VOLUME 7 □ NUMBER 7 IEEE spectrum JULY 1970

features

23 Spectral lines: What we must do

Many will say that survival problems are social and political, and that technical solutions already exist and need only to be applied—and it is the scientist and engineer who understand best how to accomplish this

24 Electric generating prospects for nuclear power

Manson Benedict

The breakeven cost, at which electricity produced from either nuclear or coal-burning plants would be the same, is in the range of 27 to 31 cents per million Btu

31 Science in the seventies—the policy issues

Hubert Heffner

As we enter the 1970s we find that the economic climate has changed. If we want to do something new, we must stop doing something old. In other words, we must make choices, and to do so rationally, we must develop a science policy

35 Circuit breakers: physical and engineering problems

I—Fundamentals

W. Rieder

The main goal of circuit-breaker development is to reduce the thermal time constant of the arc column and to reduce thermal and electric stresses on the arc

44 Reflection and crosstalk in logic circuit interconnections

John A. DeFalco

The high speeds of today's integrated circuitry have given the interconnections of digital logic circuits the characteristics—and the problems—usually associated with transmission lines

51 The role of the engineer in public affairs

Maurice Ponte

We are now living in an engineer-dominated age, that of the computer, which is at least as important as the Renaissance. Its consequences are not only technological and economic, but they are cultural and political as well

63 Least-squares estimation: from Gauss to Kalman

H. W. Sorenson

Since Gauss was very concerned with the computational aspects of least-squares applications, one can imagine that he would appreciate the practical benefits of the Kalman filter, which lends itself to digital-computer solutions

the cover

Signal reflections on transmission lines may be encountered from both linear and nonlinear terminations. The article on page 44 describes a graphical technique that calculates not only reflections arising from linear terminations, as depicted on this month's cover, but those from nonlinear terminations

departments

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

IEEE SPECTRUM EDITORIAL STAFF Ronald K. Jurgen, *Managing Editor;* Robert E. Whitlock, *Senior Editor;* Gordon D. Friedlander, *Senior Staff Writer;* Marcelino Eleccion, *Staff Writer;* Evelyn Tucker, Paul Hersch, Alexander A. McKenzie, *Associate Editors;* Stella Grazda, *Editorial Assistant;* Ruth M. Edmiston, *Production Editor;* Herbert Taylor, *Art Director;* Janet Mannheimer, *Assistant Art Director;* Barbara Gordon, *Assistant to the Art Director;* Morris Khan, *Staff Artist*

IEEE PUBLICATIONS BOARD M. E. Van Valkenburg, *Chairman;* F. S. Barnes, *Vice Chairman;* 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; David Slepian; 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*

World Radio History

NEW

COUNTS AND MARKS FREQUENCY WHILE SWEEPING!

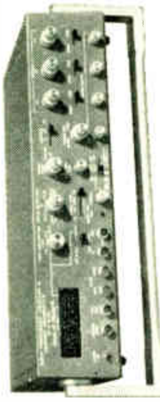
MODEL 8323A

The Model 8323A is a DC to 115 MHz frequency counter and frequency marker system. In this unit, a counter and a variable PULSE marker are combined in a completely new way to mark and count the frequencies of a VCO sweeper, while it is sweeping. The PULSE bypass markers provide settability and readability to better than ±0.25% of sweep width—no matter how narrow the width. The frequency counter provides counter accuracy of better than ±0.0002% and readout to the 7th significant digit.

- Three independent markers are generated to mark and count any point in the sweeper range —and to set the ends of the sweep with the same accuracy.
- A versatile sweep drive (VCO control) generator is part of the counter-marker system. It provides sweep rates from 100 sec/sweep to 100 Hz/sec, manual and start-stop sweep controls, zero blanking or variable reference lines during retrace, and a one-shot sweep.
- In addition to its system role, the frequency counter provides general purpose performance over the full DC to 115 MHz range.

PRICE $1295.00

KAY ELEMETRICS CORP, Maple Ave., Pine Brook, N. J. 07058 (201) 227-2000 TWX 710-734-4347



Circle No. 3 on Reader Service Card

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

BOARD OF DIRECTORS, 1970

J. V. N. Granger, *President*
J. H. Mulligan, Jr., *Vice President*
Harold Chestnut, *Vice President, Technical Activities*
M. E. Van Valkenburg, *Vice President, Publication Activities*
R. H. Tanner, *Secretary*
R. W. Sears, *Treasurer*
F. K. Willenbrock, *Junior Past President*
S. W. Herwald, *Senior Past President*

H. P. Bruncke
Werner Buchholz
Seymour Cambias, Jr.
L. B. Cherry
J. K. Dillard
L. C. Hedrick
H. Heffner
D. M. Hodgin
D. G. Lampard
C. A. J. Lohmann
D. C. Ports
G. A. Richardson
C. F. Savage
George Sinclair
J. E. Storer
W. H. Thompson
J. G. Truxal
R. P. Wellinger
J. R. Whinnery

A. N. Goldsmith, *Editor Emeritus and Director Emeritus*
E. B. Robertson, Sr., *Director Emeritus*

HEADQUARTERS STAFF

Donald G. Fink, *General Manager*

John L. Callahan, *Staff Consultant*
Richard M. Emberson, *Director, Technical Services*
Ralph H. Flynn, *Director, Publishing Services*
Elwood K. Gannett, *Director, Editorial Services*
William J. Hilty, *Director, Convention and Exposition Services*
William J. Keyes, *Director, Administrative Services*
John M. Kinn, *Director, Educational Services*
Leon Podolsky, *Staff Consultant*
Betty J. Stillman, *Administrative Assistant to the General Manager*
Howard E. Tompkins, *Director, Information Services*

Committees and Staff Secretaries

Awards Board: Una B. Lennon
Convention Board: W. J. Hilty
Educational Activities Board: J. M. Kinn
Electrical Engineering Newsletter: I. S. Coggeshall
Fellow: Emily Sirjane
Finance: W. J. Keyes
History: W. R. Crone
Information Services: Howard Falk
Information Systems Advisory: H. E. Tompkins
Internal Communications: Audrey L. van Dort
Intersociety Relations: J. M. Kinn
Life Member Fund: W. J. Keyes
Long Range Planning: W. J. Keyes
Membership and Transfers: Emily Sirjane
Nomination and Appointments: Emily Sirjane
Publications Board: E. K. Gannett
Regional Activities Board: W. J. Keyes
Standards: J. J. Anderson
Student Branches: Robert Loftus
Technical Activities Board: R. M. Emberson

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. **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. Printed in U.S.A. Copyright © 1970 by The Institute of Electrical and Electronics Engineers, Inc. IEEE spectrum is a registered trademark owned by The Institute of Electrical and Electronics Engineers, Inc.

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.

IEEE also copublishes, with the Institution of Electrical Engineers of London, England, *Electrical & Electronics Abstracts* and *Computer & Control Abstracts*, as well as their companion titles journals, *Current Papers in Electrical & Electronics Engineering* and *Current Papers on Computers and Control.*

World Radio History

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.

Patent rights

"A Plea for a Proper Balance of Proprietary Rights" by Robert H. Rines in the April issue of IEEE Spectrum (pp. 41–46) was very timely in view of the fact that in January, after a year and a half of legislative research, Congressman John E. Moss of California introduced in the House of Representatives H.R. 15512: "A bill to create a comprehensive Federal system for determining the ownership of and amount of compensation to be paid for inventions and proposals for technical improvement made by employed persons."

The bill would, in effect, invalidate blanket agreements requiring employees to relinquish and preassign all patent rights to employers as a condition of employment.

In brief, it would require an employee to give his employer prompt notice of his inventions; but it would also require the employer to take prompt action in declaring a claim to each invention (or else releasing it to the inventor), filing for a patent, and compensating the employee by an amount representing the fair market value of the employer's right to the invention, adjusted to reflect the position and duties of the employee and the degree to which the operations of the employer (such as furnishing laboratory facilities) contributed to the invention.

Special provisions are included for the protection of an employer's trade secrets, where public disclosure of an invention might jeopardize the employer's position in the market.

For amicable settlement of disputes, the bill would create an Arbitration Board in the Patent Office. In each case referred to it, the Board would consist of five members—three appointed by the Commission of Patents, plus one named by the employee and one named by the employer to represent their respective interests. The Board's decision would be final and binding.

To describe the need for such legislation, Mr. Moss also had placed in the Congressional Record (January 22, 1970, pp. H191 and H192) an article entitled "Patent Rights for Employee Inventors" by Robert J. Kuntz, then first vice president (now president) of the California Society of Professional Engineers. Kuntz's article covers many of the same points as Rines' paper.

The National Society of Professional Engineers has not yet taken an official position on the bill, but is currently seeking the consensus of its membership.

Keith W. Henderson, P.E.
Mountain View, Calif.
Senior National Director from California
National Society of Professional Engineers

A three-point program for retirement security

There now is a substantial awareness among engineers of the difficulty experienced in building up retirement equity. The problem is particularly acute within the aerospace industry because of its dependence upon government contracts and the tendency, usually involuntary, for engineers to follow these from company to company. With each such move, whatever has been accumulated in equity is generally left behind and the waiting period is begun anew. Consider further that there are companies that offer no pensions at all and it is clearly possible for an engineer to reach retirement age and find himself without a nickel to retire on unless he has provided it for himself out of his own taxable income.

Other segments of the working population do much better. Doctors, lawyers, and other self-employed persons enjoy the benefits of the "Keough Act," a provision of the tax law that allows them to invest 10 percent of their income up to $2500 for retirement, and this amount is tax-exempt. Teachers and college professors have portable policies provided through the TIAA.

A three-part program suggests itself for dealing with this problem. The program has the virtue that each of its parts is individually advantageous so that they may be and should be pursued in parallel.

1. *Extend the Keough plan to the salaried engineer.* This would allow anyone who does not enjoy a company-sponsored pension plan, or does not consider its benefits adequate, to buy his own out of tax-deductible income. Likewise, employee contributions would be tax-free. Federal legislation will be required to bring about this change. My suggestion is that you make a copy of this letter and enclose it with a personal note to both your Congressman and Senator.

2. *Urge the IEEE to provide a group retirement policy* similar to its present life and disability policies. This would permit those who purchase their own policies, with or without Keough plan extension, to do so at the lowest possible rates.

3. *Urge industry to provide a retirement plan cash option.* The employee would then have the options of participating or of retaining the employer's contribution as income or of applying it as premium to his own policy. The prudent engineer will have purchased his own policy early in his career and carried its costs during those periods when there were no employer contributions. If the Keough plan is extended, he enjoys its tax advantages. If the IEEE sponsors a policy, he enjoys its group rates.

Industry will not adopt such a plan overnight. Nevertheless, enlightened industry leaders will appreciate the legitimate aspirations of the engineering community and will begin to respond. That response may not be in the form of the suggested cash option, but rather by reduction of the waiting period for participation and vesting. In any case, industry must respond to the problem because the alternatives will be a greater drift toward unionization and fewer young men entering the profession.

Arthur L. Rossoff
Vice Chairman
IEEE Long Island Section

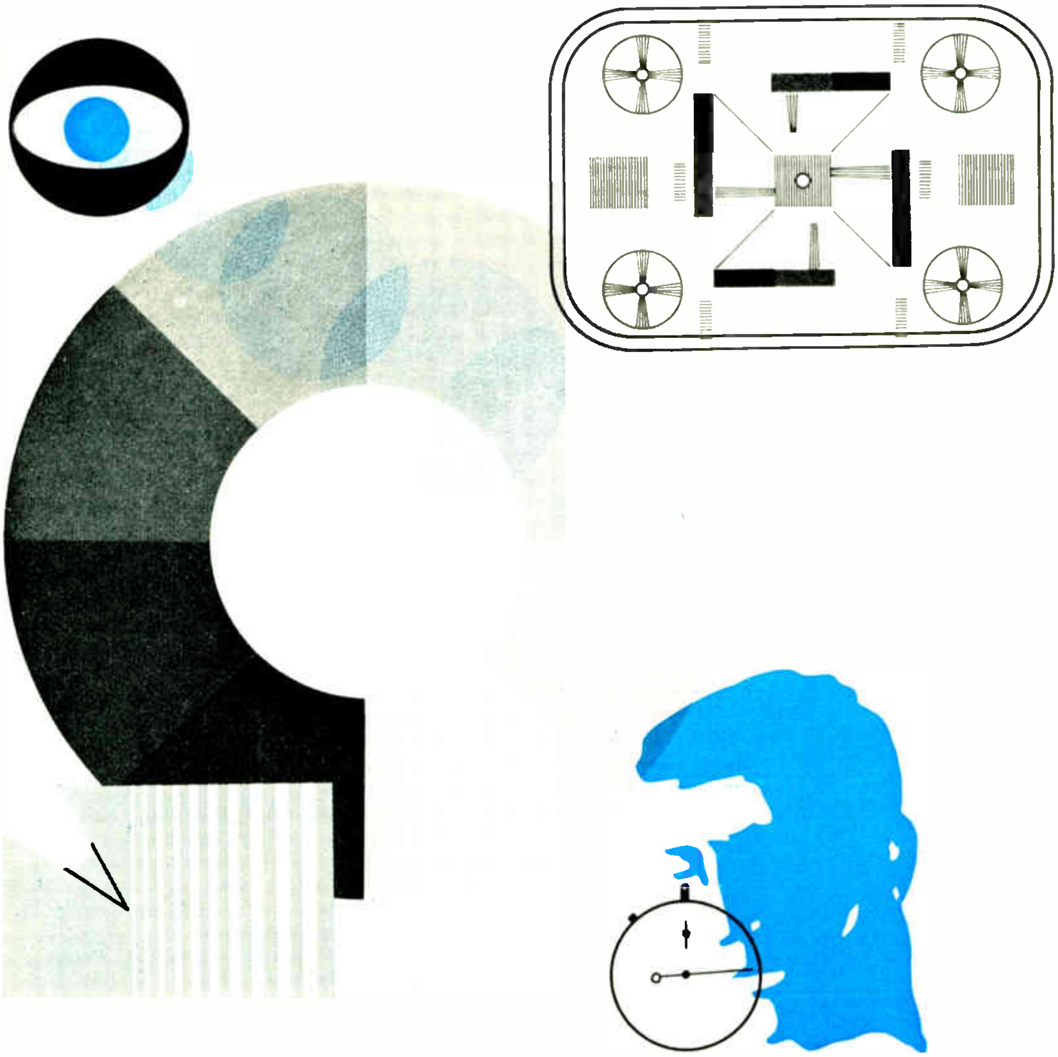
Computers available

In "Spectral Lines," April 1970, Dr. M. E. Van Valkenburg pointed out the need for modern laboratory equipment for electrical engineering education and the urgency of satisfying that need if education is to meet the challenges of the day. I know of a way partially to solve one of the problems that Dr. Van Valkenburg outlined. This problem has to do with providing laboratory equipment in the computer systems area where the products of a rapidly developing technology have a high price tag that may be out of reach of many.

It seems desirable to provide hands-on experience with a dedicated computer such as a minicomputer if the cost can be justified. These computers can be used to control laboratory instruments, record data, and perform

Circle No. 5 on Reader Service Card ⟶

Now There's A Storage Terminal That Does What Storage Terminals Can't Do!



A Revolutionary Scan Converter and Video/Graphic Storage Terminal

The PEP-400 is the only unit that combines all the desirable Video and Graphic storage functions in a single, yet, low cost terminal.

Like 1200 TV line resolution (with an 800 line option) at a cost of just $4,250.00*.

Like full gray scale.

Like retention time of 15 minutes to an hour under active "30 frame/sec" readout.

Like scan conversion from DC to 30 MHz.

Like electronic zoom capability of 50 to 1.

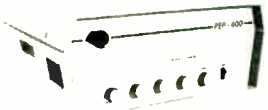
Like selective write and erase capability, a video write speed of 30 MHz in full gray scale and at least a dozen other exclusive new advantages.

No other storage media does all these things. They do far less but usually cost more. No other terminal is so well suited for video, x-y and scan conversion applications.

An important new breakthrough makes it possible. The PEP-400 utilizes the LITHICON™, a new storage tube that goes places other tubes can't.

But we've got a brochure that gives you all the details. You'll be doing yourself a favor if you send for it before you buy any storage unit now on the market.

*F.O.B. North Brunswick, N. J.



PRINCETON ELECTRONIC PRODUCTS, INC.
P.O. Box 101, North Brunswick, N. J. 08902 • 201/297-4448

Circle No. 4 on Reader Service Card

on-line computation. Hardware and software system design projects that can motivate many of today's students are also possible. But most universities do not have the funds to purchase such new and expensive equipment. Extramural sources for equipment are needed.

A potential source of more than 1000 reliable minicomputers from the Minuteman I missiles, originally costing the U.S. taxpayer $234 000 each, are scheduled to be declared excess by the Air Force during the next four years. Tulane has acquired two of the NS-10Q1 airborne guidance and control systems for use on DOD research contracts. Several other universities have also acquired identical systems for use on DOD and other government-supported research projects. Although these systems were designed by Autonetics for a specific application, the D17 general-purpose computer portion is an extremely flexible minicomputer.

The memory of the D17 consists of approximately 65 kilobits organized around a basic 24-bit word. Split-word computation is possible with an instruction repertoire that consists of 39 flexible instructions. Such operations as instruction modification are possible.

Any university with government-supported research can acquire a D17 at a cost of $40 for shipping if the Government Contracting Officer will approve the relevant use of the computer on the research project. How can an electrical engineering computer systems laboratory benefit? We acquired two D17 computers for on-line data acquisition, but it was necessary to make certain hardware modifications. Also, a complete software development was required. The hardware design and software development provided the basis for several laboratory and M.S. thesis projects. The government excess equipment provided a source of assistance to the electrical engineering laboratory, and the laboratory projects contributed significantly to making the equipment useful for the intended research. The portable, general-purpose characteristics of the D17 permit it to be shared efficiently between research and educational tasks.

The Minuteman Computer Users Group met at Disneyland Hotel on June 11–12, 1970. This meeting provided an opportunity for the sharing of hardware and software developments, education and research applications, and related questions. Both those who have obtained these computers and those who are considering acquisition attended this meeting. Government representatives discussed other available equipment items. Autonetics has provided considerable assistance to this reutilization effort, and appropriate personnel participated in this meeting.

The following contact can provide information concerning acquisition of a D17 computer: Robert E. Fink, Defense Supply Agency, DSAH-LSR, Cameron Station, Alexandria, Va. 22314, or (202) OX 4-6317.

Charles H. Beck
Tulane University
New Orleans, La.

The Esclangon diagram

In his article describing the Esclangon diagram that appeared in the February 1970 issue of the IEEE Spectrum, Michel Poloujadoff apparently overlooked W. C. Johnson's textbook, "Transmission Lines and Networks" (New York: McGraw-Hill, 1950). On pages 112–115 of his book, Johnson describes in detail both the construction and use of the Crank diagram, which, except for notation, is the same as the Esclangon diagram. It is interesting to note that Johnson does not reference his source for the Crank diagram.

As a student who has just learned about transmission lines, I would like to comment that, although the Esclangon or Crank diagrams do offer an interesting model for visualizing the variation of voltage and current down a transmission line, they do not offer any real computational aid.

John A. Latimer
Tennessee State University
Nashville, Tenn.

Mr. Latimer's comment is welcome. It is well to note that Professor Poloujadoff presented Esclangon's method not as something new, but as something old that merited wider diffusion.

The Editor

Guarantees to consumers

This letter was prompted by two items in your "News from Washington" feature in March *Spectrum.*

I believe that minimum standards are needed for guarantees on consumer products. I, and many of my friends, have had sad experiences with products on which the manufacturer refused to respect his own guarantee. Minimum standards should in fact benefit reliable manufacturers of quality products, by forcing all manufacturers to respect comparable guarantee standards.

I also believe that the consumer has a right to be informed of hazardous products. A home burned in our neighborhood recently, and as a result we are all acutely aware of the color television fire hazard.

Regarding both guarantee laws and hazard information, the government agencies should, of course, make every

TOMORROW MOSINCS +5

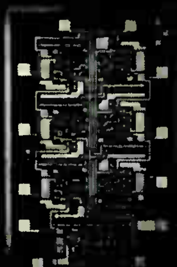
Plus 5 years in developing and perfecting MOS technology. Plus 5 new Tomorrow MOSINCS here today. All with the same advanced features that make it the optimum line to use: Low threshold voltages, low channel resistance, and faster speeds.

GENERAL PURPOSE 300-OHM SWITCH

HDGP1000. A 300-ohm diode-protected general-purpose switch. Breakdown −20 Volts, threshold −2 Volts. For long-time delay, low-frequency circuitry. 70c each (100 up).

GENERAL PURPOSE 30-OHM SWITCH/ AMPLIFIER

HDGP1001. General purpose 30-ohm diode-protected P-channel enhancement mode device. For easy interface of DTL or TTL circuits to MOS levels. Ideal for low cost applications. 80c each (100 up).

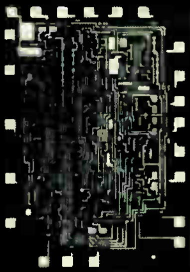


SIX-STAGE BINARY COUNTER

HCTR0206. Six circuits cascaded on a single chip. Buffered true outputs of each octave isolate cascading logic. Ideal for organ frequency divider applications. $3.30 each (100 up).

TV SYNC CIRCUIT

HSUB0525. The first IC for digitally controlling horizontal and vertical raster generation in TV sync circuit applications. Ideal for cable or closed circuit TV. $15 each (100 up).



DIFFERENTIAL DIGITAL ANALYZER

HSUB5021. For serial operation. Applied in implementing algorithims for rectangular and trapezoidal integration, or multiplication and division. $18.00 each (100 up).

HUGHES

HUGHES AIRCRAFT COMPANY
MOS DIVISION

We'll send you data on the newest MOSINCS, plus the entire line. Just write us:
Hughes MOS Division, 500 Superior Avenue, Newport Beach, CA 92663.

effort to be accurate, complete, and fair in their treatment of manufacturers.

At a time when traditional markets like defense and space are slowing in growth (be that desirable or not) and emphasis is shifting to the needs of people, the electronics industry would do well to reevaluate its attitude toward, and its standing in the eyes of, the consuming public.

J. H. Seamon
St. Louis, Mo.

Normal-mode analysis

The April 1970 issue of IEEE Spectrum contained a review of the book, "Introduction to Modern Electromagnetics," by C. H. Durney and C. C. Johnson. This generally complimentary review contained the statement that a normal-mode analysis of transmission systems is "unnecessary and unnatural."

We feel that this statement is ill-advised and needs considerable qualification. Normal-mode analysis may be unnecessary for simple ideal TEM transmission lines. However, for the analysis of more complicated multimode waveguides it is absolutely essential. Omission of this concept from a modern book would be inexcusable. We feel that one of the recommendable techniques of this book lies particularly in the inclusion of this powerful and essential concept. The authors should be commended for introducing this type of analysis.

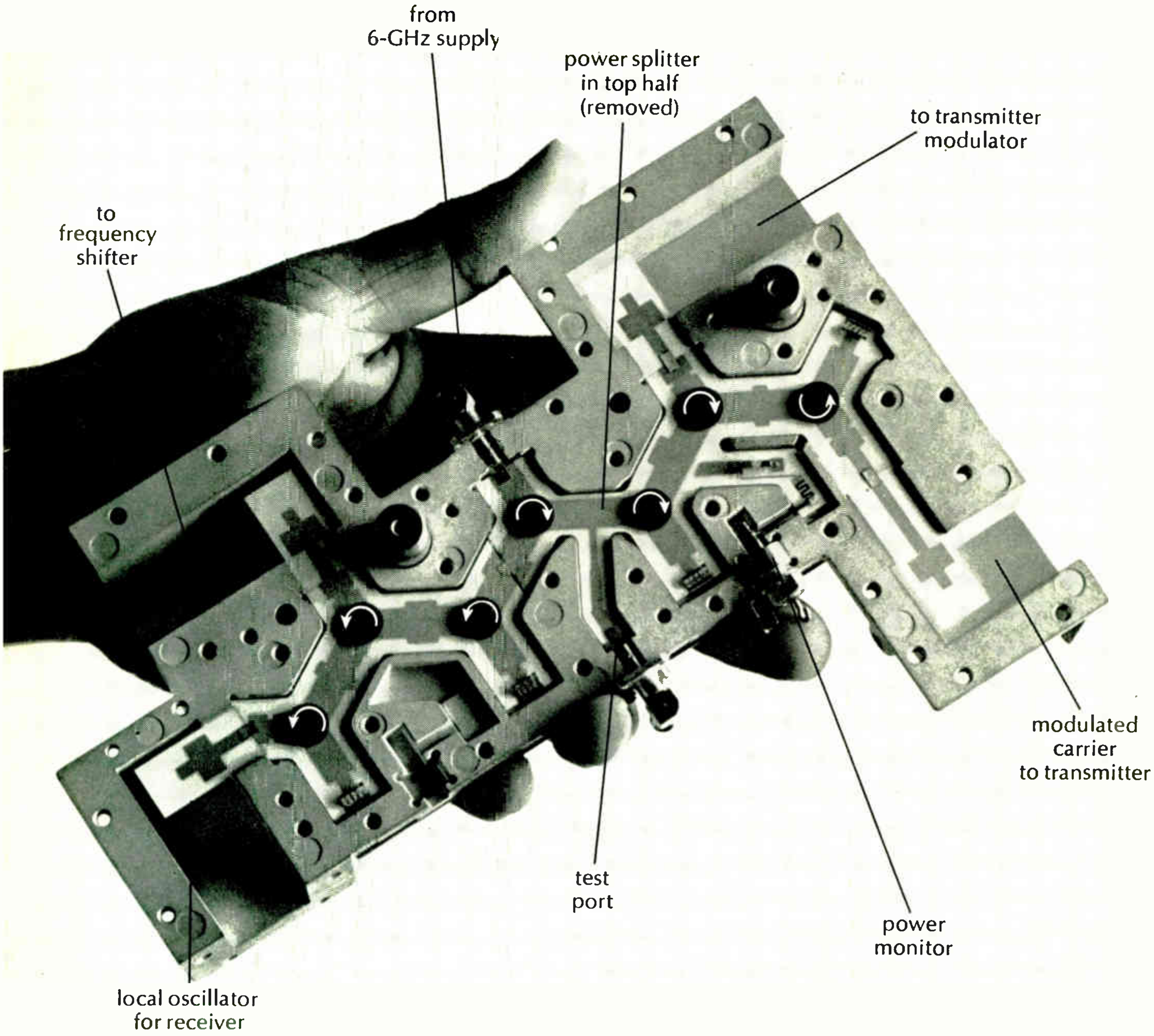
D. Marcuse
D. T. Young
Bell Telephone Laboratories
Holmdel, N.J.

Corrections

In the May article "Outlook for Binary Power Plants Using Liquid-Metal MHD," by L. L. Prem and W. E. Parkins, there is an error in the equation in text near the bottom of page 40, column 2. The equation for the overall efficiency of the binary plant should be as follows:

$$\eta_T = [\eta_M + (1 - \eta_M)\eta_S]\eta_F$$

In the May article by H. T. Hochman and D. L. Hogan, "Technological Advances in Large-Scale Integration," the flip-flops (tinted boxes) in Figs. 3 and 5 (pp. 51 and 52) were mislabeled. In Fig. 3 they should read, from top to bottom, FF_4, FF_3, FF_2, FF_1, FF_5. In Fig. 5 the first column should read FF_{41}, FF_{31}, FF_{21}, FF_{11}, FF_{51}. The second column should read FF_{4n}, FF_{3n}, FF_{2n}, FF_{1n}, FF_{5n}.



Miniature crossroads for microwaves

At every repeater station of a microwave relay system, there's a microwave distribution network—a circuit assembly that combines, divides, and directs the signals of one transmission channel. It interconnects the waveguides with coaxial cables that distribute microwave power to frequency mixers, modulators, and amplifiers.

Now, engineers at Bell Laboratories' Allentown, Pennsylvania, location have developed an integrated-circuit version. This one structure, smaller than a cigar box, has only a tenth the weight and a fifteenth the volume of the previous assembly. And, it costs less.

The network is shown above with its top half removed. The paths for the microwave signals are "stripline" —small rectangular channels with a copper-strip center conductor, electrically much like coaxial cable. The conductor strip is plated over an evaporated thin gold film on a ceramic substrate. Terminations and resistors are made by depositing tantalum nitride on the substrate. The four cross-shaped stubs (at the ends of the stripline) are stripline-to-waveguide transducers.

The seven black disks on the center conductor are ferrite microwave circulators, three-port devices which let microwave power flow from any port to the next one in the indicated direction only. This controls signal flow and isolates circuitry. The power splitter in the conductor feeds the test port.

Bell Laboratories engineers and their colleagues at Western Electric carefully selected this combination of modern materials and the techniques for working with them—including precision aluminum die casting and tantalum and gold thin-film technology. Analytical studies defined the geometry of the various circuit components to meet the rigorous standards of long-distance communications. This resulted in a superior component for our radio relay system and, at the same time substantial reductions in cost, size, and weight.

From the Research
and Development Unit
of the Bell System:

Bell Labs

VOLUME 7 □ NUMBER 7 IEEE spectrum JULY 1970

Spectral lines

What we must do. We would probably all agree that the simplest description of what the electronics engineer has done is to affect communication, making it better and faster. As I would like to use the term, communication means the transfer of information; it includes the printed page, the motion picture, the telephone and cable, radio and television, and the transport of people and materials when this transport involves information, as witness our space travel. Transfer of information also takes place inside an inanimate system, such as the electronic computer. The computer, as it modifies, manipulates, stores, and condenses information before passing it on to the user, becomes an integral part of communication also. The explosion during the last 50 years in the speed, and in the quantity, of information flow has largely been the result of electronics. The effect has been profound and not all good. There are things we must do.

Modern communications has potentially expanded man's life by a large factor. The expansion is not measured in terms of clock time, but in terms of experience, or what we might call subjective time. For a living being, time seldom passes according to the clock but is judged in terms of experience. There is an interesting paradox here, in that time spent in interesting pursuits passes rapidly, whereas time spent in idleness or illness drags. Yet, in recollection, we exactly reverse this, and we remember time largely in terms of the interesting experiences. But it is recollection that counts, since that is when stored information is used. Thus, the paradox is resolved in that we may conclude that more experience, and more information, are subjectively equivalent to a longer life.

To carry the thought further, consider a man who lived 70 clock years from, say, 1850 to 1920, and add up his potential experiences as a measure of 70 experience years. His great-grandchildren, living from 1910 to 1980, could, if they fully absorb their potential of information (experience), easily exceed 400 experience years. Their children, 1930 to 2000 A.D., may exceed 700 experience years. This expansion is what improved and speeded-up communication has made possible.

For those as yet unborn, the effect will be even more dynamic. They will jump, so to speak, on an express train that we older people saw accelerate from a low speed. It's small wonder that there's a "generation gap," and that it's difficult to know how to handle the rapid widening of this gap now taking place. Margaret Mead has one way to express it (*Culture and Commitment*. New York: Natural History Press—Doubleday, 1970): In the past, as she puts it, the young learned from the old, but we are now passing through a time when both learn together, and into a period when the old will learn from the young.

Unfortunately, a longer life, whether it be temporal or experiential, is not necessarily a happier one. The extension of our senses via radio, television, and recording mediums have often brought us information that we might well wish we had never heard or seen. Our children, even by age 5, with their hours of exposure to the outside world through television, have had vicarious experiences whose effects we still cannot fully evaluate. When they are of college age and their total experience of the world has already surpassed that of their great-grandparents, they enter a sociological world planned by those great-grandparents, which is, all too slowly, being modified by grandparents and parents. Isn't it time now for those of us who caused the enhanced information flow to consider the quality of the information, and not just its rapid and accurate transfer? Cannot our "spaceship earth," or our "world village," be modified to bring us happier, rather than only *more*, information?

Professor John Platt (*Science*, vol. 166, pp. 1115–1121, Nov. 28, 1969) wrote a challenging and provocative paper for fellow scientists and engineers, which he called "What We Must Do." His message is that those of us in technology should temporarily neglect our longer-range interests, and the less urgent short-range ones as well, to spend a decade of concentrated attention on immediate world problems. His analogy is to the way we postponed our normal pursuits during World War II, to concentrate on radar, weapons, operations research, etc., to insure survival of our freedoms. Today, the challenge is greater because, if we read the information flow correctly, we're facing the imminent extinction of mankind through overpopulation, pollution, nuclear war, and internal anarchic breakdown by a generation gap no longer passable. Platt argues that a decade may be all the time we have left before the unfavorable processes become irreversible.

Many will say that survival problems are social and political, and that technical solutions already exist and need only to be applied. *Exactly so;* we don't need research, we need *application*. Although no segment of society can be excused from helping, it is only the engineer and the scientist who know what technical solutions there are and how to apply them. It is our job to recognize the problem, to show that a solution is feasible, and to persuade our "management," i.e., our government, our industries, and our people, to provide the support and the will to do the job. Above all, as Platt indicates, engineers know that exponential increases eventually convert to S-shaped curves, and that stability will come. We can become the beautiful people, *if we survive*.

E. W. Herold
Editorial Board

Electric generating prospects for nuclear power

In many sections of the United States at the present time thermal nuclear reactors compete with coal-fired electric plants. But for nuclear plants to achieve full potential it will be necessary to develop an economical fast breeder reactor

Manson Benedict Massachusetts Institute of Technology

Most of the nuclear power plants in the U.S. today are of the light-water variety. In many parts of the U.S. these plants are competitive with plants burning coal, but the electricity that they generate will be more costly in the future as uranium supplies deplete. A promising possible answer to the cost problem is the fast-neutron reactor, which produces more fuel than it consumes. Such a plant should also be a more efficient generator of electricity and this should produce less thermal pollution than a water-reactor plant. However, it takes time to breed the fuel to make these fast reactors possible, and until more is known about the performance of such reactors, final judgment about their economic feasibility must continue to be held in abeyance.

Nuclear fission of the heavy elements, uranium and plutonium, is a process that, developed to its full potential, can provide mankind with a practically limitless store of energy. The types of nuclear reactors currently in use for electric-power generation—gas-cooled reactors in France and England, heavy-water reactors in Canada and India, and light-water reactors in the United States—make use of only part of this enormous energy supply. These reactors effectively utilize only the scarce isotope uranium-235, which appears as one part of 140 parts of natural uranium. To achieve full utilization of natural uranium, it will be necessary to develop effective means for completely utilizing the abundant isotope uranium-238, which constitutes 99.3 percent of the element. This technology will require development of a different type of reactor—the fast breeder reactor.

Light-water reactors, the type built extensively in the United States today, as I shall show, can compete economically in many parts of the country with generating plants burning fossil fuel. However, their uranium needs are so great that these reactors cannot satisfy long-term energy needs unless much larger supplies of uranium are discovered than are known at present.

Light-water reactors

There are two important types of light-water reactors.

Pressurized-water reactor. The type of light-water reactor first used in the United States for power production is the pressurized-water reactor.

Figure 1 is a schematic of a nuclear power plant utilizing this type of reactor. The fuel supply (black bars) consists of bundles of stainless-steel or zirconium-alloy tubing filled with pellets of uranium dioxide, enriched to about 3 percent in uranium-235. When a sufficient mass of this uranium is surrounded with water (in color), it constitutes a critical mass capable of sustaining a steady nuclear-fission chain reaction. Heat from fission is liberated from the fuel and transferred to the water, which is heated to about 315°C as it is pumped through the reactor. The steam pressure, around 1450 N/cm^2 (2250 lb/in^2), is above the vapor pressure at this temperature, and so the water leaving the reactor is below its boiling point. The water is pumped to a steam generator, where it transfers its heat to water boiling at a pressure of about 690 N/cm^2 (temperature about 285°C). The resultant steam flows through a turbine that drives an electric generator. The steam is then condensed and the condensate is preheated and returned to the steam generator by a boiler feed pump, as in a conventional Rankine cycle.

Boiling-water reactor. A second type of light-water reactor, developed a few years after the pressurized-water variety but now used extensively in the United States, is of the boiling-water variety. Fuel for this reactor differs only slightly from that for the pressurized reactor. The main difference is that water in the boiling-water unit is held at a lower pressure, around 690 N/cm^2, and is allowed to boil inside the reactor. Steam and water flowing past the fuel are separated—the water is returned to the reactor by the circulator and the steam flows directly to the turbine, after which it is condensed and the condensate is preheated and returned to the reactor by the boiler feed pump. The boiling-water reactor needs no separate steam generator, as the reactor performs this function.

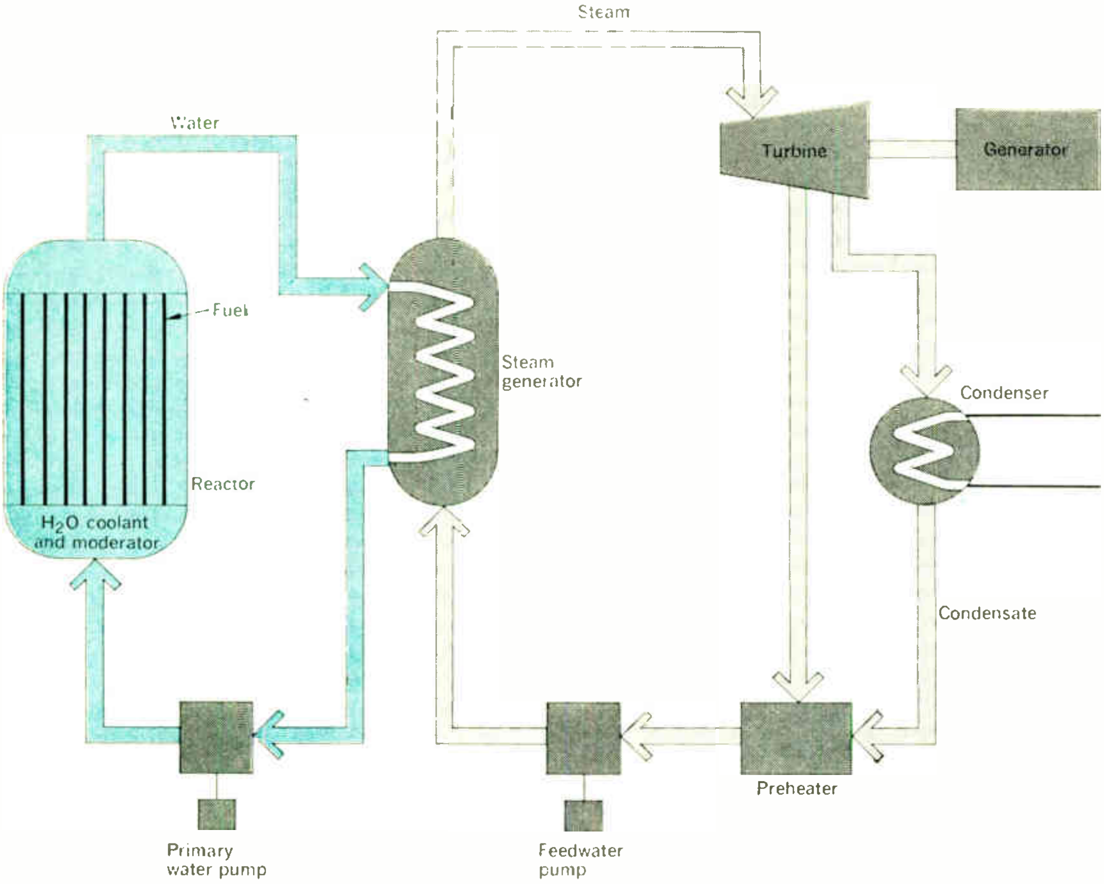


FIGURE 1. The essential elements of a pressurized-water reactor power plant.

