VOLUME 3 □ NUMBER 7 JULY 1966

IEEE spectrum

features

Departments: *Please turn to the next page*

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

The last decade has seen renewed interest in dc transmission as a result of our developing electron tube technology, although thus far, as pointed out in the article starting on page 75, present-day dc installations have been mainly underwater. The cover shows a cross section of the 200-kV dc submarine cable used in the Sardinia–Italy project.

Spectral lines

Several letters have been selected from the many received in response to "Spectral lines" in the May issue to kick off a new department in SPECTRUM *called "IEEE forum" (see page 8) where members may express their opinions on any phase of IEEE affairs. This is your department. Please use it. Letters pertaining to strictly technical matters will continue to be published in the Correspondence department, now called "Technical correspondence."*

What are we paying for? Must we read more in these columns about engineering education, and the financial problems of the Institute?

When I undertook the new task of editor, I realized that the "Spectral lines" had considered educational topics, or Institute activities and policies. I was determined that I would break the pattern and find other subjects for discussion. Since I am not an educator, and have not (until now!) been deeply involved in Institute activities, you would think that this would not be difficult. However, a number of events have caused me to change my mind.

Recently, two facts have been forceably driven home to me that I had not fully realized before.

1. *IEEE is predominantly educational in purpose.* We hear much about continuing education, retreading, renewal. In the past few weeks the subject was called to my attention by no less than four independent sources: a commencement address, several brochures on university summer programs, an excellent book,[1] and the final report of the Joint Advisory Committee on Continuing Engineering Studies.[2] The last item, a most timely report, gives a fine summary of the problem, some very illuminating statistics on engineering education and manpower, and, in addition to some two dozen recommendations to government and industry, nine specific recommendations to the engineering societies. The first of these is to: "*Actively promote as policy, through repeated publication and practice, the fact that one of the prime responsibilities of the engineer is to continue the learning process throughout life.*"

It is no news that the ever-increasing rate of change in our technology puts great strain upon the individual to maintain and improve his technical competence, and most of us will agree that we need repeated reminders and a great deal of help from publications, libraries, courses, and technical meetings in order to keep up.

I was impressed by the implication in the report that formal education has an effectiveness of only about five years. This agrees with what J. R. Pierce once said: "Education is like vaccination, it has a fine effect, but must be repeated periodically." It is certainly true that to be effective, education must be a way of life. While government, industry, colleges, and professional societies can and should do much more to provide the means for intellectual advancement, the principal responsibility falls on the individual. The main problem is one of motivation.

John W. Gardner put it well in his book: "*The ultimate goal of the educational system is to shift to the individual the burden of pursuing his own education.*"[1] Even within the university, it is not what the student does inside the classroom but what he does individually outside of the classroom that makes the difference between excellence and mediocrity.

To get back to the IEEE, isn't it true that just about every facet of our work is directed toward continued education? Meetings, publications, conventions, and committees all serve the purpose of advancing and disseminating knowledge. Can you name a major part of our activity that does not come under this heading? I can't. Now let us see if, following the recommendations of this report, we can't do our work better.

2. *It is often difficult to correlate what we get with what we pay.* This is obviously true of taxes and tuition, but perhaps even more so of IEEE dues. Some readers of this journal seem to think that the dues are a magazine subscription price, and others that it amounts to tribute imposed by "system." Actually the benefits for individuals are widely varied; they are quite different for student, educator, production engineer, research worker, and manager—and no one can claim that at any time one receives as much direct benefit as another. Indirect benefits are another matter. Each is benefited by the fact that he has the opportunity to attend meetings, take courses, peruse periodicals, and use references, whether he often does so or not. Few of us could put a price on the fact that engineering knowledge is made part of the public domain through professional society activities. Our dues are a small price to pay for this, and indeed are a small part of the total cost if you include the value of the freely donated time of active members by the thousands—from paper referees to officers.

You are *not* buying a subscription with your dues. In fact, if SPECTRUM were to become a little more commercial in its coverage and relax its advertising policy, it could be given away at a profit, like some of our friendly rival journals. It is true that part of your dues helps to pay for journals that you don't receive, but aren't they nearly as valuable to you on the library shelves as they are at home? They are to me; I can never find my own copies anyway.

C. C. Cutler

1. Gardner, John W., *Self Renewal.* New York: Harper, 1964.
2. This 100-page report is available from ECPD, 345 E. 47 St., New York, N.Y. 10017, $1 per copy.

Authors

Community antenna television systems (page 58)

Rodney D. Chipp (F), consulting engineer, Rodney D. Chipp & Associates, Bloomfield, N.J., received the B.S. degree in physics. In 1933 he joined the National Broadcasting Company, working in engineering operations and television development. He served in the U.S. Navy during World War II and received a commendation from the Secretary of the Navy for his work in radar design. Later, he was director of engineering at the DuMont Television Network and, subsequently, the Allen B. DuMont Laboratories. From 1957 to 1961 he held several important positions at ITT, such as director of engineering of ITT Communication Systems, Paramus, N.J. Throughout his career, his professional activity has been centered in broadcasting, television, radar and communications, and engineering management, and in 1961 he entered private practice in these fields.

Some current views on holography (page 67)

Robert J. Collier received the B.S., M.S., and Ph.D. degrees in physics from Yale University. In 1954 he joined Bell Telephone Laboratories, Murray Hill, N.J., working on high-power microwave oscillators and amplifiers, such as the first coaxial cavity magnetron, a forward-wave crossed-field amplifier with re-entrant electron beam, a high-power CW traveling-wave tube used as the power amplifier in Telstar ground stations, and, most recently, a pulsed multikilowatt traveling-wave tube for a multifunction-array radar defense system. He is now in charge of a group exploring the field of holography and investigating its potential for storing high-density information in an associative manner. Their experimental work has emphasized the volume nature of the hologram, and they have demonstrated that this aspect can be employed to produce multicolor holograms.

Submarine dc cables (page 75)

J. M. Oudin (SM), technical and scientific director of Les Câbles de Lyon, France, a branch of Compagnie Generalé d'Electricité, received the diploma of Ingénieur de l'Ecole Navale from the Brest Naval College. He served as an officer in the French Navy during the war and, from 1945 to 1952, was engaged in naval research. During this period he performed studies on information processing which were novel at the time. As a result of this work, the Institut de France, acting upon a proposal of the Académie des Sciences, awarded him a prize as an "originator of new submarine detection techniques." In 1952, he joined the Câbles de Lyon branch of the Compagnie Générale d'Electricité as technical and scientific director. Since then he has been engaged in studies on dielectrics and work on the development of the first submarine telephone and dc power cables and the development of very-high-voltage (up to 750 kV) ac cables.

R. A. Tellier (SM), head of the Lines and Cables Section of the Electricité de France System Equipment Department, Paris, France, received the Ing. E.S.E. degree from the Ecole Supérieure d'Electricité, Malakoff, France, in 1946. After a year of service in the French Navy, he joined the research organization of Electricité de France and was appointed to the System Equipment Department. In 1949 he was put in charge of the preparation and execution of a research program on cables for all voltages up to 380 kV at the Fontenay Research Center, and he is still responsible for all research work in the cable field. From 1950 to 1960 he was actively engaged in the preliminary work on the cross-channel project, a submarine interconnector between French and British high-voltage systems. Since 1957 he has had the additional responsibility for Electricité de France research work relating to overhead lines equipment at all voltages and is now head of the Lines and Cables Section.

Optical communications in the earth's atmosphere (page 83)

Bernard Cooper (M) received the B.E.E. degree in 1950 and the M.E.E. degree in 1956, both from the Polytechnic Institute of Brooklyn. In 1953 he began his association with ITT Federal Laboratories, Nutley, N.J., where his work was initially concerned with the design and development of missile guidance and radar subsystems. Since 1958 he has been a member of the technical staff of the Advanced Systems Analysis Section of the ITT Federal Laboratories Space Communication Laboratory. As Section head of the Advanced Systems Analysis Group in 1960, he directed operations and data analysis activities at the ITT Federal Laboratories Space Communications Research Ground Station. More recently, he has directed advanced study activities in coherent optical communications. At the present time he is engaged in communication satellite system studies in connection with controlled data storage techniques.

Frequency relations of parametric interactions (page 93)

H. Seidel (M) received the B.E.E. degree in 1943 from the College of the City of New York, the M.E.E. degree in 1946, and the D.E.E. degree in 1954. From 1947 to 1953, he worked in succession at the Microwave Research Institute, Polytechnic Institute of Brooklyn, Arma Corporation, Brooklyn, N.Y., and the Federal Telecommunications Laboratories, Nutley, N.J. He has been with the Bell Telephone Laboratories in Murray Hill, N.J., since 1953, except for the period 1961–1962 when he left to help reorganize the Merrimac Research and Development Company, Irvington, N.J. His major technical activity while at the Merrimac Company was the development of miniature high-frequency couplers. At Bell Telephone Laboratories, his work has centered on solid-state devices with particular activities in ferrite propagation studies, parametric amplification, and certain relevant forms of microwave network synthesis.

Roadblocks in the path of controlled fusion (page 106)

A. S. Bishop, assistant director of research, U.S. Atomic Energy Commission, received the B.S. degree in physics from California Institute of Technology in 1943 and the Ph.D. degree in nuclear physics from the University of California, Berkeley, in 1950. From 1943 to 1946, he was associated with the M.I.T. Radiation Laboratory. He became a research fellow at the Swiss Federal Institute of Technology in 1950. Three years later, he joined the U.S. Atomic Energy Commission and headed its program on controlled fusion. From 1956 to 1958 he provided liaison between the AEC and European laboratories working on peacetime uses of atomic energy and, later, headed the U.S. delegation to EURATOM studying a joint program for nuclear power development. In 1961 he joined the Princeton University Plasma Physics Laboratory, performing research on controlled fusion. Last February, he returned to AEC to head a program on controlled fusion.

Future role of satellites in air traffic control (page 112)

Raymond E. Spence received the B.S. degree in physics from Ohio State University in 1951 and joined the Industrial Nucleonics Corporation as a research project engineer. He became associated with the Philco Corporation in 1954, working with the Army Signal Corps on the European microwave system and with the Air Force on engineering communications and navigational aid systems. In 1958 he became chief of the Air Force Central GEEIA Regions Avigational Systems Engineering Branch. He has been with the Federal Aviation Agency since 1960. At present, he serves as chief of the Research and Development Services Voice Communications Systems Branch and is concerned with such programs as the application of solid-state switching techniques to air traffic control facilities, modernization of the air–ground–air communications system, and application of satellite technology to international air traffic control facilities.

Since their inception in 1949, community antenna television systems have experienced rapid growth. They usually are found in areas where conventional antennas do not provide adequate television reception. The basic equipment consists of a "head-end," which picks up the off-air signals, and the distribution network, which delivers the signals to the subscriber's television or FM receiver. The central antenna is erected at the most advantageous site for best reception of desired channels. Today, it is estimated that CATV serves approximately 2 million homes, and this number is increasing constantly. As a result, there is now much controversy over such factors as franchises, rates, and standards.

Communit

The use of community antenna television systems (CATV) began on a small scale some 15 years ago. The subsequent rapid growth of these systems—in number, in size, and in capacity—has been the result of public acceptance, and has been greatly assisted by advances in the electronic technology. This relatively new business is now the subject of considerable discussion and controversy among regulatory and legislative bodies ranging from town councils to the Congress itself. The present article discusses some major aspects of CATV systems.

A CATV system is a means whereby a paying subscriber obtains television signals, and often FM, from an advantageously located central antenna system and a direct cable connection. Such systems are almost a necessity in some sections of the country; in other areas they provide improved reception without the need for costly individual antennas. Also, many of these systems now offer a wider choice of programs than are available directly from local stations.

Basically, a CATV system consists of a "head-end" and a "distribution network." The head-end is that part of the system that picks up the off-air signals, amplifies them, and sometimes converts them to another television channel. The distribution system carries these signals by coaxial cable and repeater amplifiers from the head-end to each customer's home, where they are received on a conventional television or FM receiver through a "subscriber drop."

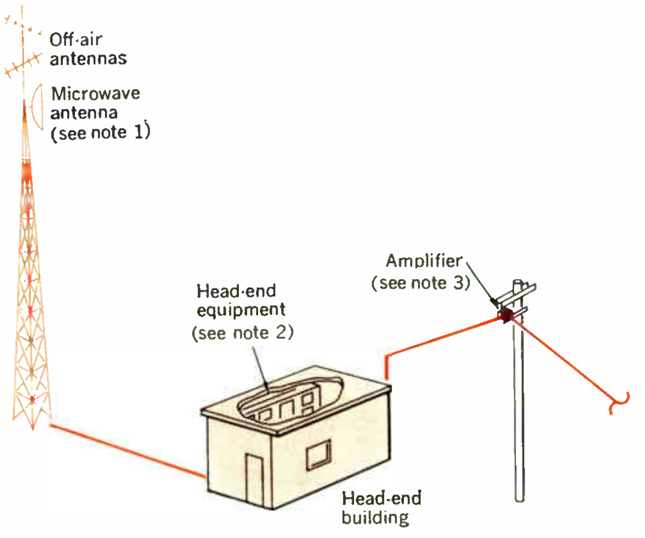
Basic principles

Generally, community antennas are found in areas where distance from stations or topographic factors preclude good reception of television signals by conventional antennas. With the aid of elaborate and expensive elevated antennas, reception can be improved. Alternatively, a community antenna can be erected at a location where the signals of desired stations have sufficient field strength to produce high-quality pictures.

Most systems today provide more than one channel to the subscriber. Generally speaking, sites are selected where a number of nonduplicating and noninterfering channels are received, each with sufficient field strength to produce pictures relatively free of noise and interference. Directional antennas are oriented for best reception of the desired channels. This is basically the function that a homeowner performs when he erects an antenna, which has been designed to give optimum reception on the channels he desires, on the roof of his house, orienting it so as to receive the desired channels and to reject those that might duplicate or interfere with those he prefers.

A typical CATV system, shown pictorially in Fig. 1 and schematically in Fig. 2, will usually have one or more antennas for each channel to be received. Rather elaborate arrays of stacked antenna elements are often used to achieve higher gain and to reduce interference. The very-high-frequency signals received by these antennas may be preamplified directly at the antenna to improve signal-to-noise ratio, while ultrahigh-frequency signals are usually converted to VHF channels directly at the antenna to reduce losses in the transmission line between the antenna and head-end equipment. Such amplifiers and converters normally receive their power through the transmission line.

Where microwave systems are used to bring distant channels to the CATV system, the microwave receiving antenna and associated equipment are usually located at



ntenna television systems

A means of providing improved television reception in fringe or shadow areas, the CATV system utilizes an advantageously located central antenna to serve its subscribers. Such systems have grown rapidly in popularity—and have also engendered much controversy

Rodney D. Chipp   Rodney D. Chipp & Associates

the head-end. At this point the received signals are converted to VHF television channels for distribution through the system. The head-end equipment may also convert higher channels to lower channels (e.g., channel 13 to channel 2), in order to minimize transmission losses in the 75-ohm distribution cables. Table I shows typical transmission-line losses; Fig. 3 shows cable types.

Among the various units of equipment used at the head-end to perform these functions—in addition to the previously mentioned antenna system, preamplifiers, and converters—are single-channel or "strip" amplifiers, tuners, carrier generators, demodulators, modulators, filters, traps, and mixing networks.

Strip amplifiers with automatic gain control (AGC) are tuned to a specific television channel and provide output levels on the same channel of the order of one volt (CATV systems use the convention that one millivolt across 75 ohms is 0 dBmV; thus one volt is 60 dBmV). Traps and filters are used to reduce adjacent channel or other interference, and variable or fixed pads to equalize levels. Tuners and associated equipment may produce modulated radio frequency on a higher channel (up-conversion) or on a lower channel (down-conversion).

Finally, the channels to be transmitted are mixed and fed to the wide-band distribution system. As is the case with all multichannel transmission systems, detailed system planning as to frequency conversions, gains and losses, filtering, amplifier location and spacing, optimum cable size, routing, etc., is extremely important. Too high a level will cause overloading and interference between channels; too low a level will decrease the signal-to-noise ratio. As shown in Table I, a typical 7/8-inch trunk line cable will have a loss of 0.4 dB per 100 feet at channel 2 and a loss of 0.8 dB per 100 feet at channel 13. Thus, 4000 feet of such cable will have a loss of 16 dB at channel 2 and 32 dB at channel 13 and the cable system must be equalized for these large differences. Uusually the trunk line amplifiers are compensated for these attenuation differences by controllable "slope" or "tilt." Most modern amplifiers also use temperature-compensating devices, since cable attenuation varies with temperature. Many large systems feed pilot tone, as is done in carrier telephone plant, and use AGC amplifiers to hold the level constant. The amplifiers may be mounted in weatherproof cabinets on utility poles or be attached directly to the messenger cable; some systems use underground cable.

The following are some typical values for trunk line

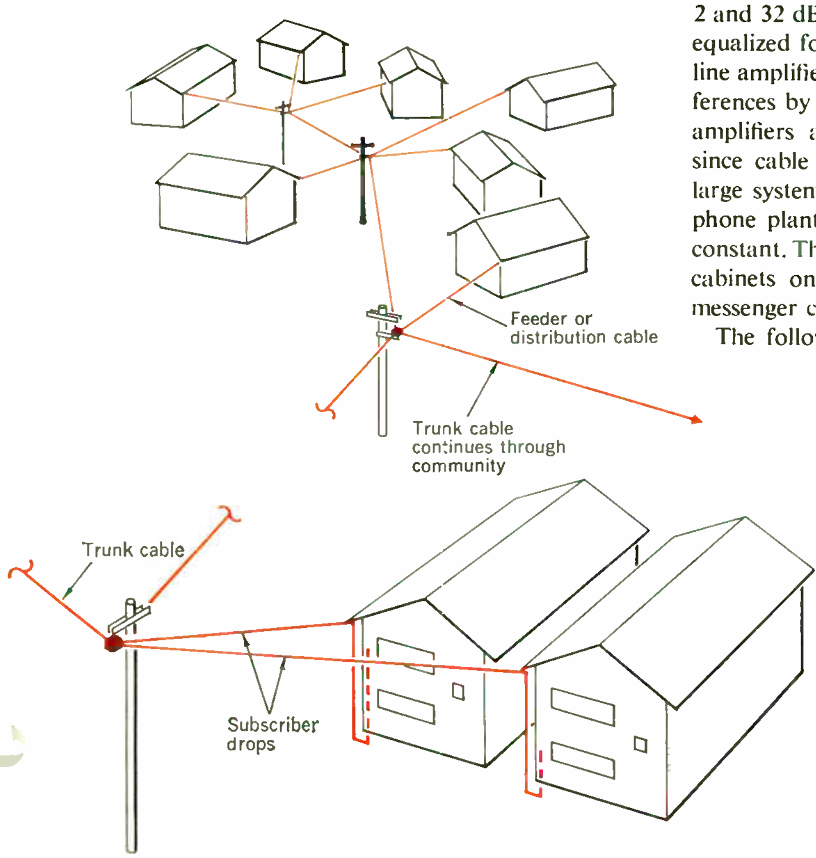


Fig. 1. A CATV system. Notes: (1) Some CATV systems use microwave for stations too far away to pick up off the air. (2) The head-end equipment may include local origination of time, weather, etc. (3) Types of amplifiers used are determined by function; e.g., trunk line, bridging, extender.

I. Approximate transmission-line losses in dB per 100 feet for typical CATV cables

Type of Cable	Use	Channel No. and Picture Carrier Frequency, Mc/s: 2 / 55.25	6 / 83.25	7 / 175.25	13 / 211.25	14 / 471.25	83 / 885.25	Outside Diameter, inches
RG59/U type with foam dielectric	House drop	2.2	2.6	3.8	4.2	—	—	0.25
RG11/U type with foam dielectric	Distribution	0.9	1.1	1.7	1.9	—	—	0.40
Aluminum outer conductor, foam dielectric	Feeder	0.7	0.9	1.5	1.8	2.5	3.5	0.50
Aluminum outer conductor, foam dielectric	Trunk	0.4	0.45	0.65	0.8	1.4	1.9	0.87
Aluminum outer conductor, foam dielectric	Trunk	0.25	0.3	0.45	0.55	0.9	1.5	1.62

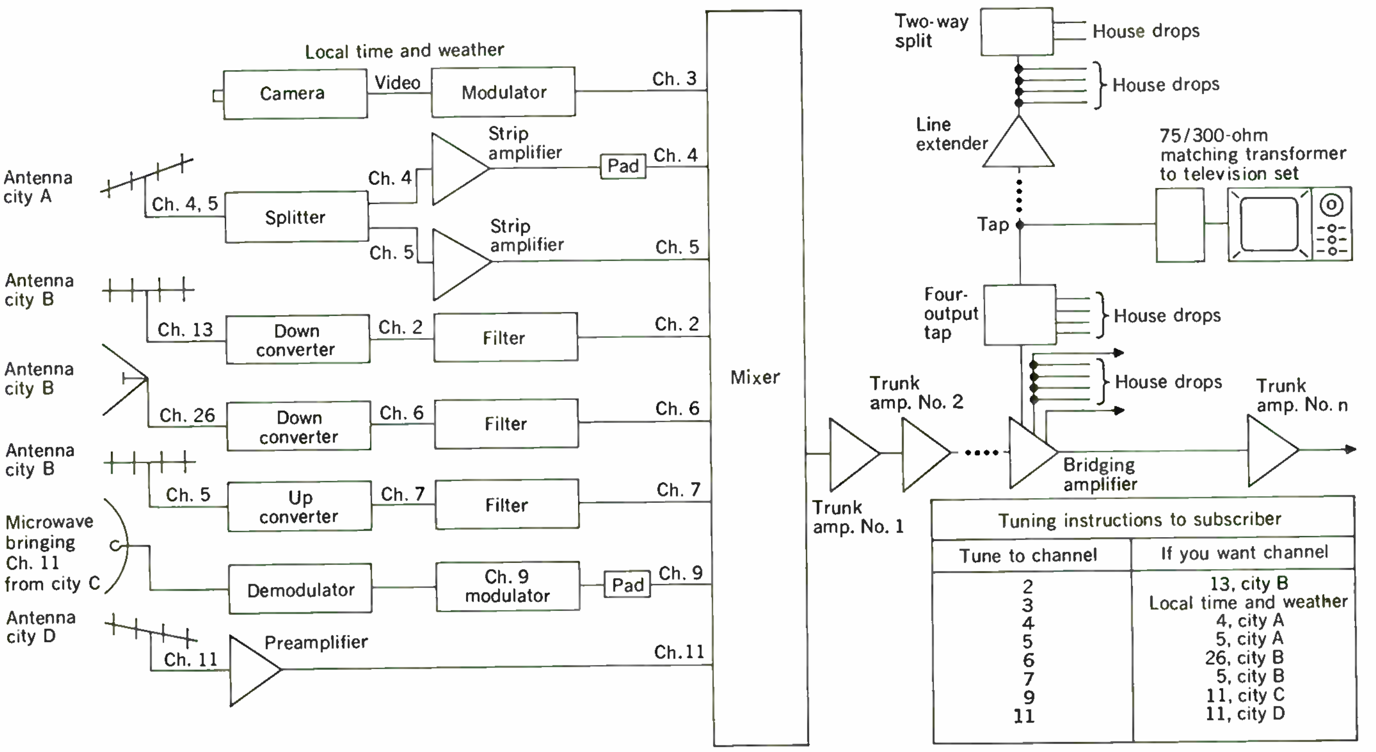
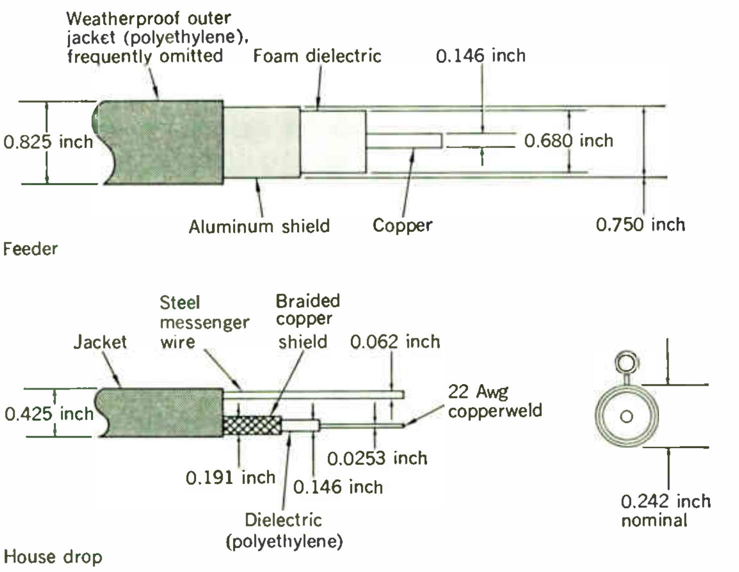


Fig. 2. Hypothetical CATV system schematic showing types of equipment used.

Fig. 3. Typical CATV cables.



amplifiers to be used in cascade:

Input	+10 dBmV
Output	+35 dBmV for 0.5 percent cross modulation
Bandwidth	54–216 Mc/s
Noise figure	12 dB
Tilt	12 dB over frequency range
Response	±1 dB over the band
Gain	25 dB

As the trunk cable passes through the community to be served it is split into feeder lines by distribution amplifiers that have one input and as many as four outputs. If the feeder line is long, an extender amplifier may be used. Feeder lines are distributed throughout the community and connected to subscriber homes by "taps" and "house drops." Many varieties of taps are available (resistive, capacitive, directional); their function is to bridge the feeder line with a high impedance, thus keeping impedance discontinuity to a minimum, and to deliver a source-terminated (75-ohm) signal of suitable level to the line feeding the subscriber's television receiver. A typical directional tap will have an insertion loss of about 1 dB on the feeder cable and will provide about 20 dB of attenuation to the subscriber drop.

Most television receivers have an input impedance of 300 ohms, to permit the use of inexpensive twin-lead transmission line, whereas, as noted previously, coaxial cable of the type used for CATV systems has a characteristic impedance of 75 ohms. Therefore, a 1:4 matching transformer must be installed at the back of the television set or as part of the wall receptacle to which the set is connected. "Splitters" are used if additional television sets or FM radios are to be fed from the CATV system.

History

Community antenna television systems were introduced in 1949 and it is believed that the first installations were in Astoria, Oreg., and Lansford, Pa.[1-3] These systems were developed in response to the public demand for acceptable television reception in fringe or shadow areas. The early systems carried one, two, or three channels, depending upon which stations could be received at the antenna location. Since these systems answered a public need, their use spread rapidly throughout the country. The systems increased in size as they moved from small communities into larger towns and, simultaneously, wide-band amplifiers were improved so that the channel capacity was expanded to five (channels 2–6).

It is estimated that by 1960 there were 800 CATV systems in the United States, serving 750 000 subscribers. By 1962 some systems were carrying all of the 12 VHF television channels as well as FM. Because the added capacity permitted viewers a wider choice of programs, systems were installed in larger cities in which viewing might otherwise be limited to one, two, or three local channels. Another aspect of CATV growth has been the use of microwave relay to interconnect systems and to bring programs originating in faraway stations to the local viewers. Figure 4 shows some of the details of such a system[4]; also see Appendix I. Finally, many CATV systems have provision for local origination of time, weather, and even live programs of local interest. In fact, some franchises require a certain amount of "local public service" programming. For example, the CATV company in Winchester, Va.,[5] must provide 30 hours a month of news, local sports, and discussion groups.

Today it is estimated[6,7] that there are more than 2000 operating systems in the United States and Canada, more than 1000 franchise holders not yet in operation, and more than 2000 applications pending. It is also estimated that present CATV systems serve approximately 2 million homes, or more than 6 million viewers, based on an average of 3.3 persons per home. This represents over 3 percent of the estimated U.S. television audience, and the number of people served is growing at an estimated rate of 15 percent annually. According to a recent report,[8] an average system now has 655 subscribers; 90 percent of the systems studied had less than 300.

A typical large system, in Vancouver, B.C., provides 12 channels to about 30 000 homes and uses about 250 miles of cable; and a newly opened system in Harrisburg, Pa., is even larger. The use of CATV is presently contemplated in metropolitan centers such as Los Angeles,

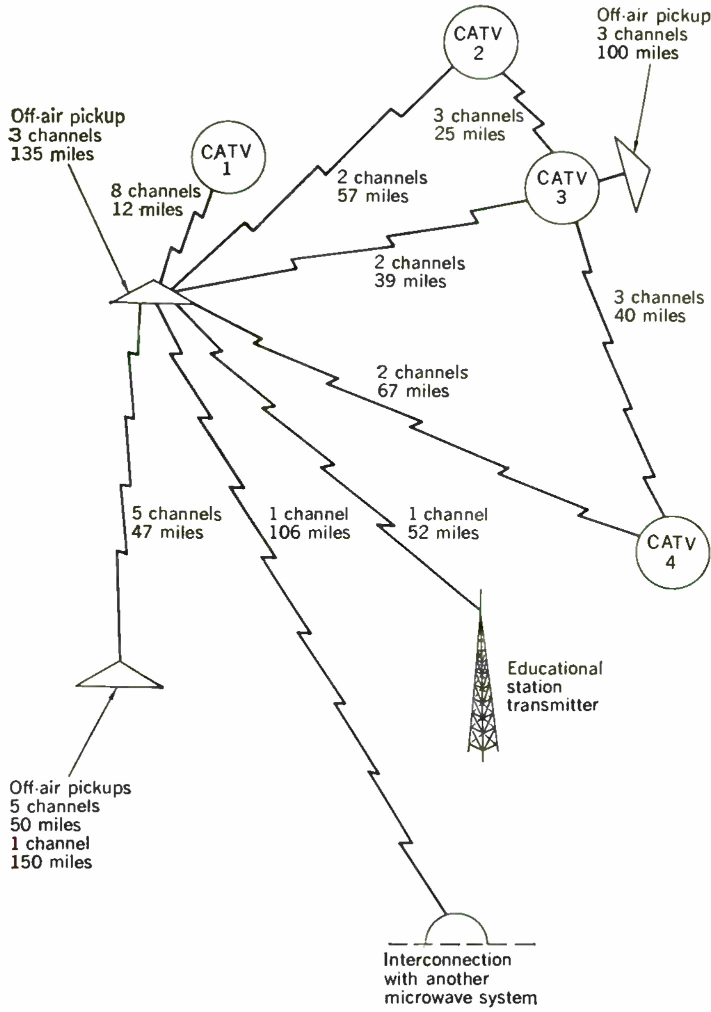


Fig. 4. A CATV microwave system.

Detroit, and Philadelphia, and franchises have been awarded in New York City. Many moderate-sized communities with CATV systems have a wider choice of programs than is available in large cities. For example, in Keene, N.H., with a population of about 17 500, CATV subscribers have access to three network stations in Boston, stations in Hartford, Manchester, Albany, and Greenfield, educational stations in Boston and Durham, a time and weather channel, and FM. In Utica, N.Y., with a population of about 100 000, the subscribers have a choice of three network stations in the Albany area, three independent stations in New York City, a local station, educational stations in New York and Schenectady, a time and weather channel, and FM. (See Table II.) A time and weather channel is a locally originated signal, sometimes called "weather-eye"; it consists of continuous scanning in sequence of a clock and several meteorological instruments. Messages from local merchants can be placed on this channel, and devices have recently become available that present news as well.

Costs

In the early days of CATV, charges for service[9,10] varied widely, but as the systems became more widespread the cost to the subscriber tended to stabilize. At present, the average initial installation fee is about $20 and the average monthly charge is about $5 for one television connection. The following is a typical rate card:

Installation charge, first TV set	$29.95
Installation charge, each additional TV set	9.95
Installation charge, FM radio with TV set	4.95
Installation charge, FM radio without TV set	19.95
Monthly rental, first TV set	4.95
Monthly rental, each additional TV set	1.95
Monthly rental, FM radio with TV set	1.00
Monthly rental, FM radio without TV set	2.95

The CATV system is not "pay TV"; the charges are not related to the programs available, the programs viewed, or the amount of time the television set is operated. The cost to the system operator will vary depending upon the size of the system, the number of channels, the complexity of the antenna system and head-end, etc. Some rules of thumb have been developed that permit reasonably accurate estimates to be made for an average system:

Cable cost	$4000 per mile (wide-band system) $3000 per mile (low-band system)
Number of utility poles	Cable miles divided by 50
Pole rental	$3.50 per pole per year
Antenna and head-end	$750 per channel
Tower	$25 per foot

In some areas, Bell System and independent telephone companies have constructed wide-band distribution systems. These services are offered to CATV operators at rates filed with the Public Utility Commissions having jurisdiction in the areas covered. The distribution system rentals, which vary with the number of channels, are of the order of $60 to $80 per mile per month. When a common carrier operates the distribution network, the CATV operator usually provides the head-end and the subscriber connections.

II. Cable channels for two typical communities

Channel	Station Received
	Keene, N.H.
2	WGBH (Boston), educational
3	WTIC (Hartford), CBS
4	WBZ (Boston), NBC
5	WHDH (Boston), CBS
7	WNAC (Boston), ABC
8	Time and weather
9	WMUR (Manchester, N.H.), ABC
10	WTEN (Albany, N.Y.), CBS
11	WENH (Durham, N.H.), educational
12	WHXR, FM music only
13	WRLP (Greenfield), NBC
	Utica, N.Y.
3	Time, weather, and background music
4	WKTV 2 (Utica), NBC (ABC)*
5	WNEW 5 (New York City), independent
6	WRGB 6 (Schenectady), NBC, or closed circuit
7	WAST 13 (Albany), ABC
8	WTEN 10 (Albany), CBS
9	WOR 9 (New York City), independent
11	WPIX 11 (New York City), independent
12	WMHT 17 (Schenectady) or FM, educational†
13	WNDT 13 (New York City) or FM, educational†

*WKTV 2 is primarily NBC, but when it switches to ABC, the NBC programs will be seen on cable channel 6.
†When the educational stations are not broadcasting FM music will be heard on cable channels 12 and 13.

Regulation

The spread of CATV into larger communities, the extension of program choice and availability by means of microwave, the fact that local programs may be originated and that CATV operators do not pay copyright fees or program costs have all raised questions in the minds of some broadcasters.[11,12] The opposing arguments may be summed up by considering recent comments of the presidents of the National Association of Broadcasters and of the National Community Television Association.

Vincent T. Wasilewski (NAB): "Broadcasters' concern centers basically around the question of whether CATV will be a supplementary service—as it started out to be—(and many of its proponents still assert it has no ambitions to be otherwise) or whether it will destroy the present system of free broadcasting, using as its principal and most lethal weapon our own programming. For, if CATV systems are permitted to bring distant radio and television stations into a community, then the ground rules which have governed broadcasting since its inception will have undergone radical changes.

"If Chicago television stations, for example, can be brought in via CATV to Rock Island and Moline, into Rockford and Peoria, into Quincy, Springfield, Decatur, and Harrisburg, the impact on local stations in those cities would be severe—so severe, perhaps, that stations in those towns would have to curtail local service."[13]

Frederick W. Ford (NCTA): "Out of the welter of all the conferences, committees, and controversy, one factor must stand out crystal clear. There is and must be a natural alliance between the CATV industry and the broadcast industry. They are in fact both parts of the same entity—CATV cannot exist without a sound TV broadcast structure—and the broadcast industry cannot withstand the public demand for additional service. I, for one, can assure you that the CATV industry is and must be dedicated to a strong, healthy, competitive television industry, and will do nothing to weaken it.

"We owe it to the public of this country, which, after all, supports both industries, to cooperate with each other and with the Government in arriving at solutions to the problems attendant upon the growth of any new industry. We must, in this manner, assure through an orderly process, that CATV becomes an integral part of the overall broadcast and communications structure of the United States which is so vital to our economic, social, and political system.

"Finally, I am convinced that both television and CATV will grow and expand and out of the conflict that exists today, and out of the fair competition between stations which CATV fosters, will emerge a stronger total television system, with better programs and clearer pictures. This is the very essence of the public interest, convenience and necessity which CATV systems seek to promote."[14]

As a result of this growing controversy, the FCC has issued regulations and recommended proposed legislation covering CATV systems. In a public notice[15] issued March 8, 1966, the FCC announced its decision. Major items covered by the new regulations are:

1. The FCC has assumed jurisdiction over all CATV systems whether or not they use microwave facilities.

2. All CATV systems will be required to carry, without material degradation, the signals of all local television stations within whose grade B* contours the CATV system is located.

3. The CATV system shall avoid duplication of the programs of local television stations on the same day that the programs are carried by the local station. A local television station will not be entitled to such protection when it carries in black and white a program offered in color by the network.

4. In the top 100 markets (see Appendix II), in the absence of a waiver, prior FCC approval after an evidentiary hearing will be required before a CATV system can bring in the signals of a television station in another market where this would extend the signals of the station beyond its grade B contour. This applies after February 15, 1966, and will not affect present service.

In the smaller television markets, below the top 100, neither prior FCC approval nor an evidentiary hearing will be required.

The recommended legislation will include:

a. Clarification and confirmation of FCC jurisdiction over CATV systems generally, along with such specific provisions as are deemed appropriate.

b. Prohibition of the origination of program or other material by a CATV system with such limitations or exceptions, if any, as are deemed appropriate.

c. Consideration of whether, to what extent, and under what circumstances CATV systems should be required to obtain the consent of the originating broadcast station for the retransmission of the signal.

d. Consideration of whether CATV systems should or should not be deemed public utilities. In this connection, Congress will be asked to consider the appropriate relationship of federal to state–local jurisdiction in the CATV field, with particular reference to initial franchising, rate regulation, and extension of service.

Franchise agreement

Most CATV systems presently operate under some form of franchise issued by the local governing body. In most cases the franchise is exclusive, meaning that the holder of the franchise has no competition in that locality, so there is often competition among prospective CATV operators for these franchises. Although the franchise agreements vary widely in complexity and degree of completeness, they usually have four basic and different areas of coverage: conditions of use, finance, liability, and service.

