplexed on each of seven channels, reducing the required number of trunks to eight. Through the use of a multiplexer–concentrator, all the terminals can be carried by a single trunk with a

I. System comparison

System	Number of Channels	Probability of a Message in Error*
Direct connection	15	1.000×10^{-5}
TDM	8	1.000×10^{-5}
Multiplexer–concentrator with 20-character buffer	1	1.001×10^{-5}

*Approximate error rates. The exact calculation of probability of error in a message would depend on the details of the transmission system.

FIGURE 5. Probability of error vs. queue length for 15 independently operated input terminals and three utilization factors.

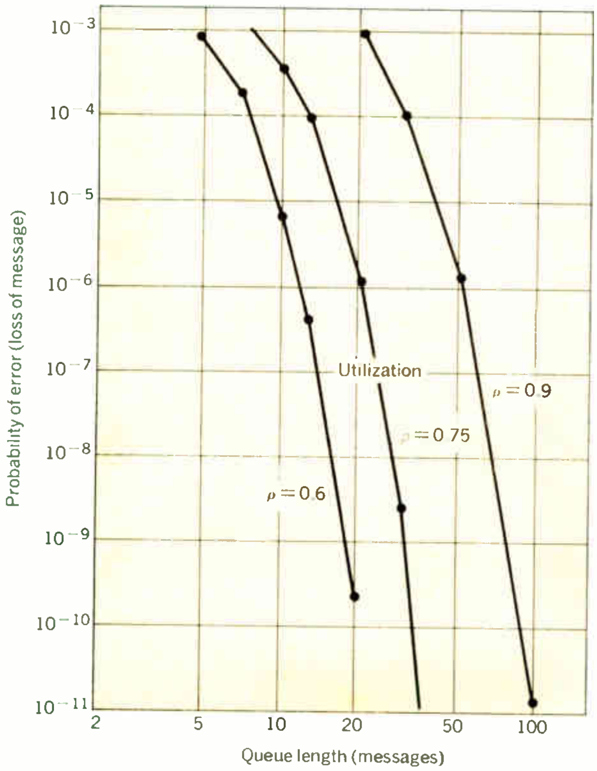
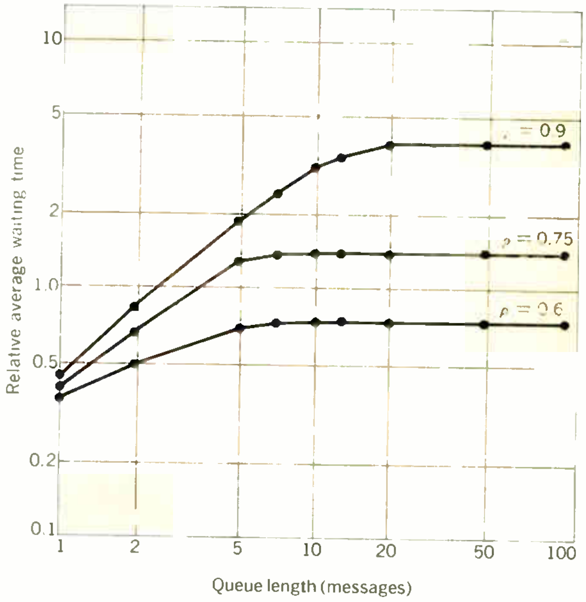


FIGURE 6. Average delay vs. queue length for 15 input terminals and three utilization factors.



* It has recently come to the author's attention that a study similar to that in Ref. 29 has been made—using somewhat different assumptions—in a paper by W. W. Chu.[31]

negligible increase in error rate, if the buffer storage has a capacity of some 20 characters. (Characters, eight bits in size, are given four-bit addresses so that the channel utilization is 0.675.*) Table I provides a comparison of the systems.

This example is only an indication of what can be achieved by a very simple system under ideal circumstances. Efficiency, it must be added, is not by any means the only criterion of goodness for such a system. Ideally the entire system should be optimized, say from the point of view of cost, taking careful account of the traffic statistics. There is room for much effort in this area.

Conclusion

The purpose of this article is not so much to present a solution (it is not a new solution) as it is to encourage discourse and further work in the data-transmission area. Considerations indicate that there are tremendous gains that may sometimes be made in adding traffic considerations to the popular signal-design considerations, which represent the bulk of the current effort in data-transmission research.

When pulse-code modulation becomes widely available, thereby increasing the data-transmission capacity of the voice channel by a factor of roughly ten, the mismatch between low-speed terminals and channel capacity will become even greater and the use of multiplexing–concentrating techniques may well be even more desirable.

* The total number of message bits that each second flow into the trunk from the output of the 15 terminals is: 15 terminals × 1000 b/s per terminal × 0.06 utilization = 900 b/s. Since each eight-bit character is accompanied by an additional four-bit address, the trunk must carry an extra 450 b/s. The total trunk load is 1350 b/s whereas the total space available is 2000 b/s. The channel utilization is, therefore, 0.675.

The author is deeply indebted to many of his colleagues for exciting discussions on this subject. Particularly, the author wishes to thank H. R. Mueller, E. Port, and D. Seitzer.

REFERENCES

1. Becker, F. K., "An exploratory, multi-level vestigial side-band data terminal for use on high grade voice facilities," *Conf. Record First IEEE Annual Communications Conv.*, pp. 481–484, June 1965.

2. Critchlow, D. L., Dennard, R. H., and Hopner, E., "A vestigial-sideband, phase-reversal data transmission system," *IBM J.*, vol. 8, pp. 33–42, Jan. 1964.

3. Kretzmer, E., "Binary data communication by partial response transmission," *Conf. Record First IEEE Annual Communications Conv.*, pp. 451–455, June 1965.

4. Howson, R. D., "An analysis of the capabilities of polybinary data transmission," *IEEE Trans. Communication Technology*, vol. 13, pp. 312–319, Sept. 1965.

5. Schreiner, K. E., Funk, H. L., and Hopner, E., "Automatic distortion correction for efficient pulse transmission," *IBM J.*, vol. 9, pp. 20–30, Jan. 1965.

6. Rudin, H. R., "Automatic equalization using transversal filters," *IEEE Spectrum*, vol. 4, pp. 53–59, Jan. 1967.

7. Di Toro, M. J., "Communication in time-frequency spread media using adaptive equalization," *Proc. IEEE*, vol. 56, pp. 1653–1678, Oct. 1968.

8. Melas, C. M., "Reliable data transmission through noisy media—a systems approach," Conference Paper CP 61-377, AIEE Winter General Meeting, Feb. 1, 1961.

9. Burton, H. O., and Weldon, E. J., "An error control system for use with a high speed voiceband data set," *Conf. Record First IEEE Annual Communications Conv.*, pp. 489–490, June 1965.

10. Kohlenberg, A., "9600 bps—a magic speed for data transmission," *Telecommunications*, vol. 2, Nov. 1968.

11. Farrow, C. W., and Holzman, L. N., "Nationwide field trial performance of a multilevel vestigial-sideband data terminal for switched network voice channels," *Conf. Record 1968 IEEE Annual Communications Conv.*, pp. 782–787, June 1968.

12. Shannon, C. E., *The Mathematical Theory of Communication.* Urbana: University of Illinois Press, 1949.

13. Bennett, W. R., and Davey, J. R., *Data Transmission.* New York: McGraw-Hill, 1965.

14. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication.* New York: McGraw-Hill, 1968.

15. Andrews, M. C., Cookson, B., Halsey, J. R., Lissandrello, G. J., Mueller, H. R., and Port, E., "A PCM-compatible switched data network study," *Conf. Record of Switching Techniques for Telecommunication Networks—London*, pp. 329–332, Apr. 1969.