Thermal efficiency

Since the steam entering the turbine in both the pressurized- and boiling-water reactors is at only 315°C, the thermal efficiency of both types of nuclear power plant is rather low (about 33 percent). (Compare it with an efficiency of around 40 percent for a modern fossil-fuel plant whose turbine receives steam at around 540°C.)

This low efficiency puts light-water nuclear plants at a slight disadvantage relative to fossil-fuel plants in disposing of waste heat, thus aggravating the so-called thermal pollution problem. (A power plant having thermal efficiency e must dispose of $1 - e$ units of heat for every e units of electricity generated. For a modern fossil-fueled plant with $e = 0.4$, the corresponding ratio is 0.6/0.4, or 1.5–25 percent less than for the water-reactor plant.)

Cost. Table I compares the cost of generating electricity in a 1000-MW modern coal-burning power plant and in a light-water nuclear power plant of the same capacity. The unit cost of the coal-burning plant is around $140/kW; costs for the nuclear plant are in the range of $180 to $200/kW, with little difference between pressurized- and boiling-water plants.

The cost of electricity, in mills per kilowatthour, is made up of fuel costs, capital charges, and operating, maintenance, and insurance costs. Fuel costs for the nuclear plant at today's uranium price (around $8 per pound of U_3O_8) is 1.5 mills/kWh. This cost is independent of location since the cost for shipping the few tonnes of fuel consumed per year by a nuclear plant is low. But fuel costs for a coal plant depend on the price of coal, which depends on how far the coal must be shipped. Near low-cost strip mines, coal prices provide heat at the low cost of 15¢ per million Btu. Far from coal mines, as in New England, heat from coal costs 35¢ per million Btu.

The cost of electricity from coal, then, varies with the fuel cost, from a low of 4.1 mills/kWh to a high of 5.8. Although capital charges for a nuclear plant, with its higher unit-plant cost, are somewhat more than for a coal plant, the total cost of electricity from a nuclear plant is between 5.1 and 5.5 mills/kWh. The break-even heat cost (at which the cost of electricity in both nuclear and coal-burning plants would be the same) is in the range of 27 to 31 cents per million Btu. (The average cost of heat from fuel burned in power plants in the United States today is around 27 cents per million Btu.)

I. Cost of electricity from new 1000-MW station*

	Coal			Nuclear	
Heat rate, Btu/kWh	8500	8500	8500	10 600	10 600
Unit plant cost, $/kW	140	140	140	180	200
Fuel cost, cents/10^6 Btu	15	25	35		
Electricity cost, mills/kWh					
Fuel	1.27	2.12	2.97	1.50	1.50
Capital charges	2.60	2.60	2.60	3.34	3.71
Operating, maintenance, and insurance	0.24	0.24	0.24	0.30	0.30
Total	4.11	4.96	5.81	5.14	5.51
Break-even fossil-fuel cost, cents/10^6 Btu				27.1	31.5

* 80 percent capacity factor, 13 percent fixed-charge rate. In this present period of exceptionally high interest rates, the fixed-charge rate is about 16 percent per year, but this high rate is not considered representative of long-term economic conditions.

Growth rate of nuclear plants

Because light-water nuclear plants can produce electricity at as low a cost as plants burning fossil fuel in regions where the power plant is far from gas fields and coal mines, they are being built in large numbers in the northeastern, southeastern, and Pacific Coast sections of the U.S.

A factor favoring the large-scale adoption of nuclear plants is that they cause practically no air pollution, since they emit no fly ash, carbon dioxide, sulfur dioxide, or nitrogen oxides. Opponents of nuclear plants may express alarm about radioactive effluents, but the fact is that the amount emitted from the light-water plants now in operation is less of a hazard than the minute amounts of radium in fly ash from plants burning coal.

The total capacity of nuclear power plants now operating in the United States is around 4000 MW, and a total of about 80 000 MW is under construction or being planned. These totals compare with the present total U.S. electric-generating capacity of over 300 000 MW.

Table II shows the growth of U.S. nuclear-power capacity as projected by the Atomic Energy Commission based on the relative costs of nuclear and conventional electricity and the projected increase in U.S. population and per capita consumption of electricity. The estimate through 1980 is likely to be accurate; for later periods, it is more uncertain.

Fuel requirements for light-water reactors

Before translating nuclear-power growth predictions into uranium requirements, it is important to understand a bit more about how the fission reaction is propagated and how nuclear fuel is utilized in a light-water reactor. Fission of uranium-235 in a light-water reactor is caused by absorption of a slow neutron in thermal equilibrium with the water in the reactor. Fission, however, produces fast neutrons. Therefore, in this type of reactor the neutrons must be slowed by repeated collisions with hydrogen nuclei in the water so that they can cause additional fissions.

II. Projected growth of U.S. nuclear-power capacity*

Year	Megawatts
1975	60 000
1980	145 000
1990	390 000
2000	735 000
2010	1 180 000
2020	1 725 000

* AEC's equation from WASH-1082. Extrapolated after 2000.

For every atom of uranium-235 reacting with a neutron, approximately two fast neutrons are produced. On the average, about four tenths of a neutron is consumed unproductively by water and other materials, and six tenths of a neutron is absorbed by uranium-238 to produce plutonium. One neutron remains to propagate the fission chain reaction.

The plutonium produced from uranium-238 is fissionable like uranium-235. Some of the plutonium undergoes fission before the fuel must be discharged after it is depleted. (However, the discharged fuel still contains enough plutonium and uranium-235 to justify reclamation and recycling to a later charge of fuel.)

To provide the initial charge of uranium, slightly enriched in uranium-235 needed by a light-water reactor, it is necessary to mine about 0.45 tonne of U_3O_8 for every megawatt of power-plant electric capacity. If subsequent charges of fuel are supplied in part by recycled uranium-235 and plutonium, and in part by additional fresh uranium, it is necessary to mine about 0.135 tonne of U_3O_8 for every megawatt-year of electricity produced.

Based on these material requirements and on the projected growth in nuclear capacity, shown in Table II, the cumulative uranium consumption of U.S. light-water reactors will be as shown in Fig. 2. By the year 2000, over 1.5 million tonnes, and by 2020, over 5 million tonnes, would be consumed.

These consumption figures are to be compared with known U.S. resources of uranium.

Uranium resources

Table III gives current estimates of U.S. uranium resources that can be produced at less than $8, $10, $15, $30, and $50 per pound of U_3O_8. Figure 3 is a plot of these quantities of uranium against price, on semilog paper, with the five given points arbitrarily connected by straight lines.

By combining this uranium-quantity-versus-price relationship with the uranium-quantity-versus-time relationship given in Table II, the uranium-price-versus-time trend shown by the dashed line of Fig. 4 is developed. The cost of electricity in a light-water reactor increases by 0.048 mill/kWh for each dollar per pound increase in the price of uranium. The solid line of Fig. 4 shows how the price of electricity from light-water reactors would be affected by the indicated increase. This plot suggests that, unless substantially more low-cost uranium can be found or a reactor developed that uses uranium more efficiently, nuclear power will be priced out of competition with coal in the U.S. before the year 2000.

Fast breeder reactor

Of the several possible types of reactor that make more efficient use of uranium than the light-water reactor, the liquid-metal fast breeder reactor is thought to have the greatest promise. In a fast reactor, water and other materials that are effective in decelerating neutrons are excluded; fission is caused by fast neutrons having nearly the same high energy as when produced.

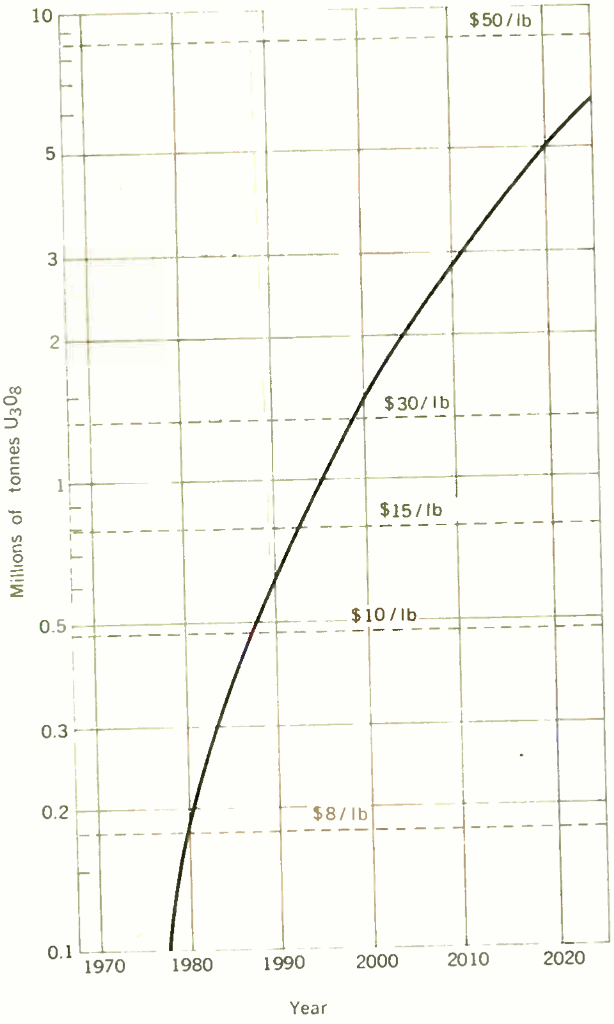


FIGURE 2. Predicted uranium consumption by light-water reactors in the United States.

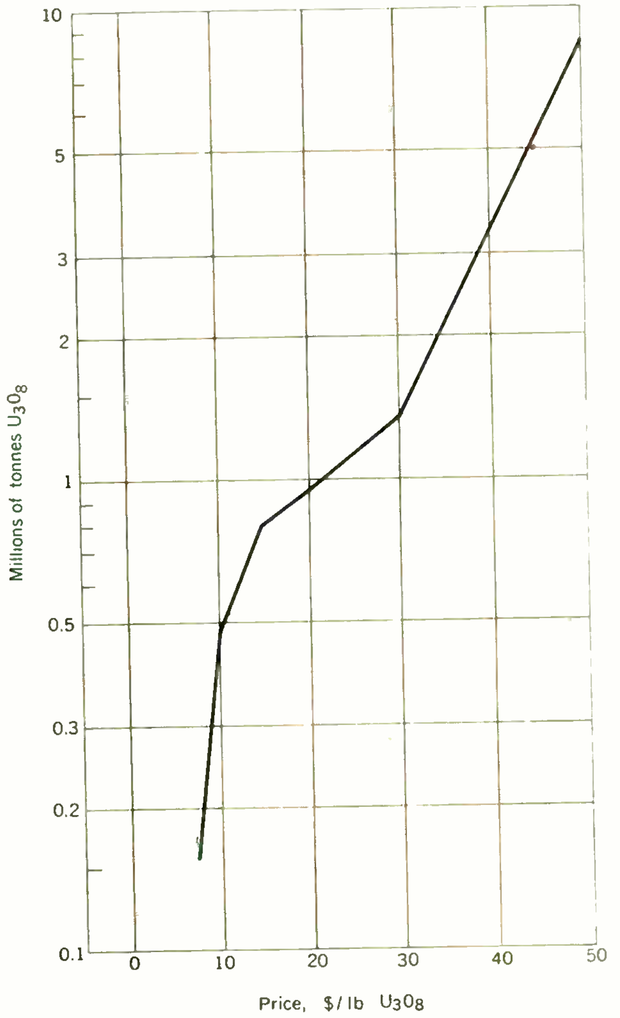


FIGURE 3. As uranium supplies become more difficult to come by, the price of uranium increases.

Because the rate of reaction of fast neutrons is much lower than thermal neutrons, a fast reactor, to be critical, needs a higher concentration and greater mass of fissionable material. This is a disadvantage. But a fast reactor also has an important advantage. When fueled with uranium and plutonium, about 2.5 neutrons are produced per neutron absorbed by the plutonium. Because water and other parasitic fast-neutron absorbers are absent, almost all of these neutrons can be used productively. One neutron continues the fission chain reaction and 1.5 neutrons convert uranium-238 into plutonium. In this way, a plutonium-fueled fast reactor can produce more plutonium than it consumes. It is said to breed plutonium. (In the example cited, the breeding ratio is 1.5.) The feed material for one of these breeder reactors is uranium-238, which is 140 times more abundant than uranium-235, the material for light-water and other thermal reactors.

This multiplication of fuel resources is a great attraction of fast breeder reactors. Another attraction is that fast breeders can use high-cost uranium without significantly increasing the cost of electricity—a point to be developed later.

Figure 5 is a schematic diagram of a liquid-metal fast breeder reactor. The fuel (black bars) consists of a mixture of 15 percent plutonium oxide and 85 percent uranium-238 oxide and is clad in stainless steel. Heat from these rods is transferred to liquid sodium, which leaves the reactor at about 650°C and atmospheric pressure. This sodium is intensely radioactive owing to the presence of sodium-24—the result of neutron capture. Because of the danger from radioactivity, should a leak develop in the steam generator (a leak that would intensify because of the violent reaction generated by sodium in water) heat from the primary, radioactive sodium is first transferred to secondary, nonradioactive sodium in the intermediate heat exchanger. Heat from the secondary sodium finally generates 540°C steam in the steam generator.

III. Estimated U.S. uranium resources

Price, $/lb U_3O_8	Tonnes U_3O_8 Mineable at This or Lower Price
8	174
10	477
15	794
30	1310
50	8525

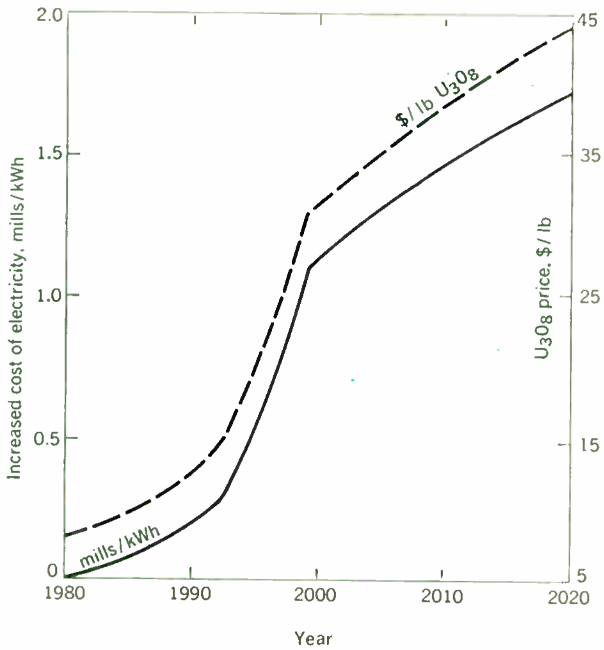


FIGURE 4. Trends in the price of uranium and the consequent cost of electricity produced by light-water reactors that incorporate plutonium recycling.

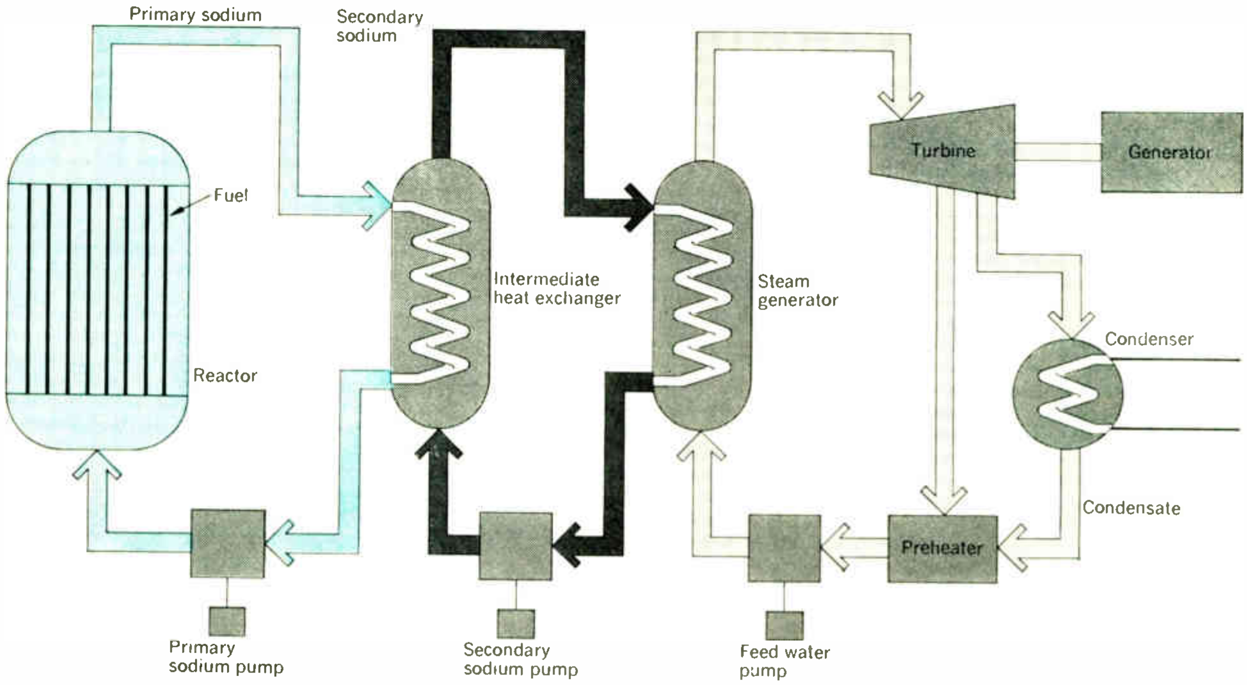
Because of the high temperature at which steam can be produced and because sodium is an excellent heat-transfer fluid that consumes little power to circulate, a thermal efficiency of 42 percent is projected for a liquid-metal fast breeder. This gives the liquid-metal fast breeder a substantial advantage over light-water reactors with respect to thermal pollution. It also has some advantage over coal-burning plants. Table IV compares the three plants.

Fuel requirements. Although the uranium-fuel requirements of a liquid-metal fast breeder reactor is easily satisfied with abundant uranium-238, plutonium-fuel requirements limit the rate at which such fast reactors can be brought into operation. The best fast-reactor power systems now being designed have a specific plutonium inventory of about 3.8 kg/MW electric output. To fuel the 735 000 MW of U.S. nuclear capacity predicted for the end of this century would require almost 3 million kg of plutonium.

The best fast reactors now being designed are expected to produce around 0.5 kg of plutonium, net, per full-power megawatt-year of electricity generated. Thus, in 3.8/0.5 = 7.6 years of full-power operation, the best fast reactor would produce enough plutonium to fuel another reactor of the same capacity. This is called the doubling time of the reactor system. Obviously, the shorter the doubling time, the sooner fast reactors can replace fuel-consuming water reactors.

Figure 6 shows the cumulative uranium consumption of a system of light-water reactors and fast breeders under two different assumptions. For the more favorable case, it is assumed that fast breeders with a doubling time of 9.5 years and a specific inventory of 3.8 are introduced in 1980 and are built as fast as plutonium availability permits. By the year 2000 no additional natural uranium will be required, and the total uranium consumption would have been only 600 000 tonnes, almost within the low-cost uranium resources known today. For the less favorable case, fast breeders with a doubling time of 19 years are introduced in 1990 and are then built as fast as plutonium availablity permits. In this case, natural uranium must be supplied until the year 2020, at which time the total uranium consumption would have reached 2 million tonnes—much greater than for the more favorable case, but still much less than the uranium needed for only light-water reactors of equivalent capacity. The actual

FIGURE 5. Neutron-moderating water is absent in the liquid-metal fast breeder reactor.



IV. Heat from power plants

	Light-Water Nuclear	Fossil-Fueled	Liquid-Metal Fast Breeder
Thermal efficiency	0.33	0.40	0.42
Heat rejected per unit electricity generated, (1 − e)/e	2.0	1.5	1.4

uranium consumption of a system of light-water reactors and fast breeders will probably fall between these values of 600 000 and 2 000 000 tonnes.

Cost of electricity

The cost of generating electricity in a liquid-metal fast breeder power plant is hard to predict. The doubling time and other performance characteristics achievable in a practical full-scale plant are quite uncertain. Yet, because they influence uranium requirements, they have an important effect on the cost of electricity. The value of plutonium affects the cost of electricity in two ways: (1) through interest charges on the value of the plutonium inventory, and (2) as credit for the value of net plutonium produced by the breeder. When the interest-charge rate equals the fraction of the plutonium inventory bred per year (the reciprocal of the doubling time), the plutonium credit just offsets the plutonium inventory charge. The cost of electricity is then independent of the value of plutonium and hence is independent of the price of natural uranium on which the value of plutonium depends. Since a typical interest-charge rate is 10.5 percent per year, a fast-reactor system with a doubling time of 9.5 years (the reciprocal of 0.105) would be required for electricity cost to be independent of the price of natural uranium.

Another, more significant contributor to the uncertainty in cost of electricity from a fast-reactor power plant is the capital cost of the plant. This cost will not be known accurately until a few large plants have been built.

Rather than predict the cost of electricity from a fast-breeder plant, I show in Table V what the unit cost of a nuclear plant would have to be to generate electricity for 5.14 mills/kWh and thus compete with a conventional plant burning coal at 27 cents per million Btu. Results are given for (1) a light-water plant with two extreme uranium prices of (\$8 and \$30/lb U_3O_8) and (2) fast reactors with the two possible doubling times of 9.5 and 19 years.

The first column, for a light-water reactor with uranium at \$8/lb, repeats the breakdown of the cost of electricity given earlier, leading to a cost of 5.14 mills/kWh with a unit capital cost of \$180/kW. The second column shows the strong effect that an increase in the price of uranium has on a light-water reactor: An increase in uranium price from \$8 to \$30/lb increases the contribution of uranium to the cost of electricity from 1.0 to 2.06 mills/kWh. This increase requires the unit capital cost to drop to \$123/kW to keep electricity cost at 5.14 mills—probably unattainably low.

The third and fourth columns in Table V are for a fast-reactor system with a doubling time of 9.5 years: The contribution of plutonium to the cost of electricity is zero, because, as I have explained, its doubling time equals the reciprocal of the interest rate of 10.5 percent per year assumed for fuel-inventory charge. Moreover, the cost contribution of uranium to the cost of electricity for this and other breeders is practically zero, because so little uranium is consumed.

The fuel-cycle cost of a fast reactor, excluding costs for uranium and plutonium, is expected to be somewhat higher than for a light-water reactor. Despite this higher fuel-cycle cost, the unit capital cost of a fast breeder reactor with a doubling time of 9.5 years can be \$221/kW and still provide electricity at 5.14 mills/kWh.

In columns 5 and 6, for a fast reactor with a doubling time of 19 years, the contribution of plutonium to the cost of electricity is around 0.25 mill/kWh for uranium at \$8/lb and 0.91 mills for uranium at \$30/lb. The increase is due to the increase in the unit value of plutonium that accompanies an increase in uranium price. For this reactor system to compete with coal at 27¢/million Btu, its unit capital cost would have to be \$207/kW for a uranium price of \$8/lb or \$172/kW for \$30/lb.

Groups that have studied the economics of fast-reactor power systems believe that unit costs in the range of \$220 to \$260/kW may be attained by 1985, with cost below \$220/kW possible a few years later. Thus, it seems that fast-reactor nuclear plants may become competitive with plants burning coal if their doubling time can be made as

V. Cost of nuclear power plants to generate electricity at 5.14 mills/kWh to compete with coal at 27¢/10^6 Btu

Characteristic	Reactor Type: Light Water		Liquid Metal,		Fast Breeder	
Doubling time, yrs	—	—	9.5	9.5	19	19
U_3O_8 price, \$/lb	8	30	8	30	8	30
Electricity cost, mills/kWh						
Operation and maintenance	0.30	0.30	0.30	0.30	0.30	0.30
Fuel cycle excluding uranium and plutonium	0.50	0.50	0.74	0.74	0.74	0.74
Uranium	1.00	2.06	0.00	0.00	0.00	0.00
Plutonium			0.00	0.00	0.25	0.91
Capital charges*	3.34	2.28	4.10	4.10	3.85	3.19
Total	5.14	5.14	5.14	5.14	5.14	5.14
Unit cost, \$/kW, of nuclear plant to bring cost of electricity to 5.14 mills/kWh	180	123	221	221	207	172

* To bring total cost to 5.14 mills/kWh; 13 percent/yr fixed-charge rate on plant, 10.5 percent/yr fixed-charge rate on fuel, and 80 percent capacity factor.

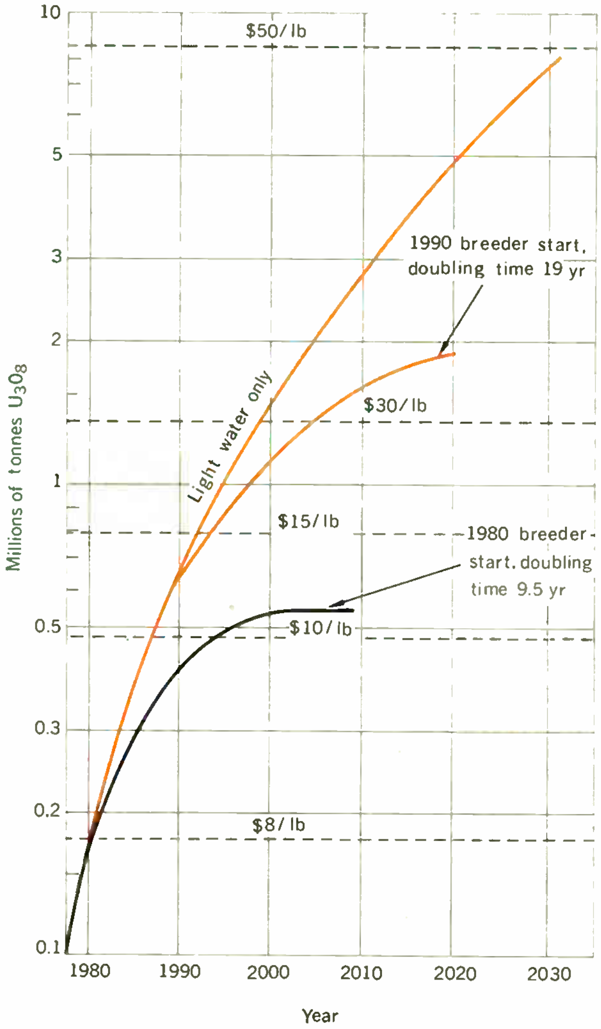


FIGURE 6. Uranium consumption varies with the time that the fast reactor is brought into service.

low as 9.5 years, or if enough uranium is found to keep its price around $8/lb until fast breeders are well established.

Status of fast reactors

Because of the key role fast reactors play in achieving the full potential of nuclear power, I'd like to summarize where they stand today.

The U.S.S.R. has operated several small, sodium-cooled, plutonium-fueled test reactors. A 350-MW (e) fast-reactor power plant is being constructed on the shores of the Caspian Sea and is expected to be in operation next year. A full-scale commercial fast reactor is expected to operate by 1976.

The United Kingdom has been operating the 15-MW (e) Dounreay fast reactor since 1959 and is building a 250-MW (e) prototype fast reactor, to operate in 1972. By 1985 the U.K. expects to have 15 000 MW of fast-breeder power plants in commercial operation.

France has been operating the 20-MW (t) Rapsodie test reactor since 1967 and is building the Phenix 250-MW (e) test reactor, to operate in 1973.

Germany, India, and Japan also have active fast-reactor programs.

The United States was a pioneer in fast-reactor development, but its schedule now lags some other countries. In the U.S. the EBR-I 1-MW (t) fast-breeder reactor began operation in 1951 and was the first reactor in the world to generate token amounts of electricity. The 20-MW (e) EBR-II experimental fast reactor has been operating since 1963 and is the principal U.S. facility now available for testing fast-reactor fuel. A larger fuel-testing facility, the FFTF costing $80 million, will be in operation by 1973. Three demonstration plants, with capacities around 350 MW (e), have been proposed for construction in the late 1970s, with the first commercial units in the U.S. in operation in the early 1980s.

Four key questions to be answered to determine how to generate electricity in the next century are:

1. How much more low-cost uranium can be found?
2. How soon can practical fast breeder reactors be developed?
3. How good will be the performance of fast breeder reactors?
4. What will fast breeder reactors cost?

Because of the multinational fast-reactor development, it seems clear that by the early 1980s there should be enough knowledge to determine whether the fast breeder reactor will provide the world with low-cost electricity in practically unlimited amounts or whether some other concept will be needed to fulfill our future energy needs.

My personal opinion is that the answers to all four questions will be moderately favorable, and that in the year 2000 the electric-power industry will be operating a mix of light-water and fast breeder reactors (and, of course, fossil-fueled plants) with substantially increased resources of somewhat more costly uranium, and with nuclear electricity costing a little less than it does today.

This article is based on a paper presented at the NEREM Session on "Perspectives in Energy Utilization," Boston, Mass., Nov. 6, 1969.

Manson Benedict returned to the Massachusetts Institute of Technology in 1951 after having received the Ph.D. in chemistry sixteen years earlier. In 1958, he was appointed head of M.I.T.'s Department of Nuclear Engineering, a position that he continues to hold. His principal fields of research at M.I.T. have been in isotope separation and fuel cycles for nuclear reactors. Dr. Benedict received the B. Chem. from Cornell (1928). After earning the Ph.D. in 1935 he did some postgraduate work at M.I.T. and Harvard and then entered the commercial world with positions at the M. W. Kellogg Company and Hydrocarbon Research, Inc. Also, during World War II, Dr. Benedict was head of the Process Development Division of the Kellex Corporation and was in charge of the process design of the gaseous diffusion plant for separating U-235. He eventually received the Perkin Medal (1966) for this separation work. His other awards include a Guggenheim Fellowship (1968), an American Chemical Society award (1963), and the William H. Walker and Founders awards from the American Institute of Chemical Engineers. Dr. Benedict is a member of the National Academy of Science and the National Academy of Engineering. He has been president of the American Nuclear Society (1962–63) and is a past director of the American Institute of Chemical Engineers. He is now serving the Atomic Industrial Forum as director.

Science in the seventies—the policy issues

Up until the last few years
there has been no coherent science policy in the United States,
because there has been never been a need for it.
The situation has changed

Hubert Heffner Office of Science and Technology, Executive Office of the President

The generosity that characterized research grants in the 1950s and 1960s is apparently at an end. Not only have the purse strings on scientific monies been tightened, but the economic climate and the general nature of scientific spending have eroded the value of the monies received. The scientific community is therefore faced with the problem of how best to allocate the funds that it does get. And although this community can't make policy regarding these allocations, it can help shape it. No effort is made in this discourse to state the specific steps to be taken—only to suggest the alternatives that are open and to stimulate the reader into accepting what is, inevitably, one of the important tasks to be faced in the 1970s.

I want you in the scientific community to do some hard thinking about the future of science, of technology, and of education in the coming decade. Conditions within the scientific area have changed within the last few years and urgently demand our attention. The problems are particular to no single nation: The answers may differ; the questions remain the same. However, I am most familiar with the problems of the scientific community in the United States, and I address myself—and the reader—to them.

This discourse is about "science policy." The first word, of course, comes from the Latin word for "knowledge"; the second has the same root as the word "politics." In a sense, then, I want to discuss the politics of knowledge. And make no mistake, the forging of a science policy is a political action. Scientists can strongly influence that action; alone, they cannot set it. My purpose here is to raise questions relevant to what is becoming a problem of increasing concern. I won't presume to offer any answers.

The United States has not had a coherent science policy. (Until recently we really have not noticed the lack.) The reason is not hard to find. During most of the postwar period, federal expenditures rose rapidly and it was easy, once a new need was identified, simply to add a new federal program to a burgeoning base. Thus, federal expenditures for R&D grew from $4.5 billion in 1957 to $16.8 billion in 1967—a growth rate of almost 15 percent per year. Policy making tends to be unnecessary unless choices have to be made.

Since 1967, we have experienced a radically altered picture. Rising science budgets abruptly leveled off. Many readers need no reminder that level budgets in times of inflation mean an actual reduction in scientific effort. Moreover, in research, there always seems to be an inflation, regardless of what the national economy is doing. As the total body of scientific knowledge grows, it seems to require more sophisticated, and hence more costly, techniques to obtain successive increments of knowledge.

It is a mistake to assume that current budget leveling off is temporary

A number of us who grew up in the heady expansion period of the 1950s and 1960s assume that the current budgetary leveling off is a temporary aberration—that once the Vietnam war is concluded, or once inflation is overcome, federal expenditures for science and technology will resume their upward climb. I do not believe that will happen; certainly it will not happen without a conscious policy decision.

To understand why I make this statement, it is only necessary to peruse the postulated national cash flow situation for the next half decade. Our gross national product is projected to grow at a rate of 4.3 percent per year, rising from $932.3 billion in 1969 to $1.2 trillion in 1975. Assuming no increase in taxes, federal funds available for expenditure will grow from $189 billion in 1969 to $206 billion in 1975. Although this appears to be a healthy increase, it rapidly evaporates when prior claims on the budget, in the form of Medicare, veterans' benefits, and so on, together with the President's proposed programs of Family Assistance and Revenue Sharing, are taken into account. As the report of the Council of Economic Advisors puts it, "The projected claims, which assume no addition to present Federal non-defense programs beyond those already proposed by the Administration, would approximately absorb all available resources through 1973 and leave room for significant additions only by 1975. The basic lesson of the estimates is that the country is already at a point where, despite prospective rapid growth of output, a decision to satisfy an existing claim on a larger scale or to satisfy a new claim will require giving up something on which people are already counting."

This, then, is to be the economic climate of the early and middle seventies. There is no room for "add-ons." If we want to do something new, we must stop doing something old. We must make choices, and to make rational choices we must develop policy. The issues are real and they are pressing.

Federal budgetary decisions must be made each year. Without generally agreed-upon policy, those decisions will be impelled by the crisis of the moment, leading to action and reaction without discernible long-term goals. And the decisions, as I have stated, must generally be agreed upon. That is how politics comes into policy.

The first issue is perhaps the most encompassing and the hardest to attack. It is simply, "How much research is enough?" Of course, a rejoinder to that question is, "enough for what?" Why does a society spend its resources on research?

Too often we as recipients of public funds for research have failed to consider the motivations of the suppliers of those funds. Too often we have assumed that our own motivation, the belief that the acquisition of new understanding is good in itself, is also the motivation of the public to supply billions for research.

I believe that the public has been relatively generous in its federal support of research for two—possibly three—reasons. One reason provides strong motivation; another is considerably weaker; the third is of possible future importance.

The first reason is the belief that research is essential if we are to carry out successfully our governmental responsibilities, first in defense, and then later in health care, space exploration, and education. The payoff in defense and space has been enormous. It has been somewhat less in health care, and, in my view, almost nonexistent in education. As national priorities change from defense and space expenditures to the domestic problems of environmental pollution, adequate health care, education, transportation, and urban regeneration, research will be on trial.

A second, much weaker, motivation has been the uplift to the human spirit and to national pride that comes from new fundamental discoveries about the nature of our world. Quasars, pulsars, DNA, and quarks not only are new additions to our semipopular vocabulary but are reminders of human achievement of high order. Such achievement is accorded public recognition, but it would be a mistake to overestimate the fraction of the gross national product that society would be willing to devote to champion science. After all, more people in the U.S. are familiar with the name "Joe Namath" than with the name "DNA."

The payoff in defense and space has been enormous

To my mind, we who have been engaged in research have been almost too successful in using federal funds appropriated under the first motivation for work that could be justified only under the second. Some of us now read with embarrassment the early justifications by scientists for appropriations to construct the first accelerators. Visions of great advances that would aid military- and commercial-energy purposes were placed before the eager Congress. Perhaps that shaky foundation explains in part why the budget for high-energy physics is now in trouble. Our success in obtaining federal research funds for certain projects and then diverting part to satisfy our own motivations is partly responsible for the present requirement of Section 203 of the Military Authorization Bill, which demands narrow relevance of defense research. This law may have repressed research for personal motivation, but in doing so I believe that it has made the criterion of relevance too narrow. And, excessive relevance can often prejudice the success of a mission-directed research program more than too much diversity.

In any event, we still face a problem resulting from the unequal weights attached to the different motivations and their differing proportions by scientists and by the public. As we enter the seventies, I believe we would be wise not to tout too loudly that research in unified field theory or radio astronomy will help solve urban transportation or water-pollution problems.

I suspect that a third motivation, another public purpose of research, is the need for research to ensure technological innovation.

There is general agreement that technological innovation is a component of economic growth—some economists believe it is the primary factor. Yet, there is certainly no general agreement that the assurance of technological innovation is the responsibility of the federal government. I believe it may come, however.

When looking at the high-growth industries, we find as outstanding examples, aircraft, computers, and electronics. When looking at the industries with a favorable balance of trade, we see virtually the same list of high-technology industries. These technologically intensive industries account for only 14 percent of the gross national product but they perform about 85 percent of all of the industrial R & D. These high-growth, favorable-trade-balance industries are precisely those where industrial research and development have been subsidized by NASA or DOD. As the R & D funds of NASA and DOD decline, we might worry whether there will also be a decline in the overall health of the economy as the pace of innovation is slowed. At that point, if not before, we may decide that government does indeed have a responsibility for industrial research. It will simply be an extension of a responsibility for economic well-being already reflected in the exercise of tariff and tax powers.

If government does take on the responsibility of stimulating technological innovation, presumably it can do it more efficiently than by relying on the "trickle-down" effect of space and defense expenditures. But there still remains the question, "How much research is enough?" And I leave that answer to you.

Let me turn to another major issue of policy. To what extent does the federal government have a responsibility for the education of scientists and engineers? The United States has not adopted the policy of many European countries, which accept the full responsibility for higher education in all fields. However, we have subsidized either directly or indirectly, in one form or another, a large fraction of the doctoral students in engineering, mathematics, and the physical sciences—the so-called EMP fields. We did this because of a general belief that we needed an increasing pool of highly trained scientific manpower, not only for our defense and space efforts, but also to cope with an increasingly technologically based society. There is a recent disposition by governmental agencies, however, to question whether needs have been met. There is, consequently, some thought being given to reducing direct grants to graduate students in these fields.

Thought is being given to reducing direct grants to graduate students

Before reacting to this possible cutback, it is wise to consider a few facts. First, the U.S. has been adding to its EMP doctoral manpower at about 9 percent per year. The attrition through retirements and deaths is estimated to be about 1 percent per year. Doctorate holders entering and leaving the United States are roughly in equilibrium. Thus, the doctoral manpower pool is increasing by about 8 percent per year—something like six times the growth rate of the general population. Moreover, reductions in the defense and space budgets are likely to reduce job opportunities in the near term.

On the other side of the coin is the need for highly trained scientists and engineers in activities where they have been in short supply. Examples include teachers at two- and four-year colleges, industrial management, applications engineering, government, and so on. Many of these positions have not been attractive to new Ph.D.s, but as the number of Ph.D.s increases and alternatives disappear their attractiveness will increase. A further factor complicates the manpower-supply picture. Over the last several years, the absolute number of bachelor degrees awarded in science and engineering has held constant; it may even be declining in some fields. This has occurred despite a rising overall college population. Since this is the pool from which advanced-degree candidates are drawn, it raises the question of whether this country may not in time be in trouble because of too few scientists and engineers.

There are other educational issues beyond simple supply and demand

The situation is further confounded because of the long delay between funding actions taken now and the effect on a Ph.D. population four to seven years later. We all know that electronic feedback systems with long time delays in the feedback loop can be unstable. Governmental systems are no exception. Fortunately, a panel of the President's Science Advisory Committee is analyzing the manpower problem.

