VOLUME 3 □ NUMBER 2 **FEBRUARY 1966**

IEEE spectrum

features

Departments: *please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

The Northeast power failure of November 9–10, 1965, is the subject of this month's cover design. The feature article, beginning on page 54, presents a comprehensive chronological review of the actual event, the problems involved in the restoration of service, and, finally, the recommendations that are proposed by the Federal Power Commission to ensure a greater degree of service reliability for interconnected systems.

Spectral lines

Intersociety Relations. Probably relatively few IEEE members are familiar with the many relationships that the IEEE has with other societies. Yet there are more than 50 different committees, councils, and boards to which the IEEE appoints representatives, and a lesser number of national and international bodies with which it provides an IEEE interface.

These relationships fall into several general categories. The largest single category involves committees concerned with arrangements for conferences or symposia. Such committees provide the means by which societies having interests in a single area of technical activity can work cooperatively and thus prevent unnecessary duplication of effort.

Another relationship arises through the provision of technical advice and assistance to international or national bodies. The fact that the IEEE has within its membership highly competent people in a variety of technical specialties is recognized, and the Institute provides the service of locating and nominating an appropriate person for a particular assignment. In so doing, the Institute in effect endorses the individual as one who is competent to render sound technical judgments.

In a somewhat similar category are representatives to joint committees concerned with standards and awards. Again, the IEEE by the act of nomination effectively endorses the judgment of its nominees. In these cases, the individual makes decisions primarily on the basis of the technical facts involved. While undoubtedly influenced by the fact that the IEEE has designated him and that the Institute has a nonprofit, nonpolitical, nonnational nature, he does not need to consider, and, in fact, may not be well informed about, all aspects of IEEE policies.

However, there are other cases in which the position of the IEEE's nominees is not so clear cut and in which the individual is actually representing the Institute. Then the question of the extent to which he is free to act on the basis of individual judgment and the extent to which he is bound to represent the IEEE's large and diverse membership is not easy to determine. For example, if the Institute's representative is on the governing board of a council or federation of engineering societies, his position on various questions can be heavily influenced by what he feels are the best interests of the IEEE whether or not these coincide with his viewpoints as an individual member of that board or council. Thus he may be faced with the dilemma of whether he is an instructed or a free delegate. The immediacy and seriousness of this question vary according to the circumstances, but it is clear that in some cases careful delineation between the position of the individual and the position of the Institute is necessary. Representatives of this type would certainly need a detailed familiarity with the IEEE's policies to operate effectively.

In order to advise the Board of Directors on intersociety matters, there is an important standing committee—the Intersociety Relations Committee (ISRC)—which reports directly to the Board. It consists of nine members plus a chairman, who is usually the Junior Past President of the Institute. The members are selected on the basis of their prior experience in intersociety federations or groups. This committee, with the assistance of the Technical Activities, Awards, and Publication Boards, has the difficult task of submitting to the Executive Committee nominations for IEEE representatives. If this job is to be done well, a knowledge of the competence of the individual, and also of the characteristics of the organization to which the appointment is made, is required.

The ISRC recently undertook a study of the various organizations with which the IEEE has relations and is making an attempt to categorize them in a meaningful way. An attempt also is being made to clarify the position of appointed IEEE members so that there are no misunderstandings about the extent to which they speak as individuals or for the Institute.

Despite the obvious problems involved, it is desirable, and indeed essential, for the IEEE to take an active role in association with other organizations. There are matters of great importance to the IEEE that overlap the areas of concern of other professional societies in significant and diverse areas. Continuing education, uniform indexing and abstracting procedures, standardization of procedures and nomenclature, and career guidance in secondary schools are examples of these. To contribute effectively in such areas, the IEEE through its representatives must be able to handle questions of both technical and policy natures. The development of appropriate means for forwarding IEEE objectives in conjunction with other organizations is a continuing problem that needs a continuing good solution.

F. Karl Willenbrock

Authors

The Northeast power failure—a blanket of darkness (page 54)

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

New horizons in semimetal alloys (page 74)

Leo Esaki (F) received the B.S. degree in 1947 and the Ph.D. degree in 1959, both in physics, from the University of Tokyo. He joined the staff of the Sony Corporation, Tokyo, in 1956, where he led a semiconductor research group. During this time his tunneling studies in p-n junctions resulted in his discoveries of the backward diode and, subsequently, of the tunnel diode (in 1957), which is also widely referred to as the Esaki diode. In 1960 he came to the United States to join the the IBM Corporation's Thomas J. Watson Research Center, Yorktown Heights, N.Y. Since then he has been active in work on fast-switching devices, such as heterojunctions, and has conducted studies of semimetals and narrow-gap semiconductors. He is now in charge of the Applied Physics Group in the Solid State Electronics Department. His work in semimetals led to the discovery of the "kink" effect and to successful results in tunneling spectroscopy.

Dr. Esaki has received numerous awards and prizes, chiefly for his discovery of the tunnel diode. His Japanese honors include the Nishina Memorial Award (1959), the Asahi Press Award (1960), the Tokyo Rayon Foundation Award for the Promotion of Science and Technology (1961), and the Japan Academy Award (1965). In the United States he received the IRE's Morris N. Liebmann Memorial Prize (1961) and the Ballantine Award of the Franklin Institute (1961). He is a Fellow of the American Physical Society.

International developments in controlled nuclear fusion (page 87)

Arwin A. Dougal (SM) received the B.S. degree in electrical engineering from Iowa State College in 1952. He previously had served in the Airways and Air Communications Service of the U.S. Air Force, in which capacity he contributed to the development of radar-controlled and scheduled all-weather transport flight, and to the development and evaluation of the USAF-RCA Teleran system of television-radar air navigation. In 1955 and 1957, respectively, he received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois. He served as research assistant, research associate, assistant professor, and associate professor in electrical engineering at the University of Illinois from 1952 to 1961. In 1961 he was appointed professor of electrical engineering at the University of Texas, a position he currently holds. In addition, he is the director of the Laboratories for Electronics and Related Science Research.

Dr. Dougal has contributed original research publications in the fields of microwave gaseous electronics, electron-ion plasma studies, physical electronics, electromagnetic theory and boundary-value problems, MDH and controlled thermonuclear fusion, and quantum electronics-optical masers. He has served as a consultant to the University of California's Los Alamos Scientific Laboratory, the University of Illinois' Coordinated Science Laboratory, Texas Instruments Inc., and the General Dynamics Corporation. He is a founding member of the Society of Engineering Science and a member of the American Physical Society. This past summer he was the guest of a number of national research institutes and universities in western Europe, where he lectured on his current research work.

Radar separation of closely spaced targets (page 94)

A. Golden (M) received the B.S. degree in electrical engineering from the City College of New York in 1958 and the M.S. degree, also in electrical engineering, from the Drexel Institute of Technology, Philadelphia, Pa., in 1962. In 1959 he joined the staff of the Missile and Surface Radar Division of the Radio Corporation of America, in Moorestown, N.J. His initial work there was concerned with the design and development of solid-state circuitry to be employed on radar receivers. He subsequently became involved in the analysis and design of radar systems.

In 1964 Mr. Golden joined the faculty of Drexel Institute of Technology, where he had received an appointment as an instructor in the Department of Electrical Engineering. Upon his return to RCA in 1965 he became a systems engineer in the Air Defense Systems group. At the present time he is working on the synthesis of an advanced tactical air defense system.

Correlative level coding for binary-data transmission (page 104)

Adam Lender (SM) received the B.S. and M.S. degrees in 1954 and 1956, respectively, and the Electrical Engineer degree in 1960, all from Columbia University. From 1954 to 1960, as a member of the technical staff of Bell Telephone Laboratories, Murray Hill, N.J., he was concerned with the exploratory development of high-speed digital data transmission systems. From 1960 to 1961 he was a project engineer at ITT Laboratories, Nutley, N.J., in charge of a group working on time-division multiplexing, pulse code modulation, and delta modulation for secure voice communications. He is now with Lenkurt Electric Co., Inc., San Carlos, Calif., where he heads the Digital Techniques Section of the Advanced Development Laboratoy and is engaged in research in digital data transmission techniques, error control, and digital voice communications.

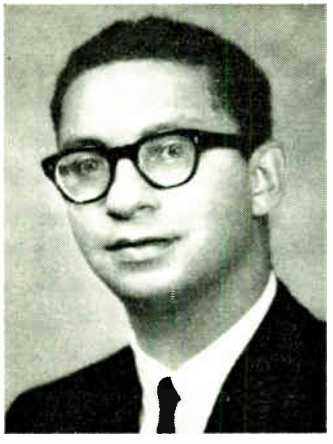
Mr. Lender has written several papers on analog-to-digital conversion and digital data communication. He is a member of Tau Beta Pi and Eta Kappa Nu.

Search methods used with transistor patent applications (page 116)

June Roberts Cornog is a research psychologist, with a "home base" at the National Bureau of Standards. She is a consultant to several federal agencies on research problems involving the interface between men and computers, and has worked in the areas of keyboard design, coding of materials for computer storage, and the mental processes that people go through in using computers for various purposes.

Dr. Cornog was graduated from Pennsylvania State University in 1955, with a major in industrial psychology and a minor in clinical psychology and labor relations. She was formerly a college professor at various small colleges in Pennsylvania and subsequently was a test instructor in the U. S. Naval Air Training Command for two years. In addition, she worked for three years with National Analysts in Philadelphia as a specialist in consumer motivation.

Herbert L. Bryan, Jr., joined the U.S. Patent Office, Washington, D.C., in 1962 as a student trainee patent examiner, assigned to the Research and Development Branch, while he was attending Howard University. After he received the B.S.E.E. degree from Howard in 1963 he was promoted to the position of patent examiner. During 1963–1964 he attended the Patent Office Academy for an advance course in patent examining. Since the completion of this course he has been involved in the application of his knowledge of patent examining to the development of a nonconventional patent search system for transistor and nonlinear-conductor systems. In addition, he is engaged in work leading to the eventual automatic indexing of the full text of patents, as well as to the automatic indexing of schematic diagrams, by electronic computers. He is on the executive board of the Patent Office Society.



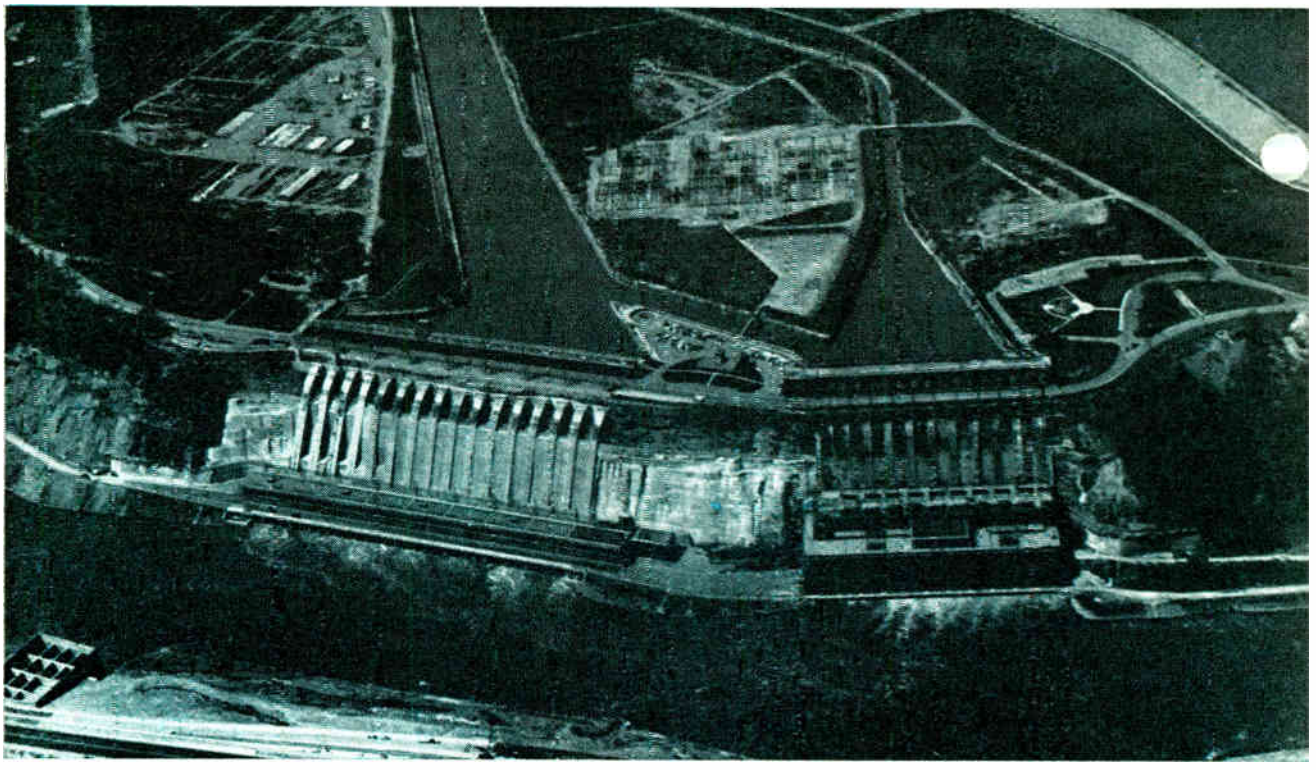


Fig. 1. Aerial view of Ontario Hydro's Sir Adam Beck-Niagara Generating Stations Nos. 1 and 2, situated on the Niagara River near Queenston, Ont. The Beck No. 2 plant, where the extensive Northeast blackout was triggered, is in the lower left foreground. This station has a capacity of more than 1200 MW. The switchyard is shown in the center of the picture and the associated pumping-generating station may be seen in the upper right background.

The Northeast power failure—a blanket

The massive outage in the Northeast on the night of November 9–10 was a critical episode in the annals of electrical engineering. Here, in brief, is a chronicle of what happened, and a preliminary view of the issues involved

Gordon D. Friedlander *Staff Writer*

The massive blackout of November 9–10 has triggered some prompt and expeditious action, spearheaded by the Federal Power Commission's list of 19 affirmative recommendations that should be implemented by the utilities to preclude the possibility of a recurrence. These proposals include the need for fully coordinated power pools, more extensive stability studies for the operation of interconnected systems, the need for mixed generation facilities for quicker response to power swings, and a review of the adequacy of existing automated equipment.

The IEEE has a unique role in the wake of the November 9 power failure. Among its members are those qualified to determine the technical changes which should be made to the nation's power supply system in the light of this occurrence. Its technical meetings and publications will provide a forum for a full discussion of power supply problems. The Institute also has a responsibility to report the essential facts of the incident and to indicate—to the extent possible at this time—the issues involved.

The principal published source of information on the events of November 9 is the Federal Power Commission's report to the President, *Northeast Power Failure*, dated December 6, 1965, and available from the United States Government Printing Office. Other authoritative background material includes the *Minutes of Hearing of November 20, 1965*, before the New York State Public

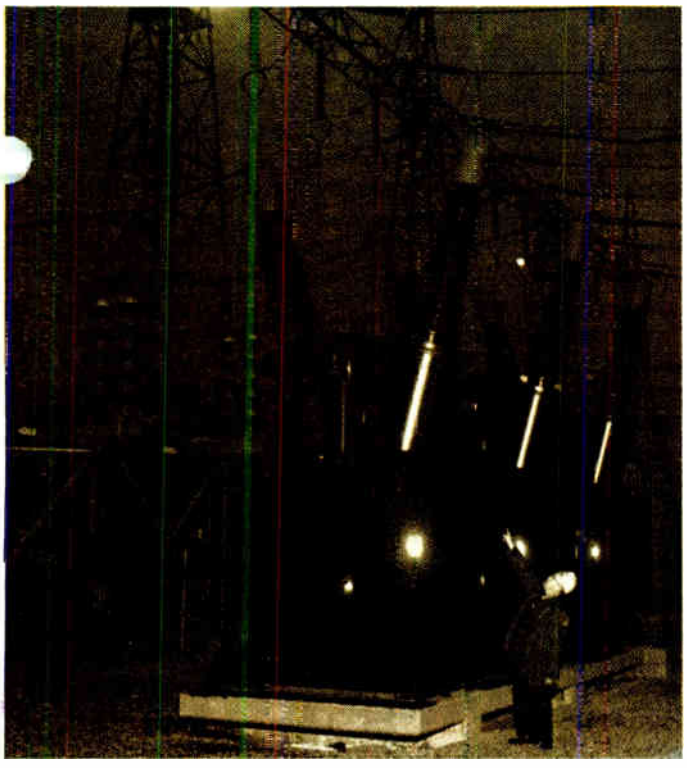


Fig. 2. Circuit breakers, of the type that tripped out during the disturbance on November 9, are shown in the Sir Adam Beck No. 2 generating station switchyard.

# of darkness

Service Commission, and the official statement, dated November 19, 1965, of the Power Authority of the State of New York.

The writer has drawn heavily on the FPC report, has talked with many knowledgeable persons in the power industry, and has had his final manuscript reviewed by more than 15 experts in the power field. There are legitimate areas of controversy about what can and should be done to modify the nation's power supply system. The purpose of this article is to provide a point of origin for the professional discussion that is required. To this end, responses from the membership of IEEE are invited and will be published ensemble in a forthcoming issue of SPECTRUM.

Chronology of a disturbance

After five days of intensive investigation by United States and Canadian officials, an advisory panel of power experts, special consultants from the electric power industry, the FBI, and the Department of Defense, following the great Northeast blackout that affected 28 of the operating utilities (see map on p. 56) of the Canada–United States Eastern Interconnection (CANUSE), circumstantial and recorded evidence pinpointed the disturbance, which triggered the biggest power failure in North American history, at the Sir Adam Beck No. 2 power generation complex (Figs. 1 and 2) in Queenston, Ont.

The CANUSE network. By way of background for the description of the sequential events of November 9, an introduction to the CANUSE network operation and facilities is essential. At the outset, it should be noted that this network, according to the FPC report, is not yet a fully integrated power pool. The term *power pool* can be misleading since it sometimes loosely describes a group of member utilities that are interconnected but have not necessarily attained the optimum economies of power interchange or the increased service reliability that is realized by a fully integrated power pool. In the CANUSE service area (CANUSE map and Fig. 3), 73 percent of the power generation is produced by fossil-fuel thermal stations and 26 percent is produced by hydroelectric plants. The remainder of the generation—one percent—is from three nuclear plants, and gas turbine and diesel-electric generating equipment.

The bulk of the hydroelectric generating capacity is concentrated in the Niagara Falls area in plants of The Hydro-Electric Power Commission of Ontario (Ontario Hydro) and the Power Authority of the State of New York (PASNY). In addition, at Massena, N.Y., on the St. Lawrence River, PASNY and Ontario Hydro (Fig. 4) separately operate hydro plants, each having a firm capacity in the 700–800-MW range.

A major portion of the hydro plant power generation at Niagara and St. Lawrence is transmitted to load centers that are located a considerable distance from these stations. For example, the primary load center of the Ontario Hydro system is near Toronto, the largest city in the province. Power generated by the PASNY hydro stations is transmitted to a major extent over twin 345-kV transmission lines (there is a single circuit, however, as shown in Fig. 7, paralleled by two 230-kV circuits, between Edic and New Scotland) that extend across New York State from Niagara eastward to Albany and thence south to New York City. These EHV lines—which are, at once, the backbone and the weakness of the interconnected systems in New York—overlay a number of 115- and 230-kV transmission lines from the PASNY plants at both Niagara and Massena. It should be noted that the capacity of a transmission line increases approximately as the square of the voltage. Thus a 345-kV line has about 2.3 times the capacity of a 230-kV line.

A single 345-kV line, a 230-, and five 115-kV lines (Fig. 3) form the transmission interconnection between Niagara Mohawk and PASNY and the four utility companies of the Connecticut Valley Electric Exchange (CONVEX) and the three utilities of the New England Electric System.

At the southern end, the CANUSE system has interties with the Pennsylvania–New Jersey–Maryland (PJM) pool by seven transmission lines, ranging from 115 to 230 kV. One of these tie lines—which figured prominently in maintaining some isolated service in New York City during the massive outage—links the Consolidated Edison system with those of PJM through the Arthur Kill generating station on Staten Island.

The largest of the 28 CANUSE utilities include Ontario

Hydro, PASNY, The Niagara Mohawk Power Corporation (Niagara Mohawk), and the Consolidated Edison Company of New York (Con Edison). The generating capabilities, system loads, and other quantitative data, just prior to the disturbance of November 9, for the principal CANUSE utilities, plus PJM, are shown in Table I.

Beck generating station—5:00 p.m., November 9. According to Ontario Hydro, when the new power schedule was set at 5:00 P.M., the utility was meeting its system load of about 6400 MW (Table I) by generation of 1335 MW from the Beck No. 2 plant and the associated pumping–generating station, plus an inflow of about 500 MW on two tie lines with New York (Fig. 5). One line, designated PA27, runs to the Moses-Niagara plant of PASNY (296 MW); the other, called BP76, connects to the Packard station of Niagara Mohawk (200 MW). Of the 500-MW inflow, 200 MW was a clockwise power circulation (Fig. 4) that was being returned through the tie line from Cornwall, Ont., to New York State. By referring to Fig. 5, it may be seen that power flows north from the Beck No. 2 station over five 230-kV lines that connect the plant with the load center in the Toronto area.

The incident. At 5:16:11 P.M., a backup relay, protecting line Q29BD (indicated in Fig. 5), operated normally and caused the circuit breaker at Beck to trip the unfaulted line. The power flow on the disconnected line shifted to the remaining four lines, each of which then became loaded beyond the critical level at which its backup protective relay was set to function. Thus the four remaining lines tripped out in cascade in 161 cycles' time (2.7 seconds).

The relay that triggered the disturbance was one of five backup sensing devices (one backup relay per line) that protect the lines against failure of the Beck primary relays, or of circuit breakers at remote locations. According to the FPC report, the five backup relays were installed in 1951 and, in 1956, a breaker on one of the 230-kV lines failed to open (reason not explained) following a fault. In January 1963, as a result of a re-evaluation study of its backup protection requirements, Ontario Hydro modified these relay settings to increase the scope of their protective functions.

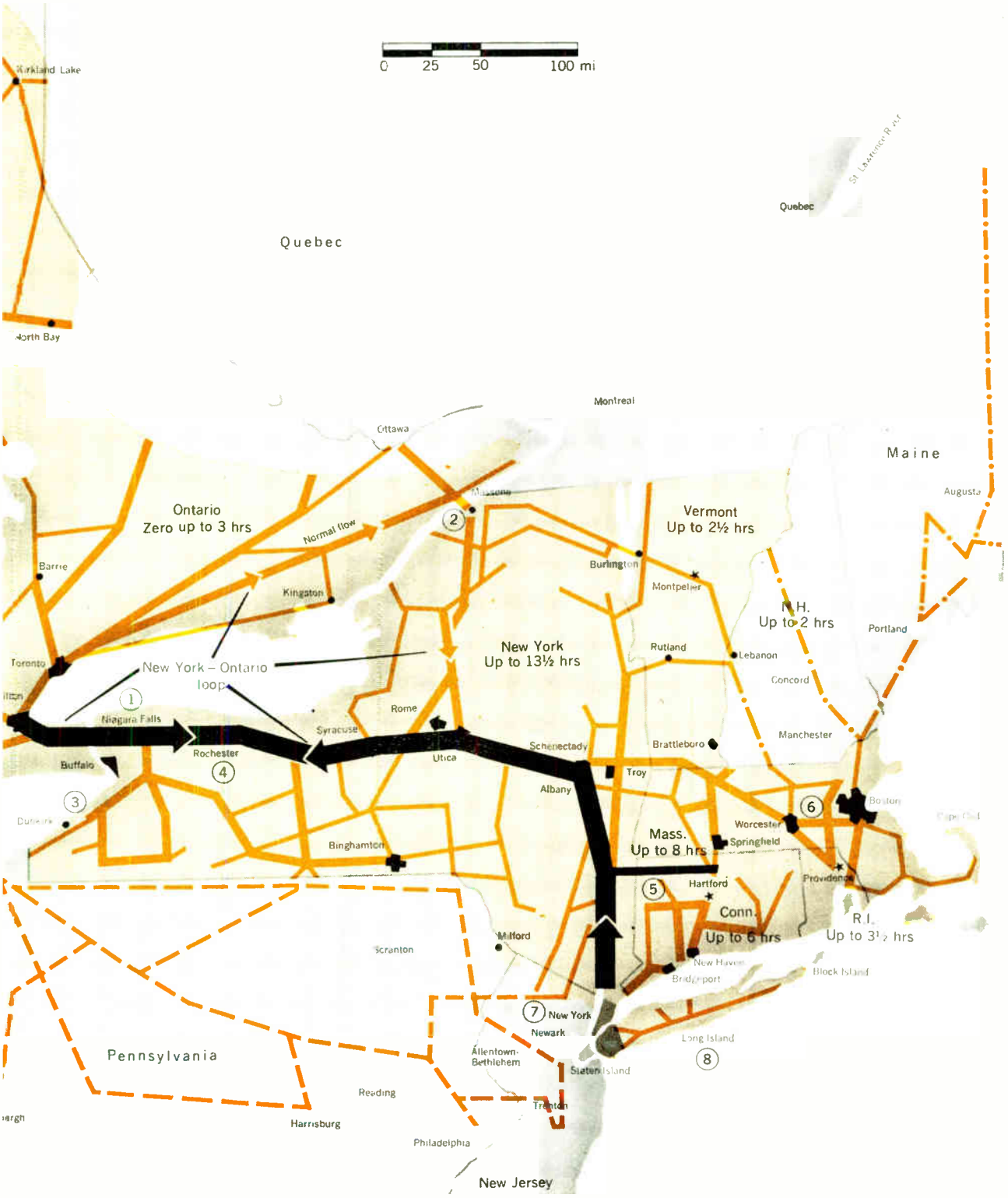
Figure 6 indicates the set of conditions under which this type of relay would trip. The evidence suggests that, at 5:16:11, the load and generation characteristics of the Canada–United States interchange caused such a condition to be reached.

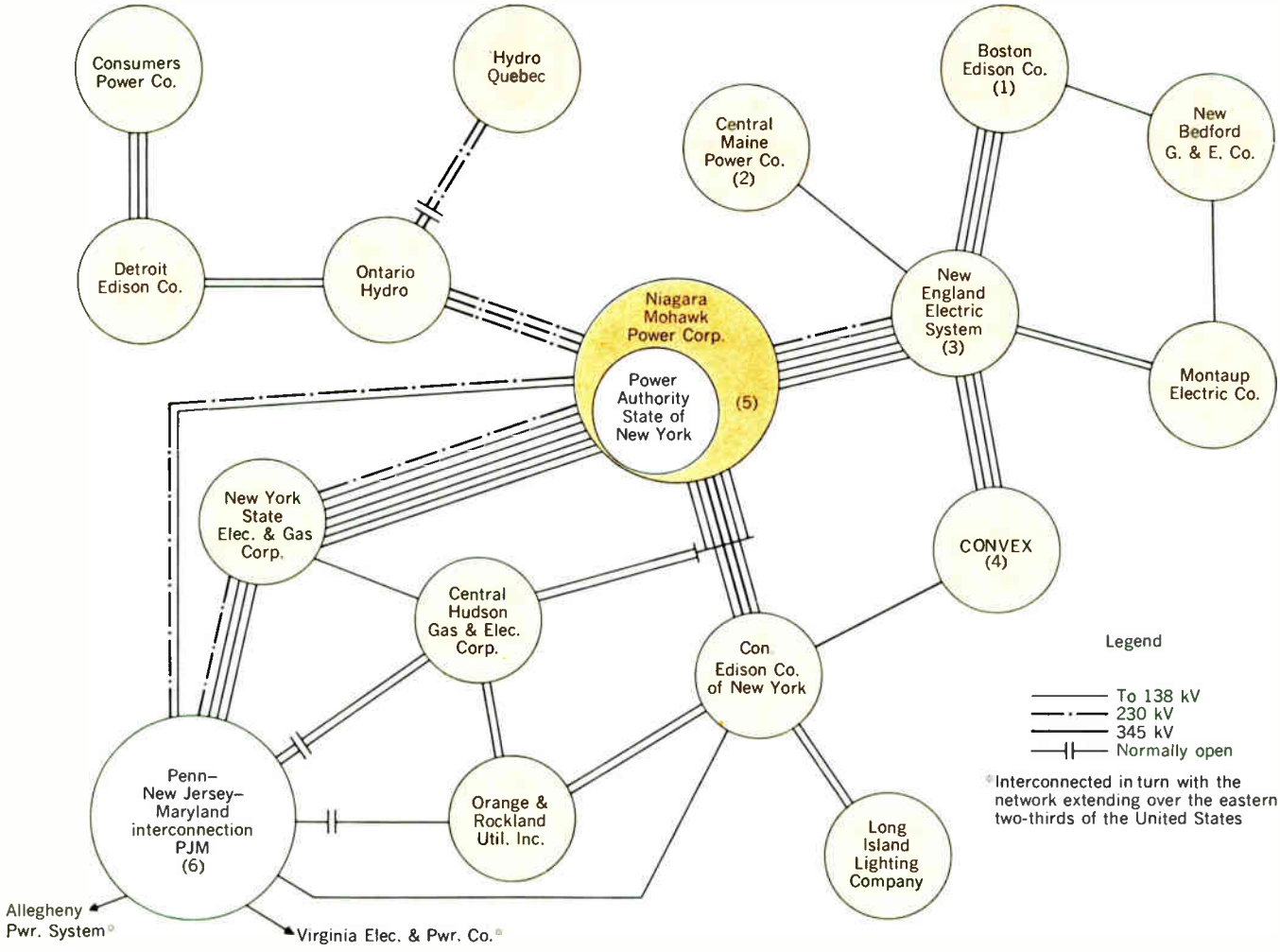
The FPC report further states that the relay settings made in 1963 at the Beck plant were in effect at the time of the November 9 power failure. The backup relay on line Q29BD was set in 1963 to operate at about 375 MW and 160 Mvar at a bus voltage of 248 kV and, although the load-carrying capacity of each of these lines is considerably higher, it was necessary to set each backup relay to operate at a power level below the line's capacity to provide the desired protection and to achieve coordination with other relays on the system. This setting was believed to be sufficiently high to provide a safe margin above expected power flows.

When the backup relays were modified and the power levels were set in 1963, the load on the northbound lines from Beck No. 2 was appreciably lower than the trip setting of the backup relay. Recently, the megawatt and megavar loadings on the transmission lines from Beck

Simplified map of most of the CANUSE system, plus the PJM pool (Michigan interconnections, with the exception of the ties at Windsor and Sarnia, are not indicated). The heavy black lines show the 345-kV EHV transmission trunk routes (the arrows indicate the initial eastward surge following the disturbance, then the flow reversal toward th north and west as the power demand hit the southern utilities in the CANUSE area), while the colored lines represent 115-, 138-, and 230-kV transmission. Also shown is the New York–Ontario clockwise circulation loop. The

dot-and-dash colored lines show portions of the New England systems unaffected by the blackout. Dashed colored lines indicate the principal high-voltage (up to 230 kV) routes of the PJM pool. The shaded area denotes the extent of the blackout, and approximate outage times for various areas are given. Circled numbers indicate chronological events, beginning with the original disturbance (time divisions are—hours:minutes:seconds): 1—Initial disturbance is triggered at 5:16:11 P.M. by backup relay on line Q29BD at the Sir Adam Beck No. 2 switchyard. 2—At 5:16:15 the two 230-kV lines, connecting the PASNY St. Lawrence plant (Massena), trip out. 3—Thermal plants in the Niagara–Dunkirk area shut down at 5:18. 4—Rochester loses its load at 5:19. 5—Generators in CONVEX group shut down between 5:19 and 5:30 because of inadequate power supply for station auxiliary equipment. 6—Power fails in the Worcester–Boston area from 5:18 to 5:21. 7—Con Edison system, with exception of Arthur Kill station, trips out at 5:27. 8—Long Island Lighting Co., last to be affected, shuts down at 5:28.



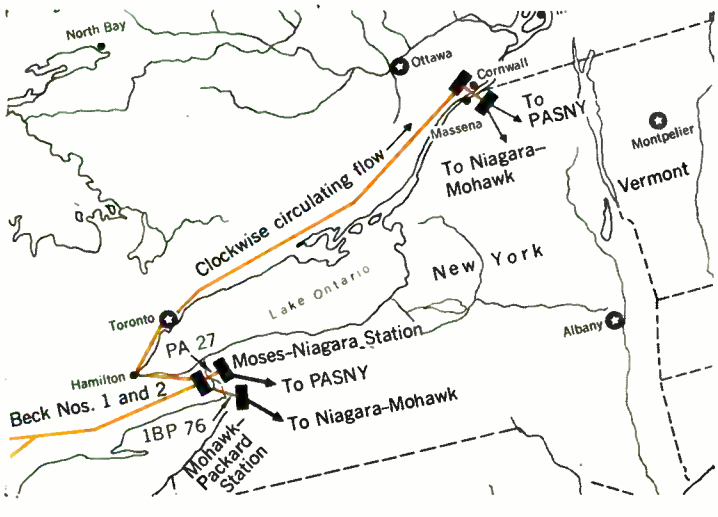


Numbers refer to following list of "systems normally controlling sublisted systems"

1. Boston Edison Co.
   Cambridge Electric Co.
2. Central Maine Power Co.
   Bangor Hydro Electric Co.
3. New England Electric System
   Central Vermont Public Service Corp.
   Citizens Utilities Co.
   Green Mountain Power Corp.
   Public Service Co. of New Hampshire
   Fitchburg Gas & Electric Co.
   Yankee Atomic Electric Co.
4. Connecticut Valley Electric Exchange–CONVEX
   Connecticut Light & Power Co.
   Hartford Electric Light Co.
   United Illuminating Co.
   Western Massachusetts Electric Co.
5. Niagara Mohawk Power Corp.
   Rochester Gas & Electric Co.
   Power Authority State of New York
   New York State Elect. and Gas Co. (Part)
6. Pennsylvania–New Jersey–Maryland Interconnection–PJM
   Atlantic City Electric Co.
   Baltimore Gas and Electric Co.
   Delaware Power & Light Co.
   Jersey Central Power & Light Co.
   Lucerne Electric & Gas Div.
   United Gas Improvement Co.
   Metropolitan Edison Co.
   New Jersey Power & Light Co.
   Pennsylvania Electric Co.
   Pennsylvania Power & Light Co.
   Philadelphia Electric Co.
   Potomac Electric Power Co.
   Public Service Electric and Gas Co.

Fig. 3. Diagram of load control areas and power system interconnections, CANUSE and PJM.

Fig. 4. Map showing simplified portion of Ontario Hydro's southern transmission system and the major interties to the United States at Niagara and Cornwall.



to the north, because of emergency outages in a new Ontario Hydro steam-electric plant, have been very heavy. This temporary situation produced a deficiency in Ontario generation, with the result that a heavier inflow of power from the United States interconnections was necessary.

According to Ontario Hydro spokesmen, the average flow had reached 356 MW (and approximately 160 Mvar) in the line that tripped out first, but momentary fluctuation in flow is normal. Therefore, at 5:16 P.M., as already mentioned, the power flow apparently reached the level at which the relay was set; it functioned in accordance with its setting, and its circuit breaker tripped out the

I. Northeast power loads and resources just prior to disturbance of November 9

Utility	System Load, MW	Generating Capability, MW	Resources Used to Meet Load: Generation, MW	Resources Used to Meet Load: Net Received, MW	Resources Used to Meet Load: Total, MW
Niagara Mohawk Power Corp.	3 405	2 681	2 556	+849	3 405
Rochester Gas & Elec. Corp.	500	350	296	+204	500
PASNY[1]—Moses-Niagara	0	2 500	2 274	−2 274	0[2]
PASNY[1]—Moses-St. Lawrence	480	800	700	−220	480[2]
New York State Elec. & Gas Corp.	1 035	569	547	+501	1 046
Central Hudson Gas & Elec. Corp.	335	309	282	+53	335
Consolidated Edison Co.	4 770	5 896	4 555	+215[4]	4 770
Long Island Lighting Co.	1 289	1 442	1 197	+90	1 289
Orange & Rockland Utilities	232	144	121	+117	238
Hydro-Elec. Pwr. Comm. Ont.	6 400	6 750	6 100	+300	6 400
CONVEX	2 626	2 685	2 583	+43	2 626
Vermont Elec. Pwr. Co., Inc.[3]	306	0	0	0	0[2]
New England Elec. System	1 300	1 804	1 642	−342	1 300
Public Service Co. of N. H.	410	385	370	+40	410
Boston Edison Co.	1 222	1 578	1 405	−154	1 351
Central Vermont Public Service Co.	150	36	26	+124	150
Pa.-N. J.-Md. Interconnection	13 600	14 451	13 355	+245	13 600
Detroit Edison Co.	3 196	3 280	3 050	+146	3 196
Consumers Power Co.	2 161	2 668	2 294	−121	2 173
Central Maine Power Co.	471	581	490	−19	471
Total	43 582	48 909	43 843	—	—

[1] Power Authority State of New York.
[2] Wholesale supplier.
[3] Distributes PASNY power in Vermont. Not included in totals to avoid duplication.
[4] Outflow to Long Island not deducted.

line. Ontario Hydro also informed the FPC that its operating personnel were not aware that the relay on line Q29BD was set to operate at a load of 375 MW.

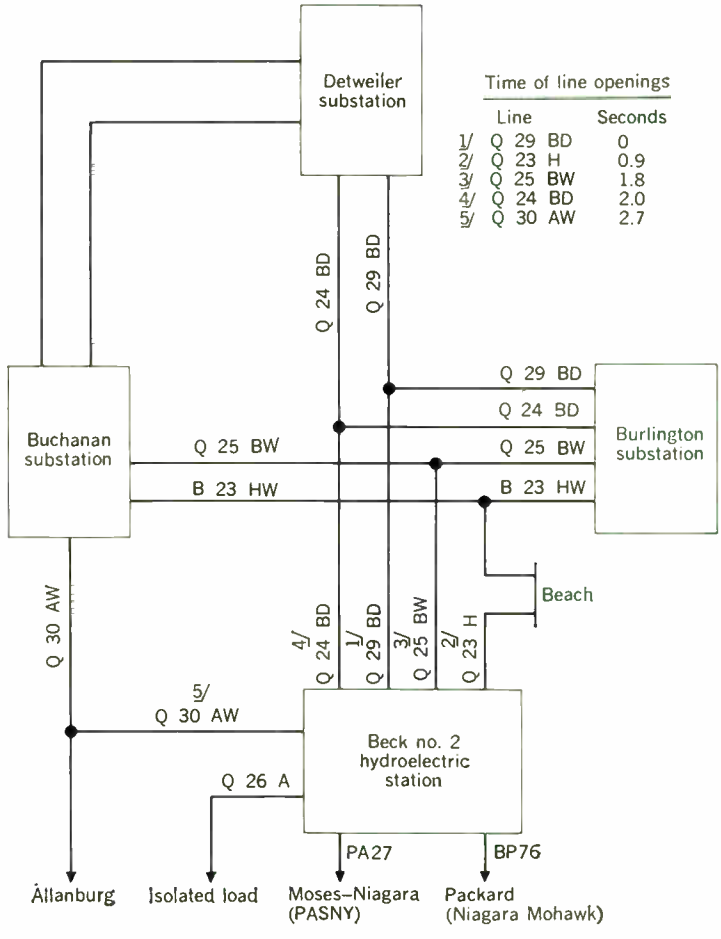
The initial power surge

The outage of the five 230-kV lines from Beck No. 2 separated the Ontario generation along the Niagara Frontier from the loads in the province. Immediately before the disturbance, about 1800 MW of power, generated at PASNY's Niagara plant, was flowing east and south over the New York State transmission system.

Quoting from the FPC report: "With the dropping of the lines to Toronto, the power being generated at the Beck plant and at PASNY's Niagara plant, which had been serving the Canadian loads around Toronto, amounting to approximately 1500 MW, reversed and was superimposed on the lines to the south and east of Niagara. It was this tremendous thrust upon the transmission system in western New York State which exceeded its capability and caused it to break up. (According to Niagara Mohawk, this massive surge of power was due to the loss of a very large block of load—a most unusual occurrence on a power system. Normally, protection is designed to guard against a loss of generation, which is a more usual occurrence.)

"The instantaneous result of the tripping of the lines from Beck to the Toronto area was the acceleration of the generators at Beck and PASNY-Niagara, with a sharp drop in their electrical output, but as the speed increased the electrical power output at the two plants rapidly increased. The instantaneous drop in generation at Beck and PASNY, followed by the rapid buildup, resulted in putting the Beck and PASNY generation out-

Fig. 5. Block diagram of the 230-kV transmission system in vicinity of the Beck No. 2 station. Relay designations, starting with Q29BD and time of line openings from original disturbance, are indicated.



of-phase with most of the other generation attached to the interconnected transmission system. . ."

This simple power-limit pullout was the prime cause of the ensuing chain reaction of power failures.

Impact on the East. One-half second after the last of the five lines at Beck tripped out in cascade, the Cornwall–Massena intertie between PASNY and Ontario Hydro became overloaded and was opened by its protective relay. Thus the Ontario system was isolated from the New York systems, except at Niagara where the Beck plant remained connected to the New York State system but was isolated from the rest of Ontario.

The PASNY statement. The Power Authority of the State of New York issued an official statement, dated November 19, of the events that affected its generating stations and transmission lines on the evening of November 9. Quotations from this statement may be of interest as additional information.

"At 5:15 P.M. . . Authority facilities were operating normally with no schedule changes, no switching and no relay tests in progress. Authority's generating facilities were producing 2977 MW with loading on all lines well within established limits.

"At Niagara Falls we were producing 2275 MW at the time of the disturbance. Our tie with The Hydro-Electric Power Commission of Ontario at Niagara was carrying 260 MW from the Authority's switchyard into [the Ontario Hydro] system, our two 345-kV lines to Rochester and eastward were delivering 840 MW to interconnections served by these lines. . .

"At St. Lawrence (Massena) we were producing 702 MW at the time of the disturbance. Our tie with [Ontario Hydro] . . . at Massena was carrying 190 MW [from Canada] into our system. . .

"Immediately following the original disturbance the load on the [Ontario Hydro] tie [at Niagara] reversed from 260 MW toward Canada to over 400 MW toward New York. The loading on the two 345-kV lines increased from 840 MW to more than 1200 MW; loading on New York State Electric & Gas Corporation's Stolle Road line [New York southern tier] decreased from 155 MW to 75 MW while delivery to Niagara Mohawk Power Corporation over the 230-kV and 115-kV ties decreased . . . from 1019 MW to 812 MW.

"As a result of the disturbance the circuit breakers in the Authority's 230-kV circuit to [Ontario Hydro] tripped by . . . relays, reclosed once and then relayed to lockout . . . Eight of the nine operating units at the Lewiston [Niagara] plant tripped off by overspeed relay operation, dropping a total of 218 MW. All of the twelve operating units at the Moses [Niagara] plant remained in operation but governor action reduced the generated output from 2030 MW to a low of 400 MW. The plant load was manually stabilized at approximately 1500 MW in about ten minutes."

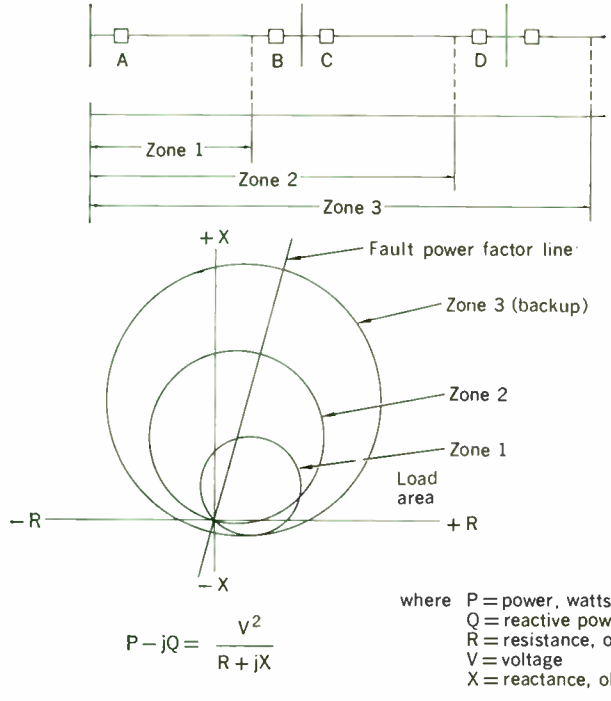
The chain reaction continues. Since the output from Beck No. 2 and the PASNY-Niagara plant could not be handled by the remaining transmission system after the tripout of all five of the lines at Beck, the EHV transmission system was severed 0.9 second later with the stability-limit opening of the two 345-kV lines between Rochester and Syracuse, N.Y. Almost simultaneously, tie lines to the PJM pool were broken both in the Niagara Mohawk area (Fig. 3) and in Brooklyn at the Con Edison side (Greenwood Station) of a connection with PJM (through Public Service Electric and Gas Company) at Staten Island. Figure 7 shows the lines that were forced out because of the initial disturbance. Typical relays in use in the CANUSE area are shown in Fig. 8.