Under *conditions of use* are such items as the right to construct and use poles, repair of damage to city or private property, interference with the public, obstruction of traffic, term of the franchise, and renewal or cancellation provisions. In some cases there are requirements to carry local programs and to carry educational programs into schools free of charge. There may also be prohibitions against duplication of programs carried by local stations and against pay television.

* The "FCC Rules and Regulations," par. 73.683, define the grade B contours as: channels 2–6, 47 dB above 1 μV/meter; channels 7–13, 56 dB above 1 μV/meter; channels 14–83, 64 dB above 1 μV/meter. These are statistically derived values, and are based on assumptions that a quality acceptable to the median observer is expected to be available for at least 90 percent of the time at the best 50 percent of the receiver locations at the outer limits of this service contour.

Under *finance* are provisions for payment to the municipality and requirements for inspection of books and periodic audit. There may also be references to the installation charges and monthly rental rates.

Under *liability* are requirements for a performance bond, indemnification for claims for damage, and public liability insurance. In addition, there are instances in which the municipality has required protection against copyright infringement, since there are pending court actions on this matter.[12]

Under *service* are requirements as to number of channels, picture quality, nondiscrimination regarding service and rates among subscribers, and the like. Typical clauses in this category are

"The Company guarantees that it will furnish reception from at least__television channels and__FM channels. Service shall be continuous during the operation hours of the stations whose signals are being received."

"The Corporation shall at all times establish and maintain reasonable standards of service and quality of products consistent with those established and maintained generally in the field of providing television transmission to viewers by coaxial cable, and shall not unjustly discriminate in its service or rates among its subscribers."

"All installation of equipment shall be of a permanent nature, durable, and installed in accordance with applicable portions of the National Electrical Code."

Technical standards

As noted previously, the generally used franchise provisions for service, as compared with those covering conditions of use, finance, and liability, are something less than rigorous. None of the franchise documents reviewed included detailed technical standards, which is in contrast with the documents describing power and telephone service. For example, a typical power company agrees to maintain the voltage at ±4 percent of its nominal value and the frequency at ±3 percent of its nominal value. Similarly, if a class A line is ordered from the local telephone company, it will usually provide a nominal bandwidth of from 100 to 5000 c/s, accept a level of 0 dBm and deliver a level of 0 dBm, and have a noise level of not more than 36 dB above reference noise. Therefore, it seems reasonable to expect that there will be common acceptance of minimum standards in future CATV franchise agreements. This is particularly true in cases in which an exclusive franchise is granted, and customer service quality may have to be assured by proper technical auditing based on objective measurements.

In the work of the National Television Systems Committee[16] both technical and subjective factors were considered. This work was promulgated in the various rules and standards issued by the FCC, the Electronic Industries Association, and IEEE. Further work by the Television Allocations Study Organization[17] has added to our knowledge of propagation and the subjective effects of various kinds of interference.

Since the point of origin of a CATV system signal is the broadcast signal taken from the air at the CATV antenna, we can say that a CATV operator, in effect, supplies an antenna, amplification, sometimes frequency conversion, and an RF transmission system consisting of amplifiers and cable. Although it would be technically possible for the operator to modify critical parameters such as the sync timing relationships and waveforms, there should be no reason to do so. Also, such parameters as geometric distortion and contrast range are a function of the pickup cameras in the studio and the receivers in the home. What elements of picture quality, then, can the CATV system affect?

An individual putting up an antenna system would align it for the desired stations and locate it so as to minimize interference. A low-loss transmission line would be used to provide sufficient signal strength at the receiver input terminals. In an area of low signal strength, a preamplifier might be located at the antenna to offset the line losses; the amplifier would need to have low noise input, appropriate bandpass and phase-shift characteristics, and sufficient dynamic range. It would also be necessary to match the impedances of the various parts of the system and see that they did not introduce any extraneous interference. Basically, these are the elements of a CATV system that can affect the received signal that are within the control of the system operator. In practice, they are determined by antenna configuration and placement, by selection of equipment having suitable characteristics, and by adequate system design and layout.

A system specification, especially one that might become part of a legal document, should describe performance, not attempt to set forth the details of equipment design. Further, system performance should be described as simply as possible and in terms of only those characteristics that are affected by the system design.

In CATV the customer is concerned primarily with the quality of the pictures and sound, the safety of the installation, and continuity of service. The system operator, on the other hand, will be concerned with mechanical and electrical items, such as connectors, mounting hardware, temperature compensation, slope control, and automatic level control. With this distinction in mind, and with full recognition that weather conditions and terrain irregularities have a significant effect on television propagation, that interference is not always subject to control, and that optimum site selection is not always possible, we can develop minimum performance standards that are adequate, reasonable, and simple. One version of such a specification[18] is outlined in the following:

1. The system shall be designed to transmit NTSC monochrome and color signals over the 54–88-Mc/s and 174–216-Mc/s frequency bands and to transmit FM signals over the 88–108-Mc/s band.

2. The signal level at the picture carrier frequency for each desired television channel shall be at least 1000 μV for 95 percent of the time at the input to the distribution system.

3. The following values, with respect to the desired carrier level at the subscriber receiver terminals, should not be exceeded: random noise (4-Mc/s bandwidth), -34 dB (which is approximately equal to -40 dB measured on a standard CATV field strength meter); hum, -30 dB; cross modulation, -40 dB. With regard to cochannel and adjacent channel interference, such signals are subject to propagation vagaries, the interference takes several forms, and the effects on picture quality are significantly different. Simple and inexpensive methods of measurement need to be developed before standards are set.

4. At the input to each customer television receiver the maximum amplitude variation across any television channel shall be less than ± 2 dB.

5. At the input to each customer television receiver the signal level for each desired channel shall be between 750 and 3000 μV referred to 75 ohms. The accompanying aural carrier shall be between 7 and 10 dB below the visual carrier.

6. The entire system shall be designed for maximum return loss (i.e., minimum VSWR). At any subscriber location reflected signals shall be 20 dB or more below the desired signal.

7. Isolation between any two customers shall be at least 35 dB.

8. The system shall be designed for continuous operation.

9. These performance specifications shall be met in outdoor temperatures ranging from -50 to $+130$°F, and with variations in supply voltage from 105 to 130 volts.

10. Spurious electromagnetic radiation shall fall within the limits specified by the Federal Communications Commission in vol. II, pt. 15, subpart D, of their rules.

11. The installation shall conform to Articles 800 and 810 of the National Electrical Code.

The NCTA Engineering Committee has been active in considering industry standards. In a recent statement[19] before the FCC, the NCTA took the position that "the CATV operator will do his utmost to make available the best reception possible," and it was suggested that no technical rules be adopted. However, certain recommendations were made as "technical guidelines" should the rules be adopted. Abstracts of the guidelines follow.

a. Signal voltage (defined as the rms voltage, during synchronizing pulse interval, of the visual carrier modulated with FCC standard composite video signal) measured at the input to the first CATV amplifier, shall be not less than 150 μV on channels 2–13, and not less than 300 μV on channels 14–83.

b. The frequency response at the system input between 0.5 and 4.2 Mc/s shall not differ by more than 2 dB from the response observed on a half-wave test dipole, located at the CATV site, tuned to the visual carrier frequency of station X.

c. At the subscriber drop the visual carrier level shall be not less than 500 μV across 75 ohms. Aural carrier level shall be at least 7 dB and no more than 20 dB below visual carrier level.

d. Visual carrier of the upper adjacent channel shall not be more than 6 dB greater than the visual carrier of station X.

e. Aural carrier of the lower adjacent channel shall be at least 10 dB less than the visual carrier of station X.

f. The ratio of rms voltage of the visual carrier during synchronizing pulse interval to the rms thermal noise voltage, within a 4-Mc/s bandwidth, measured at the system output, shall be not less than 34 dB.

g. Picture degradation (ringing, smear, streaking, loss of resolution or contrast, hum, reflections, or beats from spurious or adjacent channel signals) shall be tested by comparison of pictures of test patterns photographed at the antenna output and the subscriber drop.

h. Cross modulation of a continuous-wave signal substituted at the system input terminals for the visual carrier of station X, with all other station signals adjusted for normal operation, shall be not less than 46 dB.

i. The possibility of color degradation will be investigated by visual comparison of the picture received at the antenna output and at the output of the single-channel equipment at the head-end. This is based on the supposition that, generally speaking, a wide-band RF system will not cause color degradation.

The complete recommendations are more explicit; they include suggested methods of measurement, limitations, and explanatory material. However, the foregoing abstract indicates, in general, the scope and the minimum requirements as viewed by NCTA.

Both sets of suggested technical standards described here are performance standards, and do not include much of the detail that would be required in a specification for construction of a system. In obtaining bids for a system, a CATV operator would need, in addition to overall electrical performance, details on type of cable, type of hardware, physical requirements, installation requirements, channels desired, system drawings, etc.[20]

Conclusion

In summary, it may be said that CATV performs a service to a large segment of the population and that, at present, it depends on broadcasting for its primary source of programs. Regardless of the scope and extent of current or future regulation, the formation of technical standards and other suitable ground rules will improve service to the subscriber. Wire communications, if technically and economically feasible, are often superior to radio services in that they are not subject to mutual interference and propagation vagaries. It is not unreasonable to surmise that in years to come many homes and business establishments, particularly in countries with highly developed communication systems, will handle transmission and reception of voice and data, as well as visual and aural entertainment, on a single coaxial line, thus widening the use of the radio spectrum by land, sea, air, and space vehicles.

Appendix I. Common carrier microwaves serving CATV

The following lists the services extant as of March 1.

Alabama Microwave, Decatur, Ala.
American Microwave Communications, Alpena, Mich.
American Mobilphone Co., Inc., Bisbee, Ariz.
American TV Relay Inc., Phoenix, Ariz.
Andrews Tower Rentals Inc., Fort Worth, Tex.
Arizona Microwave System Co., Phoenix, Ariz.
Bell Telephone Co. of Nevada, San Francisco, Calif.
Black Hills Video Corp., Little Rock, Ark.
Brentwood Co., Bakersfield, Calif.
California Interstate Telephone Co., Victorville, Calif.
Carter Mountain Transmission Corp., Cody, Wyo.
Central Microwave Inc., Moscow, Idaho
Cerracche and Co., Inc., Ithaca, N.Y.
Columbia Basin Microwave Inc., Ephrata, Wash.
Columbia Communications Co., New Orleans, La.
Dorate Microwave Corp., Sayre, Okla.
East Texas Transmission Co., Tyler, Tex.
Eastern Microwave Inc., Oneonta, N.Y.
Eastern Shore Microwave Relay Co., Washington, D.C.
Electronics Inc., Vero Beach, Fla.
First TV Corp., Salisbury, Md.
Garden State Micro Relay Inc., Wildwood Crest, N.J.
Golden West Communications, Los Angeles, Calif.
Great Plains Microwave Co., Perryton, Tex.
H & B Microwave Corp., Beverly Hills, Calif.
Hi-Desert Microwave Inc., Burns, Oreg.
High Point Relay Co., Hazleton, Pa.
Houston County Telephone Co., Inc., Crockett, Tex.
Idaho Microwave Inc., Twin Falls, Idaho
Laredo Microwave Inc., Dallas, Tex.
Mesa Microwave Inc., Oklahoma City, Okla.
Micro-Relay Inc., Dublin, Ga.
Microrelay of New Mexico, Inc., Roswell, N.Mex.
Microwave Communications Ltd., Los Angeles, Calif.
Microwave Service Co., Tupelo, Miss.
Mid-Kansas, Inc., Junction City, Kans.
Midland Telephone Co., Moab, Utah
Midwest Microwave Inc., Peru, Ill.
Midwestern Relay Inc., Stevens Point, Wis.
Minnesota Microwave Inc., Willmar, Minn.
Modern Electronics Co., Cleveland, Miss.
Mountain Microwave Corp., Denver, Colo.
Mountain States Tel. and Tel. Co., Denver, Colo.
New England Microwave Corp., Athol, Mass.
New York–Pa. Microwave Corp., Corning, N.Y.
Newhouse Microwave Inc., Syracuse, N.Y.
North Canadian Microwave Co., Inc., Woodward, Okla.
Northco Microwave Inc., Laconia, N.H.
Northern Microwave Service Inc., Caribou, Maine
Pacific Telatronics, Inc., Medford, Oreg.
Penn Service Microwave Co., Hazleton, Pa.
Pilot Butte Transmission Co., Inc., Rock Springs, Wyo.
Potomac Valley Telecasting Corp., Cumberland, Md.
Racom Inc., Lewiston, Maine
Sekan Microwave Inc., Iola, Kans.
Service Electric Co., Mahanoy City, Pa.
Southern Bell Telephone and Telegraph Co., Atlanta, Ga.
Southwest Microwave Inc., Ozona, Tex.
Southwest Texas Transmission Co., Del Rio, Tex.
Southwest Transmission Co., Inc., Fort Myers, Fla.
Superior Communications Co., Inc., Marinette, Wis.
Telecommunications Inc., Seattle, Wash.
Telecommunications of Oregon, Inc., Seattle, Wash.
Telephone Utilities Services Corp., Killeen, Tex.
Teleplex Microwave Systems, Inc., Beverly Hills, Calif.
Teleprompter Transmission of Kansas, Inc., Liberal
Teleprompter Transmission of New Mexico, Inc., New York, N.Y.
Teleprompter Transmission of Oregon, Inc., Eugene
Teleview Networks, Sheridan, Wyo.
Television Microwave Inc., Martinez, Calif.
Tex-Mex Communications Co., Dallas, Tex.
Tower Communications Systems Corp., Coshocton, Ohio
Trans-Muskingum Inc., Parkersburg, W.Va.
United Video Inc., Bloomington, Ill.
Upper Peninsular Microwave Inc., Iron Mountain, Mich.
Video Service Co., Atlanta, Ga.
West Texas Microwave Inc., Denver, Colo.
Western Microwave Inc., Denver, Colo.
Western TV Relay Inc., Elk City, Okla.

Appendix II. Television market rankings

The following cities are ranked according to the American Research Bureau's total net weekly circulation figures for 1965.

1. New York, N.Y.
2. Los Angeles, Calif.
3. Chicago, Ill.
4. Philadelphia, Pa.
5. Boston, Mass.
6. Detroit, Mich.

7. San Francisco, Calif.
8. Cleveland, Ohio
9. Pittsburgh, Pa.
10. Washington, D.C.
11. Baltimore, Md.
12. St. Louis, Mo.
13. Hartford–New Haven, Conn.
14. Providence, R.I.
15. Dallas–Fort Worth, Tex.
16. Cincinnati, Ohio
17. Minneapolis–St. Paul, Minn.
18. Indianapolis, Ind.
19. Atlanta, Ga.
20. Miami, Fla.
21. Seattle–Tacoma, Wash.
22. Buffalo, N.Y.
23. Milwaukee, Wis.
24. Kansas City, Mo.
25. Houston, Tex.
26. Toledo, Ohio
27. Sacramento–Stockton, Calif.
28. Dayton, Ohio
29. Charlotte, N.C.
30. Columbus, Ohio
31. Wheeling, W.Va.–Steubenville, Ohio
32. Tampa–St. Petersburg, Fla.
33. Harrisburg–Lancaster–Lebanon–York, Pa.
34. Memphis, Tenn.
35. Syracuse, N.Y.
36. Portland, Oreg.
37. Albany–Schenectady–Troy, N.Y.
38. Grand Rapids–Kalamazoo, Mich.
39. Birmingham, Ala.
40. Denver, Colo.
41. Johnstown–Altoona, Pa.
42. Nashville, Tenn.
43. New Orleans, La.
44. Greenville–Spartanburg, S.C.–Asheville, N.C.
45. Charleston–Huntington, W.Va.
46. Flint–Saginaw–Bay City, Mich.
47. Lansing, Mich.
48. Louisville, Ky.
49. Greensboro–Winston Salem–High Point, N.C.
50. Raleigh–Durham, N.C.
51. Oklahoma City, Okla.
52. Salinas–Monterey–Santa Cruz, Calif.
53. Manchester, N.H.
54. San Diego, Calif.
55. Norfolk–Portsmouth–Newport News–Hampton, Va.
56. Wichita, Kans.
57. San Antonio, Tex.
58. Tulsa, Okla.
59. Portland–Poland Spring, Maine
60. Omaha, Nebr.
61. Roanoke, Va.
62. Phoenix, Ariz.
63. Salt Lake City–Ogden–Provo, Utah
64. Green Bay, Wis.
65. Richmond, Va.
66. Quad City (Davenport, Iowa–Rock Island–Moline, Ill.)
67. Orlando–Daytona Beach, Fla.
68. Rochester, N.Y.
69. Shreveport, La.
70. Wilkes Barre–Scranton, Pa.
71. Little Rock, Ark.
72. Jacksonville, Fla.
73. Champaign–Decatur–Springfield, Ill.
74. Cedar Rapids, Iowa
75. Mobile, Ala.–Pensacola, Fla.
76. Des Moines, Iowa
77. Spokane, Wash.
78. Springfield–Holyoke, Mass.
79. Jackson, Miss.
80. Knoxville, Tenn.
81. Madison, Wis.
82. Binghamton, N.Y.
83. Columbia, S.C.
84. Columbus, Ga.
85. Baton Rouge, La.
86. West Palm Beach, Fla.
87. Cape Girardeau, Mo.–Paducah, Ky.–Harrisburg, Ill.
88. Evansville, Ind.
89. Greenville–Washington–New Bern, N.C.
90. Sioux Falls, S.Dak.
91. Fresno, Calif.
92. Chattanooga, Tenn.
93. Lincoln–Hastings–Kearney, Nebr.
94. Rockford, Ill.
95. Youngstown, Ohio
96. Augusta, Ga.
97. South Bend–Elkhart, Ind.
98. Peoria, Ill.
99. Fort Wayne, Ind.
100. Albuquerque, N. Mex.

REFERENCES

1. L'Heureux, Robert D., "The history, nature, and scope of CATV," *TV Commun.*
2. Beisswenger, Robert H., "History and status of CATV," *View*, vol. 2, Feb. 1966.
3. *CATV Primer.* Washington: Nat'l Community Television Assn.
4. Collins, Cliff, "Transistorized microwave," *TV Commun.*, Mar. 1966.
5. *Broadcasting Mag.*, p. 83, Mar. 14, 1966.
6. *Television Digest* (*CATV Addenda*), vol. 6:13, Mar. 28, 1966.
7. *Commun. News*, Aug. 1965.
8. Seiden, Martin M., *An Economic Analysis of CATV Systems and the Broadcasting Industry*. Washington: U.S. Gov. Printing Office.
9. Kirtland, C. M., "Financial aspects of CATV," *View*, vol. 2, Feb. 1966.
10. Howman, Ralph W., "CATV economics and microwave," *TV Commun.*, Nov. 1965.
11. Matthews, John D., "Permits and licenses," *View*, vol. 2, Feb. 1966.
12. *Broadcasting Mag.*, p. 38, Feb. 21, 1966.
13. Wasilewski, Vincent T., "Broadcasters see dangerous change in CATV objectives," *Commun. News*, Aug. 1965.
14. Ford, Frederick W., "CATV and the public interest," *Commun. News*, Aug. 1965.
15. "Second report and order," Fed. Commun. Comm., Mar. 8, 1966.
16. *Television Standards and Practice*, D. G. Fink, ed. New York: McGraw-Hill, 1943.
17. Special Issue on Television Allocations, *Proc. IRE*, vol. 48, June 1960.
18. Chipp, R. D., "CATV technical standards," *IEEE Trans. on Broadcasting*, vol. BC-12, pp. 28–30, June 1966.
19. NCTA Comments in Opposition to Proposed Rule Making, exhibit A, docket 15791, Fed. Commun. Comm.
20. Richey, Milford G., "Turnkey CATV system specification," *TV Commun.*, Mar. 1966.

Some current views on holography

Considerable interest has developed recently in the relatively new science of holography, a form of photography in which images are produced without cameras or lenses. Investigations are now being made to determine the suitability of holographic techniques to a number of different applications

Robert J. Collier *Bell Telephone Laboratories, Inc.*

Holography, the science of producing images by wavefront reconstruction, has been receiving increased attention during the past few years, chiefly because of achievements made possible by the development of the laser as a coherent light source. By holographic techniques, three-dimensional objects may be reconstructed as three-dimensional images, in full color, and magnified or reduced in size, as desired. This article describes the basic principles of holography, as well as some of the work being done in this field, including experiments with volume holograms and hologram interferometry.

A number of fine review articles on holography[1-3] have succeeded in presenting and explaining the properties of holograms in a manner stimulating to readers with even the most casual interest in science. We read about images having three-dimensional properties,[4-6] magnification obtained by reconstructing with a wavelength greater than that used in forming the hologram,[4,5] diffuse[6] holograms which, even when broken, produce whole images, multicolor images[7] obtained from emulsions which normally produce only black and whites; in fact, there seems to be something for everyone from adman to zoologist. But the tale has not yet run out, and work in holography during the past year can perhaps best be explained by retelling the story with a shift in emphasis. Through a more general approach to the basic concepts of holography, the newer work as well as some of the older, well-known, often conveniently ignored experimental effects can be made more plausible. The experimental work described in this article should not be construed to represent a cross section of the results obtained by the many workers in the field; it is, however, the work with which the writer is most familiar, having been performed at Bell Telephone Laboratories.

Some definitions

Figure 1 indicates a general arrangement for forming a hologram. Two arbitrary objects, 1 and 2, are illuminated by parallel laser light. In the general case, the light reflected by these objects will be diffuse and the reflected wavefronts will proceed to interfere in the photosensitive medium where the interference pattern can be recorded. After the photosensitive medium has been exposed and processed it is called the *hologram*,[4] which may be defined as the recorded interference pattern of two or more coherent wavefronts. When the hologram is illuminated by one of the *original* wavefronts used to form it, the remaining wavefronts are reconstructed. Thus in Fig. 2 the hologram has been replaced in its exact original position, object 1 and the laser beam have stayed as they were during the forming of the hologram, and only object 2 has been removed from the scene. Under these conditions, the wavefront from object 1 illuminating the hologram is the original one. Then if we look through the hologram (from the right in the figure) in the direction of the former location of object 2, we can observe the reconstructed wavefront from object 2. Observation of these reconstructed wavefronts is nearly equivalent to observing the objects from which they were originally derived. The hologram recording technique records not only the spatial variation of the amplitude of wavefronts coming from the object but also the spatial variation of phase. This means that not only information on light and dark areas of an object are recorded but also its topology.

A hologram isn't much to look at; in many cases, it is merely a smudged or darkened photographic plate. (There is a fine grating structure to this smudge, however, as one can see by holding the plate up to a white-light lamp and, with a little practice, observing a rainbow-type spectrum.) The hologram record is formed as a permanent change in the transmittance of a photosensitive medium caused by the incident light interference pattern. The change in transmittance may be an increase in the absorption of the medium, as when a photographic emulsion is exposed and developed; or, if the emulsion were bleached, the record might be the resulting localized changes in the index of refraction of the medium. For most photosensitive media, an exposure and development process can be chosen to make the transmittance change proportional to the *intensity* of the incident light pattern. In the general case, the interference pattern is recorded throughout the *volume*[7-10] of the medium; only in certain cases can the record be considered essentially planar.

The elementary hologram

The simplest hologram results from the interference pattern produced by two plane waves derived from the same coherent source. In Fig. 3 the directions of the wave normals corresponding to plane wave 1 having an amplitude A_1 and plane wave 2 having amplitude A_2 are given by the arrows. The wavefront planes, F_1 and F_2, are

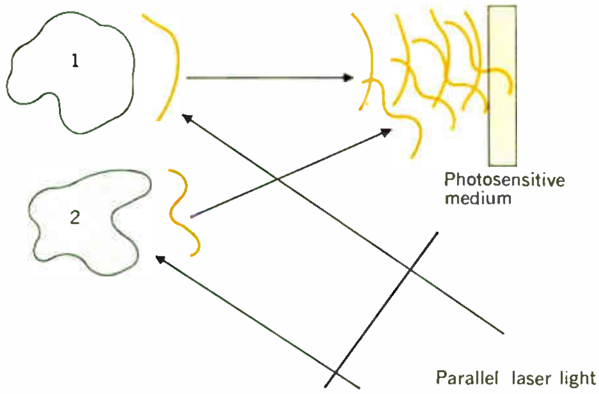


Fig. 1. General arrangement for hologram formation.

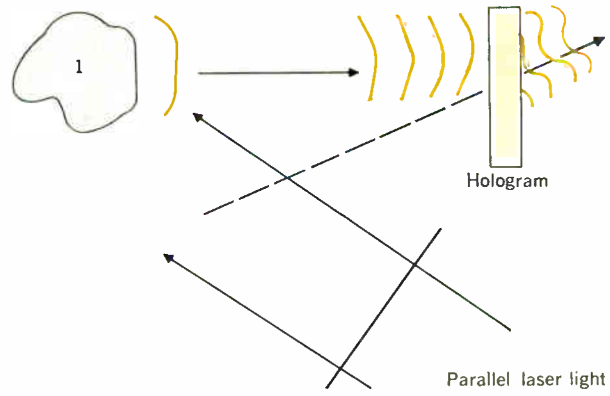


Fig. 2. Wavefront reconstruction.

normal to these directions. If we assume that the wavefront planes drawn represent the positive maximum amplitude or antinodal fronts, then they will be spaced periodically by the wavelength λ. The planes F_1 and F_2 intersect in the photosensitive medium of thickness t, the intersection being a line perpendicular to the plane of the figure. As the waves progress in the direction of their wave normals, these lines will generate planes in the medium (seen edge-on in the figure). Thus for the case of the antinodal fronts being considered, planes of maximum light amplitude will be generated in the photosensitive medium by the interference of the two plane light waves.[8] These generated planes are perpendicular to the figure, and some simple geometry using the figure shows that they bisect the angle θ between the wave normals, are parallel to each other, and occur periodically throughout the exposed medium. Of course, planes intermediate to these will be generated in the medium by the intersection of wavefronts not in phase. If the photosensitive medium were a photographic emulsion, the exposure and development process would result in silver atoms developed out over the surface of the generated planes in proportion to the time average of the square of the amplitude, or intensity, of the light in these planes. This would result in a more or less sinusoidal density variation in the medium, as indicated by the density of vertical lines in Fig. 3. This sinusoidal density variation is the hologram recording of the elementary interference pattern. Examination of the heavy-lined triangle in Fig. 3 shows that the period y of the density variation is given by

$$2y \sin \frac{\theta}{2} = \lambda \qquad (1)$$

which can be recognized as having the form of Bragg's law. When the elementary hologram is illuminated for reconstruction purposes, Bragg's condition must be obeyed by the incident and the exiting light. It should be noted that the angle $\theta/2$ is the grazing angle made by the wave normals with the generated planes and does not necessarily refer to the photosensitive surface normal, which need not coincide with these planes.

In order to see that these concepts may be extended to more complicated interference patterns, consider the illumination of a 35-mm slide by parallel laser light. The slide may be characterized by its transmittance variation over the slide surface. This complicated function may be Fourier-analyzed into a sum of periodic transmittances, or, in other words, into a sum of sinusoidal transmission gratings. Each sinusoidal grating component will diffract the incident parallel light into plane parallel waves in characteristic directions. Thus, the arbitrary wavefront, such as that transmitted by the slide, can be considered to be a sum of plane waves, which will interfere in a manner related to the simple hologram formation.

Response difference between volume and planar recording

We have shown the simple hologram as being formed through the volume of the photosensitive medium. In Fig. 3, if the angle θ between the interfering wavefront normals is small enough, the periodic distance y can be large compared with the medium thickness. Here the recording can be considered to be a surface transmission grating. There is an essential difference in response to incident light between the surface grating and the grating of Bragg planes extending through a volume. In the case of a surface grating, coherent parallel light incident at some arbitrary angle to the grating will result in diffracted parallel beams at angles determined only by the condition that the transmitted light must constructively interfere. In the case of the volume grating, however, the incident and exiting light must obey the equiangle law of reflection with respect to the silver planes generated in the emulsion as well as meet the condition for constructive interference, in order for a directed light output to be obtained. This is just another way of stating Bragg's law, as given in Eq. (1).

From Fig. 3 it is evident that either of the original forming beams can satisfy the Bragg condition, since each makes the Bragg angle with the generated planes. In fact, if we do illuminate the hologram formed as in Fig. 3 with the original plane wave 1, light will be reflected from the Bragg planes (diffracted) in the direction of the other original plane wave 2 so that incident and exiting light both satisfy Bragg's law. A similar result holds for illumination by wave 2. Another pair of light waves, those traveling in the antidirections of plane waves 1 and 2, will also satisfy the Bragg condition. These latter waves may be considered the conjugates of the original waves, essentially retracing the original ray paths. In this case, the conjugate of wave 1 illuminating our simple hologram will result in the reflection (diffraction) of a wave in the direction of the conjugate of wave 2, so that again the grazing angle has the Bragg value for light into and out of the hologram. For constant λ and y, only the original wave directions and their conjugates will exactly satisfy the first-order Bragg condition.

Basic holography equations

It may be helpful to formulate the ideas we have been developing for the general case, in which objects 1 and 2 are diffuse reflectors and the hologram is formed through the volume of the photosensitive medium. Keeping in mind the general hologram-forming arrangement shown in Fig. 1, let us assume the complex amplitude of the wavefront arriving at the photosensitive medium from object 1 to be A_1 and that from object 2 to be A_2. The amplitude of the interference pattern is their sum

$$a = A_1 + A_2 \quad (2)$$

The change in transmittance of the medium will be proportional to the intensity

$$aa^* = (A_1 + A_2)(A_1^* + A_2^*)$$
$$= A_1A_1^* + A_2A_2^* + A_1A_2^* + A_1^*A_2 \quad (3)$$

The first term $A_1A_1^*$ is the intensity I_1 of the light arriving from object 1 while the second term $A_2A_2^*$ is the intensity I_2 of the light arriving from object 2. If the objects, in the general case, are diffuse reflectors, then the unfocused intensities I_1 and I_2 can be regarded as fairly uniform over the hologram surface and thus can be considered constants. Therefore, we can write for the change in transmittance ΔT, which is the hologram record,

$$\Delta T \propto I_1 + I_2 + A_1A_2^* + A_1^*A_2 \quad (4)$$

Suppose we wish to reconstruct the wavefronts from object 2 by illuminating the hologram with the wavefront from object 1. The light amplitude transmitted through the hologram will be the product of the incident amplitude A_1 times the transmittance of the hologram. We shall be interested in only that portion of the transmittance of the photosensitive medium modified by the interference pattern recorded during hologram formation. Therefore, the diffracted or reconstructed wavefronts will be given by the product

$$R = A_1\Delta T = A_1(I_1 + I_2) + A_1A_1A_2^* + I_1A_2 \quad (5)$$

Recalling that I_1 and I_2 are constants, we see that the first term is a reconstruction of a wavefront proportional to the original wavefront from object 1. The second term, where A_1 is a diffuse light wavefront, will not give a distinct image and will not satisfy the Bragg condition; it can best be regarded as a noise term. The third term is the interesting one because it indicates that a wavefront proportional to the original one from object 2 has been reconstructed. Since this wavefront is one of the original forming waves, it will satisfy Bragg's condition. Looking through the hologram, the reconstructed wavefront diverges from a ghost image of object 2, which appears to be located exactly where object 2 originally was. The image is called *virtual* because the reconstructed waves appear to *diverge* from it. It should be noted that illuminating the hologram with the original waveform A_1 implies that it will strike the hologram exactly as it did during formation. This in turn means that the spatial

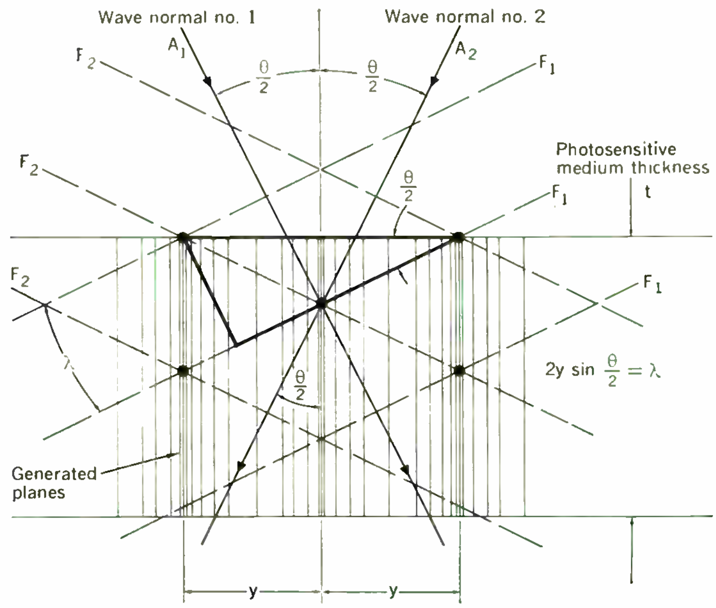
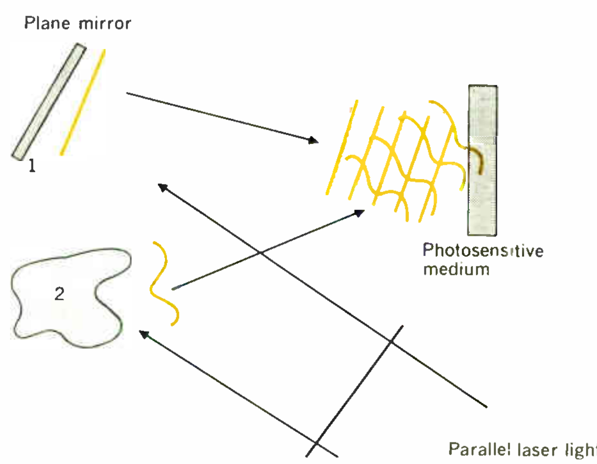


Fig. 3. Elementary hologram.

Fig. 4. Hologram formation with object 1 a plane mirror.



relations that existed between object 1, the hologram, and the laser beam, during formation of the hologram must be maintained during reconstruction.

An image of object 2, called the *real image*, may be produced by illuminating the hologram with the conjugate waveform from object 1. (If the original wavefront diverges from an object, the conjugate waveform *converges* to a real image.) In this case the reconstruction is given by

$$R' = A_1^*\Delta T = A_1^*(I_1 + I_2) + I_1A_2^* + A_1^*A_1^*A_2 \quad (6)$$

The first term is proportional to the conjugate waveform from object 1 and the second is a reconstruction of the conjugate waveform from object 2. This term leads to the real image of object 2. Note that both A_1^* and A_2^* satisfy Bragg's angle condition but that the third term cannot satisfy the condition. Also in this term, factors $A_1^*A_1^*$, being diffuse wavefronts, will serve to prevent any distinct image of object 2 from being formed. This process of obtaining the real image will be discussed next for a special case of object 1.

Object 1 as a plane mirror

Up to this point we have not limited the kind of objects used in the hologram formation. Let us now consider the important special case in which object 1 is a plane mirror. As shown in Fig. 4, the wavefront from object 1 will now be a plane wave and is what has been normally called the "reference" beam. We can proceed, of course, as before and illuminate the hologram with the original reference wave from object 1 to obtain the reconstructed terms. In this case, however, it is really not necessary to use the same plane wave as in the formation of the hologram; one plane wave is as good as another. Thus, if the surface of the hologram originally facing the objects during formation is called the front surface, the virtual image can be obtained by illuminating this surface with any parallel laser beam and rotating the hologram to satisfy the Bragg angle condition. As shown in Fig. 5, the reconstructed wavefront will diverge from the virtual image of object 2 and proceed in a direction relative to the plane wave as in formation.

Illuminating the back surface of the hologram with a parallel laser beam is in effect illuminating with the conjugate of the original plane wave. Again, as in the case of front-surface illumination, any parallel beam will do. As the hologram is rotated so as to make the illuminating beam incident at the Bragg angle, the conjugate waveform of object 2 will be reconstructed and will converge to the real image. If the parallel illumination is antiparallel to the original forming plane wave then the real image will lie exactly where object 2 had been.

The advantages of using a plane wave as one of the forming wavefronts are that (1) no critical distances need be maintained from formation to reconstruction, (2) a simple rotation of the hologram can locate the Bragg angle, and (3) the intensity of the illuminating wave is nearly constant over the hologram.

We should also note the particular special case in which object 1 is a plane mirror and the hologram recording is essentially planar. For this case the second term in Eq. (5) for R will yield the conjugate wave from object 2, so this wave and the reconstructed original wave from object 2 will proceed in directions symmetrically disposed about the illuminating beam direction. This case, however, has already received more than its due attention, and even for modest angles between forming beams the real image is better obtained by illuminating the back side of the hologram if high resolution is required. This occurs because the effects of the medium thickness (i.e., the Bragg condition) can nearly always be detected to some degree, and the image will not be reconstructed exactly unless the Bragg condition is met.