There are other educational issues beyond the simple supply-and-demand questions. Even if those questions can be answered, issues still remain with regard to the degree and form of federal support required. Some have suggested that the benefits of education accrue overwhelmingly to the individual and that the federal government should therefore restrict its support to providing loans that enable students to attend college or the university. This view may have merit for the undergraduate. For the Ph.D. candidate, however, the facts are otherwise. On the average, the present value of the future earnings of the Ph.D. in engineering, mathematics, and physical sciences over his bachelor's degree colleague is said to be about $12 000, considering work-period differentials and possible interest accruals. This figure is less than his cost to attend postgraduate school. (If it were only an economic matter, a graduate student would do better to invest borrowed funds in securities rather than use them to acquire a doctorate.)

So far, I have mentioned two of the most important policy issues: How much research is enough, and what responsibility does government have for the education of scientists and engineers? I shall mention one or two others, perhaps not so central, but nevertheless important.

The first has to do with the extent of government involvement in carrying out research, and to what extent research should be conducted on contract outside of governmental entities. The issue is particularly pressing as attention shifts to problems of environmental pollution, urban blight, transportation, and housing. Should we attempt to form new government laboratories to do the necessary research and development, should we retread existing governmental laboratories, or should we look chiefly to the industrial, nonprofit, and university sector?

As in any tough issue, there are factors to argue on both sides. The federal government now operates well over 100 laboratories. They range in quality from good to awful, with—in my prejudiced view—more in the awful than good range. The reasons for that distribution are not hard to discern. A major one is the Civil Service system, which makes it virtually impossible to fire a marginally competent employee. However, unlike the universities that have much the same problem with tenure, very little attention is given to hiring only the best people available. It takes an exceptional manager to get around the federal bureaucracy and create a competent, dedicated, and effective research team. There, therefore, have evolved the federally financed research centers (FFRCs) and the nonprofit laboratories to circumvent these problems.

On the other hand, although the FFRCs, the nonprofits, and the universities may have more competent people on the average, they tend to be far less well coupled to the decision process of the agency that supports their research. What is the proper course for mounting new governmental research efforts?

Another problem that must be settled in the next year or so is how to organize the responsibility for research funding within the federal government. Most of us of my generation thought we had settled this issue. We believed that we had gained acceptance for the view that each mission agency had a responsibility to fund research that was broadly, not narrowly, relevant to its mission. We looked upon the National Science Foundation as the agency that would survey all of the fields of science, determine where mission agency support left gaps, and serve as the "balance wheel" to ensure that a satisfactory level of research was being pursued in all fields. It is possible that this concept is not so widely held as we had thought. Section 203 of the Military Authorization Bill, which directs that "none of the funds authorized to be appropriated by this Act may be used to carry out any research project or study unless such project or study has a direct and apparent relationship to a specific military function or operation," is possibly, though not certainly, an attack on this principle. Whether Section 203 is a straw in the wind or not, it and various proposals for a centralized Department of Science do contest the views of those who advocate agency responsibility for funding research. During the next few years, we must either confirm our earlier vision or reorganize our federal support of research.

I could go on at length listing the policy options now open and where decisions must be made. Perhaps I would do best, though, to leave you with a story that was first told to me by Congressman Morris Udall.

As you may have heard, Washington is divided into two classes—those who write speeches that they never give and those who give speeches that they never write. This story concerns a high Washington official of the latter category. He had a superb speech writer, who somehow was able to capture just the right mood for every occasion and produce speeches that his boss read to loud acclaim and contributed to his ever-increasing political influence.

After a time, this superb speech writer convinced himself that he was grossly underpaid. On the eve of the year's most important speech he went to the politician and asked for a raise. His boss said, "Now, now, son, you know we must fight inflation, and I want you to do your part. Forgo a raise in the knowledge that you are doing your country a service." The speech writer began to argue, but the boss was adamant, telling him that he could write his own speeches, and that if the writer weren't careful, his job would disappear.

Furious, the speech writer stamped back to his typewriter to prepare the evening's speech. He handed it to the politician that evening as he mounted the rostrum. The politician began on the first page.

"My friends," he read, "I bring you tonight a bold new program. It is a program that will not only reverse the tide of inflation, but will also solve our great environmental problems."

He turned the page. "It is a program that will bring low-cost medical care to every man, woman, and child in this glorious land. It is a program that will not only cure our agricultural surplus and improve the income of the farmer, but also will eliminate subsidies."

He turned to the next page, knowing that he had the complete attention of his audience. "My program is a program of full employment with declining prices for the consumer. It is a program that will bring law and order to our cities, peace to our campuses, and a guaranteed income for all. Moreover, ladies and gentlemen, my program will not cost one additional cent in taxes."

By this time his audience had straightened up, and hushed anticipation filled the auditorium. The politician turned to the last page—and saw a handwritten scribble, "All right, big boy, you are on your own now."

I relate that story to make my last point. As scientists and engineers we have promised a lot. It is now up to all of us to point the way whereby we can continue to deliver on those promises.

This article is based on a speech presented at the 1970 IEEE International Solid-State Circuits Conference, Philadelphia, Pa.

Hubert Heffner (F) began his college training in the field of physics (B.S. 1947), then changed tracks to receive the M.S. (1949) and the Ph.D. degrees (1952) in electrical engineering—all from Stanford University. For the following two years he worked for the Bell Telephone Laboratories, researching electron-beam dynamics. Returning to Stanford in 1954, he accepted a position as assistant professor in electrical engineering, was promoted to associate professor in 1957, and in 1960 was given a full professorship in electrical engineering and applied physics. That same year, he took leave from Stanford to serve one year in London as Scientific Liaison Officer of the U.S. Office of Naval Research there. Back again at Stanford, and until becoming Deputy Director of the Office of Science and Technology last July, Dr. Heffner assumed positions as acting chairman of the Applied Physics Division, and associate provost and dean of research. Dr. Heffner is a Fellow of the American Physical Society. He is on the IEEE Board of Directors and is a member and past chairman of the intersociety Joint Council on Quantum Electronics. He also served as chairman (U.S.) of Commission VII of the Union Radio Scientifique International (URSI), and (from 1961 to 1967) of the Working Group on Microwave Devices of the Defense Department's Advisory Group on Electron Devices. (He is currently a member of the parent group.) He had been a member of the NASA "Tycho" Space Science Study Group and the American Council on Education Committee on Sponsored Research. Dr. Heffner is a consulting editor for the McGraw-Hill Physical and Quantum Electronic Series and is the author of numerous technical articles on electron-beam focusing, noise theory, parametric amplifiers, and quantum electronics.

Circuit breakers
Physical and engineering problems

I—Fundamentals

Though an integral element of our vastly expanding worldwide power system complexes, switchgear represents one area that is largely taken for granted, even by the majority of technically oriented engineers

W. Rieder Brown, Boveri & Co., Ltd.

In this three-part survey article, some of the demands that traditionally have been made upon switchgear—and how these demands are being met—are described. Part I covers the basic problems associated with current interruptions, arc phenomenons, arc mediums, and testing. In Part II, to be published in August, some of the more general design aspects will be discussed. Part III will treat specialized problems connected with the conception of certain physical principles of current interruption.

If you ask the man in the street what a switch is, he will probably mention the switch on the wall of his living room, which he takes very much for granted. Switches of this type do not involve major physical or engineering problems, since they carry only a few amperes and disconnect some 100 volts, are switched on and off only a few times a day, and do not have to work extremely fast or reliably.

Even most electrical engineers know very little about switches, inasmuch as far less attention is paid to them at the universities than to electronics or rotating electric machinery. Since switchgear is not conspicuous and is not taught in detail, the problems that it poses are seldom appreciated. Books on it are few. But in a power network there are millions of circuit breakers, tap changers, isolators, motor switches, contactors, controllers, etc.—all widely different from our light switches at home. Their sales are by no means negligible, but their responsibilities are even more impressive. Thus the question arises of why they are assumed to be either too simple or too complicated to write or teach about. Indeed, in the beginning of the art, switching and circuit breaking seemed to be just a matter of common sense and experience. But soon one began to ask, "Why do electrical contacts become worse as their surfaces are made larger and flatter?" This problem was not solved until scientific investigations shed new light on the physics of electrical contacts, and led to the conclusion that one should aim not at a single, large contact area, but at one—or several—well-defined points of contact.[1]

With increasing demands upon interrupting capacity, life, reliability, size, and economy came the need for more knowledge of the physics of arcs. At present, trends in opposite directions confuse even specialists, since gas pressures both high and low, gas molecules both light and heavy, are used in circuit breakers, all with increasing success.

Fundamental duties of switchgear

Any switch or circuit breaker, regardless of its special application, is required to accomplish four fundamental duties:

1. When closed, it should be an ideal conductor.
2. When open, it should be an ideal insulator.
3. When closed, it has to be able to interrupt its as-

signed current promptly at any instant, without causing dangerous overvoltages. In some cases it must limit a fault current to a tolerable value.

4. When open, it has to be able to close promptly at any instant, possibly under short-circuit conditions, without being impaired by contact welding or the like.

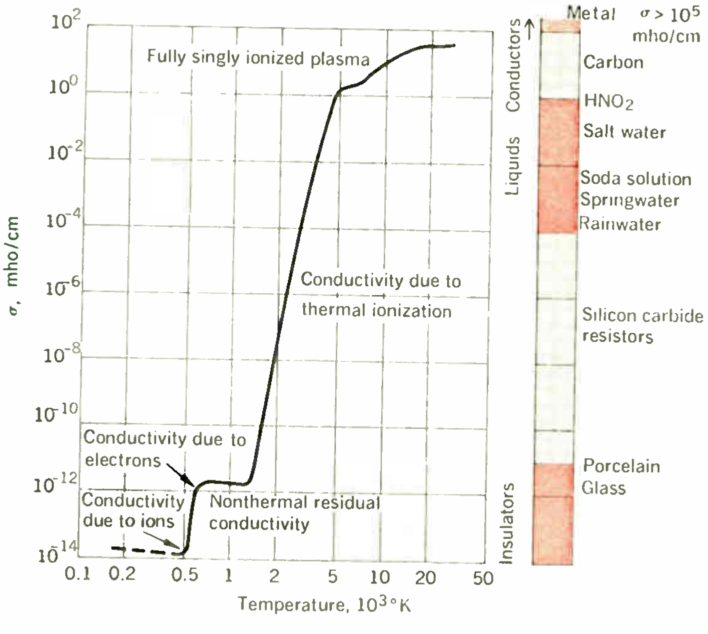
These duties lead to the contradictory requirements that make development of switchgear so difficult, and so uniquely fascinating. Fortunately, no single switch or breaker is stressed to extremes in every respect.

Telecommunication switches (such as telephone relays, selectors, etc.), which have extremely high switching frequencies (up to a hundred operations per second) and long life (up to a billion operations) but have to handle only very low currents and voltages (milliamperes, tens of volts), are carefully protected in clean rooms and are tended by specialists.

Control switches (for motors and other electric equipment) have to carry heavier currents (amperes to kiloamperes), but still at low voltages (hundreds of volts). As a rule, switching frequencies are reduced to a few operations per minute or less, but these switches have lives extending to tens of millions of operations. Often they are exposed to rough conditions and subjected to rough treatment by nonspecialists.

Switches and circuit breakers in the supply systems have to carry continuous currents up to 4000 amperes* and to interrupt service voltages up to 765 kV. Short-circuit currents as great as 70 to 90 kA must be interrupted extremely reliably within two to three cycles (33 to 50 ms) after tripping. Generator breakers may carry as much as 36 kA continuously and may be subjected to short-circuit currents up to 150 kA. For outdoor breakers the ambient conditions are very severe, and although the switchgear in supply systems is maintained by specialists, such maintenance work must be reduced to a minimum.

FIGURE 1. Electrical conductivity σ vs. temperature of air at atmospheric pressure.[10]



This article deals mainly with high-voltage circuit breakers; however, some of the considerations are more general, especially those in the section that follows.

The switching arc

The engineering aspects of arcing. In our switches at home we have merely to open the contacts a fraction of a millimeter. A small arc, of which the user is not often conscious, causes the current to continue until the next zero of the service-frequency current. Within about a microsecond after current zero, a space-charge layer able to withstand about 200 volts is built up at the new cathode.[2-4] For higher voltages, however, either more gaps in series must be used, or the arc has to be lengthened until the voltage it requires for survival is more than that available.[3] In this way the current is continuously reduced to zero by the increasing arc resistance.[5] Thereafter, the contact gap keeps the supply voltage isolated from the disconnected part of the circuit.

For voltages of the order of kilovolts, however, the arc lengths needed for interruption would be too great for practical applications (about one meter per kilovolt). Therefore the switching arc is often called the archenemy of switching, but this attitude is quite superficial.[6] If the arc did not prevent discontinuous current interruption at the moment of contact separation, an overvoltage $L\,di/dt$ would be produced that, with infinite di/dt, would kill the people and destroy the systems and be the end of all electrical engineering. But actually the arc preserves the current i until the energy $Li^2/2$ stored in the inductance L disappears by itself, without any damage, at the next current zero.

Avoidance of arcing. The arc could be avoided if the contacts were opened exactly at current zero. But it is extremely difficult to predict this moment with sufficient accuracy, and to move the contacts with sufficient accuracy and speed. To avoid an arc, the current must be less than 0.3 ampere. A sinusoidal 50-Hz current of 100 kA rms reaches this value 10 nanoseconds before its zero. Thereafter the restriking voltage rises at about 10 kV/μs; to avoid ignition due to field emission, a 10^{-5}-cm contact gap is needed after 1 ns. This requirement implies an acceleration of 10^{-5} cm/10^{-18} s^2 or 10^{10} g. Even for weaker stresses, the arc solves the problem of synchronization more cheaply and reliably, because it ceases spontaneously at the next current zero of the power frequency—if reignition caused by thermal inertia is avoided. This consideration does not exclude a reduction of arcing time and power and thus a rise of the interrupting capacity of a switch by synchronous timing of contact separation. Although such quasi-synchronous switches have not been manufactured yet for 60-Hz duty, some attempts are being made in this direction.[7,8]

Basic plasma physics. Since synchronous switching by mechanical means is not feasible, the question remains whether the arc is the only alternative and, if so, why.[9] The arc plasma is the only substance able to change its conductivity by orders of magnitude from that of a good conductor (10 mho/cm) to that of a reliable insulator (10^{-12} mho/cm) merely when the temperature changes by factor of ten (10 000 to 1000°K); see Fig. 1. Moreover, the change from the conducting to the isolating state occurs extremely rapidly. The conductances of semiconductor diodes and thyristors vary only by eight orders, and no faster than that of an arc. They are neither true conduc-

---

* In Japan, 8000-ampere breakers are under consideration.

tors nor true insulants, and they are extremely sensitive to current or voltage overload (as will be discussed in Part II of this article). A plasma has never been destroyed by overload.

The crucial physical problems of the switching arc are the mechanism of the change of conductance, the means of reducing its time constant, and the avoidance of reignition after current zero.

The conductivity of the arc column is affected by its temperature only, which is about 6000°K in open-air arcs of a few amperes, but probably up to 25 000°K and more in high-power circuit breakers at the current maximum.[11-13]

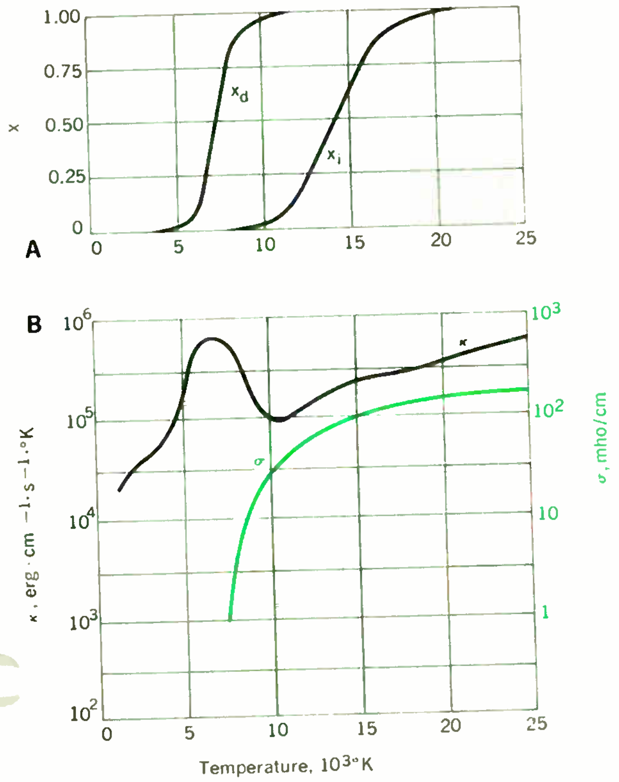
It is common knowledge that air molecules at room temperature (300°K) move randomly with remarkable velocities (averaging about 500 m/s) and collide very frequently at 10^{10} times per second (roughly equivalent to meeting every person in the world three times per second). Nevertheless, the kinetic energy of these molecules (the average of which is proportional to the absolute temperature) is too low to enable even the fastest particles to damage those with which they collide.

At higher temperatures, however, it happens that the molecules break down at the most severe collisions and dissociate into their atoms. Energies of 9.7 eV and 5.1 eV, respectively, are needed to dissociate an N_2 or O_2 molecule. The degree of dissociation x_d is described by the formula[12-17]

$$p\left(\frac{x_d^2}{1 - x_d^2}\right) = KT^{3/2}\exp\left(\frac{-W_d}{kT}\right) \tag{1}$$

and illustrated in Fig. 2 (where K = a constant, W_d = energy of dissociation, p = gas pressure, k = Boltzmann's constant, and T = absolute temperature).

FIGURE 2. (A) Degrees of dissociation x_d and ionization x_i and (B) thermal conductivity κ and electrical conductivity σ, all as functions of temperature.[14]



At still higher temperatures some molecules and atoms are deprived of an electron and the hot gas, now called plasma, becomes a conductor (Figs. 1 and 2). The degree of ionization x_i due to thermal collision is described analogously to Eq. (1) by

$$p\left(\frac{x_i^2}{1 - x_i^2}\right) = KT^{5/2}\exp\left(-\frac{W_i}{kT}\right) \tag{2}$$

and also illustrated in Fig. 2 (where W_i = energy of ionization).

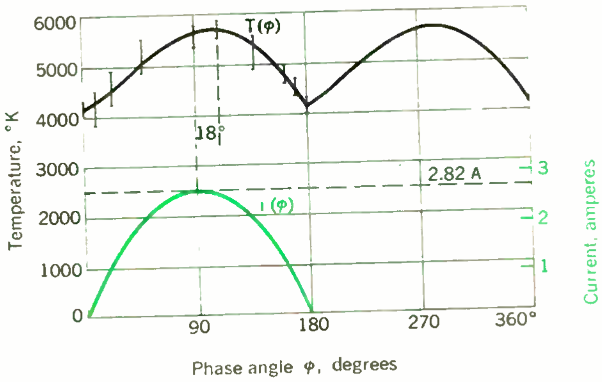
This thermal ionization, resulting from random collisions in a hot gas, must be distinguished from impact ionization, which is caused by electrons accelerated in an electric field between two collisions. The latter cause dielectric breakdown, even in a cold gas.

Arc dynamic problems. In opening contacts, the last remaining contact point melts due to Joulean heat and finally evaporates, initiating a switching arc, which remains stable as long as the Joulean heat of the arc current is equal to the power losses from thermal conduction, radiation, and convection.[11-13, 15-17] At the beginning of a current loop, additional power is needed to increase both arc temperature and cross section; at the end of the loop, however, this energy content must be dissipated. Therefore, the arc temperature does not decrease to room temperature without delay in relation to the current, but a residual temperature (and often even a residual conductivity) remains at current zero; see Fig. 3. This effect, of course, enables the recovery voltage appearing across the arc gap after current zero to sustain the arc either by dielectric reignition (if the conductivity has ceased but the density is still lower than at room temperature) or by Joulean heating (if there was any residual conductance).

The main goal of circuit-breaker development is to reduce the thermal time constant of the arc column and to facilitate thermal and electric stresses on the arc around current zero. Thermal time constant is defined as the time needed to reduce the conductance of the arc by a factor $1/e$ (= 1/2.7) due to temperature decay, according to Eq. (2) and Figs. 1 and 2, assuming no current is flowing. This thermal time constant,[5, 13, 17-20] given by

$$\vartheta = \text{const.} \times \frac{A}{D} \tag{3}$$

FIGURE 3. Current and axial temperature of a 2-ampere, 50-Hz arc in free air vs. phase angle ϕ.



(where A = arc cross section, D = diffusion coefficient or temperature conductivity) is of the order of milliseconds for low-current atmospheric arcs, but about one microsecond in modern circuit breakers at current zero.

ϑ is always much greater than the time (10^{-8} second) needed to obtain the degree of ionization corresponding to the momentary temperature [Eq. (2)] and therefore is decisive for the current interruption.

If there is some power input P while the arc decays (as in ac arcs before current zero) the conductance G decreases more slowly as power input increases and as power losses P_0 decrease, according to Mayr's theory[5,17,19,20]:

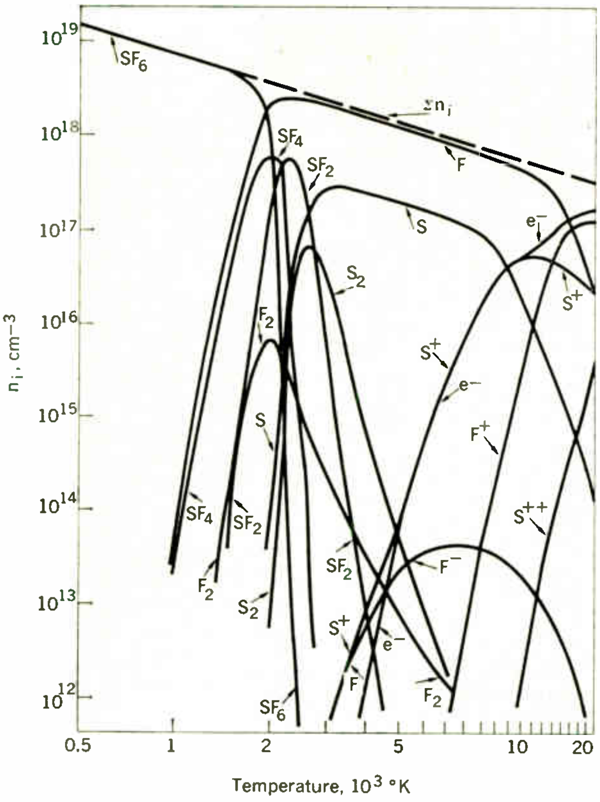
$$\frac{d(\ln G)}{dt} = \frac{1}{\vartheta}\left(\frac{P}{P_0} - 1\right) \tag{4}$$

From integrating the arc power balance it is known[11,13] that the arc cross section A (and, hence, ϑ) decreases with increasing power losses—for example, by forced convection in a high-pressure nozzle. Thus, even before the crucial current-zero interval, the boundary conditions of the arc can be optimized. Around zero P_0 seems to be influenced mainly by heat conduction.

The intermediate regime of thermal nonequilibrium reignition—the missing link between thermal reignition according to Eq. (4) and dielectric breakdown—is outside the scope of this survey.[13,20,21]

In molecular gases, heat conductivity is attributable not only to the transport of kinetic energy of the molecules by their random motion, but also to the transport of their energy of dissociation W_d, which was fed to decompose the molecule in hotter zones and is released to heat the ambient gas, if the single atoms recombine in cooler zones according to Eq. (1). This effect is predominant in the interval where the degree of dissociation depends on the temperature (around x_d = 50 percent; Fig. 2).[13,15,17]

FIGURE 4. Thermal dissociation and ionization of SF_6. Particle densities at one atmosphere are plotted as functions of the temperature.[26]



Arc mediums in circuit breakers. To simplify understanding, let us first attempt to characterize an imaginary ideal arc medium.

In current-limiting and dc breakers the arc medium should cause a high electrical arc gradient.[5,6,11,22,23] This demands a high ionization energy and high thermal conductivity (small molecular mass and small collision cross section).

In the usual HV ac breakers, however, minimum energy release is the primary objective; therefore, low ionization energy and low thermal conductivity are wanted during the high-current interval of the current loop.

Near current zero, at any rate, a critical temperature interval must be reached, during which the gas changes from a good conductor to an excellent insulant. To avoid current chopping,[17,24,25] the insulating state must not be attained before the natural current zero. Hence, if the power supplied by small momentary currents has to be able to maintain the arc above the critical temperature, the losses have to be low, because to reduce the time needed for the change at current zero, the critical temperature interval should be small and should coincide with a prominent maximum of heat conductivity.

After deionization (electron–ion recombination) a high dielectric strength may be provided either by fast removal of all the ionizable particles (establishing an extra-high vacuum) or by a high density of strongly electronegative molecules, which are difficult to ionize.

Obviously, even these idealized requirements—which so far have not allowed for restrictions due to the actual gases—show serious contradictions (such as high and low heat conductivity, high and low ionization energy) and contradictory alternatives (high vacuum or high pressure), characterizing the fundamental problems.[6,9]

Nitrogen and air. The dissociation energy of N_2 is extremely high (W_d = 9.7 eV). At 7000°K it causes a maximum of heat conductivity nearly one order of magnitude above the usual value. This happens just in that temperature range in which electrical conductivity changes exponentially according to Eq. (2) and Fig. 2. Furthermore, N_2 molecules need more energy for ionization (W_i = 15.5 eV) than single N atoms in the hot-arc zone (W_i = 14.5 eV). According to the claims of the preceding paragraphs, these advantages contribute notably toward the effectiveness of the air-blast circuit breakers. But the effect is not perfect, since the thermal conductivity due to dissociation is no longer important below 5000°K, while at this value ionization is still many orders above its normal level and hence the gas is still electrically conducting instead of insulating. Fortunately, at lower temperatures the thermal conductivity of oxygen reaches its maximum (W_d = 5.1 eV); on the other hand, however, the ionization energy of O and O_2 is lower (W_i = 13.6 and 12.5 eV, respectively) and still lower (9.5 eV) for NO, which appears between 2000 and 6000°K.

Hydrogen (oil). Hydrogen, which is the ambient gas of oil-immersed arcs, generated by thermal decomposition of hydrocarbons, produces an outstanding classical heat conductivity (due to the extremely low values of atomic mass and diffusion coefficient) causing high electrical gradients and thin arc columns with low time constant, Eq. (3), but poor dissociative heat conduction (W_d = 4.4 eV only) and dielectric strength demanding higher pressures than air.

Sulfur hexafluoride. The outstanding features of the SF_6 arc medium were not understood for many years. Among these features are the extremely high ionization energy of the SF_6 molecule (W_i = 19.3 eV) and its electronegativity (electron attachment), both of which dominate at quite low temperature, when the proper arc has ceased.

The 22.4-eV bonding energy of the SF_6 molecule ($SF_6 \rightarrow S + 6F$) is 2.3 times that of N_2 and produces a corresponding dissociative heat conduction maximum; but fortunately, in contrast to the situation for air, Eq. (1), the maximum thermal conduction appears at rather low temperatures around 2000°K, because dissociation occurs in six energetically equal steps at six successive collisions, each of them needing only one sixth of the total amount of the bonding energy (Figs. 4 and 5).

In this range of temperature, neither SF_6 nor F, F_2, or any radical containing fluorine is ionizable. Hence, at about 2000°K, free electrons appear due only to ionization of free sulfur. But as soon as the sulfur recombines with fluorine below about 2000°K, the free electrons disappear instantaneously and so does the electrical conductivity; the previously conducting plasma changes into an excellent insulant in a manner that cannot be described by the usual thermal time constants [Eqs. (3) and (4)].[28]

At higher temperatures, however, due to the comparatively low ionization energy of the free sulfur (W_i = 10.4 eV) and the large masses of S and F atoms, a plasma of high electric and poor thermal conductivity (Figs. 6 and 7) yields extremely low values of both power dissipation and arc voltage.

In many respects, the advantages of SF_6 are impressive. However, there is no other superiority of SF_6 arcs at high currents, because (in contrast to H_2) the desired low power losses result in large arc cross sections and high time constants [Eq. (3)] if the arc is not constricted by forced convection.

Vacuum. The foregoing paragraphs have shown that various gases, as used successfully in circuit breakers, possess different advantages and disadvantages, even though only the physical properties around current zero have been discussed. However, a factor common to all breakers using these arc mediums is the need of a high gas pressure to cause a forced gas flow in a nozzle (to intensify convection cooling) and a high gas density (to increase the dielectric strength). On the other hand, extremely low gas pressures (10^{-6} torr) allow uninhibited

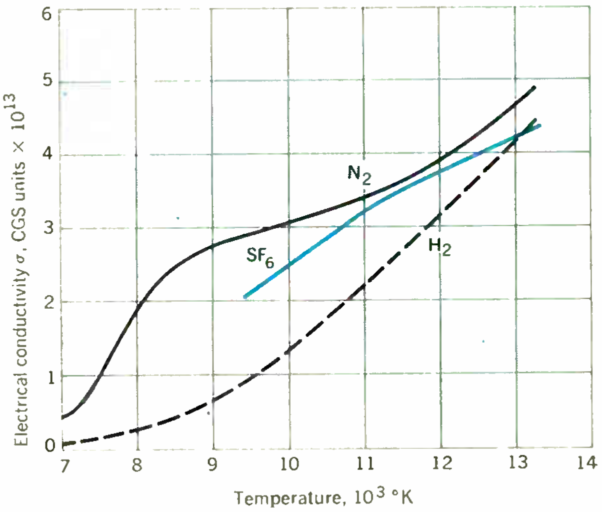


FIGURE 6. Electrical conductivity of SF_6, H_2, and N_2 as a function of temperature.[29]

FIGURE 7. Thermal conductivity of SF_6, H_2, and N_2 as a function of temperature.[19]

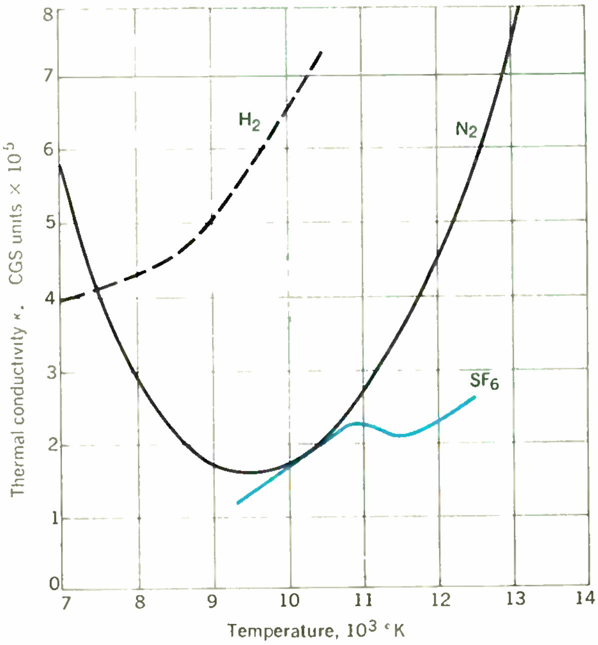
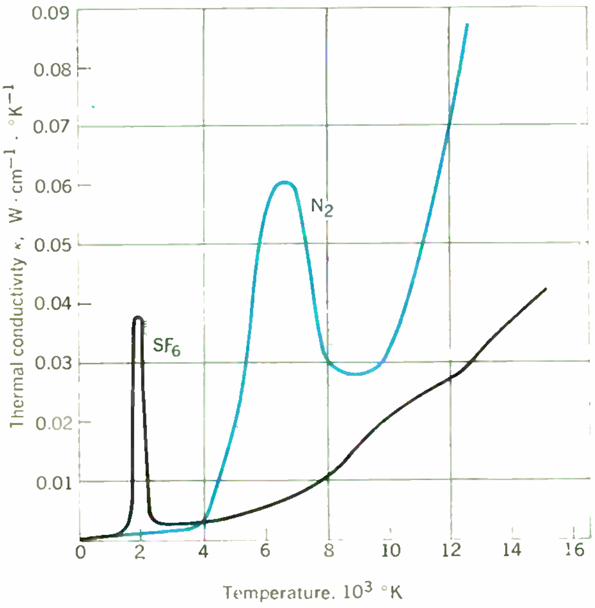


FIGURE 5. Thermal conductivity of SF_6 and N_2 as a function of temperature.[27]



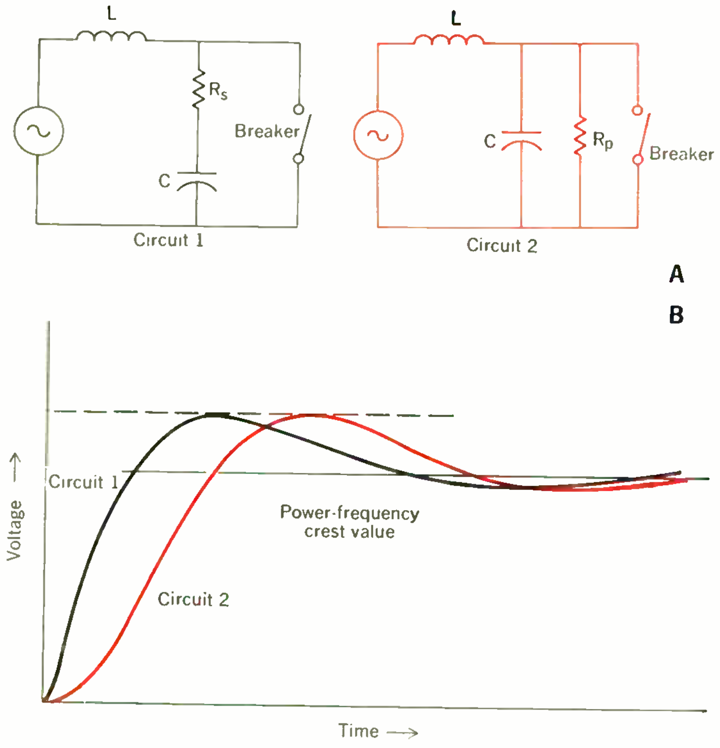
diffusion to the walls and cause a high dielectric strength by the lack of ionizable molecules.

Indeed, in the early days of electrical engineering, physicists expected to find an ideal switch if a pair of contacts were opened in vacuum, since if there is no gas, no arc should appear. However, because of the monomolecular gas layers adsorbed on the electrodes, the arc did appear and did not cease, and the designers were therefore disappointed. Later, when high-vacuum techniques enabled the manufacturers to avoid gas layers, the arc still appeared, because of vaporization of the electrode metal. But if, for electrodes of low vapor pressure, this metal-vapor arc became unstable at low current values, the designers were disappointed once more by the extremely high overvoltages caused by current chopping. Electrodes of high vapor pressure, on the other hand, cause high metal-vapor densities at high currents, and the arc loses the advantageous features of a vacuum arc: its voltage and power increase and it reignites after current zero at comparatively low voltages.[30-32] In intermediate ranges, however, such an arc acts almost ideally, with an extremely low arc voltage (about 20 volts, nearly independent of current and gap length) and fast recovery to a high, though limited, dielectric strength.[33]

Special duties of circuit breakers in service

Under different circumstances circuit breakers may be subjected to extremely different stresses.[17,34] First of all, the current varies from the no-load current of a transformer—a few amperes—up to the heaviest short-circuit currents, which may amount to a hundred kiloamperes. After current zero, the initial rate of rise and the peak value of the recovery voltage stressing the contact gap depend on the configuration of the network—not only on its natural frequency and damping, but also on the relative position of the resistances (in parallel or series with the main capacitance of the circuit); see Fig. 8. Fortunately, as a rule the natural frequency decreases with increasing current. Often it is not possible to describe the recovery voltage by a single frequency and amplitude factor (Fig. 9) and the circuit impedance itself may change within the first 10^{-4} to 10^{-3} second because open lines or cables connected to the bus bars at the breaker behave initially like resistances (having the values of their surge impedances), but later like capacitances. Whereas load currents are more or less ohmic, short-circuit currents are purely inductive and the currents of capacitor banks and unloaded lines are mainly capacitive.

FIGURE 8. A—Typical switching circuits, with resistances in series and in parallel with the circuit capacitance. B—Recovery-voltage curves.



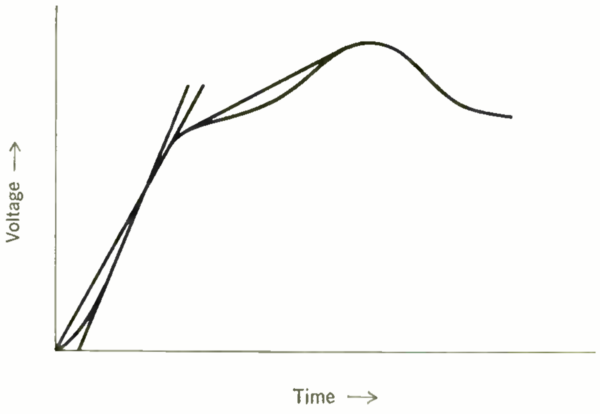
As mentioned at the beginning of this article, circuit breakers not only must interrupt but also must close the circuit. This may cause some trouble, especially if the breaker closes on a short circuit, because then the voltage breakdown that bridges the contact gap before the contacts touch produces a high-current arc, which melts the contacts before closure. Of course, this causes ideal welding conditions, but such a situation is not welcome since the breaker must be able to open the contacts again.[1]

Often automatic reclosing is required. In about 80 percent of the cases the cause of the short circuit no longer exists when the current has been interrupted, so as soon as the fault arc has ceased (i.e., after a fraction of a second) the breaker may close again before the two parts of the grid fall out of phase. About 20 percent of the short circuits persist, however, and immediately after reclosing, the breaker again has to interrupt the short-circuit current just made. This is a very severe duty, especially if there are extremely high currents requiring heavy contacts to be accelerated and decelerated in both directions within hundredths of seconds without bouncing, which would result in contact welding and wear.

When a dead EHV line is to be connected, high overvoltages may occur, especially if the line is loaded (because of the foregoing interruption), and often the breaker has to limit this overvoltage, even though it is not to blame.[36]

Besides making and breaking of no-load, load, and short-circuit currents, there are some rare but extremely

FIGURE 9. Recovery voltage after a short-circuit current. This voltage cannot be described in terms of a single frequency and amplitude factor.[35]



severe duties that often are required by users. Phase opposition may occur if the breaker recloses after a fairly long pause, during which the disconnected parts of the networks fell out of phase. In the worst case, the maximum fault current is only half the normal short-circuit current but the recovery voltage is double.

Evolving faults occur if the breaker, while interrupting the no-load current of a transformer, produces an overvoltage due to current chopping, which in turn induces a sparkover at the transformer terminals, followed by a short-circuit current, which immediately has to be interrupted by the breaker itself.

One mysterious problem encountered was the "short-line fault," occurring when a branch line is short-circuited about 1 km behind the breaker. Although the inductivity of this length of the line lowers the short-circuit current (compared with the breaker terminal fault immediately at the beginning of the line), the breaker is nevertheless much more severely stressed. At first this stressing seemed to be due solely to the very high rate of rise of the line oscillation after current zero (Fig. 10), but later it was found that the processes around current zero are much more complex and can be explained only in terms of the interaction between the arc and the circuit during an interval of about 100 μs around current zero.[20, 37]

Interaction between breaker and network

Since in many breakers the measured arc time constant ϑ is of the order of 1 μs, and arc recovery is furthermore retarded by heating of the arc by the current even during the last microseconds (for 100 kA rms the rate of decrease to zero is 50 A/μs), it is obviously not a question of why the arc reignites when stressed after current zero by a rate of rise of recovery voltage of about 1 kV/μs, but of why it does not do so every time. One reason is that the interaction of the varying arc voltage and the capacitance in parallel with the breaker influences the shape of the arc current.[17, 20, 24, 37] Figure 11 shows in schematic form how the current across the capacitance *C* aids the interruption by lowering the current immediately before zero. The same happens if a resistance *R* is parallel with the breaker—but not if it is connected in series with the capacitance, thus reducing the bypass current through *C* (Fig. 8).