At 1.33 seconds after the separation of the Beck generators from their loads, the two 230-kV lines (CANUSE map and Fig. 7), connecting the PASNY plant at St. Lawrence (Massena) with the 345-kV trunk lines that run from Niagara to downstate New York, and New England, tripped out. This action simultaneously tripped out five of the 16 generators at PASNY's St. Lawrence installation. The automatic tripout of these generators occurred in accordance with preplanned operating procedures, since this hydroelectric station was designed so that the remaining generators could continue to supply, by independent radial supply circuits, the large industrial loads to the industrial plants in the Massena vicinity.

The generators at the Beck plant were not provided with relays to trip them out upon loss of a transmission line, since such a scheme would have been ineffective in the face of the large increase in inflow over the Niagara ties which would follow and offset the loss of the tripped generation. The simultaneous loss of the five transmission lines to the north was considered to be an improbable contingency.

Within four seconds after the initial tripout at the Beck station, most of the CANUSE area east of Michigan

Fig. 6. A—Line diagram showing primary and backup circuit breakers in three protective zones for controlling breaker A. In operation, either the primary protection (breaker A) trips instantaneously for fault in Zone 1, or, in case of faults in the backup protection Zones 2 or 3, breaker A will trip after predetermined time delays, if the fault is not cleared by the remote breakers. B—Typical mho (impedance) distance relay characteristic is shown plotted for Zones 1, 2, and 3. General equation indicates reactance "seen" by relay.



(Maine and a portion of New Hampshire did not lose power) was broken into four isolated segments:

1. The Ontario Hydro system was completely separated from New York and was badly deficient in generation capability.

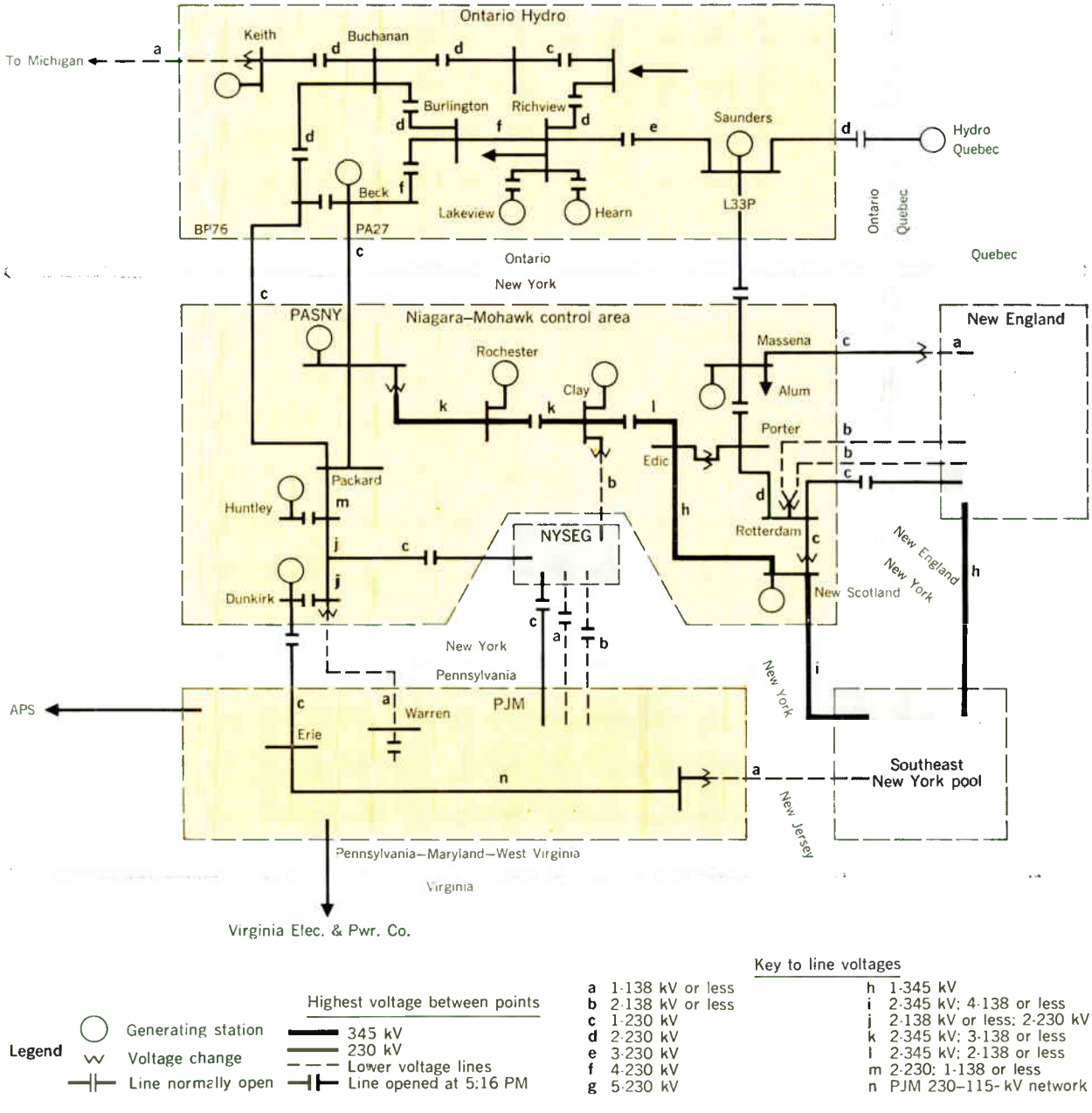
2. The northern New York region, supplied by PASNY-St. Lawrence and Niagara Mohawk's northern hydro sources, was isolated, but the remaining generation was able to carry loads in the Massena, Potsdam, Watertown, and Oswego areas.

3. The region in the vicinity of Niagara, on the American side of the international boundary (Niagara–Dunkirk area), including the New York State Electric and Gas south central area, was separated from the remainder of the interconnection and had large excesses of generation.

4. The balance of the CANUSE area, including a part of upper New York, the New England systems, and the systems in the southeast New York region, was separated from the rest of the group but remained interconnected within itself. Michigan was temporarily separated from all but a small section of the Ontario Hydro system. Figure 9 indicates the definitive areas of separation.

Actually, area 1—the Ontario system—divided into three subsections and dropped about 3800 MW of load in this process. As it was now isolated from New York, it no longer had any effect on the systems in the northeastern area of the United States. After a separation of 17 minutes, the Detroit Edison Company reclosed its interconnection to Ontario Hydro at Windsor. Its other interconnection at Sarnia remained closed to assist in maintaining service to the southwestern peninsula of Ontario. Also, a large section of eastern Ontario remained intact.

Fig. 7. Initial transmission line outages, Northeast blackout of November 9.



Shutdown of the Niagara area generating plants

The excess of generation in area 3 (Niagara–Dunkirk) caused all of its hydroelectric and steam-electric generators to accelerate, accompanied by a rise in frequency. The steam plants, including the large Niagara Mohawk stations at Huntley and Dunkirk, shut down by governor action because extensive mechanical damage could have been done to turbine blades by overspeed.

The thermal plant shutdowns in area 3 were followed in about 1½ minutes (at approximately 5:18 P.M.) by the tripping of ten generating units at Beck because of low governor oil pressure (the result of excessive governor operation), and five pumping–generating units in the PASNY-Niagara station were closed down by overspeed governor control.

Because of these sequential shutdowns, the load in the area now exceeded the remaining power supply. The frequency dropped to 58.5 c/s, and two 230-kV lines between Beck No. 2 and PASNY-Niagara opened by underfrequency relay action.

The two lines had remained closed during the initial surge of power to the United States because they had inverse time relays that were set to trip out at 864 MW, and they could not function within the 0.9-second interval before the breakup of the 345-kV transmission system. And after the initial disturbance, the southward flow from Beck *did not exceed 864 MW* for the required time to trigger the relays.

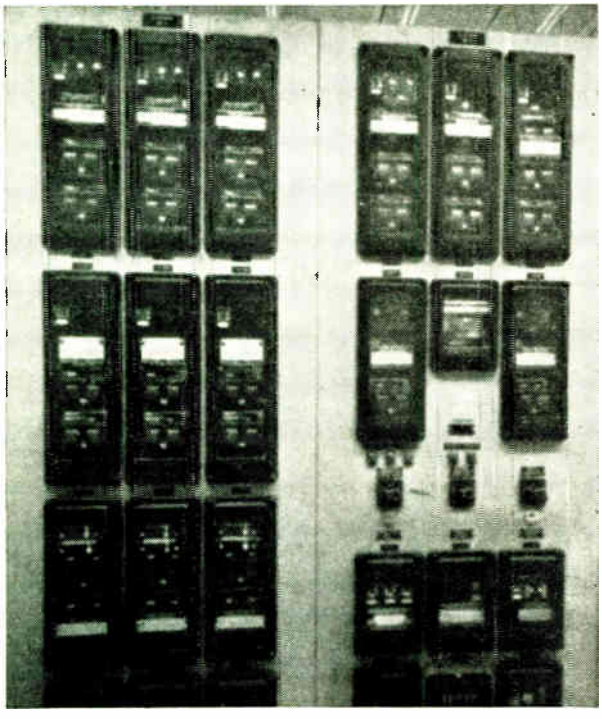
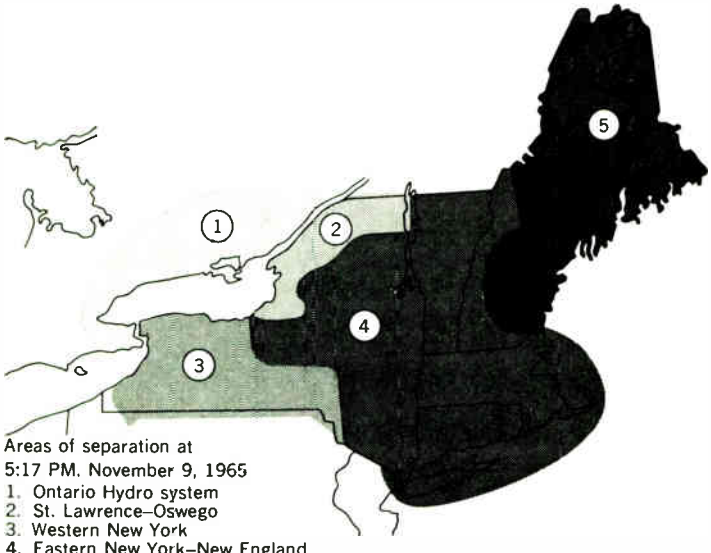


Fig. 8. Relays on the Clay–Edic 345-kV transmission line of the Niagara Mohawk Power Corp. are typical of those used throughout the area.

Fig. 9. Map showing the service area separations during the Northeast blackout.



Areas of separation at
5:17 PM. November 9, 1965
1. Ontario Hydro system
2. St. Lawrence–Oswego
3. Western New York
4. Eastern New York–New England
5. Maine and part of New Hampshire

Note: Some intertie lines in some of the areas were open

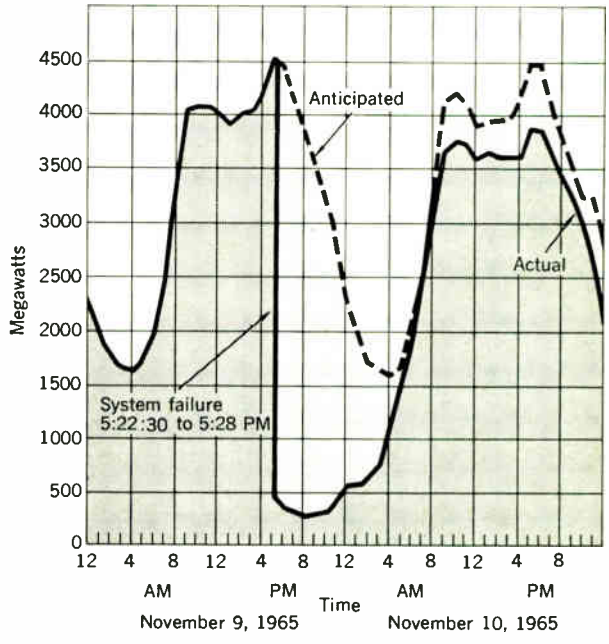
Breakdown in area 4

Referring to Fig. 9, it may be seen that area 4 encompasses the largest region of the massive outage, from the north central part of New York to the state's southeastern extremity and the major portion of a group of five New England states. Just before the disturbance, power inflows were being drawn from the CANUSE system in the amounts of 140 MW by New England and 400 MW by the downstate New York area.

In the four-second time interval following the initial disturbance at Beck, in which the PASNY and Niagara Mohawk systems had split, the upstate New York area, east of this point of rupture, was still tied to the New England and the southeastern New York areas. This situation created an instantaneous deficiency in generation for area 4 of approximately 1100 MW.

Minutes of a hearing. On November 20, the vice

Fig. 10. Graphic plot of Con Edison's actual and anticipated net system load during the 48-hour period of November 9 through 10, 1965.



president of operations for Con Edison testified at a hearing before the New York State Public Service Commission. According to this spokesman, Con Edison, at 5:00 P.M. on November 9, was generating about 4550 MW from its own system and drawing 220 MW from its interconnections to meet a load of 4770 MW. At 5:16 P.M., the Energy Control Center in Manhattan noted a sharp surge of power into the system, which lasted somewhere between 50 and 90 cycles' time. This was immediately followed by a high outward surge of approximately 1300 MW.

Simultaneously, the New Jersey intertie to PJM (through Public Service Electric and Gas) added a 1000-MW surge, lasting about 167 cycles' time, into the Con Edison system at Staten Island. This overload caused the tie circuit breakers at the Greenwood (Brooklyn) station to open, thereby leaving the Jersey intertie intact and the Arthur Kill station (Staten Island) still generating power to supply Staten Island, plus the Brooklyn load that had been carried by Greenwood.

For a brief period, there was the impression that the system was stabilizing back to normal, but the Con Edison generators—as well as those throughout area 4—were unable to respond quickly enough to the enormously increased demand upon them from the north. There was a drop in voltage and frequency and, in a matter of minutes, one system after another in southeastern New York went down as generators tripped out. Finally, at about 5:27 P.M., almost all power in four boroughs of New York City was lost.

The FPC version. According to the Federal Power Commission's report, at approximately 5:16 P.M. on November 9, Con Edison was operating with a system load in the range of 4800 MW. Figure 10 shows the net system load for November 9–10, while Fig. 11 represents a block diagram of the net interchange between Con Edison and other utilities just prior to and during the disturbance.

The capacity of Con Edison's 47 turbogenerators on the line at the time was 5900 MW. The arithmetical difference between this figure and the 4800-MW load is 1100 MW, the amount of spinning reserve. (Con Edison has stated that their spinning reserve was about 1350 MW. The company's spokesmen feel that apparently the FPC report took the total Con Edison load and subtracted it from the generating capacity, and that this does not take into account that some of the utility's load at the time was being supplied by ties from neighboring utilities.)

Con Edison was receiving power inflows, according to schedule, of 360 MW from Niagara Mohawk and 40 MW from the PJM pool. The utility was exporting power in the amounts of: 35 MW to CONVEX, 80 MW to the Long Island Lighting Company, 115 MW to Orange & Rockland Utilities, and 35 MW to the Central Hudson Gas & Electric Company.

In a matter of seconds, following the initial disturbance, Con Edison found itself in a situation in which

1. The inflows from both Niagara Mohawk and PJM ceased.
2. The intertie with PJM had broken in Brooklyn, but the 345- and 138-kV circuits to Niagara Mohawk remained closed.
3. It was obliged to export more than 560 MW of power to the isolated eastern segment of the Niagara Mohawk system (area 4).
4. Transmission to CONVEX reversed and, for a period of 2–3 minutes, an inflow of 180 MW was received from this New England group.
5. Long Island Lighting Co. dropped its inflow and

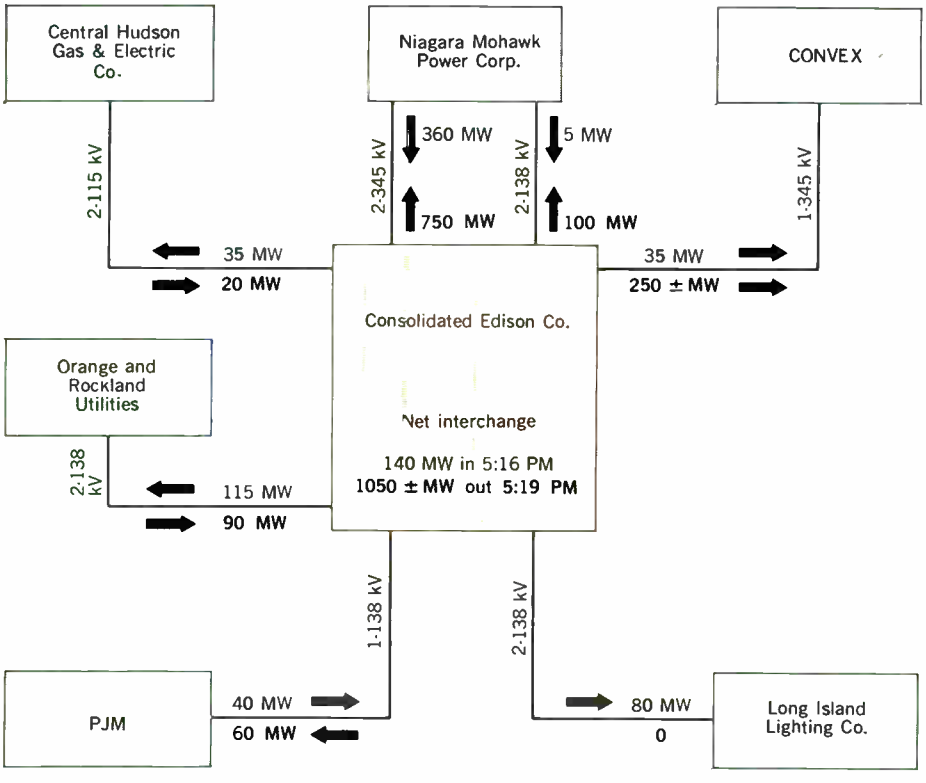


Fig. 11. Block diagram of approximate power transfers of Con Edison with other utilities before and during the disturbance of November 9. Power flows at 5:16 P.M. (before surge) are indicated in lightface type; power flows at 5:19 P.M. (peak of outflow) are shown in boldface type.

began to transmit 20 MW for the same time interval as CONVEX, after which the transfer, for one minute, dropped to zero, and then resumed as a transmission of 50–120 MW for the next three minutes.

Essentially, the total load on Con Edison's generating equipment increased rapidly, at 5:16 P.M., by 600 to 800 MW, maintained this range for about 2 minutes, and then increased further to substantially more than 1000 MW when, at 5:19 P.M., the 180-MW CONVEX inflow reversed to an outflow of 240 MW. An additional load was placed on this utility when, at 5:23 P.M., Long Island Lighting cut its tie in a vain effort to save itself. At 5:28 P.M., Long Island Lighting fell like the last domino in the line.

The FPC analysis. According to the FPC report, the spinning reserve capacity in area 4 was adequate—amounting to some 1650 MW—but this reserve was mainly in thermal plants in which there is always a time lag in the response of turbogenerators to large changes (more than 16 percent of rated capacity) in output. Further, the responsive capability of the spinning reserve was affected by the manner in which the total quantity of this reserve was distributed among the various generating units and plants at the time.

For example, the basic problem was complicated by the fact that a large percentage of the reserve capacity (about 350 MW) was contained in one unit, Con Edison's giant 1000-MW Ravenswood no. 3 cross-compound machine (Fig. 12) that was loaded, at the time of the disturbance, to about 650 MW. Although it was able to increase its generation by 100 MW just prior to system shutdown, apparently some 240 MW of its potential spinning reserve was unable to go on the line in time.

Since the system reserves in area 4 were inadequate in responsive capability to meet the sudden demand imposed, the frequency on these systems continued to fall. This created a cascade tripout effect within the thermal plants, since the drop in frequency also diminished the output of pumps and other electric auxiliary equipment required for steam-electric generation. Thus the thermal plants were shut down to prevent their destruction.

In review, there were apparently only two possible alternatives by which the increased demand on the Con Edison system could have been met:

1. By shedding some of its own load.
2. By increasing the generation output of its units.

A large repair bill. Considerable damage to journals, gland seals, turbine blades, etc. (Fig. 13), was done to three of the large Con Edison generators, including the huge Ravenswood no. 3 unit, when the electric pumps supplying the lubricating oil failed, oil pressure dropped, and the shaft bearings burned out. These three machines, at Astoria, East River, and Ravenswood (with a total generating capacity of about 1500 MW), down for an extended period, resulted in the loss of a considerable block of power not only to New York City but also to the entire interconnection.

Communications crisis. Of the 24 AM broadcasting stations in metropolitan New York, ten of those with studios in Manhattan have their transmitters in New Jersey, and thus receive power from Public Service Electric and Gas. With all television down, these ten radio stations maintained full-power service on normal frequencies to thousands of transistor sets and car receivers.

Another bright spot was the availability of telephone service, although it was subject to overloads, priority calls, and queues at pay stations. Automatic switching to trickle-charged batteries, thence to fallback generators, paid off for the phone companies. But newspaper and other teletype machines were casualties, except for police and fire department installations with emergency power facilities.

By the light of the moon, glaring deficiencies

It was perhaps providential that the regional power failure occurred during relatively mild weather, with a clear sky, and under the light of a full moon—almost the sole source of auxiliary illumination for 30 million people—rather than during blizzard conditions and freezing temperatures.

The helplessness of people trapped in underground transit systems and in buildings, plus the intolerable situation of muted air raid sirens, nonfunctioning traffic signals, and unlit stairways and emergency exits, suddenly revealed a massive blind spot in our thinking and planning. Except for some hospitals and isolated instances where design foresight paid off for commercial or industrial buildings, most of New York City was caught entirely without any auxiliary power sources.

The issues are drawn

As a result of the incident of November 9–10, public confidence in the power industry was shaken. Engineers and laymen are asking some rather pointed questions.

A list of some of the pertinent queries might be enumerated as—

Question one. *Are there deficiencies in the overall planning of fully coordinated power pools which could trigger another incident that will cascade, in a matter of minutes, into a massive system outage?*

Question two. *In our sophisticated power technology, which is producing supersized generators and increasingly complex interconnections, have we in some instances neglected to provide adequate controls and backup equipment that can assist in an emergency situation?*

Question three. *As one possible approach, have our utility companies, participating in heavy power pools, planned adequately for scheduled load shedding by automatic and manual controls in emergency situations in which underfrequency develops?*

Question four. *Can a maior city any longer afford to have complete dependence upon slow-reacting thermal plants, without available reserve generating facilities that will provide faster emergency response?*

Question five. *Have we in the engineering profession become so self-satisfied by our overall technological and scientific achievements that we have lost sight of the potential dangers lurking in emergency or unexpected situations?*

Restoration of service—a mammoth problem

As of 5:28 P.M. on November 9, for all intents and purposes, a major portion of the northeastern United States was without electric power. But at the same moment, the utilities initiated the Herculean task of restoring service. Some underground grid circuits in major cities, which had tripped out in a fraction of a second, would require hours to re-energize. Thermal generating stations throughout the affected area were shut down. In most cases, even the auxiliary equipment—such as coal

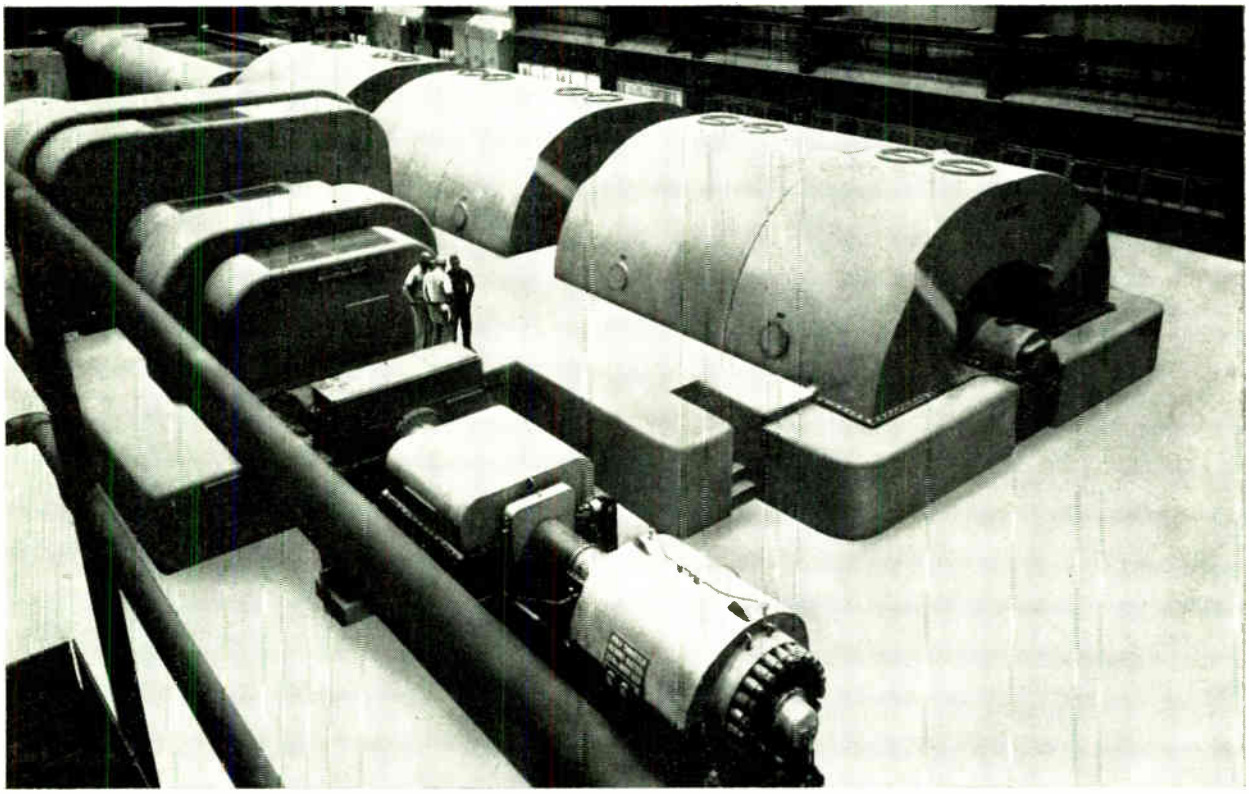


Fig. 12. Overall view of Con Edison's 1000-MW Ravenswood no. 3 generating unit. This cross-compound machine is the largest in the world. It began supplying the system on June 7, 1965.

Fig. 13. Damage to the 1000-MW Ravenswood no. 3 unit occurred when the oil pressure failed and the machine was taken off the line. The bearings were damaged and the journal, as shown, was scored.

conveyors and oil pumps, needed to fire the boilers for the restoration of system generation—was dependent upon the power supply that had failed. Thus, before some circuits could be energized, power from outside sources was a prerequisite.

Restoration speed depends upon the characteristics of the system's generation sources (thermal or hydro), transmission facilities, the availability of auxiliary or external power sources, the nature of the disturbance causing the outage, the availability of qualified personnel for unit and station start-up, and the system's load requirements. Generally, overhead transmission lines can be restored to service more readily than the cables and complex circuit arrangements of underground networks.

Part of the general problem. Whenever a transmission system has lost its power, the first step beyond the isolation of the affected system is the examination of relays, circuit breakers, and switches to determine—if possible—the cause of the disturbance. It must be borne in mind that during the night of November 9–10, *the cause of the massive power failure was not known.* Therefore, each utility company realized the possibility that its system may have been responsible for triggering the blackout, or had sustained damage during the incident. In view of this, extra precautions had to be taken—with a corresponding loss of time—to ensure that relays, circuit breakers, and switches were in proper operating condition. Furthermore, as power once more became available, the load had to be picked up carefully by a sequential procedure. As each portion of a system was brought up to normal load, frequency had to be synchronized with that of any energized part of the system. Only then was it possible to restore the interconnections.

Difficulties in station start-up. When a steam-electric station loses its load, the automatic control equipment may shut down the plant. Thereafter, the boilers, turbines, generators, and all auxiliary machinery must be inspected. All necessary repairs must be made and the

automatic controls reset before the start-up sequences can begin.

Hydroelectric stations that have been shut down temporarily, however, have a distinct advantage over thermal stations in that they can usually be started without auxiliary power other than emergency battery equipment for lights and control circuits. But if the outage time is for long duration, auxiliary power will be needed for governor- and lubricating-oil pumps and for compressors. By contrast, since about 2 to 5 percent of each turbogenerating unit's output is channeled into running its vital auxiliary equipment, steam-electric station start-up procedures are complex.

Unfortunately, many of the affected utilities had made no provision for the unlikely possibility that their entire system would shut down simultaneously and, hence, there were no independent auxiliary power sources for such an eventuality. Intricate circuits had to be established, some from remote sources, to feed in the essential auxiliary power.

Service restoration in upstate New York. The key utility in service restoration in upstate New York was the Rochester Gas and Electric Corporation (Rochester), situated at the strategic western boundary of the disturbance area. Although Rochester lost its load at 5:19 P.M., after separating from Niagara Mohawk, its tie lines with PASNY (according to the FPC report) remained energized. This fortuitous circumstance enabled Rochester to draw upon an auxiliary power source immediately at hand. Also, the utility had its own hydroelectric generation sources as a spinning reserve, and the hydro was quickly put on the line to supplement the auxiliary power requirements of starting up its steam-electric units. When these thermal units were brought up to load, Rochester's system requirements were stabilized. By 10:30 P.M., 80 percent of its entire system was back in service.

A delay was encountered in attempting to close the tie line from Rochester to the Clay (N.Y.) substation on the PASNY system. This closure was a vital link in restoring EHV transmission from west to east along the 345-kV lines. Apparently, the circuit breakers on the Rochester–Clay line would not close because repeated early efforts (when the disturbance was first assumed to have originated at this substation) to reactivate the breakers exhausted the reserve air pressure in the pneumatically actuated mechanism. This condition, like a vicious circle, could not be quickly corrected because power was unavailable to operate the air compressors.

Despite the delay in restoring the Rochester–Clay 345-kV circuits, immediate restoration of Niagara Mohawk's 115-kV system was initiated, and by 9:30 P.M. only 1000 of its 1 124 000 customers were still without power.

New York State Electric & Gas Corporation (NYSE&G), a system with a predominantly internal overhead transmission network and access to external power sources through interconnections with other utilities, was able to restore service with reasonable speed and efficiency.

The company managed to maintain service, following the disturbance, to 71 percent of its customers because its 230-kV connection to PASNY-Niagara held with almost no interruption. During the first minute of the disturbance, machines operating at the Greenridge, Hickling, Goudey, and Milliken thermal generating stations were severely overloaded and tripped off the line. One unit at the Jennison station continued to operate isolated from the system and one unit at Milliken was manually shut down to protect it from damage.

Swings in frequency and voltage caused difficulties and some delays in the restarting of generators. The first unit was restored at 5:35 P.M. and the last unit at 11:02 P.M. Within two hours from the start of the disturbance all but 6 percent of the NYSE&G customers had service, and by 9:52 P.M. all customer service was restored.

Service restoration in New England

The CONVEX group offered a good example of service restoration procedures in an integrated system that utilizes both hydroelectric and gas turbine generation for peaking loads. The entire generation and transmission facilities of its participating companies (see list in Fig. 3) are controlled from dispatch centers at North Bloomfield and Southington, Conn., without regard to the distribution of property ownership by the group members.

Between 5:19 P.M. and 5:30 P.M., steam-electric generators, with a total net capacity of 1588 MW out of the pool's combined capacity of 2685 MW, were shut down because of an inadequate power supply to auxiliary machinery. But the remaining plants, including Hartford Electric Light's 10-MW gas turbine generating units, stayed on the line to carry local loads.

Immediately after the CONVEX pool was separated from CANUSE, restoration procedures commenced. In the next 2½ hours, the energized transmission system was used to reactivate the essential auxiliary equipment in the thermal stations and for reclosure of the interconnections within CONVEX. By 10:30 P.M., the system was apparently restored to full service, but at 10:35 an insulator failure caused another outage in the tie lines between Connecticut and Massachusetts. Despite this setback, the CONVEX system was intact by 11:00 P.M., and normal service was achieved by midnight.

The NEPCO experience. The New England Power Company (NEPCO), serving Worcester and other towns in east central Massachusetts, lost its 50-MW unit in the Webster Street steam-electric station in Worcester at 5:18 P.M., but with the assistance of the Harriman hydroelectric station at Whitingham, Vt., it was able to obtain auxiliary power by 6:00 P.M. By 6:45 P.M., the underground network in Worcester was energized, and at 7:33 P.M. the Webster Street generator resumed full load.

NEPCO's Worcester experience, however, was not typical. The company's Brayton Point Plant, with a net main generating capacity of 482 MW, shut down at 5:17 P.M. Although station service restoration was completed at 6:25 P.M., and the start-up procedures for the 241-MW no. 2 unit were begun at 6:35 P.M., it was not until 12:25 A.M., November 10, that this machine could be put on the line. It was operating only at about 50 percent capacity by 2:00 A.M.

In Boston, wheels were rolling. The Boston Edison Company was carrying a system load of about 1375 MW in the Boston metropolitan area when its service went down at 5:21 P.M. All of the metropolitan area was affected, except the Metropolitan Transit Authority, which has its own independent power plants. Thus subway and elevated transportation continued to roll during the outage.

By 6:30 P.M., the necessary station power was obtained

from the NEPCO intertie at Edgar and at Boston's L Street station. Simultaneously, Boston's Mystic generating station at Everett received auxiliary power from the U.S. Naval Shipyard in Charleston. From 7:40 to 9:20 P.M., generation loads were picked up by these stations and, in sequence, additional units went on the line until approximately 2:30 A.M.

The underground networks. Electric power service in the downtown Boston area is provided by a complex, cable-fed, underground network system similar to that used by Con Edison in New York City. Boston, however, is much smaller than New York City, and its cable mileage and loads are considerably less.

The advantages of an underground transmission system normally are

1. Insulation from natural disturbances.
2. Protection from feeder failures.

Under normal circumstances, therefore, the high-voltage grid is more reliable for vital services than are other systems.

Under most normal and abnormal conditions, the low-voltage underground distribution networks offer the highest degree of reliability; however, when de-energized, there is the inability to provide a priority of service restoration to vital municipal services such as hospitals, traffic control signals, street lighting, and rail transportation (except where transit power is obtained from independent sources).

Further, the rapidity with which auxiliary power from external sources can be brought into the system is hampered by the *capacitor effect*, which occurs as the underground transmission cables are energized. This effect produces a voltage rise on cables that have been energized without sufficient load to place on them. It can reach hazardous proportions if improper loading increments are applied.

New York City's revival—a formidable task. In New York City, Con Edison's major operating areas are served by 305 miles of 138-kV and 59 miles of 345-kV transmission cables. The utility also provides a 25-c/s network for urban transit systems, plus 60-c/s service to some railroads and subways which, at customer's option, may be converted to 25-c/s service or rectified to dc.

Con Edison's system is divided into 42 networks, which are electrically isolated and can be independently energized (but with some attendant problems). Except for suburban sections, each city network serves a geographical area whose loads range from about 25 to 300 MW.

The network area is served from one substation which, in turn, is served by the transmission network from two or more directions. And two or more areas may be served by a substation.

Two of the reasons why it required up to 13½ hours to regain complete service in New York City were that

1. The huge physical area of the city and the load apportioned to each of the centrally situated substations is as large as the entire electrical load of many big cities.
2. If the central substation is forced out, there is no alternative source from which to supply the network area. This is generally true in the design of network systems.

Chronology of Con Edison's reclosure. Con Edison's service restoration procedures were achieved in accordance with a plan which was originally established in 1938; but the scheme has been kept up to date and it was available to the energy control center and at every start-up point. The two logical points to initiate service restoration were: at the Staten Island end of the system, where power from the Arthur Kill station was maintained throughout the disturbance; and at the Pleasant Valley substation, near Poughkeepsie, N.Y., where power became available through restoration of the ties to the Central Hudson Gas and Electric Corporation. The primary objective of this scheme was to obtain adequate auxiliary power in minimum time.

At the southern end of the system (Brooklyn and Staten Island), switches, relays, and circuit breakers were reset as soon as maintenance and repair crews could reach the substations. By 6:50 P.M., auxiliary station power was available at the Hudson Avenue (Brooklyn) generating station. The 80-MW no. 4 unit at this plant was restored to service at 9:17 P.M., and the other seven units, in sequence, were thereafter placed on the line. Some time between 9:17 and 10:36, some radial load was picked up from this station, and service to the 25-c/s subway system was restored at 11:30 P.M., when Hudson Avenue's frequency converter was put on the line.

By 9:15 P.M., station service auxiliary power was regained by the Astoria plant, and station power was restored to Ravenswood at 11:20. The Sherman Creek and Hell Gate generating stations, in northern Manhattan and the Bronx, regained power at 1:08 A.M. and 1:57 A.M. respectively.

With auxiliary equipment power available, generation began at the Waterside station by 10:36 P.M., and by 11:34 P.M., at the 59th Street station. From shortly after midnight to about 1:50 A.M., station power was available to the remainder of the system's generating stations.

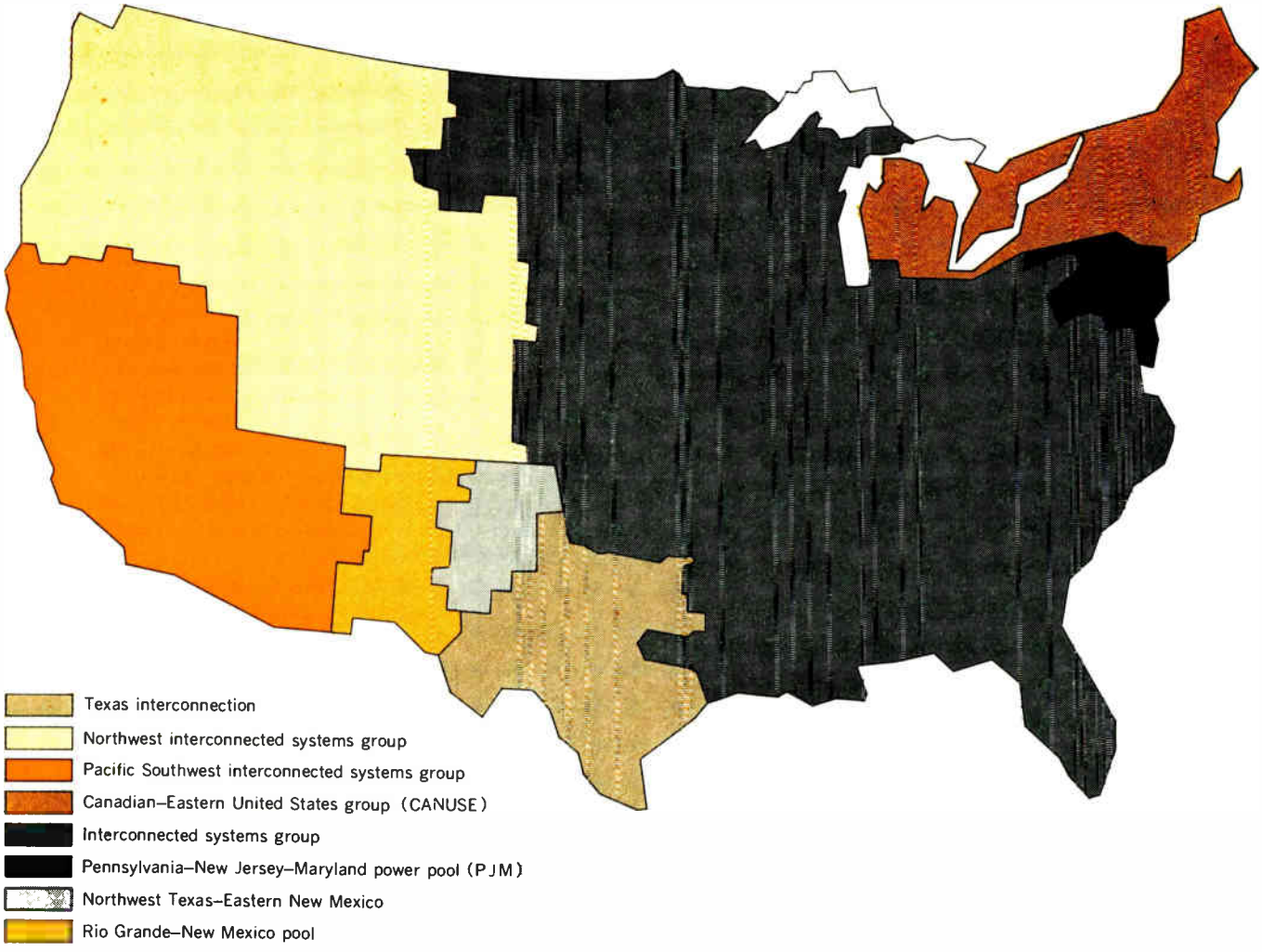
As soon as power was available from the north, Con Edison's crews reactivated the Pleasant Valley substation and, proceeding southward, re-energized the line between Pleasant Valley and Millwood by 8:25 P.M. From there, power was fed to the Dunwoodie (Yonkers) switching station, from which point it was made available to the Sherman Creek and East 179th Street stations. Auxiliary power at the Hell Gate plant was obtained from the latter station.

While this work was under way, portions of the 345- and 138-kV lines were being energized for customer service. By 10:36 P.M., customer services were being restored to the Borough Hall area of Brooklyn, and, to the north, areas of central Westchester County regained service by 1:30 A.M. Thereafter, the networks were energized as rapidly as power became available, with the last network—the West Bronx—back in service at 6:58 A.M. on November 10.

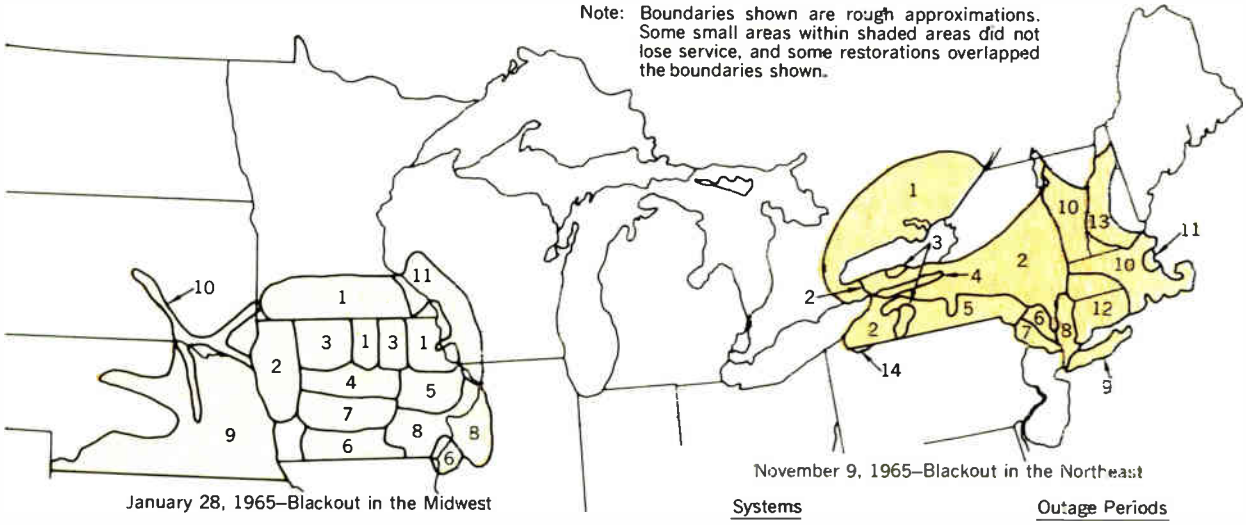
The FPC report, in analyzing Con Edison's difficulties, indicates that

1. Dependence upon steam-electric generation was a restrictive factor in the speed of service restoration.
2. Lack of emergency power sources for the operation of generating station auxiliary equipment contributed to the length of the outage period.