Association

The general-case hologram formation process can be considered to be an example of a rudimentary associative memory. In the hologram, objects 1 and 2 are recorded in association. When, under the necessary conditions for reconstruction, object 1 serves to illuminate the hologram, the associated member of the pair, object 2, is imaged and vice versa; Fig. 6 illustrates this point. The left and center photos[11] were made with room light as well as laser light illumination and show the subject scene

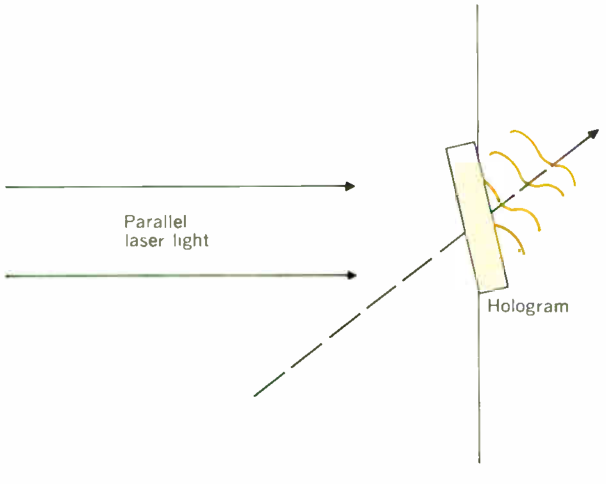
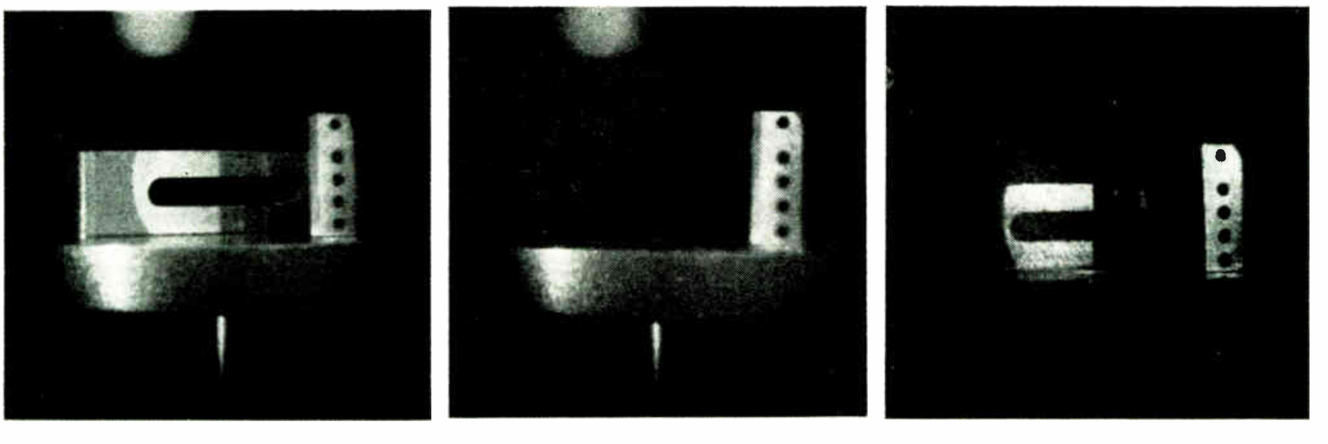


Fig. 5. Reconstruction with a plane wave.

Fig. 6 (below). Ghost imaging in the near field. Shown are the original subject scene (left), the illuminating object (center), and the ghost image (right). The front surface of the vertical bar and the brighter portions of the horizontal bar are the only parts of the subject scene illuminated with laser light; these parts alone participate in the hologram process. The background noise under the ghost image of object 2 (observable as well in Figs. 8 and 9) results from the fact that I_1 in Eq. (5) is not truly a constant. If light from object 1 is diffuse, the noise is fine-grained; if it is uniform (object 1 a mirror), the noise disappears.



before and after hologram formation. The photo on the left shows two brass bars: a vertical bar with small holes, which is designated object 1; and a horizontal bar with a slot, which is designated object 2. The laser illumination is coming from the right and falls on the vertical bar and on the brighter areas of the horizontal bar. In fact, the shadow of the vertical bar cast by the laser light can be seen on the horizontal one. Only the laser-illuminated portions of this scene will serve as the subject of the hologram. The hologram is formed and replaced exactly in position, and object 2 (the horizontal bar) is removed from the scene as shown in the center photograph in Fig. 6. Only object 1 is left to illuminate the hologram. The resulting ghost image of the laser-illuminated portions of object 2 and the image of object 1 can be observed in the reconstruction shown in the right-hand photo of Fig. 6. The shadow noted earlier is preserved in the ghost image.

If object 1 were to be moved from its original position, the ghost image of object 2 would disappear. In certain applications—for example, character recognition—this disappearance is not desirable. An invariance with respect to translation of object 1 can be obtained by forming what is known as a Fourier-transform hologram.[12]

The Fourier-transform hologram

Up to this point we have been discussing holograms made in the *near* field of the subject; these are called *Fresnel* holograms. The Fourier-transform hologram records, by use of the hologram technique, the Fourier transform of a subject placed in an input plane. When forming such holograms one generally is dealing with planar subjects such as a transparency; one can then speak of an input plane, where the subject is placed, and also of a Fourier-transform plane, where the transform is displayed and recorded. The Fourier-transform hologram has the desirable property that if objects 1 and 2 are the subjects of the hologram, object 1 can be translated in the input plane and can still reconstruct the wavefront of object 2 upon illuminating the hologram. Thus, if object 1 were the letter *A* and object 2 some suitable recognition signal, then the letter *A* would reconstruct its recognition signal despite the fact that *A*'s location in the input plane might differ from the one it occupied during formation.

The Fourier transform of a subject can be produced by a lens. Consider a slide illuminated by parallel coherent light and backed by a lens, as shown in Fig. 7. The plane-wave illumination insures that the light transmitted by the slide will be proportional to the slide transmittance. Further, each Fourier component of the slide transmittance will diffract the parallel incident light as parallel light in a characteristic direction or, in current terminology, with a characteristic *spatial frequency*. Now the lens has the property of focusing parallel light to a point in its focal plane and thus will gather all the light transmitted through the slide emitted in *one* direction (i.e., with one spatial frequency) and plot it as a single point on its focal plane. This summation of the light amplitude diffracted with one spatial frequency is the Fourier transform, which is plotted, as a function of spatial frequency, by the lens on the focal plane. The Fourier transform can be recorded by placing the photosensitive medium at the focal plane of the lens. When object 1 is allowed to illuminate the completed hologram through the lens, the Fourier transform of object 2 will be reconstructed.

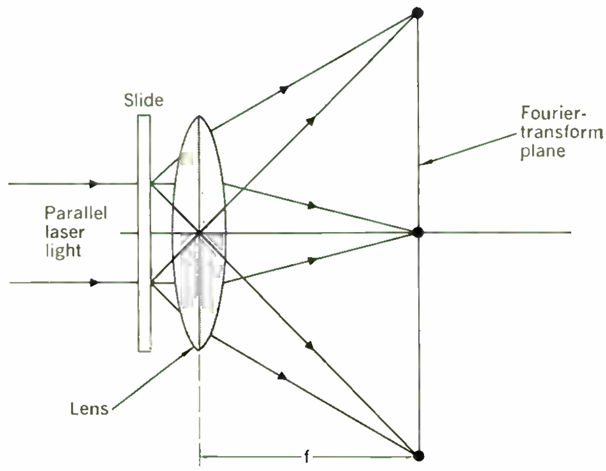


Fig. 7. Fourier transformation by a lens.

A real image of object 2 can be obtained by the use of an auxiliary lens to perform the inverse transformation, or the eye itself, viewing from a sufficient distance, can accomplish the same thing.

Figures 8 and 9 are examples of Fourier-transform ghost imaging.[13] In Fig. 8 the subject of the hologram is shown at the top. The word DEMODULATOR can be arbitrarily chosen to be object 1 and the letters FM can be chosen to be object 2. When light passing through DEMODULATOR illuminates the completed hologram, the ghost image of FM is obtained. The bottom photos in Fig. 8 are of the images obtained when DEMODULATOR was translated about ½ inch across the input plane. The left side of Fig. 9 is again a photo of the ghost image of the same subject, but in this case the subject slide was illuminated by laser light passing through a diffusing ground-glass screen. The result was to make the noise under FM somewhat more uniform. The photo on the right of Fig. 9 is of the hologram-produced image of a continuous-tone slide. The upper portion of the head and hat were considered to be object 1. When only this portion was allowed to illuminate the completed hologram, the ghost image of object 2, the facial features, was observed and photographed.

Experiments with volume holograms

The first experiment to demonstrate the usefulness of the volume hologram was the formation of the multicolor hologram by K. S. Pennington and L. H. Lin of Bell Laboratories.[7] They took advantage of the Bragg condition to insure that there was no color crosstalk. In the formation process object 1 was a mirror and object 2 a color slide, and the laser beam was a beam containing blue light from an argon-ion laser and red light from a helium–neon laser. The angle θ between the light passing through the slide and the plane-wave mirror beam or reference beam was made large (of the order of 90°) to insure the formation of a high-frequency grating of Bragg planes. The photographic plate, which was the sensitive medium, could then be considered to be a thick medium. When the completed hologram was replaced and illuminated by the blue and red beam coming from the mirror and incident at the Bragg angle, then the multicolor image photographed in Fig. 10 was observed; red, blue, and nearly white shades are all maintained in register. The Bragg condition insured that the grating originally produced by the blue part of the forming beam would not reconstruct spurious images when illuminated by the red part, and vice versa.

More recently, workers at the University of Michigan[14] have shown that when the reference beam and the subject beam are nearly 180° apart (the wave from object 2 is incident on the front surface of a photographic plate while the plane-wave reference is incident on the back surface) the images can be reconstructed by white-light illumination. Since, as we have shown earlier, the Bragg planes must bisect the angle between the interfering waves, the planes generated in this hologram will be nearly parallel to the surface. The reconstructing beam incident at the Bragg angle on the back of the hologram will reflect from these Bragg planes, giving rise to the reconstructed waveform of object 2 coming out the same back side. Thus such wide-angle holograms may loosely be termed *reflection* holograms (although all volume holograms involve reflection from the Bragg planes). The hologram behaves as a very-high-selectivity filter, and when white light is shown onto the hologram at roughly the Bragg angle only the proper colors and angles are utilized by the hologram to reconstruct the object 2 waveform. The remainder are rejected by the filter and cannot reconstruct any interference. As pointed out by Pennington, the selectivity of any Bragg filter is characterized by the dispersion relation

$$\frac{d(\theta/2)}{d\lambda} = \frac{1}{\lambda} \tan \frac{\theta}{2} \qquad (7)$$

obtained by differentiating Bragg's law, given in Eq. (1). In the experiment under consideration, $\theta/2$ is tending toward 90°, where the tangent is very large. This strong variation of Bragg angle with wavelength tends to reject all of the wavelengths in the white-light illuminating beam except the proper ones.

Figure 11 is a photograph of a multicolor reflection hologram image reconstructed by white-light illumination[15] and demonstrates two of the advantages associated with volume hologram formation. The result, obtained at

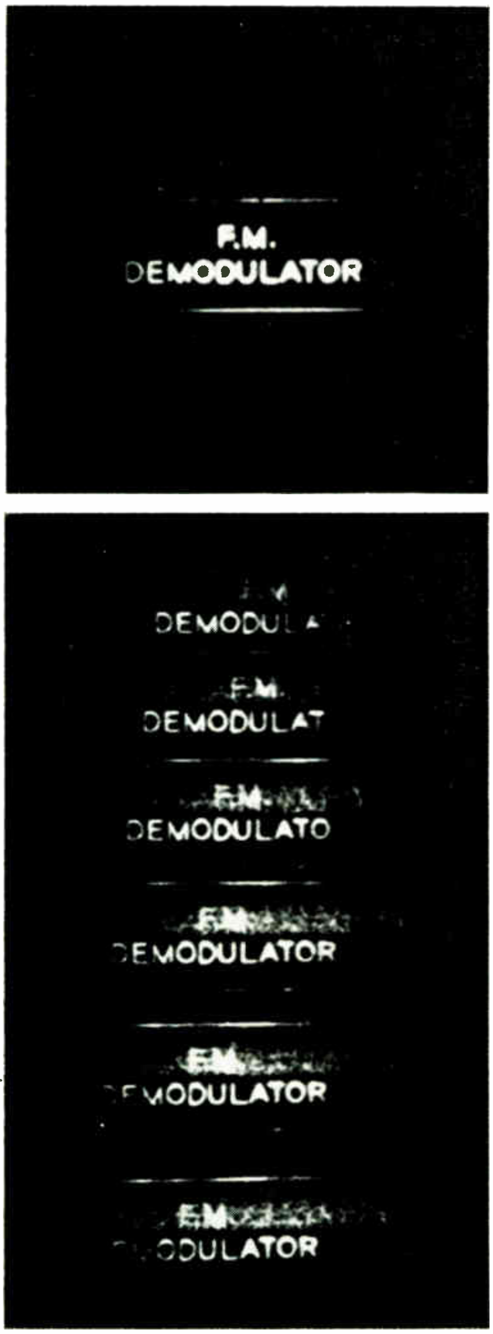


Fig. 8. Ghost imaging with the Fourier-transform hologram. The original subject scene was a slide bearing transparent letters on an opaque background and is shown in the photograph at the top. At the bottom are photos of ghost images of FM (object 1) produced by illuminating the finished hologram with light passing only through the DEMODULATOR portion (object 2 of the slide). The six frames show FM reconstructed from the same hologram when the illuminating object DEMODULATOR is translated through six different positions in the input plane. Note that FM remains in register with DEMODULATOR.

Fig. 9. Ghost imaging using the Fourier-transform hologram and diffuse illumination. At the left is a photograph of the ghost image of the subject of Fig. 8. To make the background noise more uniform, a ground-glass screen was interposed between the laser light source and the slide, resulting in diffuse illumination of the slide. FM is again reconstructed when light passing only through DEMODULATOR illuminates the hologram. At the right is a photo of the ghost image of a continuous-tone picture, again using a ground-glass screen to provide diffuse illumination. The light passing only through the upper portion of the slide and illuminating the hologram reconstructs the remainder of the slide.

the Bell Telephone Laboratories, is a consequence of combining the two techniques described. Multicolor holograms can now be viewed in the home by those homeowners not yet possessing argon and helium–neon lasers.

Hologram interferometry

The hologram can be used as a unique type of interferometer, capable of detecting differential motions of the order of $\lambda/10$ in diffuse reflecting irregular surfaces.[16-19] The visible interference fringes observed on the surface of a thermoelectric device, used as subject, are shown in Fig. 12. One surface of a thermoelectric device was illuminated, as object 2, while a mirror served as object 1. The developed hologram was replaced; the mirror, object 2 (the thermoelectric device), and the laser beam all remaining as in the forming process. Upon looking through the hologram one sees the reconstructed virtual image of the thermoelectric device cast exactly back on the device itself. The light from the device and from its virtual image can interfere; this will result in interference fringes observable in real time if any differential motion perpendicular to the surface of the device subsequently takes place. This same effect can be observed by means of a double-exposure hologram technique; the photosensitive medium is first exposed to the illuminated device and the reference beam, then covered while voltage is applied to cool the subject surface. A second exposure is taken of the resulting slightly domed surface to give two superimposed holograms. When the plane-wave reference beam illuminates the developed hologram, the two virtual-image reconstructed wavefronts will interfere and produce the fringes observed in the figure.

Conclusion

The preceding discussion has passed over many interesting theoretical and experimental aspects of holography (e.g., hologram imaging properties[20-22]), which have been the subject of current papers. As a result there is an abundance of material waiting for other reviewers of this bounding subject. The present review has, I believe, placed the emphasis where it currently should be, that is, on the

Fig. 10. Photograph of the first multicolor hologram image. The hologram was formed with a beam containing red and blue laser light, and the reconstruction was obtained with the same beam. The effect of a dust particle can be observed near the boy's cheek. Some granularity is introduced into the image, probably by the diffuse screen used in the hologram formation.

Fig. 11. Photograph of the first multicolor hologram image reconstructed with white-light illumination. The forming beam again combined red and blue light, but the image from the reflection hologram formed for this case can be reconstructed by illuminating the hologram with any white-light source ranging from flashlight to sunlight. The colors in the reproduction of the image have shifted toward the short-wave end of the spectrum. The shift depends on such factors as emulsion (which causes the Bragg-plane spacing to change), spectral content of the illuminating source, observation angle, and color film used to photograph the image. The pinholes were present in the original hologram photographic emulsion.

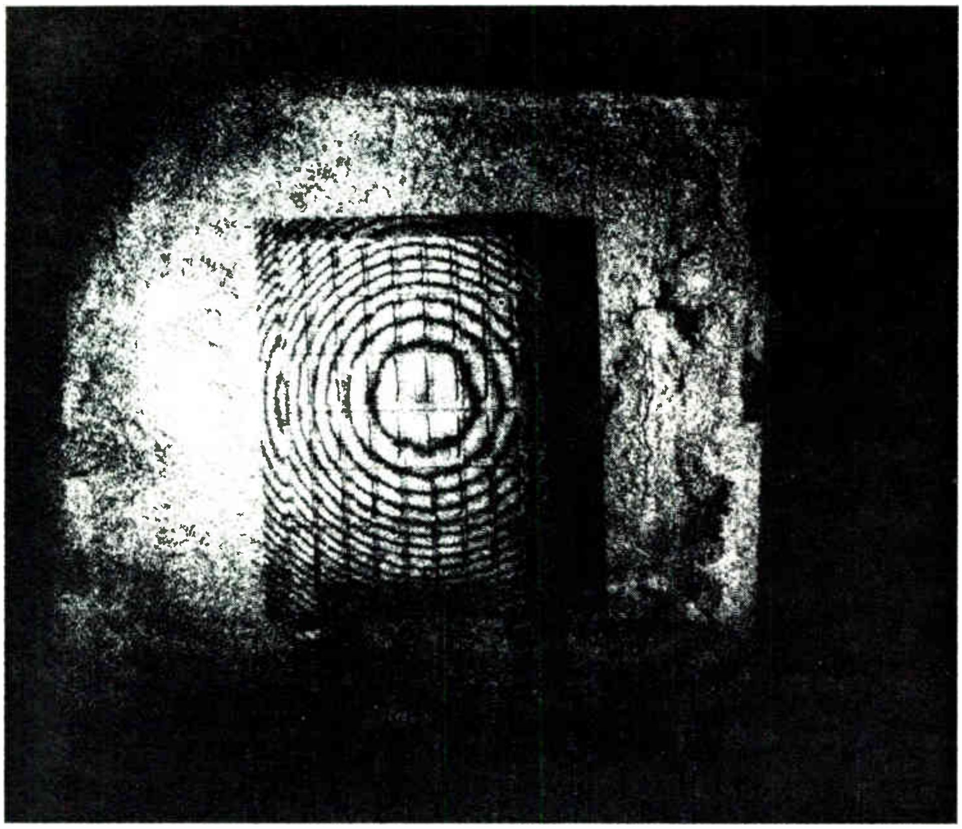


Fig. 12. Interference fringes on a thermoelectric device surface produced by hologram interferometry. The photograph is of the superimposed images and resulting interference fringes reconstructed from a hologram formed by the double-exposure technique. The normal differential displacement of the surface is of the order of λ/2 per dark ring. The surface appears to have domed in a symmetrical manner. The displacement of the cooled surface relative to the room-temperature surface is greatest at the edges, where the frequency of the fringes is highest.

volume hologram and on the idea of information stored in associated pairs. It remains to be seen whether van Heerden's concept of a high-density associative memory[8] can be realized through the use or partial use of holograms.

REFERENCES

1. Stroke, G. W., "Lensless photography," *Internat'l Sci. Tech.*, pp. 52–60, May 1965.

2. Leith, E. N., and Upatnieks, J., "Photography by laser," *Sci. Am.*, vol. 212, pp. 24–35, June 1965.

3. Pennington, K. S., "How to make laser holograms," *Microwaves*, vol. 4, pp. 35–40, Oct. 1965.

4. Gabor, D., "Microscopy by reconstructed wave-fronts," *Proc. Roy. Soc. (London)*, ser. A, vol. 197, pp. 454–487, July 7, 1949.

5. Rogers, G. L., "Experiments in diffraction microscopy," *Proc. Roy. Soc. (Edinburgh)*, sect. A, vol. 63, pp. 193–221, 1952.

6. Leith, E. N., and Upatnieks, J., "Wavefront reconstruction with diffused illumination and three-dimensional objects," *J. Opt. Soc. Am.*, vol. 54, pp. 1295–1301, Nov. 1964.

7. Pennington, K. S., and Lin, L. H., "Multicolor wavefront reconstruction," *Appl. Phys. Letters*, vol. 7, pp. 56–57, Aug. 1, 1965.

8. van Heerden, P. J., "Theory of optical information storage in solids," *Appl. Opt.*, vol. 2, pp. 393–400, Apr. 1963.

9. Denisyuk, Yu N., "Representation of optical properties of an objective by means of wave pattern of light scattered by it," *Soviet Phys. Doklady (English Transl.)*, vol. 7, pp. 543–545, Dec. 1962.

10. Denisyuk, Yu N., "On the reproduction of the optical properties of an object by the wave field of its scattered radiation," *Opt. Spectr. USSR (English Transl.)*, pt. I in vol. 15, pp. 279–284, Oct. 1963; pt. II in vol. 18, pp. 152–157, Jan. 1965.

11. Collier, R. J., and Pennington, K. S., "Ghost imaging by holograms formed in the near field," *Appl. Phys. Letters*, vol. 8, pp. 44–46, Jan. 15, 1966.

12. Vander Lugt, A., "Signal detection by complex spatial filtering," *IEEE Trans. on Information Theory*, vol. IT-10, pp. 139–145, Apr. 1964.

13. Pennington, K. S., and Collier, R. J., "Hologram-generated ghost image experiments," *Appl. Phys. Letters*, vol. 8, pp. 14–16, Jan. 1, 1966.

14. Stroke, G. W., and Labeyrie, A. E., "White-light reconstruction of holographic images using the Lippman–Bragg diffraction effect," *Phys. Letters*, vol. 20, pp. 367–369, 1966.

15. Lin, L. H., Pennington, K. S., Stroke, G. W., and Labeyrie, A. E., "Multicolor image construction with white-light illumination," *Bell System Tech. J.*, vol. 45, pp. 659–661, Apr. 1966.

16. Burch, J. M., "The application of lasers in production engineering," *Production Engr.*, vol. 44, pp. 431–442, Sept. 1965.

17. Collier, R. J., Doherty, E. T., and Pennington, K. S., "Application of moiré techniques to holography," *Appl. Phys. Letters*, vol. 7, pp. 223–225, 1965.

18. Brooks, R. E., Heflinger, L. O., and Wuerker, R. F., "Interferometry with a holographically reconstructed comparison beam," *Appl. Phys. Letters*, vol. 7, pp. 248–249, 1965.

19. Stetson, K. A., and Powell, R. L., "Interferometric hologram evaluation and real-time vibration analysis of diffuse objects," *J. Opt. Soc. Am. (Letters)*, vol. 55, pp. 987–992, Aug. 1965.

20. Leith, E. N., Upatnieks, J., and Haines, K. A., "Microscopy by wavefront reconstruction," *J. Opt. Soc. Am.*, vol. 55, pp. 981–986, Aug. 1965.

21. Meier, R. W., "Magnification and third-order aberrations in holography," *J. Opt. Soc. Am.*, vol. 55, pp. 987–992, Aug. 1965.

22. Armstrong, J. A., "Fresnel holograms: their imaging properties and aberrations," *IBM J. Res. Develop.*, vol. 9, no. 3, pp. 171–178, May 1965.

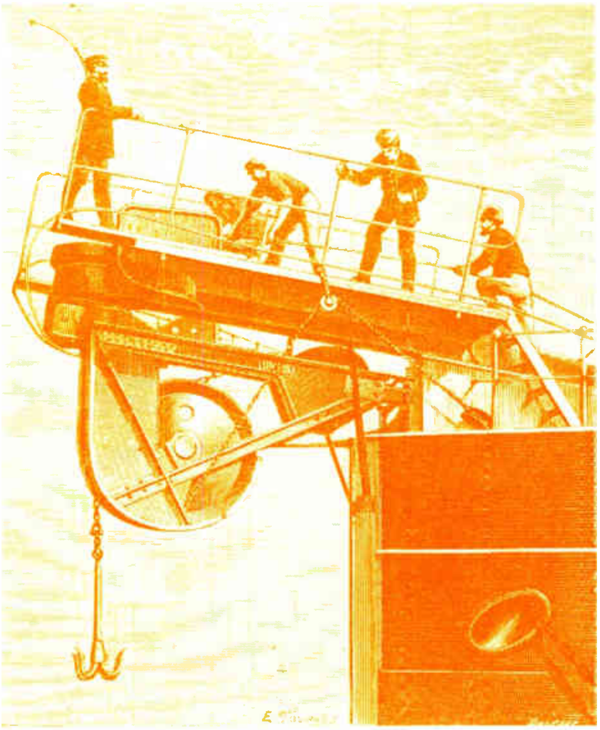


Fig. 1. Bow sheave of a large 19th century cable ship.

Submarine dc cables

Although the use of direct current had given way to alternating current by the early 1900s, recent advances in electronic component design have aroused new interest in dc transmission. This development, however, has also entailed a number of problems

J. M. Oudin *Cables de Lyon*

R. A. Tellier *Electricité de France*

Since 1900 most of our power has been supplied by alternating current, but the high-voltage electron tubes and the rectifying equipment developed during the last decade have revived our interest in direct current. Because geographical factors have made most of our present-day dc cables submarine, we find that we are beset by the same problems that were involved with their historical antecedents—transatlantic telegraph, and later telephone, cables. Mechanical problems arise because of the complex structure of a cable, and because of all of the stresses to which it is subjected. There is also the problem of electrical stress. A comparison of the various types of cable shows that rugged, lightweight equipment is desirable, and that taped insulation insures reliability.

Although power cables were first developed during a period when electricity was still being supplied by low-voltage direct current, ac distribution rapidly became predominant after the introduction of the transformer by Gaulard and Gibbs and the early paper-insulated cables were designed for ac operation, which quickly reached the 10 000-volt level.[1] Consequently, dc cables were in very limited use after the turn of the century.

Aside from some low-voltage distribution systems[2] and electric railway systems, the only dc schemes in operation during the first half of the 20th century seem to have been the 150-kV Moutier–Lyon (France) system, which remained in service from 1906 to 1937, and the Willesden–Ironbridge (Great Britain) 100-kV ring main, which gave good service for many years after its installation in 1910. Both systems involved several miles of paper-insulated lead-sheathed cables.[1] Nevertheless, during the past 12 years dc transmission has aroused renewed and increasing interest in the electrical world as a result of the possibilities offered by the newly developed high-voltage electron tubes and associated equipment for rectifying ac voltage.

For geographical reasons, however, all dc schemes installed during this recent period have involved long submarine crossings. Here the use of insulated cables is implied, since overhead transmission is not possible and charging current effects rule out ac transmission. As a matter of fact, dc submarine cables may have economical and technical advantages over ac cables for shorter crossings also (asynchronous operation of two interconnected systems, limitation of short-circuit current, etc.). Present-day dc cables are therefore mainly submarine, and their development has entailed mechanical problems similar to those involved with telecommunication submarine cables, as well as the electrical problems common to both land and submarine designs.

Transmission of information and energy

Similar cylindrical combinations of conducting metal and dielectric material, manufactured on the same type of machinery, are used for transmitting either information or electric energy. This problem of transmission is assuming a growing importance in today's world, where our needs for power and communication facilities double practically every ten years.

From the historical viewpoint, the transmission of information preceded that of energy, since it was more than a century ago that the first long submarine telegraph circuits were laid on the deep ocean beds.[3] The information carried by these telegraph cables was very "poor," and the attenuation over long distances consequently was limited. The transmission process was, to a certain extent, comparable to a dc process—insofar as direct current can transmit information. Although the concept of telephone communication over transatlantic cable dates back to the 19th century, it was only during the past decade that it was actually achieved as the result of important advances in the design of electronic components.[4] With a view to increasing the amount of information to be transmitted on one conductor, the frequency bandwidth has been raised from a few tens of cycles (telegraph cables) to some 10^6 c/s (telephone cables). This broadening of the frequency band has entailed a tremendous attenuation of the signal (some 10^{900} for a transatlantic cable at the upper frequency); the attenuation can be compensated for only by suitable means of amplification distributed along the cable. This important step forward was made possible by the diode and its later adaptation—the electron tube with grid control—which can be used for both amplification and modulation. In addition, the old telegraph cables have been replaced by polyethylene-insulated cables, with high insulation resistance and low permittivity, and in recent years deep-sea lightweight cables with steel-reinforced conductor have been introduced.

Submarine telecommunication cables, which so far have proved more reliable than high-frequency radio for long-distance communication, may even offer some advantage over the possible zenithal satellites, as they can use the shortest distance between two points of earth and do not introduce delay in transmission of a telephone conversation.

With regard to the transmission of electric energy, this development did not appear until some time after the introduction of communications via cable. At the end of the 19th century rotating machines were available to generate, from mechanical energy, electric energy for local lighting circuits and electric motors. It should be noted here that, in the beginning, these machines paradoxically were used to produce direct current in spite of the difficulties arising as a result of the mechanical commutation involved. But at that time engineering thinking was probably unconsciously still dominated by the Volta cell, the first low-power generator, which incidentally had been used for supplying the electric current necessary to carry the first messages over a distance.

The first problems to be solved by cable engineers were related to distribution rather than to long-distance transmission.

In principle, direct current is difficult to distribute economically; a constant current distribution would entail overinsulation of the receiving equipment, whereas a constant voltage distribution would necessitate oversized conductors. For a given power, there exists in fact an optimum economical relationship between voltage and current; this relationship makes it necessary to reduce the voltage as the load is divided among smaller and smaller groups of consumers. The ac generator and the transformer provided solutions to these problems. Moreover, dc circuits are difficult to interrupt as soon as the voltage or current exceeds a specified limit. For alternating current, circuit breakers benefit by the fact that current drops to zero after each half cycle.

The actual problems of long-distance transmission of electric energy arose at the beginning of the 20th century, when the use of distant hydroelectric sources was being contemplated. The possible advantage of direct current was recognized very soon when it was realized that the external insulation involved could withstand a voltage $\sqrt{2}$ times higher than the ac voltage, in a clean atmosphere, and that, in the case of insulated cables, dc voltages not much below the breakdown value could be supported indefinitely whereas a more or less rapid breakdown would occur under a much lower ac voltage. However, as is often the case in technical history, this notion could not be applied because of the lack of adequate equipment to convert alternating current into direct current, and vice versa. This is one of the main reasons why the long-distance lines built to carry hydroelectric power to possible use areas, and interconnecting lines intended for the pooling of spares and the utilization of diversities between different areas, were developed as ac lines in most instances.

In places where rivers, lakes, and harbors had to be crossed, insulated cables sometimes were necessary. The question of crossing straits and supplying islands also arose. In this field alternating current presented special drawbacks. The possible operating voltage was limited by the relative fragility of the ac cables, which, in view of their weight, were rather delicate to handle and difficult to lay, and also by the problem of the charging current, since the capacitance of a long circuit could practically short-circuit the terminal utilization impedance. Here, again, dc transmission appeared very attractive; but it was necessary to wait until the past decade, as with long-distance submarine telephone circuits, before an important advance made it possible to return to the use of direct current for the transmission of electric energy.

The main factor in this evolution has been the development of tubes with grid control, suitable for operation under high voltage and high current, and able to achieve ac/dc conversion, and vice versa, without mechanical means. As a result, submarine power cables have now extensively increased their potentialities by the use of direct current; the transmission distance is no longer limited by charging current, but only by losses, and the possible voltage and current have multiplied the power transmissible per cable by a factor of at least 3. This increase in submarine cable performance probably has not yet greatly affected the development of submarine links, although many projects that previously had been considered either unfeasible or feasible only with risk and difficulty are now in operation. However, the lack of circuit breakers suitable for dc operation still prevents the substitution of dc lines for ac lines in most interconnected systems, except for special applications.

It seems, therefore, that submarine transmission circuits have, to a certain extent, developed along parallel lines in the fields of telecommunications and power transmission, partly as a result of the progress made in tube technology. And it is probable that in the near future submarine circuits in both fields will again benefit, in this instance by the successful development of semiconductors.

Whereas telecommunication circuits have progressed from almost strictly dc to the very-high-frequency field, the evolution of power circuits has been in the other direction, from alternating to direct current. Actually, the objectives of the two transmission systems are different. The power carried by telecommunications signals is of the order of a milliwatt, and the enormous attenuation of the signal due to the very high frequency necessary for carrying a great volume of information can be compensated for by distributed amplification, with practically no limitation due to economical considerations of efficiency. On the other hand, efficiency is an important economic consideration in a power circuit, since the power carried by a single line can reach some 100 to 1000 MW.

Mechanical problems

The laying of submarine cables subjects them to a number of mechanical stresses that can weaken the cables and cause damage that sooner or later will destroy them. Several of these mechanical problems were experienced by the engineers who developed the first transatlantic telegraph circuits (Fig. 1), as can be seen from a description of their efforts.

Two dates are especially important in the early history of these cables: 1858, the year in which the first Atlantic cable was laid, which unfortunately failed; and 1866, when efforts with a second cable proved successful.

The pioneer attempt in 1858 was made by Cyrus Field, who established a link at mid-ocean by means of two ships, *H.M.S. Agamemnon* and *U.S.S. Niagara*. Shortly after the cable was put in service, transmission began to fade and within 20 days the cable was totally out of service.

There were probably a number of reasons for the failure of the circuit. For one, the quality control during manufacture was undoubtedly poor at that time, when no electric unit had even been defined. Also, the theory of cable laying was not yet well established, since it was not until 1874 that W. Siemens presented his paper on the topic before the Berlin Academy of Sciences.[5] In fact, the cable used was subjected during laying to a static mechanical tension that was not much lower than its ultimate strength. Additional stresses could have also resulted from the ship's movement due to the action of the swells and waves.

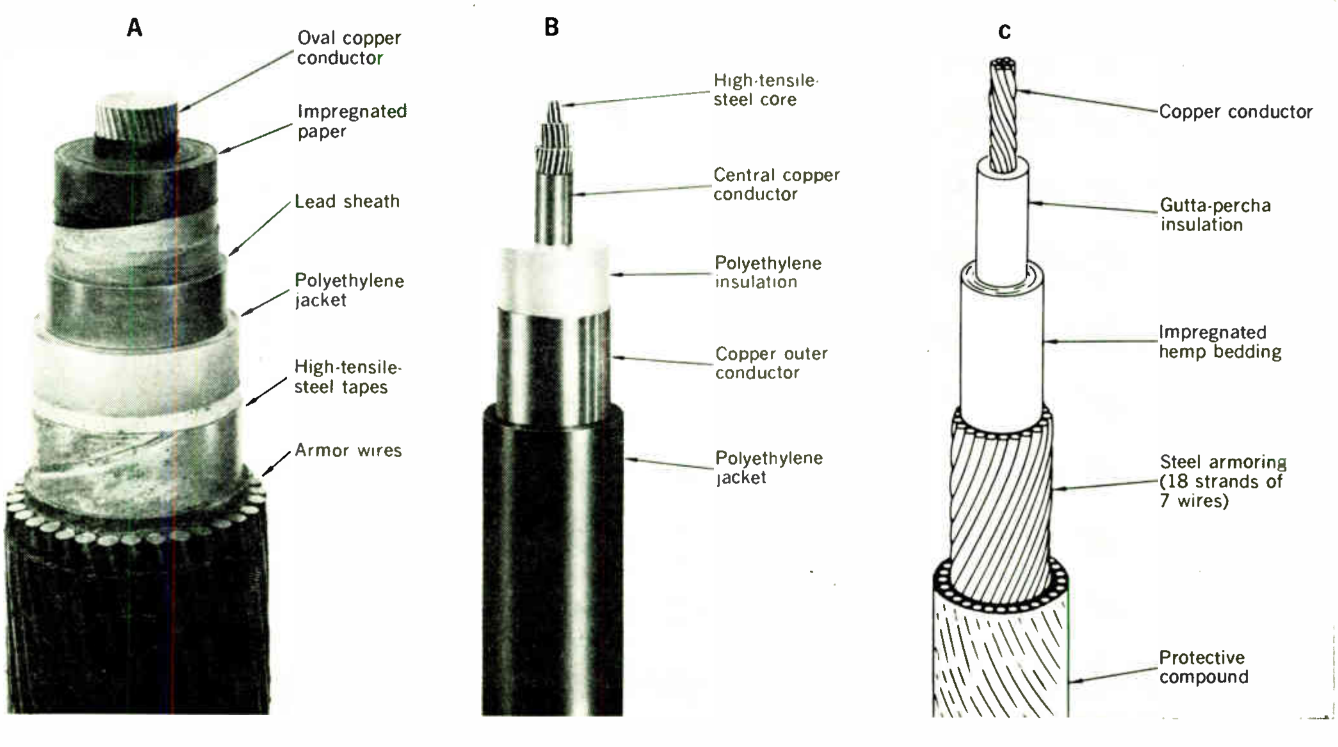
Field's success with the second transatlantic cable, in 1866, was probably partly due to his use of a larger ship, the *Great Eastern*, which had better navigational qualities than its predecessors. Moreover, the cable design was improved and much better adapted to the necessary conditions, as the cable's ultimate strength was of the order of two to three times the laying tension, a safety margin that has since been maintained for telegraph cables and for the more recent telephone cables.

In spite of this first success, submarine telecommunication cables were subject to many other afflictions. Sometimes there were concealed or incipient faults, which could grow worse during cable repair operations, because at this time the stresses are even greater than during laying, when the cable is lightened by the lift resulting from hydrodynamic forces. The theories generally accepted since Siemens[5-7] assume that a cable lies practically in a straight line and that the difference between the tensions at sea level and on the sea bed is nearly equal to the per-unit weight of the cable p multiplied by the water's depth d. During repairs, however, the hydrodynamic forces are suppressed and the cable's shape is similar to the catenary line studied by Euler in the 18th century. The difference between the tensions at sea level and on the sea bed is again equal to pd, but the necessity of insuring a sufficiently positive tension on the bottom generally leads to a higher tension than during laying.

The troubles that affect submarine cables are essentially attributable to mechanical problems resulting from the cables' complex structure, which includes helically spiraled layers. These problems are essentially: the spewing, which causes the core to penetrate the external armor; the knuckling, i.e., the buckling of the conductor that goes through the insulation; the birdcaging, or buckling of the armor wires, particularly during coiling; and the screwing and kinking of the cable on the sea bed. All of these phenomena are difficult to analyze in detail as they result from exceeding the elastic limits of the materials or the adhesion limits between different parts of the cable. Unfortunately, our capacity for analysis is limited to the linear field of our classical mechanics.

Some of the afflictions that were rather mild in the case of telecommunication cables, whose dimensions

Fig. 2. Comparison between (A) a 200-kV dc submarine cable, Sardinia–Italy; (B) a modern deep-sea telephone cable, TAT IV; and (C) the first transatlantic cable, in 1858.



and weight are generally small, appeared more serious with power cables, which are much larger and heavier; see Fig. 2. The power cables are also more delicate than telecommunication cables because their insulation must withstand a much higher electric stress; in the telecommunication cables this is limited to the supply voltage of the amplifiers.