A short line in series with both the breaker and the capacitance impedes the capacitive current and influences not only the recovery voltage curve, but also the state of the arc paths at current zero (Fig. 10).

The interaction between arc and circuit can be calculated around current zero, if the time constant and rate of power dissipation of the arc, as well as the initial conditions of the interaction interval, are known. These values can be measured and put into a mathematical arc model, which has to be combined with the differential equation of the circuit. With the aid of a computer it is possible to calculate the curves of current and voltage around current zero and to decide whether or not residual conductance and thermal reignition of the arc path occurs (even for the case of nonequilibrium mentioned previously).[20] This technique allows comparison of the behavior of a breaker in two circuits; perhaps in one of them it has been tested, whereas in the other it is to be used. However, at present it is not possible to calculate reignition introduced by dielectric breakdown of the arc path or the interaction be-

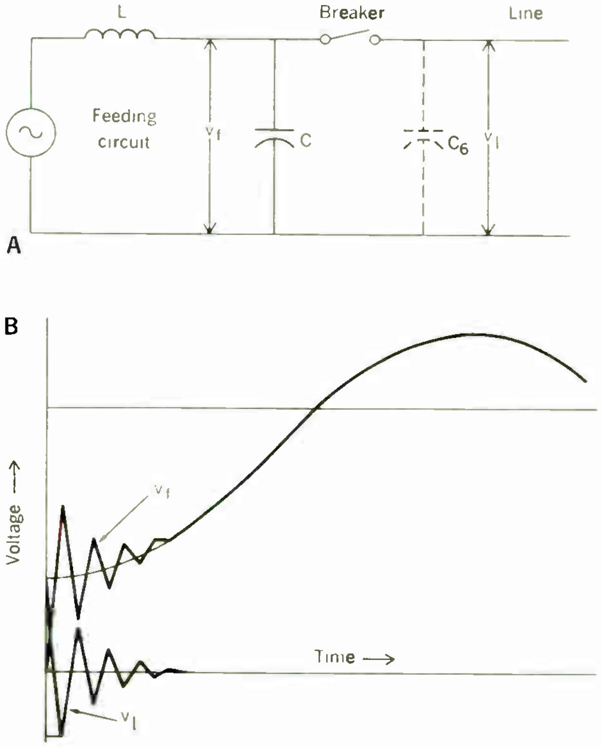
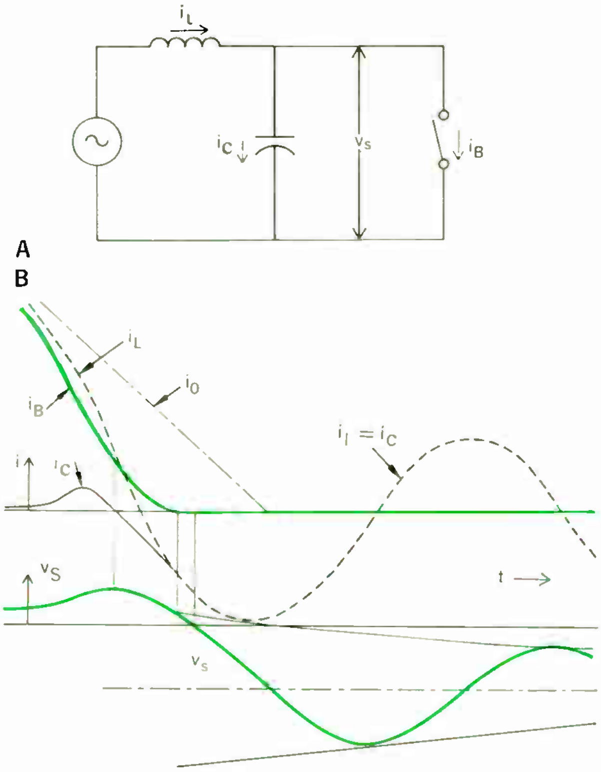


FIGURE 10. A—Short-line fault circuit. B—Recovery-voltage transient.

FIGURE 11. A—LC circuit. B—Current deformation resulting from the interaction of arc voltage with the inductance and capacitance of the circuit.[24] (i_0 = prospective current for an ideal breaker, without arc voltage.)



tween the gas flow and the arc in the breaker governing the initial conditions, time constant, and rate of power release.

Since the arc diameter shrinks continuously before current zero, the thermal time constant ϑ is not a constant for a time interval of an order greater than itself. Especially in SF_6, at the critical temperature of about 2000°K where the ionizable pure sulfur recombines with fluorine, the idea of a thermal time constant is no longer applicable.[28]

The nonlinearity of the equations and the multiplicity of the parameters are not conducive to calculating the interrupting capacity of a breaker, nor to valid testing by a small-scale model.

Because of the negligible arc voltage in vacuum breakers, no great assistance resulting from the interaction of these breakers with any circuit is to be expected. Also, for SF_6 breakers the interaction is less important than for air-blast breakers because of their lower arc voltage and time "constant."

Circuit-breaker testing

Although the arc theory makes it possible to reduce the number of tests and, within certain limits, to apply the results to other circuits, full-scale testing cannot be avoided altogether. Therefore, testing is much more important and expensive for the development of circuit breakers than for any other electric machine or equipment.

Field testing. The main problem in testing is to obtain the full short-circuit power needed. Although a few utilities have special field test stations assigned for short-circuit tests fed direct from the supply system, the full power of the system can never be made available for this purpose, because this power is needed to provide continuous operation for industry, traffic, offices, and households. Furthermore, even with reduced short-circuit power of the network, only a small number of tests is allowed, whereas the development of one single type of breaker needs thousands.

Finally, even the full short-circuit power of a system would not be enough to enable breakers to be developed for the short-circuit powers of future, as yet nonexistent, systems—for example, the world's first 1100-kV system—or for existing systems not yet interconnected or extended to full power.

Short-circuit generators. No single company—not even the entire electrical industry—is able to build power systems for circuit-breaker testing with ratings larger than those existing in the supply authorities' networks.

Therefore, an attempt was made to avoid drawing more power from the network than was available; instead, limited amounts of energy were drawn and stored over a long time (for example, 20 minutes), and then released across the test breaker during a few cycles of the service frequency.

For storage, use is made of the kinetic energy of the rotors of special "short-circuit generators" driven by asynchronous motors fed from the supply network. When the generator is at full speed it is disconnected from the supply, excited, and then short-circuited across the test breaker (in series with a safety breaker). Thus a 2000- or 3000-kW motor may drive a generator delivering 3000 or 4000 MVA of test power.

At the present time the biggest short-circuit generator produces about 2 GVA, and the biggest test station (connecting four generators in parallel) makes 5 GVA available, whereas modern circuit breakers have guaranteed breaking capacities up to 60 GVA. Therefore, indirect testing methods had to be invented to simulate more than was really available.

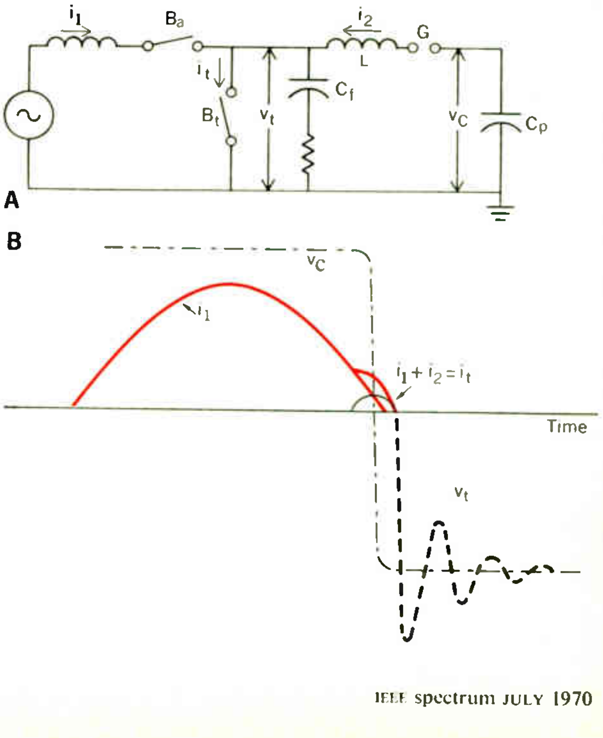
Single-phase testing. The first step was to test merely a single pole between two phases of the generator. For this the test power needed is just one half of that in the three-phase test (because the recovery voltage across the first interrupting pole is 1.5 times the phase voltage). On the other hand, the generator delivers only $1/\sqrt{3}$ of its three-phase testing power.

Unit testing. It was found that greater reductions could be made by testing only a part of a pole, if it consists of a number of equal interruptors connected in series. If uniform voltage distribution is assured, part of the breaker may be tested with full current, but with only the corresponding fraction of the supply voltage.

Synthetic testing. Unfortunately, even the power needed for the units of modern breakers is often not available, and users want at least half a pole to be tested. For this requirement we may take advantage of the fact that full test power never really appears. When the full current stresses the breaker before interruption, only the arc voltage is wanted; this is a small fraction of the supply voltage. After interruption, the total recovery voltage must appear, but only a small current is needed—just enough to feed perhaps a postarc current high enough to cause thermal reignition.

Hence two different circuits may be used: a high-current circuit with considerably reduced voltage and a high-voltage circuit with low current yield. The main difficulty arises during the critical 100 μs around current zero, when the interaction between breaker and circuit must be neither disturbed by changing the circuit or by the operation of an auxiliary breaker, nor falsified by the use of

FIGURE 12. Current-injection scheme, according to Weil-Dobke. A—Circuit. B—Waveforms.



wrong values of L and, above all, C.[20, 24, 27]

These and many other problems have been solved satisfactorily by so-called current-injection schemes (Fig. 12). The full test current is fed from the high-current alternator A. The auxiliary breaker B_a and the test breaker B_t open simultaneously at the beginning of a current loop. The high-voltage circuit is fed from a capacitor C_f loaded to the crest value of the full test voltage. About 1 ms before the prospective current zero, the gap G is triggered. Then C_f is discharged and recharged through the test breaker and the inductance L, which has nearly the same value as in a direct test circuit. The frequency of this discharge current i_d, governed by L and C_f only, is about ten times that of the supply frequency of the alternator, and its amplitude is about one tenth of that of the test current. Consequently, before zero, di/dt has the same value as in the direct test.

The arc in B_a ceases about 200 μs before that in B_t and therefore the test breaker is supplied from the high-voltage circuit only during the interaction period. After current zero the recovery voltage oscillates with a frequency determined by L and both C_f and C_p in series. The capacitance C_p parallel to B_t has exactly the same value as in the equivalent direct test. Nevertheless, for synthetic testing we need a great deal of knowledge about the physical phenomenons of circuit breaking, which have contributed to the development of a rather complex specialty.[38, 39]

REFERENCES

1. Holm, R., *Electrical Contacts*. Berlin: Springer, 1967.

2. Nöske, H., "Untersuchungen an kurzen Wechselstrom-Lichtbögen in Luft," *Z. Angew. Phys.*, vol. 10, pp. 327–336, 382–393, 1958.

3. Rieder, W., and Schneider, H., "On the reignition of ac arcs," *Proc. Internat'l. Conf. Ionization Phenomena in Gases*, vol. 4, pp. 1082–1087, 1959.

4. Erk, A., and Schmelzle, M., "Einfluss des Kontaktwerkstoffes auf das Löschverhalten kurzer Bögen," *VDE Fachber.*, vol. 25, pp. 75–80, 1968.

5. Rieder, W., and Sokob, P., "Probleme der Lichtbogendynamik," *Sci. Elec.*, vol. 5, pp. 93–112, 1959.

6. Rieder, W., "The importance of the arc in switchgear," *Brown Boveri Rev.*, vol. 49, pp. 398–403, 1962.

7. Troger, H., "Der neue 110-kV-Expansionsschalter für 16-2/3-Hz-Bahnnetze," *Siemens Z.*, vol. 33, pp. 780–786, 1959.

8. Kesselring, F., "Gesteuerte Synchronschalter," *ETZ*, vol. A 88, pp. 593–598, 1967.

9. Rieder, W., "Verfahren der Leistungsunterbrechung in Hochspannungsnetzen," *Sci. Elec.*, vol. 14, pp. 1–9, 1968.

10. Kriechbaum, K., "Hochspannungs-Schaltgeräte," *Tech. Mitt. AEG*, vol. 59, pp. 94–99, 1969.

11. Cobine, J. D., *Gaseous Conductors*. New York: Dover, 1958.

12. Hoyaux, M., *Arc Physics*. Berlin: Springer, 1968.

13. Rieder, W., *Plasma und Lichtbogen*. Braunschweig: Vieweg, 1967.

14. Burkhardt, G., "Zusammensetzung, innere Energie und spezifische Wärme der Luft bei hohen Temperaturen," *Z. Naturforsch.*, vol. 3a, pp. 603–607, 1948.

15. Finkelnburg, W., and Maecker, H., "Elektrische Bögen und thermisches Plasma," in *Encyclopedia of Physics*, vol. 22. Berlin, Springer, 1956.

16. Engel, A. v., and Steenbeck, M., *Elektrische Gasentladungen*. Berlin: Springer, 1932.

17. Trencham, H., *Circuit Breaking*. London: Butterworth, 1953.

18. Phillips, R. L., "Time constants for nonstationary arcs," *Z. Phys.*, vol. 211, pp. 113–131, 1968.

19. Cassie, A. M., and Mason, F. O., "Post-arc conductivity in gas-blast circuit-breakers," CIGRE Rept. 103, 1956.

20. Rieder, W., and Urbanek, J., "New aspects of current zero research on circuit-breaker reignition," CIGRE Rept. 107, 1966.

21. Lee, T. H., and Greenwood, A. N., "The mechanism of thermal breakdown of a high-temperature gas and its implication to circuit breakers," CIGRE Rept. 13-06, 1968.

22. Kind, D., Marx, E., Moellenhoff, K., and Salge, J., "Circuit breaker for HV dc transmission," CIGRE Rept. 43-05, 1968.

23. Oberdorfer, G., Bonin, E. v., Koetzold, B., and Thiel, H. G., "The circuit interruption in direct-current high-tension networks using commutating circuit breakers with eligible characteristics and absorbing elements of energy," CIGRE Rept. 13-10, 1970.

24. Rieder, W., and Passaquin, J., "On the decrease of the current to zero and the residual current in circuit breakers," CIGRE Rept. 105, 1960.

25. van den Heuvel, W. M. C., "Interruption of small inductive currents in ac circuits," Thesis, Technische Hogeschool, Eindhoven, 1966.

26. Frie, W., "Bestimmung der Gaszusammensetzung und der Materialfunktionen von SF$_6$," *Z. Phys.*, vol. 201, pp. 269–294, 1967.

27. Swarbrick, P., "Composition and properties of a sulfur hexafluoride arc plasma," *Brit. J. Appl. Phys.*, vol. 18, pp. 419–426, 1967.

28. Fukuda, S., Ushio, T., Ito, T., and Miyamoto, T., "Recent studies on arc behaviors around current zero in SF$_6$ gas," CIGRE Rept. 13-03, 1968.

29. Motschmann, H., "Uber die experimentelle Bestimmung der Wärmeleitfähigkeit und der elektrischen Leitfähigkeit von Wasserstoff und Schwefelhexafluorid im elektrischen Lichtbogen," *Z. Phys.*, vol. 191, pp. 10–23, 1966.

30. Reece, M. P., "The vacuum switch," *Proc. IEE (London)*, vol. 110, pp. 793–811, 1963.

31. Farrall, G. A., "Decay of residual plasma in a vacuum gap after forced extinction of a 250-ampere arc," *Proc. IEEE*, vol. 56, pp. 2137–2144, Dec. 1968.

32. Barkan, P., Lafferty, J. M., Lee, T. H., and Talento, J. L., "Development of contact materials for vacuum interrupters," *IEEE Trans. Power Apparatus and Systems*, vol. PAS-89, to be published.

33. Charbonnier, F. M., Bennette, C. J., and Swanson, L. W., "Electrical breakdown between metal electrodes in high vacuum," *J. Appl. Phys.*, vol. 38, pp. 627–640, 1967.

34. Ruoss, E., "Test methods and facilities in the high-power testing station," *Brown Boveri Rev.*, vol. 55, pp. 714–726, 1968.

35. Baltensperger, P., *et al.*, "Transient recovery voltage in high-voltage networks," CIGRE Rept. 13-10, 1968.

36. Baltensperger, P., and Ruoss, E., "Switching overvoltages in EHV und UHV networks. Methods of limitation," CIGRE Rept. 13-14, 1970.

37. Eidinger, A., and Rieder, W., "Short line fault problems," CIGRE Rept. 103, 1962.

37. Hochrainer, A., "Synthetic testing for high power circuit-breakers," *Electra*, pp. 51–62, Oct. 1968.

39. Grycz, B., "The equivalence of the fundamental schemes for synthetic and direct testing from the point of view of interaction between the arc and the circuit," CIGRE Rept. 13-08, 1970.

Werner Rieder was born and educated in Vienna, Austria. He received the Ph.D. degree in physics from the University of Vienna in 1949, and then worked as an assistant on metal physics and later on switchgear at the Institute of Technology of Vienna until 1956. From 1951 to 1956 he was also with ELIN Ltd. as research physicist. In 1956 he joined Brown, Boveri & Co., Ltd., at Baden, Switzerland, where he has been involved with switchgear physics and its application to the development and testing of circuit breakers. Since 1957 he has been teaching plasma physics at the Institute of Technology of Vienna, where he was awarded the title of professor in 1963. Since 1961 he has also been lecturing on switchgear at the Institute of Technology of Graz (Austria) and since 1958 on arc physics and switchgear at the Swiss Federal Institute of Technology, Zurich.

He has written a number of articles on arc physics, electric contacts, and switchgear problems, as well as a book on plasma and the arc. Dr. Rieder is an active member of the CIGRE Working Group on Synthetic Testing of Circuit Breakers.

Reflection and crosstalk in logic circuit interconnections

Modern transistor–transistor and emitter-coupled logic circuits have resulted in speeds of such proportions that hitherto neglected interconnections are beginning to require special attention

John A. DeFalco **Honeywell Inc.**

Digital logic circuits are now available and are being used with delay times that are comparable to the delays of interconnections used in packaging these circuits. At high speeds, however, such interconnections no longer behave as simple short circuits, but take on the appearance of transmission lines. Unless transmission lines are terminated properly, "reflections" can develop that might be of sufficient magnitude to produce false logic levels or exceed maximum circuit voltage specifications. One may choose to solve the problem by increasing the density of the system. This, however, introduces the problem of "crosstalk." The present article describes several analytical techniques for predicting the kinds of reflections and crosstalk that are typically seen in digital systems, thus enabling the engineer to determine in advance whether or not such "interconnection noise" will result, how bad it will be, and what the typical interconnection limitations are for circuits of various speeds.

The high speeds of today's integrated circuitry have given the interconnections of digital logic circuits characteristics usually associated with transmission lines. The most common method for eliminating the reflections that occur in such lines involves a procedure called termination. Terminating these transmission lines is not always practical, however, since termination networks take up space and increase system costs. The most common termination—a resistor—requires current to be supplied at least part of the time, often increasing the power supply and logic circuit requirements.

A possible solution to the problem of reflections is to decrease the length of the interconnections by increasing the system density. The trend toward greater density, however, brings about another problem—that of crosstalk between the various conductors of the system. This coupling, which often exists between two adjacent transmission lines, may be strong enough for signals to appear on both lines when they are only desired on one. The problem of crosstalk is present even if lines are terminated, and since these lines are carrying digital information, this unwanted coupling introduces false information to the system. The general category of extraneous voltages and currents due to reflections and crosstalk is commonly called interconnection noise.

Hence, in designing a high-speed digital system, both the circuits and the interconnections must be considered or system performance may be impaired.

Transmission-line basics

For the purpose of calculating reflections and crosstalk in digital systems, we shall consider Fig. 1.

A transmission line can be characterized in terms of its delay T_D (the time a waveform needs to traverse a unit length of transmission line), the characteristic impedance Z_0 (the impedance that would be presented by the line if it were infinitely long), and α (the line attenuation factor, which in most cases is neglected).

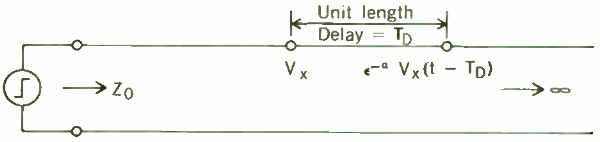


FIGURE 1. Drawing of a generalized infinite two-conductor transmission line.

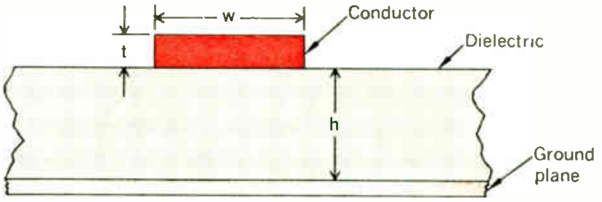
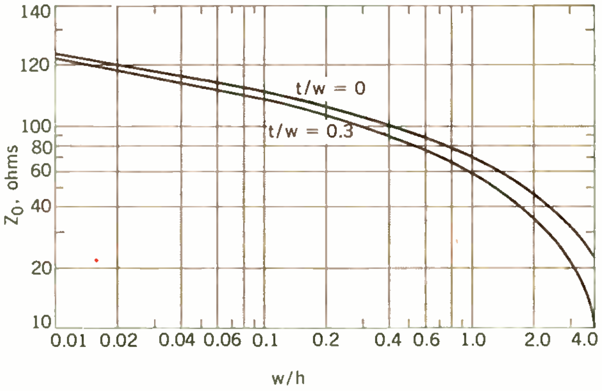


FIGURE 2. Typical microstrip transmission line.

FIGURE 3. Impedance of a microstrip line for a relative dielectric constant of 5.



When α is negligibly small, T_D and Z_0 can be defined in terms of L (the inductance of the line per unit length) and C (the capacitance of the line per unit length) as

$$T_D = \sqrt{LC} = \frac{1}{\mu} \quad \text{and} \quad Z_0 = \sqrt{\frac{L}{C}}$$

where L and C are functions of the width, thickness, and spacing of the conductors and the dielectric separating them; and μ is the wave propagation velocity.[1,2]

The microstrip transmission line

The most common type of transmission line for interconnecting high-speed circuits is the microstrip line (Fig. 2), which can be batch-fabricated in the form of printed circuit boards and thick- or thin-film integrated circuits. However, to be able to use the microstrip line for circuit interconnection, Z_0, T_D, and the various types of reflections and crosstalk that might occur must be readily determinable.

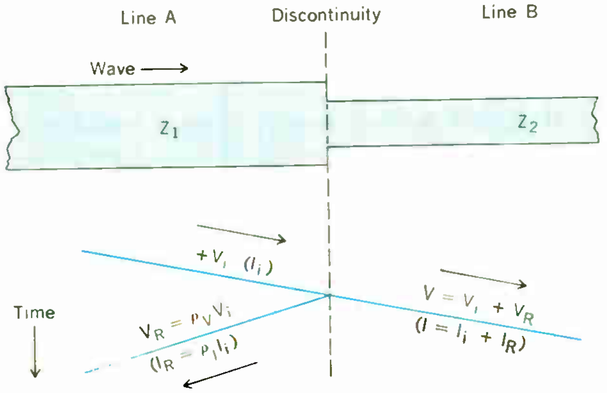
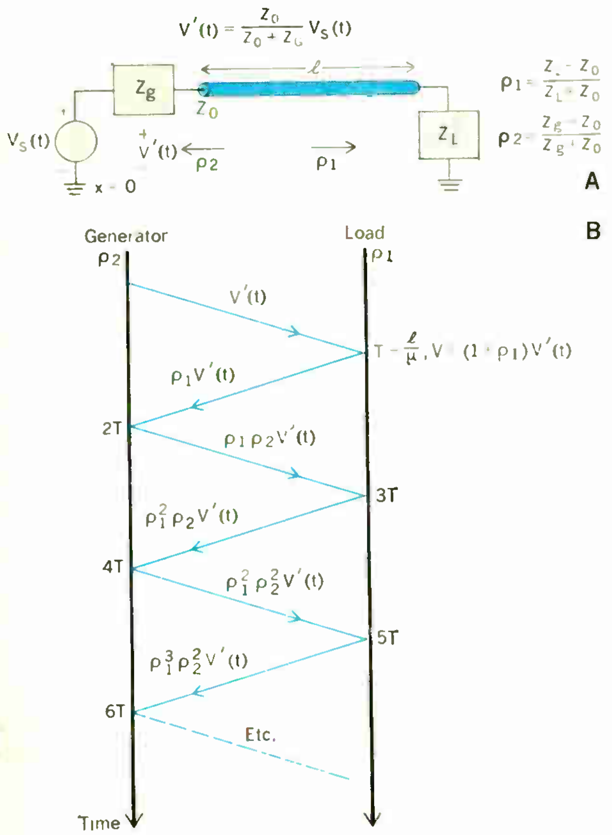


FIGURE 4. An example of transmission-line reflections at a discontinuity.

FIGURE 5. Lattice diagram for the purpose of computing transmission-line reflections.



Such calculations are complicated because the microstrip line has a compound dielectric (air and a board that is usually a fiberglass epoxy or ceramic material). This compound dielectric causes the microstrip line to be faster than if it were completely buried in one dielectric material. The value of the effective dielectric constant is[3]

$$e_{re} = 0.475 e_r + 0.67 \qquad (1)$$

where e_{re} is the effective dielectric constant and e_r is the relative dielectric constant for the board material. Equation (1) is useful for specific values of e_r; that is,

$$1 \leq e_r \leq 15$$

The delay of this type of line is found to be

$$T_D = 1.017\sqrt{e_{re}} = 1.78 \text{ ns/ft } (0.06 \text{ ns/cm}) \qquad (2)$$

for a fiberglass epoxy board with $e_r = 5.0$.

The characteristic impedance of a microstrip line is[3,4]

$$Z_0 = \frac{60}{\sqrt{e_{re}}} \log_e \frac{5.98h}{0.8w + t} \text{ ohms} \qquad (3)$$

where $\log_e$ is the natural logarithm; h, w, and t are indicated in Fig. 2. Equations (1) and (3) are derived using field-mapping techniques. They have been checked against time-domain reflectometry measurements[5] and found to be accurate over the range of practical system interconnection geometries. Figure 3 is a plot of Eq. (3) for $e_r = 5.0$.

Reflections on transmission lines

Signal reflections on transmission lines are a major source of noise in digital systems. They occur whenever a transmission-line signal encounters a discontinuity—an impedance different from that of the original line. The discontinuity can be another transmission line, a circuit, or a load device.

The general case is shown in Fig. 4, where an incident voltage (V_i) and incident current (I_i) waveform traveling along a transmission line with impedance Z_1 encounter a discontinuity of impedance Z_2. The reflected voltage originates at the plane of discontinuity and travels toward the generator. In Fig. 4, the incident voltage on line B is $V_i + V_R$ of line A.

The reflected voltage $V_R = \rho_V V_i$, where ρ_V, the voltage reflection coefficient, is defined as

$$\rho_V = \frac{Z_2 - Z_1}{Z_2 + Z_1}$$

The reflected current $I_R = \rho_I I_i$, where ρ_I, the current reflection coefficient, is

$$\rho_I = \frac{Z_1 - Z_2}{Z_1 + Z_2} = -\rho_V$$

If input signals may be considered voltage steps, the problem of transmission-line reflections can be handled very simply by using a lattice diagram.[6] Such a diagram (Fig. 5) is constructed by selecting some convenient scale for the horizontal axis, which represents electrical length. A vertical line is then drawn at each end of the horizontal axis; the left-hand line represents the discontinuity at the generator and the right-hand line represents the discontinuity at the load. Hence, the vertical axis becomes the time scale, and we consider the wave to travel diagonally down the plot from the top. The voltage that starts down the line $V'(t)$ is the result of a voltage division that occurs between the source impedance and the line impedance. It proceeds down the line to the load, where it is reflected toward the generator and so on down the diagram. The voltage at any time is the algebraic sum of all the individual reflected voltages up to that time. For example, in Fig. 5 the voltage just after $t = 3T$ is $V(t) = V'(t)(1 + \rho_1 + \rho_1\rho_2 + \rho_1{}^2\rho_2)$.

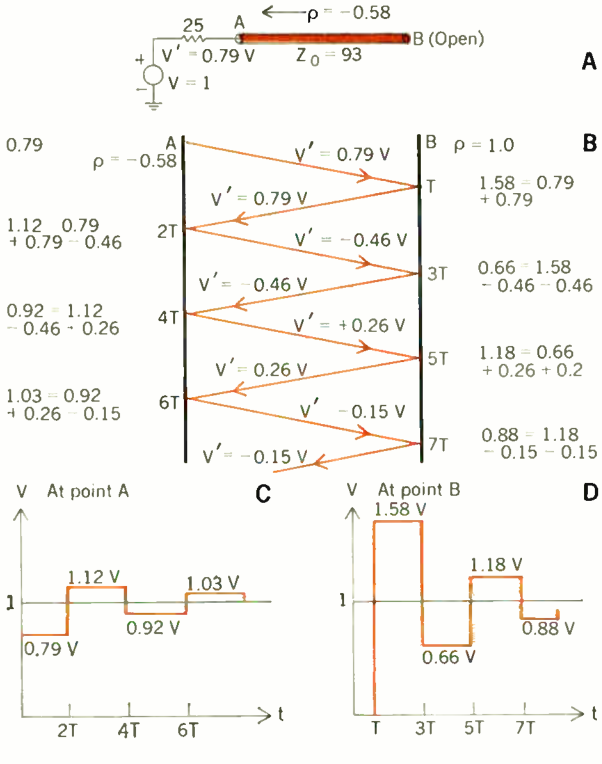
As a specific example of this technique, consider the case depicted in Fig. 6. The initial voltage division between the source and the line is computed to be 0.79 volt. This voltage travels down the line to the load, where it is completely reflected ($\rho_1 = 1.0$). In Fig. 6(B), the voltage travels back to the generator where it is again reflected. The process continues as long as the additional reflections are of interest to the engineer. The computed waveforms for both ends of the line are presented in Figs. 6(C) and 6(D).

The lattice diagram is a simple way of determining transmission-line waveforms for cases where the source and load terminations are linear. The lattice diagram may be expressed mathematically in the following form[7]:

$$V(x,s) = V'(s)\left[e^{\frac{-sx}{\mu}} + \rho_1 e^{\frac{-s(2l-x)}{\mu}} + \rho_1\rho_2 e^{\frac{-s(2l+x)}{\mu}} + \rho_1{}^2\rho_2 e^{\frac{-s(4l-x)}{\mu}} + \rho_1{}^2\rho_2{}^2 e^{\frac{-s(4l+x)}{\mu}} + \cdots\right] \quad (4)$$

where x is the distance at any point on the line and l is the physical length of the line; s is the Laplace transform variable and therefore $V(x,s)$ is the Laplace transform of the voltage at any point x. $V'(s)$ is the Laplace transform of $V'(t)$ and ρ_1 and ρ_2 are the reflection coefficients from Fig. 5. This form is particularly useful if the input waveform is not a step voltage or the signal rise time is greater than twice the line delay.

FIGURE 6. Lattice diagram and computed waveforms for a typical reflection situation.



Reflections from nonlinear terminations

Nonlinearities of the type $I = f(V)$ are very common, especially when transmission lines are driving and being driven by integrated logic circuits. In these cases, a graphical technique[8] that can also be used with linear impedances is quite useful. The technique consists of plotting load lines corresponding to the slopes of the source impedance, load impedance, and transmission line on a current vs. voltage diagram. To illustrate this technique, the problem of Fig. 6 with linear impedances will be worked out. The technique will then be extended to the case where nonlinear loads are present.

In Fig. 7, the first line of slope $-1/R_s$ (R_s is the source impedance, which is 25 ohms in this example) is drawn from a point on the x-axis corresponding to the input voltage source value (1 volt). Another line corresponding to the transmission-line impedance is drawn from the origin with a slope $1/Z_0$. The intersection of the source impedance line and the transmission-line impedance gives the initial voltage and current starting down the line after the voltage division between the source and line impedances (point 0 in Fig. 7; in Fig. 6(C), this voltage is at point A for $t < 2T$). From point 0, another line of slope $-1/Z_0$ is drawn until it intersects a line drawn from the origin with a slope $1/Z_L$, where Z_L is the impedance at the far end of the line. In this example, Z_L is an open circuit, so the slope of the line is zero, coinciding with the x-axis. The intersection of these lines corresponds to the current and voltage at the end of the line (point 1) ($T < t < 3T$).

The signal then travels from the load back to the source and corresponding to this we draw a line from the load (point 1) back to the generator source impedance with a slope equal to the reciprocal line impedance, i.e., a line of slope $1/Z_0$ from point 1 to point 2 at the $-1/R_s$ line. This point represents the voltage and current at the source end of the line for $2T < t < 4T$. The procedure is continued by drawing lines with slopes $1/Z_0$ and $-1/Z_0$ from each intersection with the source impedance line to the load impedance and back again.

If the intersections are labeled as shown in Fig. 7, then the even-numbered points are the currents and voltages at the beginning of the line for $2T$ increments of time (T is the one-way line delay). In other words, point 0 represents the current and voltage at the beginning of the line for $0 < t < 2T$, point 2 corresponds to $2T < t < 4T$, and so on. The odd-numbered points correspond to time increments of $2T$ at the end of the line. Point 1 corresponds to $T < t < 3T$, and point 3 represents the current and voltage at the load for $3T < t < 5T$. The results obtained in Fig. 7 are the same as in Figs. 6(C) and 6(D).

The technique is illustrated for nonlinear terminations in Fig. 8. Figure 8(A) shows a NAND gate (transistor–transistor logic—T²L—in this case) driving a long 90-ohm transmission line loaded by a similar T²L circuit. The worst case (causing ringing) for this type of circuit occurs when it is going from a "one" state to a "zero" state, because the mismatch between the circuit output impedance and the line is greater than the case when the output is going high. The technique is the same as shown in Fig. 7, except that certain modifications must be made because the signal is initially at a positive "one" level and going negative.

The starting "one" level is treated as the origin and a line representing the line impedance is drawn with a slope $-1/Z_0$. The intersection of this line with the output impedance line (the output and input impedance can be determined using a curve tracer) at point 1 (Fig. 8) represents the voltage at the input to the transmission line. From this point, a line is drawn with a slope $1/Z_0$ (the negative of the first $-1/Z_0$ line). Its intersection with the load impedance characteristic (point 2) is the magnitude of the voltage at the load just after the first reflection. Additional reflections may be determined by continuing the process until it is no longer possible to draw the transmission-line impedance lines (point 3 in this example, but much later in other examples). A photograph [Fig. 8(C)] of actual transmission lines and circuits shows this reflection to be approximately the 1.3 volts indicated by Fig. 8(B).

Crosstalk between transmission lines

Digital system noise is often caused by crosstalk, which is the coupling of an unwanted signal from the signal-carrying line to another nearby line. This coupling is caused by the mutual capacitance (C_M) and mutual inductance (L_M) that exist because of the proximity of the two transmission lines. To predict crosstalk, the differential equations for the coupling between two lines must be developed and then solved for various boundary conditions determined by the line terminations.[9] The resulting crosstalk waveforms for many common line terminations appear in Fig. 9. Note that although six different terminations are shown, the terms

$$\gamma(k+1)V, \qquad \gamma(k-1)\frac{l}{\mu}\frac{V}{a}$$

$$\text{where } \gamma = \frac{C_M}{C},\ k = \frac{L_M}{L} \times \frac{C}{C_M},\ \text{and } \mu^2 = \frac{1}{LC}$$

or some fraction of them appear in each waveform. V is the original signal amplitude and a is the rise time of the signal of the line that should be carrying it. This waveshape is shown in Fig. 9(A), where l is the physical length of the line.

Since V, a, and l are known, we must still find $\gamma(k+1)$ and $\gamma(k-1)/\mu$, the crosstalk constants, to predict the crosstalk between two transmission lines. Figures 10 and 11 are curves of $\gamma(k+1)/\mu$ for microstrip lines ($\epsilon_r = 5.0$) as functions of the spacing between the lines and the height of the lines above the ground plane.[10] The crosstalk waveforms shown in Fig. 9 assume that the signal rise time is less than twice the electrical delay of the line ($a < 2l/\mu$).

To use these illustrations when the rise time is greater than twice the electrical delay, consider Fig. 9(A). Note that if $a > 2l/\mu$, the crosstalk fails to achieve maximum amplitude; the amplitude that is reached is proportional to the ratio $(2l/\mu)/a$ and the crosstalk amplitudes shown in Figs. 9, 10, and 11 should be multiplied by that factor. In other words, all terms that were $\gamma(k+1)V$ become $\gamma(k+1)(2l/\mu)/a$.

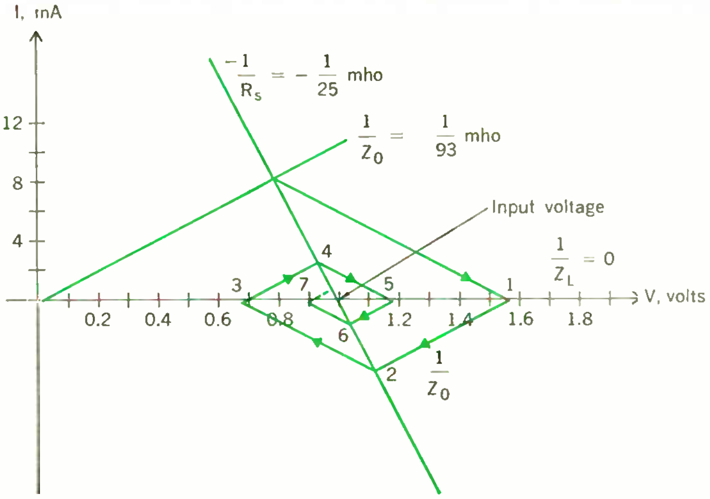
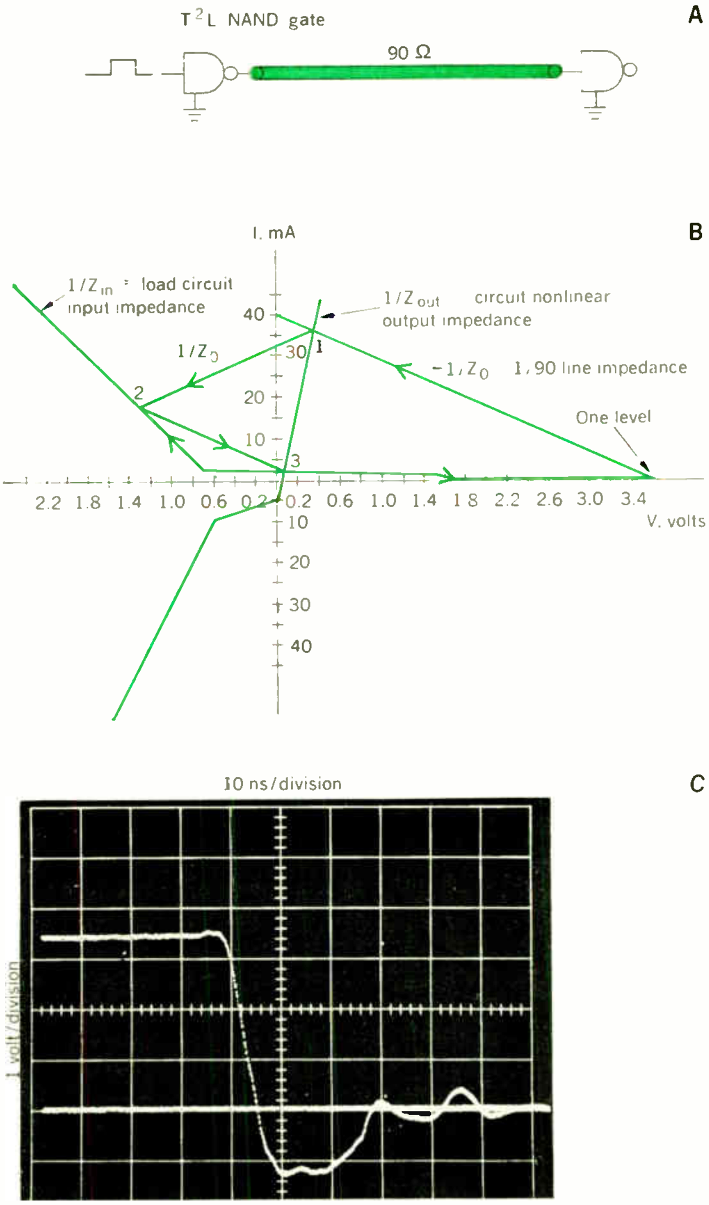


FIGURE 7. Graphical technique for computing reflections from linear and nonlinear terminations.