Canadian system restoration. While the FPC report does not contain a detailed account of the restoration of service on the Ontario Hydro system, it is known that the maximum time of outage in the province was about 3 hours. The Detroit Edison interties at Windsor and Sarnia proved to be of considerable assistance in reducing the outage time in the Ontario peninsula. Also, since the Ontario Hydro system comprises about 60 percent hy-



Note: Boundaries shown are rough approximations. Some small areas within shaded areas did not lose service, and some restorations overlapped the boundaries shown.



January 28, 1965–Blackout in the Midwest

	Systems	Outage Periods
1.	Interstate Power Co.	14 min. to 2 hrs. 14 min
2.	Northwest Iowa Power Coop.	32 min.
3.	Corn Belt Power Coop.	4 min. to 2 hrs.
4.	Iowa Public Service Co.	50 min. to 2 hrs. 30 min.
5.	Eastern Iowa Light & Power Co.	43 min. to 1 hr. 45 min.
6.	Iowa Southern Utilities	5 min. to 1 hr. 58 min.
7.	Iowa Electric Light & Power	52 min. to 2 hrs.
8.	Iowa–Illinois Gas & Electric	30 min. to 93 min.
9.	Nebraska Public Power System	7 min. to 2 hrs. 30 min.
10.	USBR	3 min. to 1 hr. 26 min.
11.	Dairyland Power Cooperative	0 to 1 hr. max.

November 9, 1965–Blackout in the Northeast

	Systems	Outage Periods
1.	Hydro Elec. Pwr. Comm. Ont.	1 hr. 31 min. to 3 hrs. 14 min.
2.	Niagara Mohawk Power Corp.	Momentary to 5 hrs. 14 min.
3.	Rochester Gas & Elec. Co.	2 hrs. 1 min. to 6 hrs. 28 min.
4.	Power Authority State of N.Y.	Momentary to 54 min.
5.	New York State Elec. & Gas	1 hr. 14 min. to 6 hrs. 4 min.
6.	Central Hudson Gas & Elec.	2 hrs. 8 min. to 4 hrs. 38 min.
7.	Orange & Rockland Utilities	1 min. to 3 hrs. 52 min.
8.	Consolidated Edison Co.	8 hrs. 33 min. to 13 hrs. 32 min.
9.	Long Island Lighting Co.	Momentary to 7 hrs. 30 min.
10.	New England Power	Momentary to 4 hrs. 22 min.
11.	Boston Edison Co.	2 hrs. 53 min. to 7 hrs. 39 min.
12.	CONVEX	13 min. to 5 hrs. 58 min.
13.	Public Service Co. of N.H.	Momentary
14.	Pennsylvania Elec. Co.	Momentary to 15 min.

droelectric generation, the essential external auxiliary power for its thermal plants was available.

The question of nuclear plants

There has been speculation as to the performance of the two atomic generating plants—Con Edison's station at Indian Point, N.Y., and the Yankee Nuclear Power Station at Rowe, Mass.—in the affected area. The questions, however, become academic, since on November 9 neither of these plants was in service. Therefore, no experience could be gained as to the effects of a loss-of-load shutdown coupled with the loss of auxiliary power. Further, there was no opportunity to evaluate the contributions of a nuclear plant toward decreasing the outage time.

Pumped storage vs. gas turbines

Some electrical engineers and utility spokesmen contend that if the proposed Cornwall pumped-storage plant had been built and in operation, either New York City's service could have been preserved or essential electric power requirements could have been restored within minutes following the regional blackout. Since the plant is nonexistent, the entire question is conjectural. Other qualified power specialists feel that gas turbines are preferable.

Public Service Electric and Gas recently dedicated the world's largest gas turbine generating unit (Sewaren Generating Station) at Woodbridge, N.J. The new station—the second gas turbine installation for PSE&G—has a rated capacity of 121 MW and is primarily designed to provide power for peak demand periods in the company's service area. The secondary function of the unit is to furnish the system with a large block of capacity, which can be started completely independently of any outside power source and brought up to full load within a few minutes' time.

A number of power engineers believe that, by placing gas turbine generating units at strategic urban locations, with independent circuits, power could be restored—at least for vital services—almost immediately following a widespread system power failure. The validity of the latter function was demonstrated in Hartford, Conn., where local service was retained during the blackout by the assistance of gas-turbine units.

It happened before—it could happen again

While the great blackout of November 9–10 was not the first interconnected systems outage (the major areas in the United States served by interconnected systems are shown in Fig. 14) in history—and probably not the last—it affected far more people than any previous intersystem failure. Also, some of the previous outages were caused by floods, tornadoes, and other natural disturbances.

Fig. 14 (top). Map showing major areas of the United States (and a portion of Ontario) served by interconnected systems.

Fig. 15. Map of areas affected by cascading electric power failures in the Midwest and Northeast.

Outage in the Pacific Northwest. More than 15 years ago, a widespread outage in the Bonneville Power Administration's (BPA) system caused considerable damage to electric motors and other equipment because the frequency was permitted to drop to about 40 c/s in some areas before tripout. The chronology of that incident follows.

At 4:55 P.M. on June 6, 1950, a phase insulator string at a tower on the Coulee–Columbia 230-kV, no. 3 line failed a few seconds after this line had been energized for the first time. Then, a circuit breaker at Coulee Dam failed to trip by relay because the switch controlling the dc circuit between the relays and the circuit breaker trip coil was inadvertently left open. The circuit breaker was tripped manually from the switchboard and the fault was cleared at the end of 190 cycles.

The prolonged fault, however, caused a general system instability in which five more 230-kV lines, and three 115-kV lines from Coulee, opened. At the time of the disturbance, the transmission lines from Coulee were carrying a total of 1328 MW.

At the Bonneville Plant, generators nos. 1 and 2 dropped out of step and were cleared by overspeed relays at 4:55 and 4:56 P.M. respectively. The no. 9 unit went out of step and was cleared by hand at 5:02 P.M.

In rapid sequence, power plants from British Columbia and Washington to Oregon, Idaho, Utah, and Montana were separated from the interconnection, and the power pool was broken into three segments.

The opening of the various transmission lines that were delivering about 900 MW from Coulee at the time of the fault, together with the loss of three Bonneville generating units, the Merwin Plant of the Portland Power & Light Company, and the Bridge River Plant of the British Columbia Electric Company, accounted for a deficiency of approximately 1260 MW of power into the Seattle–Portland–Midway areas.

The frequency in the Portland and Seattle areas dropped to about 40 c/s, as indicated by the tachometer readings on the Bonneville generators. At one town, the voltage momentarily dropped to 75 volts; within two minutes, it recovered to approximately 85 volts, then hovered around 90 volts for the next 12 minutes before returning to normal.

Since it was impossible to reduce the frequency of the Coulee machines below 50 c/s, immediate resynchronization with the system in the Portland–Seattle area was also impossible. Therefore, it was necessary to increase the frequency in these areas by shedding load and by further segregation of the system.

Considerable industrial losses were caused to paper mills and aluminum plants that were affected by this unusual outage, which lasted up to 1¼ hours.

More recent disturbances. New York City experienced blackouts in one network area on August 17, 1959, and in four network areas on June 13, 1961, caused by electric equipment failures within the bulk supply elements of the power system.

On January 28, 1965, a loose connection on a 230-kV differential relay at the Corps of Engineers' Fort Randall hydroelectric station in South Dakota isolated the generating station's bus, dropped 240 MW of generation, and triggered an outage that spread quickly through the interconnected systems to affect most of Iowa and portions of five other midwestern states (see Fig. 15 map).

Power was flowing from the northern hydro plants in the Dakotas to the south. At that time—the system was still evolving—the alternate routes could not handle the 230-kV transmission. The Fort Randall relaying scheme, however, has since been modified and three additional 230-kV lines now provide sufficient transmission capacity for emergency rerouting.

During the Midwest outage, only 2 million people were affected. Total service was restored within 2½ hours.

On April 11, 1965, tornado damage in Indiana affected the Upper Mississippi pool. From Indiana, the blackout extended to St. Louis and then into Iowa. Both Iowa Electric Light and Power and Interstate Power were affected in the eastern part of that state. The power failure also hit a portion of the Iowa Public Service system out of Des Moines.

In a somewhat parallel situation to the Northeast blackout, Iowa Power, at that time, was importing power from the interconnection and, when the outage became a chain reaction, there was a reversal of flow. Too great a load was placed on this local utility and the generators relayed out. Restoration of service required about one hour on this system.

A different situation affected the Northern States Power Company last spring during a time of record-breaking floods and ice jams along the upper Mississippi and Minnesota Rivers. One of the system's largest generators was down for maintenance and repair during a light-load period but, simultaneously, another large machine was lost because of equipment failure. The flood waters and river ice disrupted the Northern States' major transmission system in several places. Nevertheless, the utility was able to continue full service to its customers by importing power over its interconnections.

On December 2, 1965, about one million persons in sections of Texas, New Mexico, and Mexico were affected by a power failure in the El Paso area. At this writing, the causative facts are not complete, but a spokesman for the El Paso Electric Company, which serves much of the area hit by the power failure, indicated that the trouble had been traced to the company's Newman Plant, situated just south of the Texas–New Mexico border. The disruption of service hit an area from Van Horn, Tex.—120 miles east of El Paso—to Socorro, N. Mex., about 175 miles to the north. Juarez, Mexico, was also blacked out. The service disruption occurred at 8:02 P.M., and up to 2 hours were required in some sections before power was restored.

A preliminary investigation indicated that the trouble may have been touched off by the failure of a gas regulator to control the amount of fuel injected for firing the boilers at the Newman Plant.

Philosophy of the firm interconnection

As a prelude to the analyses and recommendations for future safeguards to prevent a recurrence of the unprecedented power failure in the Northeast, we must recognize that there seem to be two areas of thought among power people: one faction is in favor of the *firm interconnection*, or network; while the other, more conservative, group advocates a loose linkage that is not deeply involved in the complex interties of a heavy grid. The latter group believes that the loose linkage can be isolated more expeditiously from an interpool in an emergency situation, and service restoration may be achieved in less time.

Advocates of both philosophies, however, in accordance with the North American Power Systems Interconnection Committee (NAPSIC) instructions, subscribe to the analogy that "when a man is drowning, his associates must jump in to save him"—up to the point where "the drowning man threatens to pull his rescuers under." At this point, each system must be isolated for individual survival.

The FPC report tends to agree with the firm interconnection philosophy and stresses that measures should be taken toward strengthening the grids. It points out that weak linkages may protect the adjacent pools by quick separation from interconnected groups, as in the case of Maine and PJM, but they also lessen the ability of the adjoining systems to support each other in the event of a massive disturbance. The report cites the possibility that the admixture of strong and weak interties within the CANUSE area, and with the PJM pool, may have contributed to the inability of the affected utilities to ride out the November 9 disturbance.

Failure of equipment vs. failure of service

At this point it may be well to emphasize the difference between *equipment failure* and *service failure*. The risk of equipment failure is inherent in any mechanical or electrical device or practice. Increasing the size of generating units, use of EHV transmission, and every aspect of progress in power technology carries some degree of risk.

In a fully coordinated and well-designed system, equipment failure is not usually a major problem, provided there are sufficient safeguards and backup equipment to meet the contingency. In the previous outages just described, equipment failure was caused either by natural disturbances or human error. We may say, therefore, that if an equipment-caused outage is controlled and confined, or isolated, within or near the point of the original disturbance, it may be termed an equipment failure outage. But if the trouble cascades—as did the Northwest power failure of 1950 and the event of November 9–10—into a regional power failure, then it becomes a service outage.

Strong support for the network

The majority of the power experts with whom the writer consulted were solidly behind regional integration and power pooling. It was their feeling that, in the balance, interconnections have done more to improve service than to cause trouble (Fig. 16).

These consultants stressed, however, that adequate protective and backup equipment is a vital consideration at the outset of interconnection planning. For example, the writer was informed by the Northern States Power Company that sufficient safeguards and bus protection devices have been designed and incorporated into its system to prevent cascade tripouts. Further, a comprehensive supervisory control scheme has been established to permit the selective segregation of areas affected by outages. And frequency-sensitive relays for load shedding will automatically open switches before sudden, grinding shutdowns can seriously damage large generators.

Network advantages. The primary advantages of the firm regional interconnection are system economy and reliability, achieved by:

1. Taking advantage of time zone load diversities.

2. Applying this load diversity to different climate zones (in an interpool), as related to air conditioning or heating requirements.

3. The efficient and economic use of very large generators.

4. The reduction in installed reserves that are necessary for individual systems.

5. The relocation of generating facilities from urban to remote energy (mine mouth) areas.

6. Emergency assistance to neighboring utilities in the event of a disturbance.

. . . And disadvantages. Interconnected systems are huge and complex entities designed to maintain intersystem integrity on the basis of the most severe single disturbance contingency, but they have not usually been planned to maintain intersystem stability in the event of a combination of simultaneous multiple disturbances. This situation, however, may also be true in the operation of an independent system.

Further, each utility company in the pool has a different output capacity, dynamic inertia, speed-regulating methods, and loading systems. These parameters form part of a continuing problem that must be evaluated before and during the interconnected operation.

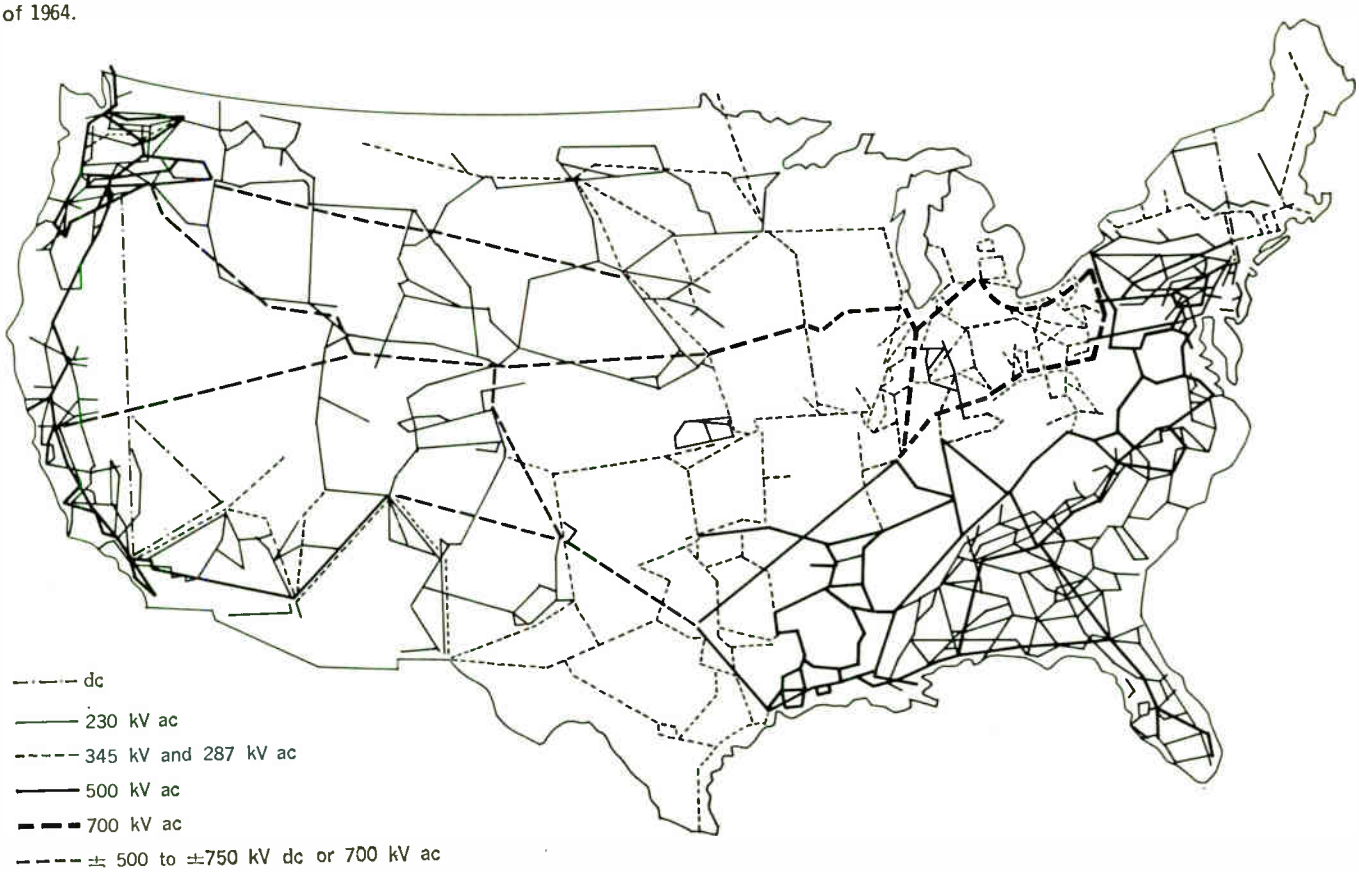
As we have seen, a severe disturbance can cause a drop in voltage and frequency which, if severe enough, can further cause one section of the interconnection to run out of synchronism with the other. This will result in a breakup of the interconnection into segments.

Interconnection power swings. The disturbance causes a power swing, with all of the utilities on the interconnection attempting to make an adjustment for the power deficiency or power surplus. Because of the oscillation resulting from the affected system's rotating inertia, the unaffected companies will actually either overshoot or undershoot the required demand on the system. This is the most critical period, since each utility's system strength is taxed as to whether it can bear the power swing.

As a hypothetical example, assume that System A, of ten machines,[1] is intertied to System B, of infinite generating capacity. If there is a loss in generation in System A, 100 percent of the lost generation will be supplied through the tie line by System B, since there can be no permanent change in frequency. Also, because of the oscillation that results from the interaction of the system dynamic inertia, there may be an overshoot. The amount of this overshoot, depending upon damping, governor action, and system reactance, may be as small as 50 percent. If the stability limit, or tie-line capacity, is within a conservative figure—say 10 percent of the system capacity—the period of the power swing should be little more than 3 seconds' duration, and the maximum transient frequency deviation will be about 0.1 c/s for a 5 percent generation loss. Theoretically, if the tie line can support the generation for a few seconds, system stability is assured. But if the tie-line strength is disproportionally smaller than the capacity of the largest generator on the system, there will be the imminent danger of a tripout.

Transient stability studies. The simulation of power

Fig. 16. Map of possible pattern of generation and transmission (230 kV and higher) by 1980. This proposed national grid system is based on the Federal Power Commission's National Power Survey Report of 1964.



system performance following faults and disturbances, loss of load, or loss of generation is a very complex problem, but there are some adequate mathematical models and machine programs that will indicate the performance of the system until a new steady state evolves.

The system will show its strength—or weakness—either during the first several seconds following the disturbance or after a new steady-state condition is attained. Thus only these two time periods are usually studied. The time-sequence periods from zero (instant of disturbance) to when a new plateau of steady-state equilibrium is established may be classed as:

1. *Zero to one second*—transient performance before voltage regulators and governors at generating stations can respond.
2. *One second to 5 minutes*—performance with regulators, governors, and changing boiler conditions; but before system load control equipment can respond.
3. *Five minutes and longer*—performance under the monitoring of the system load control equipment.

Adequate system planning

To achieve a fully coordinated interconnection, adequate system planning perhaps should include

1. Relays whose accurate settings, in themselves, will act as "memory" elements—similar to those of a centralized computer—and protective safeguards in making quick responses to check the spread of system disturbances.
2. The greater use of acutely attuned sensing devices, such as relays that will trigger supersensitive oscillographs, or the use of continuous recording equipment, at the instantaneous onset of remote disturbances.
3. Periodic re-evaluation of the dynamic performance of complex power networks to ensure that no other participant in the interconnection has altered his operational parameters to a degree that might cause instability following a disturbance.
4. The necessity for system operators, in keeping the grid updated, to be critically aware of interarea oscillations, maximum permissible power swings, automatic generator voltage regulators, excitation system characteristics, system damping, and electrical load characteristics.
5. Revision of design and operating parameters to deal successively with a combination of concurrent, or rapidly sequential, catastrophes.

Auxiliary system redundancy

Because of their function and damage-control requirements, every warship in the U.S. Navy is equipped with at least one backup system to its primary power-generating machinery and pumps. Thus, if the turbogenerators, which are driven by auxiliary steam from the main propulsion turbines, fail, the diesel generators immediately pick up and maintain the vital power loads. The electrically driven pumps are backed up by steam- or hydraulic-driven pumps and, if all else breaks down, the pumps can be manually operated.

Naturally, one would not expect utility companies, normally geared to power production for a civilian economy, to incorporate the elaborate auxiliary systems that are analogous to warships. But in view of what has happened, there is a growing consensus of expert opinion, reflected in the FPC report, that sufficient emergency auxiliary station power sources or standby generating equipment should be provided to permit at least: the operation of pumps, compressors, and other vital auxiliary machinery during the start-up following a system-wide tripout.

The FPC recommendations

A number of the nineteen specific recommendations of the FPC report, including the need for auxiliary power sources, more extensive stability studies, fully coordinated power pools, mixed generation (hydro, pumped storage, gas turbine, and others), internal load shedding, adequate communications facilities, etc., have already been discussed and reviewed in this article.

The balance of the FPC recommendations include—

Closer international cooperation. The Commission calls for an even closer working relationship between Canadian and United States operating organizations, and the governmental authorities of both nations. In this context, the National Energy Board of Canada has been fully apprised of the various phases of the continuing investigation and has extended its utmost cooperation in working with the FPC.

Strengthening the networks. Isolated systems are not well adapted to modern needs, either for purposes of economy or service. The power systems in the CANUSE area are presently in a transition period from isolated operation or weak ties to strong interconnections and full coordination. The stability of the system may also be strengthened by the strategic location of generating capacity, planned on a pool-wide basis.

Also, there are numerous additional EHV transmission facilities, which the systems in the Northeast have already agreed to construct or which are under consideration, that will

1. Strengthen the internal ties among generating stations and load centers within individual systems.
2. Strengthen the links between adjoining systems.

The FPC recommends an acceleration of the present program toward stronger transmission networks within each system and stronger intersystem connections to achieve more reliable service coupled with the greatest possible economy.

Importance of frequent relay-setting review. The massive power failure indicated the importance of careful and frequent checks of relay settings that control major facilities. The utilities should institute such an inspection immediately and establish procedures for frequent periodic reviews that will take into consideration changing conditions of power output, overloads, etc.

Economy vs. service reliability. When there is a conflict between economic and service reliability considerations in power system design, reliability and security of service should be given priority.

Generating reserves. The preliminary investigation indicates that the type and distribution of available generating reserves, in an emergency situation, may be as important as the quantity involved. The utilities in the CANUSE area should make a more realistic evaluation of the time factor involved in the utilization of spinning reserves to determine the responsiveness of the components of the total spinning reserves to emergency demands.

Automation study. The Commission recommends a two-level study—industry-wide and by individual utilities

—of the adequacy of existing automated equipment, communication facilities, recording facilities, and operational procedures in the dispatch and control centers and in generating stations during emergency conditions.

Plant crew training. Although the FPC is not in a position to pass judgment on the need for improvement in emergency start-up training for generating station personnel, it recommends a thorough review of training procedures for emergency situations.

Maintenance of vital services. Those civilian services such as hospitals, airports, tunnels, railroad and subway stations, traffic control signals, and basic communications for which anything less than 100 percent power availability is inadequate should arrange for an auxiliary power supply.

Public protection. In most cases the cost of a complete auxiliary power source may be economically prohibitive, but in many instances it is both reasonable and necessary to provide a degree of public protection when power service is interrupted. Therefore, with regard to the subways in New York City, where an alternative power supply for train operation may be impracticable, a minimal subway evacuation plan—including auxiliary lighting equipment for stations and tunnels—should be formulated to preclude the repetition of an intolerable situation in which upwards of 600 000 people were trapped for an extended period.

Building elevators. The FPC considers building elevators to be a special problem. In some cases, it may be feasible to move at least one elevator at a time, by auxiliary power, to an intermediate floor or ground floor level for passenger evacuation. Elevators should be provided, however, with minimal emergency devices such as mechanical cranks or levers that can be manually operated in the event of stalling between floors during a power outage.

Public communications—radio and television. Communication facilities, powered by auxiliary sources, should be developed so that the public may be informed promptly as to the circumstances of a power failure. (The psychological reassurance value of immediate news information to the public is a significant factor in the prevention of panic and other situations involving public safety.)

Motor vehicle fuel facilities. One of the annoying consequences of the power failure was that motorists were unable to obtain gasoline because the service station pumps were dependent upon the system power supply. The FPC recommends that the petroleum industry devise a method to eliminate this problem in the interest of public safety and the possibility of a traffic breakdown in the event of an extended power failure.

Need for additional legislation? The Federal Power Act of 1935 made no specific provision for governmental jurisdiction over the reliability of service for bulk power supply from interstate grids. Presumably, at the time this was regarded as a problem within the jurisdiction of the individual states. But the enormous increase and development of interstate power networks during the past 30 years requires a re-evaluation of governmental responsibility for reliable continuity of service. Since a single state cannot regulate the service from an interstate—or international—power pool, the question of the need for additional Federal legislation is under active study.

As this article indicates, the electrical engineering community has a major responsibility to resolve questions raised by the Northeast power failure. To this end, IEEE SPECTRUM hereby invites discussion by IEEE members in all relevant branches of electrical engineering. Please limit responses to 1000 words or less. As many as possible will be published in a forthcoming issue of SPECTRUM.

Who has the ball?

As of now, the private utilities unquestionably have been passed the ball of collective and individual responsibility to modify, revise, and reorganize their systems, interconnections, and power pools, either in accordance with the FPC recommendations or their own sound judgment. The FPC report is an interim document—the investigations of the November 9–10 blackout are still proceeding—and its recommendations are neither binding nor mandatory.

The key warning to the private utilities, however, is contained in the just-mentioned final recommendation, which is quite explicit. If the utilities wish to avoid the reality of increased Federal intervention and stringent regulatory controls, it is urgent and in their best self-interest that immediate remedial action be taken, on their own initiative and responsibility, to minimize, to the greatest possible degree, the likelihood of a repetition of the great blackout.

Positive attitudes are required

The attainment of an efficient, economic, and reliable power supply is absolutely essential to every segment of our economy, and thus it is in the vital interest of everyone to assist in the constructive and remedial action that must be taken.

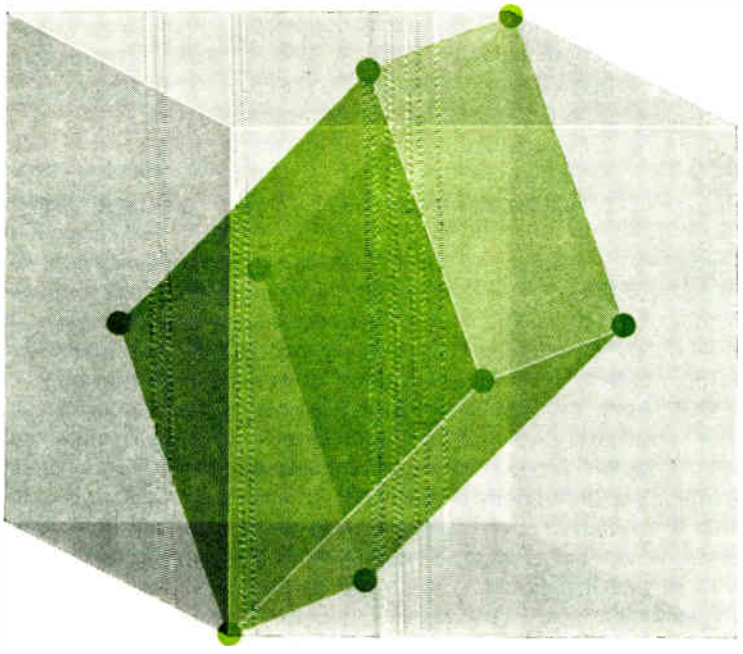
Finding scapegoats is not the answer, nor are recriminations in order. Instead, those entrusted with the development of science and technology, including the electrical engineering profession in particular, must recognize the challenge of providing for present and future power needs. Advances in communications, computer, and control techniques must be exploited where applicable; new energy sources and transmission facilities must be developed to meet rapidly growing needs for electric energy. This process must recognize the twin goals of reliability and economy in design, and produce an overall power system that meets both goals.

The power industry has made notable achievements in reliability and in lowering the cost of power to the consumer. There is every reason to believe that the important lessons learned by the November 9 event will accrue to the benefit of society in general.

The author wishes to acknowledge the valuable assistance and cooperation given to him by specialists and experts in the power industry, communications, and electrical engineering education. Figures 3, 5, 7, 8, 9, 10, 11, 14, 15, 16, and Table I are reproduced through the courtesy of the Federal Power Commission. All quantitative power values in Table I are identical to those shown in the FPC report of December 6, 1965.

REFERENCE

1. Concordia, C., "Performance of interconnected systems following disturbances," *IEEE Spectrum*, vol. 2, pp. 68–80, June 1965.



New horizons in semimetal alloys

Recent work on bismuth and antimony alloys has shown that these semimetals exhibit unique and potentially useful properties, particularly in the superconducting temperature regions

Leo Esaki *International Business Machines Corporation*

A small number of free carriers, a small cyclotron mass, a very high mobility at superconducting temperatures—these are among the properties that make bismuth, along with its alloys, the subject of intensive research efforts. The "cleanness" of these materials, for example, makes them ideal for new kinds of plasma experiments in solids. Results so far could be described as intriguing to the scientist and promising to the engineer.

Some people might consider bismuth and its alloys to be classical materials because they were leading stars among the solid-state troupe back in the 1920s and 1930s —namely, in the presolid-state electronics era. Bismuth was one of the most favored materials in classical textbooks on solid-state physics or metal physics, such as those of Wilson[1] or Mott and Jones.[2] The specific resistivity, the magnetoresistance, and the thermoelectric constant of bismuth and its alloys have been known to be unusually high compared with normal metals since the middle of the 19th century. Bismuth, then, was named the anomalous metal or the semimetal.

Early in 1930 Schubnikov and de Haas discovered an oscillatory magnetoresistance as a function of magnetic field, and subsequently de Haas and van Alphen found an oscillatory susceptibility without knowing the meaning of their discoveries. (Both effects bear their names.) In 1933 Peierls first succeeded in explaining the latter effect with Landau's quantum theory of diamagnetism developed in 1930. These effects are of significance in the history of physics because they represent the first observation of discrete levels quantized in the solid state under the application of a magnetic field. In classical electrodynamics, the diamagnetic effect of free electrons

was identically zero. The problem had been essentially solved by obtaining the solution of the Schrödinger equation with particular techniques. This was a triumph for the then-developing quantum theory and formed, in a broad sense, a basis for recent solid-state quantum electronics.

Bismuth was also one of the earliest materials tested under high pressure by Bridgman (1934), who discovered several first-order phase transitions and called this material the most versatile known metal.

Invention of the transistor in 1948 opened up the flourishing solid-state electronics era, and semiconductor physics started to grow vigorously along with other semiconductor sciences.[3] First Ge and Si were the stars, and some time later (1952) III–V compounds such as InSb and GaAs joined the group.

Meanwhile, studies on the veterans, Bi and Bi alloys, not only have persisted but have made continuous progress toward an understanding of the details of the many complex properties of these materials. These studies have been made in terms of such concepts as energy bands, electrons, holes, and phonon spectra, and the wide variety of interactions between them, and by the application of modern techniques such as cyclotron resonance and magnetoinfrared reflection, which were primarily developed for semiconductors.[4] Conversely, some typical techniques in the study of metals, such as the Schubnikov–de Haas effect or de Haas–van Alphen effects mentioned before, were applied to the study of semiconductors (1961). Indeed, theoretical and experimental studies of degenerate semiconductors began some time after the invention of the tunnel diode (1957), which requires heavily doped materials.

In semiconductors containing a large number of free carriers, studies of two kinds of collective behavior of electrons and/or holes have been carried out: solid-state plasma and superconductivity. In these aspects, one often sees Bi treated together with other semiconductors in theoretical papers, indicating that the semimetal is very close to the semiconductor.

One of the most basic and useful things that can be known about a solid is the energy of electrons and/or holes belonging to a particular band: the conduction band or the valence band in the wave-vector space, that is, the k space. The Fermi energy in metals or heavily doped semiconductors at low temperatures is defined as the highest energy that electrons can have in equilibrium, measured up from the bottom of the conduction band or down from the top of the valence band.

At this point we would like to make clear the difference between semiconductors and semimetals. The semimetal could be defined as a kind of metal in which the number of electrons is equal to that of holes, if it is pure, and usually less than 1/100 of an electron per atom. The semiconductor could be said to be a relatively narrow-gap and/or impure insulator. There is a clear distinction between metals and insulators. Where insulators have just enough valence electrons to fill an energy band completely (with the next highest band being separated by an energy gap), metals have a half-filled energy band (in monovalent metals such as Na and K) or overlapping energy bands and thus electrons can move rather freely. It is obvious that metals have conduction carriers, electrons, and/or holes, even at 0°K, whereas insulators never have them at low temperatures. The small number of electrons and holes in semimetals leads to low Fermi energy, as shown in Table I, compared with a few volts in the ordinary metal.

Thus, there is a clear line at low temperatures between the semimetal and the semiconductor; however, it is interesting to note in the BiSb alloy system that Bi makes a continuous solid solution with Sb. Galvanomagnetic measurements for some of these alloys indicates that they are semiconductors with energy gaps from zero to ~0.020 eV, depending on composition, from about 5 to about 40 atomic percent Sb in Bi; see Fig. 1. The semiconducting behavior, first noticed by Jain and later by Brown and Silverman, is now being studied at IBM's T. J. Watson Research Center. These BiSb alloys form a natural bridge connecting semimetals with semiconductors.

Large single crystals of relatively pure bismuth were grown in the 1920s and 1930s. They were probably the first large synthetic crystals of solid-state materials. It is a little surprising to see that as early as 1928 Kapitza actually practiced zone leveling or refining of Bi without recognizing the real significance of this process, which has been one of the most useful techniques for purification "reinvented" in the transistor era. Kapitza used Bi as a subject for several of his early experiments with the 300-kilooersted pulsed magnet, which is surprisingly high even at the present standard.

I. Comparison of semimetals and metals

	Number of Electrons per Atom	Fermi Energy, electron volts
Semimetals:		
Bi	10^{-5}	~0.015
Sb	10^{-3}	~0.1
As	10^{-3}	~0.2
Metals:		
Au or Ag	1	5.5
Na	1	3.2
K	1	2.1

Note: Number of atoms per $cm^3 \approx 10^{23}$.

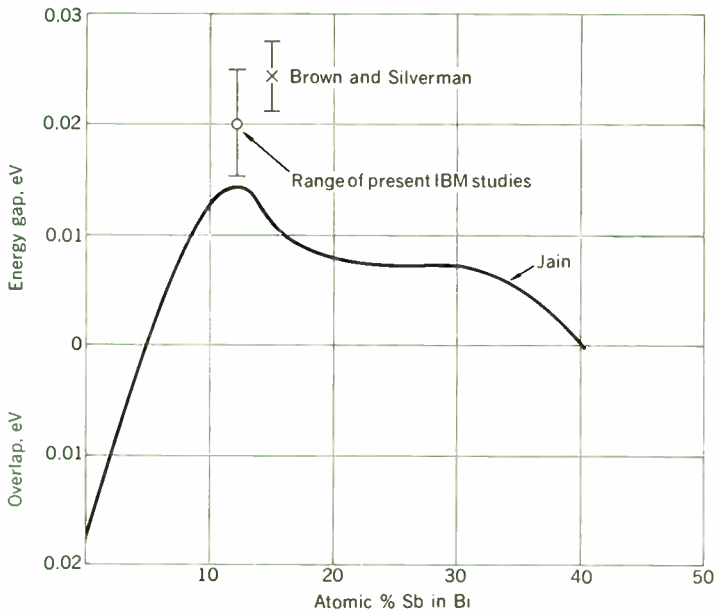


Fig. 1. Energy gap and overlapping energy as a function of antimony concentration in bismuth.

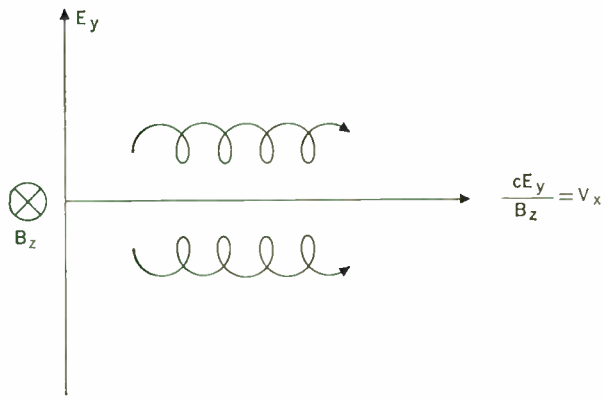
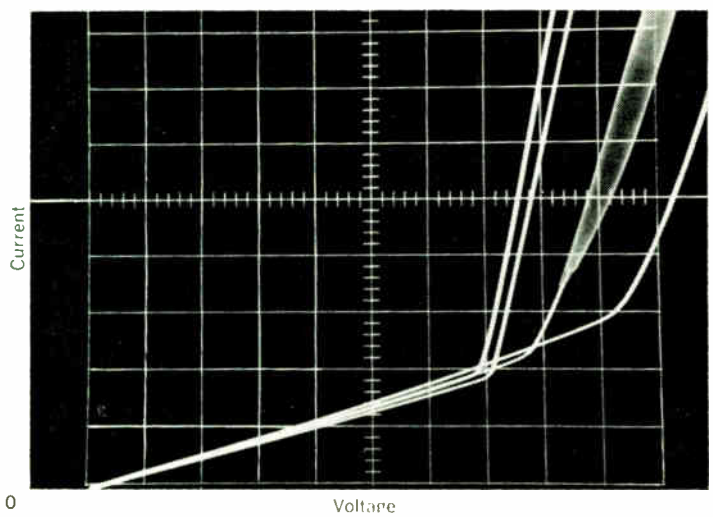


Fig. 2. Electron and hole trajectories in crossed electric and magnetic fields (in vacuo).

Fig. 3. "Kink" current-voltage curves for pure bismuth taken at 2°K, with magnetic field of 20 kilooersteds. The abscissa and the ordinate scales are 200 mV per division and 100 mA per division, respectively.



The most common semimetals are the group V elements. The group IV and group VI elements behave as acceptor and donor impurities, in the semimetals, adding free holes and electrons, respectively, analogously to the situation in the group IV semiconductors such as Ge or Si where group III and V elements are active impurities.

Bismuth has a small number of free carriers, a small cyclotron mass (less than 1/100 of the free-electron mass if the magnetic field is parallel to the bisectrix axis), a long electron-lattice relaxation time (a fraction of a nanosecond), and, hence, a very high mobility of several million $cm^2/V \cdot s$ at 4.2°K. The combination of these properties made possible the observation of quantum effects. Popularly speaking, the material is so "clean" that one can almost think of the electrons and holes as if they were in a high vacuum. The carriers travel a long distance without scattering by lattice or impurities; in other words, friction is negligible for carriers. This simplicity is obviously one of the virtues of this material. I would like to mention one example of this situation. The motion in vacuo of a charged particle of mass m and velocity v in a magnetic field B_z parallel to the z axis and an electric field E parallel to the y axis (perpendicular to each other) is classically a cyclotron rotation of angular frequency $\omega = eB_z/mc$ and radius $r = mcv/eB_z$ and a drift with velocity $v_x = cEy/B_z$ in the x direction, as shown in Fig. 2. This drift velocity turns out to be independent of both the particle's mass and its velocity, as well as of the sign of its charge. This situation can be realized in a bismuth single crystal. The current–voltage curve of bismuth at low temperatures shows a sharp increase in current observed beyond a threshold voltage, as shown in Fig. 3. It was recognized that the aforementioned drift velocity was about equal to the sound velocity of $\sim 10^5$ cm/s at this threshold voltage; hence electron and hole streams generated a sound wave and resulted in a "kink" in the curve.[5] This is a typical example of the electron–phonon interaction and also of the high degree of "cleanness" exhibited by this crystal.

The semiconducting BiSb alloy is not only one of the narrowest energy gap semiconductors, but also the most "clean-degenerate" semiconductor. A crystal doped with donor impurities of such an alloy was carefully pulled by D. F. O'Kane at IBM. The exceptionally high quality of the material was attested to by the clear observation of Schubnikov–de Haas oscillations in magnetoresistance, as shown in Fig. 4, whose amplitude was only limited by kT, even though the Fermi energy E_F is as low as 1 meV. It is remarkable that the quantum limit condition can be reached at only 2000 oersteds. Oscillations are noticeable even at 200 oersteds with this dc measurement, as seen in the figure. A necessary condition that the oscillations may be seen is that μB must be greater than 10^8, where μ is the mobility in $cm^2/V \cdot sec$ and B is the magnetic field strength in gauss. Therefore, the mobility must be greater than half a million $cm^2/V \cdot s$. An independent galvanomagnetic measurement confirmed that the mobility was of the order of one million $cm^2/V \cdot s$. As far as the writer knows, no other degenerate semiconductor shows a comparable quality in general.

This cleanness of these materials, together with their versatility in number and type of carriers, makes them ideal for new kinds of plasma experiments in solids, a most exciting field of research.[6,7] Many approaches have been tried for Bi. It might be quite desirable to investi-

gate the possibilities of developing amplifiers and oscillators using solid-state plasmas.

In 1960 Aigrain pointed out that a transverse wave of electromagnetic origin can propagate in a medium with an excess number of one type of carrier, i.e., in either n-type or p-type semiconductors, under a strong magnetic field. The counterpart of this wave in a gaseous plasma is the whistler wave, whereas equal numbers of both types of carriers are needed to support the Alfven wave. Aigrain named the wave the "helicon," and rather intuitively suggested that this wave would show an amplifying property if a dc drift field were applied parallel to the magnetic field. In 1963 Bok and Nazieres made further calculations using the Boltzmann equation, with particular emphasis on InSb, and derived conditions for instability or amplification of the wave. Misawa independently treated the same problem using the dielectric constant tensor of Bi, and concluded that instability exists, with some reservations in regard to the amplification effect (1963). The first experimental observation of this instability, as well as of amplification, was recently reported in bismuth by Bartelink, of Bell Telephone Laboratories. The effects of the helicon wave in the various BiSb alloys would certainly be an interesting subject for future study.

It should be added that the continuous solid solution of Bi and Sb and the change from metal to semiconductor will also provide a nearly ideal situation for studying specific properties of alloys in the future. There are still many theoretical and experimental problems to be answered about alloys or heavily doped semiconductors in general: for instance, how atoms of one of the constituents in alloys or impurity atoms in semiconductors are distributed among lattice points; how this distribution has an effect on the band structure or on the scattering of carriers; and whether in a specific case the band model or the localized model will provide an explanation of experimental results.

I am not going to review the whole spectrum of studies, which are probably available elsewhere in good review articles. The thermoelectric property will be entirely excluded in this article, even though this may be one of the most promising application fields for BiSb alloys.[8] Interesting studies of the magnetoacoustic effect are also omitted here because of space considerations. Instead, I would like to summarize recent studies by IBM on tunnel junctions and associated phenomena that one could call the "junction effect."

Crystal and band structures

Before pursuing the main subject, I would like to give a brief sketch of the crystal structure and band structure. All three elemental semimetals—Bi, Sb, As—and their alloys have a rhombohedral crystal structure. Although they have five electrons per atom, the atoms tend to associate in pairs, giving two ions and ten electrons per unit cell. This lattice may be constructed as follows. Look at the simple cubic lattice[9] shown in Fig. 5 as if it were made of rhombohedral unit cells, each containing two atoms, one at the corner and the other at the body center of the cell. To represent the structure of the semimetals we have to make two distortions: (1) sharpen the rhombohedral angles, reducing them from 60° to α; and (2) push the atom in the center of the cell toward its partner until the pair are separated by the fraction $2u$ (instead of ½) of the length of the long diagonal of the rhombohedron, as shown in Fig. 5.