16. Filipowsky, R. J., and Scherer, E. H., "Digital data transmission systems of the future," *IRE Trans. Communications Systems*, vol. CS-9, pp. 88–96, Mar. 1961.

17. Jackson, P. E., and Stubbs, C. D., "A study of multiaccess computer communications," *Proc. AFIPS 1969 Spring Joint Computer Conf.*, vol. 34, pp. 471–504, 1969.

18. Bacon, W. M., "Low speed data systems development in the U.S.A.," presented at 1969 Data Transmission Conference, Mannheim, Germany, Mar. 19–21.

19. Cornell, W. A., "The influence of data communications on switching systems," *Conf. Record of Switching Techniques for Telecommunication Networks—London*, pp. 342–345, Apr. 1969.

20. Hamsher, D. H., ed., *Communication System Engineering Handbook.* New York: McGraw-Hill, 1967.

21. Bell Telephone Laboratories, *Transmission Systems for Communications.* Winston-Salem, N.C.: Western Electric Company, Inc., 1964 (see especially Chap. 5).

22. Fultz, K. E., and Penick, D. B., "The T1 carrier system," *Bell System Tech. J.*, vol. 44, pp. 1405–1451, Sept. 1965.

23. Watson, J. J., III, "Concentrating and switching equipment for a real time multiple access communication system," *Conf. Record 1968 IEEE Internat'l Conf. on Communication*, pp. 123–128, June 1968.

24. Padegs, A., "The structure of system/360—part IV—Channel design considerations," *IBM Sys. J.*, vol. 3, no. 2, pp. 165–180, 1964.

25. Ossanna, J. F., Mikus, L. E., and Dunten, S. D., "Communications and input/output switching in a multiplex computing system," *Proc. Fall Joint Computer Conf.*, pp. 231–247, 1965.

26. Daley, E. A., and Scott, A. E., "IBM 7740 communication control system," *IEEE Conv. Record*, vol. 12, pt. 5, pp. 207–215, 1964.

27. Drescher, J. E., and Zito, C. A., "The IBM 7741—a communications-oriented computer," *IEEE Conv. Record*, vol. 12, pt. 5, pp. 216–224, 1964.

28. Arnold, O. E., "Automatic polling by remote multiplexers," *Telecommunications*, vol. 3, pp. 17–19, June 1969.

29. Rudin, H. R., "Statistical performance of a simple multiplexer-concentrator for communication channels," to be published.

30. Baskin, H. B., "The communications requirements of interactive computer graphics," presented at IEEE Internat'l Conf. on Communications, June 9–11, 1969.

31. Chu, W. W., "A study of the technique of asynchronous time division multiplexing for time-sharing computer communications," *Proc. Second Hawaii Internat'l Conf. on System Sciences*, pp. 607–610, Jan. 1969.

Harry Rudin (M) is currently working as a full-time consultant at IBM's European Research Laboratory in Zurich, Switzerland. His work is in the field of computer-related communications with emphasis on the application of traffic theory to the problem of data flow in communication networks. In his previous position at Bell Telephone Laboratories in Holmdel, N.J., he worked in the area of data communications, concentrating on automatic equalization techniques. From 1961 to 1964, Dr. Rudin served as an instructor in electrical engineering at Yale University. There, also, he received the bachelor of engineering, the master of engineering, and the doctor of engineering degrees in '58, '60, and '64, respectively. Dr. Rudin had served as a member of the Executive Committee of the IEEE Connecticut Section from 1962 to 1964.

Why space broadcasting?

The use of satellites offers an interesting approach to television broadcasting for both developed and developing areas; however there are a number of factors that must be taken into consideration. Not the least of these are national and international telecommunications policies

R. P. Haviland General Electric Company

This article examines some nontechnical factors that will influence both policy and technical decisions regarding operational space-television broadcasting. The reasons for using space broadcast for both community and general public services are examined. They include faster introduction of service in new areas, potentially lower cost, extension of program choice, and more efficient use of frequencies. These are considered with respect to the state of economic development and availability of existing service. Marked differences are found between developing and developed areas. It appears that developing areas could best initiate television service with a community satellite, whereas developed areas could best use the satellite for program equalization. For both areas, a combined system of both terrestrial stations and satellites appears to be the most cost-effective. Possibilities for spectrum conservation are considered along with areas needing further study.

There are four major reasons for using satellites for television broadcasting:

- Faster introduction to service in new areas.
- Potentially lower cost of providing full coverage.
- Extension of program choice.
- More efficient utilization of frequencies.

The relative importance of these factors is unique for different parts of the world, depending on the status of existing television, availability of other communication services, economic conditions, and policy decisions. This

FIGURE 1. Time history of radio receiver growth.

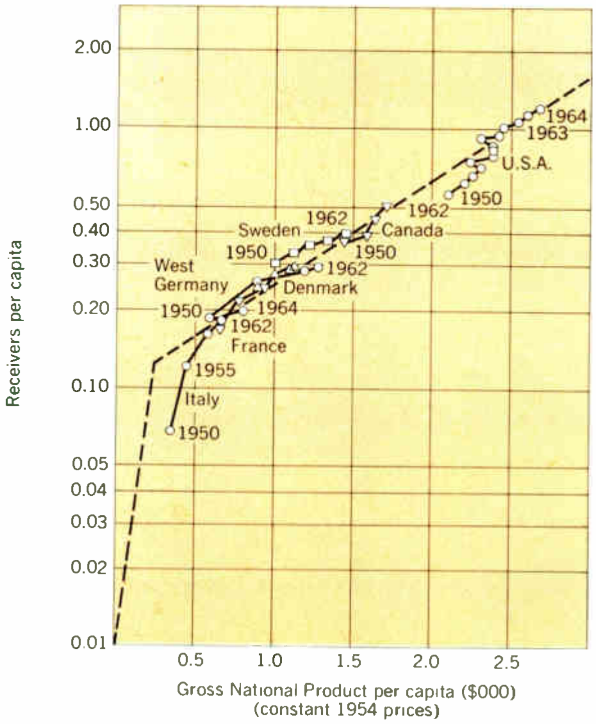
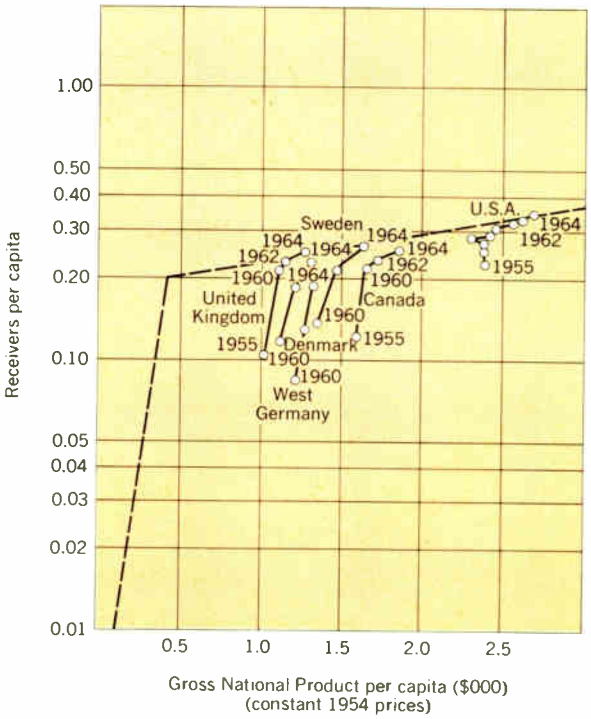


FIGURE 2. Time history of television receiver growth.



article considers these factors and their implications. For simplicity, these considerations are limited to three situations: industrially developed areas, developing areas, and emerging areas. It is believed that this restriction does not affect the validity of the conclusions reached.

