VOLUME 3 □ NUMBER 12 DECEMBER 1966

IEEE spectrum

features

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

the cover

Photochromism has been defined as the ability of a material to react reversibly to light. This phenomenon is discussed in the article beginning on page 39, which describes how photochromic glasses, for example, darken in varying degrees depending upon the intensity of the light source, and clear to original transparency when the light is removed.

Spectral lines

What has happened to specialization? Some time ago there was much talk about living in an age of specialization, much concern about the fact that our technology was getting more specialized, with ever-increasing difficulty in communication between disciplines. Surely you remember the apothegm, "the specialist is one who learns more and more about less and less, until he knows everything about nothing."

The pattern of engineering specialization was evident in most college catalogs. Out of what once were simply courses in engineering, civil engineering and mechanical engineering grew; then came electrical and chemical engineering, and later, electives in sanitation, transportation, structures, aeronautics, power, communication, biochemistry, system theory, computers, etc.—each field with its own special interest and, worst of all, its own special jargon. The situation became so bad in some places that even engineering college students could scarcely talk to each other. And the process continues, leaving some of us rather bewildered.

Even within a specialty such as ours, expansion and specialization are phenomenal. Where once two technical publications (the AIEE TRANSACTIONS and the PROCEEDINGS OF THE IRE) sufficed, now dozens appear; as many as 35 within our own Institute. One cannot begin to read all the publications, so one must select and specialize.

Must this process go on and on, like a Parkinson nightmare? Or does some unifying principle exist that can keep things within the bounds of human comprehension? There are at least two compensating effects, and perhaps several; the two that occur to me are *simplification* and *obsolescence*.

Fortunately, as understanding of a subject deepens, simplifying relations are found. Art becomes science and science becomes coordinated; common relations between seemingly diverse phenomena become evident and nonessentials are recognized. The evolution from letters to technical papers, to review papers, and to textbooks, and the continuous reorganization of educational courses, accelerate this process. If only it could keep up with the growth of knowledge!

There is comfort in the fact that the trend toward specialization in undergraduate engineering education is being reversed. Courses are becoming more basic, and some engineering schools have gone so far as virtually to eliminate specialization in undergraduate studies so that the engineer of the future will have a much broader base of common knowledge!

That this is to the good is evident when we consider what is happening in the field. In the area of power we find power plants controlled by computer, remotely observed by closed-circuit television, translated by silicon-controlled rectifiers, generated by atomic energy, plasmas, and fuel cells. We see the threat of sodium transmission lines, cryogenic transmission lines, waveguide and light-beam power distribution, and undersea power distribution. We hear of the possibility of harnessing sun power and earth heat. Power-line induced plasmas even produce UFOs. What next? Is there any future for a narrow specialist in the power industry?

For that matter, does he have a future in electronics? Look at what is happening to devices today. Resistors, capacitors, diodes, and transistors with interconnections are produced by the hundreds in one operation that might be classified as vacuum chemistry. Unless the small-device specialist starts to become integrated, he will be joining the ranks of the blacksmith, or the pre-Renaissance scribe. And the circuit specialist is no better off.

I would rather not mention what has happened to tubes; it's too painful and the point is clear enough. Old art is becoming obsolete and we are generating new lines of specialization. A specialist in any field had better keep a broad base of understanding and be ready to jump quickly. I pity the narrow specialist of yesterday and would forewarn the specialist of today. The price of excellence is specialization, but *the cost of too much specialization is obsolescence*.

What does this all mean to the IEEE? I think it means that we must be quick to see new lines of specialization, to recognize fields that are becoming obsolescent and to concentrate on the new ingredients of electrical engineering. We should seek common ground for various disciplines and find ways of unifying diverse viewpoints.

The present move to combine some of the smaller Groups is a step in the right direction. As I look back, I think that the introduction of IEEE SPECTRUM as a journal with intersubdisciplinary coverage of engineering and electronics was good. The merger too, however difficult, was probably inevitable for the same reasons.