FIGURE 8. Reflection diagram for T²L driving 90-ohm line and a photo of the actual circuit response.



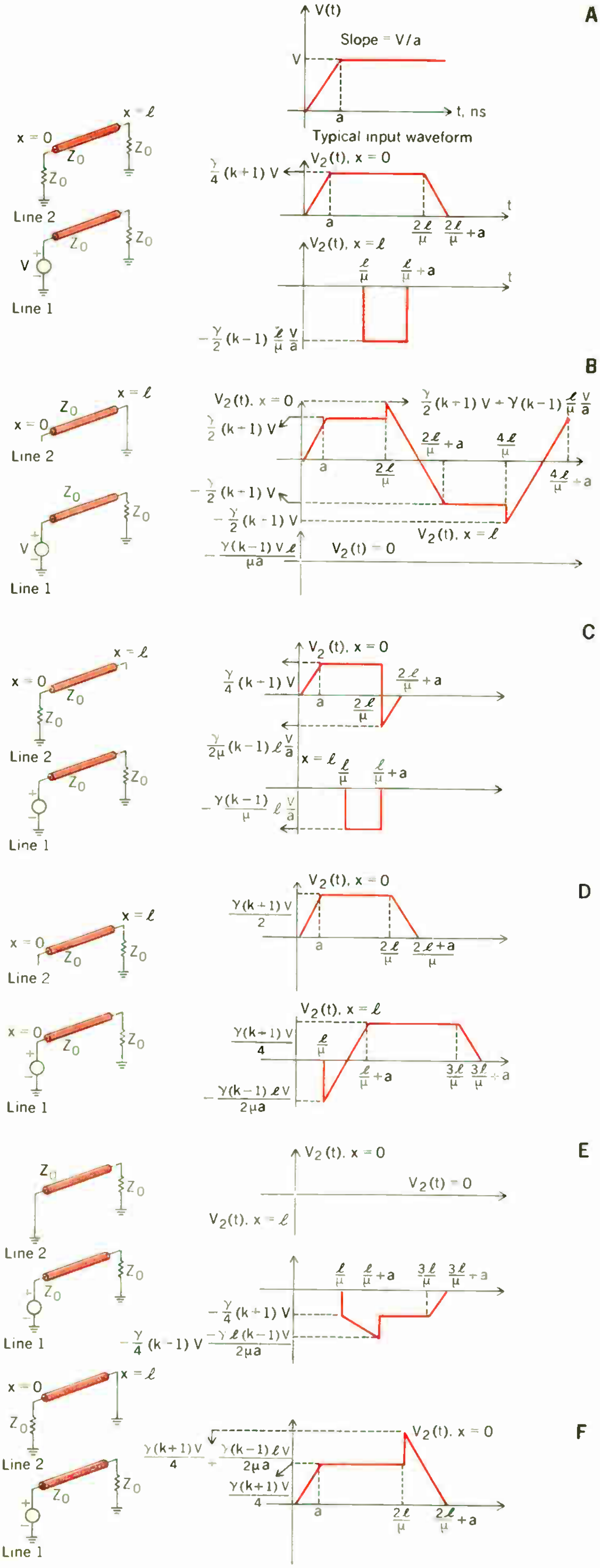
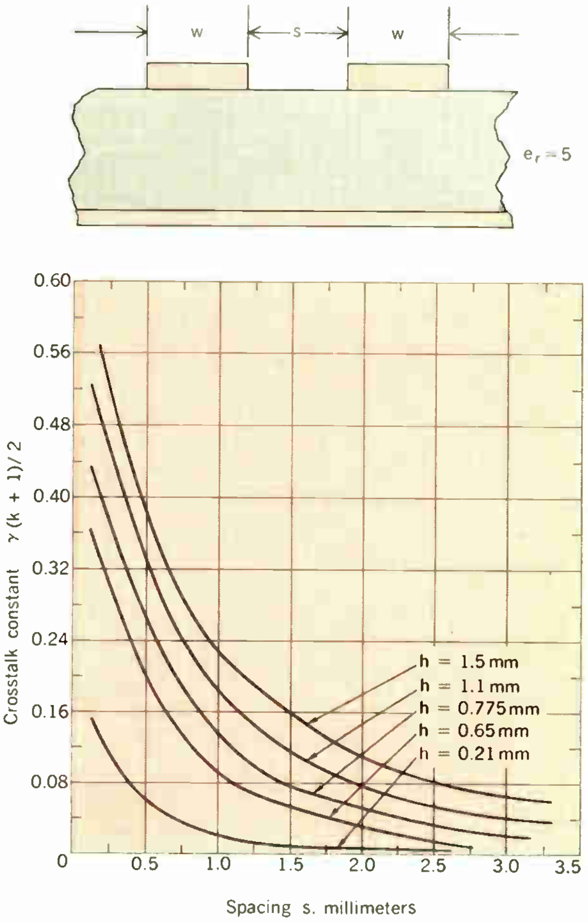


FIGURE 9. A—Case 1: Line 2 terminated at both ends and resulting crosstalk. B—Case 2: Line 2 open at near end and short-circuited at far end and resulting crosstalk. C—Case 3: Line 2 terminated at near end and open at far end and resulting crosstalk. D—Case 4: Line 2 open at near end and terminated at far end and resulting crosstalk. E—Case 5: Line 2 short-circuited at near end and terminated at far end and resulting crosstalk. F—Case 6: Line 2 terminated at near end and short-circuited at far end and resulting crosstalk.

FIGURE 10. Line spacing vs. crosstalk constant.



To illustrate the use of Figs. 9 to 11 and Eqs. (2) and (3), a board with two 38-cm microstrip lines was fabricated on glass epoxy. The width of each line and the spacing between lines was 0.3 mm. The height from the ground plane was 0.65 mm and the lines were 0.07 mm thick. Using Eqs. (2) and (3), Z_0 and T_D were calculated to be 84.23 ohms and 1.78 ns/ft (0.06 ns/cm), respectively. The results from the time-domain reflectometer measurements were

$$Z_0 = 85 \text{ ohms} \qquad T_D = 1.8 \text{ ns/ft } (0.06 \text{ ns/cm})$$

From Figs. 10 and 11

$$\frac{\gamma}{2}(k+1) = 0.3 \qquad \frac{\gamma}{\mu}(k-1) = 0.17$$

The crosstalk waveforms were calculated assuming a 5-volt, 1-ns step ($V/a = 5$ and $l = 1.25$ ft):

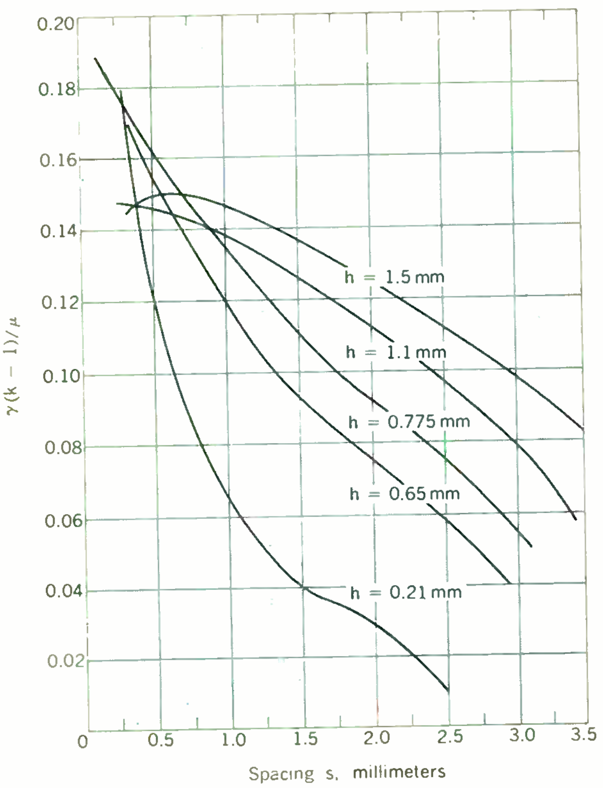


FIGURE 11. Line spacing vs. crosstalk constant. The y-axis (ordinate) is expressed in ns/ft.

$$\frac{\gamma}{2}(k+1)\frac{V}{a} = 1.5 \text{ volts}$$

$$\frac{\gamma(k-1)lV}{2\mu a} = 0.52 \text{ volt}$$

The calculated and experimental waveforms are shown in Fig. 12. For several cases, the calculated waveform shows some zero-rise-time edges. In reality, the observed trace on an oscilloscope will be slower due to scope-loading effects. Nevertheless, the agreement between predicted and actual results is very good.

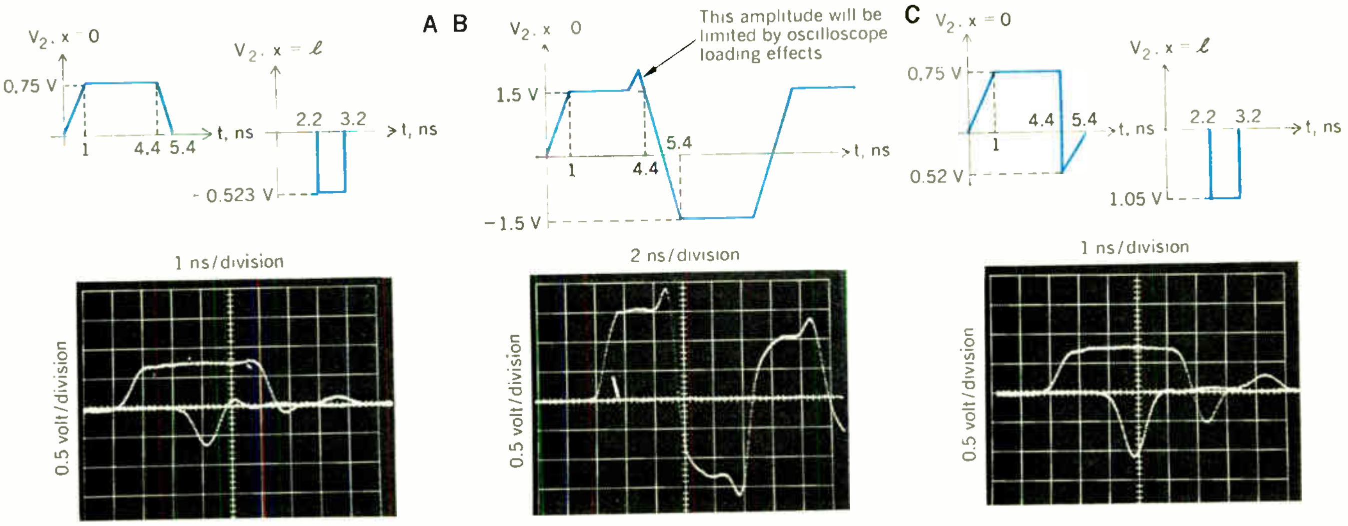
Applications to the interconnection of logic circuits

The design of an interconnection scheme for high-speed circuits is dependent on the effects interconnections have on the circuit performance. The circuit noise immunity and maximum input voltage level are of particular concern. The circuit noise immunity is a measure of the insensitivity of a circuit to spurious or undesirable signals, whereas the maximum input level is related to the circuit and device design and fabrication. In almost all cases, one of these circuit properties will impose some restrictions on the physical and electrical properties of the interconnection scheme. Various design routines can be established using the previously presented techniques to restrict the transmission-line parameters so that the crosstalk and reflections generated are within the circuit properties described in this article.

T^2L and emitter-coupled logic (ECL) have emerged as the two most widely used high-speed logic families. Both circuit types are being made with delays and rise times of 5 ns or less. Table I lists some of the essential circuit information necessary to place restrictions on the interconnection system. In general, both reflections and crosstalk can be minimized if the one-way delay of the line is less than half the circuit rise time. Since this is not always possible, both cases should be considered.

The calculation of interconnection limitations proceeds in a straightforward manner from the previously discussed concepts. Hence, limitations based on reflections follow from the lattice diagram (Fig. 5), Eq. (4), or the reflection diagrams (Figs. 7 and 8). The limitations based on crosstalk can be calculated using the noise immunity as a value of crosstalk and determining the various line properties from Figs. 9, 10, and 11. Although the limitations that result will vary somewhat depending on the exact logic family chosen and the as-

FIGURE 12. A—Calculated crosstalk waveforms for line terminated at both ends and photograph of actual waveforms (input voltage is a 5-volt, 1-ns pulse. B—Calculated and actual waveforms for Case 2 (line 2 open at x = 0, terminated at x = l. C—Case 3: Line 2 terminated at x = 0, open at x = l.



I. Typical circuit characteristics for use in designing an interconnection system

	Typical Rise Time, ns	Signal Swing	Output Impedance, ohms "Zero" Level	Output Impedance, ohms "One" Level	Input Impedance, ohms Negative Voltages	Input Impedance, ohms Positive Voltages	Worst-Case Noise Immunity	Limiting-Voltage Condition, V_{max}
T^{2}L								
Gold doped	5	3.5 volts	5	60	40–550	$\approx \infty$	450 mV	$V_{in} = -2$ volts*
Schottky diode	2	3.5 volts	5	60	40–550	$\approx \infty$	400 mV	
ECL								
Second generation	4	800 mV	10	10	$\approx \infty$	$\approx \infty$	175 mV	25% overshoot†
Third generation	1	800 mV	5	5	$\approx \infty$	$\approx \infty$	175 mV	

* Voltages more negative than −2 volts may result in a voltage breakdown in the T^{2}L input transistor.
† An overshoot of more than 25 percent may saturate an input transistor.

II. Typical limiting interconnection values for 85-ohm lines (based on information in Table I)*

		T^{2}L		ECL	
		2 ns	5 ns	1 ns	4 ns
Reflections		$V_{r(max)}$† may be kept to V_{max} from Table I if the line lengths indicated below are used			
$a > 2\frac{l}{\mu}$		$L_{max} = 16$ cm‡	$L_{max} = 40$ cm‡	$L_{max} = 5.5$ cm	$L_{max} = 21$ cm
Crosstalk	(A)	Unterminated—worst case—i.e., pickup line with a low impedance at the far end and a high impedance at the near end [Fig. 9(B)].			
		For any $S > S_{min}$; Crosstalk < noise immunity			
$a < 2\frac{l}{\mu}$		$S_{min} = 0.85$ mm	$S_{min} = 0.78$ mm	$S_{min} = 0.43$ mm	$S_{min} = 0.43$ mm
		Values of $S < S_{min}$ may be used if the line lengths are restricted as indicated below.			
		$S = 0.25$ mm	$S = 0.25$ mm	$S = 0.25$ mm	$S = 0.25$ mm
$a > 2\frac{l}{\mu}$		$L_{max} = 6.7$ cm	$L_{max} = 18.6$ cm	$L_{max} = 6.4$ cm	$L_{max} = 25.3$ cm
		$S = 0.75$ mm $L_{max} = 1.5$ cm	$S = 0.75$ mm $L_{max} = 43$ cm	$S = 0.38$ mm $L_{max} = 8$ cm	$S = 0.38$ mm $L_{max} = 31$ cm
	(B)	Pickup line terminated in Z_0 at both ends, Fig. 9(A).			
$a < 2\frac{l}{\mu}$		$S_{min} = 0.53$ mm	$S_{min} = 0.6$ mm	$S_{min} \approx 0.05$ mm	$S_{min} \approx 0.05$ mm
$a > 2\frac{l}{\mu}$		$S = 0.25$ mm $L_{max} = 13$ cm	$S = 0.25$ mm $L_{max} = 37$ cm	$S < 0.05$ mm is beyond the practical design range	

* $h = 0.65$ mm, $w = 0.33$ mm, $t = 0.07$ mm, $e_r = 5.0$.
† $V_{r(max)}$ = maximum reflection voltage.
‡ Assuming worst-case loading by a T^{2}L circuit without clamp diodes.

sumptions of the designer, the values given in Table II may be considered typical.

REFERENCES

1. Potter, J. L., and Fich, S. J., *The Theory of Networks and Lines.* Englewood Cliffs, N.J.: Prentice-Hall, 1963.
2. Matick, R. E., *Transmission Lines for Digital and Communications Networks.* New York: McGraw-Hill, 1969.
3. Kaupp, H. R., "Characteristics of microstrip transmission lines," *IEEE Trans. Electronic Computers,* vol. EC-16, Apr. 1967.
4. Springfield, W. K., "Multilayer printed circuitry in computer applications," presented at the Fall Meeting on Printed Circuits, Chicago, Ill., Oct. 1964.
5. "Time domain reflectometry," Appl. Note 62, Hewlett-Packard Company, 1964.
6. Millman, T., *Pulse, Digital and Switching Waveforms.* New York: McGraw-Hill, 1965, p. 86.
7. Potter, *op. cit.*, pp. 401–403.
8. Singleton, R. S., "No need to juggle equations," *Electronics,* p. 63, Oct. 28, 1968.
9. Jarvis, D. B., "Interconnection effects on high speed logic," *IEEE Trans. Electronic Computers,* vol. EC-12, Oct. 1963.
10. Kaupp, H. R., "Pulse crosstalk between microstrip transmission lines," *7th Internat'l Electronic Packaging Symp. Record,* Wescon '66, IECP 2/6, Aug. 1966.

John A. DeFalco (M) was born in Elmira, N.Y., received the B.S.E.E. degree from Pratt Institute, New York City, in 1964, and has done graduate work at Syracuse and Northeastern Universities. A cooperative student for the Department of the Navy from 1959 to 1965, he joined the IBM Corporation, Systems Development Division, in 1965, where he was involved in the development of high-speed integrated circuits with particular emphasis on semiconductor monolithic memories. In 1968, he joined the Honeywell Computer Control Division, where he presently is engaged as a senior engineer in the Computer Circuit Development Department. At Honeywell, he has been involved in various aspects of integrated phenomenons in digital devices, semiconductor development, and interconnection effects in high-speed ICs. He is currently responsible for the development of advanced computer devices and circuits. Mr. DeFalco has published several papers on computer-aided design, device modeling, linear network analysis, and semiconductor device development.

The role of the engineer in public affairs

A French industrialist and public servant sees engineers as needing better preparation for influencing public opinion

Maurice Ponte Agence Nationale pour la Valorisation de la Recherche

The engineer as a political person is considered in this address presented by M. Ponte, representing the Institut de France (Academie des Sciences), at the second general meeting of the World Federation of Engineering Organizations at UNESCO, Paris, France. Dr. Ponte is a former vice president of the Institute of Radio Engineers.

At a time when technological progress is throwing the world into confusion, and technocracy is subject to criticism—criticism sometimes undeserved—it is good for us to consider, dispassionately, the role of the engineer in public affairs. The concept of "engineer" is hard to pin down, but the historical evolution of the engineer's role will help us toward a definition.

The works of the ancients that have come down to us are astonishing for the sophistication (based on experience, though perhaps not on theory) that they reveal. Without even dwelling on the extraordinary monolithic statues on Easter Island, we note the spectacular works at Luxor, the pyramids, the Greek temples, the Roman aqueducts—of which the Pont du Gard is a witness that has defied time and man—and the techniques for heating homes or the baths; all of these reveal the mediation of the engineer's skills. But it is possible to date the idea of the art of the engineer proper from the Renaissance, with Leonardo da Vinci (1452–1519), and particularly from the 17th century. In the beginning, the art was associated with military tasks, and the engineer had an indirect political role because he fabricated the implements of the sovereigns, as is betokened by the mottoes appearing on the royal artillery pieces. Littré, in the 1865 edition of his dictionary, and even in later revisions, begins the definition of "ingenieur" by "he who invents, who delineates and who conducts works and devices for attacking, defending or fortifying places . . ."

This initial description stems from the fact that among the attributes of the engineer one must give prominence to the ideas of efficacy, of method, of an allotted time, and of quality. In the ancient works already mentioned,

STRASBOURG CATHEDRAL. The present structure was built from the year 1015 onward; the spire, 141 meters high, was completed in 1439.

and in the old cathedrals, the task was accomplished by a clerk of the works, surrounded by the most reputable craftsmen, some of whom came from far away; but time probably was not important, being counted in decades to such an extent that few undertakings were completed—at any rate, not in the initial style. A typical example is the marvelous facade of the Strasbourg cathedral. The designs of Erwin of Steinbach, which luckily are preserved, specify the general appearance desired, with details of the style, but nothing is recommended as to the procedures and details of accomplishing the intended result; these details were left to the experience of the builders.

It is clear that time and cost of construction counted for little. We know now that the stones of the Romanesque churches were frequently shaped on the spot, just before being set in place. After all, do we not notice, with due allowances for differences, some vestiges of these practices in present-day tensions between architects and engineers? This could not be so in military practice, where it was necessary to conceive the works as quickly as possible, to execute them and adapt them for the use of the troops: the scythe chariots of the Persians, the catapults, the siege machinery used in Roman camps, already bore the marks of the engineer's skill, and it was necessary for the techniques to be established by making best use of those available at the moment—techniques adapted to the ripostes of the enemy, or anticipating them if possible.

If I dwell somewhat on this point, it is for the purpose of defining the essential characteristics of the resulting engineer, trained to work with real objects, on the basis of precise data, whereas political man acts quite differently and, in any case, cannot be guided solely by technology because he deals with people, so that he is more like a clerk of the works. We shall return to this point later.

An engineer in trouble

A typical example, perhaps the oldest known, of an attempt of an engineer to intervene in political affairs, is found in the adventures of Vauban (1633–1707). He was clearly an engineer in the modern sense, and consequently, in his time, a military engineer. If I may say so, he even put his own shoulder to the wheel: not satisfied with having created an art of fortification and with having made the forts a reality, from 1665 to 1695 he directed 53 sieges. Louis XIV made him marshal of France, even though Vauban, born in the minor gentry, mistrusted this title; one feels clearly that this compleat engineer was afraid of the consequences of this elevation to the highest honors. The absence of any heredity in this field made him take the honors seriously. Does not the Duke of Saint Simon say about him that "it is inconceivable that with so much honesty and freedom, incapable of lending himself to anything untrue or evil, he was able to advance to a point where he won the friendship and confidence of Louvois and of the king"?

Thus it is that because his character as engineer induced him to calculate and reason honestly, Vauban considered himself capable of tackling a political matter that shocked him greatly: the tax load, which hit the French people hard, but of which little got into the coffers of the king; en route "intermediaries" sidetracked most of the funds. He set up a complete fiscal system (that early!) based on numerous quantitative studies made throughout the realm. The work was truly solid, a beautiful engineering job; but it made the mistake of ruining some powerful people—starting with the comptroller general and the chain of commissioners of finance. According to Saint Simon again, "the entire judiciary is yelling for its own interest" and the king, surrounded this time by more adroit men, saw in Vauban only "a criminal who attacked the authority of the king's ministers, and consequently the authority of the king himself." Vauban died over this. "He was nonetheless admired in all of Europe, even among his adversaries, and was regretted in France, by all who were not financiers or tools of financiers." This example should remain in our memory while we attempt to establish some conclusions regarding the role of the engineer in public affairs.

Evolution of engineering

But let us put aside these first unfortunate attempts at intervention by a true engineer in public affairs to follow the historical development of the idea of "engineer." About the middle of the 18th century, the necessity began to be felt for training specialists for nonmilitary undertakings that required an alliance of technical knowledge with elements that are conducive to the conception, production, and exploitation of things that are useful to the economy. It was probably in France that the first school of engineering was established, namely, l'Ecole Nationale des Ponts at Chaussées,* which dates from 1747 though the corporate entity goes back to 1716. One should note that its founding actually aimed at public service; the schools that followed, whose graduates swarmed across the world, emphasized the same trend: Ecole des Mines in 1778, Ecole des Travaux Publics in 1794, which became l' Ecole Polytechnique while taking on a military character now abolished, Ecole des Arts et Métiers in 1788, and Ecole Centrale des Arts et Manufactures in 1829. The last two schools greatly expanded the concept of the engineer's art beyond public services, but it is important to reiterate that since the 18th century the engineer has appeared as essential and indispensable in the technology and development of public services. In the United States, in England, in Germany, it was in the first half of the 19th century that various engineering courses were offered, either at the engineering schools or the universities. It is rather remarkable that at present the distinction between education for engineering and education for science and research is being blurred, to the regret—which is frequently well founded—of the employers of engineers.

What is the present state of this evolution? We have now to describe the essential characteristics of the modern engineer which prepare him inevitably to play a part in political affairs.

The modern engineer

The modern engineer can no longer be satisfied with just the conception and production of a certain object; he must think in systems. Examples abound. A modern aircraft, civilian or military, is used as part of a system; it cannot be designed by itself, because left to itself it could not complete any missions and could not even return to its takeoff point. Information processing is equally illustrative of this state of affairs: the central unit must be conceived on the one hand according to the state of the technology of components, and on the other hand consistently with what one wants to do, through software. It is becoming more and more unthinkable to create a product and then to determine what will be done with it. The striking proof of this modern method of industrial conception is obviously furnished by the space effort: the spectacular outcome of the space projects is the work of practicing engineers, with the scientists in the end playing only a small part.

I would like to emphasize here that in this systems approach the so-called "Cartesian" method of breaking down problems into elements luckily loses its importance, and will continue increasingly to do so. This trend is becoming more conspicuous with the increasing use of automation and its encroachment on new areas. A classic example is the automation of a product or of a service; the idea is well known and there is no point in dwelling on it, except to stress that certain products can be produced only by automated procedures, which alone can ensure the necessary uniformity and quality. This necessity implies in turn that the worked-on materials themselves must be of a particularly perfected uniformity: the dexterity of the worker can no longer be called upon to correct a faulty operation, and automation proceeds by "all or nothing."

The automation of the process of production itself can be introduced into an economic system by seeking the lowest cost price, which the experts call the automatic optimization of production. But automation is no

---

*In English, the names of this and the following schools are, respectively, National School of Bridges and Roads; School of Mines; School of Public Works, which became the Polytechnic School; School of Arts and Crafts; and Central School of Arts and Manufactures.

longer limited to the manufacture of a product once it has been conceived: automation becomes an essential instrument of design itself to reach a given objective. One employs the ability of the computers to analyze a design and draw from it technical and economic conclusions, and to propose the modifications to be used for reaching the definite economic goal.

This method of using computers is already being widely utilized in various fields—the computing of bridges, of metal structures, or of electronic components—but it also extends to much more complete projects, of which the most typical at present is the design of the "Maverick" automobile by Ford. After the U.S. designers found that the marketing of the "compacts" was unable to stem the tide of cars imported into the United States (10 percent of the market in 1968), the decision was made to bring out an American car that would be priced between the imported cars and the lowest-priced compacts. Certain conditions were placed on the system "car market, U.S. vs. imports" and in implementing the project, each engineer, designer, and technician had access to a computer, naturally including all the necessary numerical-control machines. A considerable total of designs and technological studies were made, with outcomes available almost instantly. They ranged from basic materials to the most refined devices for the passengers' comfort, always keeping within the constraints on price.

The foregoing is a characteristic example of this application of electronic computers to the planning of products at the initial stage, and not merely—as was true not so long ago—to a product worked out by the traditional office of planning. But this is not the only example. The same facilities exist in England, particularly in mechanical engineering, at the Mintech Atlas Centre (Cambridge) where a designer of a part or of a product can communicate from afar with the computer to determine the best specifications to attain a given economic goal. These methods will become widespread in the work of engineering.

And what is essential in these two great areas of automation, namely, applications to conception and to production, the engineer cannot stay alone and confront his special field. He must work as a member of a team, whose importance and versatility grow with the magnitude of the project. We have come a long way from the elementary association of specialized engineers and a few technician-designers. In a project such as the Maverick or automation of the manufacture of plastics, one needs mechanical engineers, chemists, electronic engineers, and thermodynamicists—also esthetic designers, economists, marketing specialists, and even jurists and sociologists. Consequently, during his development, an engineer—even a specialist—must receive economic and even legal education; this is a departure from the classical French tradition, which is almost entirely technical. The creation of academic institutes of technology has taken into account this obligation for training technologists; the engineering schools have to adapt accordingly. On the other hand, in view of the importance assumed by the social sciences, a category of engineers that one could call "literary" gradually will be created. The special training of the engineer thus certainly appears to be the acquiring of knowledge belonging to a certain technical field, but this must be broadened by notions that will permit him to work with a team combining other special fields, seemingly rather removed from the first. The famous Massachusetts Institute of Technology, founded in 1861, offers a fine example of this synthesis.

And thus we are in the presence of a group that will play a prominent part in public affairs, as engineers become more and more involved in this many-faceted work, needing to work in diversified teams. This modern method of working in teams—which gives the engineer unique power in the economy—is taking on, more and more, an international character. In fact, one of the barriers to communication and dissemination of the results of thinking and of action is language. The commercial and technical predominance of English does not suffice, because, even in technology, exacerbation of nationalism is developing linguistic parochialism. Translation machines are proving unable to transfer information from one language to another without degrading it. The electronic machines are beginning to give answers based on their own languages; certainly these are at present usually derived from English, but their establishment has led to study of the logic of languages and to anticipation that a universal mathematical language can be created specifically for the machines. Engineers—and scholars—thus will be able to understand each other directly, which would suggest that their work in public affairs can become international. This development would add weight to actions such as those of the World Federation of Engineering.

Missing for the modern engineer

Before drawing conclusions by "passing to the limit" of public affairs themselves, to use a mathematician's expression, one must define once more the mission of modern engineers. It has been said and written again and again that scientists should play a part in politics, and the subject of science and politics has been widely discussed. Scientists wear the halo of research, sometimes accompanied by manifestations like the Nobel Prize. They inspire admiration that is sometimes apprehensive, because what they discover or make can be used for the good of humanity or for humanity's woe. Now, in practice, what can we observe?

The importance of research for developing an economy has been widely recognized; in certain fields, products now on the market or in course of exploitation did not even exist ten or, in some cases, five years ago. It is precisely this industrial outgrowth of research that has created and enlarged the sector usually called applied research and development—its frontiers are poorly marked between basic research (that of "scientists") and production—which, in our technology-dominated era, could itself be continuously under development, particularly as an effect of automation.

Now, the resources that must be dedicated to applied research and development are incomparably greater than those demanded by basic research; the latter can really be ignored. In reckoning up the results newly acquired from space systems, we can state that those that are direct results of basic research are very limited in number—less than one percent. In many fields there is an established rule according to which, if one spends a thousand for research, development for production will cost ten thousand, and actually getting into production will cost one hundred thousand.

The example of the Ford Maverick project certainly pointed up the importance of this progression; space is another field where the cost of basic research is eclipsed by the costs of setting up a network of users—for Apollo 11, there were thousands of ground personnel. The same applies for all large projects and achievements at the present time. Subject to the remarks already made concerning the variety of specialists who must be linked for the joint undertaking, all this work is mainly that of engineers and their associates. And this is true for all political regimes: The role of the engineer—his importance in the functioning of the economic complex that constitutes a modern nation—is getting ever weightier. This preponderance is cleanly ratified in such entities as the industrial firms in a capitalist regime—so much so, that one could just as well write that a company belongs in reality more to its engineers and technicians than to its stockholders. Obviously, one must temper this opinion for the sake of the banks, but it certainly contains a lot of truth.

This study of the modern engineer's psychological makeup leads us to envisage him partaking in public affairs, in the sense of affairs with general interest, without going as far as politics. Into this classification go at the outset the fields that are completely technical, such as air pollution and water pollution, and these are primarily matters for engineers. But one could go further, into "composite" affairs, and the first of these is education—at all levels, but in particular higher education. One of the great perils at present is that education will be cut off from current life by remaining too theoretical or even obsolete. The students live intensely, like everybody else, under the influence of modern telecommunication; they are informed about the technical results that have been achieved: the exploits of the American astronauts have publicized the moon, and its attributes such as gravitation, better than yesteryear's courses in "cosmography." Teaching, to be alive and to capture the students' interest, must therefore adapt constantly to the actuality. In October 1968, at the Royal Society, very pertinent observations on the subject were presented: it was said there that "weakness of instruction is the factor which most often chills the initial enthusiasm for scientific studies," and among these weaknesses, several Fellows judged that a large part of science at school level was "boring and sterile" and that one could not really blame the students for turning their backs on it to take up arts or letters.

It is a remarkable fact that in all the reputedly free countries, although science, technology, and engineering show their power and clearly offer the greatest number of opportunities, the percentage of students who are attracted toward the scientific and technical disciplines is constantly decreasing. In France it is distinctly lower than the figures of the Plans. In 1969 there was one student in science for every two in letters and law; the need calls for equal numbers. The solution must be sought through active participation of engineers in education. This already is the case in schools of engineering and university institutes of technology, but one must enlarge this movement in all faculties—beginning, naturally, with that of science. This would not mean shoving aside the professors; through their general knowledge and their permanent contact with basic research, they must assure the foundations of education, but they ought to be linked up with engineers who would be selected for their knowledge "in real time," and for their pedagogic talents. This plan will be all the more necessary when programmed instruction—and, above all, directed instruction—is introduced. In the latter method, the computers furnish to the student personal assignments supervised by professors, who are thus freed from most of their schoolmasterly duties; the work to be done is determined by the computers as a function of the student's ability and knowledge, and also, equally, as a function of the prospective needs of the national economy. The next assignment is determined by the computer on the basis of the results obtained and the professors' observations. It is evident that in such a system, which is probably that of the future, the role of engineers must be essential in keeping up to date the economic, technical, and practical data of the programs and setting them up, and in the existing employment situation.

Henceforth, this kind of mediation of engineers in teaching is the most practical form of liaison, and the most immediately profitable. We need a university-industry complex, achieved through bilateral exchanges, with the engineers going to the university to instruct and the professors being accepted into companies for personal edification. This interpenetration constitutes an essential part of the "continuing education" that is recognized as indispensable, whose effectuation should provide a particularly pointed example of cooperation between university people and engineers. Permit me to conclude this part of my article by making reference, like others, to the present regime of Mao. The "cultural revolution" grew out of various causes, but one of them—the most pertinent to our subject, and in reality one of the most decisive for accomplishing the revolution—was that economic studies have shown that technologists and engineers have a much more decisive role in the economy than scholars do; this is how the cultural revolution, with its "scientific experimentation," has fundamentally transformed classical education by giving an essential role to the techniques of the engineer and to their mission in the education of students. This fact deserves to be stressed at a time when certain people label as Maoist some ideas that are quite at odds with this economic pragmatism.

Engineering and art

Another public affair, taken in the sense that it involves a broad domain by its definition and by its audience, is art. Certainly not everything in art is technique, and this will be a good transition between the characteristic tasks of the engineer and politics. All great historical periods have their artistic expressions—which, by the way, are more or less ahead of their times. Architecture is probably the art that, since earliest history, best symbolizes an era: it is itself dependent on the materials prevailing at the time when its great outlines take shape. The stone of our cathedrals gave them its character, as concrete and steel have modeled our modern structures. It is in this sense that technique has always influenced the architects or master builders. The founders and principal workers of the Bauhaus showed this clearly. However, in our era, there is something more: All that we have said about the working methods of engineers and their ways of thinking can be applied to architecture, and what they do is not necessarily ugly. It has even been said—

rightly—that machinery resulting from good technique is beautiful. Why should this not be the case in architecture? Certainly the architect still has his role, and a cooperative work cannot but be fruitful.

But, in fact, innovative ideas stemming from new materials or new procedures come mostly from engineers. They are in any case the best suited for laying down the basis of construction projects as a function of the allowed cost or of any other economic constraint. Madame Françoise Parturier usefully pointed out a few months ago that systematic attempts at cooperation between artists and engineers and technologists are presently under way in New York and in London.* The New York association was founded by an electrical engineer of Swedish extraction: it has the technical and human resources so that the artists are not hampered in their conceptions by technological difficulties, and it is not surprising that the American Foundation on Automation and Employment is directly interested in the matter. And, although artistic ties with the engineer concern architecture in particular, they are not absent from other fields of art, particularly music. And one should not bawl about barbarism. Music has always been way ahead of its era; let us not forget that Beethoven at first was considered as an expert on cacophony. Thus Madame Parturier is correct when she writes that in any esthetic renewal, technology can be a determining factor, and technology is certainly in the domain of the engineer.

The engineer in public affairs

Having thus traced the evolution of the idea of engineer, and dwelt on his modes of action and thought in our time, we are better able to envisage his part in public affairs. The problem must be examined with regard to two fundamental aspects: that of the help that he will provide, and that of decision making. On the one hand, the engineer is accustomed to handling precise truths, often expressed numerically; he analyzes them and concentrates his efforts toward a well-defined goal. If the objective is not defined clearly enough, his first job will be to delineate it better. Competition—and the implication of this word is not merely commercial, because it relates to all economies, even if the motivations are different—is his lot; he is no longer merely an applier but must be an innovator too. Even if intuition plays a part in his decisions, these are essentially technical and economic. In sum, the modern engineer is one who best utilizes, with a definite purpose, the scientific, technical, and technological resources of the day, by constantly adapting to them. This philosophy is clearly mentioned in the Federation plan for a code of professional ethics of the engineer. And modern facilities, notably electronics, give the engineer extraordinarily great powers, even going so far as modifying the modes of thought by languages which are his own and which are understood internationally.

Real revolutions occur in the modern age through technology, the domain of the engineer. We are now living in one of these revolutions, that of the computers, that is at least as important as the Renaissance. Its consequences, by the way, are not only economic, but also cultural and political. This is a good example of the power of the engineer's role, at least indirectly, in public affairs.

Thus, the modern engineer, with his precise ways of thinking and the previously mentioned traits that lead him to work in the style of public affairs, with the help of electronic machines, appears to be indispensable in cases where public projects must be planned and evaluated. And, as present events reinforce the idea that economics is an essential and often decisive factor in politics, the engineers must play a fundamental role since they can light the way. In that activity, I am not ascribing to them the qualifications of economists, analysts, or others.