Projection of television growth

Before investigating the reasons for satellite broadcasting, it seems desirable to estimate the extent to which television may grow. This is not a simple problem, since it involves a complex of technical, economic, and political factors. For example, at least one country has stated a nontelevision policy because of the difficulty of maintaining cultural purity. Others have adopted television as a means to achieve national unity.

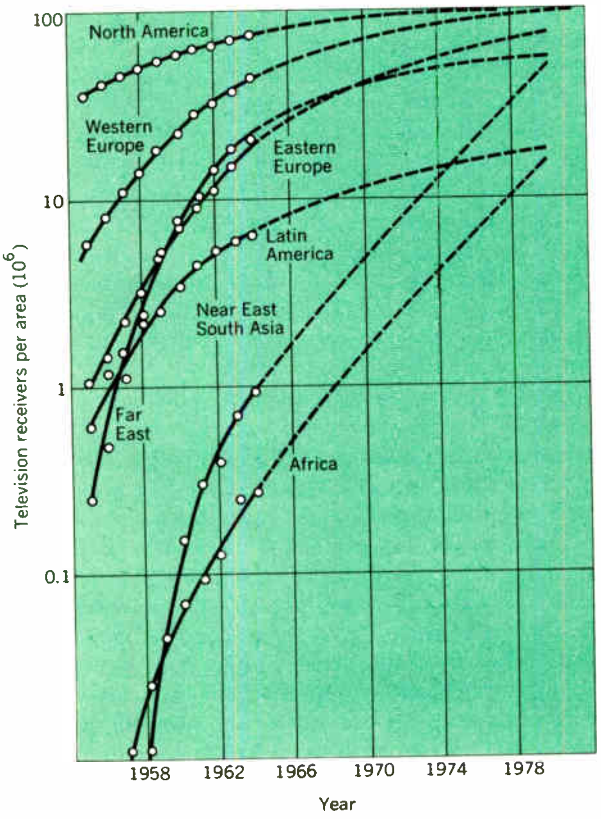
Except for outright elimination of television, it appears that the major factor governing television usage is economic. This is illustrated by Fig. 1,* which plots the time history of radio receiver growth per capita against per capita national product. The graph suggests that the receiver population may be estimated by assuming two periods: an initial growth period and a steady-state period. In both cases demand is limited by economic considerations and total population. The boundaries of these periods are indicated by the dashed lines.

For television (Fig. 2) a clear pattern is less obvious, for this is a relatively new service. The curves suggest that growth periods similar to radio occur but saturation begins at one receiver per family rather than tending toward one receiver per person in the more affluent areas. The estimated boundaries are again shown by dashed lines.

Using these trend boundaries as a projection model permits an estimate of the future world growth of television. The results of the projection are shown in Fig. 3. The circled points are historical values and therefore are used for the initial values. The forecast values are based on the above model, allowing for increases in the Gross National Product and total population. These latter values indicate that world economic conditions will permit a

* Based on studies made in connection with Contract NASw-1475. For a report of the major results of the study, see Ref. 1.

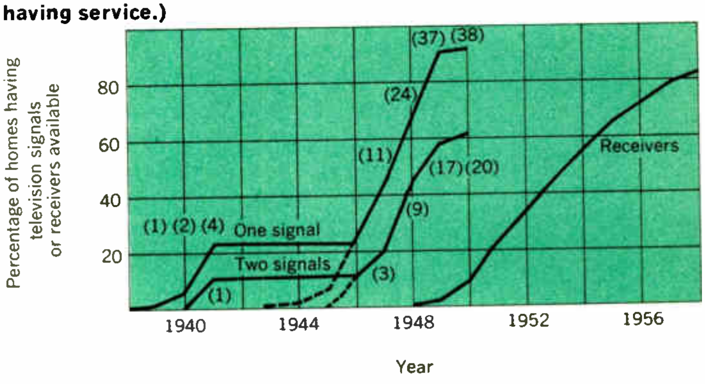
FIGURE 3. Projected growth of television.



I. Projected biannual television receiver market, millions of sets

Year	Near East/ South Asia	North America	Western Europe	Eastern Europe	Far East	Africa	Latin America	Totals
1966	0.5	12.9	11.0	6.4	6.0	0.1	1.0	36.9
1968	0.9	14.4	11.9	9.6	7.3	0.2	2.0	46.3
1970	1.1	14.6	13.9	8.9	8.4	0.5	2.2	49.6
1972	2.3	13.9	14.4	9.7	7.3	0.7	2.6	50.9
1974	3.2	14.2	14.8	11.8	8.3	1.2	2.8	56.3
1976	7.2	14.4	15.6	12.3	8.8	2.1	3.0	63.4
1978	10.9	14.7	15.2	13.3	8.5	3.1	3.2	68.9
1980	16.1	14.9	15.7	13.6	8.7	5.5	3.3	77.8

FIGURE 4. Time history of television in the United States. (Figures in parentheses represent the number of cities having service.)



growth in television receivers from approximately 200 million units at the end of 1967 to about 400 million units by 1980.

The distribution of the market may be estimated by adding an allowance for scrappage. This may be assumed to be zero initially, increasing to a mean set life of 14 years, based on experience in the United States. These models lead to the biannual projection of Table I, which shows a steadily increasing market from about 40 million receivers per year at the end of 1967 to nearly 80 million in 1980.

A surprising feature is the change in the distribution of the market with time. In addition, there is an indicated peaking of the North American market around 1970, which is a result of the balance between new-use and replacement sets.

Introduction of television service

Broadcast television has been conducted as an experimental service since the late 1920s but the modern electronic television receiver did not become operational until just prior to World War II. It was developed simultaneously in several countries, including the United States.

Figure 4 summarizes the early time history[2,3] of the introduction of television coverage and receiver usage for the United States. Neglecting the war years as shown by the dashed curves, signal coverage reached 20 percent of the population within approximately three years, 50 percent in about 4½ years, and 90 percent in six years. Ninety-five percent coverage was not attained until another ten years had passed, and even today, after 30 years, the coverage has not yet reached 100 percent.

Second program availability showed about the same initial rate, but growth slowed markedly after 50 percent coverage had been reached. In part this was due to a reexamination of policy.

Wide use of the television signal, however, was delayed, showing a lower growth rate. Sets did not reach 20 percent of the homes until approximately 13 years after the actual beginning of service, or approximately eight years after its effective start. During this period there was extensive informal community viewing at neighbors' homes, in stores, and in taverns. Growth continued, reaching 50 percent of the homes within 10½ years (effectively) and 80 percent in 15 years. Today, sets are used in approximately 92.5 percent of U.S. homes, with 22 percent of the homes having two or more sets.

Except for the availability of additional broadcasting frequencies (usually delayed for 20 or so years), the pattern in other developed countries has been similar. This is true even where television is a national service rather than a commercial enterprise.

In lesser developed areas, however, the pattern is somewhat different. This is illustrated by the growth of television[1,4] in Brazil, which began televising in 1953. Less than 15 percent of the households have receivers, and they have only 45 transmitters for a country larger than the United States. Twenty-one of these stations are in five of the more important cities.

In still lesser developed areas, it is common for service to be available only in the largest city, and even there receiver density is low. An exception to this is the United Arab Republic, where television transmitters have been installed to provide signals to some 80 percent of the population. But this is part of a program of education and national unity.

From the foregoing data, it appears clear that attainment of complete large-area national coverage is a very difficult and time-consuming process. Even the United States has not achieved it after 30 years. Reasonable coverage (90 to 95 percent) is easier, but still may require six to eight years in a country with ample resources, or longer if resources are limited.