I hear much about SPECTRUM's lack of coverage of certain technical areas, particularly power. Actually, the number of papers dealing directly with the power industry is not out of proportion to the membership in this field, but one is missing the point if he looks to SPECTRUM for papers in his own field. SPECTRUM cannot provide coverage in depth in all fields of interest; SPECTRUM *can* help one to keep abreast of the times and to be aware of the changing technology in fields other than his own. Scanning recent issues, I find that nearly two thirds of the articles contain material that should be of concern to the power engineer with a future. Most of them are called "interdisciplinary" today, but watch out for tomorrow!

C. C. Cutler

Authors

Chameleon in the sun: photochromic glass (page 39)

Gail P. Smith joined Corning Glass in 1941 after receiving the Ph.D. degree from the University of Michigan and serving there as an assistant in physics. At Corning he was concerned with the development of electronic components, especially of capacitors and of thin ribbon glass for them. He was responsible for the development of the CERCOR structure. In 1950 he was named senior research associate and, in 1961, manager of general product development. As such, he was in charge of the development of glasses for optical applications, including photochromic glasses, and of glasses and glass ceramics for structural applications. He is presently manager of International Research.

Fuel cells and fuel batteries—an engineering view (page 48)

H. A. Liebhafsky, a native of Zwittau, Austria-Hungary, received the bachelor's degree from the Agricultural and Mechanical College of Texas and the Ph.D. degree from the University of California; subsequently, he served as instructor at the latter institution. He joined General Electric's chemistry research staff in 1934 and is presently manager of the Electrochemistry Branch, Chemical Systems and Processes Laboratory, General Electric Research and Development Center. The author of over 100 scientific papers and co-author of *X-Ray Absorption and Emission in Analytical Chemistry*, he was the first chemist employed in industry to receive the Fisher Award in Analytical Chemistry.

The economics of desalination (page 63)

S. Baron received two degrees from Johns Hopkins University and then the Ph.D. degree from Columbia University. He joined Burns and Roe, Inc., in 1950 after serving three years with U.S. Industrial Chemicals Company and three years with Columbia University as a research assistant. As chief engineer at Burns and Roe, he was responsible for nuclear engineering, process engineering, and water treating. He supervised the pressurized-water design of the heat dissipation system of the new production reactor at Hanford, Wash., including shielding, equipment selection, waste handling, and water treatment. Presently, he is vice president in charge of engineering activities.

Fundamentals of proportional navigation (page 75)

Stephen A. Murtaugh has been affiliated with the Cornell Aeronautical Laboratory since 1952. Presently, he is in the Weapon Research Department. Engaged in the synthesis and analysis of missile and aircraft guidance, control, and navigation systems, he has also worked on the systems aspects of electronic countermeasures for tactical aircraft and airborne research radar systems and has directed several studies of tactical range surface-to-surface and surface-to-air missile systems, programs exploring the technical and economic feasibility of antiballistic missile defense systems, and analyses of performance requirements for optical, radar, and inertial sensors for both space navigation and homing seekers in AICBM and antisatellite applications.

Harry E. Criel joined Cornell Aeronautical Laboratory, Buffalo, N.Y., in 1955 and is presently with the Weapons Research Department. At CAL he has been concerned with systems analysis, particularly in the areas of aircraft performance, rigid body dynamics, air traffic control, navigation, weapon delivery, and sensors for space navigation. He has also been involved in extensive analytical and computer simulation studies of free-fall bombs, antisatellite missiles, and antimissile missiles. He is the recipient of two degrees in aeronautical engineering from the University of Michigan, Ann Arbor, and is a member of Tau Beta Pi, the American Institute of Aeronautics and Astronautics, and the Institute of Navigation.

The uses of a professional society

If an engineering society is to be effective, it should not be hampered by traditions handed down from societies formed to serve other objectives. Rather, it should reflect the needs of a membership that includes not only those who contribute to the field in a scholarly way but also those who implement science and technology for the public benefit

William G. Shepherd *President IEEE*

This year many of us in IEEE have had reason to ask ourselves what we expected of membership in the Institute. As president of the Institute, it was my responsibility to announce during this year the decision of the Board of Directors that an increase in dues was necessary if the health of the society and its fiscal integrity were to be maintained. When the announcement was made, the reaction of the members was sought and their suggestions as to how the Institute could serve them more effectively were solicited. Many of the responses received were discussed by the Executive Committee and the Board.