To obtain a better understanding of all the stresses that can affect a power cable during its lifetime, let us follow the career of a newly manufactured cable.

The armored cable, served overall with jute layers, is first coiled down into the factory tanks, which requires a mean 360° twist in the cable for each turn. As the armor wires are laid up with an angle α, the twist results in an elongation per unit length of the cable equal to $(r/R) \tan \alpha$, where r is cable diameter and R is coiling diameter. If the cable diameter is small, as in telecommunication cable, and R is reasonably high, the elongation is very limited and there is no problem. But if the cable diameter is large enough in comparison with the coiling diameter, the resulting elongation may damage the dielectric, especially if it is made of paper and if joints with hand-lapped insulation are included. In addition, the cable can exhibit birdcaging at any point where the outer covering is weak enough to be torn by the pressure exerted by the armor wires, which tend to open. To avoid this trouble, in some instances special armoring has been used that has no outer jute covering common to all wires.[8] The wires are protected individually by cotton tapes or extruded material, and therefore, under coiling conditions, they can open slightly without inducing a high tension in the conductor.

The cable is transferred from the cable tanks at the factory into those of the laying vessel. The same twisting is once again experienced, and can be even worse if the ship's tanks are smaller and if the coiling height is limited, for this means a possible nonuniform distribution of the twist along the cable.

The cable, now ready for laying, is passed onto the drum of the laying machine, which will hold the tension of the cable during the laying operation. Several turns are necessary to obtain a sufficient friction force at the periphery of the cable. This force produces a slippage of the conductor with respect to the armor wires, and can damage the insulation if the cable is not properly designed.

The passing of the cable over the bow sheave causes bending and possibly a degree of twisting or untwisting of the armor wires, depending upon the trend (port or starboard) of the ship, which causes the cable to press against either side of the sheave. During laying, the cable is anchored at both ends—on board the ship at the drum of the laying machine and on the shore. Between the drum and the point at which the cable reaches bottom, the cable is subjected to a mechanical tension F decreasing from a value determined at the bow sheave by the depth and per-unit length weight of the cable, to an approximately zero value on the sea bed. This tension induces a torque T due to the helically spiraled armor (Fig. 3). The relation between F and T can be written $T = Fr \tan \alpha$. Consequently, the armor tends to increase its lay on the upper part of the suspended cable but returns to an approximately normal lay on the lower part. This torsion yields an elongation, and, therefore, conductor stresses and insulation deformations that can be detrimental to the cable and its joints. In addition, if the cable tension is reduced or if the bow sheave drops suddenly, some slack is paid out and the residual forces in the armor wires near the bottom can cause the cable to form a turn that can be drawn out into a kink on the next rise of the ship's prow or subsequently during recovery for repairs. In any case, kinks are more dangerous for power cables than for telecommunication cables, since the insulation can be easily damaged by too sharp a bending radius.

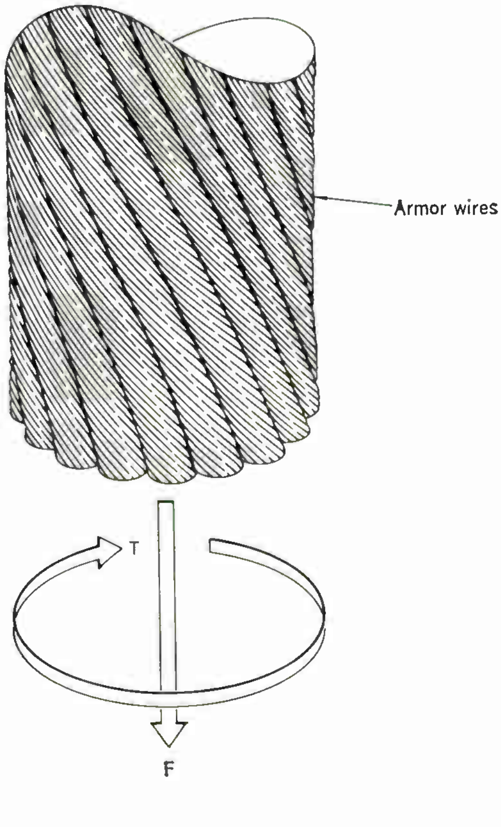
The risk of kinks appearing in power cables can be eliminated by maintaining pay-out tension high enough for the residual tension on the bottom always to be sufficiently positive. This pay-out tension therefore must be recalculated constantly, taking into account the actual depth, the bottom gradient, and the ship's speed.

The rotation and elongation of the cable can be limited by the use of a proper cable design in which the torque yielded by the armor wires under a longitudinal tension is canceled or sufficiently reduced by that of tapes helically applied underneath, in the opposite direction to the armor wires. These tapes can be conveniently combined with the reinforcing tapes of the sheath when a pressure cable is used.

When laid, the cable is subjected to the possibly high external pressure and sometimes to the action of water currents; if it is hanging on rocks, severe bending and fatigue of the lead sheath under repeated flexion may occur.

All these problems make the laying of large power cables at great depths a difficult operation, which explains some of the troubles recorded so far in the history of this technique.[9,10] Probably the most arduous task involved

Fig. 3. Torque induced in armor wires by a longitudinal force.



with submarine cables is to make repairs at sea. It is certainly at present one of the principal limiting factors with regard to submarine power cables.

Because cables are laid with practically no slack, they cannot be lifted up to the ship's bow without cutting them first, at least when the depth is greater than a few tens of meters. This operation can be difficult on a large cable lying at great depth. In addition, the jointing of the cables at sea is not only a delicate operation but also a very long one, during which the vessel must be held firmly in position regardless of the condition of the sea. This presents probably the most critical problem regarding repairs at sea. The laying of the final joint after a new length of cable has been inserted into the circuit is also a hazardous undertaking as kinks must be avoided. The development of light flexible joints[8] can simplify this phase of the job. It would also be desirable when making repairs to have available a vessel that has propulsion machinery better suited for this purpose.

Nevertheless, power cable laying at a depth of about 500 meters, with corresponding pay-out tensions of some eight tons (weight of cable, 15 kg per meter), has been possible when conditions are favorable, although the tension is higher than that of a telephone cable being laid at a depth of 5000 meters. Moreover, the latter cable can be picked up under a tension of 15 tons without damage.

Electrical problems

At first it might seem surprising, since direct current antedates alternating current, that an electrical problem could arise with dc cables. However, the slow polarization phenomena already evident with the first telegraph cables, and used for checking their insulation, prepared us to expect that the determination of electrical stress in a dc cable would probably be difficult. With alternating current, it is generally accepted that only rapid polarization phenomena are to be considered. The latter, which are the orientation of dipoles, either permanent or created, in the applied field, have a time constant that is much lower than the period of the power system frequency. Application of the Gaussian relation, with zero divergence, permits simple calculation of the local stress at any point on the cylindrical capacitor and leads to the classical hyperbolic stress distribution. This stress is macroscopic, as the insulation is nonhomogeneous (alternate layers of paper and oil, crystalline zones in polyethylene, etc.). The calculation and the assumptions made seem to correlate reasonably well with the results of breakdown tests under alternating current or under impulse. Within certain limits, the calculated maximum stress relevant to a given type of insulation is in effect practically constant at breakdown level.

With direct current, the slow charging phenomena observed when a voltage step is applied to an insulation cannot correspond to dipole movements, after the first transient, but must result from charges migrating through the dielectric. A charge density is then likely to appear, and consequently the stress divergence in the insulation is no longer zero.

This slow charging phenomenon might be considered microscopic, the result of the insulation's heterogeneity and of the absorption of electric particles on barriers distributed in the insulation, such as cellulose fibers in

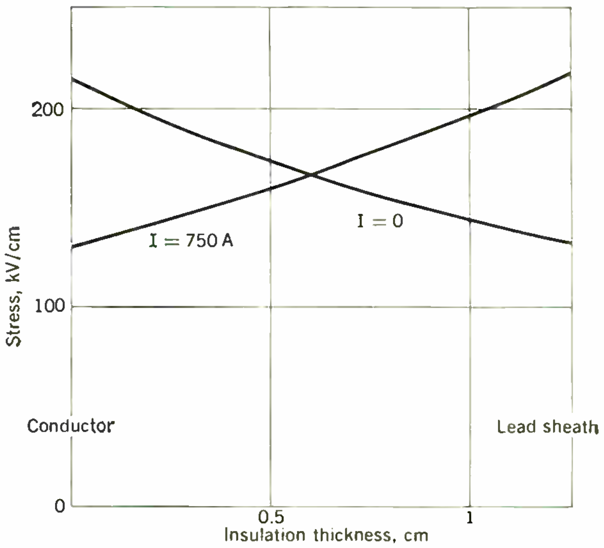


Fig. 4. Calculated stress distribution in a dc power cable (Sardinia–Italy) at no load and at full load.

paper. The phenomenon could also be macroscopic, and in this case measurements of voltage between electrodes placed inside the dielectrics should reveal abnormal voltage and stress distribution. Such measurements, rather difficult to make under high voltage, were first made at limited voltages by Pugliese[11] and showed no difference with regard to the usual ac hyperbolic stress distribution when the cable was at uniform temperature. However, when the cable was loaded, i.e., when a temperature gradient was applied to the insulation, the stress decreased near the conductor and increased at the insulation periphery. The resulting stress could be calculated from the asymptotic insulation resistance, which is a decreasing function of temperature, and thus explains the overstress observed on the outer layers of insulation of a loaded cable (Fig. 4). A possible method for checking the validity of the assumption according to which the stress could be calculated from the asymptotic conductivity of the insulation layers, was to determine experimentally this law of variation of the conductivity, then to calculate the stress, and finally to ascertain, by breakdown tests carried out under various temperature gradient conditions, that the predicted breakdown stress agreed with the theoretical value.

Measurements carried out in several laboratories on insulation samples, at uniform temperature, showed that their asymptotic resisitivity value decreased when the temperature was increased, but also when the applied stress was increased.[12] This means that the stress distribution in the dielectric is not only distorted by the temperature gradient across it, but also by the stress itself, which tends to make the stress more uniform throughout the insulation.

The most appropriate physical law for expressing the asymptotic conductivity of conventional insulating materials (oil-impregnated paper and polyethylene) seems to be given by

$$\sigma = \sigma_0 \exp\left(-\frac{W - aS}{2kT}\right)$$

I. Main characteristics of existing submarine dc power transmission schemes

Place and Date	Voltage Between Poles and Power	Maximum Depth, fathoms	Cable Length, miles	Cable Type	Conductor Size, mm²	Thickness of Insulation, mils	Weight, lb/ft
Gotland (Sweden), 1954	100 kV 20 MW	82	61	Solid (one conductor)	90 (178 MCM)	275	5.65
Cross-Channel cable (France–England), 1961	200 kV 160 MW	33	41.2	Solid (one conductor)	344 (680 MCM) and 605 (1200 MCM)	295	10.7
Cook Strait (New Zealand), 1965	500 kV 600 MW	140	25	Preimpregnated gas-pressure (one conductor)	516 (1020 MCM)	570	28.7
Konti–Skan (Sweden–Denmark), 1965	250 kV 250 MW	44	36.3	Solid (one conductor)	625 (1230 MCM) and 800 (1580 MCM)	630	16.8
			14.3	Oil-filled, flat type (two conductors)	310 (610 MCM)	488	22.2
Sardinia–Italy via Corsica, 1966	200 kV 200 MW	246	73.5	Solid (one conductor)	420 (830 MCM) (oval)	465	14.1

where

σ = conductivity for a given stress S and absolute temperature T

W = an activation energy, equal to some 2 to 3 eV as in the case of semiconductors

a = a length, from 10 to 100 Å, approximately, depending upon the material

k = Boltzmann constant, which plays its universal role and establishes a correspondence between 10 000° and 1 eV

For paper cables, the factor a is smaller than for polyethylene cables, and therefore the action of the stress is more limited. The stress is more affected by a temperature drop in the insulation than it is in a polyethylene cable, where a more uniform distribution is obtained, even with no load.

Since the time constant of the macroscopic phenomena is high, it appears that polarity reversals may entail overstresses, because of the superposition of the initial stress and any transient stress that may result from the reversal. (This overstress can be a limitation in the operation of dc cable links, where rapid polarity reversals are necessary in order to change the direction of power flow.) The variation of insulation resistance due to temperature can also entail, under extreme conditions of high conductor temperature, thermal instability with corresponding low breakdown voltage.[13]

Direct measurement of the stress distribution in a cable is a difficult operation under high voltage, especially because of the influence of terminal arrangements; but, fortunately, experimental checks have shown that the inferences drawn from the simplified theory were correct. Thus, test results obtained under various conditions of temperature and voltage gave a breakdown stress of the order of 1200 kV/cm, a value that is fairly consistent with that generally obtained under impulse. Also, failures caused by thermal instability have been predicted and then experienced under actual test conditions.

It is quite probable that other phenomena, involving charge accumulation near the electrodes, exist in the dielectric. Such charges can of course modify the stress, but they are very difficult to forecast. It seems, however, on the basis of the aforementioned test results, that these secondary phenomena do not change the picture very much at stresses close to the breakdown limit.

From a microscopic standpoint, the calculated stress can certainly be considered as an approximation open to criticism. However, what more can one ask of a physical representation than that it can enable the reasonable prediction of the results of a series of experiments? In the present instance, the results of the breakdown tests are in acceptable agreement with theory.

There is still much to be learned about the behavior of dielectric materials under dc stress. The aging and breakdown mechanisms are not yet well understood nor is the fact that the practical breakdown levels obtained for industrial materials are still much lower than those that can be measured on thin samples and those that should be attainable in theory. Statistical considerations certainly play a part in the phenomena, but it is difficult to say to what extent.

The recent development of dc transmission is opening a new field of industrial research on dielectric materials, which thus far had been studied mainly under ac conditions where only short time constants are involved. Service experience on dc cables should also be very instructive.

Present state of development

Twelve years have elapsed since the commissioning of the first submarine dc cable, between the Swedish mainland and the island of Gotland. In view of the successful operation of this pioneer link, several new schemes have since been constructed. The main features of these schemes are summarized in Table I. From the table it can be seen that paper-insulated mass-impregnated solid-type cable has been used for most existing schemes, except for one in which a gas-filled cable was used, and for part of another, in which a special oil-filled flat-type cable was retained.

It must also be realized that the maximum laying depth so far is of the order of 500 meters. At a depth of this magnitude (and even smaller) repairing the cable is not an easy operation, particularly because it is difficult to hold the ship still for a long period under such conditions. For this reason, it is sometimes preferable to cut a cable at a spot in shallower water, and to recover a rather long length of cable to reach the area that requires repair.

After these 12 years of development it is interesting to review the present possibilities of the different types of cables for dc application.

The solid-type cable, which has been used up to 250 kV, is probably not far from its limit, and it is doubtful that it will ever be used much above 350 kV, if one takes into account maximum permissible stress. The main advantages of this type of cable are its simplicity, which makes it easy to repair, and the fact that it can be produced in rather long lengths; for instance, a cable factory in France is equipped to produce a 40-km-long cable similar to the Sardinia–Italy cable. In addition, there is no length limitation, as there is for oil-filled cables. However, it does not appear possible to raise the working stress of this cable to more than some 300 kV/cm, which is approximately seven times the possible ac (rms) working stress for this type of insulation.

Where deeper water is involved, solid-type cable must be given an oval shape, and the lead sheath must be reinforced so that it can be deformed under the internal pressure built up by the heating cycle and then return to its original shape under external pressure, without damage to the sheath. Such a principle was used in the Sardinia–Italy cable. Even with low external pressure, void formation is reduced and the dielectric behavior of a reinforced oval cable is improved. A higher external pressure acts as an elastic reinforcement of the lead sheath would, and void formation need not be feared. Voltage and stress limits could be raised in this case, but the full advantage could be obtained only if special arrangements were provided in shallow water, especially at the shore ends, so that the compression effect would also exist at these points.

Oil-filled cable, which still seems to be the best for ac transmission from the electrical standpoint, is also attractive for dc transmission, although experience with this type of cable is still very limited. In addition, the necessity to feed oil at various places along the circuit, in order to reduce the transient pressure drop when it exceeds a certain length, is a limiting factor in submarine application. In this latter case, a reasonable possible limit seems to be about 30 km if special fluid oil is used. Repairs at sea also imply difficulty.

Flat oil-filled Möllerhöj-type cable, as used for part of the Konti–Skan project,[14] is certainly an interesting solution, because no longitudinal pressure drop is involved and no oil reservoirs are necessary along the cable route. However, the cable weight is increased because bipolar cables must be used, and consequently the maximum laying depth is reduced and joints and repairs are made more critical. In any case, oil-filled cables (perhaps of the high-pressure type) will probably constitute a solution for future transmission voltages in the region of ± 500 kV.

Gas-filled cables are an apparently good solution to the problem of long submarine dc links. Their electrical behavior is somewhat better than that of solid-type cables, since voids in the insulation are partially compensated by the gas pressure, and slightly higher stresses and voltages can be reached; moreover, there is in principle no distance limitation. Operating voltages of the order of 300–400 kV seem possible. Also, one must be prepared for trouble in the gas transmission if the relatively small gas passage should become blocked up by unexpected compound migration, with a correlative reduction of the withstand level.

The need for a rather high internal pressure at the deepest part of the route may necessitate a very high operating pressure near the ends, which makes the cable heavy and more difficult to handle, to lay in great depth, and to repair. This effect of overpressure at the shore ends is much more limited with oil-filled cables, because the internal head of oil compensates for 9/10 of the external head of water.

Extruded polyethylene would provide a very good solution to submarine dc cable problems, since it is a very light material and does not need a heavy impervious metal sheath. However, there has been no experience with the service reliability of this type of cable, which is delicate from the manufacturer's point of view; the extrusion of very long lengths would imply a certain risk of weak points difficult to detect. The maximum permissible stress under direct current is not well known for this insulation, and it seems that the present practical voltage limit would be of the order of ± 200 kV; however, no one has yet used this type of cable for a commercial circuit.

A gas-pressurized cable, insulated with synthetic tapes, would have some mechanical and electrical advantages over gas-filled impregnated-paper cable, without the uncertainties of an extruded insulation. This solution remains purely experimental, but a voltage of about 500 kV seems attainable.

To sum up, dc cable designs are available for the highest voltages considered so far. Although the optimistic views regarding dc cables expressed a few decades ago have been somewhat moderated by recent practical experience and experiments, the fact remains that the permissible stress for dc operation is at least some 2 to 2.5 times the ac (rms) permissible stress for modern pressure cable designs. A single EHV cable, when used with direct current, can therefore transmit power about three times as high as that possible with alternating current due to this possible increase of working stress and to the elimination of dielectric losses, skin and proximity effects, and eddy-current losses. If a circuit of two dc cables is compared with a circuit of three similar ac cables, the ratio of the transmissible power is of the order of 2 to 1. However, as only solid-type cables were considered in our initial comparison of direct and alternating current, an even more optimistic advantage was predicted for direct current, since a single solid-type cable with direct current can transmit some eight times the power transmissible with alternating current (taking into account the reduction of losses).

In the submarine field it is desirable to use simple, rugged, lightweight equipment as much as possible. To insure reliability, it is necessary to use taped insulation that must be protected by metal sheaths with their elaborate reinforcement. Consequently, the maximum possible laying depths, and also the circuit lengths, of submarine cables are limited.

Prospective considerations

In our present world, two apparently divergent trends seem likely to affect the future of electric energy transmission.

On one hand, distant hydroelectric resources (e.g., in the polar regions), which cannot be used on site because of prevailing climatic conditions, are attractive, and their use would tend to increase the transmission distances between the source and the user in an inhabited area.

On the other hand, the increasing use of nuclear energy for the generation of electricity would seem to limit the problem of power transmission. Nuclear fuel, characterized by an enormous energetic density (some 10^9 times as great as that of water stored behind a dam), poses no transportation problem. However, for economical reasons, the size of the generating units must be of the order of several thousand megawatts. In addition, a cold source is required to dissipate a certain amount of the calories yielded by the hot source, according to Carnot's thermodynamic law. Water is certainly the most economic cold source, but it is becoming a rare material in view of the extension of its multiple uses. The present trend is to locate nuclear power plants near the sea, which means an even greater transmission problem for the inland countries.

In the case of conventional fuels also, a similar consideration leads to the location of steam power plants where water is available, despite the savings made possible by transporting the fuel in pipelines rather than transmitting the product—electric energy.

Another element in the future evolution of transmission systems is the circuit itself, which can be either overhead or underground. The former is certainly the more economical, but the technique is not subject to much further improvement, as the basic insulation is air. Moreover, this kind of equipment is being increasingly rejected by modern man, who is seeking to preserve the beauty of nature. The underground system is a better solution; it is well protected and well adapted to our future civilization, which undoubtedly will be composed of troglodytes frightened by the atomic threat and the prospects of subversive war.

Improvements in this latter type of circuit can still be expected, from the point of view of both insulation and conductivity, especially in the field of direct current. Therefore, dc cables, which thus far have been practically limited to submarine links, could become an important factor in the evolution of future systems.

As mentioned previously, direct current is in principle not well adapted to the generation of electricity, since today's generators are rotating machines, nor is it well adapted to the distribution of electric energy, for the voltage must be adjusted to the load. However, new means of electricity generation now being investigated do not require the rotation of a conductor in a magnetic field to set in motion the free electrons of the metal. The new processes (MHD, for instance) use a flame as support of the electric particles. This flame, which for the antique Greeks was a "permanent shape applied to a material ceaselessly renewed," is rectilinear, and dc electric energy is therefore produced. Consequently, the possibility of a future return to direct current for the generation of energy is no longer considered to be impracticable.

In the transmission field, cooling of the conductors to temperatures much below our ambient temperature can also be contemplated, as the techniques of refrigeration, associated with those of evacuation (in this case, vacuum would ensure the necessary thermal insulation), have made great progress in the last decades. However, cryogenics seem to be essentially applicable to dc cables. In fact, the skin effect, which is an increasing function of the square root of the conductivity and of the frequency, can be troublesome at 50–60 c/s for an absolute temperature of about 300°K. It would therefore be advisable to reduce the frequency as the conductivity is improved, which again argues for the acceptance of direct current.

With such cables, the conductor losses would be considerably reduced and large blocks of power could be transmitted at a rather low voltage without excessive losses, and consequently a high-voltage step could be eliminated in the transmission. It is thus possible to imagine that in the future the use of direct current will start at the generation stage and extend to a certain distribution stage, at which conversion to alternating current probably still will be necessary.

This concept assumes of course that the existence of suitable equipment for interrupting direct current will emerge to make the interconnection of dc lines acceptable and that reactive generation problems will be solved at the receiving end. All the views expressed in this article therefore should be considered only as an indication of a possible future evolution; they should by no means be interpreted as a prognostication of what is definitely going to be.

REFERENCES

1. Hunter, P. V., and Temple-Hazell, J., *Development of Power Cables.* London: George Newnes, Ltd., 1956.
2. Malegarie, C., *L'Electricité à Paris.* Paris: Librairie Polytechnique Ch. Beranger, 1947.
3. Cahen, L., "Histoire des câbles télégraphiques sous-marins," *Bull. Soc. Franc. Electriciens*, no. 25, Jan. 1953.
4. Mottram, E. T., "Submarine telephone cables," *IEEE Spectrum*, vol. 2, pp. 96–103, May 1965.
5. Wünschendorff, E., *Traité de Télégraphie Sous-marine.* Paris: Librairie Polytechnique, Baudry and Co., 1888.
6. Zajac, E. E., "Dynamics and kinematics of the laying and recovery of submarine cable," *Bell System Tech. J.*, no. 5, Sept. 1957.
7. Besley, J. C., and Higgitt, H. V., "The recovery of deep sea cable," *J. Inst. Elec. Engrs. (London)*, vol. 72, 1953.
8. Sallard, J., Tellier, R., Cherry, D. M., and Barnes, C. C., "Problems arising from the design and construction of high voltage dc submarine cables systems," CIGRE Rept. No. 415, 1960.
9. Farnham, D. M., Shanklin, G. B., Cunha, S. H., and Short, H. D., "The St. Lawrence River high-voltage submarine cable crossing," *AIEE Trans. on Power Apparatus and Systems*, vol. 78, pp. 1098–1185, Dec. 1959.
10. Williams, A. L., Davey, E. L., and Gibson, J. N., "The 250-kV dc submarine power cable interconnection between the north and south islands of New Zealand," *Proc. Inst. Elec. Engrs. (London)*, vol. 113, Jan. 1966.
11. Pugliese, E., "La répartition transversale de la tension alternative et continue dans l'isolant des câbles à haute tension," *Bull. Assoc. Suisse Electriciens*, no. 25, 1950.
12. Oudin, J. M., Thévenon, A., and Fallou, M., "Design and development of dc cables," Paper 31 TP 65-684, presented at IEEE Summer Power Meeting, Detroit, Mich., June 27–July 1, 1965.
13. Fallou, M., "Dielectric perforation through thermal instability of dc cables insulated with impregnated paper," *Direct Current*, Feb. 1963.
14. Smedsfelt, S., and Arlgren, L., "Experiences in building and commissioning the Konti–Skan HVDC project," *Direct Current*, vol. 10, no. 4, Nov. 1965.

Optical communications in the earth's atmosphere

Although the effect of atmospheric turbulence on coherent optical radiation would seem to limit its use in earthbound applications, the advantages of bandwidth and range that it offers should prove valuable in space and interplanetary communications

Bernard Cooper *ITT Federal Laboratories*

The main reasons for the present interest in the potentialities of the laser for communication system applications lie in the coherent nature of the radiation and the short wavelength. The property of coherence, as considered in this article, is the ability of an electromagnetic wave to interfere with itself. The phenomenon is characterized by the ability to form a highly collimated beam, which leads to the ability to achieve communications over great distances. However, the perturbing effects of the earth's atmosphere on coherent optical radiation raise a serious obstacle to the use of lasers in earthbound communication systems. Space and interplanetary communications applications appear more promising as atmospheric effects are eliminated.

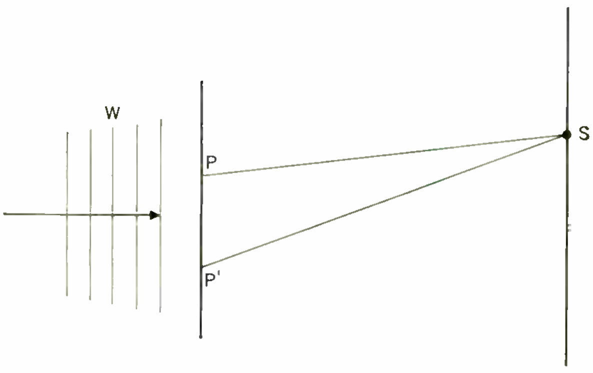


Fig. 1. Classical interference setup.

Since the first laser was demonstrated in 1960, considerable interest has developed in its possibilities for use in communication systems. The basic sources of this interest are the coherent nature of the radiation obtained as compared with all previously known extended sources of optical radiation, and the laser's short wavelength. This latter characteristic provides the potential ability to achieve bandwidths, or information capacities, that are orders of magnitude greater than anything obtained heretofore. A more realistic advantage, in terms of presently available information sources, results from the combination of high coherence and short wavelength. It is the ability to generate a highly collimated beam (limited by diffraction phenomena), which leads to the ability to achieve communications over great distances. Of equal importance is the fact that with a coherent signal, coherent detection of the information can be obtained with greatly improved immunity to natural incoherent noise sources such as the sun.

The concept of coherence is central to the following discussion. Although a thorough treatment is beyond the scope of this article, the essence of the concept can be simply presented (for an extensive treatment, see reference 1). For our purposes, the coherence property of an electromagnetic wave can be thought of as the ability of the wave to interfere with itself (constructively or destructively), which is related to the degree of determinism associated with the wavefront—or the degree of correlation between various points on the wave. An example is the classical case of monochromatic plane wave W passing through two small pinholes P, P', as shown in Fig. 1.

If the wave is perfectly coherent, then a unique (deterministic) relationship exists between the field vectors at the two pinholes. Since the intensity of the total field at any point S is related to the vector sum of the fields due to the two pinholes, it is not difficult to see that a unique interference pattern is established, a pattern in which constructive interference will occur at some points and destructive interference at others. On the other hand, if the wave were incoherent, no deterministic vector relationship would exist between various points of the wave. Therefore, for the cases considered here, no interference phenomena would be forthcoming. (Interference phenomena can be observed with incoherent stellar sources.)

Attendant upon the interference phenomenon is the ability of a coherently illuminated aperture to form a highly collimated beam or diffraction pattern. This pattern is commonly observed in connection with microwave antenna systems in which a coherent electromagnetic wave illuminates an aperture created by a parabolic reflector. The only basic difference in the process at optical frequencies *in vacuo* is the scale factor introduced by virtue of the difference in wavelengths. In the interest of mathematical tractability, two regions are defined with respect to the source aperture, as shown in Fig. 2. One region is the "near field" or Fresnel region, in which the distance normal to the aperture plane is small compared with D^2/λ, where D is the diameter of the aperture and λ represents the wavelength, in appropriate units.

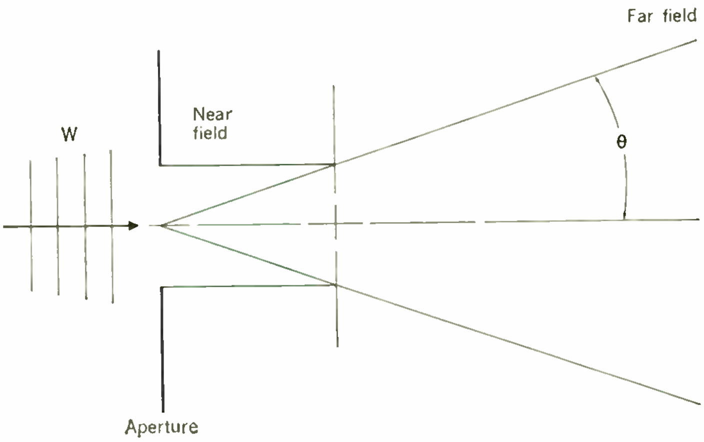


Fig. 2. Near and far fields of a coherently illuminated aperture.

In this region, the beam can be considered perfectly collimated, and a good replica of the intensity distribution across the aperture. It should be noted that even for aperture diameters of a few centimeters, this region can extend many miles at optical frequencies because of the very short wavelengths—approximately 0.6×10^{-6} meter—that are involved. The second region, "the far field" or Fraunhoffer region, is characterized by an angular distribution with respect to the aperture axis. The divergence angle of the beam in this region can be given as follows:

$$\theta \propto \frac{\lambda}{D} \tag{1}$$

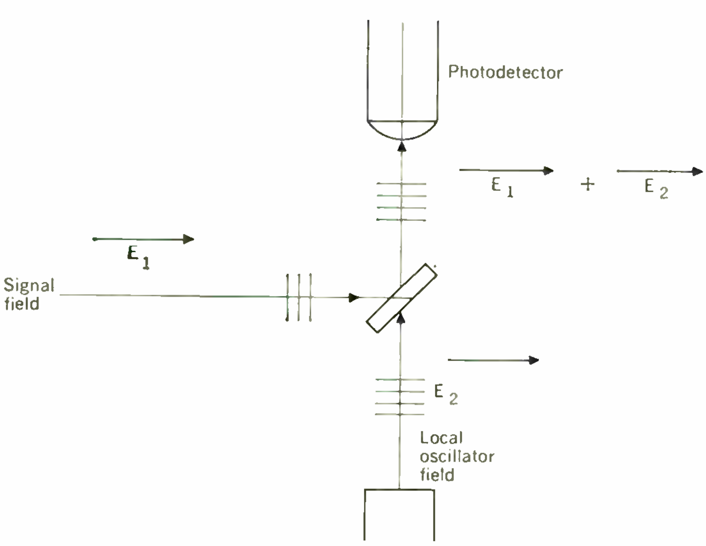
A simple calculation shows that for a wavelength of 0.6×10^{-6} meter (red light) with a 10-cm aperture diameter, the beam divergence is 0.006 milliradian. Thus, at a distance of 1000 miles, barring any anomalous effects in the propagation, the beam diameter will be only 0.006 mile or about 30 feet. The ability to concentrate energy at great distances is one of the more significant attributes of coherent electromagnetic radiation of optical frequencies.

As will be seen later, this capability is severely affected by various phenomena associated with the earth's atmospheric environment. It should be noted that the ability to focus a light beam is limited to the aforementioned angular relationship, and can be achieved only with perfectly coherent radiation.

Spectral selectivity and photomixing

The second characteristic attributable to the coherence phenomenon is that of spectral selectivity, which is the ability of an optical detector to respond to a very narrow band of frequencies to the exclusion of all others. The basis of this characteristic lies in the unique nature of the photoemission process associated with the detection of electromagnetic energy at optical frequencies. The output current is proportional to the input power.

Fig. 3. Optical heterodyning setup.



If one takes the deterministic case of complete coherence, a simple representation of optical mixing, such as heterodyning or homodyning, can be made. For example, consider the optical heterodyne system of Fig. 3. At the surface of the photodetector there are, ideally, two coplanar waves of frequency ν_1 and ν_2 and vector intensities $\mathbf{E}_1$ and $\mathbf{E}_2$ impinging on a photoemissive surface. The power, or output current, associated with these waves is given by the square of the vector sum of the two fields:

$$i(t) \propto (\mathbf{E}_1 + \mathbf{E}_2)^2 = E_1^2 + E_2^2 + 2E_1E_2 \cos 2\pi(\nu_1 - \nu_2)t \tag{2}$$

The results are a dc term and a term at the difference frequency of the two waves. The latter term is linearly proportional to each of the fields and is considered the signal term. The detected signal, therefore, has an average power proportional to $E_1^2E_2^2$.

In contrast with this, for the case of two incoherent beams of light, the output current can be described by the relationship

$$i(t) \propto E_1^2 + E_2^2 \quad (3)$$

so that no cross-product signal term results.

For the case of optical mixing one can consider E_1 as the signal and E_2 as the local oscillator, as shown in Fig. 3. In general, E_2 will be very much larger than E_1 and therefore, in the coherent case, the output current becomes

$$i(t)_{\text{coh}} \propto E_2^2 + 2E_1E_2 \cos 2\pi(\nu_1 - \nu_2)t \quad (4)$$

whereas for the case of two incoherent beams, the output current is

$$i(t)_{\text{incoh}} \propto E_2^2 \quad (5)$$

The former case provides a signal due to E_1, enhanced by E_2; the latter yields only an output essentially proportional to the square of the larger field, thereby suppressing the desired signal.

Noise considerations

Because of the quantized nature of the electromagnetic wave, a randomness is imparted to the output current of the photodetector. The power associated with this self-noise effect is proportional to the average value of the photocurrent and is uncorrelated with the heterodyne signal. In addition to the effects of "self noise," one must consider the effects of incoherent background illumination having the sun as its primary source, as well as the noise due to spontaneous random electron emission (dark current) from the photoemissive device. It can be shown[2] that if the local oscillator power is large compared with the cumulative effect of the various uncorrelated noise sources, then the detected signal-to-noise ratio approaches

$$\frac{S}{N} = \frac{QP_S}{h\nu B} \quad (6)$$

where

Q = detector quantum efficiency
P_S = mean square input signal intensity
h = Planck's constant
ν = signal frequency
B = bandwidth of receiver
S = output signal power
N = output noise power

Thus, by virtue of coherent photomixing, the effects of uncorrelated noise sources can be suppressed.

The foregoing considerations are predicated on a perfectly coherent source and propagation *in vacuo*.

Taking the practical, albeit subjective, viewpoint that useful communications must of necessity involve activities on the earth's surface, the communication engineer is immediately faced with the problem of the perturbing effects of the earth's atmosphere on coherent optical radiation. These effects can be described in terms of three types of effects on the optical beam: (1) beam spreading, (2) angular fluctuation of the beam, and (3) attenuation of the beam. To the extent that these effects are present, the coherence of the received radiation, and the ability to reap the benefits therefrom, are reduced.

In addition to the effects of the earth's atmosphere, the nonideal coherence characteristics of the laser[3] cause further departure of the performance of coherent optical detection systems from the ideal indicated by Eq. (6).

Coherent detection efficiency

The extent of degradation in communication performance can be studied in terms of the efficiency as is given by the ratio of actual to theoretical detected signal power (proportional to the square of the signal current):

$$\eta = \frac{i_S^2 \text{ (actual)}}{i_S^2 \text{ (perfect coherence)}} \quad (7)$$

The concept of mutual coherence plays a central role in the evaluation and interpretation of results of "coherent" optical communication experiments. It can be shown[4] that for optical heterodyne detection the efficiency is also given by

$$\eta = |\lambda_L| \, |\gamma_S| \qquad |\gamma_L|, \ |\gamma_S| \leq 1 \quad (8)$$

where

γ_L = real part of mutual coherence of local source
γ_S = real part of mutual coherence of received signal

Equation (8) shows that, even in the absence of external disturbance due to an inhomogeneous medium, the efficiency of the photomixing process is controlled by the coherence of the sources used for both the local oscillator and the signal. The effect of the propagation medium is reflected in an additional reduction in $|\gamma_S|$, assuming that the local oscillator path is unaffected by the short local oscillator propagation path, as shown in Fig. 3.

Measurements across the aperture (beam region) of a He-Ne gas laser with an 80-cm cavity length show that the mutual coherence drops off measurably with spacing across the aperture; that is, the correlation of the fields at two points in the laser beam right at the laser output decreases as the spacing between the points increases. The degree of this effect depends on the laser design. Some measurements of laser coherence[3] have shown a reduction of $|\gamma|$ to 0.4 and less for spacing of the order of the beam radius—so that, even before the degrading effects of the atmosphere are faced, there is a significant reduction from the ideal case of perfect coherence across the laser beam. As will be shown later, typical measurements of coherent detection efficiency show reductions to 50 percent for detection over the full aperture due to imperfect laser coherence alone.