The symmetry properties of the lattice, and hence the Fermi surface, are drastically changed by these distortions. All elements with an even number of electrons per unit cell would be insulators without the overlapping of the bands. In the semimetals Bi, Sb, and As, electrons and holes occupy only a small portion of the Brillouin zone of the k space, where bands must overlap. This slight overlap makes any theoretical prediction very difficult. Some indication of the most likely location of the overlap in the zone has been obtained from a theoretical argument. At any rate, many bands are expected to be located in a small energy range (± 0.1 eV) near the Fermi energy, as shown later. One of our objectives in studies of the tunnel junction is to locate, experimentally, the position of band edges with respect to Fermi level.

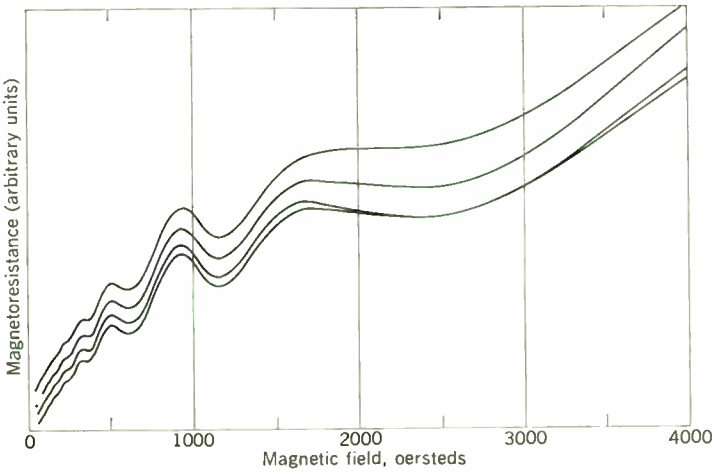
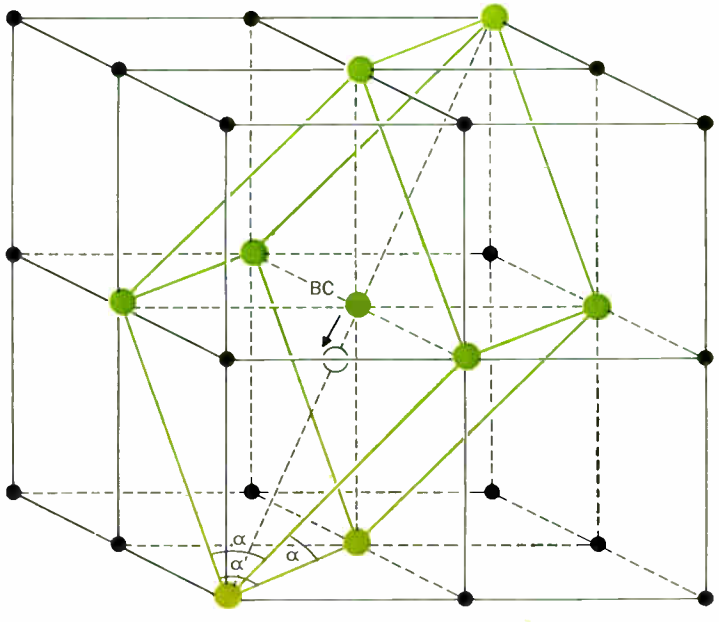


Fig. 4. Representation of the Schubnikov–de Haas effect for a semiconducting BiSb alloy.

Fig. 5. How to make the bismuth structure (1) reduce from 60° and (2) shift BC a short distance down the diagonal of the rhombohedron. (From Ziman[9])



Tunnel junctions

We have made the first attempt to make a tunnel junction on single-crystal Bi and BiSb alloys.[10] No kind of junction to these materials had ever existed before, although junctions in semiconductors are usually one of the most convenient tools to bring forth nonequilibrium phenomena such as minority-carrier or hot-electron injection. The main purpose in making tunnel junctions is to see if any fine structure exists in the tunneling current out of or into Bi. Up to now some people doubted that one could see no structure in the current–voltage curve of the metal-thin insulator–metal tunnel junction, if the metal is not superconducting. In recent studies, we constructed junctions of an insulating film several tens of angstroms thick on a cleaved surface of Bi or BiSb alloys with a counter metal electrode deposited over the insulator. Despite the previous pessimism, we have observed prominent structure in the tunneling current of this junction at low temperatures. Figure 6 shows a typical current–voltage characteristic of the BiSb tunnel junction. It is easier to see structure in a plot of the conductance of such a sample of pure Bi versus voltage at 2°K, as shown in Fig. 7. The characteristic voltages of the structure, which are much larger than any phonon energy, can be explained only by effects due to energy-band edges.

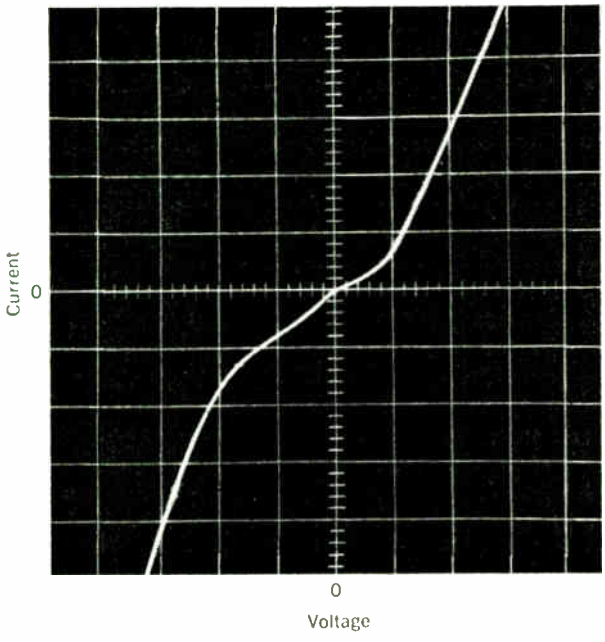


Fig. 6. Current–voltage characteristic of BiSb alloy tunnel junction at 2°K. The abscissa and the ordinate are 200 mV per division and 1 mA per division, respectively.

Fig. 7. Conductance vs. applied voltage in the bismuth tunnel junction at a temperature of 2°K.

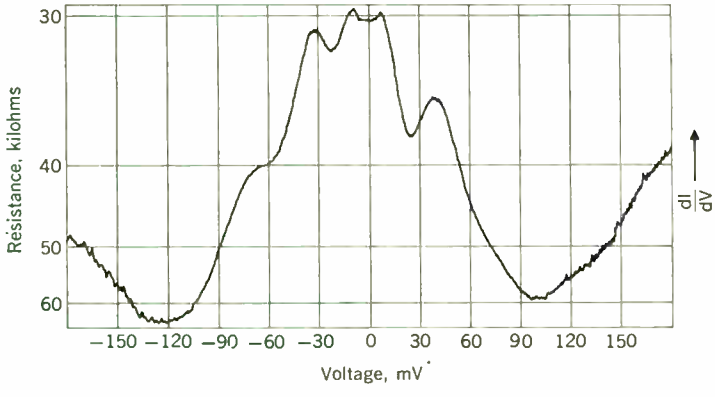


Figure 8 may be of help in understanding how band edges have an effect on the tunneling current. For the Bi tunnel junction, just eliminate the far-left Bi. In the figure, the Fermi levels in both sides are aligned, indicating no applied bias voltage. Suppose you raise the Fermi level of either side by applying voltage. Electrons, then, tunnel through the insulator, giving rise to a current. If the applied voltage attains a potential at which a band is terminated, you then might see a sudden decrease in the conductance. On the other hand, if a new band starts you then might observe a sharp increase in the conductance. Thus, this technique enables us to locate, experimentally, the position of band edges with respect to the Fermi level. We may call this process "tunneling spectroscopy."

In Fig. 7, a negative voltage shows the Bi crystal at a higher potential than the metal electrode and, hence, the electronic structure above the Fermi level in the Bi shows up. A positive voltage shows electronic structure below the Fermi level. As an approach toward interpreting the structure in this illustration, the curve was treated as a sum of the conductances from many hole- and electron-band edges, as shown in Fig. 9. In this figure, we show the four conduction and four valence bands we have observed over the range between ±150 mV. The positions of bands A, B, and C are fairly well known, and the values that we assign are in good agreement with previous estimates.

Figure 10 illustrates two rather extreme cases in the barrier tunneling: the sharp boundary and the graded boundary. In the former, the potential changes sharply as a function of position,whereas in the latter, the potential changes slowly and hence the fractional change in wavelength is small over a distance of a wavelength. In the case of the sharp boundary, electron waves are reflected by such a discontinuity, just as light waves are. With the graded boundary, however, the slowly changing potential, where the so-called WKB (Wentzel–Kramers–Brillouin) approximation can be applied, is analogous to a slowly changing index of refraction and we do not have any clear reflected electron wave, although its path may be curved as a result of refraction. The WKB approximation was used for tunneling in the semiconductor tunnel diode and tunneling across the thin insulator between simple normal metals. Because there is no dependence on the density of states in the WKB approximation, one would expect a generally smooth curve with minor fluctuations when tunneling into a semimetal. However, the observed structure is quite large. We feel that the WKB approximation is not applicable for semimetals as the electron and hole wavelength λ are large, where

$$\lambda = \frac{2\pi}{|k|}$$

The wave vector k is derived from

$$E = \frac{\hbar^2 |k|^2}{2m^*}$$

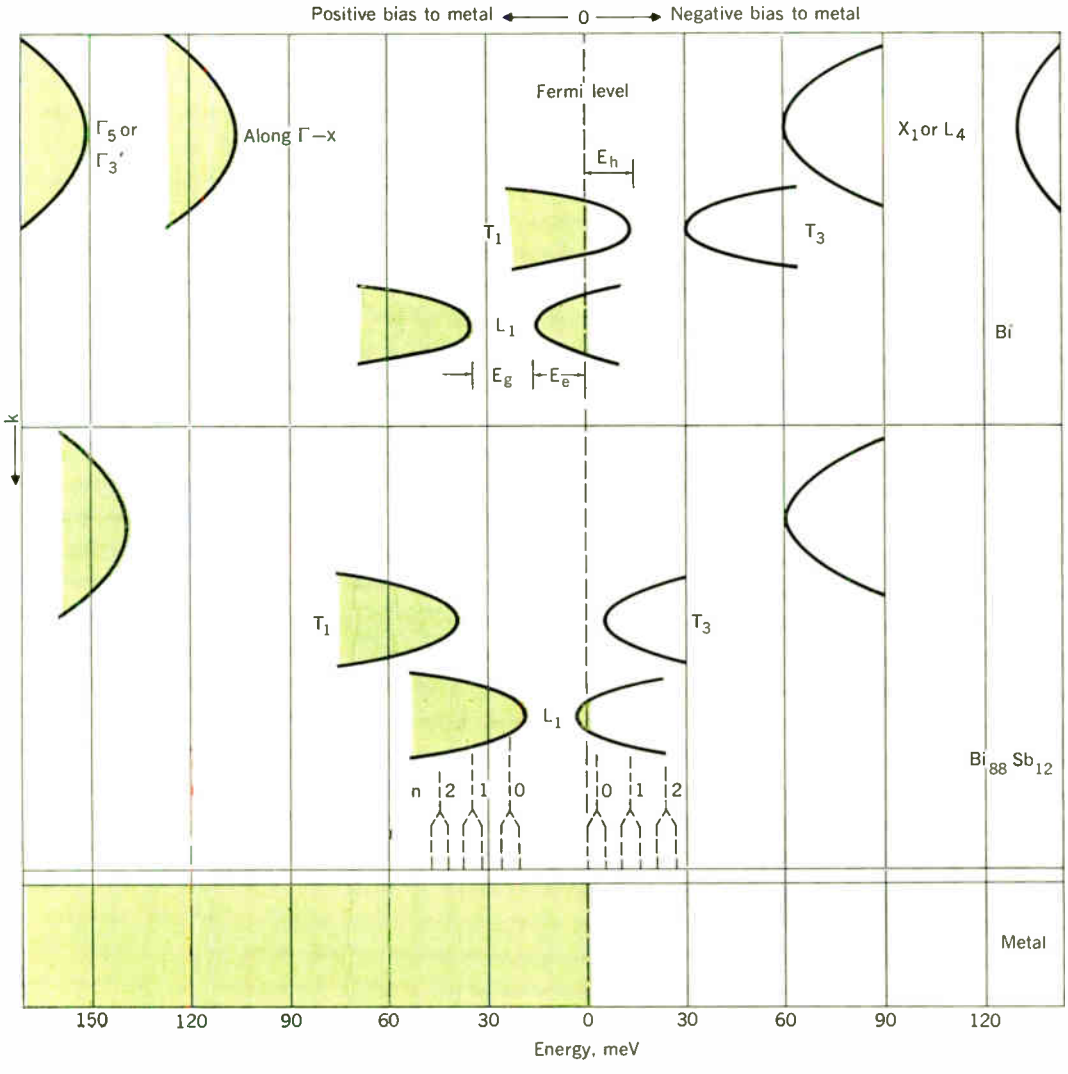


Fig. 8. Band energies for pure Bi and BiSb alloy constructed from tunneling spectroscopy data.

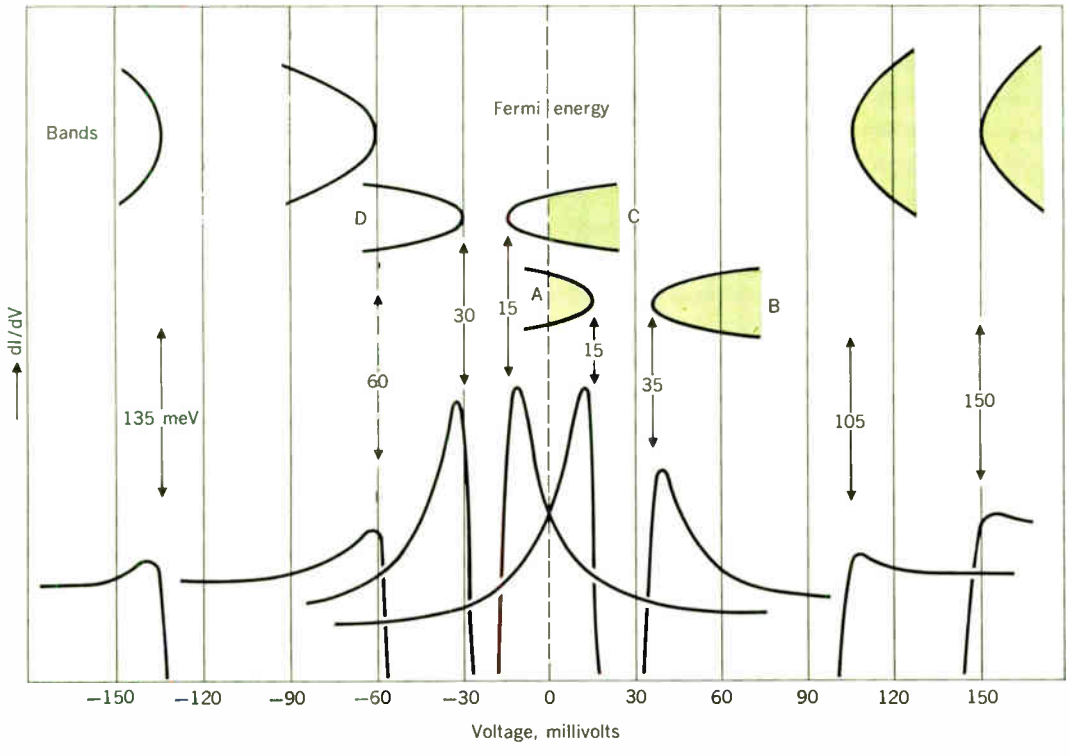


Fig. 9. Schematic illustration of each component and its related band for pure bismuth.

where E = energy measured from band edge, $\hbar$ = Dirac's constant, and m^* = effective mass. In fact, at a band edge, $E \approx 0$, the wavelengths are extremely large in any material, and one would expect a complete breakdown of the WKB approximation at this point. One reasonable explanation for the observed structure is that proposed by Harrison for the sharp-boundary case, which shows a dependence of tunneling on the one-dimensional density of states. It is quite understandable that more information can be obtained on the bulk material with the sharp boundary than with the WKB-type boundary.

In this study we have detected many bands in Bi, most of which have never been seen with other experimental methods. It is thus apparent that this technique has great potential for band-structure studies in semimetals. The optical measurements usually give the values of the intervalley and intravalley transitions; however, they are dependent upon selection rules and competing absorptions. Moreover, the values obtained from optical measurements do not have a zero voltage (energy) reference. Both methods seem to suffer considerably from surface preparation.

We now go on to the results on the tunnel junctions of semiconducting BiSb alloys. We chose the maximum energy gap alloy of 12 atomic percent Sb in Bi. The Hall-effect measurements indicated BiSb to an n-type alloy with 9×10^{15} electrons/cm^3, resulting in the Fermi energy E_F of about 1 meV. Figure 11 shows an experimental plot of the conductance versus voltage at 2°K. We have observed a large dip in conductance around −10 mV, slightly below the Fermi level. This clearly indicates the existence of the energy gap of about 20 meV in this material. It is interesting that with an increase in the magnetic field up to 4 kilooersteds a broad dip around +60 mV tends to be washed away, whereas the dip due to the energy gap is deepened, as seen in Fig. 11.

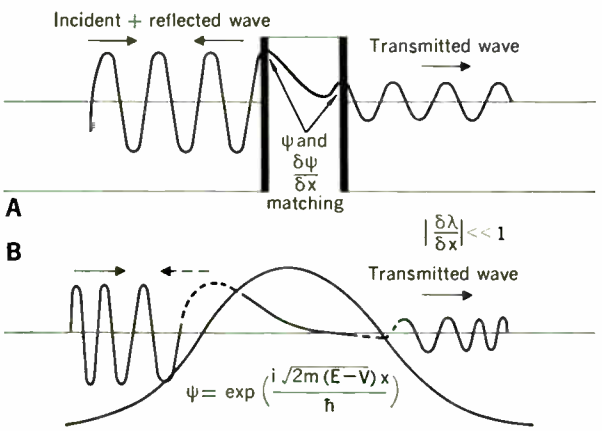
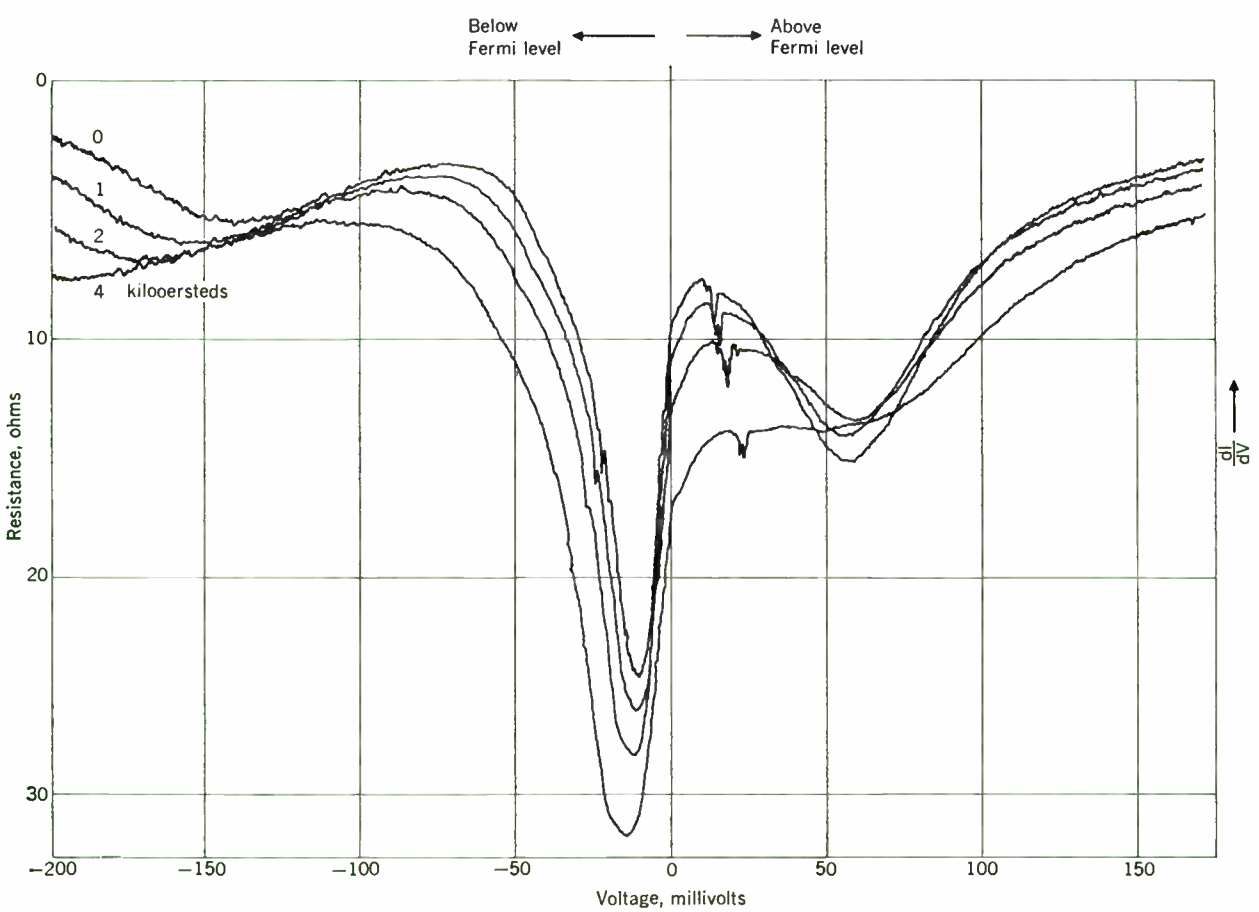


Fig. 10. Two extreme cases showing barrier tunneling of electrons. A—Sharp boundary. B—Graded boundary (with the WKB approximation used). ψ = wave function.

Fig. 11. Conductance vs. applied voltage in the BiSb alloy tunnel junction at 2°K.



(Continued on page 85)

This curve was again treated as a sum of the conductances from many hole and electron bands. We have obtained the position of many band edges with respect to the Fermi level, as illustrated in Fig. 8, together with the pure bismuth. As is shown in the figure, this alloy appears to have become a semiconductor due to widening of the energy gap between T_3 and T_1.

Superconductivity

At this point I would like to describe a new effect, recently found in the course of this tunneling study.[11] We noticed that some of our tunneling specimens have small pinholes in the insulating layer, resulting in small ohmic contacts between the semimetal substrate and the evaporated metal. We have found that an unusual property exists at low temperatures when a current flows through these pinholes. The effect is simply demonstrated in the current–voltage curves in Fig. 12. As the current is increased to a current I_c, the resistance of the sample switches from a low resistance to a high resistance. We know that a major source of this resistance is the well-known spreading resistance determined by $R = \rho/2d$, where ρ and d are the specific resistivity of the semimetal and the diameter of the pinhole, respectively. Under certain conditions, the curve shows a hysteresis as seen in the figure. This effect has three significant features:

1. The effect has no polarity; it is symmetric with respect to bias voltage.
2. The effect is strongly dependent on the temperature, as seen in Fig. 12.
3. The effect is also strongly dependent on the magnetic field, as seen in Fig. 13.

It is found that I_c is linearly dependent on magnetic field at each temperature, and H_c is defined as a critical

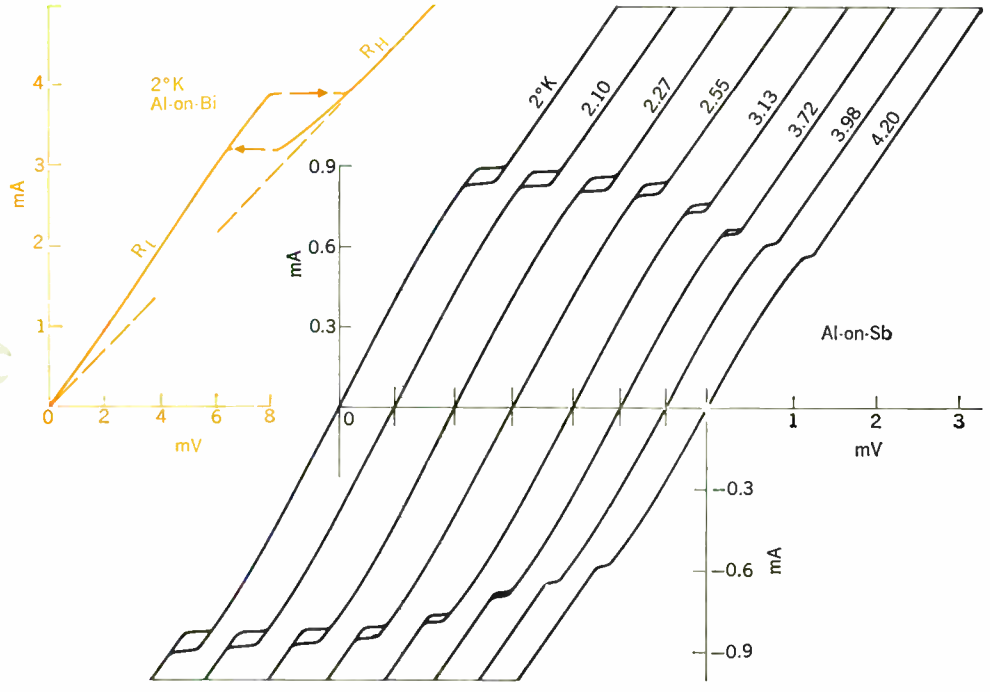


Fig. 12. Plots of current vs. voltage for Al-on-Sb at different temperatures and a plot of current vs. voltage for Al-on-Bi at 2°K, showing switching from R_L to R_H and vice versa.

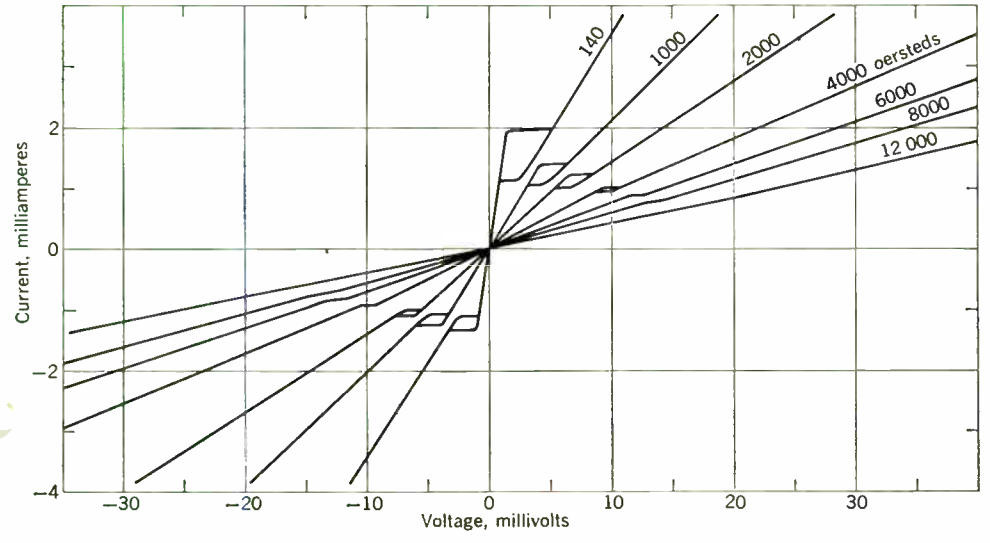


Fig. 13. Magnetic field effect on the current–voltage curve for In-on-BiSb alloy at 2°K.

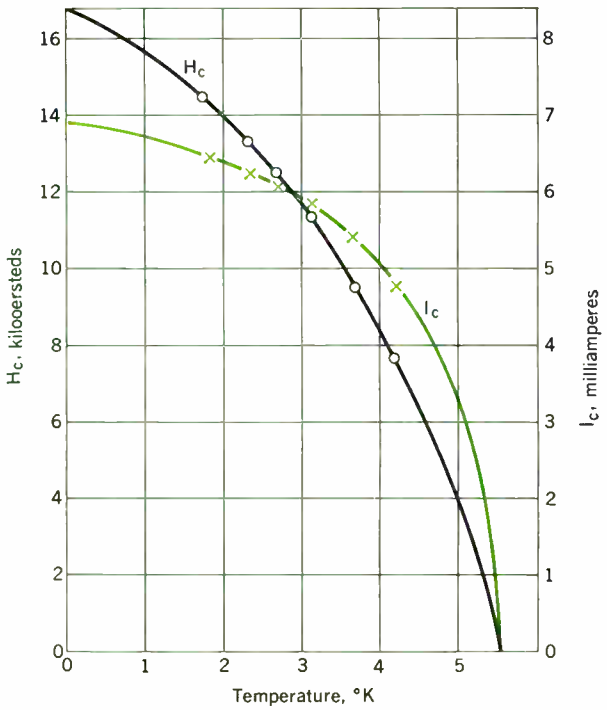


Fig. 14. Critical current I_c at zero magnetic field and critical magnetic field H_c vs. temperature for Al-on-Sb.

magnetic field at which I_c falls to zero. Measurements of H_c for different temperatures are plotted in Fig. 14 for the same sample that the temperature dependence of I_c at zero magnetic field is plotted. This figure of the Al-on-Sb case indicates that H_c and I will fall to zero at some temperature T_c in excess of 5°K. This shape of the curve is reminiscent of the BCS (Bardeen-Cooper-Schrieffer) theory of superconductivity. The general character of the phenomenon, therefore, can be well explained on a basis that a small portion of the semimetal under the contact point is superconducting over the low-current range and turns to the normal, owing to an increase in the self-magnetic field, if the current exceeds the critical value I_c. We have observed, essentially, the same effect in Bi, Sb, and BiSb alloys with the counterelectrode of In, Al, and Ag. The H_c and T_c are certainly a function of combination chosen, as seen in Table II. It is noticeable that H_c and T_c are generally fairly high.

Although the effect still needs to be studied further, we may speculate three possible explanations to account for the superconductivity:

1. *Strain effect.* It is well known that the normal phases of Bi, Sb, and BiSb alloys are not superconducting, but that the hydrostatic pressures of 25 kilobars for Bi and more than 80 kilobars for Sb cause phase transitions to high-pressure superconducting metallic phases. The possible strains involved are probably not large enough to cause these phase transitions.

2. *Alloy effect.* It is known that Bi, as well as Sb, forms a number of alloy superconducting phases. Although it is unlikely to have such alloys at the interface without further heat treatment, this possibility could not be simply eliminated because even a low current might give rise to some local heating if the pinhole is extremely tiny.

3. *Field-induced superconductivity.* There exists an electric dipole layer at the junction as the result of a difference in the contact potential, and hence excess carriers—either electrons or holes. It is possible that these excess carriers make the semimetal or the semiconductor become superconducting.

II. Critical magnetic field and critical temperature for various structures

Structure	H_c, kilooersteds	T_c, degrees Kelvin
Al-on-Sb	~20	~5.5
In-on-Sb	—	—
Ag-on-Sb	—	~3
Al-on-Bi	>40	—
In-on-Bi	~30	~4.5
In-on-BiSb	~2	~5

The current–voltage characteristic itself, shown in Figs. 12 and 13, can be suited for switching or memory application. Furthermore, if the last mechanism is true, this could lead to an interesting new device, a field-effect superconducting triode.

Conclusion

In the overall view, one might say that BiSb alloys would take an unique position among semiconductors, as Bi did and still does among metals. Some interesting results have been obtained with these materials. Although most of these results would be of scientific rather than of engineering significance, a Bi film has already been used as a magnetic-flux-sensitive resistor or a magnetic-flux meter.

The research of these materials also is of value in its own right because it offers a singular opportunity for observing intriguing properties or inspiring phenomena, under well-defined and accurately known conditions, that may or may not be predicted.

The author acknowledges his gratitude to many people at IBM for helpful suggestions, particularly to Dr. P. J. Stiles for his assistance in preparing the manuscript, and to Drs. D. J. Bartelink and R. Wolfe of Bell Telephone Laboratories.

REFERENCES

1. Wilson, A. H., *Theory of Metals*, 2nd ed. New York: Cambridge, 1953.

2. Mott, N. F., and Jones, H., *The Theory of the Properties of Metals and Alloys*. New York: Oxford, 1936.

3. Morton, J. A., "From physics to function," *IEEE Spectrum*, vol. 2, pp. 62–66, Sept. 1965.

4. Boyle, W. S., and Smith, G. E., "Bismuth," in *Progress in Semiconductors*, A. F. Gibson, ed., vol. 7. New York: Wiley, 1963.

5. Esaki, L. "A proposed new bismuth device utilizing the electron," *Proc. IRE (Correspondence)*, vol. 50, pp. 322–323, Mar. 1962.

6. Bowers, R., "Plasmas in solids," *Sci. Am.*, vol. 209, pp. 46–53 Nov. 1963.

7. Bowers, R., and Steele, M. C., "Plasma effects in solids," *Proc. IEEE*, vol. 52, pp.1105–1114, Oct. 1964.

8. Wolfe, R., "Magnetothermoelectricity." *Sci. Am.*, vol. 210, pp. 70–82, June 1964.

9. Ziman, J. M., *Electrons and Phonons*. New York: Oxford, 1960.

10. Esaki, L., and Stiles, P. J., "Study of electronic band structures by tunneling spectroscopy: bismuth," *Phys. Rev. Letters*, vol. 14, pp. 902–904, 1965.

11. Esaki, L., and Stiles, P. J., "New phenomenon in semimetals and semiconductors," *Phys. Rev. Letters*, vol. 15, pp. 152–154, July 26, 1965.

International developments in controlled thermonuclear fusion

Engineers and scientists from almost all the technologically developed nations of the world are expending a large effort to solve the problem of the controlled release of nuclear fusion energy for economic electric power production. This article is an attempt to evaluate progress toward that goal

Arwin A. Dougal *The University of Texas*

Basic processes and fundamental technological requirements of controlled thermonuclear fusion have been widely discussed. Briefly, they include: heating an ultrapure, low-density deuterium and tritium plasma to superhigh temperatures; stably containing the extreme temperature plasma by magnetic fields for a duration adequate to fuse the nuclei; diminishing particle losses occurring through diffusion and instabilities to acceptable levels; gaining useful fusion products conveying energy sufficiently in excess of thermal and radiant energy losses; and, finally, converting the energy released to useful electric power. Recent progress has been encouraging. Numerous and diverse plasmas have been produced and improved understanding and experimental confirmation of stable magnetic containment have been obtained. Also, increased plasma density, temperature, and containment times in a few experimental systems have been achieved. However, complex difficulties, such as new types of instabilities, raise formidable barriers to a workable reactor concept.

At stake in the research in controlled nuclear fusion is the exploitation of almost unlimited economic and natural resources. As an example: latent in the 0.06 pound of deuterium nuclei contained in one 55-gallon drum of tap water are 2 million kWh of energy.

The highly successful development of the hydrogen bomb to 100-megaton energies has provided impetus and motivation for the current effort. Significant advances have already been recorded; however, numerous and deep-rooted difficulties still remain to be surmounted.

Review of basic principles

In every nuclear reaction, energy is released from a change in binding energy when one or more nuclei are rearranged into others. The fusion of nuclei of the light elements into heavier ones therefore results in energy release. Reactions of interest in controlled thermonuclear fusion are

$$D + D \rightarrow He^3 + n + 3.27\ MeV \quad (1)$$
$$D + D \rightarrow T + H + 4.03\ MeV \quad (2)$$
$$D + T \rightarrow He^4 + n + 17.6\ MeV \quad (3)$$
$$D + He^3 \rightarrow He^4 + H + 18.3\ MeV \quad (4)$$

An average energy of 7.2 MeV per deuteron is realized when all four reactions are completed to convert all deuterium to helium through

$$6D \rightarrow 2H + 2n + 2He^4 + 43.2\ MeV$$

In fusion technology, it is common practice to refer to kT as "temperature." Here k is Boltzmann's constant, 0.86×10^{-4} eV/°K; a temperature of 1 eV is equivalent to 11 600°K. For a positive efficiency, deuterium plasma must be heated to temperatures of tens of keV, corresponding to hundreds of millions of degrees. Then the thermal energy of the nuclei is great enough for fusion reactions to occur at an appreciable rate. At these extreme temperatures, the deuterium becomes a fully ionized gas consisting of electrons and bare nuclei. It is called a "plasma," in which magnetic fields are used to isolate the extremely hot matter from surrounding walls.

Magnetic forces can balance the kinetic plasma pressure gradient ∇p

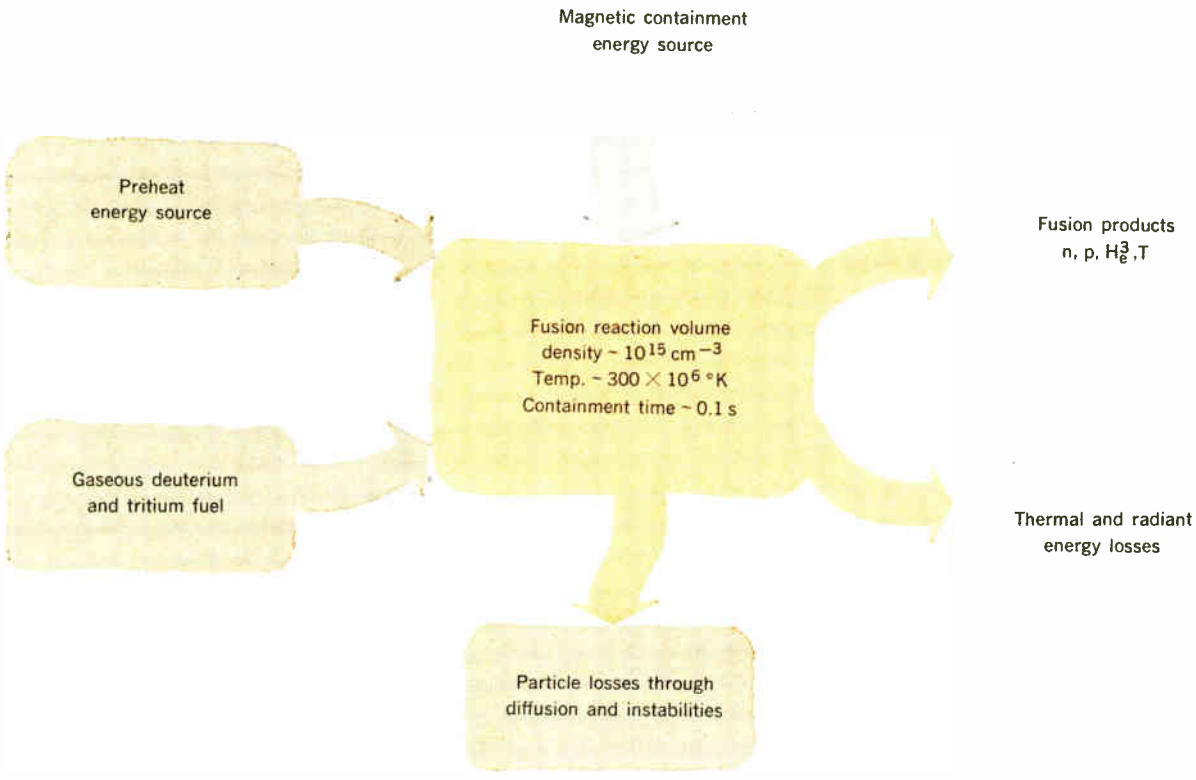


Fig. 1. Fundamental problems of controlled thermonuclear fusion: low-density plasma must (1) be heated to superhigh temperatures, and (2) be stably contained for sufficient time. Fusion products must then be converted to useful electric power.

$$\mathbf{J} \times \mathbf{B} = \nabla p \tag{5}$$

where the magnetic field **B** may be due partly to current density in the plasma and partly to external coils.

For a simple two-dimensional configuration with straight, parallel field lines, (5) gives

$$\frac{B_o^2}{2\mu_o} = p + \frac{B^2}{2\mu_o} \tag{6}$$

where B_o is the magnetic field outside the plasma where $p = 0$. A significant variable identified as β refers to the ratio between maximum plasma pressure and the magnetic pressure outside the plasma, where

$$\beta = \frac{p_{\max}}{(B_o^2/2\mu_o)} \tag{7}$$

The value of β is useful to identify various plasma–magnetic field configurations. β cannot be greater than one in an equilibrium configuration; stability considerations usually require that β be considerably less than one.

Of special interest in controlled fusion research is the product of variables $n\tau$, where n is the density of the plasma and τ is the deuteron containment time. The product $n\tau$ must range from about 10^{14} to 10^{16} s/cm^3 for the energy released through fusion to exceed the energy to heat (accounting for radiation losses). This is the basic requirement for technologically achieving a net-power-producing thermonuclear reactor.

Recent developments in controlled thermonuclear fusion research are related to the basic processes and fundamental technology depicted by the systems chart of Fig. 1. This applies in part to every known plasma source employed in fusion research, as well as to a prospective fusion reactor.

The basic processes and fundamental technological requirements of controlled thermonuclear fusion have been widely discussed.[1-5] As denoted in Fig. 1, they include: (a) heating an ultrapure, low-density deuterium and tritium plasma to superhigh temperatures in the 10^8°K range; (b) stably containing the extreme temperature plasma by magnetic fields for a duration adequate to fuse the nuclei; (c) diminishing particle losses occurring through diffusion and instabilities to acceptable levels; (d) gaining useful fusion products conveying energy sufficiently in excess of thermal and radiant energy losses; and (e) converting the energy released to useful electric power.

Specific new developments that represent significant progress, were reported at a recent international conference.* They include: (a) new methods, and improvements of existing methods, for production and heating of fusion plasmas; (b) improved understanding and experimental confirmation of stable magnetic containment; (c) identification of substantial localized in-

* The International Atomic Energy Agency's Second International Conference on Plasma Physics and Controlled Nuclear Fusion Research, held at the Culham Laboratory, Abingdon, England, Sept. 6–10, 1965.

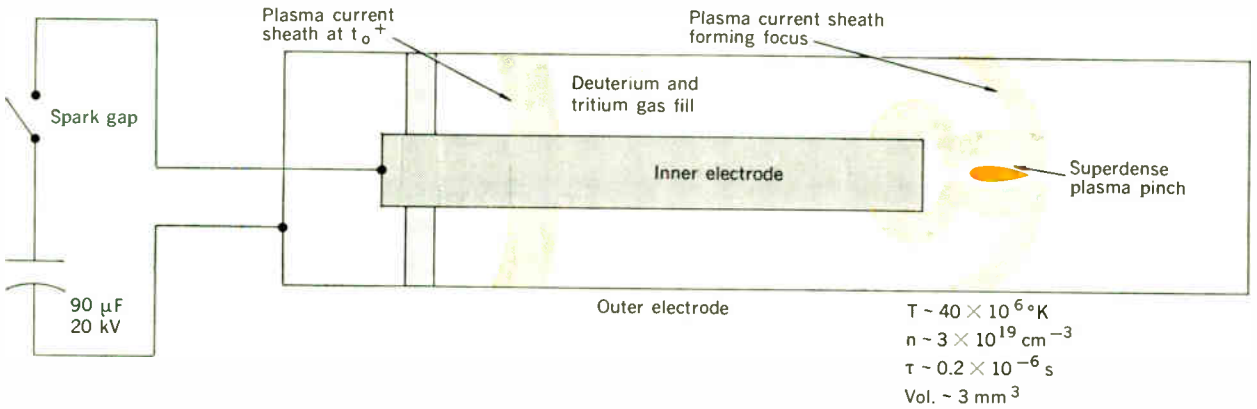
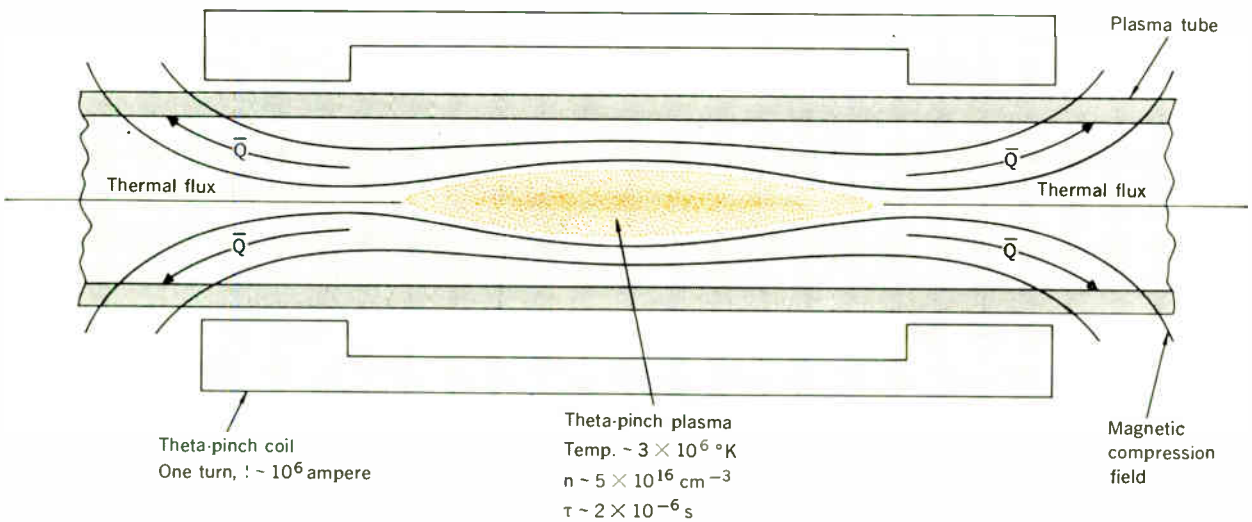


Fig. 2. Extremely dense and energetic plasma is found at focus beyond the inner electrode of a coaxial electrode discharge as a result of a spontaneous superintense pinch effect.