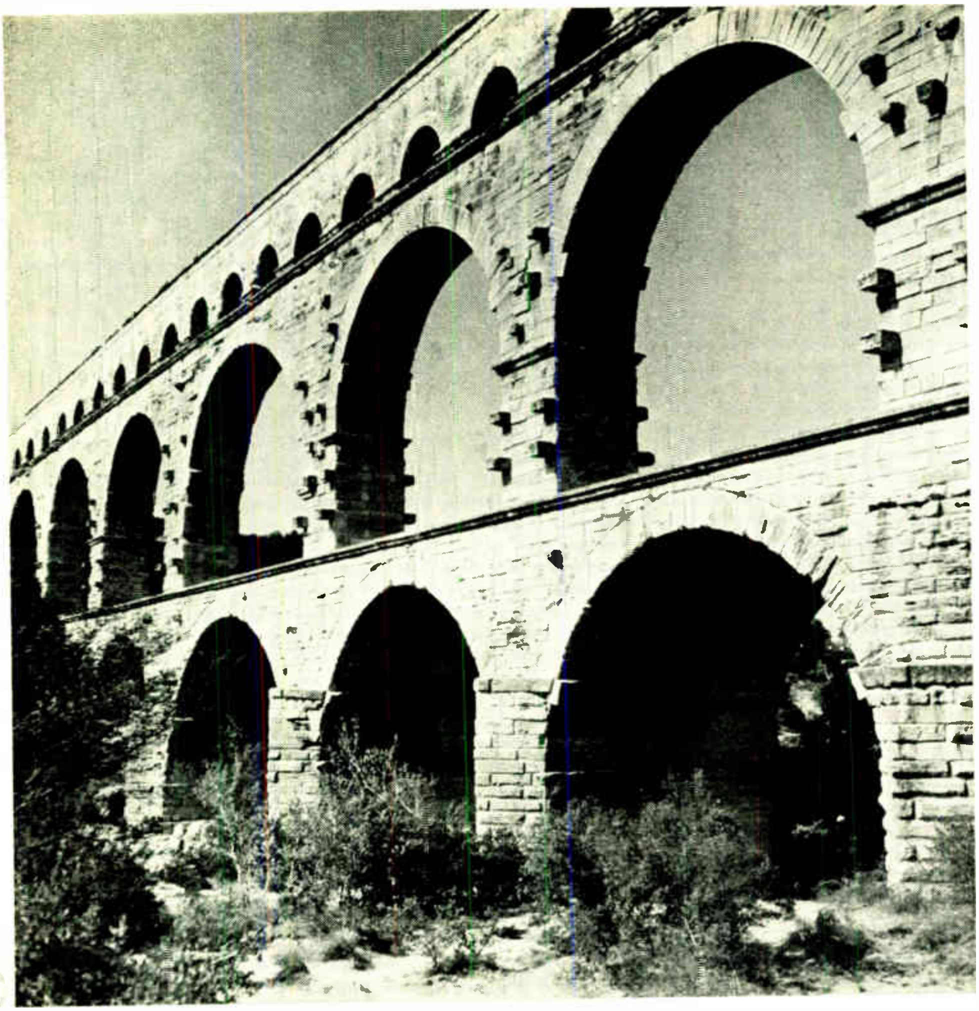
We have seen that it is necessary to have complete teams available. Many nations are finding out at this time what it costs not to have followed the imperatives of forecasting. No public budget, no precautions in foreign affairs, no long-term or short-term program, ought to be decided upon until the short- or long-term consequences have been evaluated either in financial terms or (which is the same thing) in human terms, while ascertaining the amount of knowledge and technological investment that will be needed in order to attain that goal.

L. A. Lecht and C. L. Schultz, in their recent works reviewed in *Science*,[1] did an engineering job of looking at whether the United States, even with its great economic power, would be able to attain in 1975 all the objectives that politics desires, and even promises: renewal of cities and the struggle against pollution, better housing and the fight against poverty, space exploration, great projects of public health, education, and so on. The conclusion is surprising, and is worth pausing over because it shows the danger of decisions that are made for sentimental or demagogic reasons, which is always a hazard of politics. By extrapolating the present figures, it appears that the gross national product forecast for 1975 would allow the attainment of sixteen of the goals that were conjured up, but another set of computations shows that, in order to reach these goals, there would be a shortage of about ten million able-bodied people. As there are only five years to go, it is difficult to find a remedy, and one must face the facts and limit one's aspirations.

Therefore it now remains to be seen what projects should be kept. The engineers and their associates argue, for example, that for housing or for education or for health an investment of an additional billion dollars would require almost the same number of workers—say one hundred thousand persons—but that the division of workers into various categories would be very different for different programs. For housing, 61 percent of the force would be blue-collar workers and 8 percent would be professional specialists; for certain other programs, one would need 46 percent professors, doctors, and specialists of various sorts, and only 16 percent would be blue-collar workers. Therefore, it is not sufficient to reduce or cut out some projects chosen at random; one must properly determine the consequences of the decisions—which, you will note, overlap extensively—because the training, and even the housing, of employees of all kinds must be foreseen for all the plans, including those for training and housing. Also, it will be necessary to investigate which projects will create the least amount of unemployment—or better, those that will best utilize

---

* These experiments were described in a pair of articles by Nilo Lindgren in IEEE SPECTRUM, April and May 1969.



THE PONT DU GARD, in southern France, built early in the first century of our era, is part of an aqueduct that carried water to the town of Nimes from springs on another hill 41 km away. The height from the valley floor is 50 meters; except in the canal on top, the structure contains no mortar.

the unemployed. This brings to mind the social importance of the engineer's role, on which I shall not expatiate, however.

The human area

On the other hand, in matters that are essentially political, we approach a field where the motivations are basically different, because here one has to deal with people. One is now in an area of utter indetermination, because the reactions of groups of humans are as yet uncertain. This has always been so. If Talleyrand and Briand had been engineers, it is likely that they would not have been able to be Talleyrand and Briand. But these behavior patterns have become more erratic under the effect of mass communications; the individuals who are accustomed to deluges of news and its immediate denial, subjected to actions of controlled propaganda that is shamelessly denied within a very short time, come to a point where they no longer believe anything. The announcer's image or voice represents to them only theatre, and that is what he takes it to be. However, there still remains something, particularly when television is employed and currents of opinion get started like epidemics; one knows that it is hard to provoke them, but that they develop suddenly at the moment when one least expects them, a little like human migrations. This is what explains the present imperviousness of the masses to any logical arguments, quantitative or not—this skepticism or disenchantment that can suddenly change into unforeseen currents of opinion. Actually, these are the result of public expression by a relatively limited number of individuals.

In a liberal political regime, thanks to the untrusting state of mind that has been evoked as mentioned above, opinion is divided about equally between two sides, when the number of people consulted is large and there are only two choices. Accordingly, one must know how to displace at an opportune moment a percent or so of the votes; the politician can do this by using his intuition, his capacity for human contacts, his (at least apparent) confidence in himself. The engineer cannot be at ease in such a situation.

On the other hand, would it be desirable that public affairs taken in a political sense should be governed and regulated by purely numerical and "practical" considerations? Certainly not, because people fortunately are not standardized like machines. Examples are plentiful. On the military side, all modern sophisticated armaments presume that the lives of the combatants should be protected. They become ineffectual if the adversary is unmoved by this feeling, but puts foremost a struggle for an ideal, accepting sacrifices of life or of comfort. The modern weapons then become mostly a bother, by their demands on logistics. Certain present wars furnish proof for this, wars that were launched and conducted for a time on the basis of computer outputs.

The problem of the blacks in the U.S., the troubles in Northern Ireland—the racial difficulties in general—are in the domain of politics, and the engineers cannot do anything about them, except perhaps to repair the havoc that happens. The often violent positions of some of the youth in all countries against society stem in large part from their reaction to the fear of servitude to an entirely technical and numerical philosophy, typified by "computer forecasting." For 20 years we have pointed out the dangers of defining a rise in the standard of living as an increase in commodities and conveniences; this is an adult definition, which fails to satisfy youth. The young respond intuitively in the manner of someone who wants to "stump" a chess machine; they will not play the game.

All this cannot be figured and evaluated in an exact manner, and an engineer, with his own unaided methods, would risk being as disappointed as Vauban was.

The outcome of this survey is therefore that it is necessary to seek out, in public affairs that are likely to involve politics, a harmony between the role of the politicians—who, in the final analysis, make a decision—and the role played by engineers, who bring along the elements of foresight and decision, with an understanding of powerful resources unknown until now. The apportionment of these roles will depend on the degree of economic development of the country.

It seems that this is not always an evident truth. In his renowned book *The New Industrial State*,[2] J. K. Galbraith does not mention the word "engineer" even once, and it does not have the honor of appearing at all in the index, which however includes a large number of entries under "technology" and "technostructure." Galbraith defines decision makers by their roles, rather than by their education, with emphasis on specialized knowledge, but without apparently attaching any special importance to the perceptions of the engineer. It is an error to confuse the term "technician" with "engineer" as the latter is distinguished as a man who makes things materialize. This trait is particularly important at the present time, because political decisions now waver between worship of the past and the adoption of new directions, some of the latter being nevertheless just as conservative or even more so, though having the appearance of being revolutionary. The engineer is there to affirm the potential of the innovation, and to measure its importance.

In regimes other than capitalism, this importance of the economic role of the engineer is expressed in other terms, but in fact it is just as well attested; we need only to cite the importance given to the training of engineers of all kinds, and the numbers of engineering students. But in politics the role of engineers—as such—remains weak, political philosophy still being the controlling factor in the government's actions.

Conclusion

In truth, there is no longer any public affair that is entirely political, nor is there any that is entirely technical. This is the case with the great pilot programs for research—such as space, or problems of energy. What points this up well for space is the extraordinary reply to an opinion poll made in the United States among "average Americans" after the spectacular results of Apollo 11, so brilliantly turned to account on the foreign scene: 56 percent of the people questioned replied that these expenses for space are excessive and that they must be reduced. The engineers are there to furnish the costs and the modalities of the programs. It is up to the government to evaluate and to make decisions, after weighing the economic—and therefore political—consequences. This, in fact, will be one of the incentives for action by engineers that will be felt more and more. It will be no longer suffice for them to say how to do it; faced with the accelerated progress of technology, they will have to advise the politicians on what to do.

Thus the engineers, inside or outside public affairs, must be associated with decisions, to a degree that depends on the nature of the question under debate. To ignore this manifest fact is to run into economic, and thus political, miscalculations. Engineers should not be uninterested in politics. It is not enough for them merely to carry out their technical tasks. Therefore I suggest that the Federation include in its recommendations for the development of engineers a provision concerning their preparation for public affairs, taken in a larger sense than that of mere administration.

REFERENCES

1. Wolfle, D., "Which goals to emphasize," *Science*, vol. 164, p. 249, Apr. 18, 1969.
2. Galbraith, J. K., *The New Industrial State*. Boston: Houghton-Mifflin, 1967.

Maurice Ponte (F), managing director of the Agence Nationale pour la Valorisation de la Recherche (National Agency for the Evaluation of Research—ANVAR) was born in Voiron, France. He is a graduate of the Ecole Normale Supérieure and holds the degrees of Agrégé d'Université and Docteur ès Sciences. He began his research career in 1924, working with Henry Abraham, Sir William Bragg, and Prince Louis de Broglie, and in 1929 became associated with the Compagnie Générale de T.S.F. In addition to his IRE vice presidency in 1954, he has been president of the following companies and societies: Société Francaise des Electriciens, Société des Radioélectriciens, Compagnie de Télégraphie sans Fil, Compagnie Francaise de Télévision, and Société Francaise d'Astronautique. He is a member of, among others, the Administrative Council of the Ecole Normale Supérieure, the Institut Pasteur, and the Institut de France. The author of various technical and scientific papers, Dr. Ponte is director of *Annales de Radio-électricité* and also of Téléondes. He has received a number of decorations, including Commander of the Légion d'Honneur, of the Lion of Finland, and of the Order of Merit of the Italian Republic, and has been awarded the Blondel Medal and the Christopher Columbus Prize. Dr. Ponte also has been honored as a Grand Officier du Mérite. He has been affiliated with ANVAR since 1968.

Least-squares estimation: from Gauss to Kalman

The Gaussian concept of estimation by least squares, originally stimulated by astronomical studies, has provided the basis for a number of estimation theories and techniques during the ensuing 170 years—probably none as useful in terms of today's requirements as the Kalman filter

H. W. Sorenson University of California, San Diego

This discussion is directed to least-squares estimation theory, from its inception by Gauss[1] to its modern form, as developed by Kalman.[2] To aid in furnishing the desired perspective, the contributions and insights provided by Gauss are described and related to developments that have appeared more recently (that is, in the 20th century). In the author's opinion, it is enlightening to consider just how far (or how little) we have advanced since the initial developments and to recognize the truth in the saying that we "stand on the shoulders of giants."

The earliest stimulus for the development of estimation theory was apparently provided by astronomical studies in which planet and comet motion was studied using telescopic measurement data. The motion of these bodies can be completely characterized by six parameters, and the estimation problem that was considered was that of inferring the values of these parameters from the measurement data. To solve this problem concerning the revolution of heavenly bodies, the method of least squares was invented by a "young revolutionary" of his day, Karl Friedrich Gauss. Gauss was 18 years old at the time of his first use of the least-squares method in 1795.

As happens even today (e.g., the Kalman filter), there was considerable controversy in the early 19th century regarding the actual inventor of the least-squares method. The conflict arose because Gauss did not publish his discovery in 1795. Instead, Legendre independently invented the method and published his results in 1806 in his book *Nouvelles méthodes pour la determination des orbites des comètes*. It was not until 1809, in his book *Theoria Motus Corporum Coelestium*, that Gauss published a detailed description of the least-squares method. However, in this treatise Gauss mentions Legendre's discussion of least squares and pointedly refers to his own earlier use (p. 270, *Theoria Motus*)*: "Our principle, which we have made use of since the year 1795, has lately been published by Legendre in the work *Nouvelles méthodes pour la determination des orbites des cometes, Paris, 1806*, where several other properties of this principle have been explained which, for the sake of brevity, we here omit." This reference angered Legendre who, with great indignation, wrote to Gauss and complained[3] that "Gauss, who was already so rich in discoveries, might have had the decency not to appropriate the method of least-squares." It is interesting to note that Gauss, who is now regarded as one of the "giants" of mathematics, felt that he had been eclipsed by Legendre and wrote to a friend saying,[3] "It seems to be my fate to concur in nearly all my theoretical works with Legendre. So it is in the higher arithmetic, . . . , and now again in the method of least-squares which is also used in Legendre's work and indeed right gallantly carried through." Historians have since found sufficient evidence to substantiate Gauss' claim of priority to the least-squares method, so it is Legendre rather than Gauss who was eclipsed in this instance and, indeed, in general.

The method of least squares

The astronomical studies that prompted the invention of least squares were described by Gauss in *Theoria Motus*.[1] The following quotation (p. 249) not only describes the basic ingredients for Gauss' studies but captures the essential ingredients for all other data-processing studies. "If the astronomical observations and other quantities on which the computation of orbits is based were absolutely correct, the elements also, whether deduced from three or four observations, would be strictly accurate (so far indeed as the motion is supposed to take place exactly according to the laws of Kepler) and, therefore, if other observations were used, they might be confirmed but not corrected. But since all our measurements and observations are nothing more than approximations to the truth, the same must be true of all calculations resting upon them, and the highest aim of all compu-

---

* The page numbers here refer to the English translation available from Dover Publications, Inc.[1]

tations made concerning concrete phenomena must be to approximate, as nearly as practicable, to the truth. But this can be accomplished in no other way than by a suitable combination of more observations than the number absolutely requisite for the determination of the unknown quantities. This problem can only be properly undertaken when an approximate knowledge of the orbit has been already attained, which is afterwards to be corrected so as to satisfy all the observations in the most accurate manner possible."

Let us briefly reconsider some of the ideas contained in the preceding statement and relate them to "modern" developments.

1. Gauss refers to the number of observations that are absolutely required for the determination of the unknown quantities. The problem of establishing this minimum number of observations is currently discussed in terms of the "observability of the system" and is the subject of many papers; see Refs. 4 and 5, for example.

2. Gauss notes that more observations are required than this minimum because of the errors in the measurements and observations. Thus, he notes the need for "redundant" data to eliminate the influence of measurement errors.

3. Gauss implies that the equations of motion must be exact descriptions, and therefore the problem of dynamic modeling of the system is raised.

4. Gauss requires that approximate knowledge of the orbit be available. This is currently required in virtually all practical applications of Kalman filter theory,[6] for example, and implies the use of some linearization procedure.

5. Gauss states that the parameter estimates must satisfy the observations in the most accurate manner possible. Thus, he calls for the residuals (that is, the difference between the observed values and the values predicted from the estimates) to be as small as possible.

6. Gauss refers to the inaccuracy of the observations and indicates that the errors are unknown or unknowable and thereby sets the stage for probabilistic considerations. In doing so, he anticipates most of the modern-day approaches to estimation problems.

7. Finally, Gauss refers to the "suitable combination" of the observations that will give the most accurate estimates. This is related to the definition of the structure of an estimation procedure (i.e., linear or nonlinear filtering) and to the definition of the performance criterion. These are extremely important considerations in current discussions of estimation problems.

As stated earlier, Gauss invented and used the method of least squares as his estimation technique. Let us consider Gauss' definition of the method (Ref. 1, page 260). He suggested that the most appropriate values for the unknown but desired parameters are the *most probable values*, which he defined in the following manner: ". . . *the most probable value* of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum." The difference between the observed and computed measurement values is generally called the *residual*.

To make the discussion more precise, consider the following statement of the estimation problem. Suppose that m measurement quantities are available at discrete instants of time $(t_1, t_2, \ldots, t_n)$ and are denoted at each time t_k as $\mathbf{z}_k$. Suppose that parameters $\mathbf{x}$ are to be determined from the data and are related according to

$$\mathbf{z}_k = H_k\mathbf{x} + \mathbf{v}_k \qquad (1)$$

where the $\mathbf{v}_k$ represent the measurement errors that occur at each observation time. As is seen in Eq. (1), the measurement data and the parameters $\mathbf{x}$ are assumed here to be linearly related, thereby making explicit the assumption that Gauss indicated was necessary in the foregoing quotation.

Denote the estimate of $\mathbf{x}$ based on the n data samples $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$ as $\hat{\mathbf{x}}_n$. Then, the residual associated with the kth measurement is

$$\mathbf{r}_k \triangleq \mathbf{z}_k - H_k\hat{\mathbf{x}}_n \qquad k = 0, 1, \cdots, n \qquad (2)$$

The least-squares method is concerned with determining the most probable value of $\mathbf{x}$ (that is, $\hat{\mathbf{x}}_n$). This most probable value is defined as the value that minimizes the sum of the squares of the residuals. Thus, choose $\mathbf{x}$ so that

$$L_n = \frac{1}{2}\sum_{k=0}^{n} [\mathbf{z}_k - H_k\mathbf{x}]^T W_k [\mathbf{z}_k - H_k\mathbf{x}] \qquad (3)$$

is minimized. The elements of the matrixes W_k are selected to indicate the degree of confidence that one can place in the individual measurements. As will be explained more fully in the discussion of the Kalman filter, W_k is equivalent to the inverse of the covariance of the measurement noise.

Gauss with his remarkable insight recognized that the simple statement of the least-squares method contains the germ of countless interesting studies. As he says in *Theoria Motus* (Ref. 1, page 269): "The subject we have just treated might give rise to several elegant analytical investigations upon which, however, we will not dwell, that we may not be too much diverted from our object. For the same reason we must reserve for another occasion the explanation of the devices by means of which the numerical calculations can be rendered more expeditious." Judging by the interest in estimation theory over the years, this statement must stand as one of the greatest understatements of all time. In passing, we note that the Kalman filter can be rightfully regarded as an efficient computational solution of the least-squares method.

Gauss did not merely hypothesize the least-squares method; it is interesting to consider his discussion of the problem of obtaining the "most probable" estimate as an introduction to more modern techniques. First, it is significant that he considered the problem from a probabilistic point of view and attempted to define the best estimate as the most probable value of the parameters. He reasoned that errors in the measurements would be independent of each other, so the joint-probability density function of the measurement residuals can be expressed as the product of the individual density functions

$$f(\mathbf{r}_0, \mathbf{r}_1, \cdots, \mathbf{r}_n) = f(\mathbf{r}_0)f(\mathbf{r}_1)\cdots f(\mathbf{r}_n) \qquad (4)$$

Next, he argued that the density $f(\mathbf{r}_k)$ would be a normal density

$$f(\mathbf{r}_k) = \frac{\sqrt{\det W}}{(2\pi)^{m/2}} \exp\left[\tfrac{1}{2}\mathbf{r}_k{}^T W_k \mathbf{r}_k\right] \qquad (5)$$

although he recognized that one never obtains errors of infinite magnitude, and thus Eq. (5) is not realistic.

However, he rationalized away this difficulty by stating (page 259) that: "The function just found cannot, it is true, express rigorously the probabilities of the errors for since the possible errors are in all cases confined within certain limits, the probability of errors exceeding those limits ought always to be zero while our formula always gives some value. However, this defect, which every analytical function must, from its nature, labor under, is of no importance in practice because the value of our function decreases so rapidly, when $[\mathbf{r}_k{}^T W_k \mathbf{r}_k]$ has acquired a considerable magnitude, that it can safely be considered as vanishing."

Gauss proceeded by noting that the maximum of the probability density function is determined by maximizing the logarithm of this function. Thus, he anticipated the *maximum likelihood method*, which was introduced by R. A. Fisher[7] in 1912 and has been thoroughly investigated up to the present time. It is interesting that Gauss rejected the maximum likelihood method[8] in favor of minimizing some function of the difference between estimate and observation, and thereby recast the least-squares method independent of probability theory. However, in maximizing the logarithm of the independent and normally distributed residuals, one is led to the least-squares problem defined by Eq. (3).

Kalman filter theory

Let us now leave the early 19th century and enter the 20th century. Consider, briefly, some of the major developments of estimation theory that preceded the introduction of the Kalman filter. As already mentioned, R. A. Fisher introduced the idea of maximum likelihood estimation and this has provided food for thought throughout the subsequent years. Kolmogorov[9] in 1941 and Wiener[10] in 1942 independently developed a linear minimum mean-square estimation technique that received considerable attention and provided the foundation for the subsequent development of Kalman filter theory.

In Wiener–Kolmogorov filter theory, Gauss' inference that linear equations must be available for the solution of the estimation problem is elevated to the status of an explicit assumption. There are, however, many conceptual differences (as one would hope after 140 years) between Gauss' problem and the problem treated by Wiener and Kolmogorov. Not the least of these is the fact that the latter considered the estimation problem when measurements are obtained continuously, as well as the discrete-time problem. To maintain the continuity of the present discussion, attention shall be restricted to the discrete formulation of Wiener–Kolmogorov (and later Kalman) filter theory.

Consider the problem of estimating a signal $\mathbf{s}_n$, possibly time-varying, from measurement data $(\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_n)$, where the $\mathbf{s}_n$ and the $\{\mathbf{z}_i\}$ are related through knowledge of the cross-correlation functions. Assume that the estimate of $\mathbf{s}_n$, say $\hat{\mathbf{s}}_{n/n}$, is to be computed as a linear combination of the $\mathbf{z}_i$:

$$\hat{\mathbf{s}}_{n/n} = \sum_{i=0}^{n} H_{n,i}\,\mathbf{z}_i \tag{6}$$

The "filter gains" $H_{n,i}$ are to be chosen so that the mean-square error is minimized; that is, choose the $H_{n,i}$ in such a way that

$$M_n = E[(\mathbf{s}_n - \hat{\mathbf{s}}_{n/n})^T(\mathbf{s}_n - \hat{\mathbf{s}}_{n/n})] \tag{7}$$

is minimized. It is well known[11] that a necessary and sufficient condition for $\hat{\mathbf{s}}_{n/n}$ to minimize M_n is that the error in the estimate $(\tilde{\mathbf{s}}_{n/n} \triangleq \mathbf{s}_n - \hat{\mathbf{s}}_{n/n})$ be orthogonal to the measurement data

$$E[\tilde{\mathbf{s}}_{n/n}\mathbf{z}_i{}^T] = 0 \qquad i = 0, 1, \cdots, n \tag{8}$$

This is the Wiener–Hopf equation, which is frequently written as

$$E[\mathbf{s}_n\mathbf{z}_i{}^T] = \sum_{j=0}^{n} H_{n,j}E[\mathbf{z}_j\mathbf{z}_i{}^T] \qquad i = 0, 1, \cdots, n \tag{9}$$

This equation must be solved for the $H_{n,j}$ in order to obtain the gains of the optimal filter. One can rewrite this as a vector-matrix equation whose solution, theoretically speaking, is straightforward. However, the matrix inversion that is required becomes computationally impractical when n is large. Wiener and Kolmogorov assumed an infinite amount of data (that is, the lower limit of the summation is $-\infty$ rather than zero), and assumed the system to be stationary. The resulting equations were solved using spectral factorization.[9,10,12]

The problem formulated and described here is significantly different from Gauss' least-squares problem. First, no assumption is imposed that the signal is constant. Instead, the signal can be different at each n but can be described statistically by the autocorrelation and cross-correlation functions of the signal and measurement data. Second, instead of arguing that the estimate be the most probable, a probabilistic version of the least-squares method is chosen as the performance index.

It has been found that Eq. (9) is solved in a relatively straightforward manner if one introduces a "shaping filter"[13,14] to give a more explicit description of the signal. In particular, suppose that the signal and measurement processes are assumed to have the following structure. The measurements are described by

$$\begin{aligned} \mathbf{z}_i &= \mathbf{s}_i + \mathbf{v}_i \\ &= H_i\mathbf{x}_i + \mathbf{v}_i \end{aligned} \tag{10}$$

where $\mathbf{v}_i$ is a white-noise sequence (that is, $\mathbf{v}_i$ is both mutually independent and independent of $\mathbf{x}_i$). The system state vector $\mathbf{x}_i$ is assumed to be described as a dynamic system having the form

$$\mathbf{x}_{i+1} = \Phi_{i+1,i}\mathbf{x}_i + \mathbf{w}_i \tag{11}$$

where $\mathbf{w}_i$ represents a white-noise sequence. Note that if the noise $\mathbf{w}_i$ is identically zero and if $\Phi_{i+1,i}$ is the identity matrix, then the state is a constant for all i and one has returned basically to the system assumed by Gauss. With the system described by Eqs. (10) and (11), the known statistics for the initial state $\mathbf{x}_0$, and the noise sequences $\{\mathbf{w}_i\}$ and $\{\mathbf{v}_i\}$, one can proceed to the solution of Eq. (9).

Although the weighting function for the filter can be determined, a new solution must be generated for each n. It seems intuitively reasonable that estimates of $\mathbf{s}_{n+1}$ (or $\mathbf{x}_{n+1}$) could be derived, given a new measurement $\mathbf{z}_{n+1}$, from $\hat{\mathbf{s}}_{n/n}$ and $\mathbf{z}_{n+1}$ rather than from $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_n, \mathbf{z}_{n+1}$, since $\hat{\mathbf{s}}_{n/n}$ is based on the data $(\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_n)$. In 1955 J. W. Follin[15] at Johns Hopkins University suggested a recursive approach based on this idea, which he carried out for a specific system. This approach had immediate appeal and essentially laid the foundation for the developments that are now referred to as the Kalman filter. It is clear (for example, see Ref. 16, p. 129) that Follin's work provided a direct stimulus for the work of Richard Bucy, which led to his subsequent collaboration with Kalman in the continuous-time version of the filter equations.[17] As frequently happens, the time was ripe for this approach, because several other people independently investigated recursive filter and prediction methods; see, for example, Refs. 18 and 19. Also, the method of stochastic approximation[20] was introduced and being studied for related problems[21] during this period.

Kalman published his first paper on discrete-time, recursive mean-square filtering in 1960.[2] It is interesting to note that, analogous to the Gauss–Legendre squabble concerning priority of the least-squares method, there is a difference of opinion concerning the originator of the Kalman filter. Peter Swerling published a RAND Corporation memorandum in 1958[18] describing a recursive procedure for orbit determination. Of further interest is the fact that orbit determination problems provided the stimulus for both Gauss' work and more modern-day developments. Swerling's method is essentially the same as Kalman's except that the equation used to update the error covariance matrix has a slightly more cumbersome form. After Kalman had published his paper and it had attained considerable fame, Swerling[22] wrote a letter to the *AIAA Journal* claiming priority for the Kalman filter equations. It appears, however, that his plea has fallen on deaf ears.

The developments beginning with Wiener's work and culminating with Kalman's reflect fundamentally the changes that have occurred in control systems theory during this period. In the "classical control theory," the emphasis was on the analysis and synthesis of systems in terms of their input–output characteristics. The basic tools used for these problems were the Laplace and Fourier transforms. The original formulation and solution of the Wiener–Kolmogorov filtering problem is consistent with this basic approach. More recent developments have stressed the "state-space" approach, in which one deals with the basic system that gives rise to the observed output. It represents in many ways a return to Gauss' approach, since he referred to the dynamic modeling problem as noted earlier. Also, the state-space approach makes use of difference and differential equations rather than the integral equations of the classical approach. Although the two approaches are mathematically equivalent, it seems to be more satisfying to work with differential equations (probably since dynamical systems are generally described in this manner).

At this point, let us summarize the Kalman filtering problem and its solution. The system that is considered is composed of two essential ingredients. First, the state is assumed to be described by

$$\mathbf{x}_{k+1} = \Phi_{k+1,k}\,\mathbf{x}_k + \mathbf{w}_k \qquad (11')$$

and the measurement data are related to the state by

$$\mathbf{z}_k = H_k\mathbf{x}_k + \mathbf{v}_k \qquad (10')$$

where $\{\mathbf{w}_k\}$ and $\{\mathbf{v}_k\}$ represent independent white-noise sequences. The initial state $\mathbf{x}_0$ has a mean value $\hat{\mathbf{x}}_{0/-1}$ and covariance matrix $P_{0/-1}$ and is independent of the plant and measurement noise sequences. The noise sequences have zero mean and second-order statistics described by

$$E[\mathbf{v}_k\mathbf{v}_j^T] = R_k\delta_{kj} \qquad E[\mathbf{w}_k\mathbf{w}_j^T] = Q_k\delta_{kj}$$
$$E[\mathbf{v}_k\mathbf{w}_j^T] = 0 \qquad \text{for all } k, j$$

where δ_{kj} is the Kronecker delta.

An estimate $\hat{\mathbf{x}}_{k/k}$ of the state $\mathbf{x}_k$ is to be computed from the data $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_k$ so as to minimize the mean-square error in the estimate. The estimate that accomplishes this is to be computed as an explicit function only of the measurement $\mathbf{z}_k$ and the previous best estimate $\hat{\mathbf{x}}_{k-1/k-1}$. This approach leads to a recursive solution that provides an estimate that is equivalent to the estimate obtained by processing all of the data simultaneously but reduces the data-handling requirements. The estimate of the signal

$$\mathbf{s}_k = H_k\mathbf{x}_k \qquad (12)$$

is given by

$$\hat{\mathbf{s}}_{k/k} = H_k\hat{\mathbf{x}}_{k/k} \qquad (13)$$

The solution of this recursive, linear, mean-square estimation problem can be determined from the orthogonality principle given in Eq. (8), as well as in a large variety of other ways, and is presented below. This system of equations has come to be known as the Kalman filter. The estimate is given as the linear combination of the estimate predicted in the absence of new data, or

$$\hat{\mathbf{x}}_{k/k-1} = \Phi_{k,k-1}\hat{\mathbf{x}}_{k-1/k-1}$$

and the residual $\mathbf{r}_k$. Thus, the mean-square estimate is

$$\hat{\mathbf{x}}_{k/k} = \Phi_{k,k-1}\hat{\mathbf{x}}_{k-1/k-1} + K_k[\mathbf{z}_k - H_k\Phi_{k,k-1}\hat{\mathbf{x}}_{k-1/k-1}] \qquad (14)$$

The gain matrix K_k can be considered as being chosen to minimize $E[(\hat{\mathbf{x}}_k - \hat{\mathbf{x}}_{k/k})^T(\mathbf{x}_k - \mathbf{x}_{k/k})]$ and is given by

$$K_k = P_{k/k-1}H_k^T(H_kP_{k/k-1}H_k^T + R_k)^{-1} \qquad (15)$$

The matrix $P_{k/k-1}$ is the covariance of the error in the predicted estimate and is given by

$$P_{k/k-1} = E[(\mathbf{x}_k - \hat{\mathbf{x}}_{k/k-1})(\mathbf{x}_k - \hat{\mathbf{x}}_{k/k-1})^T] \qquad (16)$$

$$= \Phi_{k/k-1}P_{k-1/k-1}\Phi_{k,k-1}^T + Q_{k-1} \qquad (17)$$

The $P_{k/k}$ is the covariance of the error in the estimate $\hat{\mathbf{x}}_{k/k}$.

$$P_{k/k} = E[\mathbf{x}_k - \hat{\mathbf{x}}_{k/k})(\mathbf{x}_k - \hat{\mathbf{x}}_{k/k})^T] \qquad (18)$$

$$= P_{k/k-1} - K_kH_kP_{k/k-1} \qquad (19)$$

Equations (14), (15), (17), and (19) form the system of equations comprising the Kalman filter.[2,6]

Kalman filter theory—a perspective

Let us relate elements of this problem to Gauss' earlier arguments. First, Kalman assumes that the noise is independent from one sampling time to the next. But it is clear from Eq. (5) that this is equivalent to assuming that the residual $(\mathbf{z}_k - H_k\mathbf{x}_k)$ is independent between sampling times and therefore agrees with Gauss' assumption. Next, the noise and initial state are essentially assumed by Kalman to be Gaussian. The linearity of the system causes the state and measurements to be Gaussian at each sampling time. Thus the residual is Gaussian, as Gauss assumed. Therefore, one sees that the basic assumptions of Gauss and Kalman are identical except that the latter allows the state to change from one time to the next. This difference introduces a nontrivial modification to Gauss' problem but one that can be treated within a least-squares framework if the noise $\{\mathbf{w}_k\}$ is considered as the error in the plant model at each stage. In particular, one can formulate the least-squares problem as that of choosing the estimates $\hat{\mathbf{x}}_{k/k}$ and the plant errors $\mathbf{w}_k$ to minimize the modified least-squares performance index.

$$L_n = \frac{1}{2}(\mathbf{x}_0 - \mathbf{a})^T M_0^{-1}(\mathbf{x}_0 - a) + \frac{1}{2}\sum_{i=0}^{n}(\mathbf{z}_i - H_i\mathbf{x}_i)^T R_i^{-1}(\mathbf{z}_i - H_i\mathbf{x}_i) + \frac{1}{2}\sum_{i=0}^{n-1}\mathbf{w}_i^T Q_i^{-1}\mathbf{w}_i \quad (20)$$

subject to the constraint

$$\mathbf{x}_k = \Phi_{k,k-1}\mathbf{x}_{k-1} + \mathbf{w}_{k-1} \quad (21)$$

Note that the first term essentially describes the uncertainty in the initial state. If one has no a priori information, then M_0^{-1} is identically zero and the term vanishes. Similarly, if there is no error in the plant equation, Q_i^{-1} is identically zero so this term vanishes. Then Eq. (20) is seen to reduce to Gauss' least-squares problem, as given in Eq. (3). The weighting matrices M_0^{-1}, R_i^{-1}, and Q_i^{-1} represent the matrix inverses of the a priori covariance matrices if a probabilistic interpretation is desired.

One can obtain a recursive solution to the problem of minimizing (20) by noting that

$$L_n = L_{n-1} + \frac{1}{2}(\mathbf{z}_n - H_n\mathbf{x}_n)^T R_n^{-1}(\mathbf{z}_n - H_n\mathbf{x}_n) + \frac{1}{2}\mathbf{w}_{n-1}^T Q_{n-1}^{-1}\mathbf{w}_{n-1} \quad (22)$$

and by then proceeding inductively starting with $n = 0$ to obtain recursion relations for the least-square estimate.[23] If this is done, the Kalman filter equations are obtained. It is then indicated that, for this linear problem, deterministic least-squares estimation theory and the probabilistically based mean-square estimation theory are equivalent. Further, the problem of minimizing Eq. (22) is seen to give the most probable estimate for this system.

Since the Kalman filter represents essentially a recursive solution of Gauss' original least-squares problem, it is reasonable to consider the substance of Kalman's contribution and attempt to put it into perspective. It cannot be denied that there has been a substantial contribution if for no other reason than the large number of theoretical and practical studies that it has initiated. I suggest that the contribution is significant for two basic reasons:

1. The Kalman filter equations provide an extremely convenient procedure for digital computer implementation. One can develop a computer program using the Kalman filter in a direct manner that (initially, at least) requires little understanding of the theory that led to their development. There are well-established numerical procedures for solving differential equations, so the engineer does not have to be worried about this problem. By contrast, the solution of the Wiener–Hopf equation and the implementation of the Wiener–Kolmogorov filter must be regarded as more difficult or there would have been no need for the Kalman filter. Since Gauss was very concerned with the computational aspects of least-squares applications, one can imagine that he would appreciate the computational benefits of the Kalman filter.

2. Kalman posed the problem in a general framework that has had a unifying influence on known results. Further, one can analyze the behavior of the estimates within the general framework and thereby obtain significant insights into the results obtained from computational studies. There has been a veritable "explosion" of theoretical papers that have recognizable roots in Kalman's work and thereby testify to the richness of his formulation.

Finally, a third reason for the popularity might be considered, although it is less tangible in character than the other two. It is worth noting that Kalman recognized the potential of his results, whereas others working in the area either did not or were not as successful in communicating the intrinsic worth of recursive filtering to others. One cannot overemphasize the value of recognizing and successfully communicating significant new results.

The Kalman filter, which assumes linear systems, has found its greatest application to nonlinear systems. It is generally used in these problems by assuming knowledge of an approximate solution (as Gauss proposed) and by describing the deviations from the reference by linear equations. The approximate linear model that is obtained forms the basis for the Kalman filter utilization. Commonly, such applications are accomplished with great success but, on occasion, unsatisfactory results are obtained because a phenomenon known as divergence occurs.[24,25]

Divergence is said to occur when the error covariance matrix P_k computed by the filter becomes unjustifiably small compared with the actual error in the estimate. When P_k becomes small, it causes the gain matrix to become too small and new measurement data are given too little weight. As a result, the plant model becomes more important in determining the estimate than are the data and any errors in the model can build up over a period of time and cause a significant degradation in the accuracy of the estimate. This happens most commonly when the plant is assumed to be error-free (i.e., $\mathbf{w}_k \equiv 0$ for all k). If the model were perfect and contained no random or model errors, then the vanishing of the error covariance matrix would be desirable and would represent the fact that the state could be determined precisely if sufficient redundant data were processed. However, it is naive at best to assume that any physical system can be modeled precisely, so it is necessary to account for model errors. Thus, it has become good practice[6] always to include the plant error or noise term $\mathbf{w}_k$. It should be emphasized that divergence does not occur because of any fault of the filter. If the system were actually linear, Kalman[4] showed that the filter equations are stable under very reasonable conditions. Thus, the divergence is a direct consequence of the errors introduced by the linear approximation.

To reduce approximation errors, the so-called "extended Kalman filter" is generally used in practice. In this case the nonlinear system is linearized by employing the best estimates of the state vector as the reference values used at each stage for the linearization. For example, at time t_{k-1}, the estimate $\hat{\mathbf{x}}_{k-1/k-1}$ is used as the reference in obtaining the transition matrix $\Phi_{k,k-1}$. This approximation is utilized in Eq. (17) to obtain the Kalman error covariance $P_{k/k-1}$. The estimate is given by

$$\hat{\mathbf{x}}_{k/k-1} = \mathbf{f}_k(\hat{\mathbf{x}}_{k-1/k-1}) \quad (23)$$

where $\mathbf{f}_k$ is used to denote the nonlinear plant equation.

In most cases it is obtained as the solution of an ordinary differential equation that describes the plant behavior. The processing of the data obtained at t_k is accomplished in a similar manner. The estimate $\hat{\mathbf{x}}_{k/k-1}$ serves as the reference for the determination of a linear approximation H_k to the nonlinear measurement equation. The matrix H_k is used in Eqs. (15) and (19) to determine the gain and error covariance matrices K_k and $P_{k/k}$. The filtered estimate is then given by

$$\hat{\mathbf{x}}_{k/k} = \hat{\mathbf{x}}_{k/k-1} + K_k[\mathbf{z}_k - \mathbf{h}_k(\hat{\mathbf{x}}_{k/k-1})] \tag{24}$$

where $\mathbf{h}_k$ is used to denote the measurement nonlinearity; that is, the measurement is assumed to be described by

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k) + \mathbf{v}_k \tag{25}$$

Through the use of the extended Kalman filter, one can hope to eliminate or reduce divergence. Note, however, that the $P_{k/k-1}$ and $P_{k/k}$ matrices are still linear approximations of the true error covariance matrices. Further, the nonlinear models $\mathbf{f}_k$ and $\mathbf{h}_k$ are themselves approximations of the actual physical system, so modeling errors can still exist. Thus, the extended Kalman filter does not insure the elimination of the divergence problem.