It is an entirely human reaction that a member's response came in terms of the immediate relevance of the Institute activities to his individual needs and the particular content of his job. Such immediate reactions understandably overlook the historical development of professional societies and the ends they were established to serve. In general, it seemed to me that the reactions and the discussion reflected some of the confusion in all of our minds as to the goals and purposes of engineering. I should like to review the historical background of professional societies and compare the traditions inherited from our past and the needs of engineering today. Hopefully this may suggest the appropriate uses of a professional society and, equally important, the obligation assumed by a member when he associates himself with the society.

The early societies

Professional societies had their origins in the scientific academies, of which perhaps the earliest in a form relevant to this discussion was the Academia Secretorum Naturae, founded by Della Porta in Naples in 1560. To become a member, an individual had to have made a discovery in natural science. The hazards of science in that day are reflected in the fact that the academy did not long survive, because Della Porta was suspected of practicing black art and summoned before the papal court to justify himself. Though acquitted, he was required to close his academy.

The academies that have had continuous histories and can most appropriately be regarded as the ancestors of our modern societies were founded in the 17th century. Notable examples are the British Royal Society and the French Academy of Sciences. The Royal Society developed from an informal organization established about 1645 composed in the words of the day of "divers worthy persons inquisitive into natural philosophy" who met weekly. For those concerned with the IEEE dues it is worthy of note that the membership fees were about seven dollars a year—a rather princely sum for the time, particularly since the members received no publications. In 1660 the organization received a royal charter and

Full text of an address presented at the National Electronics Conference, Chicago, Ill., Oct. 3–5, 1966.

assumed its present name with a membership restricted initially to 55. The scientific exchange between the members was clearly direct and the only publication was in the records made by a secretary. Interchange with members of similar societies was accomplished through correspondence. Out of the records of the secretary and the correspondence developed the *Philosophical Transactions*, which have persisted to our day. It is interesting to note that an important function of the Society at its beginning was the performance of experiments before the membership—perhaps one of the earliest organized forms of continuing education.

There were several distinctive features of these societies. Their concern was with basic science, and membership was restricted to those who had demonstrated an ability to make original contributions and who were prepared to participate directly in further contributions. The societies served as forums where new concepts could be presented and subjected to critical examination. They provided an organized means for developing archival records of scientific advances. In short, they were established to legitimize and facilitate an interchange of information between active practitioners in scientific study.

Prior to the establishment of these societies, interchange between philosophers was individual, either direct or through correspondence. The number of minds brought into interaction was obviously limited; in consequence, the pace of advance was slow. Widespread publication would not have been possible, of course, before the invention and development of printing, but this technique had preceded the founding of these societies by 200 years. Thus, the organization of the societies was in itself a significant event, providing opportunities for a wider interaction between individuals and an enhanced stimulation of further advances. The establishment of these societies was not by itself sufficient to produce our present science and technology, but it is clear that without them our present state of knowledge would be much more primitive.

James B. Conant expresses this point of view succinctly in his book, *Science and Common Sense*, as follows:

"The important thing which emerges from even a superficial study of the recent history of the experimental sciences (say since 1850) is the existence of an organization of individuals in close communication with each other. Because of the existence of this organization new ideas spread rapidly, discoveries breed more discoveries, and erroneous observations or illogical notions are on the whole soon corrected. The deep significance of the existence of this organization is often completely missed by those who talk about science but have no first hand experience with it. Indeed a failure to appreciate how scientists pool their information and by doing so start a process of cross fertilization in the realm of ideas has resulted in some strange proposals by politicians even in the United States."

This last comment of Conant deserves some emphasis. His book was published in 1951 when there was great concern over proposals to restrict publication of research findings touching on classified areas and with the stifling effect this would have on cross-fertilization of ideas. Fortunately that controversy had a favorable outcome. A concern about limitation of the cross-fertilization of ideas should, however, be before us as we consider the publication problems of our present-day societies in dealing with the specialization of our literature.