In contrast, television by satellites offers full national coverage from the beginning of service. The duration of the necessary design and construction program would depend on the system parameters chosen, but would be perhaps four years for an area of 5 million square kilometers. Thereafter, further areas could be added every three to six months. World coverage would be possible within ten more years, or in even less time if system parameters were properly chosen. In view of this, it seems clear that satellites are the fastest means for attaining national coverage. In this respect special advantages are offered to the lesser developed areas.

The cost of providing television service

The element of speed cannot alone answer the question of desirability. The economic comparison of the alternate approaches must also be considered.

In the terrestrial system, the first operational station will usually be placed in the largest city, and typically will cover some 10 percent of the population of the country. Successive stations will be placed in the second, third, etc., largest cities. Since the population of these coverage areas is decreasing, the cost-effectiveness of the individual stations is also decreasing. The order of magnitude of the effectiveness can be estimated by using the usual rule that the population of the *n*th largest city is $1/n$ times the population of the largest city. With this installation plan, the initial cost is relatively low. Total cost appears to become excessive if full national coverage is attempted. As a result of this excessive cost, no country has 100 percent coverage.

The terrestrial approach has one significant advantage. By locating the transmitter in or close to the city, the signal levels are adequate to overcome the urban noise level. Noise drops off rapidly with an increase in distance from the city, and therefore is not a limiting factor in suburban and rural area coverage. Generally, Grade 1 pictures can be obtained up to a distance of 80 km on VHF, and 55 km on UHF.[5] Outdoor antennas are almost always required to obtain this signal grade, to avoid ghosting in the cities and suburbs and to provide adequate rural signal levels to overcome set noise.

Figure 5 shows an effectiveness evaluation of this method of transmitter installation. The analysis considers the United States and utilizes "standard TV market"[2] as a data source. (Note: Since there is some overlap in the coverage of markets, the sum of individual station coverages exceeds 100 percent. This factor is neglected here. The actual coverage is approximately 97 percent of U.S. households.) The cost curve assumes only one station per market and a typical station cost of one million dollars. The curve indicates a variation in transmitter investment per household covered from $0.15 to nearly $15.00. Coverage per station ranges from 12 percent down to 0.12 percent of the total number of households. Average investment in transmission facilities is approximately $3.00 per household covered. Fifty percent coverage is attained with some 30 stations, at an average investment very close to $1.00 per household (approximately 29.5 million households covered).

The cost behavior in space broadcasting is quite different. Satellite signals readily cover a wide area. However, normal-band strong-signal satellites became expensive, so the cost of covering the highly populated areas is the highest due to higher man-made noise levels. In addition, there is a cost balance between decreasing the antenna gain at the receiver and increasing the size and cost of the satellite.

As an example of these relationships, assume that we have 60 million households in the United States, one third with urban noise conditions, one third with suburban levels, and the rest rural, an assumption reasonably close to conditions as they exist in the United States. Figure 6 shows the approximate satellite investment* needed for an optimally cost-balanced system for the three areas as a function of the number of receivers in use. Since the values above 10^7 receivers are essentially constant, the satellite investment required to cover 60 million households would be approximately $65 million for rural coverage, $67 million for suburban coverage, and $95 million for urban coverage. These values correspond to roughly $3.30 per receiver for rural coverage, $1.67 for suburban coverage, and about $1.58 for urban (i.e., total) coverage; and they compare favorably with the afore-listed terrestrial coverage costs.

If a Grade 2 picture is considered satisfactory, satellite costs are markedly lower—approximately $14 million for both suburban and rural coverage and $16 million for urban coverage. These values correspond to roughly $0.35 per household for rural and suburban coverage and $0.28 per household for urban coverage.

On a strictly transmitter-cost basis, such a Grade 2 satellite signal would appear the most advantageous. This is not the only criterion, however, for Grade 2 signals may not be considered acceptable. The satellite system does not have the program flexibility of the terrestrial system, and it does require good antennas for all receivers, whereas the terrestrial system permits low installation cost for receivers reasonably close to the transmitter. On the other hand, terrestrial stations become very expensive when full national coverage is attempted.

The foregoing factors suggest that a mixed system would be most effective if all factors are considered. To illustrate this, assume that a large country, similar to the United States in area, population, etc., has $60 million available for transmission investment. Assume that the goals are

- Full national coverage, if possible.
- Grade 1 signals, if feasible.
- Good programming flexibility.

Using the previous assumptions, this would allow

- Sixty terrestrial stations, about 90 to 95 percent coverage, average investment of $1.05 per household covered, maximum flexibility, and Grade 1 quality.
- One satellite, 100 percent coverage, average investment of $1.00 per household, approximately urban Grade 1.5 signal, and a single program.
- A mixed system of 44 terrestrial stations having 85 to 90 percent coverage, plus an urban Grade 2 coverage satellite. Some 10 to 15 percent of the thresholds would have single-channel coverage and 85 to 90 percent would have two-channel coverage. Average investment is $0.54 per channel per household.

The assumptions made are not necessarily optimum, and the cost factors may not be representative. However, it appears that the relative cost factors can change appreciably without affecting the results. For example, assume that the investment cost of the satellite is double the $16 million assumed. Then, the mixed system would use only 14 terrestrial stations covering approximately 60 percent of the households (receiving two channels).

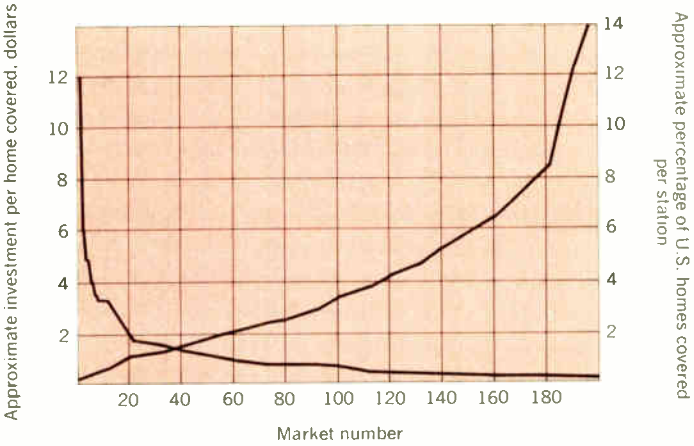
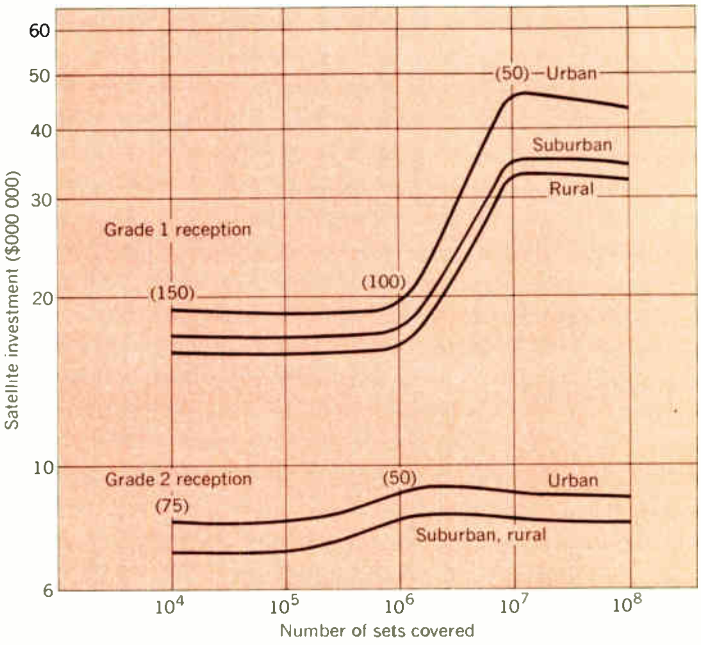


FIGURE 5. Terrestrial station effectiveness measures. Notes: (1) Television market areas assumed. (2) One station per market. (3) $1 million invested per station. (4) Interconnection costs not included.