In a practical application one does not know the error in the state estimate, so there are grounds for uneasiness in using this method. Of course, the same type of problem must be considered in any least-squares application and we can return to Gauss for the means of judging the behavior of the filter. Recall that he said that the estimates should satisfy all the observations in the most accurate manner possible. Thus, one is led to further consideration of the residuals as a measure of filter performance. Kailath[26] pointed out recently that the residual sequence $(\mathbf{z}_k - H_k\hat{\mathbf{x}}_{k/k-1})$ is a white-noise sequence. Since the residual can be computed explicitly, it can be examined at each stage to verify that the residual (or innovations) sequence has the appropriate statistical characteristics. A method of controlling the divergence, based on the residuals, has been proposed[27,28] in which the plant and measurement noise covariance matrices Q_k and R_k are chosen in a manner that is appropriate to cause the residuals to have the desired properties. But this method is essentially the same as choosing the least-squares weighting matrices as a reflection of the accuracy of the measurements (or plant). Thus, the least-squares aspect continues to dominate the practical application of the method.

REFERENCES

1. Gauss, K. G., *Theory of Motion of the Heavenly Bodies.* New York: Dover, 1963.
2. Kalman, R. E., "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82D, pp. 35–45, Mar. 1960.
3. Bell, E. T., *Men of Mathematics.* New York: Simon and Schuster, 1961.
4. Kalman, R. E., "New methods in Wiener filtering theory," in *Proc. First Symp. of Engineering Application of Random Function Theory and Probability.* New York: Wiley, 1963, pp. 270–388.
5. Sorenson, H. W., "Controllability and observability of linear, stochastic, time-discrete control systems," in *Advances in Control Systems,* Vol. 6, C. T. Leondes, ed. New York: Academic Press, 1968.
6. Sorenson, H. W., "Kalman filtering techniques," in *Advances in Control Systems,* Vol. 3, C. T. Leondes, ed. New York: Academic Press, 1966.
7. Fisher, R. A., "On an absolute criterion for fitting frequency curves," *Messenger of Math.*, vol. 41, p. 155, 1912.
8. Berkson, J., "Estimation by least-squares and by maximum likelihood," *Proc. Third Berkeley Symp. on Mathematical Statistics, and Probability*, Vol. 1. Berkeley: University of California Press, 1956, pp. 1–11.
9. Kolmogorov, A. N., "Interpolation and extrapolation of stationary random sequences," translated by W. Doyle and J. Selin, Rept. RM-3090-PR, RAND Corp., Santa Monica, Calif., 1962.
10. Wiener, N., *The Extrapolation, Interpolation and Smoothing of Stationary Time Series.* New York: Wiley, 1949.
11. Sorenson, H. W., and Stubberud, A. R., "Linear estimation theory," Chapter 1 in *Theory and Application of Kalman Filtering; AGARDograph*, no. 139, 1970.
12. Yaglom, A. M., *An Introduction to the Theory of Stationary Random Functions*, translated by R. A. Silverman. Englewood Cliffs, N.J.: Prentice-Hall, 1962.
13. Bode, H. W., and Shannon C. E., "A simplified derivation of linear least-squares smoothing and prediction theory," *Proc. IRE*, vol. 38, pp. 417–425, Apr. 1950.
14. Zadeh, L. A., and Ragazzini, J. R., "An extension of Wiener's theory of prediction," *J. Appl. Phys.*, vol. 21, pp. 645–655, July 1950.
15. Carlton, A. G., and Follin, J. W., "Recent developments in fixed and adaptive filtering," *AGARDograph*, no. 21, 1956.
16. Bucy, R. S., and Joseph, P. D., *Filtering for Stochastic Processes with Applications to Guidance.* New York: Interscience, 1968.
17. Kalman, R. E., and Bucy, R. S., "New results in linear filtering and prediction theory," *J. Basic Eng.*, vol. 83D, pp. 95–108, 1961.
18. Swerling, P., "A proposed stagewise differential correction procedure for satellite tracking and prediction," Rept. P-1292, *RAND Corp.* Santa Monica, Calif., Jan. 1958; also published in *J. Astronaut. Sci.*, vol. 6, 1959.
19. Blum, M., "Recursion formulas for growing memory digital filters," *IRE Trans. Information Theory*, vol. IT-4, pp. 24–30, Mar. 1958.
20. Robbins, H., and Munro, S., "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–407, 1951.
21. Kiefer, J., and Wolfowitz, J., "Stochastic estimation of the maximum of a regression function," *Ann. Math. Stat.*, vol. 23, pp. 462–466, 1952.
22. Swerling, P., "Comment on 'A statistical optimizing navigation procedure for space flight,'" *AIAA J.*, vol. 1, p. 1968, Aug. 1963.
23. Sorenson, H. W., "Comparison of Kalman, Bayesian and maximum likelihood estimation techniques," Chapter 6 in *Theory and Application of Kalman Filtering; AGARDograph*, no. 139, 1970.
24. Schlee, F. H., Standish, C. J., and Toda, N. F., "Divergence in the Kalman filter," *AIAA J.*, vol. 5, pp. 1114–1120, June 1967.
25. Fitzgerald, R. J., "Error divergence in optimal filtering problems," presented at the 2nd IFAC Symposium on Automatic Control in Space, Vienna, Austria, Sept. 1967.
26. Kailath, T., "An innovations approach to least-squares estimation—Part I: linear filtering in additive white noise," *IEEE Trans. Automatic Control*, vol. AC-13, pp. 646–655, Dec. 1968.
27. Jazwinski, A. H., "Adaptive filtering," *Automatica*, vol. 5, pp. 475–485, 1969.
28. Mehra, R. K., "On the identification of variances and adaptive filtering," *Preprints 1969 Joint Automatic Control Conf.*, pp. 494–505.

Harold W. Sorenson (M) was born in Omaha, Nebr., in 1936. He received the B.S. degree in aeronautical engineering from Iowa State University in 1957 and the M.S. and Ph.D. degrees in engineering from the University of California, Los Angeles, in 1963 and 1966, respectively. He worked for General Dynamics/Astronautics from 1957 to 1962 and for the AC Electronics Division, General Motors Corporation, from 1963 to 1966. After a year as guest scientist at the Institut für Luft- und Raumfahrt, Oberpfaffenhofen, West Germany, he joined the faculty of the University of California, San Diego, where he is now an assistant professor of aerospace engineering. His research activities have been centered on linear and nonlinear stochastic control system problems.

At the crossroads in air-traffic control

II. The view from the ground—instrument-landing systems, automated ground control, en route automation, introduction to collision-avoidance systems, positive-control airspace, improved communications, STOLports, and terminal navigational aids

The urgent need for greater safety in handling flight operations has spurred the acceleration of programs for the installation of numerous automated control systems and other electronic equipment

Gordon D. Friedlander Staff Writer

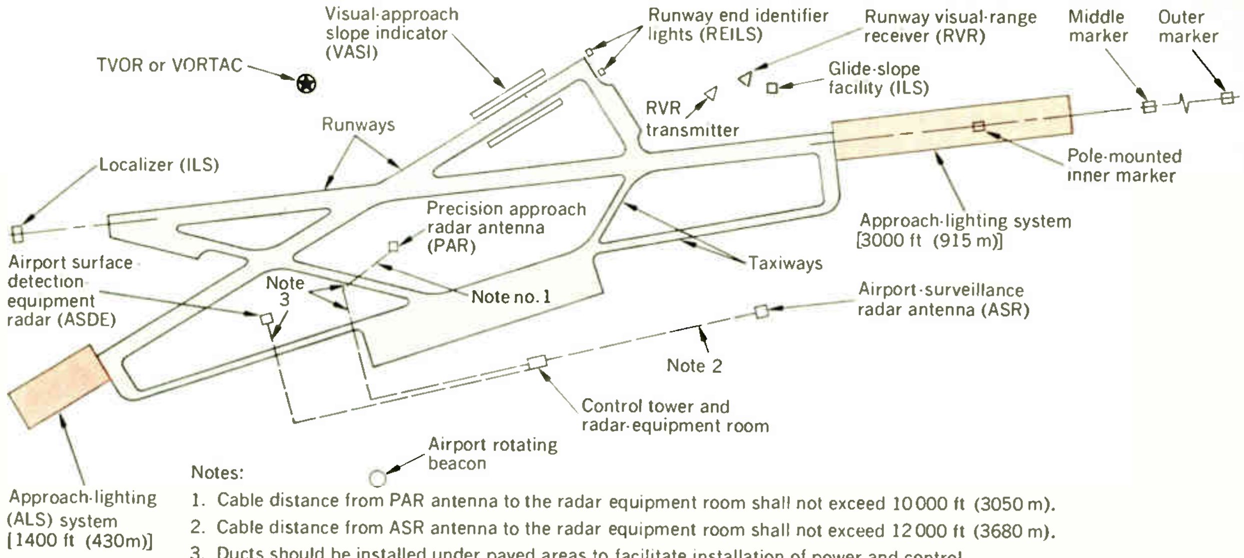
Last February, the U.S. Congress authorized the FAA to spend initially $2.5 billion for the improvement of ATC equipment. This sum reportedly includes $1.7 billion for new electronic systems. The appropriation is part of a $5 billion trust fund package for airways and airport development that is now being hammered out by a joint Senate and House conference committee. A large percentage of the total appropriation will be expended on additional en route ATC sectors, terminal area automation, en route navigational systems, and improved UHF and VHF communications channels. In addition, collision-avoidance systems (CAS), instrument-landing systems (ILS), and automatic weather-sensing equipment will either be further developed or installed. The first installment of this three-part series dealt primarily with the functions of the ARTS installations and the Common IFR Room in the New York terminal area. In this piece, we will examine other systems and aspects of the FAA's plans for the present decade.

Back in 1929, Lt. Gen. James Doolittle, USAF (Ret.), then a civilian pilot, made a blind landing at Mitchell Field. But today—more than 40 years later—civil aviation aircraft cannot land under such conditions.[1] A fair percentage of civil aircraft have been approved for ***Category I**** (200-foot, or 61-meter, ceiling decision height; 2400-foot, or 730-meter, forward visibility) airport approaches, but a much smaller percentage have been approved for ***Category II*** (100-foot, or 30-meter, ceiling decision height; 1200-foot, or 365-meter, forward visibility) landings. And although considerable advances were made during the 1960s toward the development of operational zero–zero (Category III) landing systems, the objective has not yet been attained for commercial air carriers.

As of now, only eight airports are authorized for Category II operations, which require runways equipped with FAA-approved ILS ground-guidance lighting and transmissometers (both of which will be discussed later in this installment).

---

* As explained in Part I of this series, some of the terms or acronyms used by the FAA may be unfamiliar to the reader. Thus a glossary is provided on p. 83. The term, when initially used in this installment, will be italicized in boldface to indicate its inclusion in the glossary.



Notes:

1. Cable distance from PAR antenna to the radar equipment room shall not exceed 10 000 ft (3050 m).
2. Cable distance from ASR antenna to the radar equipment room shall not exceed 12 000 ft (3680 m).
3. Ducts should be installed under paved areas to facilitate installation of power and control cable to ATC & N facilities.
4. Distance measuring equipment (DME) may be colocated at the ILS glide-slope site,where operationally required.

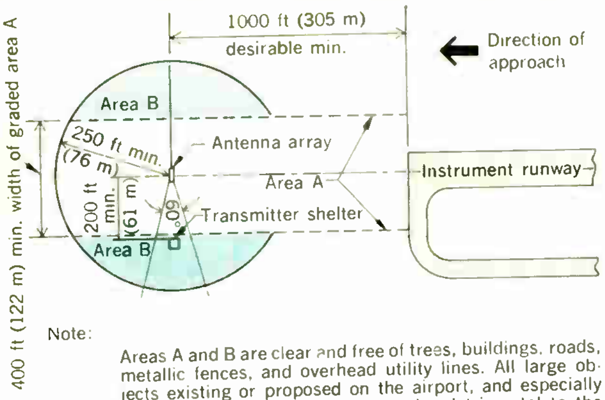
FIGURE 1. Typical location plan for terminal air-traffic control and navigational facilities.

FIGURE 2. Pilot's view of a landing approach on an illuminated instrument-landing system (ILS) runway at the Washington (D.C.) National Airport.

Instrument-landing system for precision instrument approaches

There are approximately 280 instrument-landing systems presently in service at airports in the U.S. An additional 123 ILSs are programmed for installation by the end of the current fiscal year, and a total of 1355 systems (FAA, plus military) are planned by the end of 1979.

The purpose of the ILS in a ***precision approach procedure*** is to provide electronic-instrument guidance to the pilot so that a properly equipped aircraft (a description of the on-board instrumentation will be given in Part III of this series) can attain the exact alignment and angle of descent on its final approach for landing.[2] Figure 1 shows the facilities required at an ILS-equipped airport. These include a localizer, glide-slope facility, and outer- and middle-marker beacons. Other components normally associated and used in conjunction with the system are a compass locator (COMLO), placed at the outer-marker site; an approach lighting system; and other visual aids. In addition, at certain airports, distance-measuring equipment (DME) is installed when operational requirements indicate its inclusion. Figure 2 is an aerial photo of an ILS run-



Note:

Areas A and B are clear and free of trees, buildings, roads, metallic fences, and overhead utility lines. All large objects existing or proposed on the airport, and especially adjacent to the ILS runway, may be detrimental to the operation of the localizer, and should be reported to the FAA.

FIGURE 3. Layout plan for a typical ILS localizer.

way and ancillary ***terminal navigational aids*** (NAVAIDS) at the Washington National Airport.

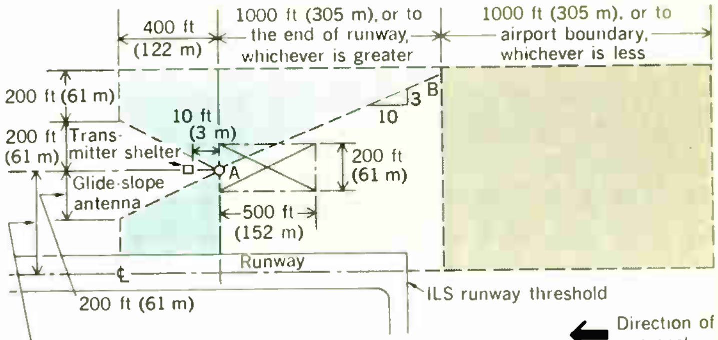
The site facilities criterions indicated in Fig. 1 are applicable to both Category I and Category II operations. The FAA recommends that such ILS facilities be installed at sites that will not be affected, for at least five years, by further runway, taxiway, or building construction that could necessitate the relocation of the system.

System components and related equipment. To give the reader a comprehensive idea of the functions and interaction of the various ILS components and equipment shown in the Fig. 1 line diagram, let's start with the—

*ILS localizer.** The localizer facility is used to provide alignment or lateral course guidance for landing an aircraft on an ILS runway. The major elements of the equipment are the antenna array, transmitting gear, monitor field detectors, and the transmitter shelter. A typical installation is shown in the Fig. 3 sketch. The present standard antenna system authorized for FAA procurement is the V-ring array.†

The location of the localizer's antenna array is of prime importance; thus it is generally placed on the extended centerline of the runway (at the opposite end of the direction of approaching aircraft). The array should be sufficiently beyond the runway to ensure that the specifications stipulated in note 1 of Fig. 3 will be met. The illustration also shows the siting criterions for the transmitter shelter, which may be placed on either side of the antenna array, depending upon local terrain conditions, access roads, or clearance requirements from other existing facilities. The antenna array may be elevated to provide better signals in the approach zone.

Glide-slope facility. The principal elements of the ILS glide-slope (or glide-path) facility are



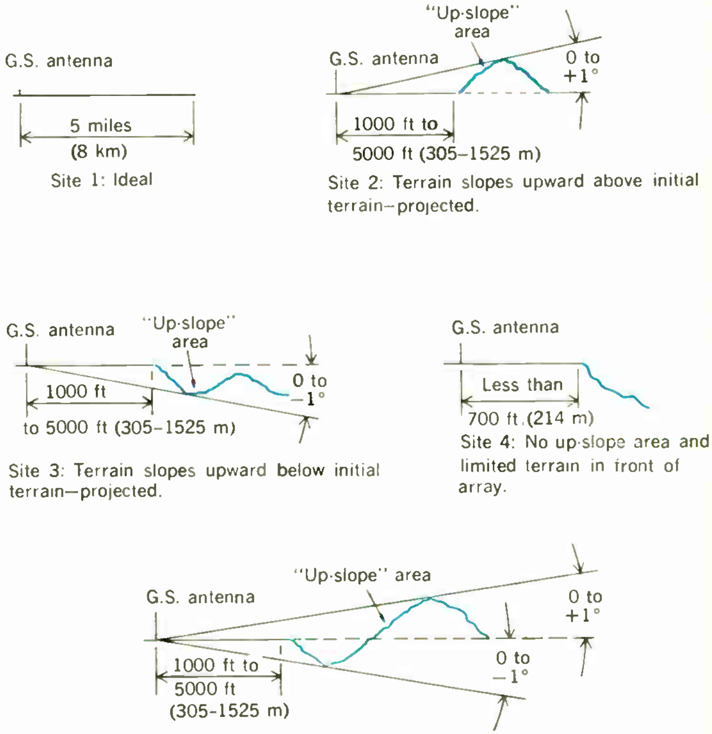
D = 250 ft (76 m) to 650 ft (198 m) for Category I–ILS
D = 400 ft (122 m) to 650 ft (198 m) for Category II–ILS

Notes:

1. It is desirable that the glide-slope transmitter, antenna, and associated equipment be installed on the runway side that is free of taxiways, intersecting runways, aprons, or where the least number of aircraft would be taxiing or warming up on the approach side of the facility.
2. All indicated shaded areas shall be clear of all objects such as buildings, trees, roads, metallic fences, overhead utility lines, and similar above-ground construction.
3. 🟦 This marked area, in addition to the requirements of note 2, must be clear of taxiways, runways, aprons, etc., to ensure monitor system stability.
4. ▨ This area shall be clear of all objects as given in note 2, with the exception of airport roads, which are under airport control.

FIGURE 4. Layout plan and specifications for a typical ILS glide-slope installation.

FIGURE 5. Sketches (in elevation) showing dimensional and angular criterions for glide-slope systems installed under various site terrain conditions.



Site 1: Ideal

Site 2: Terrain slopes upward above initial terrain—projected.

Site 3: Terrain slopes upward below initial terrain—projected.

Site 4: No up-slope area and limited terrain in front of array.

Site 5: Combination of sites 2 and 3

	Site 1	Site 2	Site 3	Site 4	Site 5
Approximate percentage of total sites	5	20	15	15	45
Null reference	all	very few	few	none	very few
Sideband reference	all	few	many	very few	few
Capture effect	all	most	most	none	most

---

* Localizers are assigned on 20 channels at odd tenths of a megahertz in the 108–112-MHz band. The course position is established by the sideband and carrier radiation patterns superimposed in space to produce 150- and 90-Hz components in the aircraft receivers for crosspointer "fly left–fly right" indications, respectively. The conventional localizer employs a 200-watt transmitter.

† The V-ring antenna installation, since it is directive, offers a reduction in course bends caused by reflections, and reduced power requirements for a given field intensity.

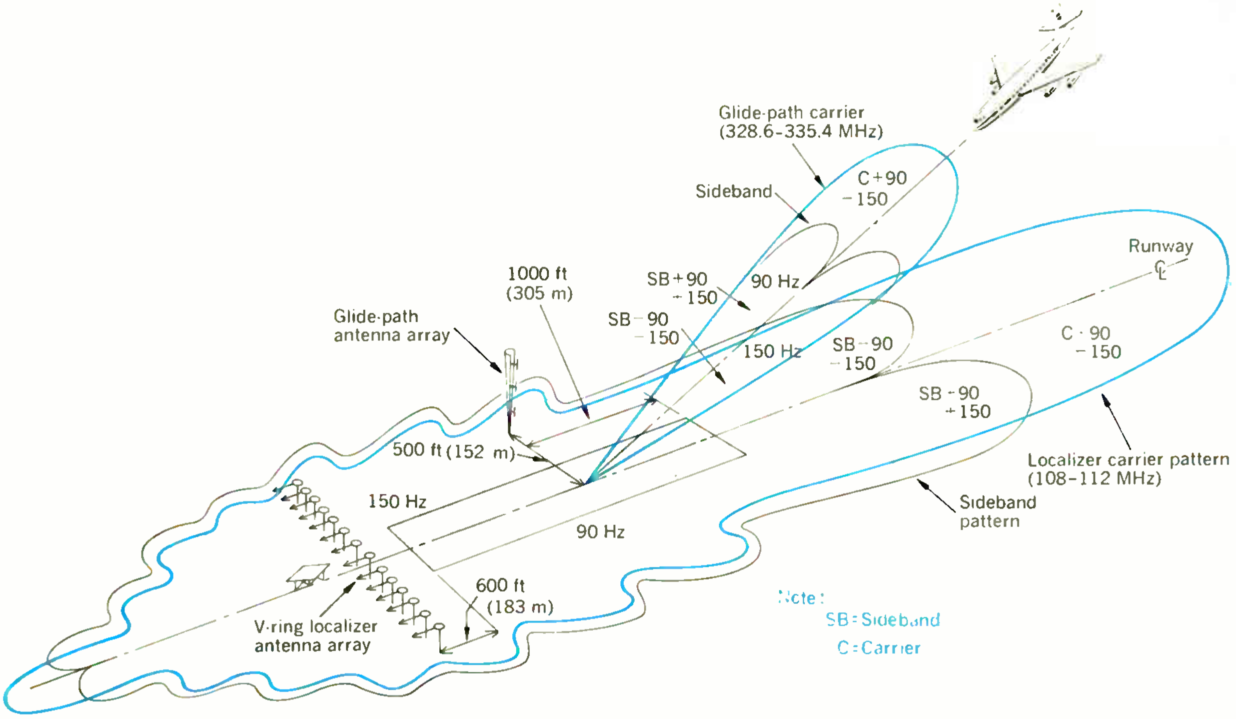


FIGURE 6. Carrier and sideband radiation patterns for both glide-slope and localizer facilities along a typical ILS-equipped runway.

1. The transmitting antenna mast.
2. A transmitter shelter.
3. A monitor antenna array.

The null reference is the basic type of glide-slope system; the sideband references and the capture-effect system are modifications made necessary by undesirable site conditions (see Fig. 4).

The glide-slope equipment is installed alongside the ILS runway, usually 500 feet (153 meters)—but not more than 1200 feet (366 meters)—inward from the landing strip threshold. Dimensional siting criterions, and layout of the antenna system, transmitter equipment shelter, and monitor equipment areas are shown in Figs. 4 and 5.

Glide-slope facilities use authorized channels[3] assigned at 0.3-MHz intervals in the 328.6- to 335.4-MHz band.

In the conventional null-reference facility, sideband radiation from the upper, horizontally polarized antenna produces a series of vertical lobes, phased alternately so that the aircraft detects 150/90-Hz "fly up–fly down" signals whenever the plane is below or above a nominal 2.5 degrees null on-path, respectively. Carrier power is fed to a lower antenna to produce a vertical lobe that is superimposed about the two lower-sideband lobes, thereby effectively reinforcing and canceling fields to generate the path-guidance components.

Figure 6 shows the carrier and sideband radiation patterns for both the glide-slope and localizer (with V-ring antenna array) facilities along a typical ILS-equipped runway.

Marker beacons. Table I summarizes the criterions

I. ILS marker beacon data and location criterions

Type	Function	Nominal Location	Location Tolerances	Operational Requirement
Outer marker	Marks intercept point of glide-slope and minimum-holding altitude	The intercept point	±800 feet (244 meters) lateral and longitudinal	Required for Category I and Category II ILS locations
Middle marker	Marks decision-height point for Category I operations	The decision-height point	±300 feet (91 meters) lateral and ±500 (152 meters) feet longitudinal	Required for Category I and Category II ILS locations
Inner marker (pole-mounted)	Marks decision-height point for Category II operations	See note no. 1	±25 feet (7.5 meters) lateral and longitudinal	Required only for Category II ILS locations

1. ILS inner marker antenna is located at the point where glide-slope elevation is 100 feet (30.5 meters) above the elevation of the highest point in 3000-ft (914-meter) runway touchdown zone.
2. Radar (ASR and PAR) and DME may be used as a substitute for outer marker under certain circumstances.
3. The outer marker may be approximately 4 to 7 miles (6.4–11.3 km) from the ILS landing threshold, while the middle marker may be approximately 3500 feet (1070 meters) from the same threshold.

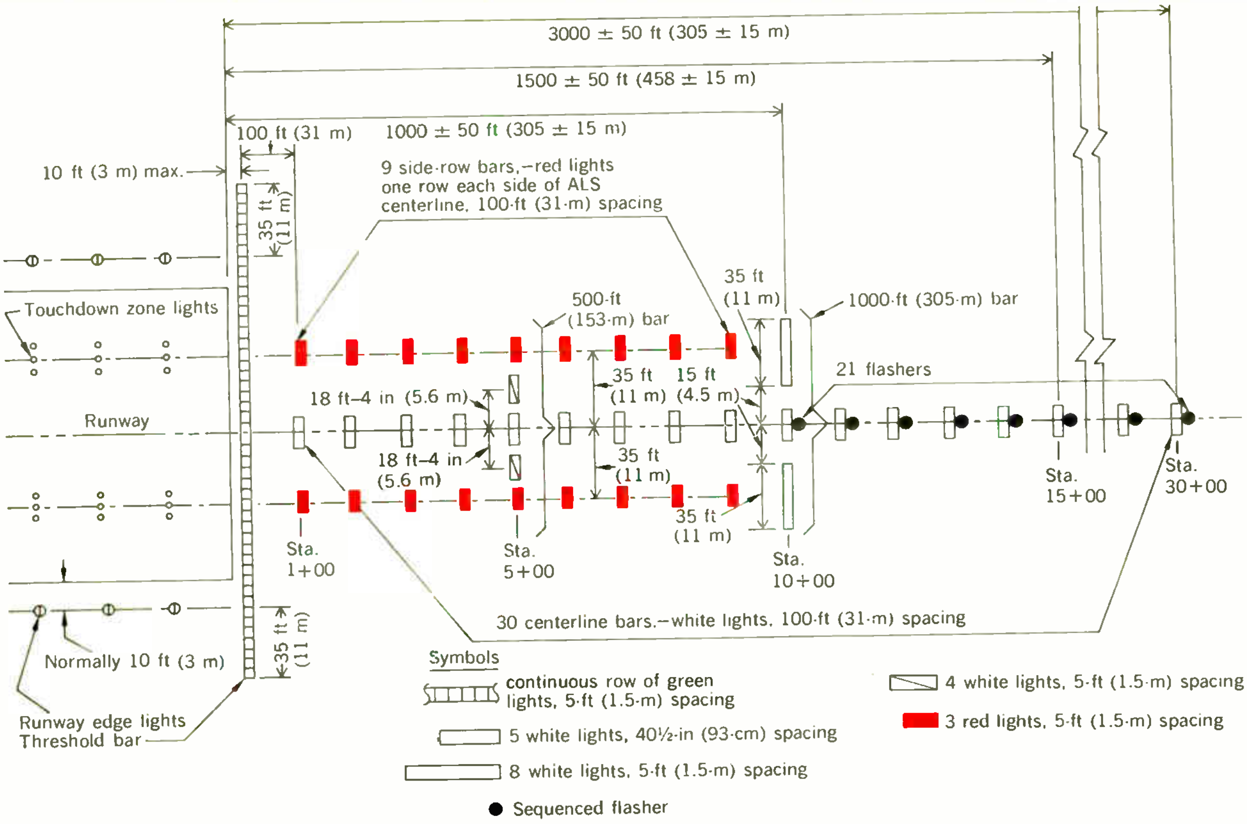


FIGURE 7. Lighting scheme ALSF-2, for principal air-carrier airports that meet the FAA requirements for Category II approach capability.

for the various ILS marker beacons for Category I and Category II locations. A simplified and compact marker beacon that has been developed recently includes a solid-state transmitter, antenna, and power supply, all of which are mounted on a single pole.

Compass locator (*COMLO*). As an auxiliary aid to the ILS system, a nondirectional radio beacon may be colocated at the outer marker (Fig. 1) to establish a final navigational fix for the approaching aircraft. Such facilities are classified as compass locators and are designated as LOM (locator outer marker).

Distance-measuring equipment (*DME*). Although the DME is not presently considered as an integral component of an ILS-equipped runway, it may be installed in conjunction with that system to provide the aircraft's pilot with instantaneous distance information during his approach to the runway touchdown point. If DME is used, it is generally located at the glide-slope site, and it may serve as a substitute for the outer marker. This is sometimes necessary in situations where the ILS approach path is over water, or at locations where no feasible site is available for the outer marker. Usually, the DME is housed in a separate shelter; its power and control circuits are provided by the glide-slope facility. The DME antenna is mounted on a mast that may be affixed to the shelter or ground-mounted. No special site-clearance requirements need be observed if the DME is located at the glide-slope site.

A precision DME for Category III operations is now under development, but detailed information on it is not yet available. Present FAA plans call for the installation of 532 additional terminal DMEs—399 colocated with ILSs, and 146 with TVORs, in the 1970s.

Approach lighting systems (ALS)

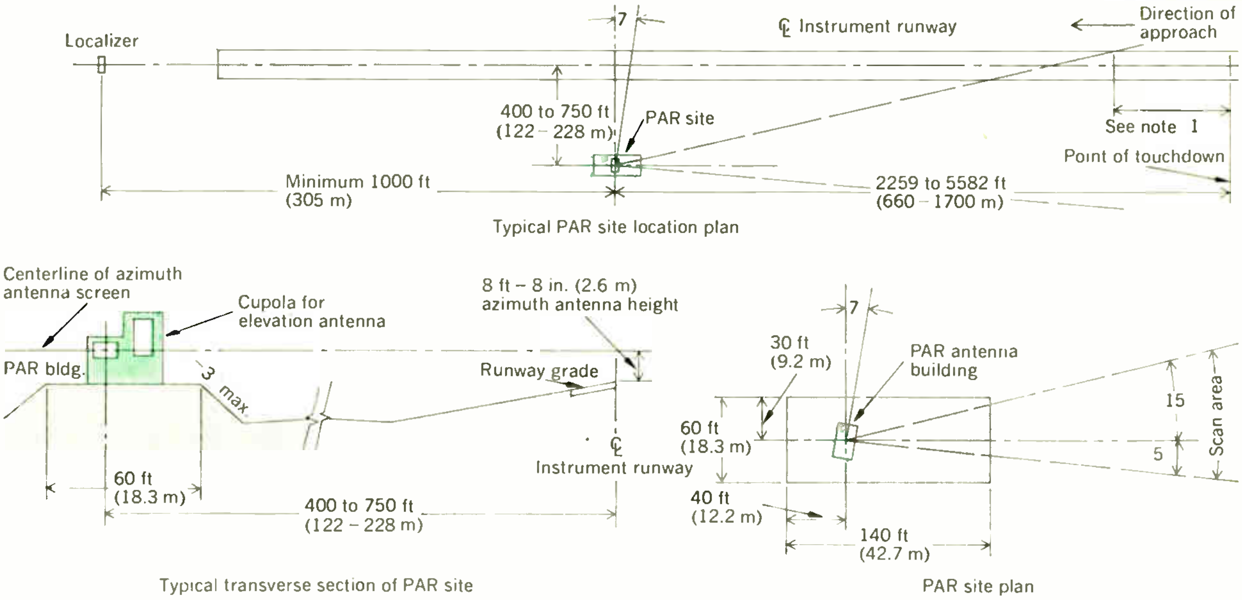
There are numerous FAA-approved lighting system layouts and configurations at the various airports in the U.S. We will be concerned in this article, however, only with those schemes applicable to Category I and Category II ILS systems.

The lighting configuration plan, designated ALSF-1, is for airports with Category I ILS capability. Essentially, these are air-carrier airports that handle jet aircraft with landing speeds of 141 knots (262 km/h) or more, but less than 166 knots (308 km/h); and whose weights are 150 000 pounds (68 000 kg), or more. This scheme is also required if an airport anticipates serving such aircraft within the next three years. The lights in this scheme extend for about 3000 feet (915 meters) beyond the approach end of the ILS-equipped runway.

Figure 7 is the lighting plan (ALSF-2) for principal air-carrier airports that meet the FAA requirements for Category II approach capability.

Air-traffic control towers and related facilities

The ATC tower is the focal point for the safe control of aircraft operations in the designated sector of its airspace and the movement of vehicles and aircraft within the maneuvering area of the airport. The specific criterions for the siting and height determination of these structures are

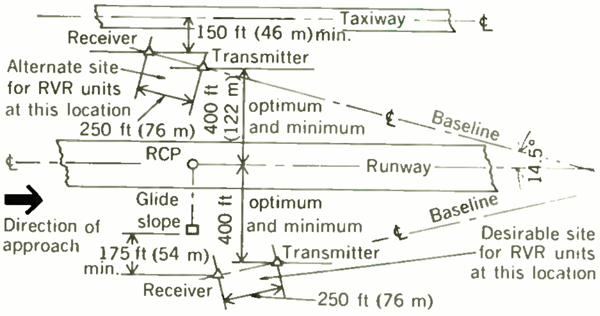


Notes:

1. The antenna will be located so that the runway will be scanned in an area back of the point of touchdown for a minimum distance of 1000 feet (305 m) for a one-approach facility and 500 feet (153 m) when the facility serves more than one approach.
2. Minimum height of azimuth antenna measured with respect to elevation of runway centerline at point directly opposite antenna shelter.

FIGURE 8. Typical precision approach radar (PAR) layout and dimensional criterions.

FIGURE 9. Siting criterions for runway visual-range (RVR) transmissometer system.



Notes:

1. Transmitter shall not be located more than 500 feet (152 m) from runway centerline, unless approved.
2. The reference-control point (RCP) is the intersection point of the runway centerline with the perpendicular intercept through the glide-slope building or trailer.
3. The RVR units shall be within 1500 feet (458 m) of the RCP when located down the runway from the ILS approach end.

1. The attainment of maximum visibility from the tower cab of airborne traffic patterns.

2. A clear, unobstructed, and direct view of the approach to the ends of the primary instrument runway and all other active runways and landing areas.

3. Complete visibility and adequate depth perception by ATC personnel of all airport surface areas utilized in aircraft ground movement.

Under the present FAA plan, 140 new ATC towers will be built, for a total of 521 towers in operation by 1980. And in this decade, 80 existing towers will be relocated to meet the aforementioned siting criterions, and 90 existing structures will be improved by the installation of more sophisticated electronic gear.

Airport surface-detection equipment (ASDE). Airport surface-detection radar is used to permit ATC personnel to observe and control aircraft and vehicular ground traffic on the airport under all weather conditions. The FAA plans to reactivate present nonoperational ASDE equipment and to modify such installations to accommodate visual-display equipment in ATC towers. The ASDE system is a vital part of the proposed automated ground-control (AGC) scheme, which will be explored a bit later in this piece.

Precision approach radar (PAR). The PAR unit enables an air-traffic controller to monitor the approach end of runways (refer to Fig. 1), and it may also be used to provide the controller with an information base for voice-communicated precision-approach instructions to an incoming aircraft. The PAR facility is housed in a special transceiver building that is situated according to the sketches and specifications shown in Fig. 8.

The PAR transmission operates in the range of 9000–9180 MHz.

Transmissometer facilities for runway visual range (RVR)

A transmissometer is essentially an optical device for the determination of RVR during adverse weather conditions. The placement of a transmissometer near the touchdown area of an ILS runway provides for the measurement of visibility in the runway touchdown zone in terms of feet of RVR. Thus the RVR value indicates to

II. Statistical data base, ground-control communications

Airport	Number of Aircraft	Number of Contacts	Avg. Length of Communications Contact Cycle, seconds	Avg. Channel Occupancy per Aircraft, seconds	Frequency Utilization, percent	Controller Utilization, percent	Guidance Information, seconds	Control Information, seconds	Other Information, seconds
LGA	191	494	10	26	70	63	766	1590	749
EWR	136	328	10	24	46	65	358	1021	697
JFK	222	523	10	27	71	59	959	1473	606
ORD	*	*	*	*	*	*	962	1078	534

* ORD data analyzed only with respect to message category and content.

Airport code letters
LGA = New York LaGuardia
EWR = Newark Municipal
JFK = New York Kennedy International
ORD = Chicago O'Hare International

the pilot the horizontal distance down the runway that may be seen upon landing. The RVR is calibrated by the sighting or visibility of the high-intensity runway edge lights (see Fig. 7) by the pilot. For example, "2400 RVR" indicates that approximately 2400 feet (730 meters) of runway edge lights are visible. For Category I ILS runways, RVR is not usually necessary, but it is required for Category II ILS runways.

A basic transmissometer system consists of a computer, transmitter, receiver, and recording instruments. The transmitter directs a beam of light to the receiver instrument that is about 250 feet (76 meters) away. The receiver unit then sends an electronic signal by cable to readout and recording instruments in the control tower cab. Operating personnel in the cab can then inform approaching pilots of the actual visibility conditions on the runway in terms of RVR in feet. Figure 9 shows the siting criterions for a typical transmissometer system.

NAVAIDS for nonprecision instrument approaches*

Although there are five systems presently in use to assist general aviation aircraft in making instrument approaches, spatial constraints permit only the description of two systems that are extensively employed.

Terminal VHF omnirange transmission (TVOR). The VHF omnirange (VOR) is the standard electronic NAVAID used throughout the airway system to provide azimuth (magnetic bearing) guidance to aircraft equipped with standard instrumentation, but without electronic glide-slope capability. The TVOR system is usually installed to establish a nonprecision instrument approach procedure assistance, which enhances the IFR potential of general aviation aircraft.

A TVOR site is preferably located so that the facility will provide guidance to the approach ends of runways. This criterion may be met by selecting a location adjacent to the intersection of the principal runways (see Fig. 1). Ideally, the TVOR site should be at the highest ground in the vicinity, and there should be no conflicting objects within a radius of 3000 feet (915 meters) of the antenna.

The TVOR transmission is authorized on even tenths of a megahertz[3] in the 108–112-MHz band. The conventional omnirange utilizes a 200-watt transmitter to radiate a reference signal, a variable signal, voice, and identification. The carrier energy is fed in such a manner as to produce an omnidirectional radiation pattern, whereas the sideband energy is fed in a way to produce a space pattern (cardioid), which rotates at a 30-Hz rate. As an aircraft flies around the VOR, it makes a continuous phase measurement of the reference signal received. The same aircraft receiver is commonly used for both VOR and localizer reception, with alternate circuitry for the different modulation frequencies and course widths.

Nondirectional beacon (NDB). The NDB transmits nondirectional signals on which a pilot may take a bearing by using the automatic direction-finder equipment in the aircraft to "home in" on the destination airport. The NDB facility consists of a small transmitter building and a special wire-antenna system that is carried by two poles spaced about 250 feet (76 meters) apart. Above-ground structures and airport appurtenances rarely cause interference with the performance of an on-airport NDB.

Automated ground control

The FAA recently initiated a program for the development of an all-weather automated ground-control (AGC) system. The objective is to afford both aircraft and ground vehicles the necessary data to maneuver safely on the airport surface. The AGC system should have the capability of locating the position of an object anywhere on the airfield. Thus ATC personnel could be provided with constantly updated positional information on all aircraft and vehicles in motion or at rest within the airport operational areas.