The effects of the earth's atmosphere on the performance of a "coherent" optical communication system are considerably more dramatic than the nonideal source characteristics that were discussed in the previous sections.

Atmospheric effects

The effects of the earth's atmosphere on the propagation of both coherent and noncoherent light have been studied by many investigators, from many viewpoints. References 5–11 give a representative cross section of this activity for readers interested in further work. The one observation that is common to all of the studies, however, is the nondeterministic nature of the processes contributing to the phenomenon. There are virtually an infinity of possible atmospheric conditions, each of which can be described only in a statistical manner with reference to the effect on propagation at optical frequencies. To help overcome this seemingly intractable analytic problem, the effects of the atmosphere can be described in terms of three disturbance functions; these are attenuation modulation, phase fluctuation, and angular fluctuation.

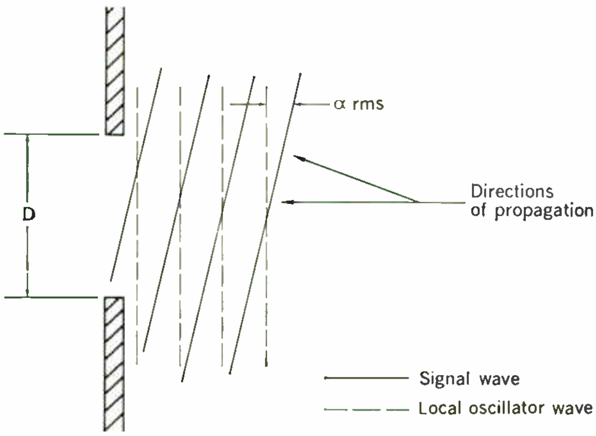


Fig. 4. Simplified representation of angular fluctuation effect.

The first two disturbance types can be accommodated by suitable detection technique[6]; however, angular fluctuation represents an irreversible degradation in the coherence of the wave and thus limits the obtainable detection efficiency. Once having been so degraded, no compensation can be made to restore the coherence. The essence of the effect of angular fluctuations on coherent photomixing can be illustrated by a greatly simplified example.

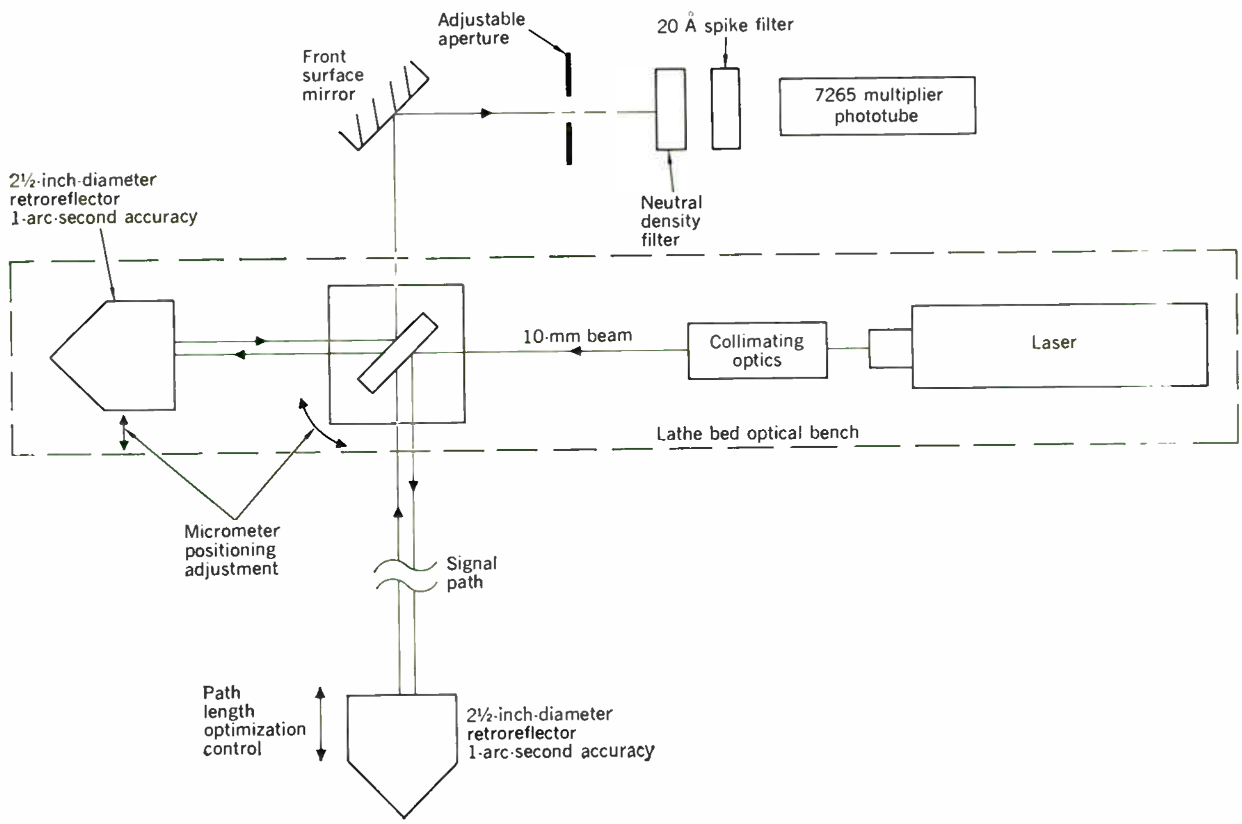
Consider plane waves from a local oscillator and a remote source arriving at a photomixer of diameter D, as in Fig. 4. Furthermore, assume a simple angular fluctuation of the signal such that the wavefront remains essentially planar as shown, but tilts with respect to the local oscillator wave in a random manner with an rms value of α. Since the total output of the photodetector is the square of the vector sum of the fields integrated over the entire aperture, it can be qualitatively seen that as α approaches $\gamma/4D$ the photomixer output will decrease. This decrease is attributable to the fact that on some portions of the aperture the local oscillator and signal fields will be orthogonal to one another, so that the cross-product term will be zero. As α increases, the proportion of the aperture experiencing orthogonal fields will increase.

An additional degradation can be predicted since the perturbed signal wave is not really planar, but undulates in a random manner. The relationship

$$\alpha_{\max} \cong \frac{\gamma}{4D} \tag{9}$$

illustrates a rather important, demonstrable point in the appreciation of the limited utility of coherent optical

Fig. 5. Coherent homodyne detection test setup.



techniques for communication through the earth's atmosphere.

First, it is evident that the longer the propagation path through the atmosphere, the greater will be the rms angular fluctuation. With this in mind, (9) can be rewritten as

$$L_{\max} \propto \alpha_{\max} \cong \frac{\gamma}{4D} \qquad (10)$$

where L is the path length. Thus, as the path length is increased, the maximum aperture diameter must be decreased in order to maintain constant detection efficiency. Decreasing the aperture diameter, however, reduces the received power over and above the reduction caused by the increased path length, which limits the maximum range capabilities of a coherent optical communication system if it is operating in the earth's atmosphere.

Now that the nature of the coherence degradation phenomenon has been introduced, in a rather elementary way, some empirical results will be given in the next section.

Empirical measurement of coherent detection efficiency

Reference 3 presents data on detection efficiency as measured in an optical homodyne system by the writer. The test configuration used is shown in Fig. 5. The signal path is created by a cube corner reflector that reflects an incident beam parallel to itself with an angular accuracy of less than one arc second (0.000 278°). Both the signal and local oscillator beams are generated by one laser, with the aidof an optical beam splitter.

If an unmodulated laser were used in such a configuration under idealized conditions, the output current from the photodetector would be dc, proportional to the product of the intensities of the local oscillator and signal fields. In actual practice, effects due to two types of phenomena can be distinguished. The first is the effect of small fluctuations in the path lengths of the local oscillator and signal fields resulting from mechanical vibration of the components in the experimental setup. Clearly, if the differential path length varies by $\lambda/2$ (0.3164×10^{-6} meter), the polarity of the dc ouput will reverse. This effect is completely independent of atmospheric effects, as measurements over very short paths where such effects are negligible indicate. Figure 6(A) shows a typical photodetector output for a measurement of this type. Note the essentially constant amplitude, variable frequency beat pattern caused by mechanically originated path length fluctuations. This somewhat anomalous behavior, unique to homodyne systems, can be used to advantage to measure coherent detection efficiency.

It is shown[3] that the efficiency η is given by the relationship

$$\eta = \frac{(I_{\max} - I_{\min})^2}{16 I_L I_S} = \frac{(\Delta I)^2}{16 I_L I_S} \qquad (11)$$

where

$I_{\max}$ = maximum photodetector current
$I_{\min}$ = minimum photodetector current
I_L = photodetector current due to local oscillator field only
I_S = photodetector current due to signal field only

Thus, for a given local oscillator and signal strength, one would expect the output current excursion to decrease as the detection efficiency decreases; and this effect is, in fact, observed when the signal path extends over a path through the atmosphere. Figures 6(B) and 6(C) show the nature of the detector output for these observations. Note that the pure "beat" character of the wave is degraded. During the periods in which the voltage excursion is significantly reduced, the mutual coherence of the wave over the aperture is also low.

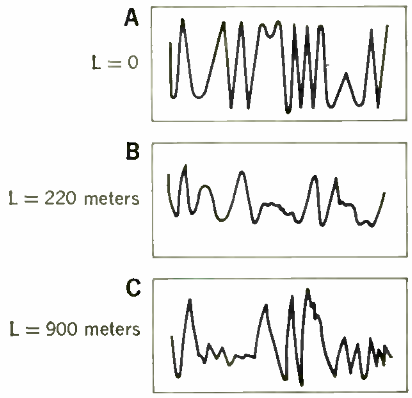
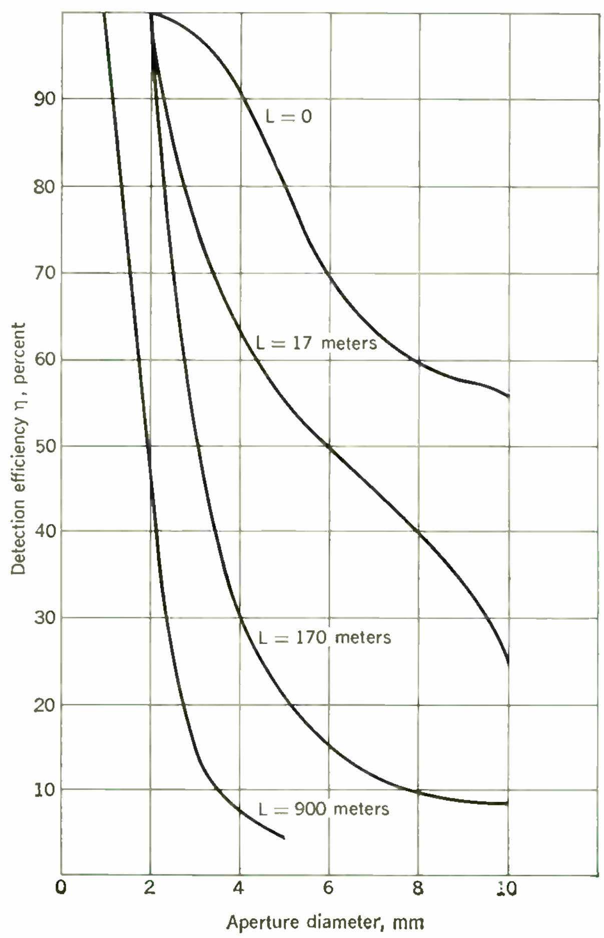


Fig. 6. Waveforms of homodyne system output for various signal path lengths.

Fig. 7. Detection efficiency vs. path length.



I. Receiver aperture diameter for 50 percent detection efficiency

Path Length, km	Average Path Elevation, meters	Maximum Diameter for 50 Percent Efficiency, mm	Reference
1	3	2	3
4	30	5	5
24	80	30	5

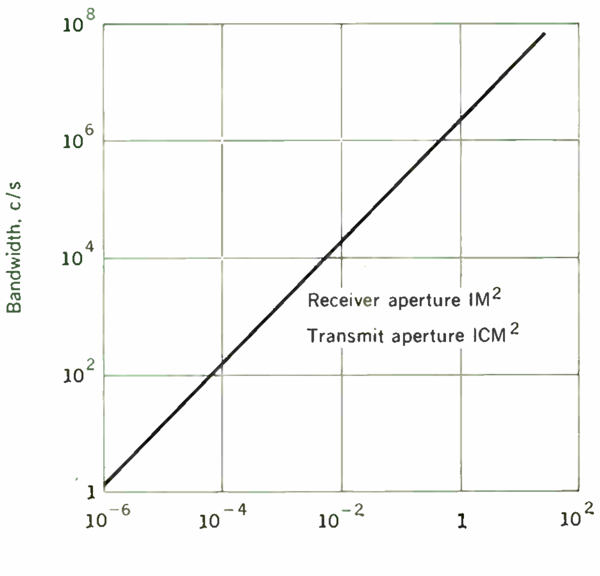


Fig. 8. Earth–Mars coherent optical link bandwidth vs. transmitter power. $\lambda = 6328$ Å.

Typical measurements of detection efficiency vs. aperture for various path lengths are presented in Fig. 7. These measurements were made over a path at an average height of 3 meters above the earth in clear air at temperatures between 80 and 90°F. Doubling the aperture reduced the short path efficiency slightly because of the lack of perfect laser coherence; the effect at 900 meters is considerably more marked. This fact demonstrates the cumulative effect of increasing path length on angular fluctuations and hence on coherence.

The effect of the path's height above the earth is also a major factor, as indicated by a number of experiments performed under varied path conditions. Some results of these experiments are given in Table I. They represent the aperture diameter at which further increase brings the detection efficiency below 50 percent. Figures are given for experiments performed over essentially horizontal paths at different elevations.

These empirical data show that the effects of atmospheric turbulence are greatest near the earth's surface, which is to be expected because of the large temperature gradients and small-scale turbulences found here. Changes in the earth's reflectivity and emissivity over relatively short distances contribute to these turbulences. These effects have a diurnal periodicity since their primary source of excitation is the sun.

Conclusions

Earthbound applications of coherent optical radiation for communications appear to be severely limited for two reasons. The first, and most significant, is the effect of atmospheric turbulence on the coherence of the radiation. The second is the effect of small vibrations on the coherent detection efficiency and signal-to-noise ratio. This can be minimized by careful design, but the first factor is beyond the designer's control. Although coherent optical detection has been demonstrated over some useful paths, the vulnerability of the link to atmospheric variations makes practical application somewhat doubtful. However, the advantages of bandwidth and range to be derived from communication at optical frequencies need not be lost, particularly in space and interplanetary applications.

Once outside the earth's direct physical environment, atmospheric effects, and many of the vibrational effects, are eliminated. The problem of beam pointing becomes somewhat greater, as does the problem of optical surface damage from micrometeorites. Nevertheless, although the turbulence of the atmosphere cannot be controlled by the designer, it is felt that he can cope with the new problems associated with space applications of optical communications.

With the solution of these problems will come the ability to establish useful communications between earth-orbiting relay stations and interplanetary probes. Figure 8 shows the theoretical capability of an optical link between an earth-orbiting relay and a Mars space probe. It is seen that with transmitter power levels of less than one watt, bandwidths of more than one megacycle per second are possible. In addition to the low transmitter power requirements, the required aperture sizes are very modest, and are quite suitable for satellite and space probe applications.

REFERENCES

1. Beran, M. J., and Parrent, G. B., *Theory of Partial Coherence.* Englewood Cliffs, N.J.: Prentice-Hall, 1964.

2. Biernson, C., and Lucy, R. F., "Requirements of a coherent laser pulse-Doppler radar," *Proc. IEEE,* vol. 51, pp. 202–213, Jan. 1963.

3. "Coherent optical propagation study," Tech. Rept. RADC-TR-65-313, AD 475743, Rome Air Development Center, Nov. 1965.

4. *Ibid.*, sect. 2.

5. Goldstein, I., Miles, P. A., and Chabot, A., "Heterodyne measurements of light propagation through atmospheric turbulence," *Proc. IEEE,* vol. 53, pp. 1172–1180, Sept. 1965.

6. "Investigation of coherent optical propagation," Tech. Doc. Rept. RADC-TDR-64-65, Aug. 1964.

7. Beckmann, P., "Signal degeneration in laser beams propagated through a turbulent atmosphere," *Radio Sci. J. Res.,* NBS/USNC–URSI, vol. 69D, Apr. 1965.

8. Gardner, S., "Some effects of atmospheric turbulence on optical heterodyne communications," *1964 IEEE Internat'l Conv. Record,* pt. 6.

9. Reynolds, G. O., and Skinner, T. J., "Mutual coherence function applied to imaging through a random medium," *J. Opt. Soc. Am.,* vol. 54, pp. 1302–1309, Nov. 1964.

10. Long, R. K., "Absorption of laser radiation in the atmosphere," Rept. 1579-3, The Ohio State Univ. Research Foundation, AD-410571.

11. Deirmendjian, D., "Scattering and polarization properties of water clouds and hazes in the visible and infrared," *Appl. Opt.,* vol. 3, pp. 187–196, Feb. 1964.

Frequency relations of parametric interactions

Some of the relationships of elementary mechanics are approached from the point of view of quantum methods, through the use of parametric frequency techniques

H. Seidel Bell Telephone Laboratories, Inc.

The parametric concept can be used to join conventional language to the language of quantum mechanics. In this article the philosophy of the parametric viewpoint is developed in terms of ordinary mechanics and the relevance of machines to energy levels is given in simple terms. Energy-level diagrams are employed to investigate various devices, which bear strong underlying relationships but have no superficial resemblance to each other. By the use of an energy-level characterization a new form of servomechanism is devised.

Two billiard balls on a collision path approach one another, collide, and rebound. With simplifying assumptions, the system of two bodies is only sectionally continuous. There is a first noninteracting region, a region of linear restoring force, and a final region of noninteraction. If we view the entire sequence of events, the restoring force of the composite system is highly nonlinear with respect to separation of the two bodies and, consequently, the system is anharmonic.

The description of this system in space and time can be transformed easily to a wavelength and frequency description. Since there is basic anharmonicity, there must be a consistency of the conventional language of dynamics with the language of wavelength and frequency mixing. Phrased conversely, the concept of mixing has inherent in it a means of characterizing dynamic particle interactions and, more generally, means of characterizing dynamics as a whole.

Let us now assume alternatively that the billiard balls had never been cut from the single embedding matrix of ivory containing them both. Though the identity of the balls becomes ephemeral we may, nevertheless, picture their motions as vibration packets, and wavelength and frequency seem less an artificial characterization. The restoring force, now, is quite linear and the packets pass through one another transparently or, alternatively, superposed. The loss of an evidence of collision is coincident with the loss of frequency admixture.

Michelangelo disclaimed the creativity of sculpture, asserting that it was no more than liberating the form frozen within the marble. So we argue that in the first example we have only released the balls that were immured within the "marble" of the second. We assert continuous existence and identity between the two states. It is not strange, within this context, that we seek a language common to both states. We claim, in the process, a greater understanding within a single language than within the two separate languages respectively limited to the two extreme states of identity.

The systemization of the study of interactions through wavelength and frequency descriptions is termed "Parametrics." This vastly inclusive title of the subject matter is, to the writer's mind, the appropriately stated distillation of the method originally entitled "The Variation of Parameters." The connection between the two is somewhat tortuous chronologically, but it may be understood in connection with the billiard ball example as the variation, in time, of the parameter of restoration between the two balls.

This article limits itself to the study of only the frequency relations of parametrics. We cannot consider finite-size particles within this self-imposed limitation, since the absence of wavelength denies us the criterion of physical size. We do, however, consider a more primitive particle not yet invested with the property of size: the photon. Photons (or, more generally, quanta) interact through a collision process, and the agency of collision is the generalization of the mechanical differential. If this metaphoric language seems forced, it has a single virtue: it works.

Here we shall dispense with the chronological and historical details in the development of parametric technology and attempt both to initiate and to motivate it from entirely different viewpoints. We shall initiate this study from considerations of elementary mechanics and show the relationship to quantum mechanics virtually from the beginning. It is hoped that the reader will find the approach neither makeshift nor tortured, but instead will find that it more clearly illuminates and unifies that which he already knows. Further, we shall use the tools of parametrics not only to improve hindsight but as a means of producing new devices. In this latter category we shall develop a new, entirely mechanical, servomechanical amplifier.

The mechanical differential

One of the great feats in the modern era of physics is Dirac's basic conceptual use of the harmonic oscillator to range over wide areas of physical study. The modest spring and mass combination becomes a tool of far greater utility and flexibility than one might have imagined. The study of parametrics is ultimately the study of reactive frequency mixing. The basic element here, much akin to the harmonic oscillator in its flexible utility, is the mechanical differential. Because of its universal applicability we shall attempt to understand its properties from fundamental considerations. Although the main result of these considerations will be what every mechanical designer knows, that all shaft torques of a symmetric differential are equal,* there are, nevertheless, subtleties that should be well understood before we make any attempt to generalize.

A mechanical differential is a three-shaft, ideally nondissipative device, which (to within inconsequential additive gearing) obeys the relationships

$$\omega_1 + \omega_2 + \omega_3 = 0 \tag{1}$$

$$P_1 + P_2 + P_3 = 0 \tag{2}$$

where 1, 2, and 3 are shaft labelings and where ω and P are shaft speeds and powers, respectively. Whereas Eqs. (1) and (2) are necessary requirements for a differential, there are two further requirements for complete specifications:

a. ω_1 and ω_2 are independent variables. This would not be true if shafts 1 and 2 were directly geared to one another.
b. The quantities P_1 and ω_1, etc., are assumed to correspond to one another such that an unambiguous, nonvanishing limit P/ω exists as ω tends toward zero.

With all of these assumptions, and as derived in the appendix, we find the purely mathematical result

$$\frac{P_1}{\omega_1} = \frac{P_2}{\omega_2} = \frac{P_3}{\omega_3} \tag{3}$$

which is the statement of equal torques on all shafts of a symmetric differential.

To show the range of validity of (3) it is desirable to show situations in which the above assumptions are violated and in which (3) may or may not apply. Let us, for example, violate the condition that ω_1 and ω_2 are necessarily independent. An interesting case in which ω_1 and ω_2 are quite often dependent is where they bear an exact rational relationship to one another. There is a strong suggestion that these frequencies are bound through harmonic relationships, since rational relations are infinitely rare in the normal frequency of occurrence compared to irrational ones. Nevertheless, ω_1 and ω_2 may be made to pass through an exact rational relationship in a differential and, during such an occurrence, the differential may be equivalently replaced by a rigidly geared system without any distinguishability of power–frequency–torque relationships. We find, therefore, that the mere fact that frequency relationships are rational does not necessarily disqualify Eq. (3), but serious questions nevertheless are encountered.

An example of inapplicability of (3) occurs in the simply contrived alternative case to that of a chance rational relationship. Consider a gear box in which there are three accessible shafts in the rational speed ratios 1:2:3 so that we satisfy the frequency condition: $1 + 2 + (-3) = 0$. Use the shaft having the relative speed of 3 as the drive shaft and load the shaft at speed 1 with a torque T_1 and the shaft at speed 2 with a torque T_2. Conservation of power requires that $T_3 = {}^1/_3(T_1 + 2T_2)$. This result is trivial, but with a major consequence: The torque equality conditions of (3) are not generally applicable. Nevertheless, we may also include (3) here as the special case in which $T_1 = T_2$, where we find that T_3 is equal to the other two. The condition of torque equality therefore represents a meeting point for both linearly dependent and independent frequency relationships.

Just as a rational frequency relationship is no guarantee that Eq. (3) is inapplicable, neither is an irrational frequency relationship a guarantee that (3) is applicable. The gear box in the foregoing example may be replaced by wheels that roll on each other without slippage. These wheels may typically produce speeds in ratio $1:\sqrt{2}:(1 + \sqrt{2})$. Here again, the shaft speeds sum to zero and power conservation insures the power condition. Nevertheless, there need not be torque equality and, again, (3) does not apply in general.

The second restriction that P/ω tends toward a nonvanishing nonsingular limit as $\omega \to 0$ has a more subtle physical consequence under violation. Actually, P is simply a variable obeying a conservation law. Momentum, for example, might have been chosen as an alternative variable. However, the momentum-to-frequency ratio is not necessarily meaningfully defined in the limit and, again, (3) might not apply. This situation occurs in the study of particle collision, where the particle, in Newtonian mechanics, has a parabolic frequency (energy) momentum dispersion and interaction occurs without preconditions on the momenta. On the other hand, when

* The shaft torques of an arbitrary differential are actually in a constant ratio. However, where the linear dependency is equally weighted among the three frequencies, as in (1), the ratio is identically unity.

there is no dispersion, (3) does indeed apply in the form of the Suhl–Tien[1] relations. Because of a necessary limitation of the subject matter of this article, we shall not pursue the study of particles within the parametric framework of description. However, it does seem important to produce this insight into the physical relevance of (3).

Within the framework of the examination of the differential we may observe certain characteristics given great prominence in the study of parametric systems. As anyone who drives knows, it is impossible to start a car with one of the rear wheels completely slipping on ice. From (3) we observe that if no power is transmitted into one of the shafts no torque may be developed at any other shaft. This is a feature characteristic of the idler. In the case of a three-frequency process, if one of the frequencies in a reactive conversion process is developed as a by-product it must, nevertheless, be loaded according to the prescription of (3) if power is to be transferred between the other two.

Let us consider power to be positive if it is delivered by one of the shafts and conversely to be negative if it is delivered into a shaft. We may then observe two cases whose designations anticipate parametric terminology: upper and lower sideband conversions.

Lower sideband (difference mode). In (1) choose $|\omega_1| + |\omega_2| = |\omega_3|$. There exist two possibilities: Either ω_1 and ω_2 are both positive and ω_3 is negative, or the reverse. In the first case, power at ω_3, which is negative by virtue of (3), feeds into the differential and power feeds out at ω_1 and ω_2. The differential is an energy sink at ω_3 and an energy source or, equivalently, a negative resistance at both ω_1 and ω_2. With a sign reversal, energy is fed into the system at ω_1 and ω_2, and ω_3 finds the system a source.

In a simple view, the drive shaft absorbs energy from the engine and emits it into the two rear wheels. Under deceleration, the rear wheels feed energy back to the drive shaft.

Upper sideband (sum mode). In (1) choose $|\omega_1| + |\omega_3| = |\omega_2|$. Here, to within sign reversal, ω_1 and ω_3 are negative and ω_2 is positive. Energy is drawn from both ω_1 and ω_3 and delivered at ω_2. As applied to the example of a car, not only is the engine supplying power to the drive shaft at ω_3 but shaft 1 is being spun in a direction opposite to that in which it would tend to be driven. Therefore, power is supplied by the wheel as well as by the drive shaft and for this reason all the power supplied is emitted at higher speed into the other wheel at shaft 2.

Torque and quanta

Equation (3) suggests a quantum transition interpretation that is generally not considered to be meaningful in application to macroscopic phenomena. We shall take the tack here that the language of quanta is entirely meaningful in this context and, indeed, the whole force of the Manley–Rowe relationships is that this view is supported. We shall explore precisely this area in subsequent sections.

Let us view the properties of the differential alternatively, assuming the applicability of a quantum language. In general, a power P is given by the number per second n of quanta of energy $\hbar\omega$ being absorbed or emitted. Therefore,

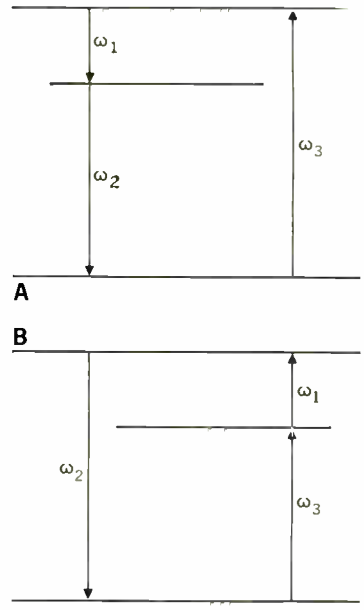
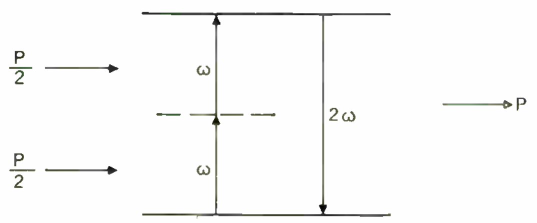


Fig. 1. Three-frequency transitions. A—Lower sideband. B—Upper sideband.

Fig. 2. Intermediate state of a 2:1 gear.



$$\frac{P}{\omega} = n\hbar \tag{4}$$

where $\hbar$ is Planck's constant divided by 2π. We may then rewrite (3) as

$$n_1 = n_2 = n_3 \tag{5}$$

The transition of quanta may be illustrated in the energy-level diagrams of Fig. 1.

The transition diagrams of Fig. 1 represent a steady-state process, so population at any energy level is a constant. For example, in Fig. 1(A), n quanta per second delivered by the pump ω_3 from the ground state to the upper level must cause n quanta to descend to the middle level and thence back to the ground state.

Equation (5) is, therefore, the statement of a steady-state quantum population, as are precisely all of the Manley–Rowe relationships. Power absorption and emission are given simply by the continued circulation, respectively, of upward and downward transitions.

Geared differentials

We have just observed that the power-to-frequency ratio of two generally unequal frequency inputs is a constant in a three-shaft differential. As mentioned previously, we expect and observe no untoward event as the two frequencies pass through identity. In this circumstance the output shaft records double the power at twice the frequency of either input shaft and the statement of torque conservation is unviolated.

Let us consider, on the other hand, that a shaft rotating at the same frequency as one of the input shafts above is applied to a 2:1 gear train. While the same power and frequency emerge at the output shaft as in the description above, torque is not conserved since the input requires twice the torque. Gearing, therefore, is not as evidently amenable to a direct quantum counting scheme as is the differential. We may, however, introduce the artifice of the intermediate state, as shown in Fig. 2, so that the gear, like the two shafts of the differential, might be made to appear to pass through two transitions. In Fig. 2 we have reconciled our quantum counting scheme, since the transition to the virtual, or intermediate, state has the same number of quanta, $(P/2)/\omega$, as does the emissive second harmonic, which produces $2(P/2)/2\omega$ transitions. Further, since twice as many quanta are absorbed from the field at ω as are emitted at 2ω, the torque at ω is twice that at 2ω.

In general, N:1 gearing may be considered to produce $N - 1$ intermediate states as the price for making quantum language applicable, as shown in Fig. 3.

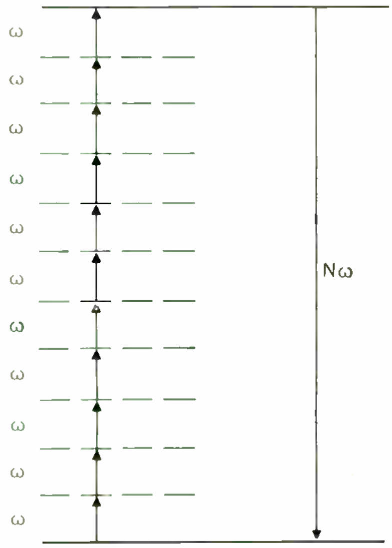


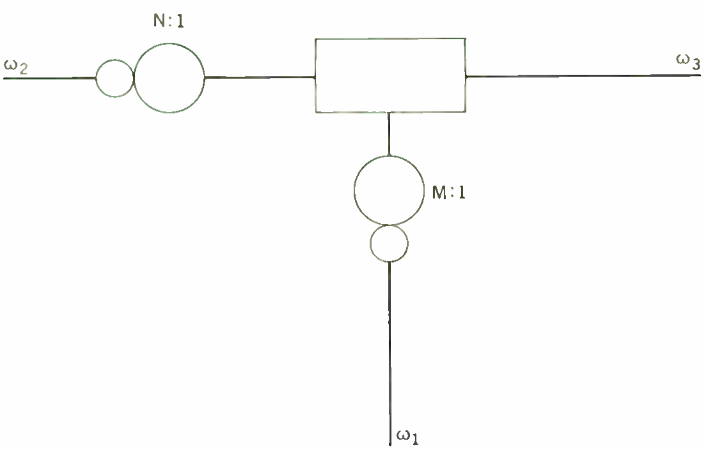
Fig. 3. Quantum picture of an N:1 gear.

Let us now consider the geared three-shaft differential shown in Fig. 4. The torque of Eq. (3) still applies perfectly to the shafts directly entering the differential but must be modified by the gearing ratios to be made applicable to the actual input shafts. Taking the gearing into account, we have

$$M\omega_1 + N\omega_2 + \omega_3 = 0 \qquad (6)$$

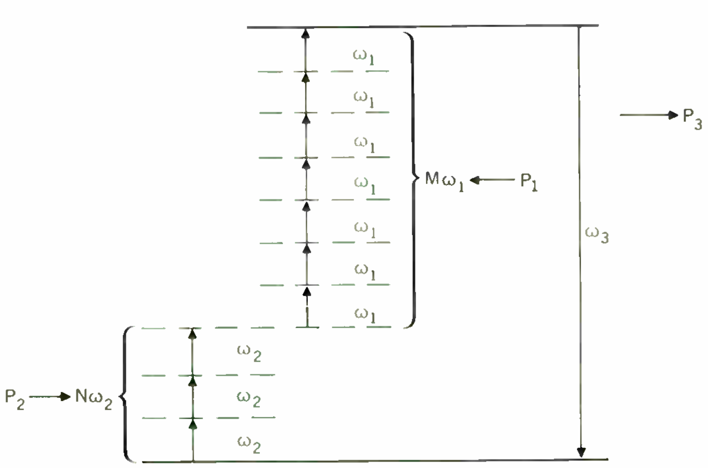
$$\frac{P_1}{M\omega_1} = \frac{P_2}{N\omega_2} = \frac{P_3}{\omega_3} \qquad (7)$$

Fig. 4. Differential with geared input shafts.



The upper-sideband diagram, as an example of a geared differential, now has the appearance shown in Fig. 5. The relationship of (6) and (7) to this figure is clear. The number of photons engaging in absorptive transitions among the intermediate levels at ω_2 is proportional to $(1/N)P_2/\omega_2$, whereas the number of absorbed photons having transition among these levels at ω_1 is proportional to $(1/M)P_1/\omega_1$. Steady-state equilibrium prevents accumulation of population at any level, so these quantities must be equal. These are equal in turn to the number emitted at $\omega_3 = -(M\omega_1 + N\omega_2)$, which is proportional to P_3/ω_3. The energy-level diagram thus continues to provide proper agreement with the torque equations.

Fig. 5. Upper sideband production in geared differential.



Multiple-shaft differential assembly

It is possible to build up a complex of three shaft differentials by first fanning out each input shaft into several directly geared branches, then interconnecting the branches from different shafts in pairs through differentials, and next fanning out the output shaft of the respective differentials in turn in a continuing process. Let us form a specific array and label it.

Employ a gear ratio $1:N_i(1)$ on the shaft of ω_1 and a gear ratio $1:N_i(2)$ on shaft ω_2 to produce an emission frequency $\omega_3 = -[N_i(1)\omega_1 + N_j(2)\omega_2]$. The torque on each shaft of the combining differential is P_3/ω_3. If to the emission at shaft 3 is added differentially some shaft at speed ω_r with a gearing $1:N_k(r)$, then a new emission takes place at $\omega_4 = -[N_i(1)\omega_1 + N_j(2)\omega_2 + N_k(r)\omega_r]$ with a torque P_4/ω_4 that is identical to P_3/ω_3 because of the constancy of the differential's torque. In general, by applying an arbitrary number of differentials we may produce an output frequency (with a small but obvious tidying up of notation)

$$\omega^{1,2,3,4,5,\ldots}_{i,j,k,l,m,\ldots} \equiv -[N_i(1)\omega_1 + N_j(2)\omega_2 + N_k(3)\omega_3 + \ldots] \quad (8)$$

with the torque contribution of shaft 1 equal to the torque at the very end of this differential chain modified only by the multiplication of its gearing constant $N_i(1)$. The total torque on shaft 1 is the sum of the torque contributions in which it is involved and this must be zero since there is no shaft acceleration if we hypothesize a steady state. Hence,

$$T_{1-\text{total}} = \sum N_i(1) \frac{P^{1,2,3,4,5,\ldots}_{i,j,k,l,m,\ldots}}{\omega^{1,2,3,4,5,\ldots}_{i,j,k,l,m,\ldots}} = 0 \quad (9)$$

Since shaft 1 is in no way physically more prominent than the others, (9) holds equally true in kind for shafts 2, 3, etc., in succession. Hence (9) represents a simultaneous set of equations over the set of all the base frequencies that are present. These Manley–Rowe[2] equations are considered to be the most basic relations of parametrics.

In deriving these relations we have shown that they are identical to those corresponding to a constant-population quantum picture; although the relationships are entirely macroscopic, they permit exact quantum phrasing.

We have built up the frequencies in this system by a very specific means. Is it possible to build some other reactive system with identical output shaft speeds emanating from a given set of independent speeds which would have power relationships inconsistent with those just developed and which would not satisfy the quantum picture? This question is examined in detail elsewhere,[3] and in all but highly contrived circumstances it would seem that the relationships are unique and that the quantum picture as described by the specialized use of gears and differentials is correct.

A point of some interest comes to light with respect to the manner of derivation. It is generally considered that the parametric process comes about through the existence of nonlinear processes, yet the foregoing derivations were quite oblivious to specific statements of nonlinearity. Rather than any statement of nonlinearity, the essence of the result was that an incoming power at some frequency made transitions to power flow at other frequencies. Anticipating later results, frequency transition is energy transition; therefore, the appropriate statement for the parametric process is that it is a language for representing collisions in terms of frequency transitions. To the extent that collision requires anharmonic processes, nonlinearity is implicit in linear motion but, as we have observed in respect to the differentials, not in circular motion.