FIGURE 6. Satellite investment required for the United States. (Figures in parentheses represent optimum receiver installation cost for antenna, etc.)



* Based on studies made in connection with Contract NAS 3-9708. For a report of the major results of the study see Ref. 6.

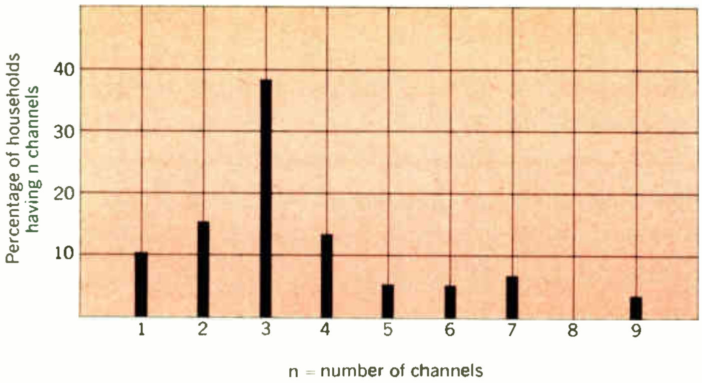


FIGURE 7. Program availability (based on "Standard TV Markets, 1966 Data").

plus 40 percent receiving the single satellite channel. The average investment would be $0.63 per channel per household, which is still lower than for either of the "pure," high-quality systems.

This cost preference for a mixed system is due to the presence of a relatively small number of densely populated areas best served by terrestrial stations, and a large total area best served by wide-coverage systems. Where this situation exists, mixed satellite–terrestrial coverage should be considered.

Evaluation of the cost factors is made difficult by the fact that many sources of investment funds may be involved. For example, in the U.S., each receiving installation is a separate budget; however, in developing countries depending on community viewing, it is likely that both transmission facilities and at least the community receivers will be paid for by the central government. Several cost optimization possibilities are discussed in Ref. 6. In these the cost of the basic receiver is omitted, since it is assumed to be a fixed value, common to any method of coverage, whether satellite or terrestrial.

At the present state of the art it appears that receiving installations will necessarily include outdoor antennas. This does not appear to be a severe limitation, in view of their acceptance for good color reception, and the extensive installations found in weak signal areas.

The use of satellite signals to feed programs to cable heads for further distribution is being studied. Such a mixed system appears to have great potential.

Program choice availability

In the foregoing example, the effect of adding a satellite to a limited terrestrial system is to provide first-coverage service to some fraction (n) of the total population, and to provide a second channel to the remainder ($1 - n$). Suppose, then, that a satellite were added to an existing large system. Figure 7 shows the distribution of the number of stations per market[2,3] normalized to the total number of households served, for the United States. The most probable number of stations is three. Some 36 percent of the households have three or fewer program choices available.

This plot neglects the fact that some 2 to 3 percent of the total U.S. households do not have television programs available, and also neglects the fact that some market areas overlap. An incomplete check of the latter factor indicates that the effect of this overlap is to provide additional program choice in densely populated areas where program choice is relatively wide. In relatively sparsely settled areas, program choice remains low. The net effect is to reduce the values over the range of the three to four channels available, and to increase the fraction having five or more channels available. This analysis, although approximate, is sufficient to show that appreciable maldistribution of program choice exists in the United States. This situation is also true in several other countries.

If a full-coverage, clear-channel satellite system were added to the terrestrial system, the effect would be to shift the entire curve of Fig. 7 to the right by an amount equal to the number of channels added. A rural- or suburban-coverage satellite would largely affect the lower part of the curve, as would a shared-channel system.

In view of these factors, it appears that satellite coverage can be used to equalize program choice. Probably the most important element is that all homes in the coverage area would have at least one program available. Most would have a choice of several, the exact number depending on the details of the terrestrial and satellite systems.

Frequency occupancy

The last element, that of frequency use, arises from propagation factors. In the terrestrial system signal quality remains high out to the critical distance, even though the signal strength is decreasing as the square of the distance. Beyond this point, quality is limited by signal strength, decreasing approximately as the cube of the distance. Typically, on UHF, the critical distance is 48 km for the higher channels. The signal is usable beyond this distance, the exact value being dependent on terrain, with values of 64 to 96 km being typical.

At still greater distances the signal, although unusable, can cause interference to stations on the same channel. Also, at closer distances interference may occur on adjacent channels, or on the image channel, which lies at twice the IF frequency away from the desired channel. In addition, it is desirable for the transmitter to be close to the largest population center served, because of urban noise.

The channel-allocation plans for the United States take these factors into account. Typically, a UHF plan[7] shows 40 to 50 stations per channel, although maximum packing would allow somewhat more. The difference is due to estimated economic infeasibility in some areas. Fifty stations would have a high-quality coverage area of about 215 000 square kilometers, or approximately 1/20 of the United States. Thus, some 20 channels would be needed to give high-quality national coverage. The corresponding figure for extended coverage of usable quality would be five channels.

In contrast to the terrestrial system, satellite signals would vary only slightly over the coverage area. A single-channel, high-power satellite would give national service; however, it would be desirable to reserve adjacent channels as guard bands so that the space service would be approximately six times as efficient in its use of the spectrum. At lower power levels (corresponding to usable quality levels) the guard bands would not be necessary, and the satellite would again be more efficient, by a factor of five.

If the satellite and terrestrial stations share the same channel,[8] complete space coverage could not be attained on a given channel; each terrestrial station would cause a gap in satellite coverage. Approximately for the existing U.S. allocation plan, these gaps would reduce the coverage of a satellite signal suitable for rural and suburban reception to roughly 60 percent of the total area covered. Two selected satellite channels would give 100 percent single-channel satellite coverage, plus 36 percent two-channel satellite coverage, plus 10 to 20 percent terrestrial coverage (always second channel). This mixed system would be the most efficient of all approaches in its use of spectrum space.

At present, there are no international allocations specifically for space broadcasting. A number of possibilities have been mentioned, for example in the documents of the CCIR, Geneva, 1969, including the 26-MHz band for AM, the 88–108-MHz band for FM, and parts of the 470–960-MHz and the 11-GHz band for television. Both exclusive and shared channels have been proposed. The matter will undoubtedly be considered at the scheduled 1971 Radio Conference. See Ref. 9 for a partial discussion of possibilities.

Applications to worldwide television

One important consideration regarding the applicability of the foregoing factors to the various geographical areas appears in Fig. 3. For the more developed areas of the world, specifically North America and Europe, the 1980 projections vary between 0.2 and 0.4 set per person, or essentially one set per family. The other areas of the world show only one set for every 20 persons. As a result, if television is to be a universal service, many countries will find it necessary to make extensive use of community viewing. This may be done informally, with sets in neighborhood gathering areas and neighbors' homes, or it may be formally organized, such as by the "Tele-clubs" in France and Japan. Considering economic status, it appears that the patterns would be

- Emerging country—depends on community viewing.
- Developing country—both community and home viewing, the latter increasing as the economy expands.
- Developed country—home viewing.

For the emerging countries, it appears that the most important factor is the attainment of national coverage at an early date. Costs should be low. Channel-choice flexibility is not a problem initially, although attention must be given to programming for diverse interests. Frequency usage is relatively unimportant. For these areas satellite broadcasting appears to merit immediate consideration due to the speed of attaining coverage at a reasonable cost.

For the developing areas the basic problem is the extension of existing television to national coverage. Cost is important, for this is the factor limiting expansion of the existing system. Second or third channel addition may be desired. Once again, frequency use is relatively unimportant. For these areas a mixed system of satellite and terrestrial stations appears to be an attractive, quick, and cost-effective solution.