As a prerequisite to embarking upon one version of the new system design, present airport surface-traffic patterns were analyzed to define the existing problems and deficiencies. One such analysis revealed some startling statistics. For example, in the area of ground-control communications, voice transcript studies from four of the FAA's most active control towers (Table II) indicated that, for the largest number of aircraft handled in a one-hour period (222 at Kennedy International), the ground-control channel-occupancy time reached 71 percent—a percentage that approaches critical conditions. Table II also shows that about 10 seconds is required for an average communication between a pilot and a controller, and about three such contacts were necessary for each aircraft. Thus there was a total channel occupancy time of about 30 seconds per aircraft. Now, if we multiply 523 (the number of contacts at JFK) by 10 (average length of communication contact cycle), the product is 5230 seconds—but there are only 3600 seconds in an hour!

---

* See glossary for definition of term, under *nonprecision approach procedure*.

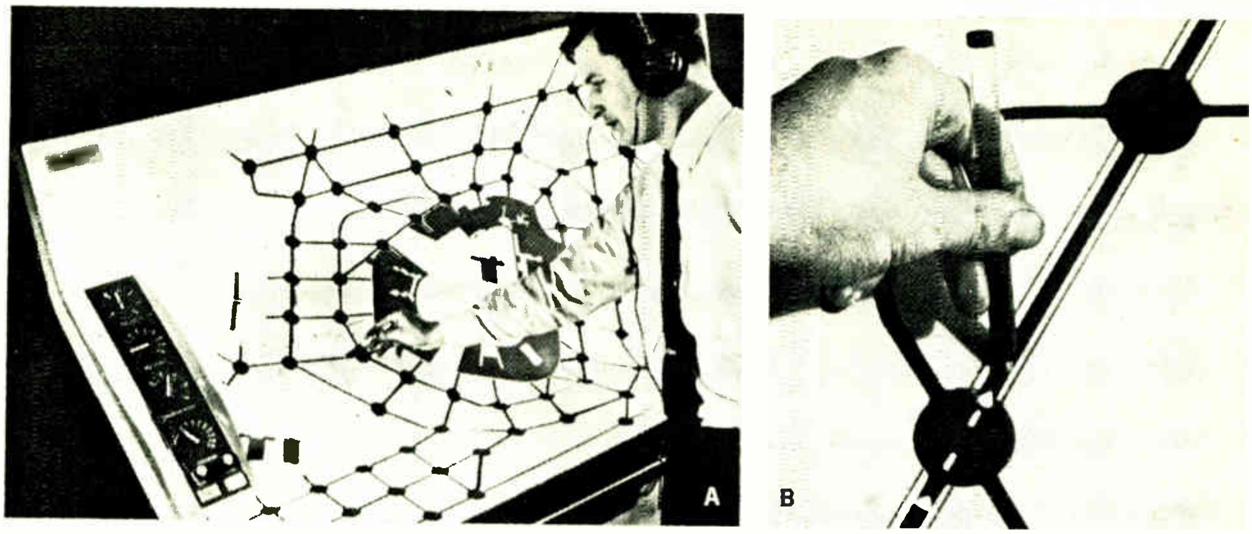
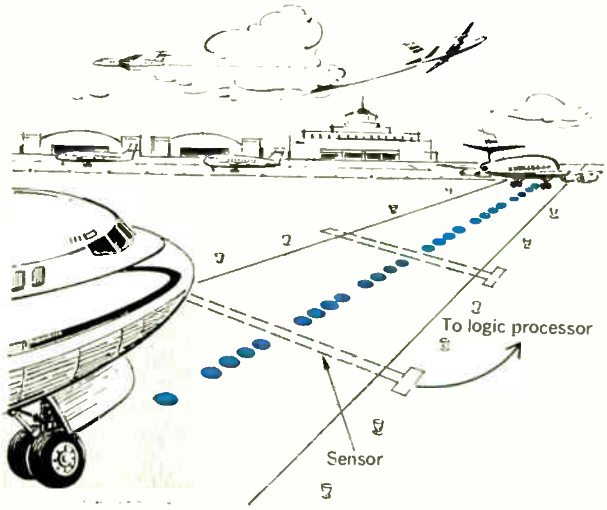


FIGURE 10. A—An airport facsimile map display serves as the master control panel for a proposed automated ground-control (AGC) system. B—Up to 50 taxiway lights can be activated when the controller moves a magnetic stylus in the grooves along a selected route on the master control map.

FIGURE 11. Sketch showing signal lights embedded along the taxiway in one version of an AGC system.



Therefore, one can readily see that the theoretical number of aircraft handled (assuming a 100 percent channel occupancy, which cannot be attained) is far below the 222 indicated in the table.

These—and other—data conclusively indicate that the present ground-traffic congestion, plus the inevitable future increase, presents a hazardous picture. Also, more and more aircraft will be operating out of the line of sight of tower cab controllers. For instance, the operation of the 747 superjets may block areas of the field either from the controller's vision or from ASDE radar detection. And the 747s, with tail structures about 60 feet (18 meters) high, will temporarily cut off portions of the airport from view.

The FAA's analysis isolated the two basic ground-control functions—control and guidance—which must be handled by any future automated system. Each of these functions consists of three subroutines: information acquisition, information processing, and response.

How the basic functions are presently handled. In applying the control function, the controller acquires information by means of visual observations, pilot and driver reports, and weather and meteorological predictions. And, at about 11 U.S. airports, ASDE radar provides assistance for low-visibility and night operations. But it is the controller who processes all of these data, and he responds by voice-communicating pertinent information and instructions to pilots.

In the guidance function, the pilot or ground vehicle driver acquires positional information by the controller's instructions and by observing airport signs, lights, and other markings. He processes these data and responds by maneuvering his aircraft or vehicle according to verbal instructions and visual signals.

Thus it is apparent that, in the present ground-control operations, human eyes and ears handle the bulk of the data acquisition, human minds do the data processing, and human response and reaction are required for the safe and efficient maneuvering of ground traffic. Unfortunately, as the volume of traffic increases, the human factor becomes overburdened and the margin for human error increases.

Some proposed versions of AGC. An article in the January 1970 issue of *The Journal of Air Traffic Control*,[4] by Walter S. Luffsey and Terence B. Wendel of the FAA System Research and Development Service, describes one of a number of proposed AGC systems (designated ARROW—air routing right of way—and developed by the E. W. Bliss Co., Canton, Ohio), which utilizes an airport facsimile map display [Fig. 10(A)] as one of its components. The master display actually serves as a lighting control panel on which a controller can activate up to 50 taxiway lights by moving a magnetic stylus in grooves [Fig. 10(B)] along the selected route. Pilots could use this information to maneuver their planes along a cleared taxiway or runway as shown in the Fig. 11 sketch. If an emergency stop is indicated, red lights, embedded along the pavement, are flashed on as a stop order. Yellow signal lights would serve as a directional command. No voice communication would be needed.

The map display, which would be analogous to a railroad's centralized traffic-control (CTC) system, shows which field signals are in use; it would also indicate outages.

In another system approach, Fig. 11 also shows the surface-embedded sensors that could detect aircraft and vehicles on the airport movement area. These sensors would activate corresponding signal lights on the control display to follow the position and route of each plane or vehicle along its assigned path. Tests conducted at Kennedy International last year indicated that induction loops, embedded across the taxiway pavement, are satisfactory for the sensing function.

The next phase in this AGC proposal is the incorporation of simple logic elements that will trigger audio and video alarms to warn ATC personnel of any irregular or hazardous situation. The logic unit would also permit the selective designation of preferred, alternate, or priority routings.

The alarm logic would activate a flashing red light at the appropriate analog point on the master map display, and this same light—plus an audio signal—would be activated if any aircraft or vehicle fails to proceed as directed. This built-in fail-safe alarm system would be triggered, for example, if an aircraft or vehicle crosses a detection point on an intersection approach where a stop order has been given; if an instruction is entered that creates a conflict with a vehicle already detected inside the affected route; if vehicle headway time decreases to indicate an "overtaking" situation; or if an aircraft or vehicle fails to activate the next detector in sequence within the required time by making an unauthorized stop.

Regular arrival and departure routes for a given runway operation could be temporarily overridden by priority logic resetting to permit abnormal or special operations.

At several of the nation's busiest airports, the volume of surface traffic would exceed the data input and processing capabilities of a simple logic system. Thus a digital computer could be added to the system for the necessary information processing, display, and field-signal operations.

The computer would

1. Accept positional inputs.
2. Determine the suitability of field signals.
3. Update the master-control display.
4. Track aircraft and vehicles, and assign routings based on speed and priority considerations.

In addition, routing conflicts could be predicted and resolved by priority queuing, and alarms would be automatically tripped in emergency situations whenever logic inputs appeared unreasonable for processing. The controller would then become essentially a monitor and fail-safe backup (by means of manual intervention) in what would be, for all intents and purposes, an automatic system.

Airport operations under extremely poor visibility conditions may necessitate the use of a reduced visibility guidance and control subsystem; one such concept is shown in Fig. 12. Here, dual forward-looking sensors would detect signals (in the 1.75–2.25-μm range) from the centerline lighting systems in the runways and taxiways. The two sensors, mounted symmetrically on each side of the aircraft, would compare the signal strength of each light; any signal-strength differential would deflect a needle on the pilot's "left–right" indicator. And as each surface light is approached by the plane, its signal strength would increase to a peak intensity. These peaks would activate a digital counter (one count per light). If the count is started from a known reference position, it could provide the pilot with data for planning his taxi route, assessing his progress, and anticipating maneuvers and turns.

The use of another sensor, operating in the 0.75–1.0-μm range, could detect stop signals and activate an audio/video alarm in an aircraft's cockpit.

A comprehensive block diagram of the total operation of this AGC system is shown in Fig. 13.

An alternative system. The "taxi-guidance" system, originally developed by Bliss for the U.S. Navy's jets, has been modified for commercial airport use.[1] This AGC scheme, unlike ARROW, requires several elements to be mounted in the aircraft: a position-error transducer, a control box, hydraulic-steering control, and hydraulic brake-manifolds.

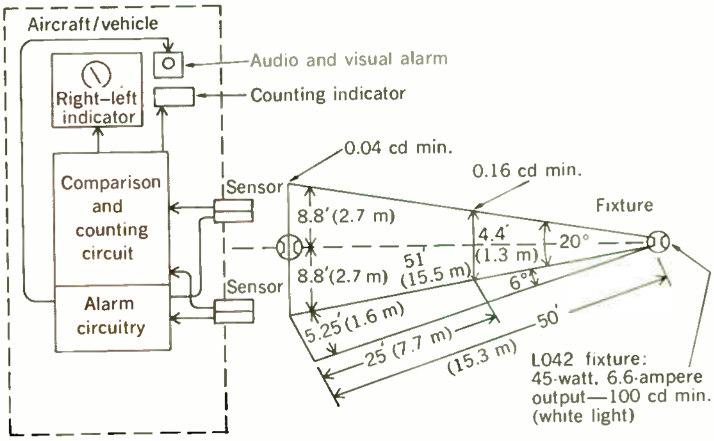
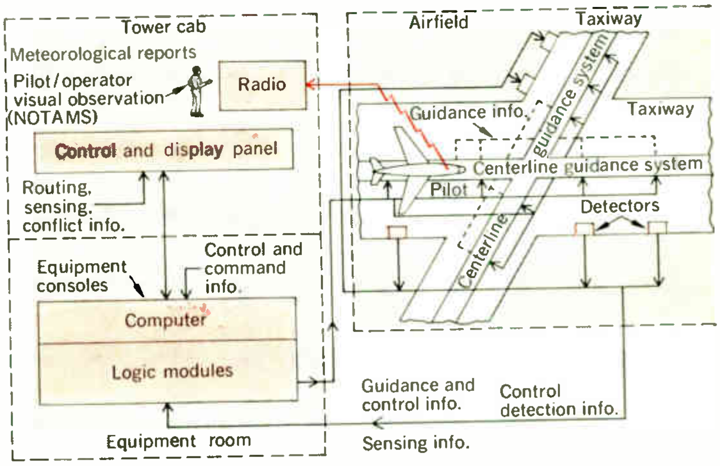


FIGURE 12. Block diagram of reduced-visibility guidance and control subsystem.

FIGURE 13. Comprehensive block diagram of the total AGS system operation shown in Fig. 12.



Two parallel wires, embedded in the taxiway, afford the primary reference path. A low current passed through these conductors creates a magnetic flux field whose strength is proportional to the perpendicular distance from the parallel wires. The magnetic field is detected by sensors mounted on the plane's nose landing wheel, which is used for steering when the aircraft is on the ground. Thus the nose wheel is guided electronically down the centerline of the taxiway as the pilot controls the aircraft's speed. If there is any deviation in the course of the plane from this centerline, the hydraulic brakes are automatically applied.

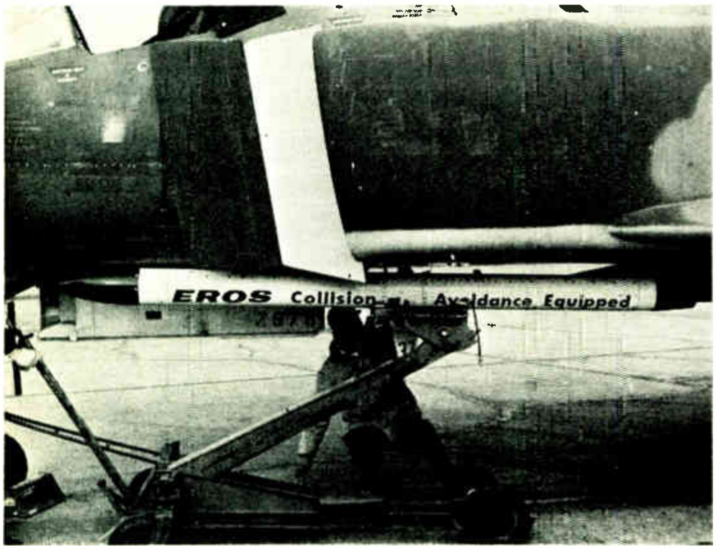
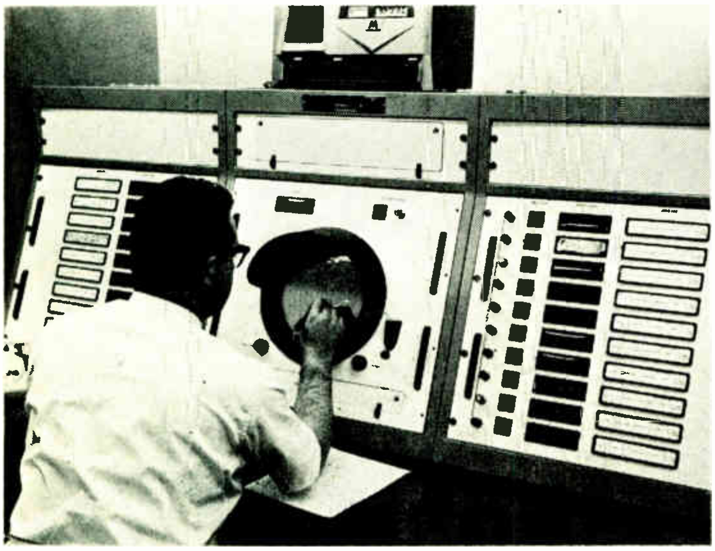


FIGURE 14. An airborne EROS pod package installed in the missile well of a McDonnell Douglas F-4 Phantom II fighter-bomber. The pod is an essential component of the EROS I collision-avoidance system.

FIGURE 15. A controller at an EROS flight-following station locates an aircraft participating in the CAS hierarchy. He identifies the target blip on his PPI scope by means of a light-pencil detector. By pressing a target blip-augmentation button on the console, the target will "bloom" on the scope for easy positive readout.



A few out of many. The AGC systems just described are but a few of a number of possible approaches to the solution of present surface-traffic difficulties. Numerous studies and tradeoffs will have to be made among many diverse concepts to arrive at the synthesis of a practicable future system. The FAA hopes that, by 1974, a fully developed ground-traffic system will be in operation at major airports.

Introduction to collision-avoidance systems (CAS)

At the present time, the FAA is initially planning the construction of 65 CAS ground stations. The final number and priority of their installation, however, has not yet been established. It is no secret that CAS has long been an elusive desideratum of the commercial airlines. The initial efforts date back to the mid-air collision of two air carriers over the Grand Canyon in 1956, and the developmental push was accelerated when two airliners collided over New York City in December 1960.

The early systems were based on the military stationkeeping concept (also called "formation flight-separation control"). And it is an airborne control concept that still has considerable validity. In fact, Sierra Research Corp., Buffalo, N.Y., believes that stationkeeping could have commercial applications if our airways become so congested that large groups of air carriers and private aircraft are required to move in convoy formations between major cities. Another adaptation of the principle might be in reducing landing intervals during peak hours, when many aircraft must be guided down safely, in close order, at a major airport under poor visibility conditions. Sierra proposes an airborne system that can display a flight formation of up to 36 aircraft on an instrument-panel-mounted PPI (plan–position indicator). The device, presently used on military C-130 transports, could be modified and simplified for installation in air carriers and general aviation aircraft.

Present-day subsonic air carriers can approach each other, on diametrically opposed courses, at relative closing velocities of more than 1200 mi/h (1935 km/h). Thus any operational CAS demands a precision, synchronized time-reference clock base.

The EROS I system. In 1962, McDonnell Douglas Corp. developed its first Eliminate Range Zero System (EROS I) for military aircraft applications ("Range Zero" being a mid-air collision!). Basically, the system consists of two packages: an airborne pod, installed in a fighter-bomber's missile well (Fig. 14); and a flight-following ground station (Fig. 15), which consists of a console, a master oscillator, and a transceiving direction-finding antenna with a parametric amplifier. The transmitter consists of a frequency multiplier and a pulse-modulated five-stage power amplifier. The multiplier accepts a basic 5-MHz reference signal input and multiplies it to obtain the transmitted frequency of 1545 MHz. The transmitter output is 1000 watts.

The ground station antenna consists of a separate transmitting antenna and receiving antenna mounted on the same tower. Bearing information is acquired by feeding a coaxial waveguide, with a discone-type receiving antenna, and the processing of two resultant orthogonal, coaxial waveguide modes.

A three-channel parametric amplifier provides amplification to permit efficient system performance up to a line-of-sight range of 140 nautical miles (260 km). The

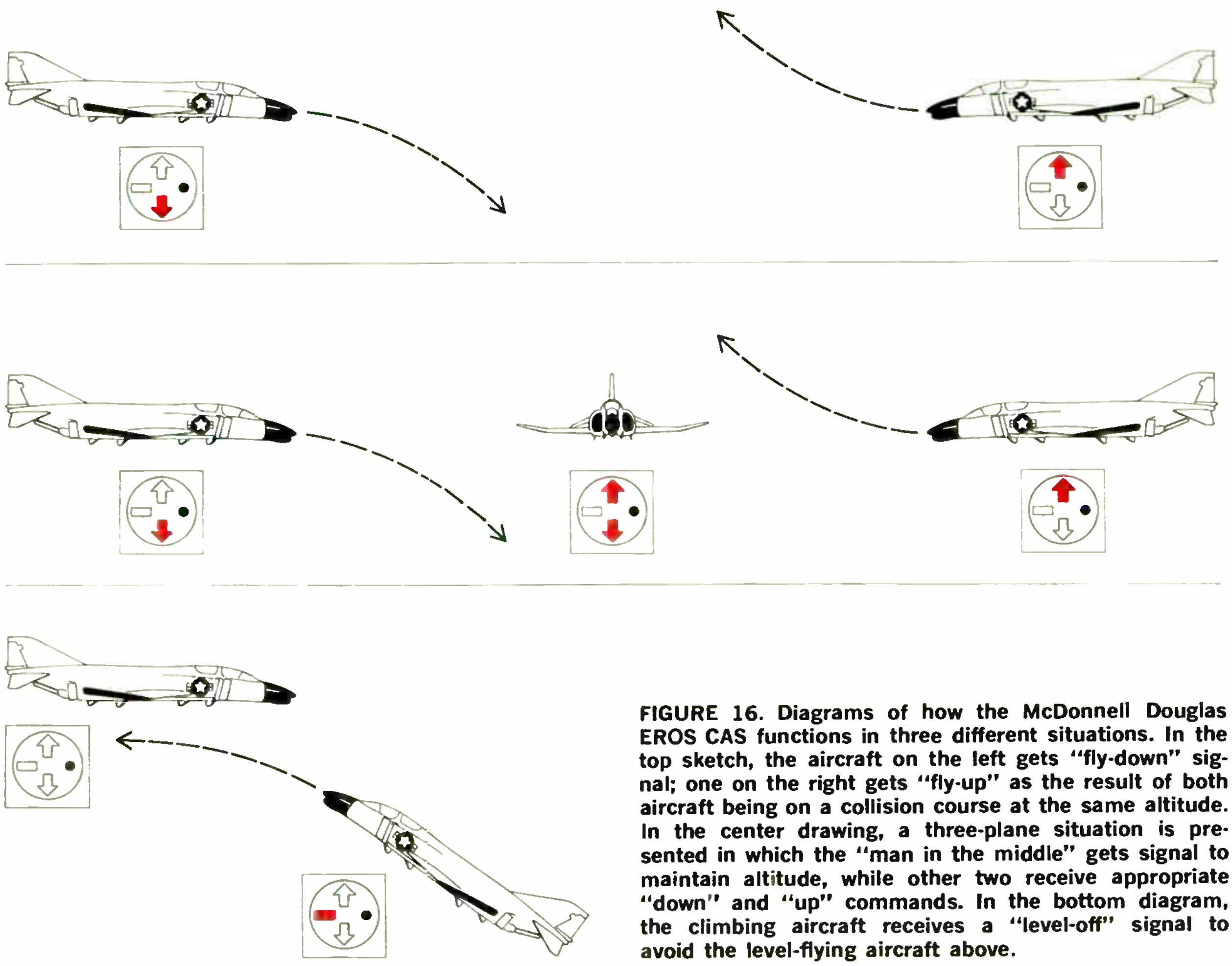


FIGURE 16. Diagrams of how the McDonnell Douglas EROS CAS functions in three different situations. In the top sketch, the aircraft on the left gets "fly-down" signal; one on the right gets "fly-up" as the result of both aircraft being on a collision course at the same altitude. In the center drawing, a three-plane situation is presented in which the "man in the middle" gets signal to maintain altitude, while other two receive appropriate "down" and "up" commands. In the bottom diagram, the climbing aircraft receives a "level-off" signal to avoid the level-flying aircraft above.

receiver accepts parametric amplifier outputs at 1545 MHz and generates from these the video signal and phase-related signals from which range, altitude, and azimuth information is derived. A central timer unit (atop console in Fig. 15) provides accurate time-base signals for use by other system components. The reference frequency is derived from a 5-MHz oscillator, with a stability of one part in 10^{10} per 24 hours. [This degree of precision measurement in defining the relative positions among various combinations of aircraft is equivalent to 0.2-μs accuracy. These one-way (ranging) radar measurements are updated every 2 seconds.]

The ground station may be operated in conjunction with the airborne pods to provide various aircraft identification and to transmit the synchronization signal from the master oscillator to airborne units. The station may also be used to generate and transmit a test sequence for preflight ramp checkout of the CAS.

EROS I in operation. The EROS I gives pilots of military aircraft 60 seconds* warning of a potential collision with another EROS-equipped plane.[5] This is followed by a simple "escape" command: either "fly up," "fly down," or "level off." The system functions when aircraft are turning, climbing, descending, or in level flight. EROS I will afford adequate warning to converging aircraft at closing speeds up to Mach 4 (four times the speed of sound).

Figure 16 illustrates the function of the EROS I system in three classic situations. In the top sketch, the plane on the left gets the "fly-down" signal; the one on the right receives "fly-up" instructions as the consequence of the two aircraft meeting at the same altitude. System fail-safe techniques preclude any possibility of ambiguity in signals. In the center sketch, a three-plane situation is indicated, in which case the "man in the middle" gets the signal to maintain his altitude, while the other two receive appropriate down and up commands. The bottom diagram shows the climbing aircraft on the right receiving a "level-off" signal to avoid the level-flying aircraft on the left. Note that the level aircraft receives no signal command since the climbing aircraft has been diverted prior to entering the "guarded airspace" zone of the level-flying plane.

To explain and illustrate the instrumentation in more detail (Fig. 17), let us assume the first of the three classic situations just mentioned. At the 60-second warning time mark, the pilots receive an audio alarm signal. Simultaneously, the pilot of one of the involved aircraft gets a command signal (in the form of a flashing red arrow) telling him to climb, while the pilot of the other aircraft is instructed to dive, thereby avoiding a potential mid-air collision.

Evaluation of airborne CAS

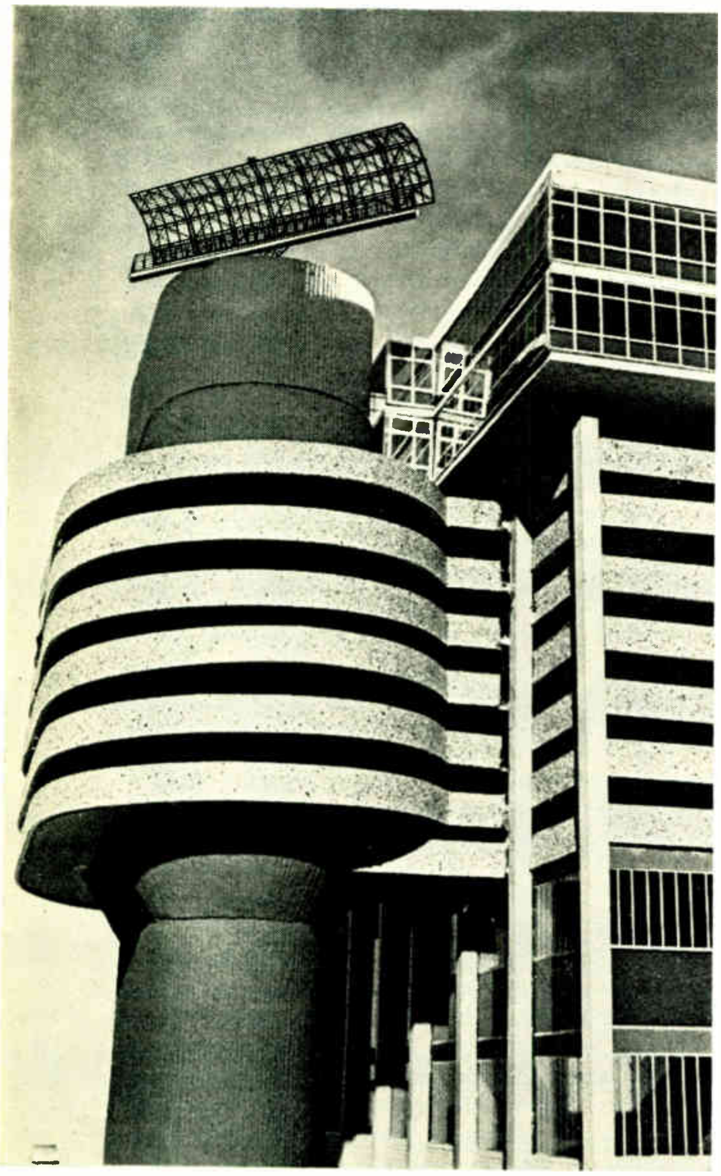
On March 20, 1970, the "Flight Test and Evaluation of Airborne Collision Avoidance System" summary report, ments for safe operation under IFR or VFR conditions.

Acceleration of automation. An FAA bulletin, released last March, states that "semiautomated ATC systems, now being implemented nationwide, must be substantially upgraded if they are even to accommodate the aviation growth of the 1970s." The basis of the bulletin was the report of the Secretary of Transportation's Air Traffic Control Advisory Committee.

The report identifies three critical problem areas that must be resolved if aviation growth is to continue without constraints: the shortage of air-terminal capacity in handling the present rate of increase in air operations, the need for new means of ensuring separation between airborne aircraft (fully dependable and operational CAS systems), and the limited capacity and increasing cost of air-traffic control.

In noting the increasing public resistance to new airport construction near metropolitan areas, the report indicates that it is feasible to "more than double the capacity of existing airports through the use of parallel runways, high-speed turnoffs, advanced terminal automation, and reduced longitudinal separation between aircraft on final approach for landing." The document optimistically predicts that the computer technology available by 1975 "will be adequate to cope with twice the projected 1995 traffic." (This forecast and other observations contained in the report will undoubtedly be considered open to debate by the air-traffic controllers and other interested parties.)

FIGURE 18. Multistory structure at London's Heathrow Airport serves as carrier for high-mounted 50-cm all-weather radar system.



A glimpse of some ATC equipment for foreign installations: English, Swiss, Canadian

Rooftop radar for London airport. The multistory structure, shown in Fig. 18 (a combination aircraft workshop and car park), at London's Heathrow Airport will house a Marconi 50-cm radar system. Installed 100 feet (30.5 meters) above grade, it will have a clear, unobstructed electronic view for the surveillance of aircraft leaving the field. The radar will also serve as an approach control under adverse weather conditions. The set is designed to perform without pronounced echo or clutter effects. It has a power range of approximately 160 nautical miles (256 km).

The radar, designated S650, is one of Marconi's fourth-generation 50-cm models designed for use in monitoring aircraft departures and holding-approach patterns. It is an all-solid-state system and operates in the frequency range of 580–610 MHz. The transmitter's peak power output is 500 kW (at up to 550 pulses per second). Structurally, the 7- by 9¼- by 2¼-foot (2.15- by 2.85- by 0.7-meter) antenna is designed to withstand a wind velocity of 120 knots (222 km/h).

Visual range equipment at Gatwick. The first fully automated overseas system for the assessment of visibility on airport runways was recently installed at Gatwick (England). The GEC-Marconi Electronics Instrumented Runway Visual Range (IVR) system will make Gatwick the United Kingdom's first airport to have the continuous fully automatic reporting of visual range, as recommended by the fifth Air Navigation Conference. At the present time, the system is operating in an evaluation mode, but when it becomes fully operational it will probably eliminate the human assessment of visual range (which is presently accomplished by an observer who counts the number of runway lights he is able to see under varying weather conditions).

The observer method cannot achieve the data rate necessary for IFR approaches under minimal Category II conditions. By comparison, the automatic IVR system makes such observations every 1.5 seconds, and collects data from several points along the runway. From this information, a mosaic pattern of the overall weather situation emerges.

The Gatwick IVR system is based upon photoelectric measurements of the atmospheric transmission of light at three points along a runway: at each end, and at the midpoint. The equipment consists of three field sensor sites alongside the runway, a data-processing complex in the electronics area of the control tower, and a number of readout indicators in consoles in the control center.

The field sensors measure the intensity of light transmittance by using optical transmissometers. Each transmissometer has a transmitter and receiver colocated at one end of a base line, with a tetrahedral mirror at the other end. This arrangement enables reference values of transmission to be established by passing light internally from transmitter to receiver. The constancy of the external path is assured by a unique air-flushing system, which maintains a shield of filtered dry air in front of the lens and mirror. Measured transmittance is then digitized in the field so that it can be relayed to the computer in the control center over ordinary telephone lines.

Two other blocks of information must also be provided to the computer in making the calculation of visual range: the level at which the runway lighting is set, and the general background brightness. The latter is measured in the field by a background luminence monitor (a digital photometer developed for this IVR system).

It should be emphasized, however, that these equipments are counterparts of RVR systems presently in use in the United States and thus do not necessarily represent more advanced techniques or a higher degree of automation than domestic installations.

Radars for the Swiss government. ITT Gilfillan, through its subsidiary, Standard Telephone and Radio of Zurich, has a contract to supply its "Quadradar" all-weather system to the Swiss government. The Quadradar is a complete terminal-area air-traffic-control system with four principal capabilities: terminal-area air surveillance, precision-approach control, height determination, and aircraft taxi information. The equipment has been use-tested under the extreme weather conditions of the Arctic and Antarctic and, like the U.S. Post Office motto, "neither sleet nor snow nor gloom of night" seems to bother this radar system in helping to land aircraft in the most adverse climatological conditions. Each of its four display modes can be individually selected. Thus, in low-density traffic situations, a single controller can perform services equivalent to those of larger-capacity equipment with multiple operator positions.

All-weather system for Canadian armed forces. A new air-traffic-control radar system, consisting of a Marconi 23-cm radar, has been ordered by the Canadian Department of Supply and Services to provide all-weather approach-guidance and terminal-area monitoring facilities for the Canadian Forces base at Comox, B.C.

Comox is a difficult site for a radar installation as it has one of the highest yearly rainfall rates in Canada and it is surrounded by a ring of mountains. The Canadian Department of National Defence decided to replace the existing 10-cm radar equipment with the 23-cm radar to reduce both ground and weather clutter on the video screens. Performance in bad weather is generally improved at longer wavelengths; however, frequency allocations prevented the use of the unique 50-cm radar.

The new equipment uses a dual-beam aerial system to provide an overlapping aerial pattern by combining the high and low cover beams from a single antenna. This feature is designed to reduce ground clutter and "ghosts" on the display, while simultaneously maintaining a high signal strength over a very wide vertical sector. The lower beam only is used for transmitting the radar pulses, but both the upper and lower beams are employed for return echo reception. The signals are returned to a common receiving system by means of a high-speed switch that operates at a predetermined time after each transmitted radar pulse.

The system is arranged to accept radar returns from short-range targets, via the high beam, to eliminate ground clutter. For longer-range targets, the low-beam antenna is selected. The changeover could be set, for example, at about 15 nautical miles (28 km).

An alibi in advance of Part III

The writer (from previous experience) anticipates comments from some readers taking him to task for failing to mention certain systems in use by the FAA for air-traffic control. Unfortunately, this is inevitable, since the length limitation of this installment precludes a discussion of all such systems or equipment. To do so would fill more than a book—and indeed it does, in the FAA's manuals that describe the full range of electronic systems presently in use.

Further, if anyone has been left suspended in midair with insufficient visibility as to the function of the airborne instrumentation that is required in an IFR approach to an ILS runway, or CAS, grab onto a skyhook and be patient for another month, when the complete explanation of on-board systems will bring you safely to earth in Part III—appropriately titled: "The View from the Cockpit."

Appendix. Glossary

Category I (operational performance). Operations down to minimums of 200 feet (61 meters) decision height and 2400 feet (730 meters) runway visual range (RVR), with a high probability of approach success. When RVR is not available, one-half mile (0.8 km) is intended. (With respect to the operation of aircraft, *decision height* is the altitude at which a decision must be made, during an ILS or PAR instrument approach, either to continue the approach or abort. This height is expressed in feet above mean sea level, and for Category II ILS operation, the decision height is additionally expressed as a radio altimeter indication.)

Category II (operational performance). Operations down to minimums below 200 feet decision height and 2400 feet RVR, and to as low as 100 feet (30 meters) decision height and 1200 feet (365 meters) RVR, with a high probability of approach success.

Nonprecision approach procedure. A standard instrument approach in which no electronic glide slope is provided.

Precision approach procedure. A standard instrument approach in which an electronic glide slope is provided, either by the ILS or by precision approach radar (PAR).

Terminal navigational aids (NAVAIDS). The electronic facilities established on or in the immediate vicinity of an airport that enable a pilot, when using compatible airborne equipment, to execute the instrument approach or approaches authorized for the airport by the FAA.

REFERENCES

1. Wood, N., "The automated sky," *Mach. Des.*, vol. 41, pp. 19–32, Oct. 30, 1969.
2. "Airport design standards—site requirements for terminal navigational facilities," U.S. Dept. of Transportation, Federal Aviation Admin., Washington, D.C., Oct. 8, 1969, p. 8.
3. "Frequency management principles, spectrum engineering measurements," U.S. Dept. of Transportation, Federal Aviation Admin., Washington, D.C., May 29, 1969.
4. Luffsey, W. S., and Wendel, T. B., "Automated ground control," *J. Air Traffic Control*, vol. 12, pp. 8–11, Jan. 1970.
5. "Air traffic control and the airborne collision avoidance system" (a follow-up on the report made in *J. Air Traffic Control*, Sept. 1962).

New product applications

Graphical system aids design of circuit mask layouts and enables automatic artwork generation

The Design Assistant is a new interactive graphical system that allows a user to generate and work with a computer representation of circuit mask layouts. When a composite layout is complete, the system automatically produces data describing the individual mask levels. These data are used directly for automatic artwork generation and design documentation. The new system increases the efficiency of the design and layout process, and eliminates the need for laborious and error-prone mechanical data capturing and the need for manual preparation of data for automatic artwork generation.


FREEHAND input to design assistant.

The new system includes a graphics terminal—a Computek storage tube display, keyboard, and data tablet—and an IBM 1130 computer. No user programming is required and software interfaces are available for a variety of artwork equipment.

In a typical application of the system, the user begins by retrieving his layout from the disk and displays it on the storage tube. He then manipulates the displayed layout by specifying a particular composite that he wishes to view. Thus, he can display an arbitrary composite of individual levels, any portion of the composite, or some of the composite or any degree of detail. The user-specified view of the layout is displayed on the storage tube within seconds after the display command is given.

The user interacts with the system with symbols drawn freehand on the data tablet with an electronic stylus. These symbols give commands to the system and indicate positions of elements and components. The hand-drawn symbols may be of each user's own choosing and design. The correspondence between symbols and layout editing commands is established with a portion of the program.

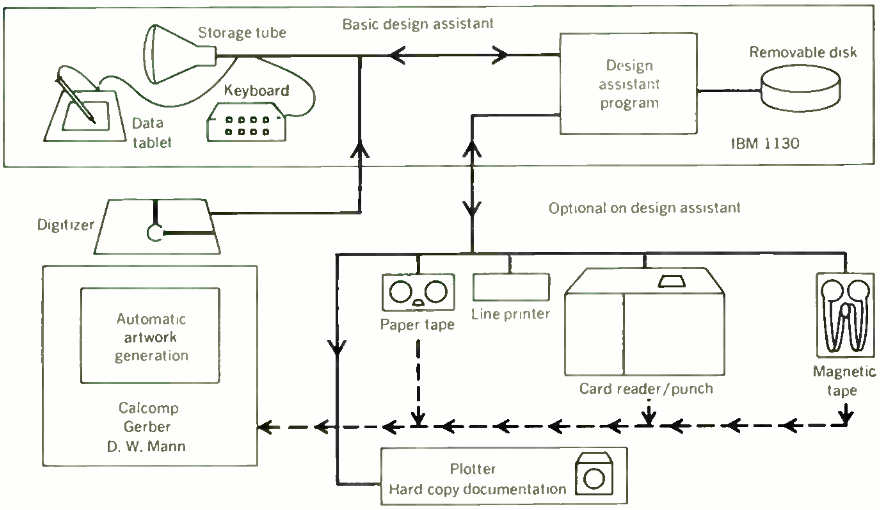
The user can add to a layout's previously defined library components. When a library component is added to a layout, the system quickly updates the display and all appropriate individual levels. Layout editing capabilities include adding, deleting, stretching, shrinking, rotating, flipping, and moving of selected components. Changes to a component are reflected as changes to all individual levels on which the component is defined. For example, when a transistor is deleted from the layout, its emitter, base, base insert, contacts, and contact cuts are all deleted from the appropriate levels.

The library is the heart of the system. Components that are placed in the actual layout are selected by the user from one of his previously defined libraries. The first step in using the system is to define a library. The system has facilities for several different libraries, each of which has its basic components and previously defined cells. The multiple library capability allows the user to use the system for many different types of work. For example, one library may be used for construction of layouts for MOS circuits and a different layout used for bipolar circuits. The system will allow over 32 000 components to be included in a layout.

More information is available from Applicon Incorporated, 83 Second Ave., Burlington, Mass. 01803.

Circle No. 85 on Reader Service Card

BASIC system components.



DISPLAY of a three-input ECL gate.