There is yet another feature of interest in the development. All frequencies ultimately absorbing or emitting power were arrived at by a systematic combination of two frequencies at a time (Ritz combination principle). The final transition relationships were independent of the algorithm of construction. This construction through partial steps has the physical description of the "temporary" existence of an intermediate or a virtual state. Although the fiction is often employed that the system resides in an intermediate state for a while, one can associate no actual chronology with such an occupation, since there is actually no time sequence in what we observe to be a steady-state phenomenon. The major virtue of the virtual state is that, from the quantum rate viewpoint, a multibody interaction may be viewed as a system of two body collisions.

We now assert via the statement on uniqueness that the relationships among the various differentials are universal in describing interaction dynamics since, in principle, one may employ transducers that conserve power and frequency in their translation from some particular energy to the rotational mechanical form used in the differential. The importance of this assertion is that one may perceive relationships relating to the rate of transition between states without ever precisely defining what the states are. This is the essence of the principles of second quantization, and the Manley–Rowe relationships are the general relationships produced by the use of second quantization in quantum mechanics.

Engineering examples of parametric language

We next seek not only to encourage confidence in the use of parametric language by examining devices of hard-headed practicality but also to reinforce earlier statements that the parametric approach produces greater insight and simplicity. We shall examine both the three-phase and single-phase induction motors. As a special case of the induction motor, we shall consider the synchronous motor and give practical meaning to the distinction between commensurable and incommensurable frequency relationships.

Three-phase induction motor. In a three-phase motor the three windings on the stator are progressively phased so that a rotating magnetic field will be produced. We assume for simplicity that the electrical angle between windings is the same as the geometric angle, and thus the magnetic field rotates at the same frequency as the stator excitation. The rotor is a short-circuit winding, which is free to rotate and which couples to an output shaft.

The rotating field of the stator induces short-circuit currents in the rotor; these currents, interacting with the stator field, produce an output torque. As the shaft rotates, the rotating magnetic field intercepts the rotor winding at a rate reduced by the shaft speed. If ω_{st} is the stator radian frequency, ω_r the slip frequency induced in the rotor winding, and ω_{sh} the shaft frequency, we observe that

$$\omega_{st} - \omega_r - \omega_{sh} = 0 \quad (10a)$$

Assuming a dissipative power P_r in the rotor coils, we have

$$P_{st} - P_r - P_{sh} = 0 \quad (10b)$$

Equations (10a) and (10b) describe a three-level system to which the relations (3) are applicable, and we have the familiar result that

$$\frac{P_{st}}{\omega_{st}} = \frac{P_r}{\omega_r} = \frac{P_{sh}}{\omega_{sh}} = T_{sh} \quad (11)$$

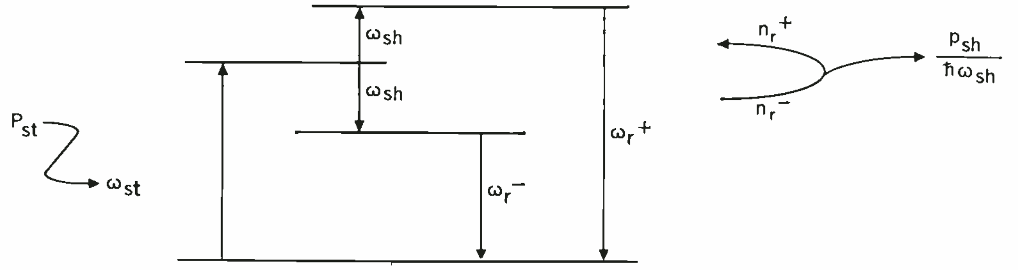


Fig. 6. Diagram of single-phase induction motor.

Equation (11) shows a relationship that, while now common in conventional parametrics, is nevertheless quite startling in its embodiment in a piece of rotating machinery. It shows that the induction motor operating with some slip ($\omega_r > 0$) must have a dissipative rotor to develop shaft torque.

Equation (11) shows that only when there is extremely small slip, namely, when $\omega_r \simeq 0$, can a low-resistance rotor develop nominally large output torque. This, however, does not explain how a well-designed machine can develop starting torque; i.e., where $\omega_r \simeq \omega_{st}$. The answer is that the idler must be loaded much as any other difference-mode three-level system—resistance must be inserted in series with the rotor for starting purposes.

The use of parametric language here produces so facile a result that one feels impelled to seek to understand it with conventional, if more clumsy, language. If the rotor is short-circuited, it produces an effect quite identical to that of a short-circuited transformer. If the short circuit is resistance-free, the primary winding presents a dead short circuit and no input power may be developed. If, on the other hand, the rotor were open-circuited, one could still produce no primary power. The only means of delivering primary power to move the shaft off its stall position is by terminating the rotor in a resistance, and since the shaft torque maximizes with rotor dissipation in (11), the rotor resistance must produce a match for maximum torque.

Single-phase induction motor. The single-phase motor is more complex than the three-phase machine, and the former must be represented as a four-level system in contrast to the three levels of the latter. Here a stator frequency ω_{st} is applied, and the rotor, interacting with the clockwise and counterclockwise rotating field resolutions of the stator field, produces slip currents at $\omega_{st} \pm \omega_{sh}$. We thus have the diagram of Fig. 6.

Direct inspection of Fig. 6 reveals that the single-phase motor must yield power to rotor dissipation at frequencies ω_r^- and ω_r^+ defined, respectively, as $\omega_{st} \mp \omega_{sh}$, if torque is to be obtained at the shaft. Let n_r^+ be the number of quanta per second expended at ω_r^+ and let n_r^- be that for ω_r^-. Assume that the rotor armature reaction dominates its resistive effects. The induced current I corresponding to + or − is proportional to the ratio of a frequency-linear induced voltage and a frequency-linear reactance and is, therefore, a constant. The single-phase motor acts, therefore, as a current transformer and a frequency changer simultaneously with respect to both the upper and lower rotor levels. For either upper or lower rotor level, assuming a frequency-independent rotor resistance,

$$P_r \sim \text{const.}$$

and

$$n_r \sim \frac{P_r}{\omega_r} \sim \frac{1}{\omega_r}$$

Therefore, especially,

$$n_r^+ \sim \frac{1}{\omega_r^+}$$

$$n_r^- \sim \frac{1}{\omega_r^-}$$

The output torque at the shaft is directly proportional to the number of quanta that are emitted into the mechanical field.

$$n_{sh} = \frac{P_{sh}}{\hbar\omega_{sh}} = n_r^- - n_r^+ = n_r^-\left(1 - \frac{n_r^+}{n_r^-}\right)$$

Therefore,

$$T_{sh} = \frac{P_r^-}{\omega_r^-}\left(1 - \frac{\omega_r^-}{\omega_r^+}\right) \tag{12}$$

As we observe from (12), the single-phase induction motor is twice bedeviled in developing the starting torque. It must produce dissipation to an adequate degree in the stalled rotor and the degenerate equality of ω_r^- and ω_r^+ must be split by some other starting mechanism to produce a nonvanishing factor.

The efficiency of the motor η is the ratio of the shaft power $(n_r^- - n_r^+)\omega_{sh}$ to the stator power $\omega_{st}(n_r^- + n_r^+)$, providing

$$\eta = \frac{\omega_{sh}(n_r^- - n_r^+)}{\omega_{st}(n_r^- + n_r^+)} = \frac{\omega_{sh}(\omega_r^+ - \omega_r^-)}{\omega_{st}(\omega_r^+ + \omega_r^-)} = \left(\frac{\omega_{sh}}{\omega_{st}}\right)^2 \tag{13}$$

Equations (12) and (13) are, of course, ideal. One assumption that fails, for example, as stator frequency and shaft frequency approach one another is that the reactance corresponding to ω_r^- dominates the resistance. Nevertheless, the simplicity of calculating the approximate result is apparent irrespective of higher-order effects.

If in Fig. 6 the shaft had been used to drive power into the system, as opposed to receiving power from it, the diagram would be that of an induction generator emitting into upper and lower sidebands ω_r^+ and ω_r^-, respectively. To this extent the single-phase induction motor and the double-sideband modulator are highly related in their essential descriptions, though with some modification in the assumptions concerning the magnitudes of resistive and reactive effects. Since a double-sideband modulator might be, typically, a magnetic amplifier, relationships may be appreciated within the energy level description that would not be apparent in other contexts since particularly a magamp and an induction motor are not generally considered to be of the same genre.

Physical meaning of commensurable frequencies

In our examination of the three-frequency differential we observed that the proofs almost always failed for the various frequencies bearing a rational relationship. Since one would not necessarily expect physical laws to be essentially different in going from a rational to an irrational relationship, we must assume the difficulties of a rational relationship to be more subtle than an actual failure of power-frequency relationships.

We shall initiate our examination by studying the synchronous motor as representing the limit of the induction motor as it approaches synchronism. At synchronism there exists a rational relationship between all the frequencies of the system. We shall find the limit to be a nonuniform process, one requiring considerable pains to define. This difficulty, in turn, may be related to certain difficulties in the Manley–Rowe developments, which we find to bear strongly on the indeterminacy of simultaneous quanta and phase measurements. Finally, we shall return to the synchronous motor and determine what nature itself does to produce what we might consider to be an acceptable compromise of the limiting difficulties.

Synchronous motor. Let us take either the single- or three-phase motor of the previous section and gradually reduce its rotor resistance. To operate within the bounds of the parametric language we are developing here, we must take care to allow adequate equilibrium time after each adjustment to permit a substantially steady state.

Let us assume operation close to synchronism so that rotor resistance dominates rotor reactance. The power dissipated in the rotor winding is proportional to the square of the voltage induced and inversely proportional to its resistance. Since, for constant stator voltage, the voltage induced is proportional to the slip frequency we find that

$$P_r^- \sim \frac{(\omega_r^-)^2}{R}$$

For ω_r^- sufficiently small, the level ω_r^+ enters trivially in the quantum transition balance. We need only consider n_r^-, which yields

$$T_{eh} \sim \frac{\omega_r^-}{R} \qquad (14)$$

If we assume that the limit of $R = 0$ can be achieved, Eq. (14) shows that an output torque may be yielded without slip, assuming the indeterminacy in turn to yield a generally nonvanishing limit. Since R is indeed zero, the rotor currents persist infinitely and the rotor becomes, simply, a synchronously rotating magnetic dipole of constant moment. Evidently this constant-field rotor may be replaced with any equivalent nonsuperconducting dipole, and we are then able to realize the conventional synchronous motor having a permanent magnetic rotor.

It is all too clear that, however correct the end results, the questions of the existence of a limit in (14) are major. Let us construct a process for attempting to attain this limit physically. Let us assume that the steps in changing R are viewed on a vastly magnified scale but, nevertheless, assume the number of steps to be finite. In the last step, R goes from a finite value to a value of zero. In this last finite resistance position, ω_r^- had attained some finite value consistent with the torque it was called upon to produce. When R is finally switched to zero, ω_r^- cannot instantaneously switch to zero because this would require infinite torque. Infinite torque, as seen from the results of the section on torque and quanta, is equivalent to an infinite quantum transition rate in the stator. Under the constant input voltage assumptions, then, a short-circuited stator input is implied. Thus, since the stator presents a short circuit, given a finite source impedance of the power bus, the result is that no torque at all is produced, as opposed to the infinite torque demanded.

The variation to zero resistance of the rotor must then be made in even smaller steps and each step takes a successively longer equilibrium time because of the ever-increasing severity of the reflection transients at the stator input. There is, therefore, no uniform limiting procedure in the simple model and, indeed, a synchronous motor is bistable, it is either synchronized or it does not run at all. It is well known, however, that a synchronous motor can be brought to speed through an induction rotor winding, although the latter must slip and cannot produce exact synchronism. The lock-in mechanism demands the appearance of other "energy levels," which we discuss at the end of this section.

Noncommensurability and Manley–Rowe. The failure to produce a uniform limiting procedure in viewing the synchronous motor as an end result of the induction motor represents a feature of a commensurable frequency relationship. This failure, which is clearly measured in the unlimited equilibrium time required as commensurability is approached, is perhaps most clearly brought out by examination of the original Manley–Rowe development.

The Manley–Rowe treatment concerns the mixing of two frequencies, ω_1 and ω_2, in a nonlinear reactance. In evaluating interaction terms one is required to perform a typical average

$$\langle \exp i(m\omega_1 + n\omega_2)t \rangle$$

For the most part this average is clearly zero, and its vanishing takes place in a time that is of the order of $1/(m\omega_1 + n\omega_2)$. The vanishing of the exponential on the average is important in defining power and, equivalently, quantum flow.

Suppose, however, that there is an exact rational relationship between ω_1 and ω_2 given in terms of two relatively prime integers p and q such that $p\omega_1 + q\omega_2 = 0$. Since m and n can take on, respectively, those particular values p and q, there is an exponent value of zero, and the average clearly does not vanish. This provides an equally clear error in our quantum bookkeeping in our manner of counting the m quanta associated with ω_1 and the n quanta with ω_2.

If we know that p and q are actually integers we may know enough to take this fact into account. This is precisely what we did in the section on geared differentials. A trickier problem exists if we are not sure that p and q are perfect integers and, indeed, they may not be. To find out we have to wait long enough to see if there is any phase variation in $(m\omega_1 + n\omega_2)t$. Even if some phase variation does develop in time, perhaps we have not taken p and q to a large enough number of places to have found that commensurability actually does exist.

To sum up, our indeterminate knowledge of the exact number of quanta involved in transition is inversely proportional to the accuracy to which we know our developing phase picture. This statement comes close to a statement of indeterminacy, prominent in quantum mechanics, that

$$\Delta n \Delta \varphi \sim 1 \qquad (15)$$

where Δn represents the indeterminacy in quantum count and $\Delta \varphi$ represents the phase indeterminacy of each quantum.

A physical insight. The limiting case of the induction motor was but one situation leading to a dilemma. The use, by both Von Neumann[4] and Goto,[5] of the $360°/n$-phase indeterminacy of the nth-order subharmonic oscillator as the parametron logic element is a case where the limiting difficulties were put to use. Another well-known case in point is the difficulty in using a strictly degenerate parametric amplifier; there is an infinite range of output power dependent on the phase of each lower sideband. The difficulties of defining the equilibrium properties of the commensurable frequency parametric system are far from a purely mathematical conundrum.

Still, it is questionable at this point whether the distinction between commensurability and noncommensurability is not just a quibble, since one would not ultimately expect physical phenomena to relate strongly to a purely mathematical question of transfinites. Is there a meaningful physical statement one can make? The reactive mixer provides a good insight as to a means for forming practical evaluations. In most reactive mixer applications harmonics are generally not strongly generated for values of m and n beyond the order of 10; certainly not beyond 100. If higher m and n values are required to produce a zero beat or close enough to it, we can safely discard our concerns for the order of indeterminacy involved.

In the case of the synchronous motor, $\omega_1 + \omega_2 = 0$ $(p = q = 1)$; in the case of the degenerate parametric amplifier, $\omega_1 + \omega_2 = 2\omega_1$ $(p = -q = 1)$. For these cases the order of the commensurable relationship is quite low, and there are very serious consequences in lightly considering the amount of indeterminate power involved. The decision, therefore, as to the physical significance of commensurability lies entirely within the orders of the transitions involved.

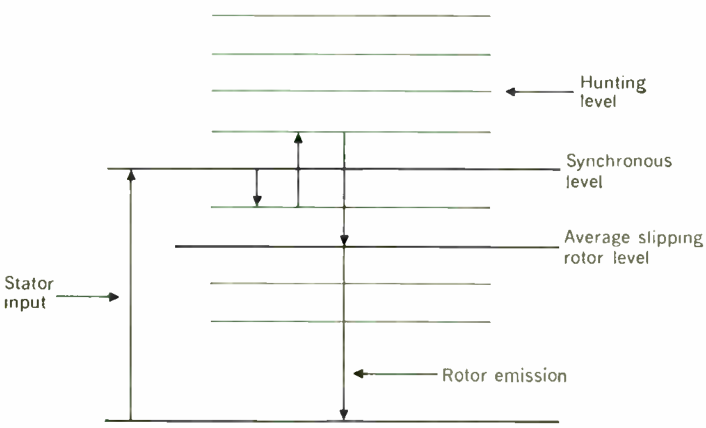
Lock-in: a multilevel process. Having expended all this effort to prove that there is a problem in defining a rational frequency-ratio limit, we have the same problem as that of showing that the seemingly aerodynamically unstable bumblebee can indeed manage to fly. We are saddled with the embarrassing knowledge that induction-start synchronous-run motors do work—and synchronously at that.

The key to the stability of the synchronous motor lies in its ability to hunt. If we regard the synchronous rotor relatively, as an observer traveling at synchronous speed, we will observe a magnetic dipole, the rotor, aligned in its minimum-energy state with the now-stationary stator field. Since the minimum-energy state is, however, only a zero-restoring force equilibrium, there exist nonequilibrium harmonic oscillations about this state. Because these harmonic oscillations relate to sinusoidal oscillations of the rotor frequency, these levels represent in actuality FM sidebands about the synchronous frequency.

In ordinary language we say that the synchronous motor locks in by a process in which the slowly slipping rotor manages to grab onto the synchronously rotating stator field as it passes by. However, with our description of the synchronous motor as a multilevel system because of its hunting, we may provide a more complicated, but more accurate, parametric picture. From the physical picture, we know that the stator frequency initially exceeds the rotor angular speed, but that after lock-in there is an overshoot of rotor speed. In Fig. 7, the synchronous stator excitation has a high "probability" of transition to a nearby, but lower-lying, hunting energy level, and emits quanta into it. By harmonic transitions, this lower-lying state feeds energy into a higher-lying state and produces the observation of overshoot. From this higher-lying state, quanta are emitted into the slipping rotor by a three-energy-level process and its speed mounts. As its speed passes through synchronous, the rotor may receive energy directly from the synchronous excitation. Although the hunting levels no longer serve a vital function, they do provide a means of stabilizing the system under fluctuations, even under equilibrium conditions.

Certain features of the foregoing description require greater elaboration. The use of the word "probability" in describing a transition that is a significant process is merely one of language convenience and in no sense implies that the process is statistical or indeterminate. The phase "transition strength" might well be more descriptive and accurate. Moreover, in connection with

Fig. 7. Synchronous-motor levels.



the transitions between the hunting levels, absorption processes must be incurred, particularly in the final three-level process, which adds energy to the rotor to permit its synchronization. The need to account for this absorption results in the contention that there are large induced low-frequency circulating currents in either the stator or the rotor, operating into a low residual resistance. Inadequacy of such resistance may well provide an inability of the system to synchronize.

We observe that the mathematically difficult limiting process is resolved by the appearance of incommensurable frequency levels not postulated in the simplified theory. These levels serve to perform a nontrivial and complex function in that they produce the final synchronization.

Since it is the intent here to show the general applicability of the parametric process, we employ for exposition a parallel problem in the synchronization of a subharmonic oscillator. For example, a degenerate parametric oscillator is one in which a varactor is pumped at ω and the system oscillates at $\omega/2$. There is nothing sacred about one half the frequency; a saturated magnetic system, describable by Duffing's equation, emits at $\omega/3$. However, let us be concerned here only with the half-frequency system.

Let us consider a degenerate parametric amplifier in which there is some negative resistance bandwidth about $\omega/2$. A resonant reflection in this frequency region placed in the amplifier load path creates a regenerative and oscillatory system in which two frequencies are present, $\omega/2 + \delta$ and $\omega/2 - \delta$, where δ is small and is determined by the reflection resonance. Experiment shows that this split oscillator is indeed constructable and that the split is controllable by varying the resonance frequency of the obstacle. Nevertheless, there exists a frequency range, for which δ is less than some limiting quantity, where the system displays hysteresis; with small perturbation the unequal frequencies merge and lock together at $\omega/2$.

The situation here bears strong resemblance to the case of the synchronous motor. The locked state is correlated in phase and forms a preferred minimum-energy state. Nevertheless, the process of synchronization is complex and, in identical language, the activation energy to produce this state is obtained only through the FM or hunting states of the oscillator.

In summary, the commensurable frequency difficulties in the Manley–Rowe developments relate to the physical distinguishability between phase correlated and uncorrelated states and the large energy jumps intervening between them. The energy jump requires an activation process, which requires, in turn, a multilevel hunting procedure for synchronization.

A concluding example of frequency relations

It is always an author's hope in writing that his exposition has been so clear and persuasive that it sufficed to provide understanding. Although this hope is still nurtured, at this point it seems desirable to provide one last example, which contains the essence of all that

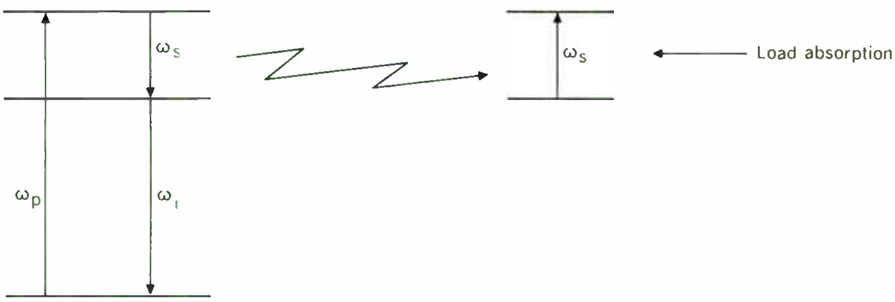


Fig. 8. Three-level diagram of mechanical amplifier.

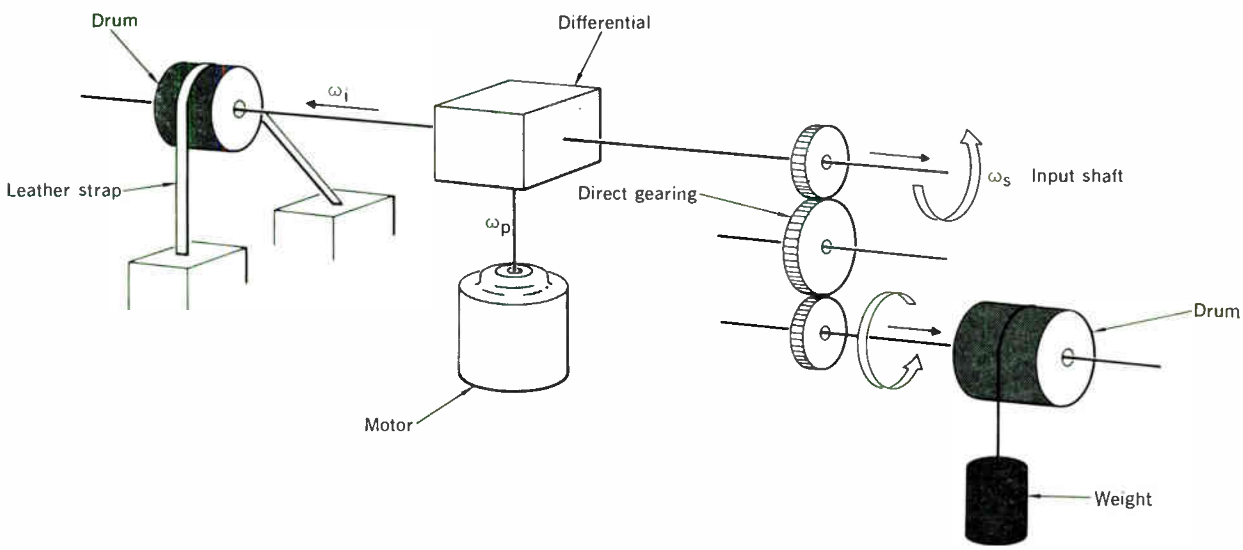


Fig. 9. Mechanical equivalent of Fig. 8.

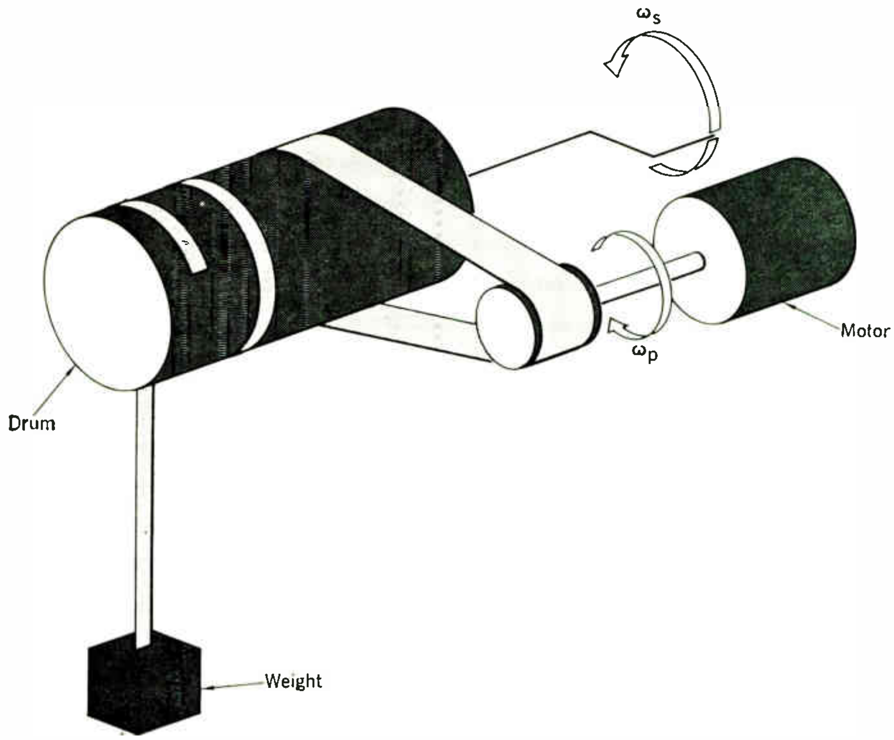


Fig. 10. Drag belt amplifier.

has been said before. The simplicity of the example and the essential logic of its operation should provide the final link between the structure of the parametric formalism and its physical meaning.

A three-energy-level mechanical amplifier. We shall first consider the development of an amplifier from a strict use of the formalism we have assembled thus far. We shall then reconsider it from a purely mechanical viewpoint and show the interrelationship of the various languages of description.

Let us seek to implement a purely mechanical amplifier using a motor as a pump. The signal will be represented as a torque applied to an input shaft. We may use the energy-level scheme of Fig. 8 as the basis for construction.

Figure 8 is a conventional three-level diagram used in a similar fashion to our past usage. An applied pump creates equal quantum emission into the frequency field at ω_s and into the idler field at ω_i. The emission into ω_s is used to furnish torque into a load at ω_s. Figure 9 shows an implementation of Fig. 8.

In Fig. 9 an output load is provided by a weight hanging on a drum, providing a torque T_s. The idler is loaded by a Prony brake arrangement in which a leather strap provides a frictional torque T_i. If we recall that torques are proportional to quantum transition rates, the three-level amplifier requirement for equal idler and signal transition rates, or $T_i = T_s$, must be met. The idler frequency is provided through a differencing of the motor-shaft frequency ω_p and the signal-shaft frequency ω_s in a differential.

The implementation in Fig. 9 is a simple one, and a laboratory model was constructed showing it to work in exactly the fashion contemplated. Very little input torque was required to lift relatively heavy weights suspended on the drum when the Prony brake tension was appropriately adjusted.

A little further thought shows that Fig. 9 might be simplified in a manner shown in Fig. 10, in which a motor drives a belt wrapped around a cylindrical drum. Suspended from the drum is a weight, which produces a load torque T_s. The input signal at ω_s is applied to a coaxial shaft.

The torque on the slipping belt is made equal to the load torque T_s as in the case of the mechanical system of Fig. 9. We see in Fig. 10 all the features of the three-level diagram of Fig. 8. There is a signal frequency ω_s, a pump frequency relative to the contact region of the drum of ω_p, and a slip or idler frequency ω_i. The torque provided by the slipping belt T_i is equal to the motor torque T_p—which, in turn, is equal to the load torque T_s at equilibrium.

We arrived at Fig. 10 by the formal language of parametrics, yet to understand it is a trivial matter in the conventional language of mechanics. The belt provides a drag torque, which counterbalances the load torque. The power to lift the weight, while stemming from the motor, nevertheless requires the expenditure of frictional energy losses to provide the requisite drag. If we assume, in general, that there exists no a priori relationship between the motor and signal speeds, then the drag principle is the only way in which energy may be transferred between the motor and the load. There must be a loss of information relating to the relative shaft positions of the motor and the load, since in a steady-state relationship there is no way of accumulating the ever-changing

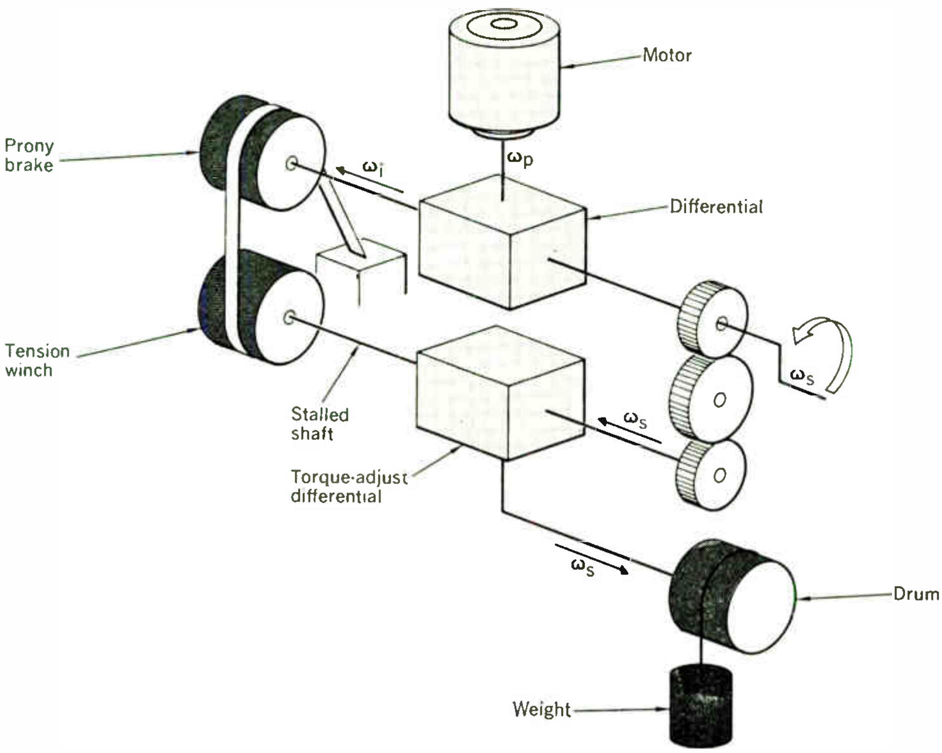


Fig. 11. Modification of three-level amplifier of Fig. 8 to provide for arbitrary load torque.

slack between them. This destruction of phase must be taken up in some disordering process that increases entropy, and hence provides heat. Within this framework the idler loses its mystery, and it becomes evident that idler power has to be expended in order to provide load power under an incommensurable frequency relationship.

If, on the other hand, there existed a dependent frequency relationship between signal and pump, then some geared relationship might have been effected to transfer energy without the need of a drag mechanism. An entirely different mechanical embodiment, bearing no relationship whatever to the drag system, could then have been employed. Nor would this concept have denied that a drag system could still have been employed with dependent frequency relationships. It is quite clear why the mathematical analysis went into gyrations in the dependent frequency case.

Modification for arbitrary load torque. The amplifiers of Figs. 9 and 10 have a failing characteristic of many parametric amplifiers in that they operate only for a specified value of load torque. They fail to adjust the idler torque to equalize to a changing load torque.

To provide this added feature we again make use of the equal torque (equal quantum rate) property of a mechanical differential. If one arm of a differential is stalled, power and frequencies are conserved in the other two shafts but the stall torque of the third shaft is equal to the dynamic torque of the other two. The stall torque may now be applied through possible torque transformations to provide the Prony brake belt tension. If we assume the existence of a constant, dynamic, Prony brake belt coefficient of friction, then the idler torque may vary with, and be made equal to, the load torque. The result is the load-independent amplifier shown in Fig. 11. The amplifier was constructed and operated well according to the principles developed here, with a torque gain of approximately 20. The nature of the device permits simple cascading.

Although this amplifier was developed for expository purposes, it nevertheless possesses certain significant features:

1. It makes no use of electrical amplifiers and has no implicit power limitations characteristically associated with their use.
2. The position error is contained entirely within the straining of the brake material, which may be made negligible.
3. The frequency response is flat over the entire range $|\omega_s| < |\omega_p|$, permitting poor speed regulation of the drive frequency ω_p.

The bandwidth restriction of the third feature is implicit in two different languages. If $|\omega_s|$ exceeds $|\omega_p|$, we no longer have a lower-sideband situation, and regeneration ceases. Alternatively, if the belt slip is reversed, the load torque is not counterbalanced but, on the contrary, is doubled.

It would trespass the purposes of the exposition to go much further into detail in the matter of fleshing out the bare bones of the energy-level description. The major point of the matter is that the energy-level description is not an esoteric evasion of the laws of mechanics but rather is an extremely elegant formulation of mechanics. Nevertheless, we shall dwell a moment longer on the discussion of idler power.

We have shown idler power to be purposely uncorrelated to the signal. It might be argued that although it is indeed uncorrelated to the signal, we might yet recombine it with the existing pump, recorrelate it to the signal, and expend it usefully. Schemes attempting this have not succeeded, probably for fundamental thermodynamic reasons. This expenditure as heat, however, does not necessarily show up as a complete loss of efficiency, since we may employ a Carnot cycle to regenerate some of the pump.

Let us assume that there exists some minimum idler power $P_{i\,\min}$ and some maximum bandwidth $\Delta\omega_{i\,\max}$. If the idler power passed through this filter it would be able to supply power to a load reservoir whose temperature T_0 were such that

$$kT_0 \frac{\Delta\omega_{\max}}{2\pi} < P_{i\,\min} \tag{16}$$

where k is the Boltzmann constant, equal to 1.38×10^{-23} watt-second per °C.

We should then be able to recoup the Carnot efficiency $(1 - T_0/T)$ of P_i, which is, minimally,

$$\eta = 1 - \frac{k\Delta\omega_{\max}T_0}{2\pi P_{i\,\min}} \tag{17}$$

where we have taken $T = 2\pi P_{i\,\min}/k\Delta\omega_{\max}$. With realistic values, Eq. (17) is always negligibly close to unity and much of the idler power may be recovered.

A simple means of feeding back idler power to the pump is provided by means of a universal motor fed from an ac line source. Such a motor possesses a commutator, which destroys any internal conservation of phase. We form the idler load by causing the idler shaft to drive the motor backwards against its natural rotation, feeding power back into the ac bus. The torque on the motor may be varied by any of a number of feedback schemes in which the motor drive voltage is varied by the load torque. The universal motor is chosen over an induction motor, which might equally have been used, because of the far smaller complications in the torque–speed characteristics over large speed ranges.

Some conclusions

We have found that the frequency relations of Manley and Rowe provided an important link between the conventional understanding of mechanics and the secondary quantization methods usually attributable only to quantum mechanical procedures. Nevertheless, as we have attempted to show, these methods are enormously practical to the extent of producing new viewpoints and new devices in what might be considered the most practical and most understood of all technologies: geared mechanics.

We shall not here pursue further the applications of parametric language to simplify, unify, and innovate. It is important to indicate, however, that it seems to have important significance in formulating classical, relativistic, and quantum mechanical particle mechanics, ray optics, and various aspects of diffraction theory. Although it does not necessarily produce startlingly new results, it seems, as stated at the beginning of this article, to provide the format of a single language that can be used to explore many facets of physical and engineering understanding.

General comments

Several years ago W. W. Mumford published an interesting historical survey of the field of activities lately designated as that of parametric technology.[6] While the mathematical singularities of the Hill equation were well known and exploited by mathematical physicists of stature in the 19th century, Mumford traces the engineering beginnings to Alexanderson, Hartley, Peterson, and others, starting with a 1916 publication by Alexanderson. Intensive work continued through the 1920s, much of it stimulated by Hartley. The basic concepts were known and understood, although a total formulation was lacking.

The main thrust of the present activity stems from the more complete formulation produced by the Manley–Rowe paper. This paper is not a divorce from the work of the earlier investigators but is, instead, a bridge to the present through J. M. Manley's long-term interest in magnetic amplifiers under the influence of R. V. L. Hartley.

It is not my intention to enter a controversy here as to the relative merits of the papers to date covering the field of parametrics; the bibliography is incredibly lengthy. The publications that have been most meaningful to me, aside from those already indicated, are those by Weiss[7] and Bennett.[8] Although the essential seeds of these papers are contained in the present article, I have attempted here to make the quantum formulation less an exercise in operationalism than does the Weiss paper and more physically intuitive a concept than is evident in the mathematical detail of the Bennett paper. Further, the operationalism of quantum mechanics is held here to lead to an exact result classically. This approach goes a step beyond the reserved position maintained by the aforementioned authors, whose works fail to assert parametrics to be a rigorous bridge between the two dissident views of mechanics.

Some of the examples chosen are paralleled in a book by Penfield,[9] which I found out about belatedly. Almost without exception, however, the treatments stressing quantum viewpoints in this article differ in major fashion from those of Penfield, whose objective is to show compatibility of various branches of classical mechanics with Manley–Rowe. The different objectives of the two treatments, mine and Penfield's, produce consequential differences in the systematization of end results and thus, ultimately, produce substantially differing insight. I believe this difference to be immensely profitable and highly recommend Penfield's book as an excellent work.

Appendix: Three-shaft differential

We seek to investigate the properties of a three-shaft differential using only its mathematical and symmetry properties as defined in the section on the mechanical differential. Power and frequency conservation demand that

$$P_1 + P_2 + P_3 = 0 \tag{A-1a}$$

$$\omega_1 + \omega_2 + \omega_3 = 0 \tag{A-1b}$$

Under rearrangement we find

$$\frac{\omega_1}{\omega_3} + \frac{\omega_2}{\omega_3} = -1 = \frac{P_1}{P_3} + \frac{P_2}{P_3} \tag{A-2}$$

which, after further manipulation, yields

$$\omega_1\left(1-\frac{\omega_3}{\omega_1}\frac{P_1}{P_3}\right)+\omega_2\left(1-\frac{\omega_3}{\omega_2}\frac{P_2}{P_3}\right)=0 \qquad \text{(A-3)}$$

From (A-3),

$$\frac{\left(1-\dfrac{\omega_3}{\omega_1}\dfrac{P_1}{P_3}\right)}{\left(1-\dfrac{\omega_3}{\omega_2}\dfrac{P_2}{P_3}\right)}=\frac{-\omega_2}{\omega_1} \qquad \text{(A-4)}$$

if the bracketed quantities in (A-3) are nonvanishing. Equation (A-4) is equivalent to

$$1-\frac{\omega_3}{\omega_1}\frac{P_1}{P_3}=\gamma\omega_2 \qquad \text{(A-5a)}$$

$$1-\frac{\omega_3}{\omega_2}\frac{P_2}{P_3}=-\gamma\omega_1 \qquad \text{(A-5b)}$$

where γ is some function of the various variables.