Fig. 3. Plasma temperature in the dynamical theta pinch with no reverse bias field is limited because of excessive heat loss through thermal conduction along magnetic field lines to the cold plasma and tube walls external to the pinched region.



stabilities (microinstabilities) and turbulence phenomena; (d) confirmation of excessive particle losses through anomalous diffusion and instabilities; and (e) achievement of increased plasma density, temperature, and containment times in a few experimental systems.

Superdense plasma pinch

A most remarkable and historical achievement in the laboratory production of fusion deuterium–tritium plasmas is the superdense plasma pinch reported by Mather,[6] and by Filippov and Filippova.[7] Figure 2 shows how a superdense plasma pinch forms at a spontaneous focus beyond the end of the inner electrode of a coaxial hydromagnetic gun. A current sheath originates at the breech and progresses toward *the end* of the center electrode due to the $\mathbf{J} \times \mathbf{B}$ force. As the sheath departs from the end, most of the capacitor energy that is stored behind the sheath is then rapidly converted into plasma energy by the inherent forces that produce the dense plasma focus.

Extensive diagnostics by Mather have established the following plasma properties: (1) a plasma temperature of 2 to 5 keV; (2) plasma volume of 1 to 5 mm^3; (3) time duration of 0.2 to 0.3 μs; and (4) neutron yields to 5×10^{10} per pulse. Neutron production has scaled almost linearly with energy.

A critical test with a D–T mixture was made by Mather. It was expected that the D–T neutron rate would increase by a factor of 100 over the D–D neutron rate because of the much larger fusion cross section. This expected increase in neutrons was experimentally confirmed and lends credence to the description of neutron production as a thermal nuclear fusion process.

Thermal losses limit peak plasma temperature

The passage of sudden, extreme currents through a single-turn coil in theta pinches, as shown in Fig. 3, primarily heats the ions.[8–11] Induction of an equal and opposite current in the plasma takes place. Heating occurs through ohmic processes and shock waves, and through fast compression of the plasma in the rapidly rising magnetic field.

Kolb *et al.*[10] extended the duration of the current in a large theta pinch to 100 μs. Measurements of electron density and temperature show that the plasma pressure is nearly equal to the magnetic pressure at 10 μs, where

$\beta = 0.8 \pm 0.2$. The heating rate is considerably higher than would be expected for shock heating and adiabatic compression. Calculations of energy balance show that the high plasma temperatures are attributable to the conversion of magnetic energy in the trapped bias field into plasma kinetic energy.

Quinn *et al.*[9] achieved peak theta-pinch magnetic fields to 180 kG with a 55-μs half period. The plasma containment was limited by end losses and the onset of hydromagnetic instabilities.

Bingham *et al.*[11] measured the energy distribution of the ions escaping through the mirrors of a theta pinch. For zero-trapped field, both the mean ion energy and spread of the distribution increased with time up to peak field. For reversed bias field experiments, the heating occurred at a much earlier instant in the cycle, with the peak occurring well before peak field. A mean ion energy of 5 keV was deduced for a peak field of 44 kG.

H. Bodin, T. Green, *et al.*[8] have analytically deduced and experimentally confirmed a basic effect that severely limits the maximum attainable temperature in theta pinches with no reverse bias field. Appreciable thermal conduction transports energy from the pinched plasma region along compression field lines to cold plasma and cold walls, as shown in Fig. 3. In the center of a theta-pinch coil there is a maximum attainable temperature T_{max} given by

$$\frac{T_{max}^{7/2}}{L^2} = \frac{\omega B^2}{\alpha 4\pi} \tag{8}$$

where L is the length of the coil, α is a numerical constant, ω is angular frequency of coil current, and B is the peak magnetic field. In the representative theta pinch with no reverse bias field, the maximum attainable temperature is of the order of 300 eV, which is far below the desired fusion temperature range. However, as was discussed earlier, much higher temperatures are achieved in theta pinches with trapped reverse bias field, although these are subject to the onset of hydromagnetic instabilities, which limit the reaction time.

Ion cyclotron resonance heating

Deposition of power into a thermonuclear plasma through radio-frequency fields at an angular frequency ω requires consideration of characteristic frequencies associated with the plasma and its magnetic field configuration. The characteristic frequency Ω_i of an ion in a magnetic field is given by

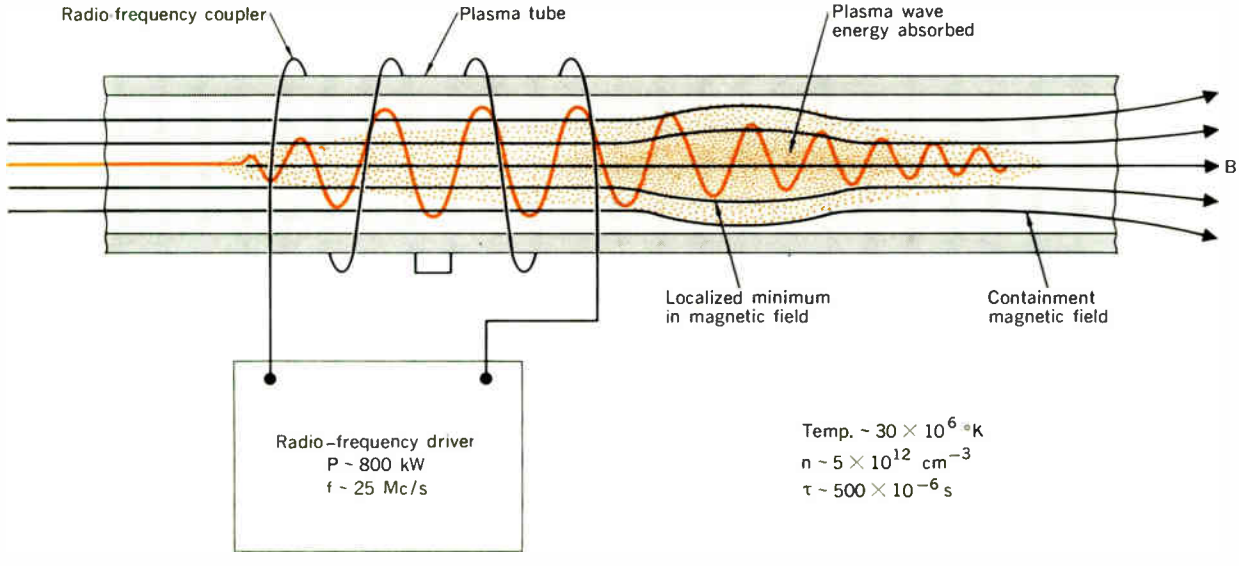
$$\Omega_i = \frac{q_i B}{m_i} \tag{9}$$

where q_i and m_i are the ionic charge and mass respectively. Earlier theory and experiments showed that effective heating of plasma ions occurs through ion cyclotron wave interactions. A simplified depiction of ion cyclotron resonance heating experiments as reported by Yoshikawa *et al.*[12] and by Matsuura *et al.*[13] is given in Fig. 4. Superhigh-power radio frequency is coupled to cyclotron motions and waves in the immediate plasma column. The ion cyclotron waves propagate to regions beyond the coil, and in doing so transport energy along the plasma column. At localized minima in the containment magnetic field, strong resonance absorption of the wave occurs with intense thermal heating of the plasma. Yoshikawa *et al.* achieved ion cyclotron resonance heating in the Model C Stellarator to ion temperatures of 250 eV throughout the working volume with a period of one millisecond. Alternatively ions were heated in local magnetic traps up to 3 keV with a period of one millisecond. Approximately 800 kW of RF were employed in these experiments in which the nominal plasma density was 5×10^{12} cm^{-3}.

Plasma instabilities and turbulence

Remarkable progress is evident from analytical and experimental reports on plasma instabilities and turbulence. Figure 5 shows the general character of plasma

Fig. 4. Medium-density plasma is effectively elevated into a moderately thermonuclear temperature regime through ion cyclotron resonance heating with superhigh power RF. Ion cyclotron waves transport energy throughout the plasma column with strong absorption and thermalization at localized minima in the containment field.



column behavior due to hydromagnetic instability and microinstability and turbulence in a hydromagnetically stabilized configuration.[14–23]

A substantially complete theory of hydromagnetic instability has evolved with evidence of plasma containment configurations realized with the minimum B configurations, which are characterized by an increase in B with some power of the radius. A major result is that the low-frequency instability associated with the hydromagnetic behavior disappears and inherent high-frequency instabilities are reduced.

Significant observations and conclusions regarding the onset of microinstabilities and turbulence have been reported. As depicted in Fig. 5 for a hydromagnetically stabilized column, passage of current and motion of electrons along the B field gives rise to localized disturbances, microinstability, and turbulence. The findings are of a contradictory nature. In the first instance, the localized disturbances contribute to enhanced particle losses perpendicular to the magnetic field. In the second instance, turbulent heating[16] occurs whereby ions are conveyed thermal energy from the electrons. This provides a most effective and unique method for heating plasma.

Kadomtsev and Pogutse[14] examined instabilities with respect to their overall deleterious effects on plasma confinement. Relatively few instabilities are considered dangerous; some can be stabilized as their effect is not too great. The worst instabilities appear to be due to temperature gradients; these cannot be stabilized by known methods, including the minimum-B configurations that are hydromagnetically stable.

An anomalous loss of particles in the containment field of the Model C Stellarator is observed to persist even in the complete absence of ohmic heating currents.[20,21] An anomalous loss rate existed for four different operating conditions: (1) ohmic heating, (2) electron cyclotron resonance heating, (3) resistive microwave heating, and (4) no heating at all. The existing overall plasma decay time is unexplained in the face of extensive, definitive, experimental data.

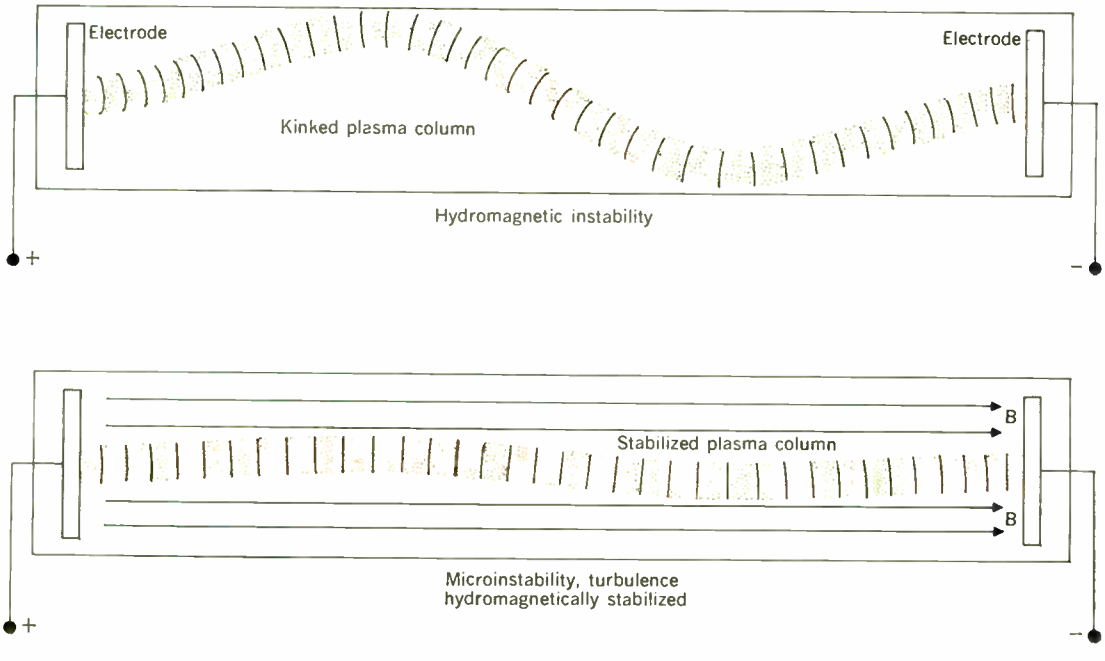
In contrast, plasma densities to 10^{10} cm^{-3} are achieved in a multiple-pass molecular ion injection experiment[23] with no evidence of hydromagnetic instabilities.

Electron or ion beam–plasma interaction

Injection of directed electron or ion beams with kinetic energies in the range from a few tens to several hundreds of kilovolts into gas-filled volumes results in effective ionization of the background gas. A striking feature of the beam–plasma interaction is the onset of intense plasma oscillations—there is effective conversion of the beam's kinetic energy to oscillations and resultant plasma heating. Figure 6 shows a representative experimental arrangement whereby an electron beam is formed and injected into a plasma-filled volume. It is noted that disturbances originate within the plasma and have a rapidly growing amplitude along the stream. Ultimately, the directed energy of the beam electrons is dissipated and they tend to lose their identity.

Alexeff *et al.*[24] report the use of electron beam–plasma interaction to generate fully ionized plasma in magnetic mirror machines. The work demonstrates that a 5-keV electron beam can heat plasma electrons to temperatures over 100 keV, and some ions to 20 keV. Electron densities of 5×10^{12} cm^{-3} are realized.

Fig. 5. Time for fusion to occur is limited by instabilities of plasma columns in containment magnetic fields. Realized are substantial theoretical and experimental findings of hydromagnetic instabilities and the means to achieve hydromagnetically stable systems. However, prevailing even in hydromagnetically stable arrangements, are localized instabilities on a microscopic scale, which cause intense turbulence that contributes to diminished containment. Also realized are new means of increasing plasma temperature through turbulent heating.



Gabovich and Kirichenko[25] report analytical and experimental work on the interaction of ion beams with plasma. Their results are relevant to improved methods of ion injection into magnetic mirrors and to collisionless thermalization of powerful ion beams. They conclude that it is in principle possible to achieve thermalization of powerful ion beams.

Smullin and Getty[26] report experiments whereby beam–plasma discharges are produced by pulsed 10–15-keV 1–10-ampere electron beams injected along the axis of a magnetic mirror into a drift region containing rarefied gas. Electron densities and temperatures of 10^{13} cm^{-3} and 10^3 eV, respectively, are achieved, while X-ray measurements indicate the existence of electrons with energies up to 100 keV. Three types of instabilities are observed—one arises during the beam pulse and is a rotating flute; another appears as a fast loss of energetic electrons associated with 50–100-keV X-ray bursts; and the third is an axial breakdown that effectively short-circuits the cathode in the electron gun to the beam collector.

Irradiation by giant pulse laser

Intense bursts of coherent light can be produced by *Q*-spoiled ruby lasers and brought to a sharp focus;

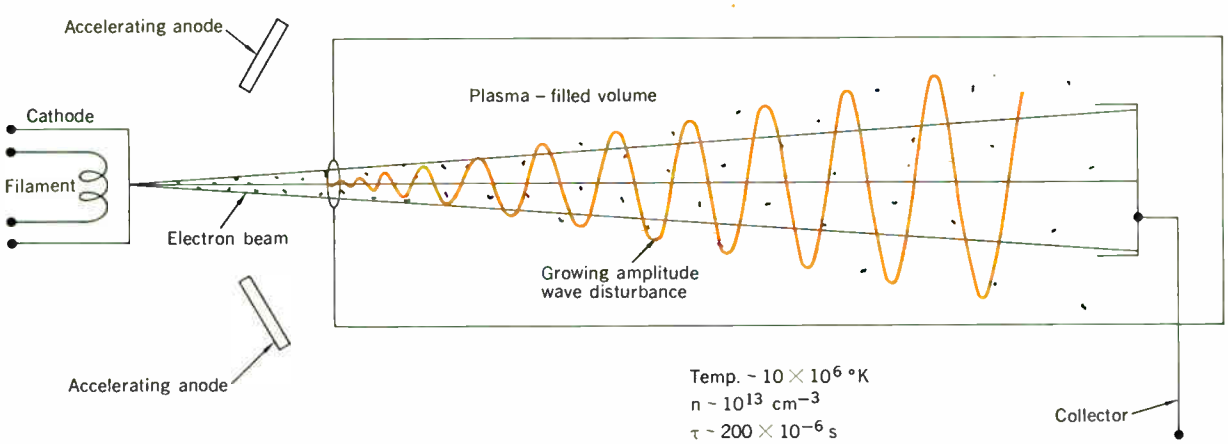
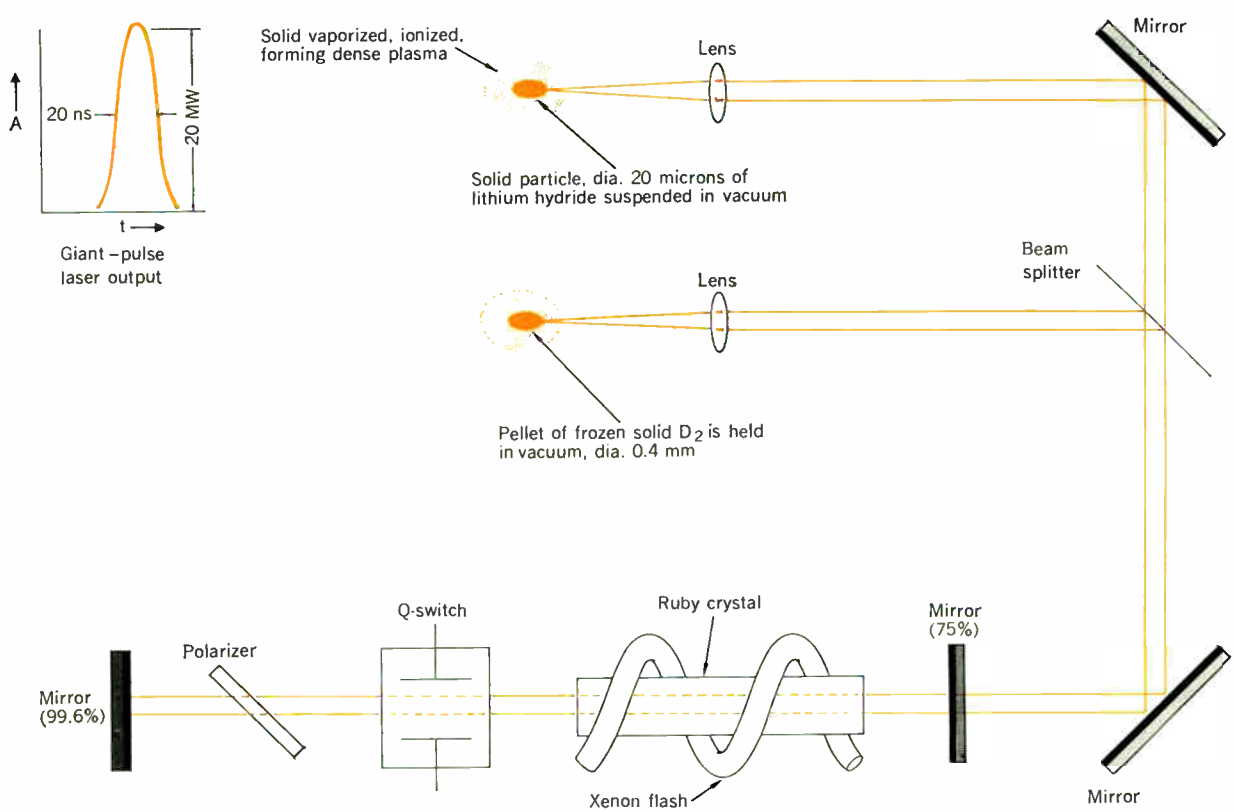


Fig. 6. Through beam–plasma interaction, directed kinetic energy of injected electron or ion beams is converted to plasma wave energy, instabilities, and thermal plasma heating.

Fig. 7. Dense, energetic fusion plasmas are produced by irradiation of minute solid particles of fusion elements by the superhigh radiation intensity at the focal spot of a giant pulse laser.



Peak powers from 10 to 100 MW are readily achieved, with pulse durations from 10 to 50 ns. An experimental arrangement for production of plasmas from solid particles is shown in Fig. 7. The giant pulse laser consists of a pair of multilayer dielectric interference mirrors, a ruby crystal illuminated by a xenon flash lamp, a polarizer, and a Q-switch that typically is an electronically actuated Kerr cell. The laser radiation is brought to a focus on a small solid particle of lithium hydride, or frozen solid D_2.

Haught and Polk[27] report producing an extremely high-density high-temperature plasma by irradiating a single, solid particle of lithium hydride suspended in vacuum. The focused 20-ns 30-MW giant pulse beam effected complete single ionization of the 10^{15} atoms in the suspended particle. The plasma temperature is believed to be in the range from 10 to 100 eV. The plasma was observed to undergo rapid radial expansion.

Ascoli-Bartoli[28] produced deuterium plasma by irradiating a frozen solid D_2 pellet with a ruby laser of 1000 MW peak output power. A dense hot plasma was obtained and measurements of density and temperature were made.

These new developments, with related techniques very recently developed for production of laser-induced discharges in superhigh-pressure gases as reported by Gill and Dougal,[29,30] open new avenues for exciting research on controlled fusion plasmas.

Conclusions

Marked progress has been made toward controlled thermonuclear fusion especially in the ability to produce in the laboratory interesting plasmas with diverse physical characteristics.

However, a workable reactor concept still remains to be demonstrated. Unfortunately, the encountering of new instability types—namely, universal instabilities and microinstabilities—is discouraging, and moves the prospective date of success further away. It is foreseen that only a tremendously dedicated effort by engineers and scientists for decades in the future can lead to ultimate success and provide society with an unlimited energy resource.

This article is based on a paper presented at the Third Annual Conference on Energy Conversion and Storage, held at Oklahoma State University, Stillwater, Okla., Oct. 28–29, 1965.

The author is indebted to the Controlled Thermonuclear Branch of the U.S. Atomic Energy Commission, the International Conferences Branch of the U.S. Department of State, the conference secretary of the International Atomic Energy Agency, and the director of the United Kingdom's Culham Laboratory; also to Otto M. Friedrich, who assisted in the preparation of the manuscript.

Research in The University of Texas' Plasma Dynamics Research Laboratory is supported by the Texas Atomic Energy Research Foundation, NSF, NASA, the USAF Aerospace Research Laboratories, and the Department of Defense's Joint Services Electronics Program.

REFERENCES

1. Bishop, A. S., *Project Sherwood, The U.S. Program in Controlled Fusion.* Reading, Mass.: Addison-Wesley, 1958.

2. Post, R. F., "Controlled fusion research—an application of the physics of high temperature plasmas," *Proc. IRE*, vol. 45, pp. 134–60, 1957.

3. Glasstone, S., and Lovberg, R. H., *Controlled Thermonuclear Reactions.* Princeton, N.J.: Van Nostrand, 1960.

4. Artsimovich, L. A., "Review of experimental results on controlled fusion research," Nuclear Fusion Supplement, pt. I, International Atomic Energy Agency, Vienna, Austria, 1962, pp. 15–20.

5. Dougal, A. A., "Problems and progress in control of thermonuclear fusion for electrical power production," *Proc. 2nd Ann. Conf. on Energy Conversion and Storage*, Oklahoma State University, Stillwater, Okla., pp. 9-1–9-8, 1964.

6. Mather, J., "High density deuterium plasma," Paper CN-21/80, *Proc. II Intern. Atomic Energy Agency Conf. on Plasma Physics and Controlled Nuclear Fusion Res.*, International Atomic Energy Agency, Vienna, Austria, 1966.

7. Filippov, N. V., and Filippova, T. I., "Phenomena accompanying the formation of a dense plasma focus during the cumulation of a non-cylindrical z-pinch," Paper CN-21/250, *Proc. II Intern. Atomic Energy Agency Conf.*, 1966.

8. Bodin, H., *et al.*, "Plasma containment and stability in a megajoule theta pinch experiment," Paper CN-21/34, *Proc. II Intern. Atomic Energy Conf.*, 1966.

9. Quinn, W., *et al.*, "Stability, heating, and end loss of a 3.5 megajoule theta pinch (Scylla IV)," Paper CN-21/92, *Proc. II Intern. Atomic Energy Conf.*, 1966.

10. Kolb, A., *et al.*, "Plasma confinement, heating, and losses in Pharos with an extended current pulse," Paper CN-21/98, *Proc. II Intern. Atomic Energy Conf.*, 1966.

11. Bingham, R., *et al.*, "Energy distribution of particles leaving a theta pinch," Paper CN-21/87, *Proc. II Intern. Atomic Energy Conf.*, 1966.

12. Yoshikawa, S., Sinclair, R., and Rothman, M., "Ion heating in the Model C Stellarator," Paper CN-21/121, *Proc. II Intern. Atomic Energy Conf.*, 1966.

13. Matsuura, K., *et al.*, "Ion cyclotron resonance heating of the QP plasma," Paper CN-21/24, *Proc. II Intern. Atomic Energy Conf.*, 1966.

14. Kadomtsev, B., and Pogutse, O., "Instability and the macroscopic effects in toroidal discharges," Paper CN-21/127, *Proc. II Intern. Atomic Energy Conf.*, 1966.

15. Dunlap, J., *et al.*, "Severe micro-instability driven losses in an energetic plasma," Paper CN-21/100, *Proc. II Intern. Atomic Energy Conf.*, 1966.

16. Bobykin, M., "Turbulent heating of plasma by a direct current discharge," Paper CN-21/154, *Proc. II Intern. Atomic Energy Conf.*, 1966.

17. Galleev, A., "Anomalous ion drift into the cone of losses," Paper CN-21/214, *Proc. II Intern. Atomic Energy Conf.*, 1966.

18. Aamodt, R., and Drummond, W., "Thermalization and anomalous diffusion of turbulent plasmas," Paper CN-21/83, *Proc. II Intern. Atomic Energy Conf.*, 1966.

19. Vedenov, A. A., *et al.*, "Oscillations and instability of a weakly turbulent plasma," Paper CN-21/155, *Proc. II Intern. Atomic Energy Conf.*, 1966.

20. Bishop, A., and Hinnov, E., "Power balance and particle loss rates in ohmically heated discharges in the C Stellarator," Paper CN-21/119, *Proc. II Intern. Atomic Energy Conf.*, 1966.

21. Stodiek, W., *et al.*, "Plasma confinement in low density Model C Stellarator discharges," Paper CN-21/120, *Proc. II Intern. Atomic Energy Conf.*, 1966.

22. Eldridge, O. C., and Harris, E. G., "Collisional and anomalous diffusion," Paper CN-21/108, *Proc. II Intern. Atomic Energy Conf.*, 1966.

23. Bell, P. R., *et al.*, "The Oak Ridge multiple-pass injection experiment DCX-2," Paper CN-21/112, *Proc. II Intern. Atomic Energy Conf.*, 1966.

24. Alexeff, I., *et al.*, "Plasma heating and burnout in beam–plasma interaction," Paper CN-21/102, *Proc. II Intern. Atomic Energy Conf.*, 1966.

25. Gabovich, M., and Kirichenko, G., "Two-beam ion instability in the case of interaction between an ion beam and plasma," Paper CN-21/222, *Proc. II Intern. Atomic Energy Conf.*, 1966.

26. Smullin, L., and Getty, W., "Characteristics of the beam–plasma discharge," Paper CN-21/122, *Proc. II Intern. Atomic Energy Conf.*, 1966.

27. Haught, A., and Polk, D., "Plasma for thermonuclear research produced by laser beam irradiation of single solid particles," Paper CN-21/110, *Proc. II Intern. Atomic Energy Conf.*, 1966.

28. Ascoli-Bartoli, U., *et al.*, "On the production of plasma by ruby laser radiation," Paper CN-21/77, *Proc. II Intern. Atomic Energy Conf.*, 1966.

29. Gill, D. H., and Dougal, A. A., "Laser induced discharges in super-high pressure gases," *Bull. Am. Phys. Soc.*, vol. II, no. 10, 1966.

30. Gill, D. H. and Dougal, A. A., "Breakdown minima due to electron-impact ionization in superhigh-pressure gases irradiated by a focused giant-pulse laser," *Phys. Rev. Letters*, vol. 15, pp. 845–847, 1965.

Radar separation of closely spaced targets

The capability of a hypothetical radar against attacks by closely spaced aircraft may be examined in geometric terms and the results can be plotted graphically. It is seen that a pulse Doppler type radar, which provides information about target velocity as well as range and angle, is essential in this kind of environment. When the targets are so closely spaced that all else fails, then separation or at least knowledge of the number of targets present is still possible if sufficient Doppler resolution is available to permit the measurement of the small changes in velocity that will be required for the aircraft to maintain formation.

A fire-control weapon system basically consists of radar, defense missile, and guidance equipment. When such a system is operating against closely spaced targets, it is often difficult for the targets to be separated (or resolved) by the radar. In cases where the separation cannot be made, the target measurements become a weighted average of the multiple target reflectors, resulting in an apparent target image that is the RF "centroid" of the unresolved group. A fire-control system operating against the centroid image will probably miss all targets of concern. In order to avoid this kind of confusion, the system obviously should be designed so that it can resolve these closely spaced targets and properly separate the measurements required for guidance against individual targets.

Radar resolution

In the present context, resolution is the ability of the radar to separate targets so that they may be tracked individually. Targets can be separately distinguished when differences occur in any one or combination of the three dimensions—angle, range, and Doppler (or radial velocity)—as measured by the radar. The resolving capability of the radar in these dimensions is a function of the characteristics of the radiating antenna, the type of signal radiated, and the length of time spent in observing the targets.

The angle resolution is inversely related to the antenna beam width; thus, the narrower the beam width the more closely the targets may be spaced and still appear separately in the antenna pattern. Similarly, the range resolution is directly related to the signal bandwidth; that is, the fineness of the signal structure (or the pulse width) is proportional to the bandwidth. It is this structure that determines the ultimate time or range resolution of closely spaced targets.

In Doppler resolution the resolving power is directly related to the precision within which the signal spectrum can be measured. This precision, in turn, depends on how long the radar observes the target; that is, the longer the target observation time the more precisely the target spectrum can be determined and therefore the higher the Doppler resolving power.

Angle resolution. As the radar antenna beam is swept across a target in space, the reflected signal will generate at the receiver a pattern that corresponds to the antenna radiation pattern; that is to say, the reflected energy will build up to a maximum when the center of the beam falls directly on the target. The energy will decay as the beam passes beyond the target, as illustrated in Fig. 1.

Consider now the return from a beam swept across two targets in space at angles θ_1 and θ_2. As the angle separa-

Fig. 1. Signal return from swept beam. A—Transmit beam. B—Receive beam.

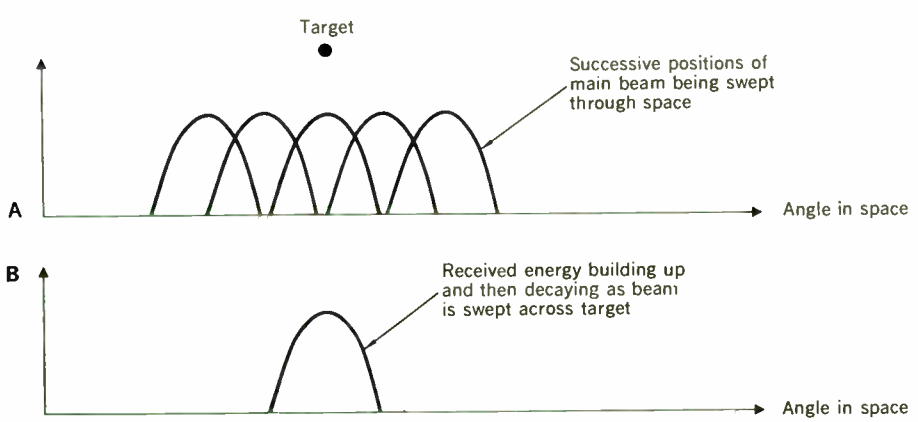
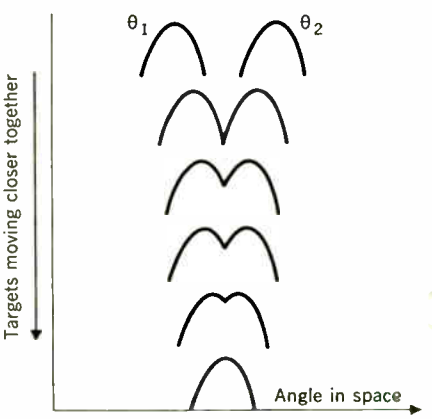


Fig. 2. Return overlap from two targets.



A common difficulty encountered by fire-control weapons systems is the so-called centroid problem, which may arise when targets are closely spaced. Pulse Doppler radar is an especially sensitive means for resolving these targets

A. Golden *Radio Corporation of America*

tion between targets is reduced, the returns will appear as shown in Fig. 2. As the targets move closer together it becomes more difficult to distinguish the two until finally they appear as one target. The narrower the radar beam, the closer the targets can come and still be separately distinguishable. Equal-energy target signals can readily be resolved in angle if they are separated by at least an antenna beam width (the angle across which the radar beam reaches half of its maximum power). Let us hypothesize that the radar antenna has been designed to provide a beam width of 1.5°. Then the angular resolution can be depicted as shown in Fig. 3.

Range resolution. Range is measured by translating the time delay between the transmission and reception of energy into distance traveled by the radiation. A direct range measurement is made by transmitting an RF pulse and measuring the time delay between the transmitted pulse and the received target echo. Two targets within the same antenna beam width but at different ranges will reflect energy from the same transmitted pulse at different times. As the targets get closer the returns appear as illustrated in Fig. 4.

Finally, as the targets approach one another the elapsed time differences cannot be distinguished. It is apparent that the resolution improves when the pulse is made narrower. (This requires a proportional increase in the signal and receiver bandwidth.) In a simple pulse system, the range resolution cell may be approximated by the effective pulse width of the received target signal. It would appear that range resolution could be extended indefinitely by decreasing the pulse width. However, there are practical limitations in that building equipment to process extremely wide-band signals is difficult and that decreasing the resolution cell significantly below the length of the target would merely make each of the targets appear as an extended source rather than as a point source.

Let us assume, then, that a pulse width of 0.2 μs (requiring a bandwidth of 5 Mc/s) is used to achieve a range resolution of 100 feet, as shown in Fig. 5. The combined resolution cell in range (100 feet) and in angle (1.5°) is shown in Fig. 6.

Doppler resolution. An additional means of resolution is available through the measurement of the Doppler shift in the target signal caused by the relative velocity of the target along the radar line of sight. Thus a continuous-wave signal transmitted as f_0 would be received at a slightly different frequency, $f_0 + \partial f$, where ∂f is propor-

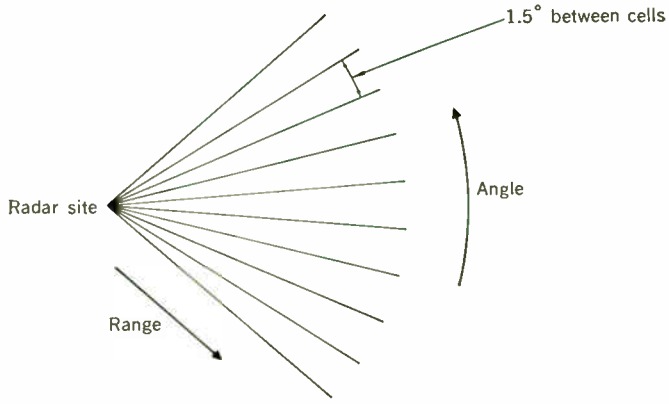


Fig. 3. Angular resolution.

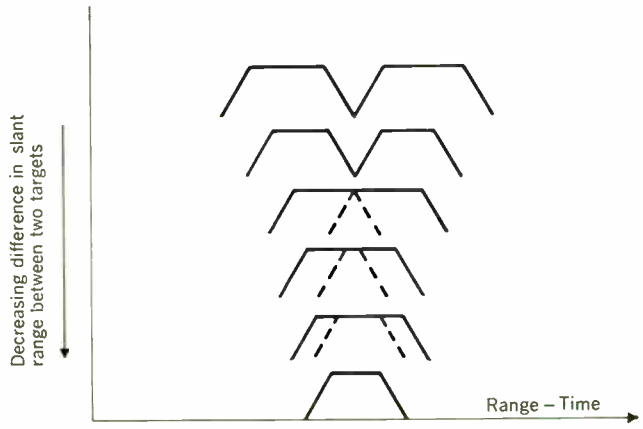
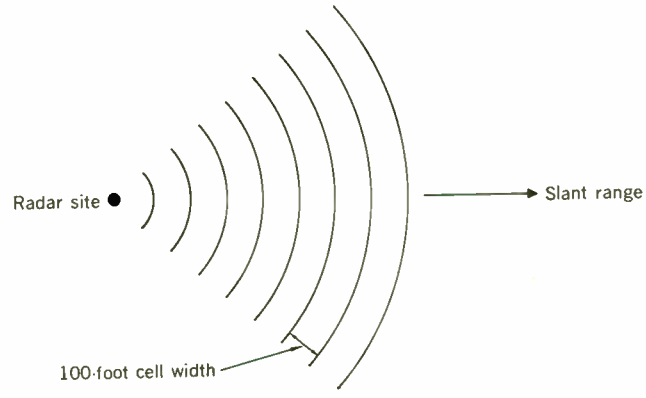


Fig. 4. Targets overlapping in range.

Fig. 5. Range resolution.



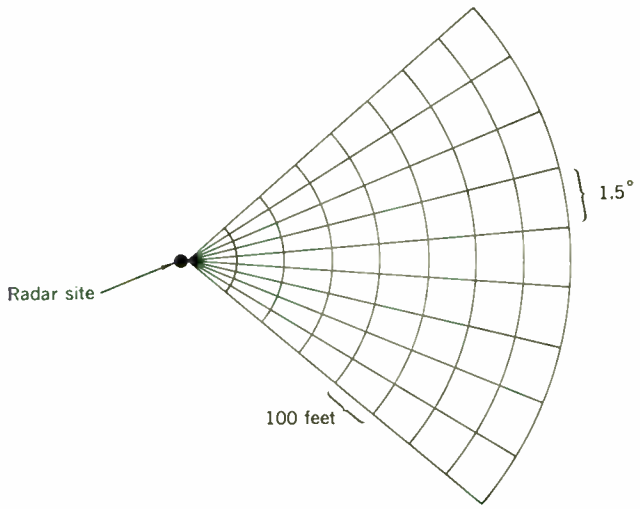


Fig. 6. Angle-range cells.

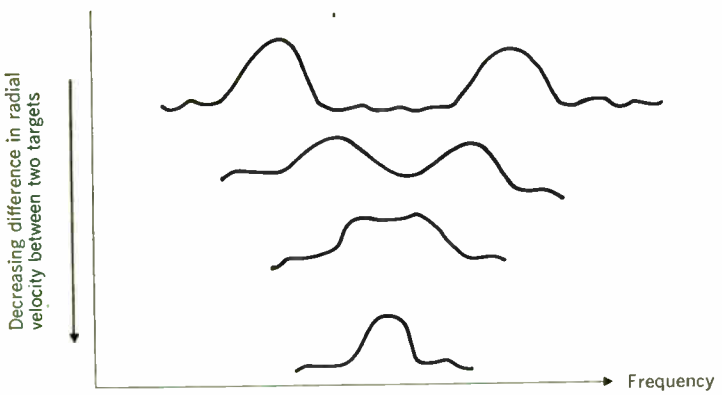


Fig. 7. Targets overlapping in Doppler frequency.

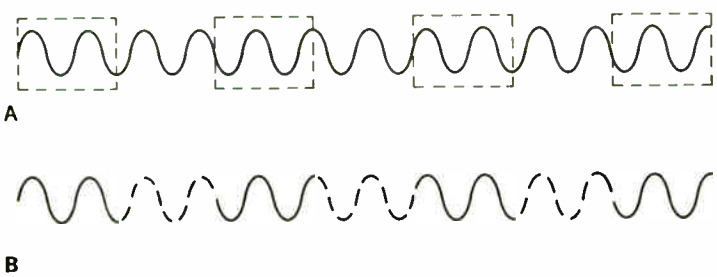
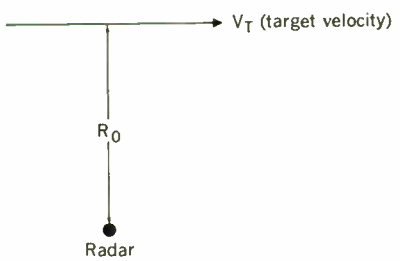


Fig. 8. Phase relationship of (A) CW and (B) pulsed signals for coherent signal processing.

Fig. 9. Typical crossing flight path.



tional to the target's closing velocity. Targets closing at different rates can therefore be resolved by separating or filtering the returned signal as a function of frequency. The resolving power in Doppler shift is reminiscent of the range dimensions for closely spaced targets (i.e., within the same range-angle cell) whose relative velocities approach one another, as shown in Fig. 7. In other words, as the difference in Doppler velocity between the two targets becomes less, resolution of the targets becomes more difficult.

To take advantage of the resolution in both range and Doppler, pulsed Doppler radar systems have been developed that provide direct range and Doppler measurements simultaneously. In essence, the fine structure in the continuous wave is preserved for Doppler measurements even though the signal is pulsed; that is, the phase structure in the sine wave is maintained from pulse to pulse, as shown in Fig. 8.

Note that the pulsed signals are shown in the same relationship as established by the CW signals. When this phase relationship is preserved and made use of in the radar receiver, the technique is referred to as "coherent" processing and is used for pulsed Doppler measurements of the target's range and relative velocity. If during an observation interval of 30 ms the radar returns can be coherently processed, a Doppler resolution capability of 33 c/s is available. The target velocity to which this corresponds is a function of the carrier frequency. If, for example, "C band" (4000 to 8000 Mc/s) is used, the 33-c/s resolution capability would correspond roughly to differentiating between targets whose radial components of velocity were more than 3 feet per second (about 2.0 mi/h) apart. Thus, assuming that targets could maintain position to escape range resolution (within approximately 100 feet), random changes in their relative velocity (greater than 2.0 mi/h) that occur as they jockey to stay in position will be detected by the ground radar.

The Doppler resolution of a pulsed Doppler radar is ultimately limited by the target's dynamic characteristics and the quality of the signal-processing equipment. If the target moves too quickly through the radar's Doppler cells the signals will not have time to build up for detection and measurement. If a filter bank is employed to separate two targets closely spaced in Doppler, as described in Fig. 7, the time required to achieve the Doppler resolution is the filter response time. The response time of the filter is equal to the reciprocal of the filter bandwidth:

$$\tau = \frac{1}{\beta}$$

where β = filter bandwidth, c/s.