The missions of the scientist and the engineer

Although the engineering societies came into existence nearly two centuries later than the scientific societies, and brought together individuals having different objectives, the fundamental purposes in organizing were the same. Thus the constitution of the IEEE states:

"Its purposes are scientific, literary and educational, directed toward the advancement of the theory and practice of electrical engineering, electronics, radio, allied branches of engineering or the related arts and sciences. Means to these ends are the holding of meetings for the reading and discussion of professional papers, the publication and circulation of works of literature, science and art pertaining thereto and any other activities necessary and proper to the fulfillment of these objectives."

There is a major difference between the missions of the scientist and of the engineer, which has an important bearing on the organization and mission of an engineering society. The scientist has as his basic concern the expansion of the fundamental laws of nature without regard for the immediate usefulness of his findings. The engineer, in the words of the charter of the oldest engineering society, has as his declared mission the development of the "art of directing the great sources of power in nature for the use and convenience of man." His objectives, as is often overlooked by the humanists, are therefore more humanitarian than are those of the scientist.

If the engineer is to perform his assumed task effectively he must have a solid foundation in the basic sciences relevant to his field of concentration and keep abreast of advances in them. He must also be prepared to deal with practical economic considerations, reliability, and aesthetics, which together will insure the viability and acceptability of the products of his efforts. These requirements are broad indeed, and a professional engineering society ought to reflect this breadth and complexity. I have found rather sobering the lack of appreciation within some segments of the membership of the Institute for the breadth of this mission. It generates intolerance for the range of interests that must be served.

Goals of engineering education

If one considers the full range of activities required to put the forces of nature into the service of man at a time when technology has achieved its present degree of complexity, one begins to appreciate the difficulties of defining a single role for the engineer. One might better think first of the engineering mission, and then consider the major subdivisions of effort that must cooperatively relate if the mission is to be fulfilled. This concept has been very well expressed in "A Statement on Goals of Engineering Education," a position paper prepared by members of the Harvard faculty of Engineering and Applied Physics.

That paper divides engineering activity into three major categories—although it is carefully pointed out that one is, in fact, dealing with a continuous spectrum. Three categories delineated are:

1. Engineering technology
2. Engineering practice
3. Engineering science

The paper then goes on to summarize the activities of each category:

1. Engineering technology refers primarily to the application of well-established technology in production and service as well as in some of the supporting aspects of research and development. Training for this role has until recently been primarily done in two-year institutes. An individual engaged in this activity should be (a) well-versed in the current state of the art of a particular technology, capable of utilizing handbooks and other forms of codified information with skill and discrimination; and (b) sufficiently versed in mathematics and the sciences related to the particular technology in order to distinguish sound procedures from unsound ones and to keep up with the current innovations in his special field as they occur.

2. Engineering practice refers to the creative application of existing knowledge to the solution of specific engineering problems. It is not concerned primarily with the development of new knowledge or of generic solutions extending beyond the particular problem attacked. Individuals involved in these activities would normally have completed a four-year B.S. program and in some cases the M.S. program. These individuals are characterized by (a) an ability to handle mathematics and science related to a general area and to handle problems not in handbooks; (b) a greater concern with finding a needed solution to a specified problem than with an understanding of all respects of the science or mathematics involved; (c) an ability to synthesize practical designs that satisfy a number of requirements, several of which may be in conflict; (d) a sensitivity to economic factors and an ability to effect trade-offs between partially conflicting objectives; (e) an ability to utilize formal technical background and practical experience to solve problems that are new in detail, but not new in concept; and (f) an ability to direct large-scale technical operations by coordinating and supervising the efforts of appropriate specialists.

3. The engineering scientist is concerned with those fields of science that are of interest because they have existing or potential application. (Conant in the reference cited earlier suggests that the role of the engineering scientist aims at a reduction of the level of empiricism in engineering.) Thus he seeks an extension of the understanding of basic phenomena or the development of generic solutions so that their practical implications can be fully exploited. Individuals engaged in such activities generally would have been educated through the Ph.D. degree. Only in the motivation behind them would their activities and methods differ from those of the basic scientist.