The question arises of what, indeed, the independent variables are. To determine one possible set of independent variables, let us construct a program based on physical insight, in which we consume all the degrees of freedom, though the following steps:

1. Fix ω_1 by driving the differential at that frequency.
2. Load shaft 3 with some constant load torque T_3 by means of a Prony brake.
3. Adjust the loading on shaft 2 so that it rotates at ω_2.

After completion of these steps we find that no adjustments remain and that we have exhausted all the independent variables. We then find that a proper functional representation for γ in Eqs. (A-5a) and (A-5b) is $\gamma = \gamma(\omega_1, \omega_2;\ T_3)$.

The subscripts 1, 2, and 3 are, of course, arbitrary labels and, because of an assumed symmetry between 1 and 2, we should be able to interchange them without consequence. Rewriting (A-5a) in its full functional form,

$$1-\frac{\omega_3}{\omega_1}\frac{P_1}{P_3}=-[\gamma(\omega_1,\omega_2;\ T_3)]\omega_2$$

Under the transformation $1 \rightleftarrows 2$ we obtain

$$1-\frac{\omega_3}{\omega_2}\frac{P_2}{P_3}=-[\gamma(\omega_2,\omega_1;\ T_3)]\omega_1$$

Since, by (A-5b), the right side is also $[\gamma(\omega_1, \omega_2;\ T_3)]\omega_1$, we find

$$\gamma(\omega_1,\omega_2;\mathrm{T}_3)=-\gamma(\omega_2,\omega_1;T_3) \qquad \text{(A-6)}$$

The left-hand side of either (A-5a) or (A-5b) is dimensionless, so that γ must be dimensionally of the form $1/\omega$. Since (A-6) shows antisymmetry in the ω_1 and ω_2 variables, γ must be expressible in the form*

$$\gamma(\omega_1,\omega_2;\ T_3)=\frac{1}{\omega_1-\omega_2}E(\omega_1,\omega_2;\ T_3) \qquad \text{(A-7)}$$

where E is an even function in the ω_1, ω_2 variables. Equation (A-7) shows a singularity for $\omega_1 = \omega_2$ that cannot exist for the quantities involved in the differential from the presumption of analyticity. The dimensionless, even function E must, therefore, supply an appropriate zero to reduce the singularities. Because it is dimensionless, E has the form

$$E=\left(\frac{\omega_1}{\omega_2}-\frac{\omega_2}{\omega_1}\right)^{2n}F(\omega_1,\omega_2;\ T_3)$$

where n is an integer and F is another even dimensionless function of ω_1 and ω_2. We then have

$$\gamma(\omega_1,\omega_2;\ T_3)=\frac{\left(\dfrac{\omega_1}{\omega_2}-\dfrac{\omega_2}{\omega_1}\right)^{2n}}{\omega_1-\omega_2}F(\omega_1,\omega_2;\ T_3) \qquad \text{(A-8)}$$

We next find that the right side goes to infinity for ω_1 or $\omega_2 = 0$. Since P_1/ω_1 is a limiting finite stall torque as $\omega_1 \to 0$, and similarly for ω_2, no singularities may arise at zero frequencies. We have at this point run out of fixes on F and must conclude that $F = 0$, and consequently $\gamma = 0$.* Inspection of (A-5a) and (A-5b) yields

$$\frac{P_1}{\omega_1}=\frac{P_2}{\omega_2}=\frac{P_3}{\omega_3} \qquad \text{(A-9)}$$

Equation (A-9) states the seemingly innocuous result that the torque on each shaft of the differential is a constant. More important, this result was obtained only from the mathematical symmetry of (A-1a) and (A-1b) without substantial reference to the detailed description of the differential. It is, however, important to emphasize the significance of the assumptions made in the derivation of (A-9), and stated in the earlier discussion of the mechanical differential, that (a) ω_1 and ω_2 are independent variables and (b) the quantities P_1 and ω_1, etc., are assumed to correspond to one another such that an unambiguous nonvanishing limit P/ω exists as ω tends toward zero.

* Implicit in (A-7) is the presumption that the differential is dispersionless and, therefore, contains no internal natural frequencies. The only way that γ can have an inverse frequency dimension is within the form shown in (A-7).

* Although this result stems from an assumption that the coefficients in (A-3) do not vanish, the very fact that γ vanishes shows this assumption to be false. We are thus still left with the result for (A-5a) that $\gamma = 0$ is consistent.

One man's hard-fought-for picture of truth is another man's patchwork of triviality. Under these circumstances, I hesitate to identify Dr. E. O. Schulz Du Bois with all of the views in this presentation. Nevertheless, I am particularly indebted to him for his skillful criticisms and his frustrating ability to demolish unrefined argument.

REFERENCES

1. Tien, P. K., and Suhl, H., "A traveling-wave ferromagnetic amplifier," *Proc. IRE*, vol. 46, pp. 700–706, Apr. 1958.

2. Manley, J. M., and Rowe, H. E., "Some general properties of nonlinear elements, I—General energy relations," *Proc. IRE*, vol. 44, pp. 904–913, July 1956.

3. Seidel, H., "Equivalence and uniqueness in parametric systems," submitted for publication, *Proc. IEEE*.

4. Von Neumann, J., "Nonlinear capacitance or inductance switching, amplifying and memory organs," U.S. Patent No. 2 815 488, filed Apr. 28, 1954, issued Dec. 1957.

5. Goto, E., "New parametron circuit element using nonlinear reactance," *Kokusai Den. Den. Kenkyu Shiryo*, Nov. 1954.

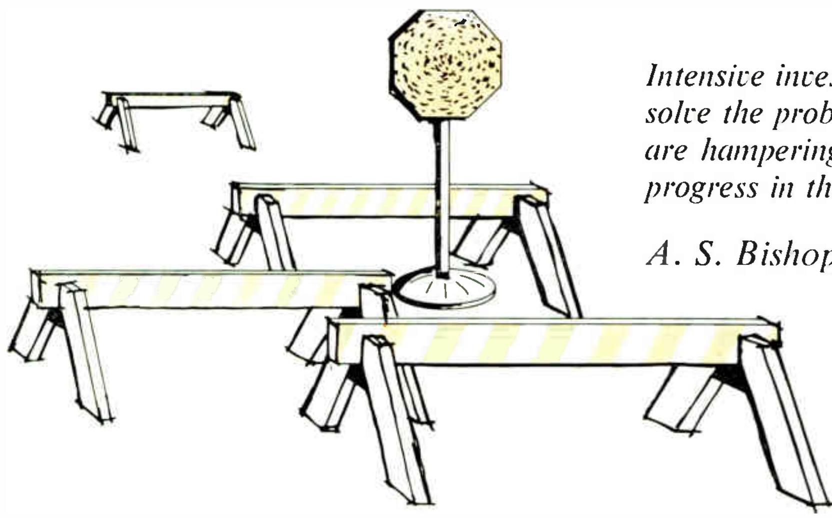
6. Mumford, W. W., "Some notes on the history of parametric transducers," *Proc. IRE*, vol. 48, pp. 848–853, May 1960.

7. Weiss, M. T., "Quantum derivation of energy relations analagous to those for nonlinear reactances," *Proc. IRE*, vol. 45, pp. 1012–1013, July 1957.

8. Bennett, W. R., "Amplification in nonlinear reactive networks," *MRI Symp. Proc. on Active Networks and Feedback Systems*, vol. 10, pp. 645–658, 1960, published by Interscience, New York, N.Y.

9. Penfield, P., Jr., *Frequency Power Formulas*. Cambridge, Mass.: M.I.T. Press, and New York: Wiley, 1960.

Roadblocks in the path of controlled fusion



Intensive investigations are under way to solve the problems of instabilities, which are hampering efforts to achieve substantial progress in the field of controlled fusion

A. S. Bishop *U.S. Atomic Energy Commission**

Progress in the field of controlled fusion is being slowed by inadequate confinement of the hot plasma, due in most cases to instabilities of one type or another. The key to the problem appears to lie in the systematic reduction of the various energy reservoirs available to drive these instabilities.

Although it is assumed that most readers are somewhat familiar with the basic principles of controlled fusion, it may be useful to summarize briefly the goals of the U.S. controlled fusion program,[1] the criteria for reaching them, and the major difficulties that must be overcome.

All of the controlled fusion efforts to be described are based on the following four requirements:

1. The fuel to be used is deuterium, or a mixture of deuterium and tritium.
2. The ions of the plasma must be heated to temperatures of 10^8 to 10^9°K (corresponding to 10^4 to 10^5 eV). Temperatures of this magnitude are required in order that the ions, moving randomly in a confined volume, have sufficient energy to overcome their Coulomb force of repulsion and undergo fusion.
3. Ion densities must be of the order of 10^{15} particles/cm^3 (corresponding to about 1/10 000 of an atmosphere under standard conditions).
4. The corresponding confinement time (τ) must have a value of about a tenth of a second.

The basic difficulties that must be overcome are threefold:

1. Impurities in the plasma must be kept to a rather low value (< one part in 5000, say). The reason is that, in contrast to the case for hydrogen isotopes, it is very difficult to strip all of the electrons from impurity atoms; therefore, they radiate very strongly. The result is an energy sink, which impedes the heating of the plasma.
2. Some appropriate method must be found to heat the plasma to thermonuclear temperatures. Even in the absence of impurities, this task is difficult, requiring much ingenuity and, usually, sizable equipment.
3. By far the greatest problem, however, is that of achieving adequate confinement of the hot plasma within a given region. All approaches under active investigation use magnetic fields (in one configuration or another) for confinement. If the plasma would remain quiescent and behave "classically," the rate of particle loss from such configurations would be very slow and the plasma could readily be confined for periods of thermonuclear interest. As we will see, the problem is that the plasma particles, through the development of instabilities or other cooperative phenomena, are extraordinarily clever in finding ways of escaping rapidly across the magnetic field to the vessel walls.

It is interesting and important to note that these three problems are all interdependent: the impurity level can determine the rate of heating; the method of heating can influence the onset of instabilities; and the presence of instabilities can, in turn, result in the rapid loss of energetic particles to the walls, with a consequent release of impurities which flood back into the discharge. It is fair to say, however, that the key to the entire problem is unquestionably that of adequate plasma confinement, and it is thus to this topic (and the closely related instability question) that I wish to devote most of this discussion.

Experimental attacks on the fusion problem

First, however, let us look quickly at the various methods of attacking the fusion problem. These methods effectively fall into two categories:

Low-energy injection. One basic technique is to start

*At the time this article was written, Dr. Bishop was on the research staff of Princeton Plasma Physics Laboratory.

with a cool or lukewarm plasma of moderate density in a confined region and then heat the plasma by some means toward thermonuclear conditions. These "low-energy injection" experiments are shown schematically in Fig. 1.

The axial pinch effect, historically the first studied, has been extensively investigated in linear and toroidal geometry at the Los Alamos Scientific Laboratory (LASL) in New Mexico and at the Lawrence Radiation Laboratory (LRL) in Livermore, Calif. Its distinguishing feature is that the confining magnetic field is produced by a very high current, which flows within the plasma itself. Heating is achieved by fast magnetic compression. The simple pinch is well known for the violent instability that develops in just a few microseconds and dashes the plasma against the vessel walls. Numerous ingenious attempts have been made to stabilize the pinch. Although none has been very successful, one of the more interesting modifications has been the so-called "hard-core" pinch, involving a solid metal conductor within the tube along its minor axis.

An orthogonal version of the axial pinch is the theta pinch, which is being pursued at the General Electric Company, at Los Alamos, and at the Naval Research Laboratory (NRL). Here, by passing a large current suddenly through a single-turn coil surrounding the plasma, a countercurrent is induced to flow azimuthally in the plasma (i.e., in the theta direction, rather than axially). The axial magnetic field resulting from these currents serves to confine the plasma. Heating is again achieved by fast magnetic compression.

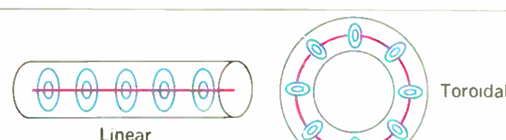
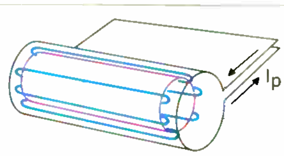
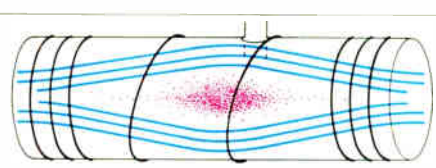
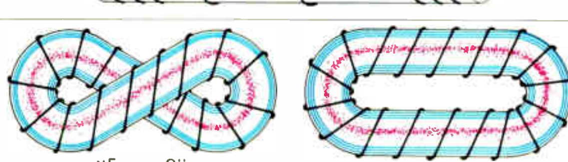
In contrast to the pinch technique, the confining field in the Pyrotron (or magnetic mirror) concept is produced wholly by currents flowing in external windings, rather than within the plasma itself. Loss of plasma out the ends is reduced by means of intensified fields (i.e., "magnetic mirrors") there, which serve to reflect and confine those particles that are moving primarily in the transverse plane, with comparatively small axial velocity. Heating of the plasma is accomplished by adiabatic compression of the magnetic field.

The last in this category is the stellarator concept at Princeton, which again uses externally imposed magnetic fields for confinement but avoids end losses by using a closed configuration (either a figure 8 or a racetrack). An important feature for confinement in the racetrack configuration is the use of helical windings in each U-bend section: in one U-bend, the lines of force are caused to rotate uniformly about the minor axis, thus providing the rotational transform required for plasma equilibrium; in the other U-bend, the result is a similar rotation of the magnetic lines, but with a pitch that increases with radius, thus providing shear fields, which should theoretically tend to stabilize the plasma. Two successive methods of heating are available. In the first method, ohmic heating, a small current (of a few kiloamperes) is induced to flow within the plasma. However, since it is not feasible to obtain ion temperatures above a million degrees by ohmic heating, a powerful technique for ion heating has recently been developed.

A close relative of the stellarator—but not shown separately in Fig. 1—is the Levitron device at Livermore. Interestingly enough, this device started out in life as a hard-core pinch experiment in toroidal geometry. Because of instability considerations to be discussed later, it was converted from a pinch-type to a stellarator-type device, with external windings added to provide confinement. Current passed through the levitated hard core at the center provides rotational transform and shear to the magnetic field configuration. While ohmic heating has been used until recently, modifications have now been made to heat the plasma by microwaves.

High-energy injection. In the second category of experimental attacks, one initially starts with a beam of particles having energies already adequate to permit fusion reactions to occur, and then—by ingeniously trapping them in a suitable magnetic field and converting

Fig. 1. Four techniques for controlled fusion involving low-energy injection, with subsequent heating.

Device	Location	Schematic Configuration	Heating
Axial pinch	LASL, LRL		Fast magnetic compression
Theta pinch	LASL, NRL		Fast magnetic compression
Pyrotron	LRL		Adiabatic compression
Stellarator	Princeton		Ohmic and ion heating

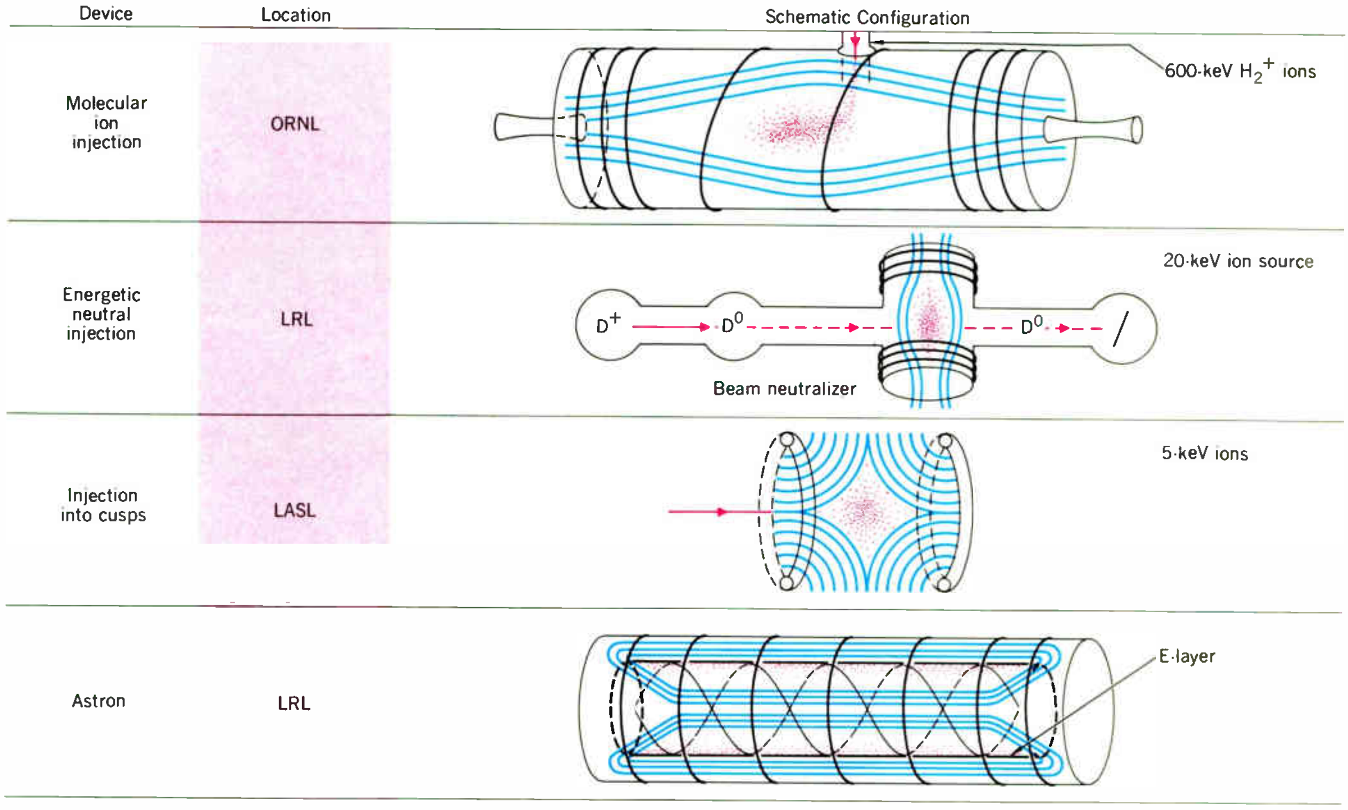


Fig. 2 (above). Four techniques for controlled fusion involving injection and trapping of a beam of high energy, with subsequent conversion of the directed particle velocities into the required random motion of a hot plasma.

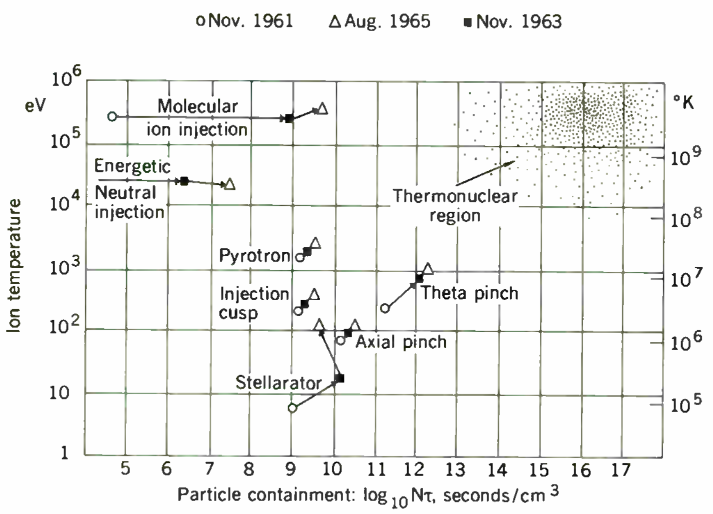
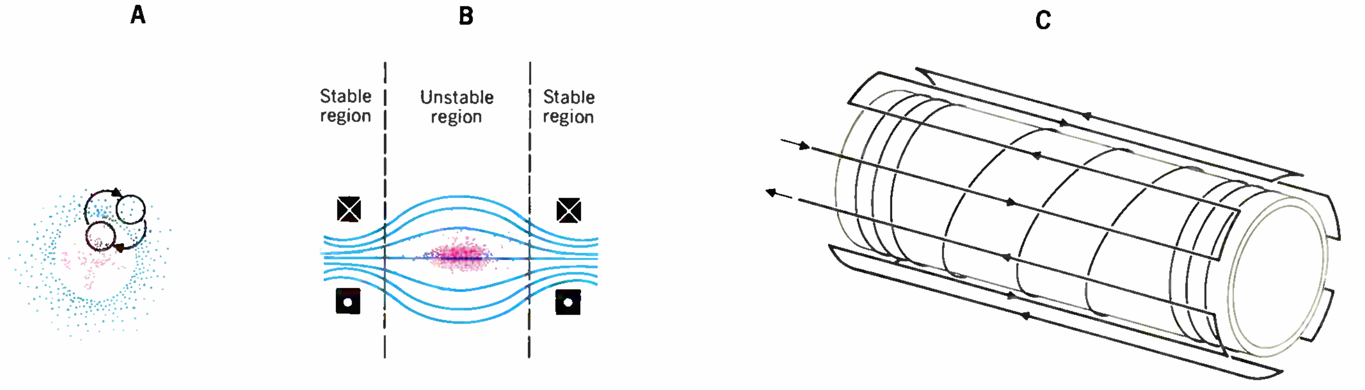


Fig. 3 (left). Chart showing progress made toward the thermonuclear goal over a four-year period. Most of the techniques described in the text are included.

Fig. 4 (below). Magnetohydrodynamic interchange instability for a magnetic-mirror configuration. A—Schematic representation of the instability. B—Simple magnetic-mirror configuration. C—Minimum-B configuration, achieved through the use of Ioffe bars symmetrically spaced around the plasma.



their directed velocities into random motion—builds up the density of this thermalized plasma to values of thermonuclear interest; see Fig. 2.

The first of these involves injecting high-energy molecular hydrogen ions (up to 600 keV) into a mirror geometry and trapping them by any of several dissociation processes. This work is being carried out at the Oak Ridge National Laboratory (ORNL).

An alternative to molecular ion injection is that of injecting high-energy beams of excited *neutral* particles, which then become trapped within the discharge chamber through the process of Lorentz ionization. This work is being pursued at both Livermore and Oak Ridge.

Because of instability difficulties associated with magnetic mirrors, work has also been carried out on the injection of high-energy particles into a magnetic cusp configuration. The initial investigations were made at Los Alamos. Several other groups, however, have investigated various modifications of the cusp technique. One of the more interesting is a toroidal multipole device at General Atomics in La Jolla, Calif.

Last, but by no means least, is the so-called Astron concept of N. Christofilos, being pursued at Livermore. Since this is a comparatively new and relatively unfamiliar technique, I should like to describe it in some detail. In this ingenious and complex device, the confining field is produced by two oppositely directed currents. Currents flowing through an externally wound set of coils produce an axial field within the straight tube, with very slight mirrors at each end. Then a beam of relativistic electrons is injected and trapped in this field to form a concentric layer of electrons (or E-layer, as it is called). The current in this counterstreaming E-layer is built up to a sufficient value that the net magnetic field in the center actually reverses in direction, resulting in a set of closed lines of force surrounding the E-layer. If cold deuterium gas is now admitted to the system, it will (in principle) be ionized and heated by the relativistic E-layer to thermonuclear temperatures, and the resulting plasma confined on the closed lines of magnetic field.

An experimental model of the Astron was recently completed, the main purpose of which is to test the feasibility of injecting and trapping relativistic electrons in this configuration without developing instabilities. The main tank, into which the electrons are injected, is 90 feet long and 2½ feet in diameter. A surrounding solenoid, 4 feet in diameter and fed by some 64 independently energized power supplies, provides the initial axial confining field. Electrons are accelerated to 4 keV, injected, and trapped to form E-layers. Experimental results, while embryonic, are encouraging: to date, no destructive instabilities have been observed.

Status of the various experiments

Where do all of these techniques stand with respect to one another and to the goal, and how rapid has the progress been in that direction? Figure 3 answers both parts of this question. The Astron is not shown, however, since it has only recently come into operation and has not yet developed a plasma of significant density. As seen, probably the most significant progress during the last two-year period has been in the stellarator program, primarily as a result of the recent addition of ion cyclotron resonance heating, which has raised the ion temperature around the machine by a factor of nearly 10.

I. Types of instabilities

Class of Instability	Driving Energy	Configurations Most Susceptible
Universal	Temperature and density gradients in plasma	All
MHD	Lines of force concave toward plasma	Those not possessing minimum B
Velocity-space	Anisotropic pressure due to truncated velocity distribution	Those using magnetic-mirror confinement
Stream	Directed beam of monochromatic particles	Those using high-energy injection
Kink or pinch	Energy associated with imposed currents in the plasma	Those with imposed plasma currents

It is apparent from this diagram, however, that the pace of progress has slowed somewhat during the last period. Indeed, it now seems clear that, except for the newly born Astron, each of the projects (including, as of now, the stellarator program) has come up against a major roadblock that is impeding further progress toward the goal. Although this roadblock manifests itself in somewhat different ways in each of the projects, the basic feature common to all of them is inadequate confinement of the plasma, apparently caused, in most cases, by the presence of instabilities of one type or another.

Classes of instabilities

In view of this situation, I now want to turn the discussion to the specific topic of instabilities. This is a very difficult and complex field (both theoretically and experimentally), but if you will bear with me, I will present some key features of this important problem. One of the main difficulties is that hot plasmas confined by magnetic fields can be subject to a wide variety of instabilities, ranging from the violent kink-type instability found in a linear pinch to the comparatively mild outward flow of particles across magnetic field lines, as found in the stellarator. Furthermore, the plasmas created in the laboratory tend, in general, to have a *mixture* of these instabilities, so that (except in rare instances) it is difficult to isolate a single type of instability in a plasma to study its properties experimentally and compare them with theoretical predictions.

In order to put the instability problem in proper perspective, it will be useful to divide instabilities into five distinct classes (shown in Table I), based upon the origin of the energy reservoir available to drive them[2]:

Universal. The "universal" class of instability is of importance primarily because, as its name implies, all practical fusion plasmas are subject to it. It arises from the fact that the confinement of a hot plasma in a localized region will necessarily result in density gradients within the plasma. Such a departure from true thermodynamic equilibrium provides a source of free energy, which can drive this class of wave-type instabilities.

Magnetohydrodynamic. The MHD class of instabilities, which is shown schematically in Fig. 4(A), involves a gross interchange in position between plasma and confining field. It may be expected to occur any time that such an interchange would result in a net release of energy. As an example, let us consider the case of a plasma in a simple magnetic mirror configuration, shown in Fig. 4(B). The plasma always acts (in whole or in part) like a diamagnetic body and hence can lower its potential energy by moving into a region of decreasing magnetic field. Note that in this simple mirror configuration, there are two distinct regions. Near the coils the curvature of the lines of force are convex to the plasma, and the field strength increases with distance from the axis; in the central region, however, the curvature is concave toward the plasma, and the field decreases with distance from the axis. The former is a region of "good" curvature, in which the plasma is stable; the latter is a region of "bad" curvature, for which a gross interchange between portions of the plasma and the magnetic field may take place with an associated release of energy which serves to drive the instability.

It was the Soviet physicist Ioffe who stressed the importance of inhibiting the development of this instability. Basically, his method was to modify the magnetic field configuration so that the plasma finds itself in a true potential well, with the field strength increasing everywhere outward with radius from a nonzero value at the center. In practice, such a "minimum-B configuration," as it is called, can be achieved in a straight system by the addition of a set of longitudinal bars symmetrically spaced around the plasma, with currents flowing in alternate directions in adjacent bars; see Fig. 4(C). In this way, the region of bad curvature can be eliminated, and the system should be inherently stable against the development of MHD-type instabilities.

Experimentally, the improvement in confinement that resulted from the application of these Ioffe bars to straight systems was rather dramatic. Indeed, much of the progress that occurred between 1961 and 1963 can be attributed specifically to this technique.

It is important to note that although a true "minimum-B" configuration can be created relatively easily for a straight system, it turns out to be topologically impossible to achieve for a system that closes on itself. This point will be discussed again later.

Velocity–space. The next class, velocity–space instabilities, can arise any time that the velocity distribution of the plasma particles is non-Maxwellian. Specifically, the driving force is an anisotropic pressure due to the truncated velocity distribution. Since, as we have seen, magnetic mirror systems inherently have a loss cone in velocity space, all devices employing magnetic confinement are subject to this type of instability.

Stream. A fourth class (related to the third) consists of the so-called stream instabilities, which arise from the presence of a directed beam of energetic particles passing through a plasma. The kinetic energy of the particles serves as the energy reservoir for driving the instability, and all systems using high-energy injection are, in principle, susceptible to it.

Kink or pinch. Finally, we have instabilities that result from the presence of imposed currents within the plasma itself. If the currents are large, as in the case of certain pinch devices, the magnetic fields associated with these currents constitute a huge reservoir of free energy, which is available to drive the instability.

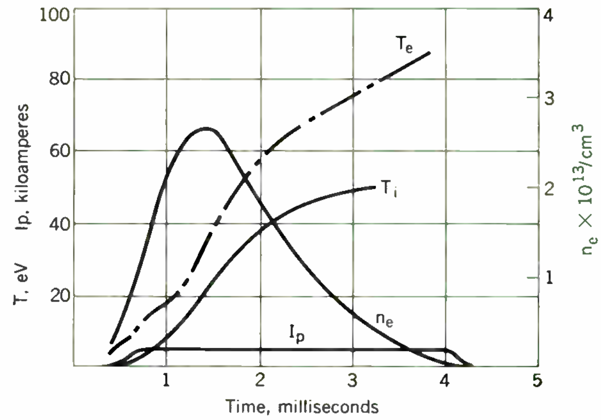


Fig. 5. Time behavior of plasma current I_p, electron density n_e, and electron and ion temperature T_e and T_i for a typical helium discharge in the C stellarator.

Systems with few reservoirs

Which are the systems possessing the fewest number of reservoirs for driving instabilities? Most of the experiments discussed possess three or more such reservoirs. However, several of them—such as the Pyrotron, stellarator, and Levitron—can be reduced to only two. Specifically, the Pyrotron approach, which incorporates a minimum-B field, is subject in principle only to universal and velocity–space instabilities. And experimentally, the evidence indeed seems to support the theory: With MHD instabilities eliminated by use of Ioffe bars, the roadblock that is presently impeding further progress in this approach has all the earmarks of a universal class of instability. Needless to say, an intense theoretical effort is currently under way to determine what modifications are required to suppress this type of instability.

We turn, therefore, to the stellarator concept, which we shall discuss somewhat more fully. To give better insight into plasma behavior during a typical ohmically heated discharge pulse, Fig. 5 shows the time variation of the imposed plasma current I_p, the electron temperature T_e, the ion temperature T_i, and the electron density n_e for a 4-kA discharge in helium, with a confining field of 35 kG. Under these typical conditions, the plasma is confined for only a few milliseconds at most. Careful measurements of the true plasma loss rate have been carried out for the entire discharge pulse, taking into account the spectroscopically determined increase of electron density due to ionization processes. These measurements show that, over a wide range of operating conditions, loss rate can be represented quantitatively as

$$\left(\frac{dn}{dt}\right)_{\text{true}} = \frac{10^8 n_e}{\pi r_0^2}\left(\frac{T_e}{B}\right)$$

where r_0 is the plasma radius. As seen, this loss rate varies directly with the electron temperature T_e and inversely with the confining field B. In contrast, the rate of plasma loss due to classical diffusion processes should vary as $1/B^2\sqrt{T}$; indeed, under the conditions shown in Fig. 5, a classically behaving plasma would be well confined for many hundred milliseconds.

What, then, is the cause of this anomalously rapid loss of plasma across the magnetic field? The most widely (but by no means unanimously) held view on this matter is that instabilities are responsible. According to theory, an ohmically heated plasma in a stellarator configuration should be subject to three classes of instability: universal, MHD, and kink (or current-driven). Of these, the last has long been the most suspected candidate. Over the past several years, however, an important series of experiments has been carried out on the stellarator, in which low-density low-temperature plasma was produced by microwave heating, without the presence of any plasma current whatsoever. Under these conditions, the current-driven class of instabilities would necessarily be eliminated. The measurements show, however, that the rate of plasma loss is independent of the method of heating and continues to be described by the same T/B relationship.

Thus the number of known instability candidates has now been narrowed to only two classes, which we now examine in somewhat greater detail. But when we turn the spotlight on these universal and MHD instabilities we find that the mode of instability that can develop will depend rather strongly on whether the plasma conductivity is considered to be infinite or finite; see Table II. Consider first the MHD case of instability for a toroidal configuration. Early theoretical work was carried out for the comparatively simple case of an ideal high-temperature plasma with infinite conductivity. For this case, the theoretical conditions for stability are well known. Indeed, an infinitely conducting *diffuse* plasma in a toroidal configuration can be stabilized against the MHD instability by shear fields alone.

In the past several years, however, theoretical treatments of instabilities have become sufficiently sophisticated that now one can confidently treat plasmas having *finite* conductivity—that is, plasmas that correspond more closely to those in stellarator experiments. Here, conditions for stability become much more limited, and shear fields turn out to be inadequate for this purpose.

We saw earlier that for a straight mirror machine the MHD instability could be suppressed by converting it to a minimum-B configuration. In a toroidal geometry, it turns out to be topologically impossible to achieve a true minimum-B field. The best that one can do is to achieve a minimum *average B* (or $\bar{B}$) field, in which there are regions of favorable curvature followed by regions of unfavorable curvature, the relative importance of the favorable sections increasing with radius. Although they cannot completely stabilize the MHD mode in a finite-conductivity toroidal plasma, they should slow its growth rate enough that the instability will no longer be serious.

In the case of universal instability, the situation is also reasonably encouraging. Recent theoretical studies show that, for a diffuse infinitely conducting plasma, universal instabilities will probably be suppressed by a modest amount of shear fields. For a finite-conductivity plasma, however, the criteria for stabilization are again more severe. Yet even for this case it appears that if sufficient shear is incorporated into the field configuration, the instability should be completely suppressible.

Thus, despite the additional complication that results from finite conductivity, the outlook for stabilization is encouraging. Moreover, any increase in the plasma temperature results in an increase in conductivity and, hence, should help to relax the criteria for stabilization.

Although I have spoken at some length about the stellarator program, I want to point out that a powerful tool for studying the universal and MHD instabilities in toroidal geometry is the Levitron. An important advantage of this device is that the shear fields that can be achieved via its hard core are much stronger than those produced in the stellarator. Composite results from the two devices should soon give an answer to the critical question involving the feasibility of obtaining adequate confinement of the plasma in toroidal configurations.

II. Instability suppression in the stellarator

Class	Conductivity	Name of Mode	Stabilized by High Shear and/or Minimum $\bar{B}$?
Universal	Infinite	Resonant drift	Yes
	Finite	Resistive drift	Yes
MHD	Infinite	MHD interchange	Yes
	Finite	Resistive interchange	No (but growth rate is slow)

Conclusions

1. With the exception of the Astron program, which has only recently come into operation, it appears that most (if not all) of the experimental efforts in controlled fusion are now being limited in their progress by inadequate confinement of the plasma, due in most cases to instabilities of one type or another.

2. These instabilities can be categorized according to the source of the energy reservoirs available to drive them. In our attempt to understand and overcome our difficulties, it would seem desirable to reduce the number and importance of reservoirs as much as possible.

3. In the Pyrotron program, the stabilization of MHD instabilities by Ioffe bars has reduced the number of such reservoirs to only two. At present, universal-type instabilities are clearly the next hurdle to be overcome.

4. In the stellarator program, the presence of plasma currents has been ruled out as a cause of the anomalously fast loss of plasma there. At present, the most likely of the known instability candidates are the resistive universal or the resistive MHD modes.

5. In an effort to eliminate (or at least to reduce the importance of) these remaining modes, intensive studies are under way toward the development of stellarator and Levitron systems having both a minimum-$\bar{B}$ configuration and also adequate magnetic shear. Simultaneously, through the recent addition of ion heating to the stellarator, attempts are being made to increase the plasma conductivity and thereby ease the requirements for achieving complete stabilization.

Essentially full text of a talk presented at WESCON, San Francisco, Calif., Aug. 24, 1965.

The work on this survey was performed under the auspices of the U.S. Atomic Energy Commission, Contract AT(30-1)-1238. The author is indebted to Drs. E. A. Frieman, M. B. Gottlieb, D. J. Grove, J. L. Johnson, R. F. Post, and L. Spitzer, Jr., for their helpful suggestions.

REFERENCES

1. Bishop, A. S., "Recent developments in controlled fusion," *Phys. Today*, vol. 17, pp. 19–26, 1965.
2. Furth, H. P., and Post, R. F., Rept. UCRL-12234, Univ. of California Radiation Lab., Berkeley.

Future role of satellites in air traffic control

Raymond E. Spence *Federal Aviation Agency*

For the past several years the United States Federal Aviation Agency has recognized the potential of artificial earth satellites for improved air traffic control. This potential lies primarily in the areas of communications, navigation, and data acquisition, with communications being the area closest to realization. Projects completed by the Agency to date indicate that a communications satellite in synchronous orbit with the capability of several VHF channels employing voice modulation is feasible for implementation by 1970 or earlier. These projects also indicate that there is less certainty about realizing the navigation and data acquisition potential. Efforts are currently under way to define further the air traffic control requirements for satellite systems and to synthesize a communications satellite system. These efforts are designed to consider such major factors as system implementation timing, cost versus benefit, circuit discipline requirements, and system technical parameters. Experimental use will be made of the National Aeronautics and Space Administration ATS-B satellite in late 1966 to support this effort and some nonsatellite hardware development will also be undertaken.