If the target trajectory is such that its Doppler is changing rapidly during this period, the filter will not respond properly and thus the target may go undetected. The Doppler resolution and the filter bandwidth must be selected to achieve a maximum resolution without running the risk of not detecting the target. In order to determine the maximum Doppler rate of change that can be tolerated without degrading the filter response, let us consider a change in Doppler frequency equal to β occurring within a time interval T,

$$A_m = \frac{\beta}{T} = \beta^2$$

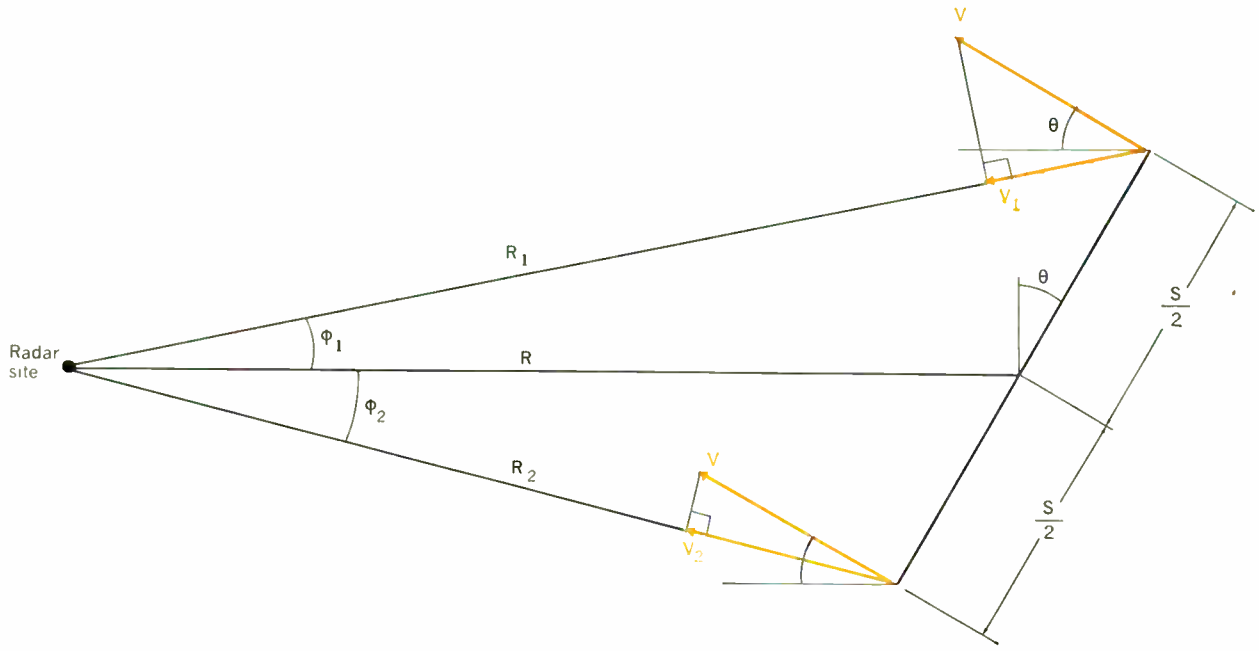


Fig. 10. Two targets approaching the radar site at a crossing angle θ.

where A_m = maximum rate of change of Doppler frequency. This maximum Doppler rate, which occurs when the target is on a crossing flight path (Fig. 9), is

$$A_m = \frac{2}{\lambda} \cdot \frac{V_T^2}{R_0}$$

It occurs at the minimum distance between the target and the radar. The actual Doppler frequency goes through zero in this limiting situation. The minimum range R_0 at which Doppler resolution of β can be achieved may be expressed as follows:

$$R_0 = \frac{2}{\lambda}\left(\frac{V_T}{\beta}\right)^2$$

In summation, we have described the three basic dimensions used in a pulsed Doppler radar system for resolving multiple targets: two space-coordinated measurements (range and angle, as shown in Fig. 6) and a dynamic measurement (Doppler). Based on the radar parameters chosen, if a pair of targets is flying a tight formation, the targets must satisfy the following restrictions simultaneously in order not to be resolved by the ground radar:

1. They must fall within a 1.5° sector when observed from the radar site.
2. They must be separated by no more than 100 feet along a radial range line measured from the site.
3. They must maintain their relative velocities with respect to the site within 3 feet per second.

Loci of resolvable targets

As a means of describing the system performance against real targets it is useful to translate the resolution capability in range, angle, and Doppler into geometric terms. To do this the combinations of target parameters (range, angle separation, velocity, and crossing angle) that yield angle separation greater than 1.5°, slant range separation greater than 100 feet, and/or Doppler velocity differences greater than 3 feet per second are plotted.

Consider aircraft targets flying in a parallel formation with the geometry shown in Fig. 10 and the general characteristics identified as follows:

S = spacing
V = velocity
V_1, V_2 = velocity components of targets 1 and 2 along radar line of sight
R = range to center of formation
θ = crossing angle of formation

From Fig. 10 it can be shown that for $R >> S$ the following relations hold:

$$\Delta\phi = \phi_1 + \phi_2 = \frac{S \cos\theta}{R} \qquad (1)$$

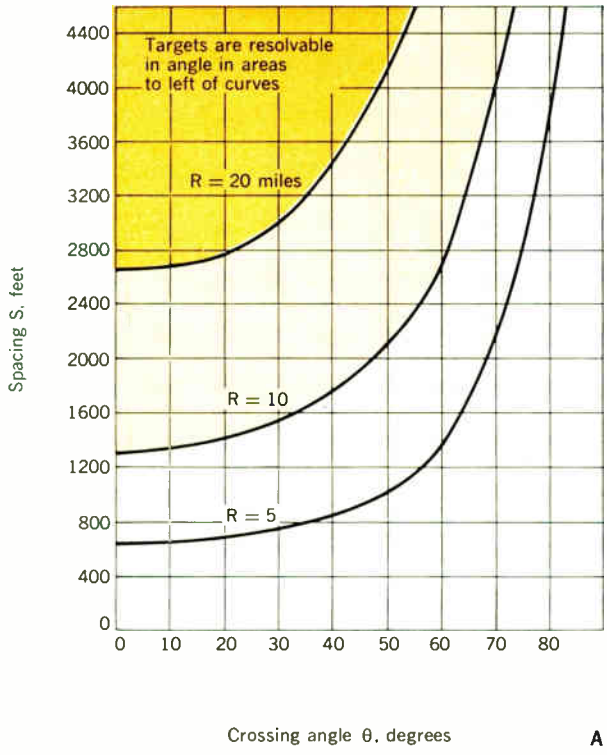
$$\Delta R = R_1 - R_2 = S \sin\theta \qquad (2)$$

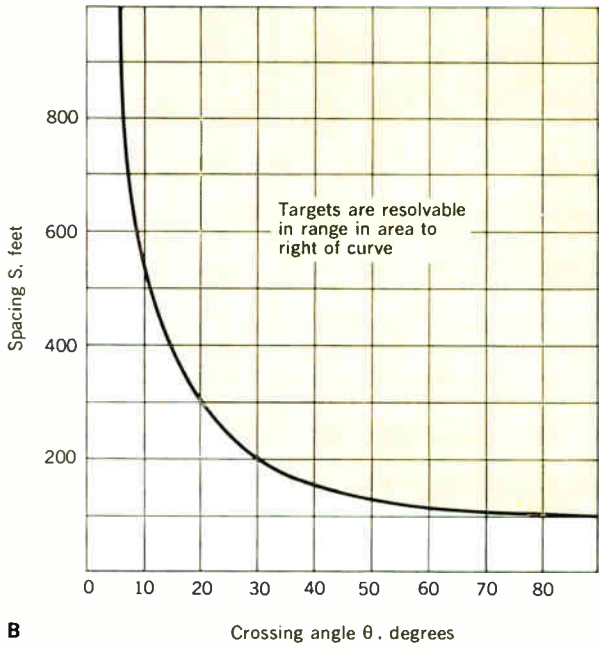
$$\Delta V = V_2 - V_1 = \frac{VS}{2R} \sin 2\theta \qquad (3)$$

By setting $\Delta\phi \geq 25.3$ milliradians (1.5°), $\Delta R \geq 100$ feet, and $\Delta V \geq 3$ feet per second (2 mi/h), these relationships are plotted to indicate which targets are resolvable as a function of range, velocity, angle separation, and crossing angle.

Figures 11(A) and (B) show the minimum spacing, as a function of crossing angle, that allows targets to be resolved in angle and range. For example, consider the curve in Fig. 11(B) relating to range resolution. If the aircraft are flying a tight formation separated by 300 feet, their course must be directed at the radar within an accuracy of 20° in order for them to remain unresolved; if the separation is 600 feet, then an accuracy of 8.5° is required.

Figure 11(C) is similar, but it applies to Doppler





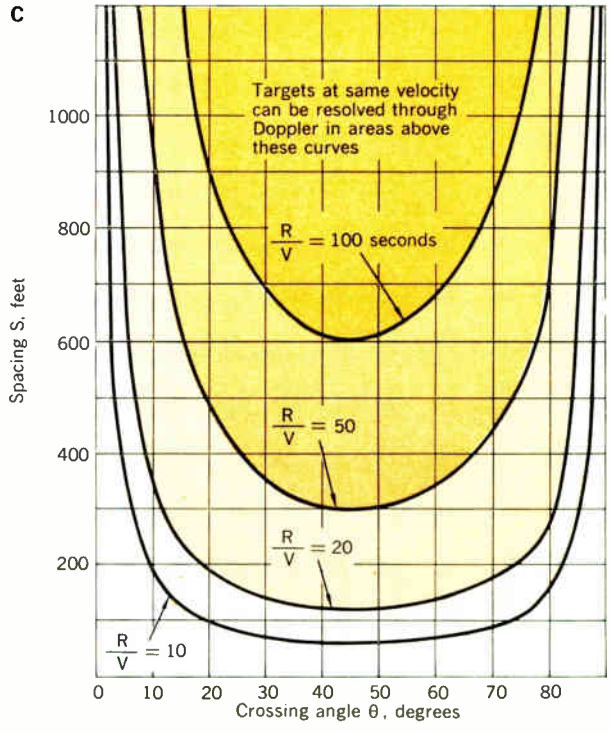


Fig. 11. Curves showing resolvability of targets in (A) angle, (B) range, and (C) velocity.

velocity resolution and assumes that all targets are at the same velocity.

Realistically, it would be unlikely that targets could keep station with sufficient accuracy to maintain their relative velocities within 3 feet per second. Thus, the number of targets in the attack group could be estimated during the early phase of the engagement because of the random wandering of aircraft speed about their assigned value.

Resolution through the missile

An additional avenue for resolution can be made available after the defense missile is launched. Up to this point reference has been made only to data gathered by the ground radar. However, if sufficient radar equipment is provided, valuable data can be gathered by the missile itself. The amount of equipment provided depends upon the type of guidance system used. Most systems operate by commanding the missile from the ground until a moderate proximity between missile and target is achieved. At this time, when more accurate data for terminal guidance and fuzing are required, the data accuracy decreases because of the ever-increasing range. This situation can be remedied through use of an antenna and simple receive/transmit system, which gathers electromagnetic energy from the target and relays it to the ground, or through use of a complete radar system and computer, by means of which the missile can seek out the target without information from the ground. In either case the closing geometry at the end of flight can be controlled and can be used to advantage in attacks against multiple targets. For example, if a group of attack craft are coordinated in an attempt to create a centroid problem against the ground radar they must fly an extremely tight formation (to a degree that would be difficult to achieve). If the defense missile approaches the attack group from any direction other than along the ground radar line of sight the individual attack craft will probably be resolved. As an example of terminal phase resolution, assume an attack group flying directly at the ground radar, where their relative positions are held within ±100 feet and their relative velocities controlled within 2 mi/h of one another, as shown in Fig. 12.

The approach of the defense missile can be controlled in order to resolve the attack craft individually for final engagement, as shown in Fig. 13. This capability can be

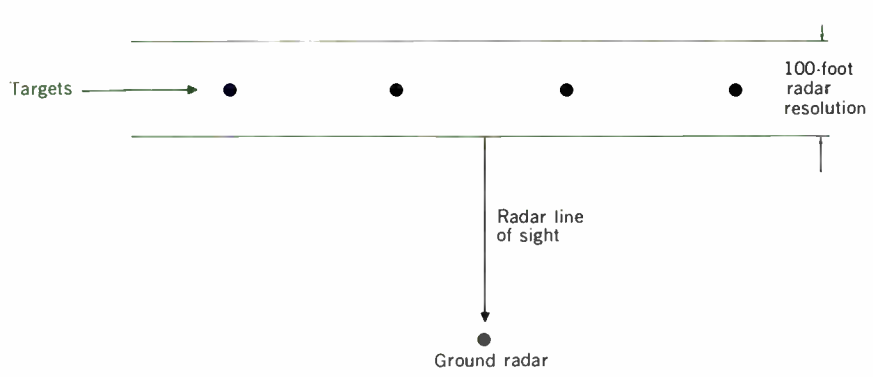


Fig. 12 (above). Terminal phase attack.

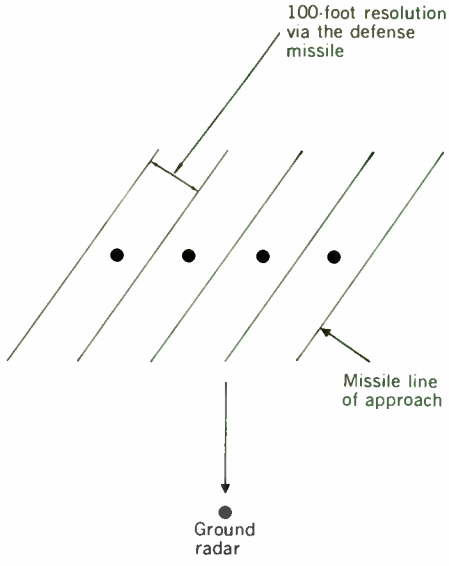
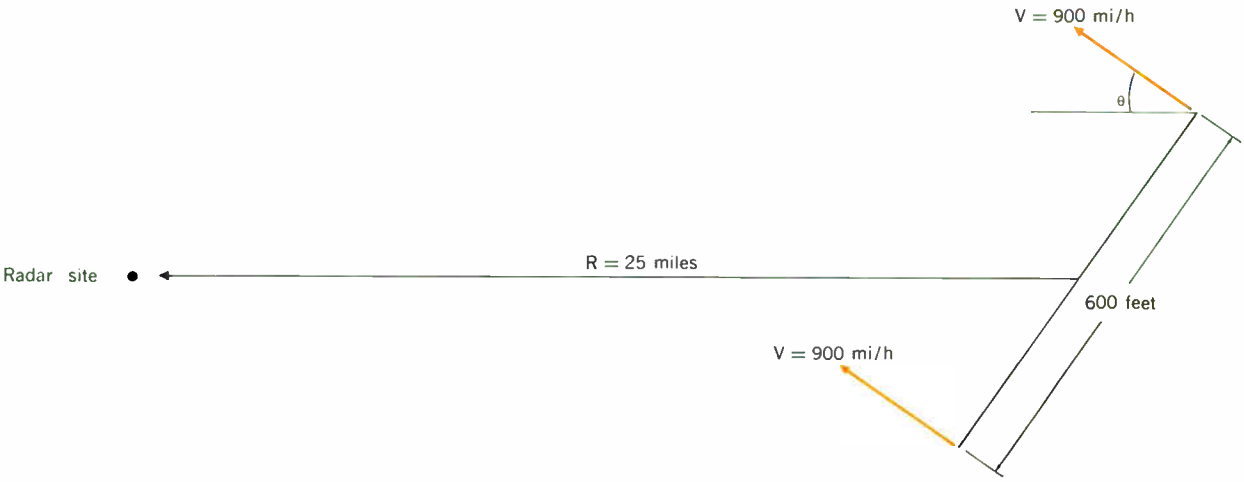


Fig. 13 (right). Terminal phase attack resolvable through missile.

Fig. 14 (below). Target model geometry.



extended for coordinated salvos where several missiles are directed to the intercept area along different lines of approach.

Performance against a specific target formation

As an example, let us investigate a target model in which two aircraft spaced 600 feet apart are flying at an altitude of 200 feet at a velocity of Mach 1.2 (about 900 mi/h). At this altitude the targets are first visible at a range of about 25 miles. The geometry is as illustrated in Fig. 14.

Angular resolution. From Fig. 11(A) it is seen that these targets cannot be resolved in angle for any θ.

Range resolution. From Fig. 11(B) it is seen that the targets can be resolved in range for $\theta \geq 9°$.

Doppler resolution. Under the assumption that the targets are always at the same velocity, it is seen from Fig. 11(C) that they can be resolved in Doppler only at $\theta = 45°$ $(R/V = 100)$.

Under the assumption that they cannot maintain a velocity spread of less than 2 mi/h, over any significant time interval, the number of targets will be immediately determined.

If it is possible to resolve the targets initially in range, angle, or Doppler, then they can be separately tracked as they approach the radar (certainly the angle and Doppler differences will increase as range decreases) and each defense missile can be guided to its assigned target.

If it is possible to estimate only the number of targets present (as would be the case if resolution could be obtained only from the station-keeping characteristics) then the defense missiles are launched and complete resolution must be obtained through the data gathered by the missile.

Conclusion

The geometric bounds have been given for which a group of targets flying against a fire-control system of specified radar parameters can be resolved. Doppler is seen to be an especially sensitive means of resolving closely spaced targets. Even if targets are successful in flying a tight abreast formation on a flight path that has a component normal to the radar face, it is likely that small variations in their relative Doppler, as the craft jockey to maintain position, will provide means for estimating the number of targets in the attack group.

A research professor leaves his classroom and laboratory to become an astronaut

It seems unlikely that many engineers and scientists would, in their first shot at the question, say yes to the idea of becoming an astronaut. But, on reflection, and with more information about how they could actually serve in the space program, they might reconsider. Dr. Owen K. Garriott, an associate professor of electrical engineering in Stanford University's Radioscience Laboratory, must be some kind of an exception. He got the bug to become an astronaut long before the first call went out for scientist–engineer applicants, and he is now a member of the first five-man group of scientists who are being trained for eventual space missions. He is, in fact, the first electrical engineer to be selected for this program, or, putting it even more chauvinistically, the first IEEE member to be so selected.

As far as his background for research in space is concerned, Professor Garriott's credentials are secure. A University of Oklahoma graduate, he earned his M.S. and Ph.D. degrees at Stanford and, after a year of ionospheric research work at Cambridge University in England on a National Science Foundation fellowship, he joined the Stanford faculty in 1961. He is project director of Stanford's satellite receiving program and a senior scientist at the Stanford Center for Radar Astronomy. His satellite monitoring station at the university's "antenna farm" has produced information about the ionosphere's electron content and rate of ionization through studies of radio signals from Sputniks, Transit, Explorer, Discoverer, and other satellites. He has taught courses in electronic circuits and electromagnetic theory, and has conducted a graduate seminar in ionospheric processes.

At the time we went to see Dr. Garriott, he was in flight training at a small airfield in Casa Grande, Ariz., where we caught him between flights. He has since gone on to jet training school and, subsequently, will go on to the NASA Manned Spacecraft Center at Houston, Tex. And some time thereafter, presumably, you will not need to read SPECTRUM to know what he is up to.—*N.L.*

Getting into the scientist–astronaut program

Dr. Garriott, when did you first think of getting into the astronaut program?

I first seriously considered it about two years ago, when I formally submitted an application to the National Aeronautics and Space Administration, indicating that I was eager to participate in the space program, and suggested that other scientists be encouraged to apply. However, NASA was not accepting applications at that time. In the autumn of 1964, a formal call went out for scientist–astronaut candidates. I then resubmitted my application to the Manned Spacecraft Center of NASA; together with a good many others, my application was processed and I was eventually selected.

What kind of tests did they put you through?

There were a number of tests. The application itself asked for a fairly extensive description of one's background and capabilities. We also took a full-day written examination of the type that applicants for graduate schools in many colleges throughout the United States take. This was a six-hour examination which basically, I believe, tested a person's aptitude and general background abilities.

Following that, the applications were sent to the Manned Spacecraft Center, which screened them regarding some of the more obvious qualifications—one could be rejected as to height, educational level, etc. The applications next were sent to the National Academy of Sciences for a more detailed screening. The National Academy compiled a list of 16 persons whom they recommended, and these 16 were invited to the School of Aerospace Medicine in San Antonio to undertake a full eight-day physical and psychological examination. From this group of 16 men, the final group of five scientist–astronauts was selected.

Do you know if others will be selected later?

The word we have from the Manned Spacecraft Center is that more will be selected. However, there has been no statement as to exactly when the call for candidates will be made.

Motivation

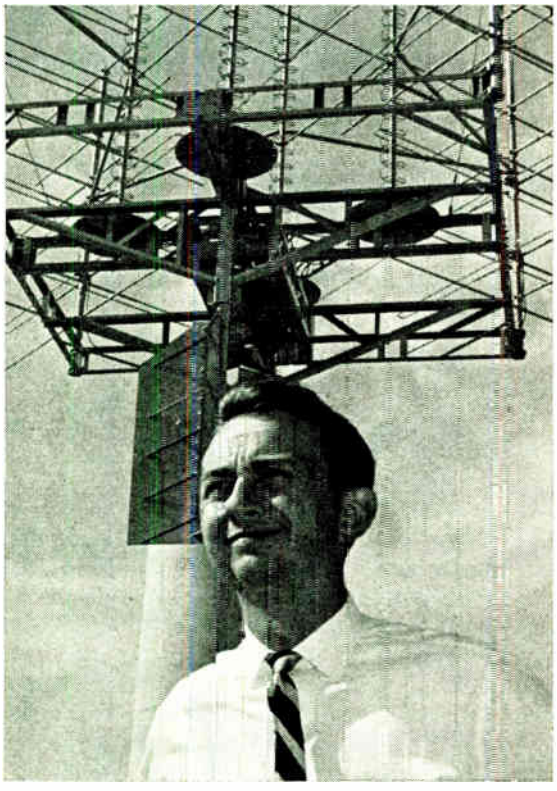
What first got you interested in the astronaut program? Did your ionospheric research lead you to think of specific experimental work that could best be carried out in space missions?

In terms of a specific experiment, the answer is no. However, I believe my general background is one that can be useful in the manned space programs; this is one of the reasons for my interest in it. There are perhaps two basic reasons why I personally am interested in the program. One is the challenge in a program like this, and the nearly unique opportunity to participate. The other is that it permits me to follow my professional objectives to some extent. The research work I had been doing, which relates to the study of the earth's atmosphere and ionosphere, is in a general area for which it would be very useful to have a man in space.

Possible in-space experiments

To do what kinds of experiments, for example?

An interview with Dr. Owen K. Garriott, the first electrical engineer to be selected for scientist–astronaut training



As one example, the intensity of the sun's rays in the ultraviolet spectrum, as they penetrate more and more deeply into the earth's atmosphere, might be studied from earth orbit with instruments on the spacecraft observing the intensity of various spectral lines. There are, of course, unmanned instruments doing just that, but there are many things that might be done better if a man were there to operate the instruments.

In a lunar orbit, which offers more interesting aspects, all sorts of electromagnetic sensing would be possible. You could study radio transmissions from the earth reflected from the lunar surface. You could yourself transmit and then receive echoes back from the lunar surface and study these as a function of position. Any number of new astronomical studies could be undertaken because you're beyond the atmosphere of the earth. In the ultraviolet and the infrared regions, you could undertake studies not possible from the earth's surface. You could study radiation from stars and from the rest of the solar system.

Do you foresee much of a direct connection between your earthbound research and your role in space?

I should imagine that the efforts of the scientist–astronauts in the space program will be necessarily very much broader than the research that we have been doing in the past. For one thing, there are very few of us, at least so far, and there are experiments to be done over a very wide range. Therefore, we're going to have to educate ourselves and to train ourselves to perform experiments other than those that we have been doing here on the ground. This is one of the reasons why a general engineering–scientific background is of more relevance to the space program than the particular specialty we followed before.

Do you feel that NASA at this point has a fairly clear idea of just how they are going to use you in this scientist–astronaut program, or do you think they are playing it by ear?

I think "playing it by ear" is a reasonably close description of the way the situation is developing. That there is a need for individuals with scientific backgrounds seems pretty clear to everybody involved, but just how we will be worked into the program probably isn't. Nor is it by any means obvious to us precisely what is the best way for us to participate.

On the proposal of space experiments*

Are the scientist–astronauts going to have a hand in proposing experiments?

Yes, we can propose experiments. If we do, however, we'll be on the same basis as any other experimenter—university, or industry, or otherwise. The proposals will have to be considered by NASA on the basis of their individual merits and then be accepted or rejected on a par with any other proposals. I hope that we do have time to generate some experiments ourselves, but most of the experiments we will be doing probably will be those proposed by others.

While on this subject, I should like to encourage researchers—engineers, physicists, etc.—who conceive useful experiments not to hesitate, but to propose them to NASA for evaluation. We are eager to see university, industry, and government agency proposals incorporated in the manned program. The participation of the scientific community has been very good in the unmanned space program, but I think it has been lagging somewhat in the manned program partly because the opportunities available simply have not been known. Now, things are becoming clearer, and NASA is anxious to push ahead in developing the best experiments possible for the manned program.

* For a review of possible manned space experiments, see Scanning the issues, page 124.

Background preparation

There might be a number of our engineer readers who for the first time are thinking of themselves as potential scientist–astronauts. For their benefit, would you mind telling us something about your own background, and more about how you see the general requirements?

As far as my own background is concerned, I became interested in engineering as a radio amateur; both my father and I are radio hams. Through that activity, my interest in electrical engineering was developed.

I entered into research on the ionosphere almost by accident. When I left the navy after three years of active duty, I applied to Stanford for a research assistantship for graduate work. This application was accepted and I was assigned to a professor who was doing research on the ionosphere; this was Prof. Allan Peterson, who incidentally is also a ham. He is one of the members of the Radioscience Laboratory; the professional staff at Stanford in Radioscience are all members of IEEE and, for the most part, radio amateurs as well. In following research of interest to Prof. Peterson and Prof. O. G. Villard, Jr., in this laboratory, my own interest developed along similar lines. As it happened, I was looking for a research topic at just about the time that Sputnik I went up, and it turned out to be very convenient to do my doctoral dissertation on the topic of ionospheric studies using radio transmission from satellites. Most of my research since then has related to ionospheric studies using satellites, which of course again relates to the space program.

In a general way, I don't see how electrical engineering could fail to be a very useful background for an individual who is interested in the space program. Almost everything related to spacecraft and to space communications comes back to electronics; an engineer's understanding of how things operate and function is bound to be of value.

For instance, the electrical engineer may be better suited than many others to work on experiments that rely on electromagnetic sensors because he is more likely to have worked with radar, radiometers, radio receivers, and what not, and he has a better understanding of how they operate. With relation to moon projects, there will be geophysical observatories which will be basically electronic devices. These need to be installed, set in operation, tested, and perhaps left for remote operation. The radar studies on the lunar surface, the infrared and ultraviolet sensors, which rely on electronics, all fall into the area that electrical engineers could well handle.

What about troubleshooting?

How much actual troubleshooting, in terms of repair, there will be is still a question to be decided. There is probably little place for a soldering iron aboard a spacecraft. But, certainly, an understanding of the system operation is going to be a very valuable asset, as is the operation of the electronic equipment related to specific experiments.

In general, the engineer whose talents and research experience are in as broad an area as possible, not only with electronic equipment but with other fields as well—geophysics, astronomy, physics, medicine, biomedical research—will be the one most likely to be of value in the space program.

Yesterday, Professor Owen K. Garriott was doing ionospheric research at Stanford University. He is shown checking radio signal recording at the satellite monitoring station on the "antenna farm."

Physical training

What other kinds of things should a potential scientist–astronaut applicant think about?

Certainly physical condition. I mention that first because that's something everyone can do something about. I think that a person who starts 12 months in advance of the time he expects to undergo a physical examination can really bring about an amazing difference in his physical condition. Personally, I think that running is one of the best exercises. It brings the whole body into action, and brings the heart and lungs, in particular, into top shape. It would be a mistake for anyone wanting to enter the program to overlook any extra physical conditioning he can do for himself. Several friends and I ran for some 6 to 12 months, prior to the examination, on a track at the university. We had quite a good time doing it as well.

Had you been running before you knew you were going to try for the astronaut program?

Not very actively. A person tends to get tied up in his other activities, and unless there is some little extra reason or motivation, he's apt to slack off. I'd certainly kept active. I've participated in sports, but nothing very regularly until perhaps a year before my selection.

Preliminary training

What about your actual training now? Do you expect to be given courses in the types of fields in which you have no prior acquaintance?

I believe so, to the extent that we have to expand our background. There will have to be some instruction program developed, or at least the opportunity for us to work with other specialists. These opportunities will presumably be made for us.

Do you get briefings on the experiences of the other astronauts?

We haven't as yet. I hope and expect that we will. Soon after the announcement of our acceptance in the program,

Today, he is in training as a scientist-astronaut, getting prepared to go up and take a closer look at what he's been measuring for the past few years.

we came to Arizona to learn how to fly jet airplanes. So our contact with the Houston group thus far has been rather slight. We assume that after we finish our jet training here, we will be working more fully with the other astronauts.

Have you flown jets?

No, I never have. That's the reason we're down here. Had I had substantial jet experience I would probably have gone straight to Houston as did two others of our group, Dr. Curtis Michel, who is a physicist, and Dr. Joe Kerwin, an M.D., who is a Navy Flight Surgeon. The others in our training group here are Dr. Edward Gibson, who worked in engineering physics at Aeronutronics, and Dr. Jack Schmitt, who was a geologist with the U.S. Geological Survey. You can see that we have a reasonable variety in our scientific backgrounds.

Some subjective views

Let's talk a bit more subjectively. Do you think you will be very scared the first time up on a space mission?

Perhaps it will not be substantially different from the first time a person flies a high-performance aircraft or from the first time he ventures into some entirely new environment such as descending far beneath the surface of the sea. I think that the first time you put yourself into a situation requiring some rather high level of performance you're bound to feel extra tension, and to some extent this is good. I believe a man performs best under a certain degree of pressure, provided it is kept at the proper level. Considering the magnitude of the space job to be undertaken, I wouldn't expect any other than a normal reaction to this new situation. My experience is that under some pressure I do perform better, very definitely.

Do you think there are many engineers who, given the chance, would think of becoming astronauts?

I think many really would. If you ask a cross section taken at random you might find a majority who'd say, "No, that's the last thing in the world I'd ever think about." But, on the other hand, if you ask a cross section of individuals who might very well be qualified for this job, and who actually are given the opportunity, I think you'd find a much higher percentage who'd be quite eager to undertake the task.

Do you find fellow teachers and engineers envious?

Some are, and would certainly jump at such an opportunity. In fact, I know some who applied for the program and, for one reason or another, had to be rejected, and they were quite disappointed. I know people in both categories: those who think that it's a silly idea and others who are envious and who would have liked the opportunity themselves.

How does your family feel about it?

I think they do not feel unduly apprehensive. Naturally, my wife, as well as I, have considered the risks involved, but neither one of us believes that there is any excessive risk.

Dr. Garriott, I understand that you are still trying to see two of your graduate students through their Ph.D.s, by commuting to Stanford, and even by the extraordinary means of ham radio. Do you regret leaving teaching?

I do leave teaching with some reluctance. There have been both joyful and taxing aspects involved. I must admit that the preparation of lectures has always been a very tedious job for me but I've always enjoyed the actual teaching itself, and this I certainly will miss. However, although I am sacrificing some things, the opportunity ahead of me more than makes up for any sacrifice. And, of course, I could very well return to teaching again at a later time.

To dig even further back for motives, do you feel that science fiction influenced you when you were very young?

No, not at all, because I very seldom read science fiction. It's not that I didn't enjoy it, but just that I never found much time for it. As far as it is concerned, I can say without any doubt that there was no influence.

What types of missions would you like to go on?

I would be very happy to participate in earth orbital missions as well as lunar missions. The scientist-astronauts will not be integrated into the program in time to work into the Gemini flights. However, we will be involved in Apollo programs and Apollo extension programs. These will involve not only earth orbital but lunar orbital and lunar landing flights, and I would be very happy to participate in any of these.

Well, Dr. Garriott, I'm sure everybody in the IEEE wishes you the best of luck. Perhaps, one day, we'll all be envying you very much.

Correlative level coding for binary-data transmission

Because of the growth in the quantity o data being transmitted over a limited number of communication channels, there is a pressing need for faster and faster systems. In addition to high speed, level-coded correlative techniques offer efficient bandwidth compression and effective error detection

Adam Lender *Lenkurt Electric Company, Inc.*

A new approach to binary-data transmission, termed correlative level coding, is presented, along with practical aspects based on research conducted during the past few years. Following a review of basic data-transmission concepts and definitions, the author discusses the general form of the level-coded correlative techniques, as well as the specific codes with and without carrier modulation. In the error-detection process, it has been found that with this type of system it is unnecessary to introduce redundant digits into the original data stream.

With the advent of computers and data-processing machines there is a growing need for the transmission of large volumes of binary data over the presently available communication channels. Simple binary transmission techniques have been and are currently being used to carry most of the data. However, the speed capabilities of binary transmission systems are limited and insufficient to meet present-day high-speed data requirements.

A logical extension of binary transmission, leading to higher speed, is the multilevel system, which uses more than two signal states; the principles were given by Nyquist as long ago as 1924.[1] Each signal level represents a group of n binary digits, rather than a single digit as in binary systems. The number of signal levels is 2^n, where $n = 1$ for binary systems. Multilevel systems have, in principle, n times the speed capability of binary systems at the expense of an increased number of levels. This implies greater sensitivity to noise and poorer error performance. Moreover, multilevel systems require complex equipment. Nevertheless, experimentation on multilevel techniques has been extensive,[2-6] and rather sophisticated systems have been developed. Since 1924, when Nyquist proposed multilevel techniques, progress has been slow in exploring other, perhaps more effective, data-transmission techniques. For this reason, new approaches have been sought so that more efficient digital communications, in terms of both performance and equipment, may be realized. Performance is usually judged in terms of speed in bits per second per cycle of available bandwidth, and the error rate is generally judged in terms of the average number of errors per bit in the presence of transmission impairments.

Review of some data-transmission concepts

A typical data-transmission system is depicted in Fig. 1. At the transmitting end the signal input consists of binary digits, all having equal duration T seconds, or, equivalently, a speed of $1/T$ b/s (bits per second). The transmission medium is a band-limited channel, and the receiver delivers a replica of the binary data in the form of binary 1's and 0's or MARKS and SPACES. The simplest form of transmission is binary, where MARK and SPACE are represented by two states in the transmission channel—for example, by two different amplitudes of a single-frequency tone or by two different frequencies or phases of a single-amplitude tone.

Suppose we assume a binary signaling system that employs a pulse to represent a MARK and no pulse for a SPACE. Such a pulse, as shown in Fig. 2(A), is identified at the receiver by the sampling process, which examines the received waveform at regular intervals of T seconds. The decision as to whether the output signal is to be a MARK or a SPACE is based on whether the waveform is above or below the slicing level at the sampling instant. The slicing level is set at a predetermined threshold, usually halfway between the steady MARK and the steady SPACE signal states. The instants of time at which the waveform intersects the slicing level represent the transition points from SPACE to MARK or vice versa. If undistorted, they are spaced exactly by T seconds, where $1/T$

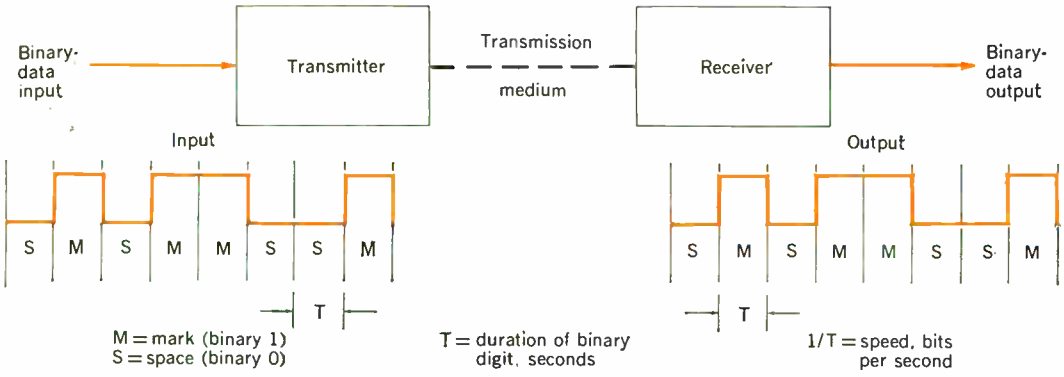
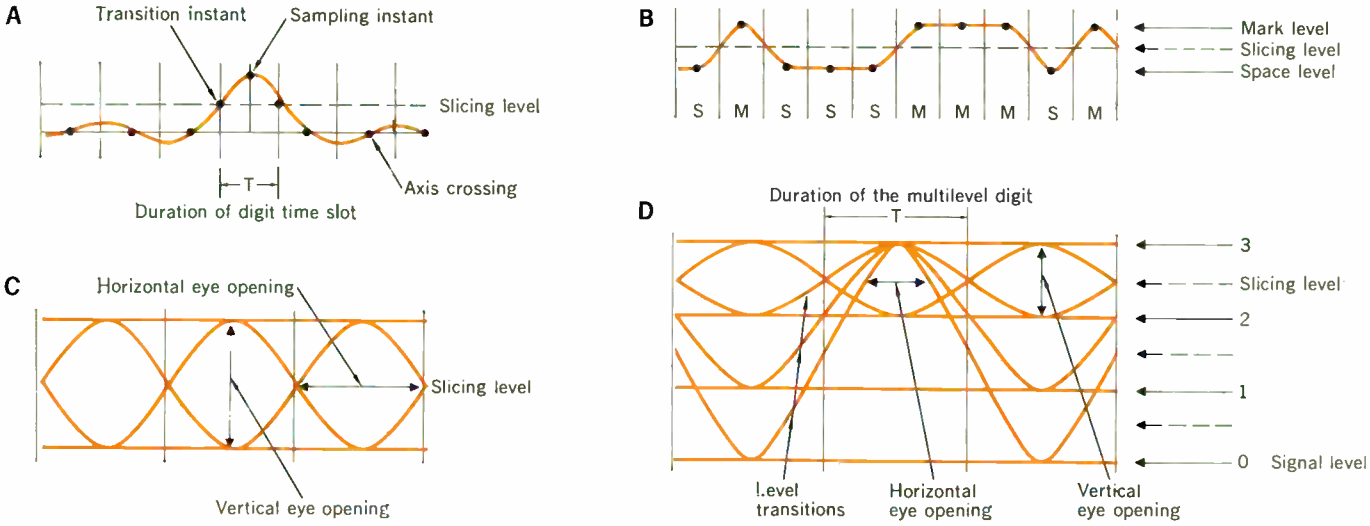


Fig. 1. Binary data transmission system.

Fig. 2. The concept of the eye pattern. A—Single mark flanked by spaces. B—Typica pulse train. C—Binary eye pattern. D—Multilevel eye pattern.



is the speed in bits per second. Distortion in the absence of noise, termed intersymbol interference, is caused by overlapping of the positive or negative overshoots of the past pulses into the time slot of the pulse being currently transmitted. Intersymbol interference is most significant at the sampling and transition instants. Suppose, for example, Fig. 2(A) is considered. The amplitude displacement of a sampling point from the MARK condition in the direction of SPACE, due to intersymbol interference, would reduce the margin to noise and transmission-line impairments. Likewise, transition points can be displaced from their correct time positions, in which case there is an increase in time jitter of a train of pulses and a corresponding reduction of tolerance to the timing imperfections in the clock-sampling pulses as well as to other transmission impairments.

A useful concept in evaluating the intersymbol interference for a long, random train of pulses is the eye pattern. It is formed by dividing a random pulse train, such as in Fig. 2(B), into segments of, for example, three bits and by superimposing as many segments as possible over the same three-bit interval. Such an eye pattern, which appears in Fig. 2(C), can be obtained experimentally by observing a random-pulse train on an oscilloscope synchronized externally with the clock pulses that drive the random data. Intersymbol interference can be evaluated in terms of the vertical and horizontal eye openings, which correspond to the sampling and transition instants respectively. These are shown in Fig. 2(C). The ratio of the actual opening to the maximum possible opening represents the degradation caused by the intersymbol interference in the absence of noise.

Nyquist[7] postulated two criteria for the elimination of intersymbol interference at the sampling and transition instants. The first criterion is that the zero crossings of the time axis, shown in Fig. 2(A), be equally spaced by T seconds to permit binary signaling at the rate of $1/T$ b/s. This assures that at the sampling instant of any par-

ticular pulse, the overshoots of all past pulses are zero and the vertical eye pattern will be a maximum. Such a condition is met by pulses representing the impulse response of an ideal, linear-phase, rectangular filter with a cutoff frequency of f_1 c/s; the binary speed is then $2f_1$ b/s. The response of such a rectangular filter is the well-known $(\sin 2\pi f_1 t)/2\pi f_1 t$ pulse, which has relatively large overshoots that fall off as $1/t$. As a result, the intersymbol interference in terms of the horizontal eye opening would be exceedingly large and make such a system useless, apart from the fact that such a rectangular filter with linear phase is unrealizable. The second Nyquist criterion is that the times between the transition instants be equal—a condition that assures a perfect horizontal eye pattern. One way of fulfilling this criterion is to have

$$Y(f) = \cos \frac{\pi f}{2f_1} \quad \text{for } 0 < f < f_1 \qquad (1)$$

and zero elsewhere

where $Y(f)$ is the channel system function. In (1) the transmission at $f = f_1$ is zero and binary signaling at the rate of $2f_1$ b/s is no longer possible assuming, for example, a steady on–off input, such as 10101010 . . . . In fact, the vertical eye pattern would nearly collapse if an attempt were made to signal at $2f_1$ b/s. If we wish to meet both criteria for binary signaling and to assure maximum eye openings in the vertical and horizontal directions at the rate of $2f_1$ b/s, one possibility is the cosine-squared low-pass filter, in which

$$A(f) = \cos^2 \frac{\pi f}{4f_1} \quad \text{for } 0 < f < 2f_1 \qquad (2)$$

and zero elsewhere

The impulse response of this filter has axis crossings spaced at half signaling intervals. It should be noted that in terms of physical channels, a binary signaling rate of $2f_1$ b/s per cycle of bandwidth would imply the use of a rectangular filter with a cutoff frequency of f_1 c/s. This would not be possible.

Although the cosine-squared filter with a bandwidth of $2f_1$ c/s permits distortion-free binary transmission at the rate of $2f_1$ b/s, this statement is not true for the previously mentioned multilevel systems. The vertical eye opening remains relatively unaffected, but the horizontal eye deteriorates rapidly with the increasing number of levels because of the increased number of possible transitions between the levels. This is illustrated in Fig. 2(D) for the case of a four-level system, where each level represents two binary digits: 00, 01, 11, and 10. Inasmuch as there are four levels, actually three eyes exist: one between each pair of adjacent levels. For simplicity, only the topmost center eye is presented in detail. The vertical opening is still near a maximum as in the binary case, but the horizontal eye opening at the slicing level is only a fraction of the digit duration T seconds, which corresponds to the maximum possible opening. Such a reduction in the horizontal eye opening is a measure of the deterioration of the signal. It is apparent that the primary contributors to this deterioration are the transitions between the ex-

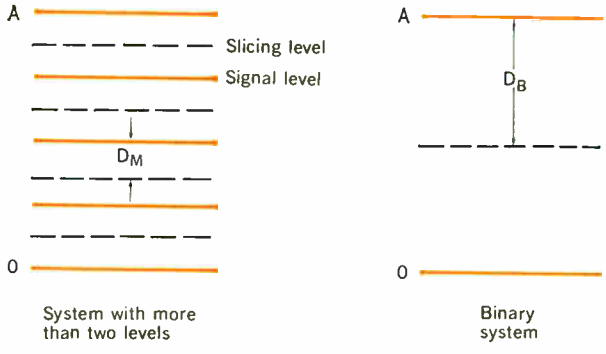
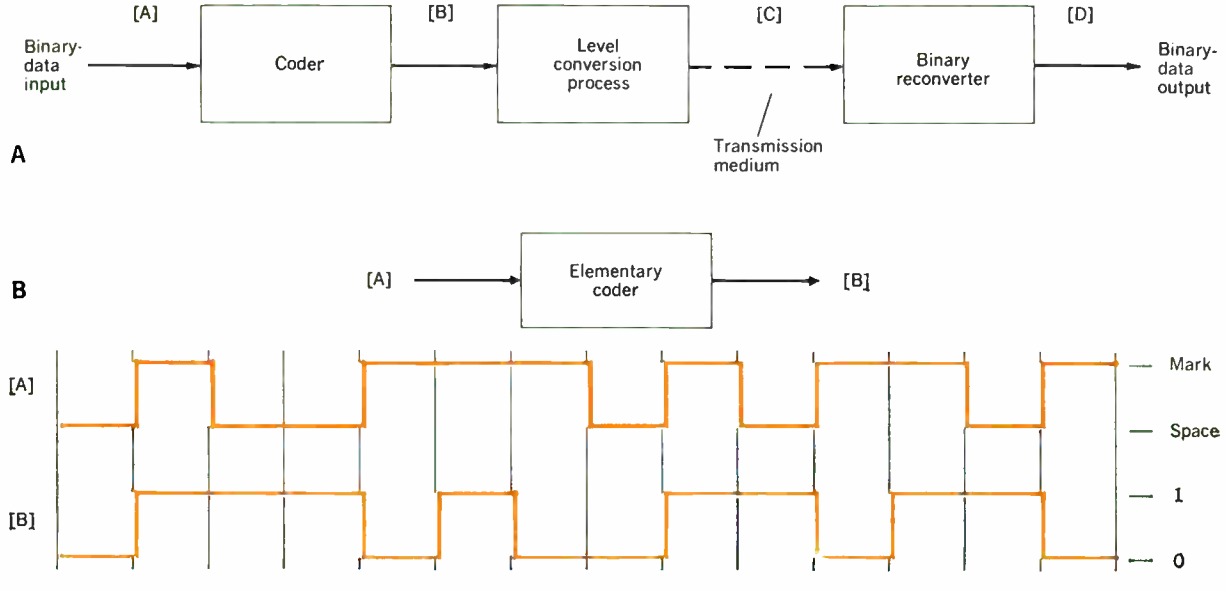


Fig. 3. Approximate noise penalty for a system in which there are more than two levels.