For the developed areas speed of coverage is not a problem since good coverage exists. In addition, cost considerations strongly favor use of existing coverage. Some equalization of program choice appears to be desirable, even though this involves a relatively small fraction of the audience. Frequency conservation, however, may be important, depending on the relative demands of nonbroadcast services. Already it is becoming a critical factor in several countries.

Although use of a satellite for national coverage would ease the spectrum problem, it is not the only approach. Cables might be used for urban and suburban distribution; or improved reception techniques could be adopted, allowing closer packing of terrestrial stations. Both national and international telecommunications policies must also be considered, and the best course is by no means clear. All things considered, it appears that satellite broadcasting offers early help to the emerging and developing areas, and is none the less important as an approach to the solution of problems in developed areas.

This article is based on a paper presented at the AIAA Second Communication Satellite System Conference, San Francisco, Calif., April 1968.

REFERENCES

1. Hesselbacher, R. W., "An evaluation of voice broadcast systems," AIAA preprint 68-423.

2. *Broadcast Yearbook*, Washington, D.C.

3. *Television Digest Factbook*, Washington, D.C.

4. *Broadcasting Stations of the World*, U.S. Government Printing Office, Washington, D.C.

5. Town, G. G., "The Television Allocations Study Organization—A summary of its objectives, organization, and accomplishments," *Proc. IRE*, vol. 48, pp. 993–999, June 1960.

6. Hesselbacher, R. W., "An evaluation of TV broadcast systems," AIAA preprint 68-1061.

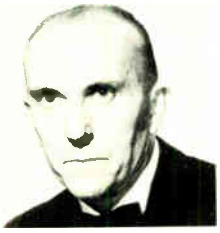
7. "In the matter of fostering expanded use of UHF television channels," Docket 14229, Federal Communications Commission, Washington, D.C.

8. Haviland, R. P., "Television broadcast from satellites—sharing considerations," *IEEE Trans. Broadcasting*, vol. BC-11, pp. 30–35, July 1965.

9. Haviland, R. P., "Choices in space broadcasting," *IEEE Trans. Broadcasting*, vol. BC-13, pp. 80–86, July 1967.

10. Haviland, R. P., "Space broadcasting—how, when, and why," presented at U.N. Conf. on Peaceful Uses of Outer Space, Vienna, Austria, 1968.

R. P. Haviland (F) received the degree of bachelor of science in electrical engineering from the Missouri School of Mines in 1939 and then entered the employ of the Schlumberger Well Surveying Corporation, Houston, Tex. From 1942 until 1947 he was on active duty with the U.S. Navy, where he was concerned primarily with the development of radar beacons and IFF equipment; he currently holds the rank of commander in the U.S. Naval Reserve. Following his naval service, Mr. Haviland joined the General Electric Company at Schenectady, N.Y., as a flight test engineer, later transferring to Key West, Fla., to supervise a field test station in connection with the development of underwater ordnance. His present work involves satellites and space systems at GE's Missile and Space Division in Valley Forge, Pa. His professional career has included many space-age "firsts"—among these are initiation of the first U.S. work on space-vehicle development and his contributions as project engineer for the first two-stage rocket (1949), the first radio transmission from space (1949), and the first launching from the Atlantic Missile Range (1951). Mr. Haviland is a member or former member of various scientific committees and delegations and his society affiliations include the British Interplanetary Society, the American Institute of Aeronautics and Astronautics, and the American Astronautical Society. He holds eight U.S. patents and has written a number of papers.



The Esclangon diagram for voltage and current along a transmission line

It is not very often that graphic representations of abstract ideas and concepts are discovered. When they are, their aid to understanding is invaluable

Michel Poloujadoff **Institut Polytechnique de Grenoble**

This article offers a simple graphic construction that shows the variation of current and voltage along an electric line for a steady sinusoidal operation. The method complements the use of formula applications and the Smith chart, thus making for a good and easy understanding of transmission-line operation.*

After having understood the propagation equation and its proof, and before learning how to apply it, electrical engineering students have to become familiar with the phenomenon of voltage and current variation along an electric line. For such a study, a graphic method would be highly desirable, since it provides a unified visual insight into some transmission-line operations.

Such a method was given by the late Prof. F. Esclangon* in a paper published in 1943.[1] Unfortunately, very little notice, if any, has been paid to this paper. Two reasons can be given for this lack of attention. First, the publication took place during the second world war, when few copies of the review were available. Second, the author failed to show clearly the advantages of his method that were important.

It is hoped that the present article will be useful to people interested in transmission lines by informing them of a graphic technique that considerably enhances the art.

Basic equations

Steady-state sinusoidal operation can be represented by the following equations:

$$\mathcal{V}(X) = \mathcal{V}_1 e^{-nX} + \mathcal{V}_2 e^{+nX}$$
$$Z\mathcal{I}(X) = \mathcal{V}_1 e^{-nX} - \mathcal{V}_2 e^{+nX} \qquad (1)$$

where $\mathcal{V}$ and $\mathcal{I}$ are complex numbers representing the sinusoidal voltage and current at point M, which is at a distance X from the generator (Fig. 1); $\mathbf{Z}$ is the characteristic impedance; $n = p + jq$ is the transmission factor; and $\mathcal{V}_1$ and $\mathcal{V}_2$ are complex constants. For lossless lines, $\mathbf{Z}$ is purely real and n is purely imaginary ($p = 0$, $n = jq$).

If the location of M is given by its distance x from the receiver (Fig. 1), Eq. (1) becomes (with $x = L - X$):

$$\mathbf{V}(x) = \mathcal{V}(L - x) = \mathcal{V}_1 e^{-n(L-x)} + \mathcal{V}_2 e^{n(L-x)}$$
$$= \mathbf{V}_1 e^{nx} + \mathbf{V}_2 e^{-nx} \qquad (2)$$
$$\mathbf{ZI}(x) = \mathbf{V}_1 e^{nx} - \mathbf{V}_2 e^{-nx}$$

As a result of the positive directions initially chosen, some authors obtain Eq. (1), whereas others obtain Eq. (2) directly. We shall use Eq. (2) for reasons that will appear clearly later.

For $x = 0$, $X = L$, the following voltage and current appear at the receiver:

$$\mathbf{V}(0) = \mathbf{V}_r \qquad \mathbf{I}(0) = \mathbf{I}_r \qquad (3)$$

Esclangon diagram for a lossless line

If there is no loss in the line, $\mathbf{ZI}_r$ and $\mathbf{I}_r$ have the same argument ($\mathbf{Z}$ being real), and $n = jq$. Hence, Eq. (2) can be rewritten as

$$\mathbf{V}(x)e^{-jqx} = \mathbf{V}_1 + \mathbf{V}_2 e^{-2jqx}$$
$$\mathbf{ZI}(x)e^{-jqx} = \mathbf{V}_1 - \mathbf{V}_2 e^{-2jqx} \qquad (4)$$

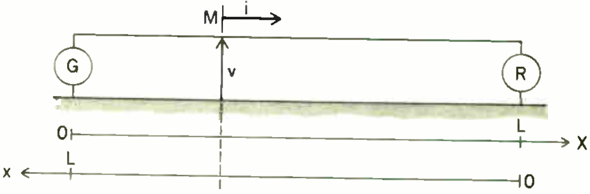
From Eqs. (2) and (3),

$$\mathbf{V}(0) = \mathbf{V}_r = \mathbf{V}_1 + \mathbf{V}_2$$
$$\mathbf{ZI}(0) = \mathbf{ZI}_r = \mathbf{V}_1 - \mathbf{V}_2$$

thus, $$\mathbf{V}_1 = \frac{\mathbf{V}_r + \mathbf{ZI}_r}{2} \qquad \mathbf{V}_2 = \frac{\mathbf{V}_r - \mathbf{ZI}_r}{2}$$

If $\mathbf{V}_r$ and $\mathbf{ZI}_r$ are represented by OA and OB (Fig. 2), respectively, and if M is the midpoint of BA, then OM represents $\mathbf{V}_1$ and MA represents $\mathbf{V}_2$. Let us now draw a circle (the Esclangon circle) at center M and of which AB is a diameter; let $A'B'$ be another diameter such that $\angle AMA'$

FIGURE 1. Positive direction for the study of the propagation equation.