As early as 1960 the concepts of satellite-supported subsystems in the future air traffic control (ATC) system were recognized and voiced by the Federal Aviation Agency. A decision was made to initiate research studies dealing with the capabilities of satellites to perform in communications, data acquisition, and navigation environments. Since that time, both in-house and contractor-supported studies have been completed in these areas of satellite application. By 1963, these studies had indicated that satellites were indeed capable of making significant contributions to future ATC systems, though with varying degrees of urgency of application. It was thought at that time, and is still thought today, that the most important and necessary area of contribution was communication over areas not suited to normal line-of-sight VHF or UHF coverage, followed in order of requirement and application by position determination and navigation.

In November 1963, R. E. Kester of the Systems Research and Development Service (SRDS) conceived and proposed a method of using the Syncom II satellite telemetry and command link in which the "down-link" on approximately 136 Mc/s and the "up-link" on approximately 148 Mc/s might be used for relaying digitally coded messages between an aircraft and a ground station. In January 1964 this concept was discussed with Dr. S. G. Lutz of the Hughes Research Laboratories and with the National Aeronautics and Space Administration (NASA), and subsequent experiments that were performed, under Hughes' sponsorship, established its feasibility. In April 1964 this concept, referred to as TC/RTTY, and tentative plans by the SRDS to implement an experiment using the TC/RTTY were discussed with Frank White of the Air Transport Association. The air transport group organized and very successfully executed a test program with the help of a number of organizations, including Pan American Airways, Bendix Radio Company, Boeing Company, Hughes Aircraft Company, Communications Satellite Corporation (Comsat), FAA, NASA, and Aeronautical Radio Incorporated. The results of this experimentation, as well as other work by Comsat, the Defense Communications Agency, and other government and industrial organizations, have been well publicized and have led to the current intense interest on the part of users and the FAA to implement a communications satellite to support North Atlantic air traffic control during 1967.

Use of sea stations

It would be incomplete not to point out that satellites have some degree of competition as a means of achieving improved ATC communications in the North Atlantic area. Two private firms have proposed the use of sea stations as a method of providing ATC as well as other services in the North Atlantic. The primary advantage of sea stations would be their capability of extending existing ground/air/ground communications across the Atlantic as well as providing secondary surveillance radar, VORTAC distance-measuring equipment, and meteorological radar. Extensions of these services by sea station would require no changes in airborne hardware and would in effect allow aircraft not now electronically equipped for ocean crossings to make the flight, assuming they are already equipped for instrument flight rule conditions and have the required flight characteristics. The

Studies have shown that satellites are capable of making significant contributions to future air traffic control systems, particularly in communication over areas not suited to normal line-of-sight VHF or UHF coverage

major disadvantage of the sea station approach seems to be the very high capital costs and recurring costs of covering large ocean areas. Because of the large areas requiring coverage in a worldwide ATC system and the very high costs associated with sea stations, the FAA is rather firmly convinced that the satellite approach is considerably more economical and achievable in the same time period, if not earlier, than the sea station approach.

At the present time the FAA is proceeding with plans to make satellite communications available to aircraft transiting the North Atlantic principal control area as soon as possible. We are currently engaging in preliminary discussions with Comsat regarding the system definition and associated costs of providing a communications satellite for this purpose in late 1967. We are also making plans to confirm the satellite characteristics through experimentation with the ATS-B satellite in early 1967. To achieve this goal, a concerted effort will have to be put forth by both the FAA and Comsat. The application of satellite communications to air traffic control is not synonymous with its application to point-to-point communications circuits. There are two major differences to be considered. First, the number of terminal stations (airborne) is not only large and subject to wide variation in number, but is somewhat randomly distributed over a wide geographical area. Second, the high powers and large high-gain antennas available for use by fixed ground terminals are not applicable to present and forecast airframes. With these two differences in mind, it seems obvious that considerable attention must be given to the system definition of a satellite-supported ATC system. The possibilities of later adding the functions of position determination and navigation must also be examined, so that undue restraints are not imposed by the initial system design.

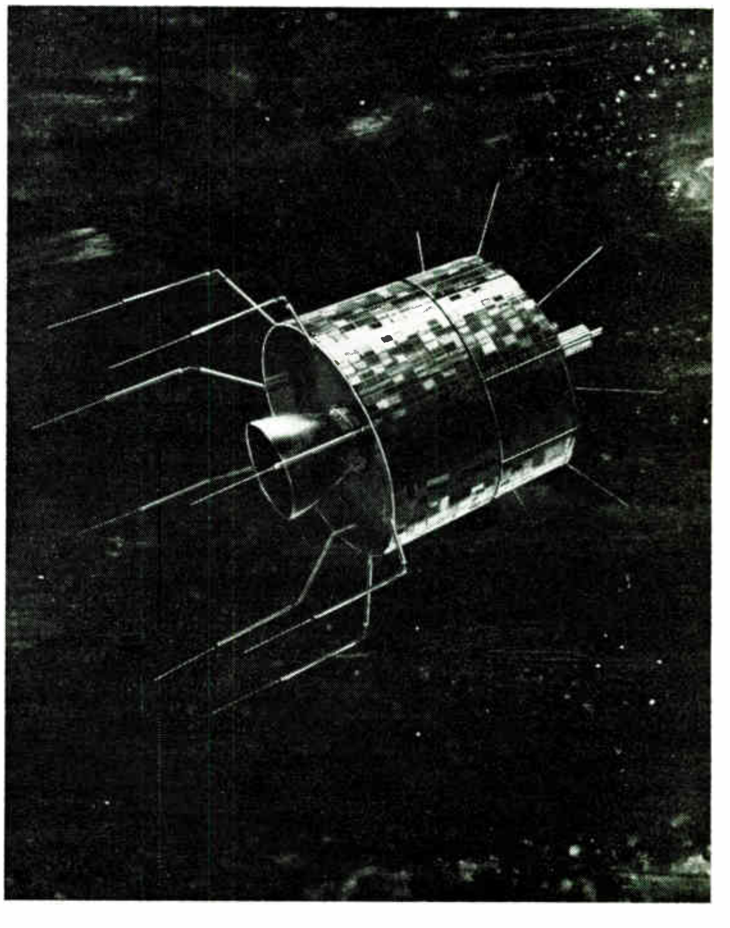
Major parameters

With the foregoing in mind, let us discuss briefly some of the important parameters to be considered in the initial system description. Nine major parameters are identified and will be discussed in order:

1. Aircraft antenna and electronic equipment
2. Satellite antenna and electronic equipment
3. Ground station antenna and electronic equipment
4. System modulation scheme
5. System frequency plan
6. System power budget
7. System organization with respect to circuit control and access
8. System interfaces with other long-range communications systems
9. Analysis of the effects the system may have on present and proposed operations and procedures

As with any major and complex system, the parameters

ATS-B satellite.



are highly interlaced and inseparable. Changing any one of them has some effect on all the others. Let us discuss them one by one with this in mind.

Antenna and electronic equipment

The aircraft antenna and electronic equipment is of major concern to all potential users of the system. Ideally, the system should be designed to accept the hardware now being used by over-ocean aircraft. Unfortunately, this practice, though technically possible, does not appear practicable at this time. Current aircraft antennas have both relatively poor gain and pattern characteristics for use in communicating with a satellite. The best average gain figure is estimated to be 0 dB over an isotropic radiator, and this is as likely to occur looking at or below the horizon as it is to occur 10 degrees or more above the horizon. The FAA is currently conducting a development effort to demonstrate the practicable feasibility of an airborne antenna with 3 dB or more of gain above $+10°$ from the horizon with several dB relative attenuation at or below $+10°$. The success of this effort is considered important in terms of adding 3 dB to the system power margin and easing of the multipath signal propagation problem. Improvements also will be required in the aircraft transceivers in terms of receiver noise figures and transmitter output power. Enough work has already been done by industry to demonstrate that 3-dB receivers and 500-watt transmitters are readily obtainable for either amplitude or phase modulation.

Considerations of the satellite antenna and electronic equipment include the antenna gain and mechanical design and the functions assigned to the electronic equipment. The ATS-B satellite, being a spin stabilized type, will use an electronically despun antenna with a probable gain of about 10 dB. Mechanical despinning also appears feasible. If a gravity gradient-stabilized satellite were used, a simple helical antenna with 14 dB to 17 dB gain seems practicable and has been proposed. Neither of these antennas has been proved in flight at this time. The functions assigned to the electronics package depend somewhat upon the system frequency plan. If the ground–satellite and aircraft–satellite frequencies are the same, a simple linear transponder may suffice on the satellite. On the other hand, if the ground–satellite link is in the microwave spectrum and the aircraft–satellite link is in the VHF spectrum the additional function of frequency translation will be required.

No significant problems are anticipated with respect to ground stations and antennas. Enough knowledge and equipment seem to be available to reduce this parameter to a matter of specifying the required terminals.

System modulation and frequency schemes

The matter of the system modulation scheme is of prime importance in achieving a workable system. Various types of amplitude, frequency, and phase modulation are technically possible and each has certain advantages and disadvantages. High-level amplitude modulation would be compatible with existing airborne equipment, but is wasteful of system power and spectrum. Single-sideband AM would be advantageous from a power and spectrum standpoint, but it requires complex frequency tracking to compensate for Doppler frequency shifts and is not compatible with present aircraft equipment. Although a proposed low-energy AM pulse position system is mindful of power and, to some extent, spectrum, it requires improvement in fidelity and also is not compatible with present aircraft equipment. Frequency- and phase-modulation systems are favorable from a power and fidelity standpoint, and are considered to be the most desirable systems at this time for satellite ATC communications. We regard the modulation system along with the frequency plan as two of the most critical parameters. In all probability, the first satellite system to provide communications for ATC will in fact determine the type of modulation and frequency of operation of aircraft-to-satellite links for some time to come. To suppose otherwise would mean either that very few users would want to equip themselves to use the first system or that extensive equipment changes would have to be made several years later; neither situation would be really tolerable.

As mentioned earlier, the system frequency plan is regarded as one of the two most critical parameters. The current consensus of thinking favors the use of microwave frequencies for the ground–satellite links and the present 118–136-Mc/s aeronautical VHF band for the aircraft–satellite link. The degree of interference between the present short-range air/ground system and the variously proposed satellite modulation schemes and power budgets is largely a matter of conjecture at this writing, and is a matter undergoing intense study on the part of FAA frequency management engineers as well as others, both domestic and foreign, including the Central Radio Propagation Laboratory of the National Bureau of Standards. Alternative frequency plans include an all-VHF system and an all-microwave system. The all-VHF system would eliminate frequency translation in the satellite and offer a "party line" to all participating aircraft. It would also compound the frequency allocation problem. At the other extreme, the all-microwave system would enjoy the same aforementioned advantages of the all-VHF system plus the opportunity to use the currently virtually unused 1540–1660-Mc/s frequency band. However, some disadvantages also appear, and these are amplified by the desire to proceed as soon as possible. All new aircraft electronic equipment and antennas would be necessary. Even though the cost of the electronic equipment might be favorably comparable to the modified and augmented VHF equipment required by a VHF satellite–aircraft link, the aircraft antenna poses a different problem. Within a given power budget, a much higher antenna gain would be called for to overcome transmission path losses at the higher frequencies. This would require a high-gain steerable array on either or both the satellite and aircraft, neither of which seems practicable in the next several years.

System power budget

Starting with a given amount of prime power available on the satellite, the determination of the system power budget is largely a matter of evaluating the various trade-offs between transmitter powers, antenna gains, modulation schemes, channelization schemes, and receiver noise figures. The effects of effective radiated powers on the illuminated earth's surface must be examined in terms of interference to any shared users of the chosen frequency plan. Some maximum effective radiated power probably needs to be specified for future system designs. Extensive studies are being conducted by and for the

FAA to determine the consequences of these various trade-offs. One of the more critical points in the power budget is the determination of the required system power margin. The definition of this margin is complex and involves defining a grade or quality of circuit to be provided for some percentage of the time. It then requires the best estimate of such transmission variables as Faraday rotation, multipath, amplitude scintillation, and ionospheric attenuation. Estimates of the margin required have ranged from 7 to 11 dB. One of the primary purposes of the ATS-B experiment is to measure these variables.

External system interfaces

The preceding six parameters might be generally classed as internal system interfaces. The remaining three could then be classed as external system interfaces. Every system must have some organization with respect to its control and accessibility. As the internal system parameters become better defined, and before they become too set in definition, one must begin to formulate answers to such questions as: How many ground stations have access to the satellites? Who has access to the ground stations? How is this access coordinated? How is airborne terminal access coordinated? Is a "party line" concept desirable? Is a "roll call" technique desirable? How are channels assigned? Many other such questions will emerge, some of which may not be apparent until the internal parameters are clearly defined.

Possibly the most important external interface consideration is the relative role that satellite communications may play in the total long-range ATC communications system. The present long-range system is primarily in the high-frequency range. It is technically possible and practicably conceivable that satellite communications could replace the HF systems and become a total long-range system in itself. However, it seems more probable at this time that both techniques will survive and together form a total system. This rationale may be developed as follows: It appears that the cost and complexity of providing polar coverage via communications satellite may be on the wrong side of the cost/benefit curve. Specifically, the cost of providing the polar coverage will probably equal or exceed the cost of providing coverage to the rest of the world. Considering current and forecast traffic routes and densities, less than ten percent of the world's international air traffic transits the polar areas. Thus, polar coverage via satellite does not appear very favorable on an economic basis. For this reason, it may be expected that HF radio will continue to be used in these polar areas for some time to come. High-frequency radio will also be used as backup communications in prime satellite coverage areas until redundant satellite coverage can be provided. Additionally, it may be a continuing requirement in those countries that use HF radio as part of their domestic air-ground communications system.

The remaining external system consideration is one of the effects a new long-range communications system may have on present and proposed operations and procedures. It should not be assumed that the organization of international ATC that is quite satisfactory today would be equally satisfactory with a new communications system. Constraints imposed by a new system would be totally different than those imposed by the present HF system. A detailed analysis will be undertaken as soon as the new system description has started to assume a more definite shape.

International considerations

Other than the formulation of a sound system description as just discussed, there are equally important steps to be taken before we see the beneficial use of communication satellites by the worldwide air traffic control systems. Because of the worldwide application envisioned for the satellite communications system, many sovereign nations are necessarily concerned with all aspects of the system. Perhaps one of the more difficult considerations from an international coordination and agreement viewpoint is the system frequency plan. International agreement must be obtained on the system description so that all ground stations and aircraft terminals are compatible. Unless some one nation is willing to assume the burden of the entire system cost, the matter of sharing system costs must be resolved. This consideration is in itself extremely complex in view of differing national policies and will not be discussed here. Also to be resolved on an international basis is the matter of system organization with respect to circuit control and access, and analysis of the system's effects on operations and procedures.

Several organizational elements of the FAA are already at work to initiate and resolve these international considerations through established channels, such as the International Civil Aviation Organization and the International Telecommunications Union. As our proposed system begins to assume a more definitive shape, international coordination efforts will be intensified.

Position determination and navigation

Although most of this discussion has been directed at the communications aspects and potential of satellites for the air traffic control system, the FAA is not neglecting the potential of satellites in the areas of position determination and navigation. There is a need today for a position determination system over the North Atlantic and some other ocean areas that is independent of the aircraft navigation system. Such a system would extend the overland control philosophy of ground-derived position information to the ocean areas and could be a significant step in the reduction of separation standards in these areas. At this time, the system parameters for such a position determination system are not as clearly defined as is the case with the communications system. Extensive work must be done before any definitive system recommendations are made, and thus it is unlikely that position determination will be a function of the first satellites to support air traffic control.

It is conceivable and technically possible that future satellites to support the ATC system may have a navigation function. Here also, a great deal of study and experimentation needs to be accomplished. Additionally, presently conceived land-based or aircraft self-contained navigation systems are giving satellite-supported systems severe competition in producing the required accuracies and cost/benefit ratio. For these reasons, the role of satellites in an aeronautical navigation system will be somewhat "iffy" for some time to come.

Full text of a paper presented at the AIAA Communication Satellite Systems Conference, Washington, D.C., May 2–4, 1966. Published as AIAA Paper No. 66-259.

Special Conference Report

National Transportation Symposium 9th Joint Railroad Conference

Gordon D. Friedlander *Staff Writer*

In what might be billed as a "doubleheader," the first National Transportation Symposium (May 2–6) and the 9th Joint Railroad Conference (May 4–6)—the former under the joint sponsorship of the American Society of Mechanical Engineers (ASME), IEEE, and the American Society of Civil Engineers (ASCE), and the latter under the aegis of ASME and IEEE—were held at the Jack Tar Hotel in San Francisco. About 700 registrants attended the unusual and significant sessions, which, for the first time, brought together people concerned with every aspect of the subject of transportation on sea, on land, and in the air. The possibility of space transportation of cargoes was touched upon in serious discussion.

The keynote

Henry N. Muller, president of ASME, sounded the keynote of the symposium by his statement that "beside the possibility of war or political considerations, the three major problems of the contemporary world are: air pollution, water pollution, and *transportation*."

Winfield H. Arata, Jr., chairman of the ASME Aviation and Space Division, gave a comprehensive presentation, with visual aids, of the tremendous growth of the transportation industry since World War II, and the new modes of packaging and "containerization" of cargoes in standard-sized modules for transhipment by rail, truck, cargo plane, and cargo vessel. Arata revealed the startling statistic that there are 90 million privately owned cars on the American highways today—one for every 2.2 people in the United States—and that, by 1975, this number would increase to about 120 million.

The special guest speaker at the opening session on May 2 was the Hon. Alan S. Boyd, Undersecretary of Commerce for Transportation. Boyd alleged that, despite the tremendous need for more and more rapid transit facilities, there is presently no existing overall or coordinated plan for transportation in the United States. He strongly supported President Johnson's recommendation to Congress that a Department of Transportation be established, to be headed by a secretary of cabinet rank.

Boyd touched upon some of the investigative efforts and studies presently under way for the development of fuel-cell-powered motor vehicles to reduce the air pollution caused by the conventional internal combustion engines, and vertical takeoff and landing (V/TOL) aircraft for use either in densely populated areas or at airports at which runway extensions for jet plane accommodation cannot be easily made. The Undersecretary also mentioned the studies under consideration for "moving sidewalks" to speed up pedestrian traffic, automated roadways for high-speed long-distance thruway travel, etc. He emphasized the necessity of "salvaging" existing transportation facilities and systems, such as railroads, to "build up upon our prior technology."

The introductory session

The opening session on May 2 (Transportation II) featured speakers who presented their views on the economic, planning and design, and engineering philosophies for meeting future transportation requirements.

Wilfred Owen, senior staff member, Brookings Institute, stressed the need for concurrent improvements in transportation facilities and methods.

Martin Wohl, National Science Foundation Faculty Fellow at the University of California, indicated that transportation system planning must be a coordinated effort by analysts, mathematicians, designers, economists, and engineers. He stated that "sometimes the best decision—for the time being—is to do nothing; sometimes society is unwilling to underwrite new programs."

John Jones, assistant to the president of North American Aviation, Inc., envisioned future hydrofoil, gas-turbine-propelled merchant ships, capable of speeds in the range of 150–300 mi/h. He also emphasized the need for coordinated advance planning techniques, coupled with computer studies and mathematical models for the extrapolation of requiremental data for all areas of future transportation vehicle and system design.

Transportation and railroad sessions

From Tuesday, May 3, through Thursday, May 5, 13 transportation sessions and four railroad sessions were devoted to the presentation of some 65 papers dealing with the planning, design, research, locomotive systems, automation, traffic control, and engineering criteria for land, sea, and airborne vehicles of numerous types and purposes. Several papers were presented by American and foreign speakers on the economics and political policies governing the construction of projected rapid transit systems, supersonic commercial aircraft, etc., both in the United States and abroad.

The BART system

One of the most interesting—and tangible—efforts toward the goal of interurban mass rapid transit, the Bay Area Rapid Transit (BART) system, is presently under construction right in the host city of the Transportation Symposium, San Francisco. Here, a $1 billion, 75-mile-long transit complex (see Fig. 1 map), scheduled for final completion in 1971, will ultimately intertie Alameda and Contra Costa counties in the East Bay metropolitan area with the city of San Francisco by means of 37 elevated, surface, and subway stations spaced an average of two miles apart.

In a paper entitled "Automatic Train Control for BART" (IEEE Paper No. 34 CP 66-206), W. P. Quintin, Jr., an engineer for Parsons Brinckerhoff–Tudor–Bechtel (PBTB), the joint-venture general consulting engineers for the project, explained the feasibility of an automatic train control system capable of controlling two- to ten-car, lightweight multiple units, operating at speeds up to 80 mi/h.

Design schedules call for train headways as close as 90 seconds during peak hours, every 3 to 4 minutes at other times during the day, and from 5 to 15 minutes late at night and over weekends. During rush hours, it is expected that 30 000 passengers per hour will be carried on a single line.

Quintin stated that "four different concepts of automatic train control for the BART system have been tested during the past ten months on a test track schedule that has run 16 hours a day, six days a week."

Automatic train control. Each of the automatic train operation (ATO) concepts, engineered and designed by four manufacturers, was installed at the 4.57-mile-long test track facility, situated in Contra Costa county (see Fig. 2), in three laboratory cars. Each manufacturer was given the opportunity of demonstrating the capability of his particular ATO system, while both the manufacturer's representatives and PBTB engineers recorded data that could be used to judge the performance efficiencies. The equipment of each manufacturer was thoroughly reviewed in the areas of train protection, line supervision, and train operation.

A generalized ATO control loop is shown in the Fig. 3 block diagram, in which the reference (V_r) and state (V_s) speeds are summarized to produce a difference, or error, signal (e). The direction of error will indicate the direction in which the system must react in order to "close" or cancel the error. The magnitude is proportional to the time rate of velocity difference $x(t)$ and bears a direct relationship to the correcting torque. The modulator in Fig. 3 is schematically a four-terminal network that translates the error signal into a signal that is suitable to

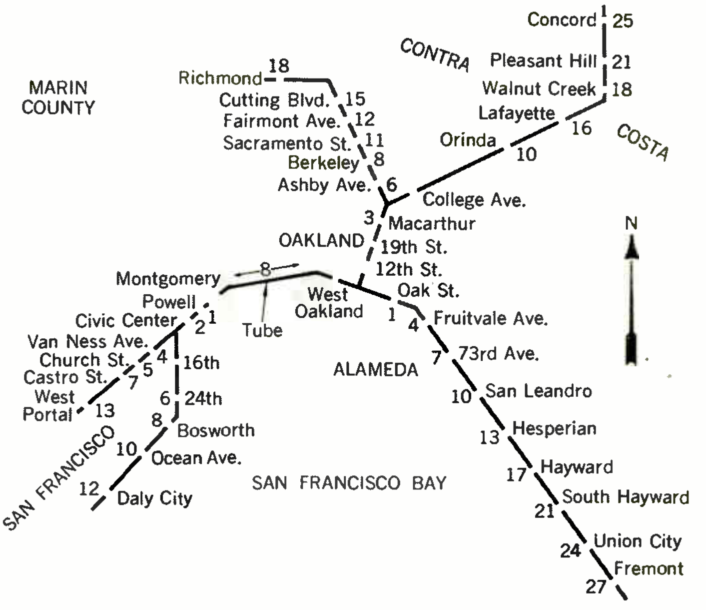


Fig. 1. Map showing proposed routes of the Bay Area Rapid Transit (BART) system. White dots indicate the proposed station locations, and the peak-hour travel time between stations is indicated in minutes.

the traction interface and the characteristics of the traction system. For descriptive purposes, the diagram refers to an analog current trainline, represented here as the P-signal. In the simplified and typical case shown, the P-signal is calibrated with respect to *pounds of tractive effort per unit weight of train.* The range of the P-signal is from *full braking* to *full propulsion* power. Loss of this signal will result in full braking of the train.

Propulsion equipment and power supply. In the paper, "Propulsion–Power Supply System for BART" (IEEE Paper No. 34 CP 66-205), Clare K. Olsen, a PBTB engineer, observed that nine electric transmission systems and 27 different concepts for train propulsion equipment either have been evaluated or demonstrated at the Diablo test track. The field has been narrowed to two transmission systems, 1000-volt dc and 3-phase 4160-volt ac (see Table I), and three types of propulsion units.

There are several different types of contact rail installed on the dc system for evaluation. These include 150-lb/yd steel rail, a copper contact surface with a steel base, a stainless-steel cap on aluminum base, a steel "T" insert on an aluminum base, and silicon-hardened aluminum on aluminum.

Two types of contact rail are being tested in the ac installation: extruded copper contact surface and a stainless-steel cap on aluminum base.

One of the characteristics of a rapid transit system, according to Olsen, is the tremendous power surge produced by the trains while accelerating. Olsen described some of the design problems that occur when a number of the BART fleet of 450 cars accelerate simultaneously. The average ten-car train will demand 8800 kW when accelerating from a standstill to a maximum speed of 80 mi/h.

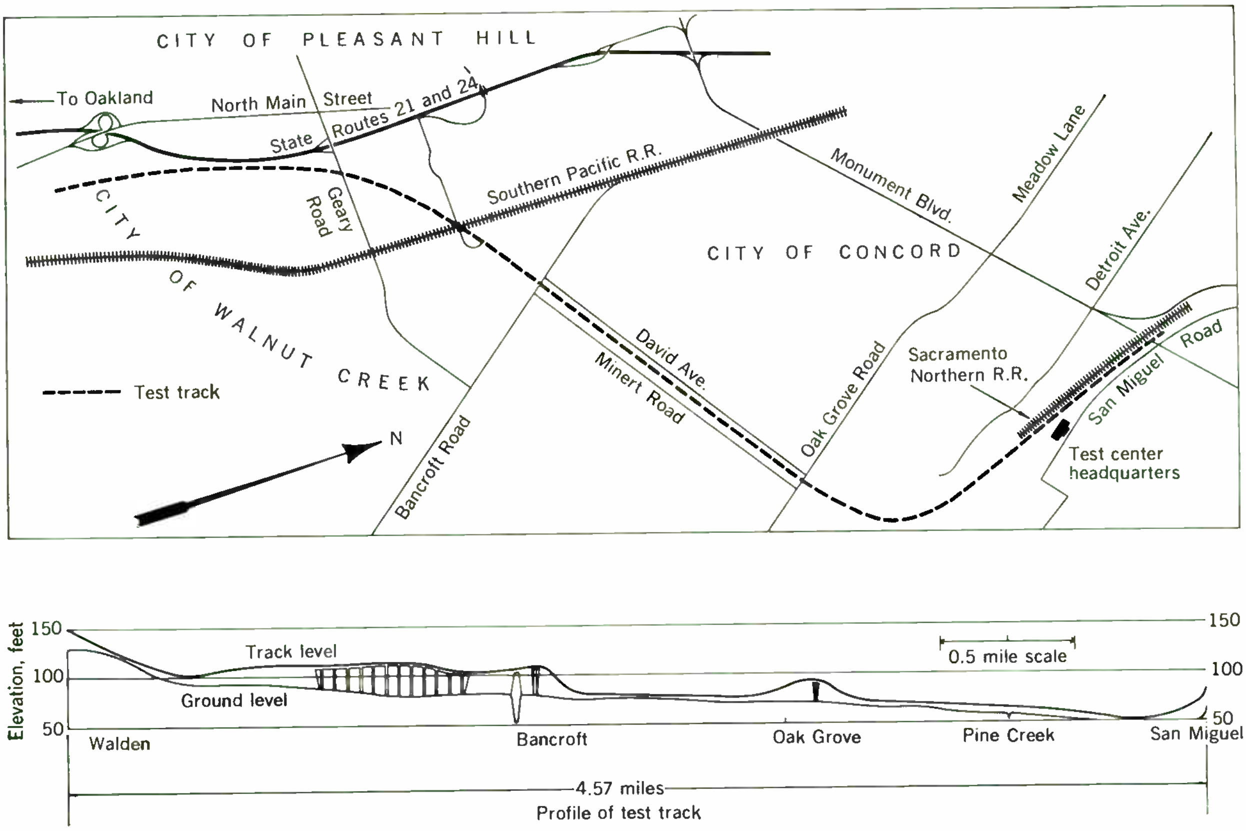
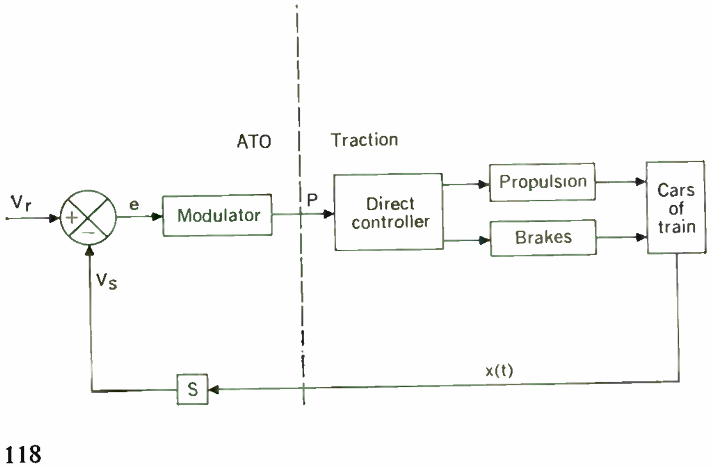


Fig. 2. Map and profile indicating route of the BART Diablo test track where exhaustive tests and trials of train components are presently being conducted.

Table II contains a description and specifications of some of the components, produced by various manufacturers, in connection with the BART testing program.

Traction systems. Four traction systems are presently being evaluated on the test track. These are a 1000-volt dc (conventional switched resistor), a 1000-volt "chopper," and two systems that receive power from the ac distribution system. One of the ac systems is installed on Car A_2 and operates dc motors by a phase-controlled rectifier; the other is a complete ac system, collecting three-phase energy (see Fig. 4), and it provides, through a dc-link design, a voltage variable-frequency source to truck-mounted high-speed ac squirrel-cage motors. This latter installation is in Car C_2.

Fig. 3. Block diagram showing configuration of a generalized automatic train operation (ATO) control loop.



Other significant papers

In the area of ATO concepts, the paper, "Automation of Heavy Freight Trains" (IEEE Paper No. 34 CP 66-202), presented by C. M. Hines and N. P. Erkel of the Westinghouse Air Brake Company, covered an area of automation that has only recently become an operational reality. This paper described the electronic equipment required (see Fig. 5) for the successful automation of a heavy ore train, operating over severe grades, for a distance of 75 miles. The descriptions included the three operating modes: automatic, manual, and remote. The results of this successful computer programming operation have been reflected in better train handling, with a minimum required equipment maintenance.

Other papers included "Computerized Weighing of Railroad Freight Cars Coupled in Motion," "An AC Electric-Drive System for Locomotive and Multiple-Unit Traction Systems," "Train Resistance, Power and Energy Requirement of M-U Cars," and "Performance and Power Consumption of Transit Vehicles."

Transportation education and safety

Portions of two sessions (Transportation V and VI) were allocated for the presentation of pertinent papers on education and on transportation safety. Professor William W. Seifert, Assistant Dean, School of Engineering, M.I.T., presented a paper on "Education of Engineers

I. Equipment assignment for BART laboratory cars

	Test Series I			Test Series II	
	Car A_2	Car B_1	Car C_1	Car B_2	Car C_2
Trucks	The Budd Co.	Le Tourneau–Westinghouse	General Steel Industries	Pullman-Standard	L. F. M.-Rockwell
Brakes	Westinghouse Brake & Signal Co. Ltd. (air brake–Budd disk)	Westinghouse Air Brake (electric brake–tread)	Westinghouse Air Brake (air brake–tread)	Westinghouse Air Brake (air brake–tread)	B. F. Goodrich (disk brake–hydraulic)
Propulsion system	Westinghouse Elec.	Westinghouse Elec.	General Electric	Westinghouse Elec.	Garrett Airesearch
Gears	Safety-Spicer	Le Tourneau–Westinghouse	General Electric	Westinghouse Elec.	Western Gear
Couplers	Westinghouse Air Brake (X-end) Ohio Brass (Y-end)	Waugh Equipment (X-end) Westinghouse Air Brake (Y-end)	Ohio Brass (X-end) Waugh Equipment (Y-end)	Waugh Equipment (X-end) Westinghouse Air Brake (Y-end)	Ohio Brass (X-end) Waugh Equipment (Y-end)
Door operator	Vapor-electric	Vapor-electric	Vapor-pneumatic	Vapor-electric	Vapor-pneumatic
Power source	4160-volt, 60 c/s, ac	1000-volt dc	1000-volt dc	1000-volt dc	4160-volt, 60 c/s, ac (regeneration)
Drive type	Phase-controlled rectifier	Chopper	Conventional switched resistor	Chopper	Frequency converter

II. Components for BART propulsion unit trucks

Manufacturer	The Budd Co.	Le Tourneau–Westinghouse	General Steel Industries	L. F. M. Manufacturing Co.	Pullman-Standard
Wheels	Tapered, lightweight steel	Cylindrical tread, standard steel	Tapered, standard steel	Cylindrical, standard steel	Cylindrical tread, aluminum centered, with steel rim
Axles	Hollow steel	Housed steel	Solid steel	Hollow steel	Steel
Motor	300 hp, car-mounted	150 hp, parallel drive	150 hp, parallel drive	150 hp, parallel drive	150 hp, parallel drive
Springs	Air	Hydraulic	Air and coil	Hanger suspension, with coil springs	Air and coil
Frames and material	Articulated frame, fabricated steel	Semirigid frame, fabricated steel	Articulated frame, cast steel	Articulated frame with ball joint and knee-action journal, support-cast steel	Articulated frame, fabricated aluminum
Brakes	Dynamic and disk (air)	Dynamic and tread (elect.)	Dynamic and tread (air)	Dynamic and disk (air)	Dynamic and tread (air)
Bearings	Inside journal (roller)	Inside journal (roller)	Inside journal (roller)	Outside journal (roller)	Inside journal (roller)
Special features	Coupled axles	Torque-proportioning differential	—	—	—

for Transportation Systems," in which he discussed the development of broad-gauge graduate programs in the technology of transportation. Seifert emphasized that, in order to attract students to such necessary and vital courses, the appropriation of subsidy funds in this area by the federal government is essential. He cited, as one illustration of the magnitude of transportation problems, the present and urgent need for trained and qualified technicians in the field of air traffic control. For example, Chicago's O'Hare Airport, one of the world's most congested, handles 1400 commercial planes a day, with a landing or takeoff occurring every 40 seconds during the hours of peak activity. Further, the United States maritime industry is faced with severe financial problems with attendant declining profits. Trained economists and technicians will find many career opportunities in this area alone.

The paper, "Air Traffic Control," by Alexander B. Winick of the Systems Research and Development Service, Federal Aviation Agency, discussed the increasing complexity of automated air traffic control by computer programming and the rapid obsolescence of hardware and software systems as the commercial air traffic density at major airports is intensified.

Captain W. C. Foster, U.S.C.G., Chief of the Merchant Vessel Inspection Division, discussed "The Future of Shipping Traffic Safety in the United States." Captain Foster alleged that "control of shipping traffic necessarily lags behind control of air, rail, and highway traffic; those three types of traffic have come into being more recently, move considerably faster, and necessitate positive control for safety purposes far more than ship traffic." He stated, however, that there are already in existence a limited number of sealane "traffic control lights" that are analogous to motor vehicle signals. These are essentially the international "Rules of the Road," universally observed

Fig. 4. View of a typical BART system car, incorporating a proposed high-speed three-phase ac conductor system. Heavy-duty crossarm-type collector will operate at speeds from zero to 80 mi/h in either direction, with vehicle acceleration and deceleration rates of 3.5 mi/h per second. The cross section to the right shows the configuration of the three-phase conductor. Each car will be 70 feet in length by 10½ feet in width by 10½ feet in height and will weigh 56 000 pounds. Each car will have a seating capacity of 72 people and will be equipped with four air-operated sliding doors on each side to permit rapid entrance and exit. An air-comfort system will provide uniform circulation of fresh air.

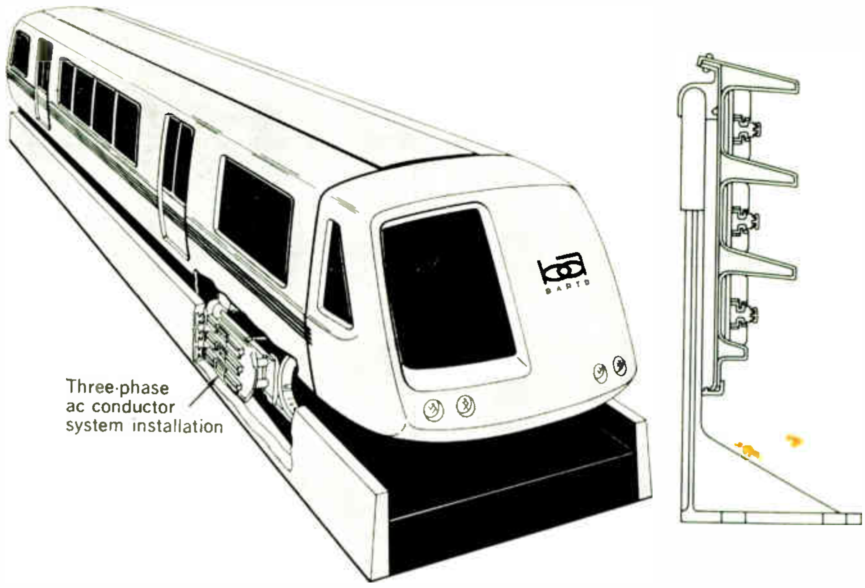


Fig. 5. System block diagram of ATO equipment used for the automation of a heavy freight train over a 75-mile-long run.