Fig. 4. Level-coded correlative system. A—General system diagram. B—Input-output relationship of elementary coder with corresponding waveforms.



treme levels 0 and 3, while the transitions between adjacent levels such as 2 and 3 do not affect the horizontal eye opening at all. Such intersymbol interference in terms of horizontal eye opening is inherent in multilevel systems and is irreducible. For comparison, in a class of level-coded correlative systems, where energy is redistributed in such a manner that it is concentrated at low frequencies, only transitions between adjacent levels are permitted. As a result, the intersymbol interference is often smaller than in multilevel systems.

Since multilevel and level-coded correlative systems employ more than two levels, they incur a noise penalty with respect to binary systems. Although exact calculations of such penalties are complicated and often cumbersome, there does exist a quick method of finding an approximate value. In Fig. 3 both types of systems are shown. Assuming an equal peak voltage of A volts for both cases, the noise penalty is merely the ratio of the distances between any signal level and the adjacent slicing level for the two systems. The corresponding distances are $D_m = A/2(b-1)$ for a b-level system and $D_b = A/2$ for binary. Consequently, the approximate noise penalty in dB of a b-level system relative to a binary system is $20 \log_{10} D_b/D_m = 20 \log_{10}(b-1)$.

The general form of a level-coded correlative signal

The common characteristic of existing multilevel codes is the absence of correlation between the levels. Lack of correlation implies that in the coding process at the transmitter every possible combination of a group of n binary digits is associated with one and only one particular level, regardless of the past history of the multilevel waveform; again, at the receiver this particular level is identified with the same specific group of binary digits. Attention has been directed to the possibility of utilizing discrete signaling levels that would be correlated in the process of generating such levels and yet be treated independently in the detection process.[8] Unlike the situation in multilevel systems, each level in a system with correlated levels represents only one binary digit: MARK or SPACE. Here correlation between the levels implies that, in the coding process at the transmitter, each MARK (or SPACE) is associated with one of several predetermined levels and the choice of a particular level depends upon the past history of the signal. However, at the receiver each level can still be uniquely associated with MARK or SPACE without examining the past history of the waveform. Owing to the coding process of the transmitter, the signal has inherent correlation properties at the receiver. These properties can be used to detect errors without the necessity of introducing redundant digits into the input binary data at the transmitter.

Generation of a signal with correlated levels permits overall spectrum shaping in addition to individual pulse shaping. It is, for example, possible to redistribute the spectral energy so as to concentrate most of it at low frequencies or, alternatively, to eliminate any energy at low frequencies. In certain instances, bandwidth compression can be applied to physical channels with a gradual cutoff as opposed to the strict Nyquist sense of nonphysical rectangular transmission channels. The fundamental characteristic of such techniques is that the level-coded signal states are correlated, and therefore they have been termed correlative.

Let a binary message[8] at point [A] of Fig. 4(A) with two signaling levels (MARK and SPACE) be transformed into a signal at point [C] with b signaling levels numbered consecutively from zero to $(b-1)$, starting at the bottom. All even-numbered levels are identified as SPACE, and all odd-numbered ones as MARK (or, equally well, the other way around). Both the original message and the level-coded signal have an identical digit duration of T seconds. The binary message is transformed into the level-coded correlative signal in two steps. In the first step, the original sequence at [A], consisting of independent MARKS and SPACES, is converted into another binary sequence in such a manner that at [B] a group of $(b-1)$ consecutive digits represents a MARK at [A] if it includes an odd number of binary 1's, otherwise the group represents a SPACE. The binary sequence at [B] has exactly the same bit speed as the sequence at [A]. However, its binary digits are correlated over a span of $(b-1)$ bits. Suppose $b = 5$ levels and the sequence at [A] is MMSMSS (where M and S stand for MARK and SPACE respectively); then a group of four bits represents each M or S. A possible sequence at [B] is 000101100. Here, for example, the first four bits (0001) represent M, the second through fifth (0010) represent M, the third through sixth (0101) represent S, and so on. Following a similar coding rule, an elementary coder with corresponding waveforms is shown in Fig. 4(B) for $b = 3$, with a correlation span of two digits. When $b = 2^k + 1$, with k an integer, a cascade of $(b-2)$ such elementary coders can provide the desired binary transformation.

The second transformation step in the block diagram of Fig. 4(A) involves the conversion of the binary sequence at [B] (in which 1 and 0 no longer represent MARK and SPACE) into the level-coded sequence with b levels. This conversion is accomplished by forming the digit sum of successive groups of $(b-1)$ consecutive digits of the sequence at [B]. Since only the binary 1's contribute to the digit sum, an odd-numbered level representing a MARK will result if the number of 1's in a group of $(b-1)$ digits is odd, and similarly for even-numbered levels representing SPACES. Using the previous example, 0001 and 0010 will result in level one, each representing a MARK, and 0101 in level two, representing a SPACE, etc. One result of the level conversion process is that SPACES and MARKS at [A] correspond uniquely to the even- and odd-numbered levels respectively at [C]. Therefore, in spite of the correlation properties, which span over $(b-1)$ digits, each level-coded digit of the waveform at [C] can be independently identified at the receiver as MARK or SPACE. The primary consequence of such properties is a redistribution of the spectral density of the original binary sequence at [A] into energy compressed near low frequencies for the new level-coded sequence at [C]. For example, when MARKS and SPACES are equally likely at [A] and represented by rectangular binary pulses of unit height in Fig. 4(A), the redistribution of energy at [C] corresponds to the scaling down of the spectral densities in the frequency domain by a factor $(b-1)$, where b represents the number of levels at [C]. This "scaling down" effect is reflected in the following two expressions, which give the spectral densities[8] at points [A] and [C] respectively:

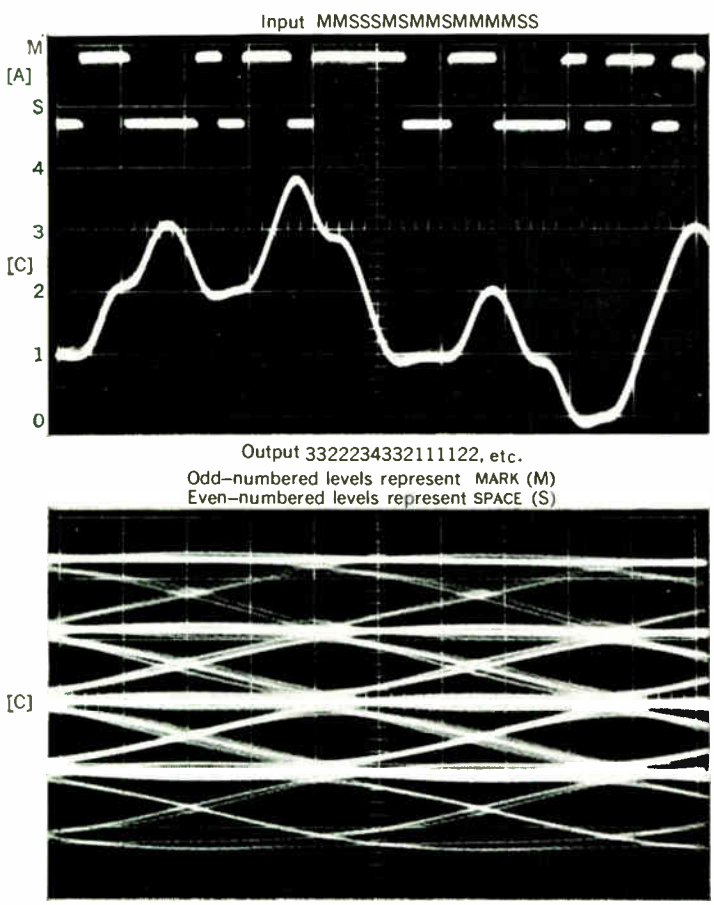
$$W_1(f) = \frac{T}{4}\left(\frac{\sin \pi f T}{\pi f T}\right)^2 \quad \text{(for binary)} \qquad (3)$$

$$W_2(f) = \frac{(b-1)^2 T}{4}\left[\frac{\sin(b-1)\pi fT}{(b-1)\pi fT}\right]^2 \quad \text{(for level-coded)} \tag{4}$$

The approximate noise penalty for the b-level system relative to binary is $20 \log_{10}(b-1)$ dB, as previously indicated. Such a system has been implemented for the case of $b = 5$, with the level conversion process accomplished by means of a low-pass filter to approximate the digit sum. Figure 5 shows experimental waveshapes, with letter designations corresponding to the points in Fig. 4(A). The intersymbol interference, particularly in terms of the horizontal eye pattern, is inherently small, since the only possible transitions in two successive digit-time slots in a level-coded system occur between the adjacent signaling levels.

The level-coded correlative technique described presumes equal weighting for each bit in forming the digit sum to generate the odd and even levels. A suggested variation in this procedure[9] leads to different forms of spectral reshaping of binary transmission if weighted coefficients are attached to each of the $(b-1)$ digits comprising the digit sum. In another proposed scheme,[10] the SPACE condition corresponds to the same level as the previous digit, whereas the MARK condition corresponds to the advancement of a level by a single step. A series of MARKS would cause steady advancement in one direction—up or down—until one of the extreme levels is reached, and then in the opposite direction; a single-step advancement always corresponds to a single MARK. In this process a one-to-one correspondence between any particular level and MARK or SPACE no longer exists.

Fig. 5. Typical level-coded correlative five-level experimental waveforms, at a speed of 4800 b/s. Top—Data pattern. Bottom—Eye pattern.



To provide a better insight into the characteristics and usefulness of the level-coded correlative techniques, the following few sections will concentrate on the various properties of specific codes, both with and without carrier modulation.

The baseband duobinary process

The duobinary process[11] is a level-coded correlative technique with three levels ($b = 3$), where "duo" indicates doubling of the bit capacity of a simple binary system. Its most significant property is that it affords a two-to-one bandwidth compression relative to binary signaling; or equivalently, for a fixed bandwidth, it has twice the speed capability in bits per second compared with a binary system. The same speed capability for a multilevel code would require four levels, each of which represents two binary digits. The approximate noise penalties of the four-level and three-level systems relative to a binary system are 9.5 dB and 6 dB, respectively, in accordance with Fig. 3.

The essence of the duobinary process is best understood by reference to Fig. 6. Suppose a low-pass filter having a bandwidth of $2f_1$ c/s is assumed, with impulsive response $h(t)$ sketched as either of the three pulses in Fig. 6. Binary signaling on an impulse or no-impulse basis to represent MARK and SPACE, respectively, is possible by sending impulses at intervals indicated by the time position of pulse number 3 relative to 1 and assuming for a moment that pulse 2 is absent. Such a rate would be $2f_1$ bits per second as in the case of the cosine-squared filter previously discussed. The intersymbol interference would be small, since the overshoots of $h(t)$ fall off as $1/t^2$. Suppose now the bit rate is doubled so that $4f_1$ bits per second are sent over the same filter with bandwidth $2f_1$. This process is depicted by adding pulse 2 halfway between pulses 1 and 3.

If we assume sampling instants at times -3, -1, $+1$, $+3$, etc., three distinct and equally spaced levels will result, rather than two as before. This may be observed, for instance, in the case of the specific instant of time $t = +1$. When neither pulse is present, the amplitude is zero; with only pulse 1 present, half of the amplitude is obtained as compared with the amplitude when both pulses 1 and 2 are present. Pulse 3 as well as any successive pulses are zero at $t = +1$. A similar argument applies to all instants of time represented by odd integers; that is, only two successive pulses and no other pulses can contribute to the formation of the three levels 0, 1, or 2. If we assume that the new three-level signal is sampled at these particular instants of time, no intersymbol interference will result and the vertical eye opening will be maximum. There will be some intersymbol interference at the instants of time represented by even integers. However, because of small overshoots of $h(t)$, the horizontal eye opening will be negligibly affected.

The three levels are numbered 0, 1, and 2, starting from the bottom. In accordance with the previous convention, the extreme or even levels (0 and 2) represent SPACE and the center or odd level represents MARK. The steady SPACE condition results when either no impulses or a steady

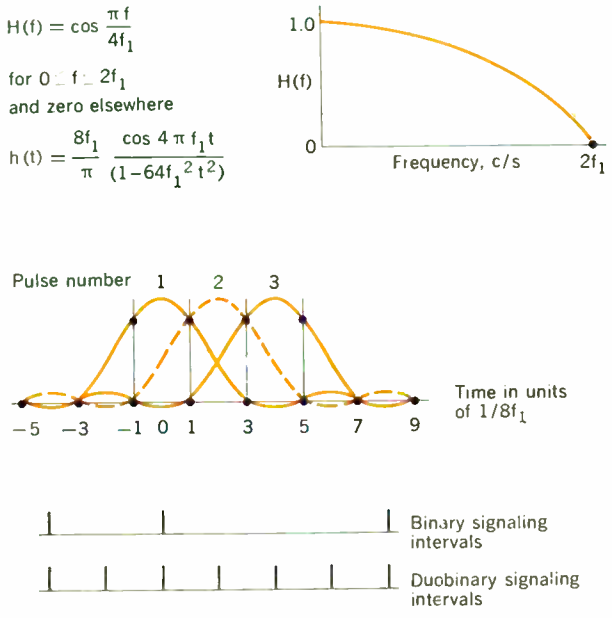


Fig. 6. Superposition of low-pass filter responses.

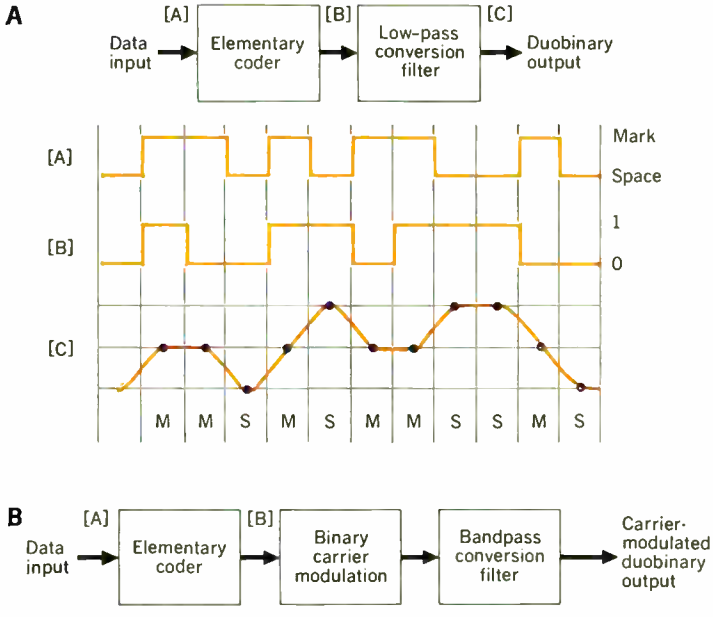
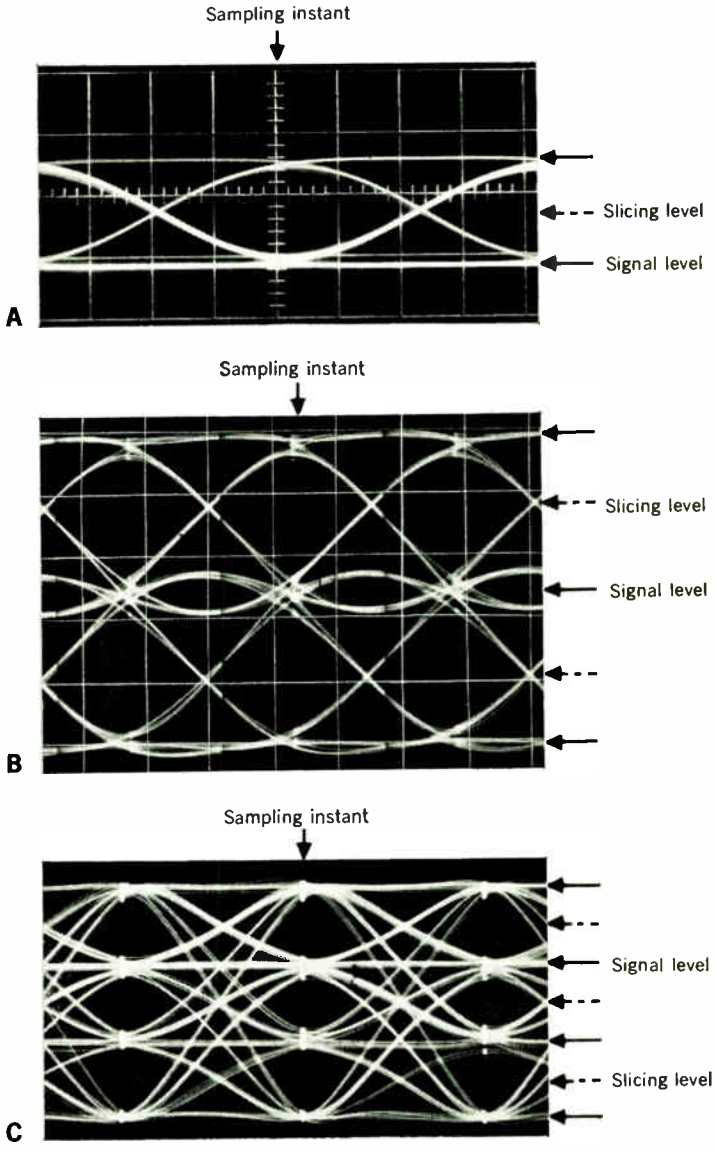


Fig. 7. The duobinary conversion process. A—Baseband process. B—Carrier process.

Fig. 8. Experimental eye patterns for (A) binary, (B) duobinary, and (C) multilevel systems.



stream of impulses at the rate of $4f_1$ b/s are sent; the latter implies that pulses 1, 2, 3, and all other pulses would be present. The steady MARK condition results when only every other impulse is sent, corresponding to pulse responses 1, 3, etc., and the absence of even-numbered pulses in Fig. 6. The net effect of inserting additional pulses, such as pulse 2 in Fig. 6 to double the bit rate, can be regarded as a deliberate introduction of intersymbol interference at times -3, -1, $+1$, $+3$, etc. This interference is quantized and interpreted in terms of the original data input.

The actual experimental duobinary process for a random input of MARKS and SPACES, along with the corresponding waveshapes, is shown in Fig. 7(A). In the baseband mode (no carrier modulation), the speed is $4f_1$ b/s, and a low-pass conversion filter is employed with cutoff frequency at $2f_1$ c/s. The same filter accommodates binary digits at the rate of $2f_1$ b/s. Waveform [C] indicates that in spite of the correlation properties, which span over each two successive digits, MARKS always appear at the center level and SPACES at the extreme levels, thus assuring that each digit can be independently detected at the receiver without resorting to the past history of the waveform. However, owing to these correlation properties, the waveform at the sampling points, indicated by the heavy dots at [C], follows a set of predetermined rules: Two successive SPACES (at the extreme levels) always have the same polarity if the number of intervening MARKS is even; otherwise their polarities are opposite. The result of the transformation from [A] to [C] is the redistribution of the spectral density of the original binary data into a highly concentrated energy density at low frequencies. Such a redistribution in the baseband process could also be accomplished in a strictly digital manner from [B] to [C] in Fig. 7(A). For example, the low-pass conversion filter in Fig. 7(A) could be replaced by a unit delay equal to one digit slot; next, the waveform [B] would be algebraically added to its delayed version, resulting in three

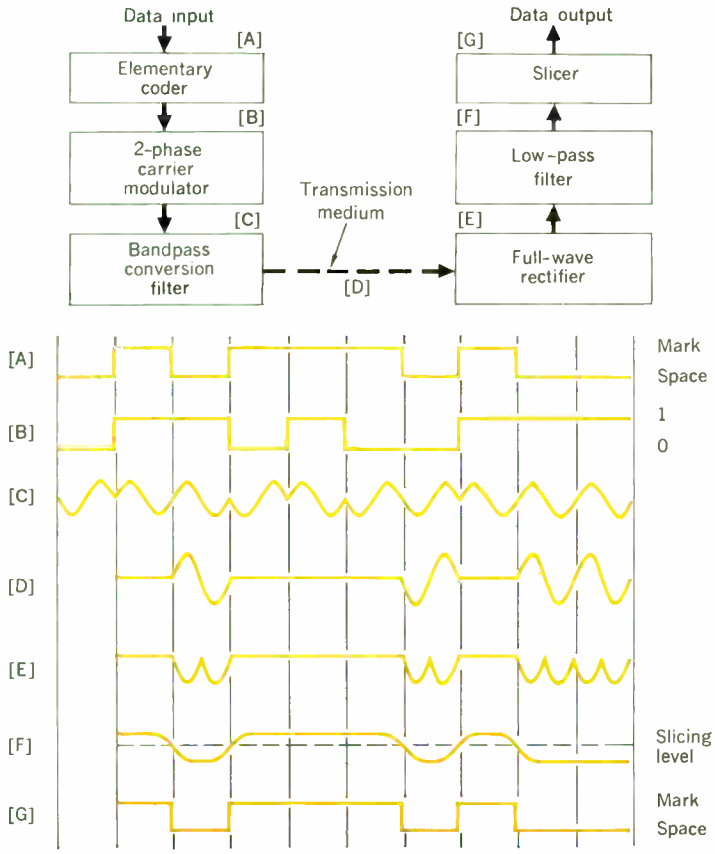
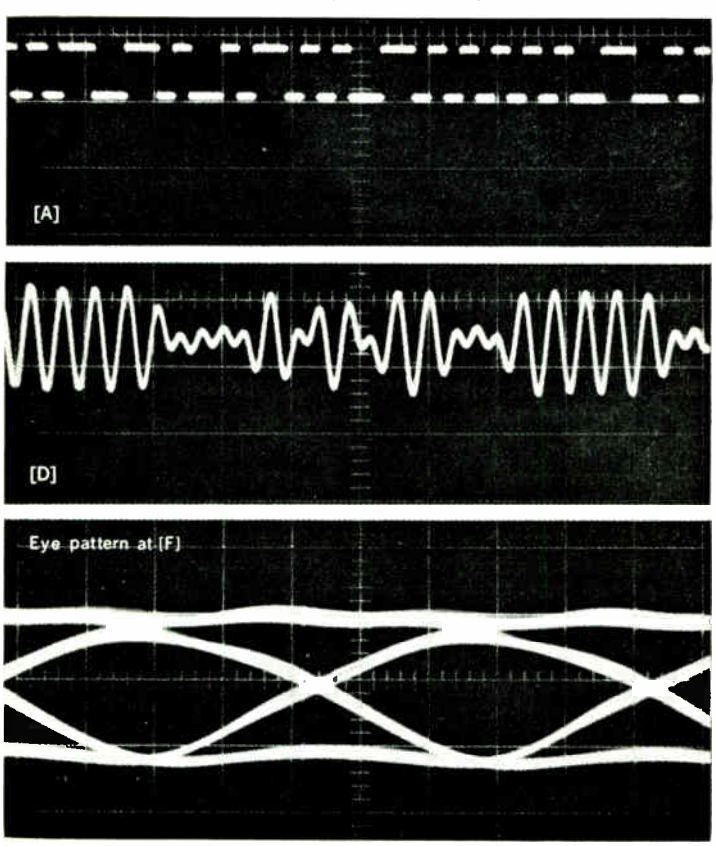


Fig. 9. Duobinary AM-PSK process, with envelope detection employed at the receiver.

Fig. 10. Experimental waveforms of the AM-PSK process at a speed of 2400 b/s.



levels at [C]. In this way the binary data are converted prior to carrier modulation and transmission. In both cases spectral reshaping takes place; however, only the analog conversion process can be arranged to constitute an integral part of carrier modulation and transmission. In the last case a strictly binary type of carrier modulation is carried out before bandpass filter conversion, as shown in the block diagram of Fig. 7(B). Here, a definite transmission benefit accrues in the sense that the data are compressed into a well-defined band narrower than the original binary pulse train. For the strictly digital method of conversion, this compression is not possible and merely spectral redistribution of energy takes place. The digital method of conversion into the level-coded correlative signal is always possible. This is not the case for the analog method, and therefore not all types of level-coding lend themselves to conversion after carrier modulation.

An interesting aspect of the duobinary process is demonstrated when the intersymbol interference criteria are considered in relation to binary and multilevel systems. All three systems were compared on the basis of identical bandwidths of $2f_i$ c/s in the absence of noise. The binary speed was $2f_i$ b/s and duobinary and multilevel (four-level in this case) speeds were $4f_i$ b/s. The corresponding eye patterns appear in Fig. 8. As expected, vertical and horizontal eye openings for the binary system are nearly perfect. Similar criteria indicate that duobinary has only a slight deterioration, primarily because the only transitions permitted are those between the adjacent levels. Finally, the vertical eye opening in the four-level pattern is reasonably good but the horizontal is considerably impaired. In this example binary transmission by the level-coded correlative technique (duobinary) is clearly superior to the four-level multilevel system, both from the point of view of noise penalty—due to an increased number of levels—and from the point of view of the intersymbol interference. The bit speed is identical in both cases.

Some interesting level-coded processes take place when the conversion process is integrated with carrier modulation and transmission. Two such processes are discussed in the following sections.

The three-level AM-PSK process with envelope detection

The three-level correlative baseband technique described in the previous section is suitable for any type of carrier modulation. However, a rather interesting signal characteristic results from the unique combination of this technique with AM-PSK modulation.[8,12] In AM-PSK modulation the carrier is amplitude modulated (AM) as well as phase modulated in a binary manner. Such a type of phase modulation is usually referred to as phase shift keying (PSK) to denote the discrete phase reversals of the carrier. Suppose the center MARK level of the duobinary signal is represented by the absence of carrier, the upper SPACE level by a constant-amplitude carrier, and the bottom SPACE level by the same carrier reversed by 180°. The key point is that this process is completely analogous to the baseband duobinary process described in the previous section, except for carrier modulation. There is a 180° reversal of the carrier if the number of intervening MARKS (in this case represented by the absence of carrier) is odd; otherwise there is no reversal. The very fact that the waveform has carrier phase reversals that follow pre-

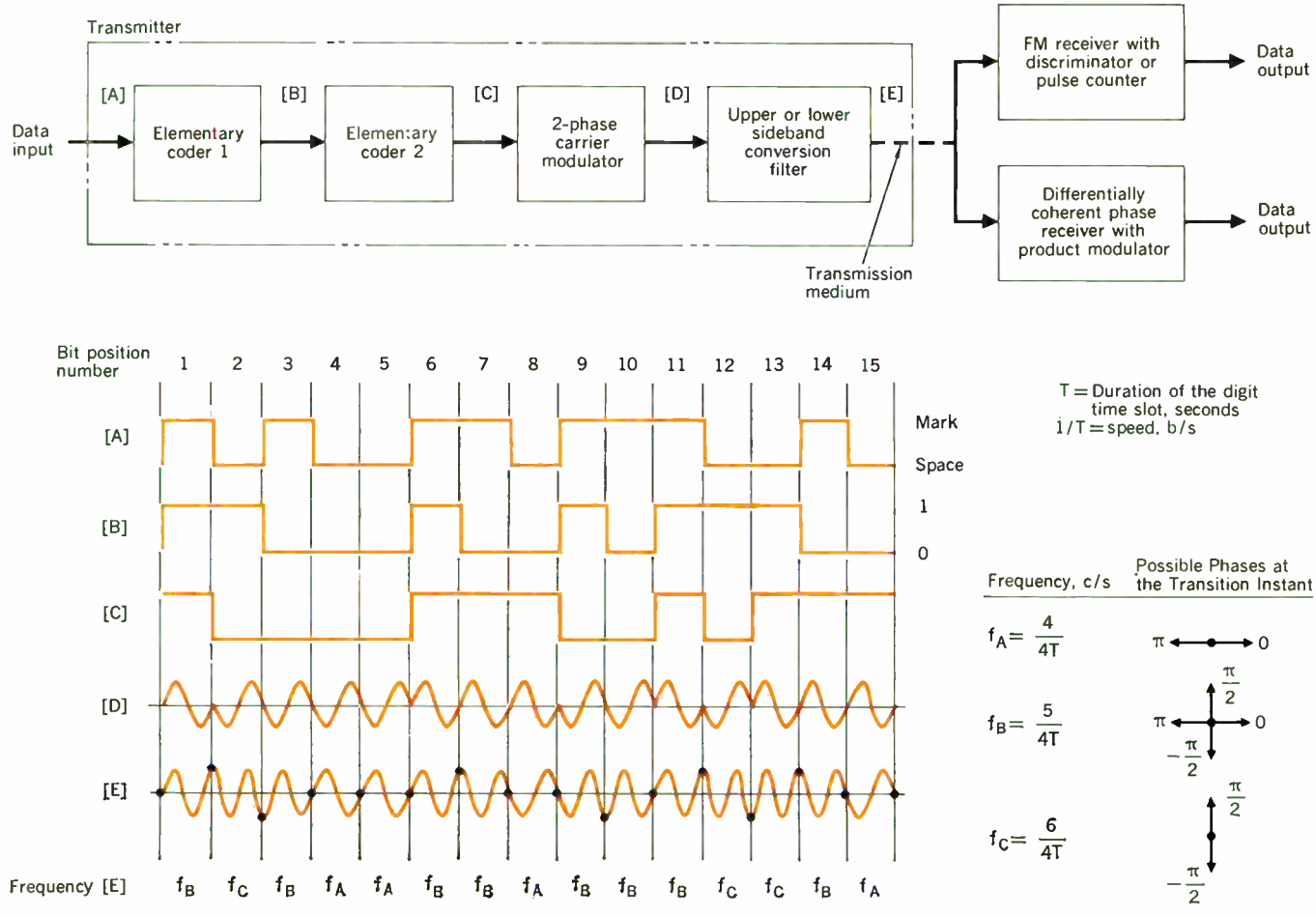
determined duobinary rules compresses the bandwidth (as in the baseband duobinary case) by a factor of two compared to a simple on–off AM, where no carrier reversals take place. In view of the fact that in the AM-PSK system the presence of the carrier in either phase represents a SPACE condition, there is no need at the receiver to distinguish between the phases to identify a SPACE. Consequently, the demodulator at the receiving end is arranged to disregard phase reversals and to detect only the envelope of the carrier. The AM-PSK duobinary system has two carrier amplitude states, and yet it requires only one half the bandwidth of a conventional on–off AM system. Conversely, for a fixed bandwidth, the AM-PSK duobinary system has twice the bit capacity of a conventional binary AM system. In fact, there is a 3-dB noise advantage over straight binary AM, since an AM-PSK duobinary system still has two amplitude levels but requires only half the bandwidth.

The entire AM-PSK process and the corresponding block diagram are shown in Fig. 9. The elementary coder again converts the original binary sequence at [A] into a binary sequence with correlation span of two digits at [B]. Carrier modulation with 180° phase reversals is accomplished in a strictly binary manner and follows the reversals of waveform [B]. Although any number of carrier cycles per bit are possible, for this particular case there is only one cycle per bit. The conversion process of waveform [C] is accomplished by a filter that is a bandpass version of the low-pass filter shown in Fig. 6, and centered at the carrier frequency. In a practical system the symmetrical bandpass filter has 3-dB points at $\pm f_i$ c/s, and 25 dB or more loss at and beyond $\pm 2f_i$ c/s from the carrier; the speed is $4f_i$ b/s. The net effect of the conversion filter is to add the waveforms at [C] in each two successive time slots, resulting in the on–off signal with phase reversals at [D]. The processing of such a signal at the receiver follows the conventional routine of envelope detection, as indicated in Fig. 9, and is self-explanatory; obviously, the phase reversals are completely ignored. The waveforms of the actual system, along with the binary eye pattern for a random-data input, are shown in Fig. 10. The letter designations correspond to those in the block diagram of Fig. 9. It should be emphasized again that inasmuch as the conversion process in the system described is accomplished after the carrier modulation, there is a two-to-one bandwidth compression relative to the conventional binary AM system.

The three-level signal with dual properties

Here we consider an unusual duobinary process again in conjunction with phase modulation of a carrier in which the resulting signal has dual properties such that the original information can be recovered either by noncoherent FM detection or by differentially coherent

Fig. 11. The three-level process with dual characteristics.



phase detection, depending upon the characteristics of the transmission medium. Being inherently duobinary, the signal again affords two-to-one bandwidth compression compared with binary systems.

The salient characteristic of this signal[13] is that the waveform has one of the three possible frequencies in each digit time slot as well as one of the four possible phases at the transition points between the digits. The three-level duobinary waveform can be extracted at the receiver either from the frequencies or from the phases at the digit transition points. The processing of the original binary message is shown in Fig. 11. It begins with digital coding using two elementary coders. The correlation property of the waveform at [C] is such that each bit depends not upon the previous digit, but upon the second digit back. For example, the first and third bits at [C] have an odd number of binary 1's and correspond to MARK in position 3 at [A]. Next, the second and fourth bits at [C] have an even number of 1's and correspond to SPACE in position 4 at [A], and so on. Carrier modulation with zero and 180° phases is governed by the reversals of the waveform at [C]. Again, the number of carrier cycles per digit slot is immaterial, but for simplicity only one carrier cycle per bit is shown at [D]. The resulting signal at [D] has two symmetrical sidebands. Finally, a bandpass conversion filter passes only one sideband—either the lower or the upper. The bit speed, the shape, and the bandwidth of the bandpass conversion filter are exactly the same as in the AM-PSK system discussed in the previous section, but the center frequency of the filter is no longer at the carrier frequency. In expressing the relationship between the carrier frequency and the center frequency of the filter, it is convenient to use the bit duration T seconds, where $1/T$ is the speed expressed in bits per second. The 3-dB and 25-dB points of the single-sideband filter are $1/2T$ c/s and $1/T$ c/s apart respectively. The position of the center frequency f_s of the bandpass filter is given by

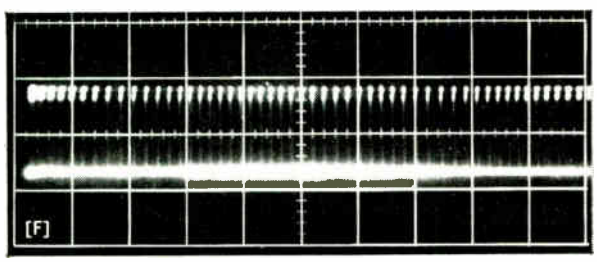
$$f_s = \frac{2k \pm 1}{4T} \text{ c/s} \tag{5}$$

where plus is for the upper and minus is for the lower sideband, and k is an integer > 2 that represents the number of half cycles of the carrier per bit. The carrier frequency is

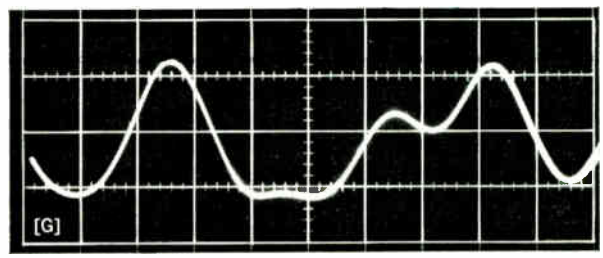
$$f_c = \frac{k}{2T} \text{ c/s} \tag{6}$$

Each time slot of the resulting upper or lower sideband signal at [E] in Fig. 11 contains one of three clearly distinguishable and equally spaced frequencies as well as one of the predetermined phases at the transition points between the time slots. The center frequency is the same as in (5) and the extreme frequencies are $(f_s \pm 1/4T)$ c/s. For the particular case shown in Fig. 11, $k = 2$ and the bandpass filter is upper sideband, corresponding to the plus sign in (5). The three equally spaced frequencies in Fig. 11 are expressed in terms of T, and their possible phases at the transition points are indicated. These three distinct frequencies have a one-to-one correspondence with the original binary data at [A]; that is, frequency f_B corresponds to MARK and f_A and f_C to SPACE. Because of the dual properties of this signal, the same MARK and SPACE information is also contained in the phases at the transition points between the time slots in differential form between each two successive digits. Consequently, the original message can be recovered at the receiving end

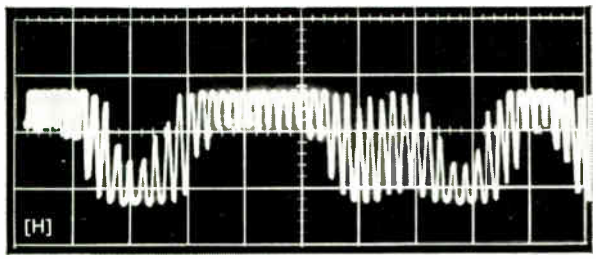
Fig. 12. Receiver waveforms for three-level process with dual characteristics. Speed: 2400 b/s. Line frequencies: 3600, 4200, and 4800 c/s. Fixed pattern: 100011101011.



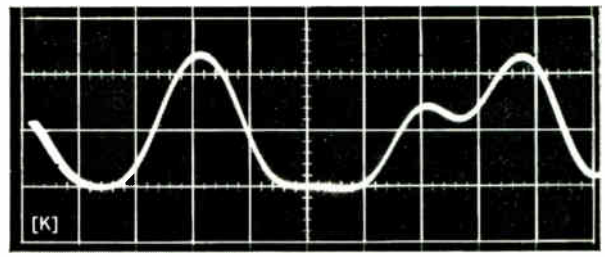
FM axis crossings



FM low-pass filter output



Differentially coherent product modulator output



Differentially coherent low-pass filter output

by two different methods, as depicted in the block diagram of Fig. 11.

Treating the waveform as an FM signal, any conventional noncoherent FM receiver with frequency or pulse-counting discriminator will convert the incoming carrier wave into the duobinary waveform. Because of its four-phase characteristic at the digit transition points, the same waveform can also be processed in a differentially coherent phase receiver. This involves the multiplication of the incoming signal at [E] by a replica of itself delayed by T seconds, which corresponds to a single-bit interval; the output of the product modulator through a low-pass filter yields exactly the same duobinary waveform as given by noncoherent FM detection. Both of these methods reconstruct the original data input; but whether one or the other, or both processes simultaneously, are employed will depend upon the characteristics of the transmission medium.

A striking example of the process depicted in Fig. 11 appears in experimental form in Fig. 12. Signal [F] represents the output of a pulse-counting FM discriminator in the form of narrow pulses occurring at the zero crossings of the incoming wave [E] in Fig. 11; signal [H] is the output of the product modulator in the differentially coherent phase receiver after the same wave is processed. These signals—[F] and [H]—hardly resemble each other, yet they contain the same information and, moreover, provide nearly indistinguishable output waveforms [G] and [K], after passing through identical low-pass filters. The waveforms [G] and [K] are, of course, duobinary, with the MARKS situated at the center and the SPACES at the extreme levels.

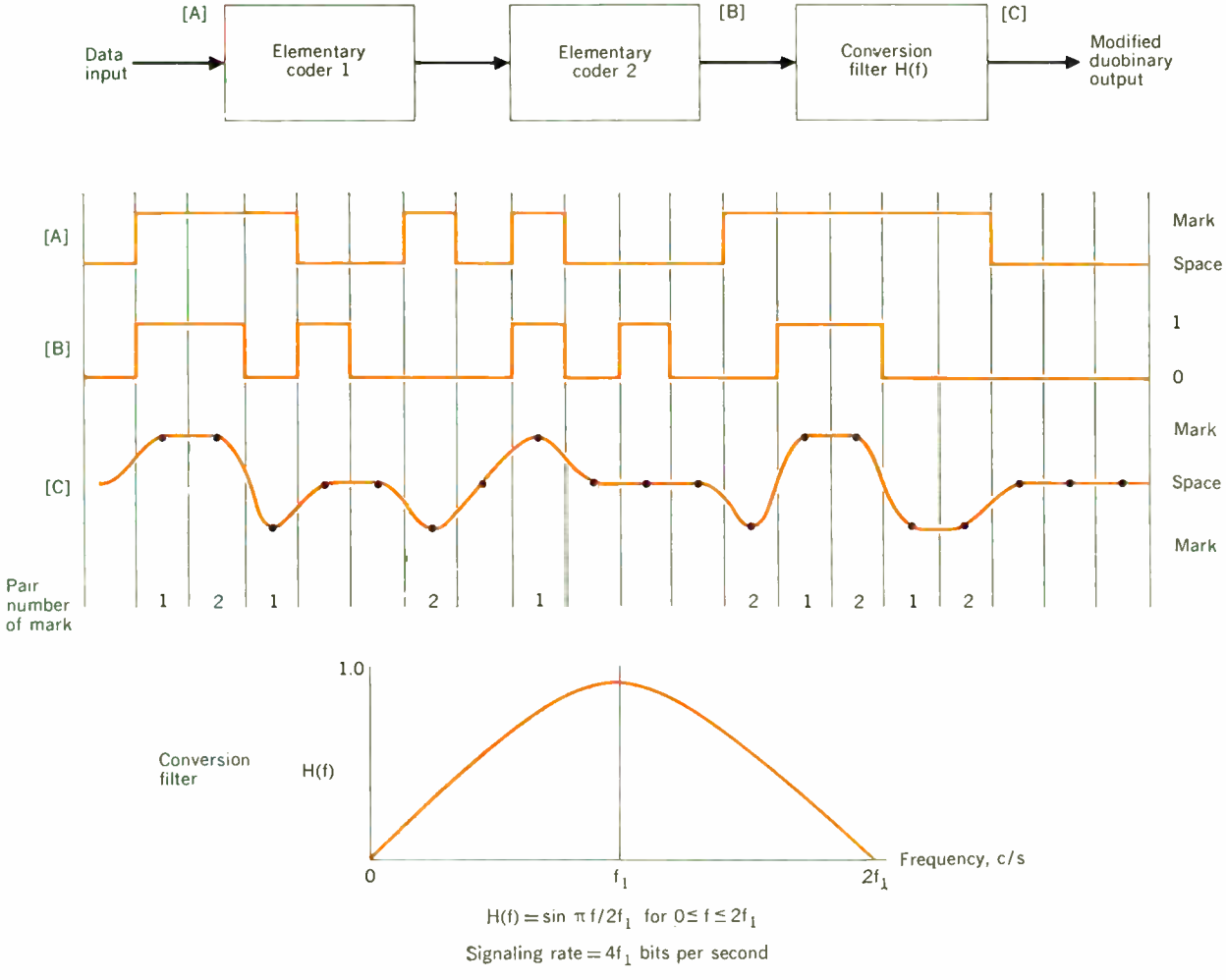
Three-level correlative process with a small low-frequency content

As stated before, level-coded correlative techniques permit redistribution of the energy of a binary pulse train consisting of MARKS and SPACES. In the previous examples, the spectrum shaping was such that most of the energy was concentrated at low frequencies—as, for example, in the duobinary process. The purpose of this section is to demonstrate a case in which the level-coded correlative method is employed to eliminate the zero-frequency component and leave only a small amount of energy at low frequencies.