* Director of the Institut Polytechnique de Grenoble, 1941–1954.

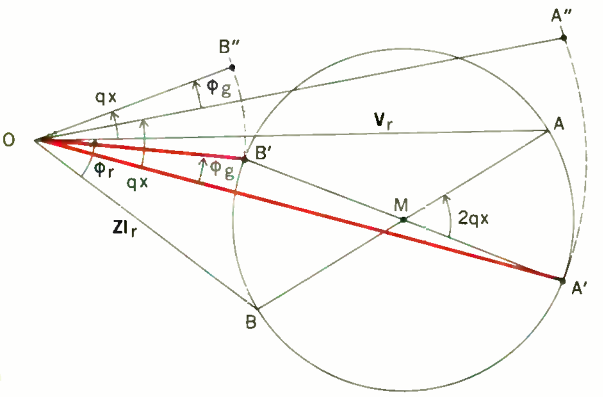
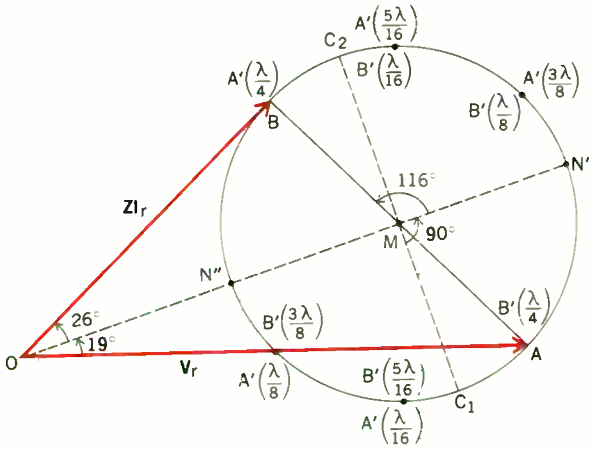


FIGURE 2. Principle of the complete Esclangon diagram for lossless transmission lines.

FIGURE 3. Example of an Esclangon diagram for a lossless line with a receiver impedance of Z(1 — j).



$= -2qx$. We can see that OA' represents $\mathbf{V}(x)e^{-jqx}$ and OB' represents $\mathbf{ZI}(x)e^{-jqx}$, from which we readily deduce $\mathbf{V}(x)$ and $\mathbf{I}(x)$.

When $2qx = 2\pi$ or $x = \pi/q$, A' is on A, B' on B, and

$$\mathbf{V}(\pi/q)e^{j\pi} = \mathbf{V}_r \qquad \mathbf{ZI}(\pi/q)e^{j\pi} = \mathbf{ZI}_r$$

so that

$$\mathbf{V}(\pi/q) = -\mathbf{V}_r \qquad \mathbf{I}(\pi/q) = -\mathbf{I}_r$$

When $2qx = 4\pi$, A' is again on A, B' on B, and

$$\mathbf{V}(2\pi/q) = \mathbf{V}_r$$

From this, it can be seen that the wavelength λ is equal to $2\pi/q$. This also gives the relationship between the angle AMA' and the ratio x/λ.

The angle $A'OB'$ is equal to the angle $A''OB''$, so that φ_g represents the phase angle at the generator when $x = L$. From Fig. 2, we see that $\mathbf{I}_g$ is lagging with respect to $\mathbf{V}_g$, but φ_g is much smaller than φ_r. If the transmission line were a little longer, φ_g would vanish, or even become negative.

It can also be remarked that $|e^{jqx}| = 1$, so that, if we are interested only in the magnitude of $\mathbf{V}(x)$, we can study just the length of OA', and do not have to then draw OA'', thus simplifying the diagram. Voltage is maximum when A' is on N'; hence, this maximum value of the voltage is $|\mathbf{V}_1| + |\mathbf{V}_2|$. In this instance, B' is on N'', so that the current is minimum and equal to $(|\mathbf{V}_1| - |\mathbf{V}_2|)/\mathbf{Z}$; at the same time, $\varphi_g = 0$. Voltage amplitude is minimum and equal to $|\mathbf{V}_1| - |\mathbf{V}_2|$ when A' is on N''; in this case, current amplitude is maximum and equal to $(|\mathbf{V}_1| + |\mathbf{V}_2|)/\mathbf{Z}$.

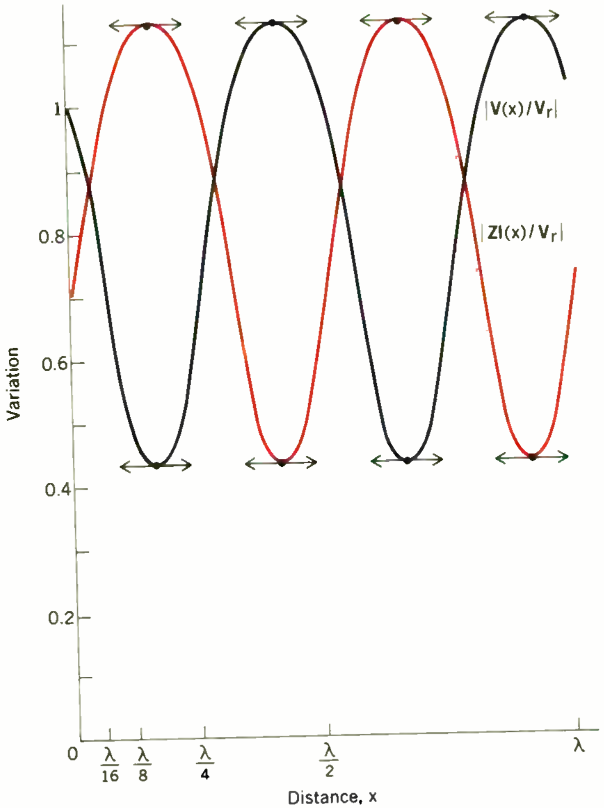


FIGURE 4. Voltage and current variation along the transmission line considered in Fig. 3. Distance is measured from the receiver end of the transmission line.

Starting on the line from a point where voltage is maximum, one can find a point where voltage is minimum, which is at a distance of λ/4 from the starting point; after another λ/4, there is again a voltage maximum . . . and so forth.

An example of lossless line operation

Let us consider a transmission line for which $\mathbf{Z}$ and λ are given; the impedance of the receiver is $\mathbf{Z}_r = \mathbf{Z}(1 - j)$, so that $\mathbf{ZI}_r = (\mathbf{V}_r/\sqrt{2})\angle 45°$. This case is very convenient for a first example, since the diagram shown in Fig. 3 has some simple geometric properties: the quadrangle $AA'(\lambda/8)\,BB'(\lambda/8)$ is a square, and the angle BOM is $\tan^{-1}(0.5)$.

Let $\varphi(x)$ be the lag of $\mathbf{I}(x)$ behind $\mathbf{V}(x)$, $\psi(x)$ be the lag of $\mathbf{V}_r$ behind $\mathbf{V}(x)$, and let us recall that x is the distance from the *receiver*.

1. For $x = \lambda/8$, we see that $\tan[B'(\lambda/8)OA] = 0.5$ so that $\varphi(x) = -26°$. Also, $\mathbf{V}(\lambda/8)e^{-j\pi/4}$ is collinear with $\mathbf{V}_r$ so that $\psi(x) = 45°$ and $|\mathbf{V}(x)| = |\mathbf{V}_r/2|$.

2. Since the angle $AMN'' = 116°$, and since two revolutions correspond to one wavelength, A' is on N'' when $x = (116/720)\lambda = 0.16\lambda$. For this value of x, $\psi = 19° + 116°/2 = 77°$.

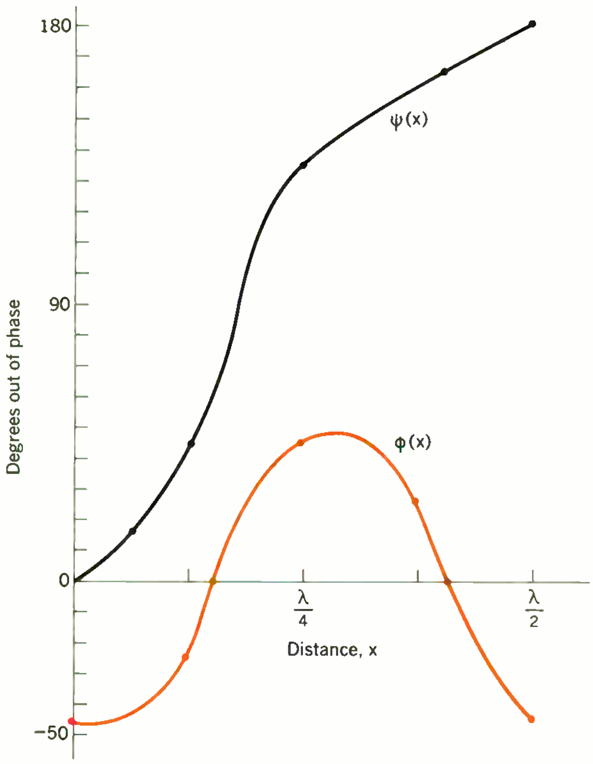
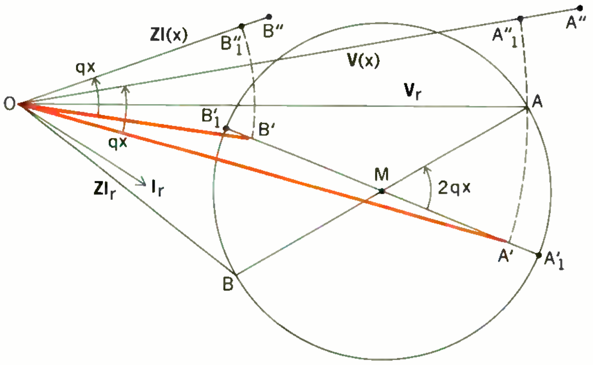


FIGURE 5. Curves of $\varphi(x)$ and $\psi(x)$ corresponding to the example of Figs. 3 and 4.

FIGURE 6. Principle of the complete Esclangon diagram for a transmission line with losses.



When x increases from 0 to 0.16λ, $|\mathbf{V}(x)/\mathbf{V}_r|$ decreases steadily from 1 to a minimum of 0.43; at the same time, $\varphi(x)$ increases steadily from $-26°$ to $0°$.

3. If $x = \lambda/4$, A' is on B and B' is on A, so that $\varphi(x) = 45°$, $\psi(x) = 45° + 90° = 135°$, and $|\mathbf{V}/\mathbf{V}_r| = \sqrt{2}/2$.

4. Since A' is on N'' when $x = 0.16\,\lambda$, it is on N' for $x = 0.41\lambda$. Obviously, $\varphi(x)$ equals 0, and we measure $|\mathbf{V}/\mathbf{V}_r| = 1.13$ and $\psi(x) = 19° + (116° + 180°)/2 = 167°$.

From these four results, one can now plot the curves of $|\mathbf{V}/\mathbf{V}_r|$, φ, and ψ against x; these are given in Figs. 4 and 5, along with the curve $|\mathbf{ZI}(x)/\mathbf{V}_r|$.

One can ask, "When do the curves $|\mathbf{V}(x)|$ and $|\mathbf{ZI}(x)|$ intersect?" This is equivalent to asking when $OA' = OB'$, and the answer is: when A' is on C_1 or C_2, C_1C_2 being normal to OM (see Fig. 3). With angle $ANC_1 = 19°$, we find that $x = 0.026\,\lambda$.

Lossy transmission lines

From Eq. (2), we obtain the following expressions for a line with losses:

$$\mathbf{V}(x)e^{-nx} = \mathbf{V}_1 + \mathbf{V}_2e^{-2nx}$$
$$\mathbf{ZI}(x)e^{-nx} = \mathbf{V}_1 - \mathbf{V}_2e^{-2nx} \tag{5}$$

These relations are quite similar to Eq. (4), and we can follow the same reasoning as that for lossless lines. We have only to remember that $\mathbf{Z}$ is no longer real.

From $\mathbf{V}_r$ and $\mathbf{ZI}_r$, we deduce $\mathbf{V}_1$ (represented by OM in Fig. 6) and $\mathbf{V}_2$ (represented by MA). The quantity $\mathbf{V}(x)e^{-nx}$ (represented by OA') is obtained by rotating MA through an angle of $2qx$ and multiplying its magnitude by e^{-2px}. Then, not only is OA' turned through an angle qx, but it is multiplied by e^{px} to obtain OA'', representing $\mathbf{V}(x)$.

A similar method is used to obtain OB'', which represents $\mathbf{ZI}(x)$.

Conclusion

Although the Esclangon diagram does not replace other methods for solving transmission-line problems, it is the most efficient way of showing the voltage and current variations along an electric line. Hence, the method complements both the Smith chart and the usage of basic formulas for a proper understanding of transmission-line problems.

REFERENCE

1. Esclangon, F., "Diagramme pour le calcul des lignes de transport d'énergie électrique," *Revue Générale de l'Electricité*, vol. 52, pp. 217–220, July 1943.

BIBLIOGRAPHY

Bewley, L. V., *Traveling Waves on Transmissions Systems.* New York: Wiley, 1933; 2nd ed., 1951.

Cahen, F., *Electrotechnique*, vol. 2. Paris: Gauthier-Villars, 1962–1964.

Kimbark, E. W., *Electrical Transmission of Power and Signals.* New York: Wiley, 1949.

Pauthenet, R., course given at the Institut Polytechnique de Grenoble (not published).

Süskind, C. S., *Electrical Transmission Lines.* New York: McGraw-Hill, 1951.

Michel Poloujadoff (M), born in Asnières, France, received the ingénieur diploma (master of engineering) in the power field from the Ecole Supérieure d'Electricité in 1955. After beginning a thesis on induction machines at the Sorbonne in Paris under the late Professor Esclangon, Dr. Poloujadoff received a one-year scholarship to study applied mathematics at Harvard University in Cambridge, Mass., where he obtained the M.S. degree in 1958. Returning to France, he completed his earlier thesis work and received the doctor of physical sciences degree. With the completion of his military service, he joined the Institut Polytechnique de Grenoble in 1961, where he has been primarily engaged in teaching the theory of electric machines and lines. His research has been oriented mainly toward the theory of induction machines and linear induction motors. Dr. Poloujadoff's recent interests have been in the computation of magnetic fields and in the theory of symmetrical components as an eigenvalue problem. A membre lauréat of the Société Française des Electriciens, he has published 37 papers and three books. His most recent opportunities to visit the American continent have been at the invitation of NSF, and as a visiting professor at Laval University.