A significant characteristic of this kind of signal is that, like the duobinary waveform, it still affords two-to-one bandwidth compression relative to the binary system and has only three levels. For that reason the technique has been termed "modified" duobinary.[14]

A block diagram, along with the corresponding waveshapes, is shown in Fig. 13. The encoding process in-

Fig. 13. Block diagram and waveshapes for modified duobinary technique.

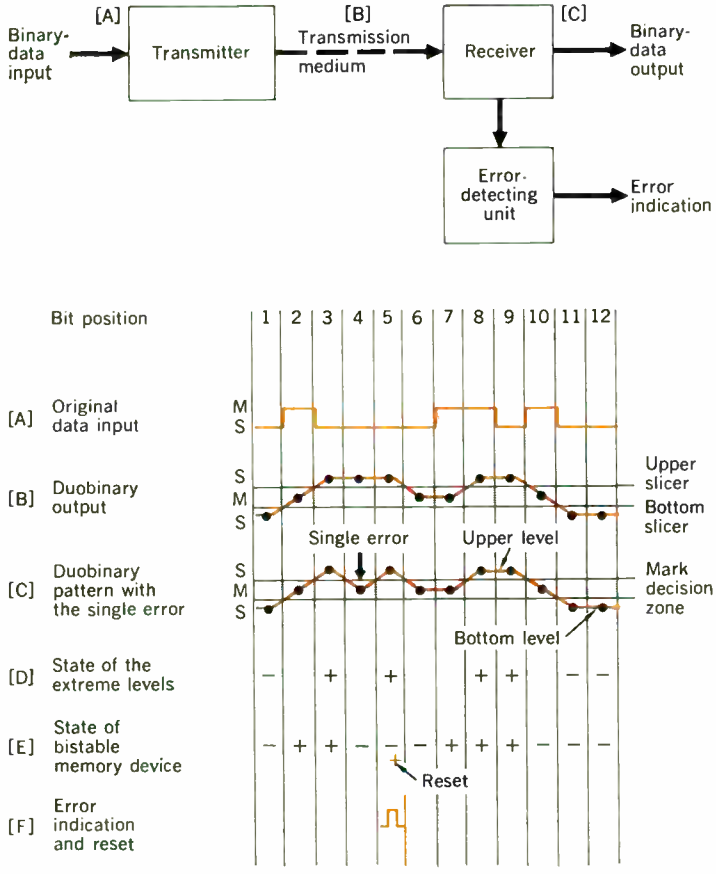


volves two elementary coders and is similar to that for the signal with dual characteristics previously discussed. Each bit in the waveform at [B] in Fig. 13 is correlated with the second bit back, rather than with the previous bit. The conversion process shown in the baseband form is carried out through a filter with bandwidth of $2f_i$ c/s, as shown in Fig. 13. The zero-frequency component is eliminated and the energy is centered at a frequency of f_i c/s. It is important to note that the signaling rate is $4f_i$ b/s, as compared with $2f_i$ b/s for a binary system having the same bandwidth. The resulting waveform at [C] has several interesting characteristics: There is a one-to-one correspondence of each sampling point (indicated by a heavy dot) with the original data at [A]; in addition, MARKS are always at the extreme levels and SPACES at the center level, so each digit can be identified independently at the receiving end. The modified duobinary signal has three levels, just as the duobinary, and therefore the approximate noise penalty relative to the binary system would appear to be the same as for duobinary. This point, however, must be qualified. As shown at [C] of Fig. 13, the modified duobinary signal permits all possible transitions between the levels; for example, from one extreme to the other. This causes irreducible intersymbol interference (just as in the multilevel case previously discussed), affects the horizontal eye opening, and somewhat reduces the margin to impairments as compared with the duobinary mode of transmission. The vertical eye opening is relatively little affected.

As would be expected, owing to its correlation properties, the modified duobinary signal follows a predetermined set of rules. These rules can easily be deduced by grouping all the successive MARKS in pairs and assigning the pair number to each MARK as shown below the waveform [C] in Fig. 13. A MARK bearing number 1 in a pair of two successive MARKS always has the opposite polarity relative to the previous MARK—which, of course, carries number 2. The polarity of the MARK that has number 2 relative to the previous MARK bearing number 1 is governed by the set of odd and even rules as in the duobinary case. Thus, if the number of intervening SPACES between a pair of MARKS numbered 1 and 2 is even, their polarities are identical; if not, their polarities are opposite. This statement is easily verified for the waveform [C] in Fig. 13.

Such rules—or more generally, the correlation properties of the level-coded correlative signals—permit the detection of errors. Error detection should not be confused with correction of errors. Error detection implies an indication (a pulse, for example) at the receiver that an error or errors occurred; however, there is no identification of the time location of the error. An error occurs when MARK is changed to SPACE, or vice versa, due to noise or any other transmission impairments. If the time location of the error were known, then error correction could be accomplished, since only binary values (1 or 0) are involved. Conventional systems employ redundant binary digits inserted into the binary data streams at the input to the transmitter shown in Fig. 1 to detect errors at the receiver. One of the key points of the level-coded techniques is that such redundant digits are not necessary, yet most of the errors are detected. Such a system can be employed for retransmitting erroneous data blocks, monitoring the state of the transmission medium, or merely discarding erroneous messages.

Fig. 14. Error detection without redundant digits.



Error detection in level-coded systems without redundant digits

As described earlier, a b-level coded time function has correlation properties extending over $(b - 1)$ digits. These properties are not utilized in the signal detection but can be employed to check pattern violations without introducing redundant digits into the original binary message. Such violations may result in errors, which are easily detected. Although error detection without redundant digits is possible in level-coded correlative systems with any number of levels,[8] the principle is best illustrated in the case of a three-level duobinary system. Here advantage is taken of the predetermined rules, namely that two successive SPACES (represented by either of the extreme levels) have the same polarities if the number of intervening MARKS is even; otherwise their polarities are opposite. The essence of the error-detecting process is a two-state memory device, which counts only the intervening MARKS and predicts the polarity of each SPACE following the last intervening MARK.[15] Whenever there is disagreement between the prediction and the actual polarity of the SPACE, an error indication is given. At the same instant of time the binary memory device is reset to its correct state so that the detection process can start all over again.

Figure 14 shows the original waveform composed of MARKS and SPACES at [A] and its corresponding duobinary signal at [B]. Heavy dots indicate the sampling instants, and the MARK decision zone is represented by the shaded

area. Should the duobinary waveform be in this zone at the instant of sampling, a MARK decision will be made; otherwise, there will be a SPACE decision. Next, the same waveform is shown at the input to the receiver at point [C], with a single error in the bit position 4. Since the waveform is in the MARK decision zone at the sampling instant, the receiver output will be MARK, whereas the original data input at [A] was SPACE in the same bit position. This error will be detected as soon as one of the extreme levels is reached and a comparison of the wave and the binary memory device is made.

The polarities of the extreme levels are indicated in [D]. The bistable memory device at the receiver, shown at [E], switches its polarity every time a MARK (or the center level) is present at [C]. The starting point is the bit position 1, where an agreement is assumed between [D] and [E]. Since the bit position 2 at [C] is MARK, the memory device at [E] changes state to positive and predicts that, if the next digit in position 3 is a SPACE, it should have a positive polarity at [C]. Since this is true, there is an agreement between [D] and [E] in bit position 3 and nothing happens. Next, the waveform [C] indicates a MARK in bit position 4, so there is a change of state at [E]. The implication is that a potential SPACE in bit position 5 should be negative. Here there is a disagreement between [D] and [E], and an error indication at [F] appears in the same time slot. At the same time, the memory device at [E] is reset to its correct state so that the process may start all over again.

It should be noted that the actual error occurred in the time slot 4, whereas the error indication was given later. Thus, although the error was detected, it cannot be corrected, since its time location is not identified. Any errors that violate the odd and even rules will be detected. For example, all bursts of odd numbers of errors in the MARK condition will result in an even number of MARKS between SPACES of opposite polarity, or in an odd number of MARKS between SPACES of the same polarity. If this should happen, the predetermined rules would be violated and the errors would be detected. Finally, the effectiveness of error detection increases with the correlation span of the level-coded digits.

Conclusions

The fundamental property of level-coded correlative techniques is the correlation between the digits introduced in the process of coding. Signals of this type permit spectral reshaping of the energy of the original binary data and are suitable for transmission over channels with gradual cutoff characteristics. Such waveforms can easily be generated with or without zero and low-frequency components. In some instances it is possible to achieve more efficient bandwidth compression than with the conventional multilevel techniques. For example, the three-level codes presented here have the same speed capability as four-level conventional multilevel codes and, at the same time, much less intersymbol interference, thus allowing greater margin to noise and transmission impairments.

Combinations with binary carrier modulation result in signals with remarkable properties. The AM-PSK system described has twice the speed capability of the conventional binary on–off AM system without incurring any noise penalty. In another mode of operation a signal with dual characteristics results, permitting recovery either by an FM noncoherent or by a differentially coherent phase process, the choice being decided by the properties of the communication medium. Such a signal also affords a two-to-one bandwidth compression relative to binary systems.

The correlation properties of the level-coded waves, representing the binary message, can be used to detect errors without introducing redundant digits into the original data stream. Finally, the equipment implementation of the techniques presented is relatively straightforward and simple, comparable in complexity to the ordinary binary systems. As a matter of fact, many systems of the type described in this article are currently in operation over voice and broadband channels—such as cable, carrier (including microwave), and high-frequency radio.

The signal design of level-coded correlative codes represents virtually a virgin field in digital communications. Of particular interest for future work are new and unconventional treatments of carrier modulation integrated with level coding to arrive at high speeds with less noise penalty than presently possible. Some of this work is now being carried out and will be published shortly. Finally, it should be pointed out that as the number of levels is increased in binary level-coded correlative techniques, intersymbol interference increases, as one would expect, and thus some practical difficulties are presented. It may, therefore, be potentially profitable to investigate other related areas, such as nonbinary level-coded correlative techniques.

REFERENCES

1. Nyquist, H., "Certain factors affecting telegraph speed," *Bell System Tech. J.*, vol. 3, pp. 324–346, Apr. 1924.

2. Groff, W. M., and Powers, R. C., "A high-speed synchronous digital data transmission modem," *Proc. 1961 Nat'l Conv. on Military Electronics*, pp. 208–213.

3. Lebow, I. L., *et al.*, "Application of sequential decoding to high-rate data communication on a telephone line," *IEEE Trans. on Information Theory*, vol. IT-9, pp. 124–126, Apr. 1963.

4. Critchlow, D. L., *et al.*, "A vestigial-sideband, phase-reversal data transmission system," *IBM J. Res. Develop.*, vol. 8, pp. 33–42, Jan. 1964.

5. Schreiner, K. E., *et al.*, "Automatic distortion correction for efficient pulse transmission," *IBM J. Res. Develop.*, vol. 9, pp. 20–30, Jan. 1965.

6. Becker, F. K., "An exploratory multilevel vestigial-sideband data terminal for use on high-grade voice facilities," *Conf. Rec. 1965 IEEE Ann. Communications Conv.*, pp. 481–484.

7. Nyquist, H., "Certain topics in telegraph transmission theory," *AIEE Trans.*, vol. 47, pp. 617–644, Apr. 1928.

8. Lender, A., "Correlative digital communication techniques," *IEEE Trans. on Communication Technology*, vol. COM-12, pp. 128–135, Dec. 1964.

9. Kretzmer, E. R., "Binary data communication by partial response transmission," *Conf. Rec. 1965 IEEE Ann. Communications Conv.*, pp. 451–455.

10. Shagena, J. L., and Kvarda, J. C., "A new multilevel coding technique for digital communications," *Proc. 1964 Internat'l Conv. on Military Electronics*, pp. 326–331.

11. Lender, A., "The duobinary technique for high-speed data transmission," *IEEE Trans. on Communication and Electronics*, vol. 82, pp. 214–218, May 1963.

12. Kretzmer, E. R., "An efficient binary data transmission system," *IEEE Trans. on Communications Systems*, vol. CS-12, pp. 250–251, June 1964.

13. Lender, A., "A synchronous signal with dual properties for digital communications," *IEEE Trans. on Communication Technology*, vol. COM-13, pp. 202–208, June 1965.

14. Lender, A., Unpublished report on modified duobinary technique, Lenkurt Electric Co., Inc., San Carlos, Calif., 1963.

15. Lender, A., "Faster digital communications with duobinary techniques," *Electronics*, pp. 61–65, Mar. 22, 1963.

Search methods used with transistor patent applications

Today's ever-expanding technology means an ever-increasing number of patent applications, which complicates the patent examiner's life. The effectiveness of two of his tools—manual and mechanized search methods—are compared as used in the transistor art

June Roberts Cornog *National Bureau of Standards*

Herbert L. Bryan, Jr. *U.S. Patent Office*

As the world has become more technically oriented, the number of patent applications has been increasing also—too rapidly for the Patent Office to assimilate them comfortably with current techniques. When an application for a patent is received, it must be evaluated as to novelty by a specialist who searches the prior art for similar patents. Previously, all such searches were done manually, which meant that the examiner had to rely entirely on his knowledge and experience. In an effort to remedy the situation the Office has instituted mechanized search methods. In order to ascertain the differences in patterns of thinking associated with manual and mechanized searches, a study was carried out in which a patent application in the transistor art was searched both manually and by a mechanized method. The mechanized search in this case permitted more patents to be analyzed more quickly but, being completely literal, it does not allow for hunches or browsing.

During the last decade, the United States Patent Office has been finding itself in an increasingly difficult position. The rapidly expanding world technology has produced inventive ideas at an uncomfortable rate. The number of applications for patents, as well as of patents granted, is continuously rising; to date some 3 000 000 patents have been issued, with an increase of 60 000 every year.

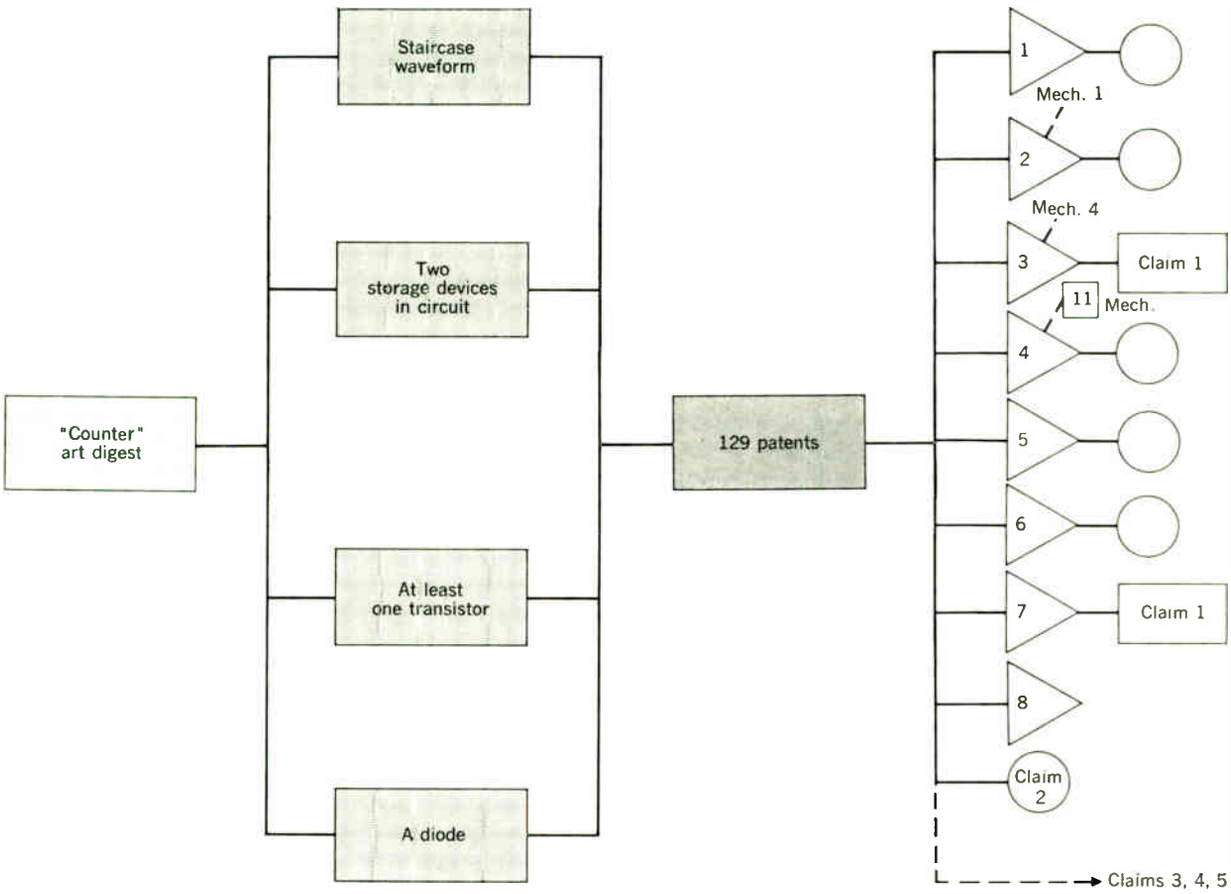
Classification of patents into categories so they may be filed, and located, by subject matter is more difficult than it would seem at first glance—inventions are inventions because they are new, and the fitting of new ideas or concepts into old classifications often means that the categories must be stretched to include unlikely patents.

When an inventor or designer has what he believes to be a patentable idea, he and his patent attorney draft an application for a patent on the invention. The application includes a specification, which usually consists of a general background and description of the inventive idea and appropriate examples of its use; a statement of claims over areas covered by the idea; and drawings, if necessary. The application for a patent is then submitted to the Patent Office where an examiner, who is a specialist in the particular art, reads and evaluates it. The examiner then searches the prior art—previously issued U.S. and foreign patents, scientific literature, etc.—to determine whether the claimed invention is novel and not obvious.

The examiner and the patent system are governed by a code of federal laws. The reason for the examiner's search is that patent statutes place limitations on what can be patented; the idea must be truly inventive. With regard to the novelty of the invention, the search presents a "yes–no" type of question, although the decision still must be made by the examiner. It is in determining whether a thing is "obvious to one skilled in the art" that complex patterns of thinking are required.

To conclude whether or not an idea is obvious, the examiner must use logical reasoning. He not only must be convinced that his judgment is correct as to whether an idea is or is not inventive, but he must prepare a statement of his decision so that he can defend it to the applicant and, occasionally, to the Patent Office Board of Appeals. He depends on the reference he finds in his search to establish the premises on which he will base his reasoning. The decision as to the obviousness of an inventive idea is considered by examiners generally to be the most delicate and professional task they perform. It requires the most sensitive handling, the finest analytical judgment, and the best possible supportive documents.

Until approximately eight years ago, all applications for patents were accorded only a manual search of the prior art to determine whether the proposed idea was truly inventive. In making such a search, the examiner first decided as to the classifications in which the general idea was likely to be filed, then found the "shoes" or file drawers for those subclasses and scanned the contents. Various short-cut helps had been devised through the years—previous searchers had penciled notes on the old patents; "briefs" or abstracts were written for patents in some files—but in the main the examiner was, and is, expected to scan the prior art in one or more categories, rapidly determine any relevance to the principle in the application, and ultimately decide upon the inventiveness of the applicant's idea. If the prior art is extensive, or if the new idea is classifiable in several different categories, then the examiner will be ahead if he is familiar enough with the art to know where the most relevant work is and what the principal disclosures in the area have been. The brain of the experienced examiner has functioned as an in-



Strategy used in manual search. Characteristics of the claimed invention used as basis for search are shown in rectangular boxes at left. Of the 129 patents in the "shoes," eight were read carefully. Numbers 3 and 7 were used to reject claim 1. Claim 2 was allowed and claims 3, 4, and 5 were said to be searchable only in other areas. Numbers 2, 3, and 4 read were the same as patents 1, 4, and 12 in the mechanized search. Mechanized no. 12, although used as a reference in that search, was passed over here as not applicable.

formation storage system for the 175 years of Patent Office existence, but the amount to be stored now doubles every few years.

The Patent Office has been attempting to meet the technical information explosion by instituting mechanical search of the prior art. The enormous complexity of this problem will not be described here except to point out what may not be obvious to everyone; that is, because of human and machine incapacity to store information from whole disciplines at once, it has been necessary to set up numerous small systems, each limited to a particular segment of the art, with attendant danger of incompatibility among them. Several good small systems have been designed, however, and one of them, the "peek-a-boo" method for the transistor art, is analyzed here.

In the transistor art of the U.S. Patent Office, the information from a somewhat limited group of patents is stored in a punched-card deck, which is searched by the peek-a-boo method. Because only U.S. patent information is filed in the deck, examiners regularly search the prior U.S. art by the punched-card method but use the manual system for foreign references and other pertinent material. The regular practice of both skills renders these examiners almost unique in the examining corps.

Problem and procedure

In the problem considered in this article, the Patent Office asked for an unbiased, objective study of the differences in patterns of thinking associated with manual and mechanized searches during an examination of the prior art. The complete processing or disposal of an application was not included in the assignment.

Designers of mechanical information search systems should consider two guiding philosophies:

1. The system should offer the user a "natural" method of querying the machine.
2. The human nervous system should be used as a model as far as possible since it appears to be the most efficient information retrieval system yet discovered.

One complaint made by examiners about mechanized retrieval in general is that it forces the searcher into such a stylized way of thinking that he feels restrained by the system, or at least unable to explore and develop new approaches to solving his problem. It is, at the least, different from the manual system he ordinarily uses to find information. The very presence of a mechanical contrivance between him and the thing he wishes to find seems in many instances to give him pause about becoming involved in extensive or repeated inquiries.

In general, the transistor punched-card file can be

through M (in squares) were used in various combinations to obtain references 1–24 (triangles). References 10 and 12 were used to meet claims 3, 4, and 5 (heavy rectangles), but most of the referred patents read were not used (circles). Search of claims 1 and 2 did not yield any applicable references.

End search of claims 1 and 2

described as a classification of the information by the terms used in the prior art and by designators of the structures in those patents—descriptors of the configurations of circuitry, types of units such as oscillators or control units. Just one term or structure designator is put on a card. Holes are then punched for the numbers of all

searches of the same application were made at least two months apart to allow time for the examiner to forget exactly what he had found in the previous search. The area of search was limited, furthermore, to U.S. patents.

The mechanized search

art; claim 2, allowable; claims 3, 4, and 5, should be searched further in subclasses other than those corresponding to the mechanized file; claim 6, statutory rejection. One of the two patents used as references in the manual search had been examined in the mechanized search and discarded, but of those used as references in mechanized search, one was read and discarded and one was not found in the manual search.

Findings of fact and inferences

The inferences stated here should be regarded only as hypotheses upon which to base more quantified research. The sample used in the investigation was inadequate to provide a basis for generalization to a larger population without quantitative verification.

1. *The mechanized search was more thoroughly planned than the manual before search activities actually began.*

As he read and analyzed the specification and claims in the mechanized search, the examiner stated that he was fitting the material into the scope of the coding terms used in the card deck. He first selected 11 of these standard terms he thought would describe the invention. With the second and third trials he added two more, for a total of 13. He used an average of 4.1 descriptors per trial after the initial inquiry.

With the manual search his plan was somewhat more flexible. He said, "Really the only search we have for this [in this limited file] is in the counter art." He intended to, and did, look for general structures, but his chief cue was the distinctive staircase waveform on patent drawings. He was also aware that one "can't always depend on the patentee's including a staircase waveform or including the term in their title, so I have to check all of them."

In both manual and mechanized searches he looked for key features. In manual searches these features were classified by his own experience, by his training and interpretation of the inventive principles involved. In the mechanized searches he was required to put his interpretations into the terms or classifications devised by the designers of the system, but in the peek-a-boo card deck the original selection of descriptors is perhaps less of a commitment than is the first classification for the manual search. In the latter instance the examiner stays within one subclass until he definitely decides it is unfruitful.

In the mechanized search of the application, he changed his combination of descriptive terms nine times. Every time he added or dropped a term card, he was changing slightly the concept of what he was looking for. It was in trial 5 that he found the first references he could use; in fact, these were the only applicable ones he located in the entire mechanized search procedure.

Change of direction of thinking occurred only twice in the manual search, but the examiner admitted that he felt artificially constrained by the fact that only one subclass of pertinent art was available to him within the limitations placed on this particular study. He said he would have used other subclasses had they been available; thus, under normal circumstances he would have had at least three changes of direction with the manual search.

2. *The basic conceptions of manual and machine search are somewhat different.*

In the manual system patent information is organized and classified on the basis of broad general concepts. Certain principles of invention are common to the patents that are filed together. Actual terminology used in a patent, or specific details of processes listed in the claims, must be translated into one of the concepts included in the classification plan. When the examiner enters the manual classification system, he must first classify the information in the application correctly and then locate the most nearly similar existing file.

The mechanized systems, particularly the peek-a-boo system, are organized to make information retrievable on the basis of either broad concepts or terms for details of structure or function. The whole system is available at all times and inquiry can be pressed in several directions simultaneously. The examiner need only be concerned with formulating a good but broad description of the claimed invention. He does not have to decide how the concept would be classified in the manual system, or physically have to locate the subfile.

Essentially, then, three factors determine the amount of effort required to prepare any search strategy: (1) the complexity of the system to be entered and its demands for specific preparation of the queries; (2) the user's familiarity with the logic of the system; and (3) the flexibility permitted the user by the design of the system.

3. *The examiner's approach to the searching of claims and to the acceptance of a reference as possibly pertinent seemed to be different in the manual and machine searches.*

With mechanized search the examiner started by looking for a "pat" reference, one that would "fully anticipate" all the claims in the application. When no such reference was forthcoming, he then began to select term cards according to the content of groups of claims. He grouped claims 1 and 2 together and 3, 4, and 5 together as being the most nearly similar. After locating the two references he ultimately used to reject claims 3, 4, and 5, he reanalyzed claims 1 and 2 but added no new term cards to those previously in use. He only tried new combinations of the old terms.

Claims 1 and 2 received priority in the manual search. Claim 3 was described as being the "broadest" and claims 4 and 5 were said to be "dependent on 3." After finding prior art which met claims 1 and 2 as satisfactorily as any that he thought, from his knowledge of the prior art, he would be likely to locate, he again checked on claims 3, 4, and 5, and said, "It practically will be impossible to find a feature like this in our digest because there is no one place to look for it."

In the mechanized search the examiner seemed to check each patent dropped out by the system with the thought, "Is this the structure I asked for or isn't it?" If it was not, he discarded it. He evidently did not expect the system to point out any equivalent concepts unless asked to do so.

"Give me something to start with," was his approach to the manual system. He treated each of the eight patents read as a definite possibility, and for the three that were "pulled" he set about analyzing each with a view to using any information that might be applicable. He seemed to be trying to stretch or twist, to match or substitute, to accommodate possibly equivalent structures or procedures. Equivalency or combination rejections seem to be more characteristic of manual searches.

It becomes evident in the restatement of the general search strategies that the order in which the patents were presented for inspection may have influenced the ultimate choice of those considered to be most nearly pertinent. One patent in the manual search (the fourth one read) was passed by as having no pertinence; the same patent

(the tenth one read) in the mechanized search was accepted as a good reference. The third one read in both searches was considered inapplicable in the mechanized search but was accepted as a very good reference in the manual search. Some influence from the order of presentation seemed to be functioning.

Although the examiner did not find the same prior art in the manual and mechanized searches, he thought he would allow claim 2 in both of them. When he had completed his search of the mechanized deck, he had no plans to search further in the files for other arts. He apparently felt he had located all patents that were likely to be pertinent. With the manual search, he indicated strongly at the end that, if he were free to do so, he would go to other subclasses. He obviously did not feel he had covered the field well enough. The finding of pertinent references by the mechanized search evidently gave him a more positive feeling of total coverage, of finality.

4. *In the manual search the examiner actually analyzed quickly and examined carefully fewer patents.*

The following will give the comparison at a glance:

	Manual	Mechanized
Number of patents possibly applicable	129	112
Number given quick analysis	8	24
Number examined	3	8
Number selected	2	2

The examiner realized that the patents turned up by the mechanized system had been retrieved because a coder had found something in them that was in some way related to his search. Because of the closer inspection of several patents in the mechanical search which he had merely leafed through in the manual search, he found material that enabled him to reject claims 3, 4, and 5. The greater attention to detail possible with mechanized search, however, did not enable him to locate patents that would reject claim 1 in the same body of art.

5. *The times at which high examiner mental activity occurred were different in the manual and machine searches.*

In the mechanized search the examiner showed the greatest concentration when he was searching his experience for possible combinations of terms that would best describe the structures he was looking for. He was, in other words, trying to find patents to look at, to guess what standard terms the coders filed the data under.

The decisions about characteristics of the structures and the classification in which to look for them took little time in the manual search. Once past this point, the examiner's chief area of concentration was on quick analyses of the art he looked through. As he glanced at each of the 129 patents in the file, he had to make a quick analysis and decide whether or not to give it an intensive examination. With the manual search he made 129 decisions about individual patents; with the mechanized, 24.

6. *In psychological terms, the "feeling tones" of the protocols for mechanized and manual searches differed.*

Both searches were tape recorded and played back later in close succession, so the difference in treatment was evident. The mechanized search produced an efficient tone, the kind that is characteristic of a person doing a well-mastered, routine job. The work was there to be done; the examiner would do it quickly and get on. The tone used in the manual search showed a higher level of interest, and later disappointment at not being able to search in other subclasses in order to finish up the job.

With the mechanized search some feeling of defensiveness was apparent, as though he must defend his interpretation and use of the system against critics. He evidently felt he was being required to employ methods devised by others, ones with which he did not entirely agree.

7. *The mechanized system does not have "hunches."*

The peek-a-boo card deck knows no short cuts. It turns up prior art only in answer to exact questions.

The examiner himself must be the source of the hunches; in the search used as an example in this article, he followed hunches nine times. Every time he tried a combination of cards and did not get a usable reference, he restructured his query and tried again. He had, of course, the benefit of his knowledge and experience to help him formulate these educated guesses as to combinations of descriptive terms that might bring forth what he was looking for. Fortunately, the peek-a-boo system makes the implementation of hunches very easy. Hunches in manual search usually mean that he must go to a different subclass, or to a different art, to continue his search.

8. *The peek-a-boo search did not appear to encourage browsing.*

Browsing—reading through adjacent documents with the hope of finding something the reader had not thought to look for—occurs with the term list (the "dictionary") in the peek-a-boo system rather than with actual patents. Documents that are "dropped" by the system but are not good references often bear little relationship to the structure under examination. Because the system was generated by assigning one term to a card, with all prior art documents employing that term or structure punched in the appropriate positions on the card, the deck drops out the numbers of patents that may use an essential descriptor in an entirely different kind of circuit than that being sought. As one of the chemical examiners pointed out: "You can't browse through 'no drops' or false drops. [The mechanical system either did not locate any references or else listed irrelevant ones.] Sometimes one of them will give you a starting point but it isn't often. The false drops are much farther from the central idea of the search than are irrelevant patents in the same subclass."

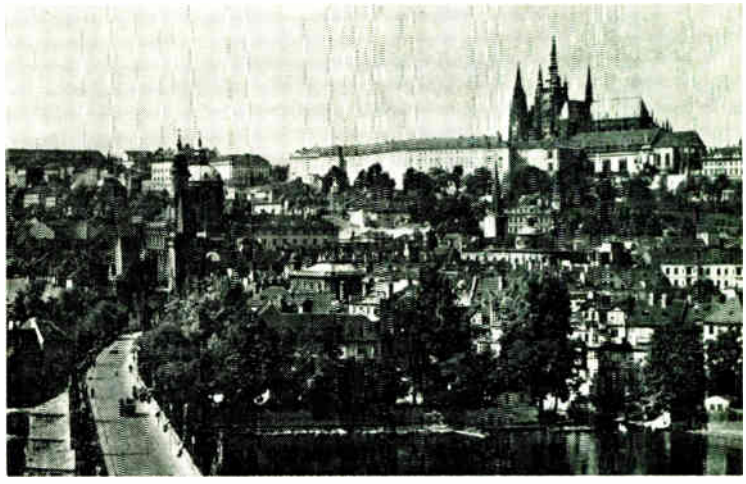
With the mechanized systems tested, in order to browse, either the questions asked must be broadly stated, so as to cover related areas, or several times as many narrowly based inquiries must be successively addressed to the system so that the examiner can probe areas as he thinks of them. With these mechanized systems an examiner has to *intend* to browse—incidental but pertinent information is not likely to arise as long as the questioner tries to locate a "pat" reference with greatest possible dispatch.

The comparison between manual and machine search in the transistor art, as presented in this article, has concentrated on the basic differences in mental activities that occur with the two systems. Further work is needed to quantify and support the hypotheses presented here. The present analysis should be treated as developmental only.

This article is based on a paper presented at the 1965 National Electronics Conference, Chicago, Ill., Oct. 25–27.

BIBLIOGRAPHY

Stiltleman, J., "Information retrieval study of electronic patents," *J. Pat. Office Soc.* vol. 46, pp. 390–404, June 1964.



Conference report

Report on Prague's Summer School on Circuit Theory

K. Géher *Polytechnical University, Budapest*

From the 6th to the 15th of September, 1965, a Summer School on Circuit Theory was conducted by the Institute of Radio Engineering and Electronics [Ustav Radiotechniky a Elektroniky (URE)]. It was held in the recently built, decorative, and well-equipped Technical University of Prague.

The following topics were covered: topological analysis and synthesis, approximation, network analysis and synthesis in terms of sensitivity and tolerance requirements, problems of integrated circuits, active and nonreciprocal networks, networks with nonlinear and time-variable parameters, and the perspective of the circuit theory.

Among the 190 participants, the great majority (124) were electrical engineers, physicists, and mathematicians of the host country. There were, in addition, engineers and research workers of the following countries: German Democratic Republic (23), German Federal Republic (12), Poland (9), Hungary (6), United States (4), Italy (3), Soviet Union (2), United Kingdom (2), China (2), United Arabian Republic (2), Cuba (1), Denmark (1), Rumania (1), Yugoslavia (1).

The Summer School, which lasted ten days, differed from the conventional conferences in its organization and methods. Each of the 14 papers, which is a relatively small number, was read at a different time, making it possible for the participants to hear all papers. Each paper was allotted almost two hours for presentation, rather than the 15 to 20 minutes usually allowed for condensed reports at conferences. This arrangement ensured thorough exploitation of the theme of each paper given—which, after all, was the purpose of the Summer School.

A review of papers read at the Summer School follows:

- S. Prokop (Czechoslovakia), "Topological Analysis and Synthesis." Simplifications usable mainly in the topological analysis of *R-L-C* networks are considered with special regard to ladder networks. The calculation of the sensitivity is solved by the use of topological equations. The relations considered are quite suitable for computation by computer.
- G. Fritzsche (German Democratic Republic), "Approximation Methods in the Frequency Domain." The greater part of the paper is based on G. Fritzsche's book *Entwurf linearer Schaltungen* (*Design of Linear Circuits*). In addition, it offers a guide to the solution of approximation by means of computer.
- D. Calahan (United States), "Computer Design of Variable Parameter Networks." The author here notes that in the design, besides the specification, optimizing conditions (e.g., minimum sensitivity) must also be taken into account; that is, the choice between equivalent circuits is determined by the optimizing criteria. The paper also deals with another problem: the design of networks containing variable-parameter active and passive elements that satisfy a given specification. Examples of the above solved by computer are shown.
- K. Géher (Hungary), "The Tolerance and Sensitivity of Linear Networks." This paper is a review of the actual

Dr. K. Géher, who prepared this report, is an associate professor at the Polytechnical University, Budapest, Hungary.

state, methods, and results of the tolerance and sensitivity calculus. (All the material of this paper has since been published in the journal *Hiradastechnika*, Oct. 1965.)

• J. Novakova (Czechoslovakia), "Distributed Parameter Circuits." The report summarizes the analysis and design of circuits consisting of distributed parameter *R-C* elements. Characteristics realizable practically with distributed parameter *R-C* filters are also included.

• V. Zima (Czechoslovakia), "Fundamental Problems of Integrated Circuit Theory." This paper discusses the current state of the art of integrated-circuit technology along with the actual possibilities of integrated circuits with respect to dimensions and complexity.

• M. Novak (Czechoslovakia), "Theory of Distributed Parameter Reciprocal Circuits." Besides presenting a complete and clear review of the analysis of distributed parameter circuits, the author (who was in the United States during the Summer School and forwarded his paper to the participants) devotes his discussion to the synthesis of circuits built of homogeneous *R-C* sections in terms of Wyndrum's works.

• A. W. Keen (United Kingdom), "Active and Nonreciprocal Networks." A method is offered for handling linear but nonreciprocal circuits built of active and passive elements. Besides the conventional passive circuit elements, this technique employs a newly defined three-terminal active element, the so-called unitor, alone. The unitor cannot be considered an abstraction of any existing active device but its use permits the analysis and synthesis of general circuits to be substantially simplified.

• J. Braun (Czechoslovakia), "Analytical Methods in Active Network Theory." The paper shows the mutual relations between the ideal transformer, the ideal gyrator, and the negative impedance converter. It demonstrates how the circuit elements can be described by means of the various types of controlled generators: voltage-controlled voltage generators, voltage-controlled current generators, etc. The notions of nullator, norrator, and unitor are introduced and a method of the systematic use of the Kirchhoff equations for circuits containing such circuit elements are given.

• J. G. Linvill (United States), "Synthesis of Active Circuits." The author as inventor of the negative impedance converter presents a summary of his articles published in recent years. He shows the method by means of which the NIC can be used in the synthesis of active circuits. Among the many practical results, the increase of the Q factor of resonance circuits by means of *R-C* elements as well as that of NIC should be noted.

• J. Cajka (Czechoslovakia), "Matrix Method Analysis of Networks with Linear Nonreciprocal Active Elements." The author describes a "matrix" method suitable also for performing the network analysis by computer.

• E. S. Kuh (United States), "Nonlinear and Time Variable Networks." The first part of the paper shows the unique mathematical representation of the characteristics of circuit elements with nonlinear or time-varying parameters. By means of the formalism introduced, the current or voltage appearing at any point of the circuit can be computed in principle in a straightforward way. The second part of the paper contains general theorems with signal flow graph representation for the case of nonlinear and time-varying elements. The conditions necessary for the unique solution are formulated.

• M. D. Karassiev (Soviet Union), "Theory of Varying Parameter Radio Engineering Systems." The author notes that close attention must be paid to the periodically varying circuit. He presents the development of parametric systems and of the theories dealing with them, and reports on laboratory works concerning a traveling wave parametric amplifier.

• N. Balabanian (United States), "The Perspective of the Circuit Theory." A historic survey of the making and development of the circuit theory is presented in detail. With reference to the latest issues of the IEEE TRANSACTIONS ON CIRCUIT THEORY, the author points out some current problems of circuit theory.

The smooth execution of the readings and discussions was due, to a large extent, to well-organized technical arrangements. Simultaneous translation facilities provided a choice of either English, German, Russian, or Czecho-Slovak. A diapositive projector and a writing episcope served for the projection of figures and equations. The written material of the majority of the papers was distributed before the readings. (Papers read at the Summer School are available for inspection at the Polytechnical University of Budapest, Chair of Wirebound Telecommunication.) Papers will be published in the journal of the Technical University of Prague (which will be edited in several languages) beginning in 1966.

In Prague, too, had been confirmed the almost traditional characteristic of conferences, namely, that personal talks held between formal presentation of papers provide greater insight into problems interesting the research worker. With this preparation, more detailed information on results can be obtained during the paper reading. The following institutions were always open for the participants: Technical University of Prague, Institute of Radioengineering and Electronics, Popov Research Institute of Radiocommunications. Many possibilities for valuable talks were given at the two all-day excursions. The well-organized agreeable bus and ship trips contributed considerably to better personal relations accompanied, of course, by the broadening of scientific contacts. The courtesy of the host country was also evident in the social program planned for the relatives of the participants as well as the pleasant closing banquet of the Summer School.

The ten days in Prague made it possible to get better acquainted with the present state of circuit theory and to obtain a perspective of the future. The trend of development shows that new problems await those engaged in circuit theory, and their solution will require the introduction of new methods as well as the use of new circuit elements.

Thanks should be extended to the organizing Czech-Slovak Academy of Sciences, especially to research workers V. Zima, R. Vich, J. Braun, J. Cizek, and to the Technical University of Prague. O. Konicek and J. Gregor of the Technical University deserve special praise for the judicious planning of the Summer School on Circuit Theory.

IEEE publications

scanning the issues
advance abstracts
translated journals
special publications

Scanning the issues

Printable Batteries: Breakthrough! "Who ever heard of *printing* batteries?" That is the excited exclamation of David B. Dobson, executive editor of the IEEE TRANSACTIONS ON AEROSPACE AND ELECTRONIC SYSTEMS, in which a new breakthrough is reported, one that could open the door to printable batteries. Such batteries could be given all kinds of interesting form factors—on circuit boards, inside instrument and device cases, or on curved surfaces. Their implications in microelectronic applications are apparent.

In their paper on the new development, the authors, J. N. Mrgudich, P. J. Bramhall, and J. J. Finnegan, of the Institute for Exploratory Research, Fort Monmouth, N. J., report that it was while some of them were studying the adaptive memory characteristics of a dime-size pellet of compressed silver-iodide powder, which carried thin evaporated films of platinum and silver on opposite faces, they observed that such Ag-AgI-Pt arrays exhibited a batterylike behavior. The data they now present concern the charge–discharge characteristics of a rechargeable solid-electrolyte battery of solid silver iodide (the electrolyte), an evaporated film of silver (the anode), and an evaporated film of palladium (the cathode). This combination behaves like a rechargeable concentration cell with unexpectedly low internal resistance and unexpectedly high capacity.

Tests show that the batteries have a service life of about 400 minutes per cm^2 of anode area across a megohm resistor while maintaining a closed circuit voltage greater than one half the original open-circuit voltage (0.5 volt per cell). Arrays have been recharged for over 100 cycles with no apparent deterioration. The cells appear operable in the temperature range $-100°F$ to $+240°F$.

An intriguing objective of the work, say the authors, is to investigate the feasibility of a battery whose components—anode, electrolyte, cathode, intercell connector, strap connector, external contacts—are all deposited by substantially conventional vacuum-deposition techniques. Specifically, they wish to make a four-cell unit consisting of two series-connected two-cell units, as shown in Fig. 1. The tantalum film (which possibly can be a carbon film) serves as an intercell connector while preventing spontaneous diffusion of silver from the silver of one cell to the platinum of the adjacent cell. The exaggerated pyramidal configuration of each two-cell stack, achieved by suitable masking to yield sequentially smaller areas, is suggested as one approach to eliminate short-circuiting caused by overlap between electrode films or between electrolyte films.

Although, as the authors point out, it is premature to discuss in detail the many application areas that might open following the demonstration of the feasibility of such printable batteries, they cite some of the dramatic possibilities, as in microelectronics. Printable batteries could exhibit all of the indicated advantages of pellet batteries (good shelf life, recharge capability, operation under expanded environmental limits, rugged and shock-resistant construction, potentially low manufacturing cost), and, in addition, should provide a great flexibility in design, permitting batteries to be deposited in many shapes. Also cited are the advantages of circuit boards with a multiplicity of batteries whose individual voltages and capacities can be tailor-made to fit the electrical requirements of particular areas of the board.

The authors conclude that possibly an even more important feature of these concentration-cell batteries is that the open-circuit voltage is a function of the state of charge. This seemingly innocuous property, they say, may permit the battery to act in a sensing capacity, possibly as an adaptive memory element or as a triggering mechanism to activate system response to a change in environment. (J. N. Mrgudich *et al.*, "Thin-Film Rechargeable Solid - Electrolyte Batteries," *IEEE Trans. on Aerospace and Electronic Systems*, December 1965.)

Manned Space Experiments. Those who have read the interview this month with scientist–astronaut Owen K. Garriott may be interested in exploring further the kinds of experiments that

Fig. 1. Tentative design of rechargeable solid-electrolyte battery composed entirely of evaporated films. Such a "printed" battery could lead to dramatic new applications.