I have paraphrased and quoted at length from the Harvard position paper because, to my mind, it so clearly delineates the total mission of engineering and its broad ranges of activities. The report aims primarily at the problems of educating engineering personnel, but I should like to use it as a springboard for discussion of the mission of an engineering society.

One of the reasons we suffer confusion about the purposes of an engineering society is that engineering in recent decades has been a profession in transition. Pre-World War II electrical engineering was characterized by a high degree of empiricism, whereas engineering today is much more firmly based on scientific understanding. Prior to World War II almost all engineers were involved with engineering practice and would have had few engineering technicians on whom to lean. There has been a marked movement in the direction of engineering science and an accompanying rapid increase in the engineering science literature. The result has been a tendency to regard engineering science as synonymous with engineering. Tensions that have developed between those in practice and those in science have beclouded the purpose of engineering societies. Indeed, these tensions are such that the titles chosen for the classification of activities are regarded by many as invidious.

I can only hope that my readers will take a charitable view of the problems of semantics with an assurance on my part that I have a wholesome respect for the validity of the full range of activities. Any attempt to categorize a complex field necessarily involves oversimplification. The categorizations chosen define roles that may be assumed by the same individual at varying times in his engineering career. The categorization of roles can be broadly applied for the totality of engineering activities of the Institute and in detail within each of the specialties.

We need to remind ourselves that the efforts of the engineering scientist would be uncalled for and sterile without the efforts of the engineering practitioner and that the efficiency of the latter would be seriously reduced if he did not have the support of well-prepared technicians.

Thus, if an engineering society is to support the mission of engineering, it should serve the full spectrum of needs of those who make the mission possible. This does not mean that every activity must be pitched at a level of intensity or abstraction that is understandable to every member. It does mean that the society should have activities that directly or indirectly serve in a meaningful way each segment of the engineering manpower complex. It requires respect for the validity of the total range of engineering endeavor and recognition that abstract contributions and the realization of useful devices are equally valid and important segments of engineering.

The very nature of the engineering mission, with all its complexities, suggests that other criteria for membership are appropriate for an engineering society than those derived from the traditions of earlier scientific societies. The latter societies properly insisted on publication of an original and scholarly work as a prerequisite for admission. However, the great majority of engineers are engaged in activities whose end result is measured by the usefulness of the product or services that is the outcome of their efforts. These outcomes may have more immediate economic importance, but they are not necessarily either more or less important than the findings of the research engineer. Thus admission to an engineering society should be open not only to those who contribute in a scholarly way to engineering science and technology but also to those who have demonstrated the ability to implement science and technology for public needs. And the society should provide for the needs of both.

The communication of ideas

The basic objective of an engineering society, however, remains the same as that of the scientific society, namely, the communication of ideas. It is appropriate that in an engineering society many of these ideas should have immediate practical relevance; but to insist, as some do, that this should be the overriding criterion in the allocation of support for the cost of publication is to ignore our responsibility to provide the foundations for the future technology of the profession. Too many forget that our present technology is founded on the heritage of abstract ideas passed on to us by our predecessors. Without the archival sources represented by the journals of professional societies, neither text and reference books nor the popular and immediately useful articles of the commercial publications could be written. The archival function of the professional societies is one of its most important features, and the acceptance of membership in the society carries with it an obligation to assist in its support. Many of our journals will closely relate in content to those published by scientific societies supporting the basic sciences undergirding our professional activities. This is both essential and proper. At the same time, we should not overlook the need to provide journals that interpret and broaden the understanding of new developments so that they can be quickly and usefully implemented.

A professional society provides the means for learning of new developments. Traditionally, the journals available for independent study and opportunities to participate were thought to be adequate. However, the vast expansion in our science and technology is making it increasingly difficult for an individual to keep abreast of the literature even in limited segments of a field. There is growing recognition that professional societies have a responsibility to develop publishing techniques that will facilitate more rapid retrieval of relevant information from the literature. Failure to do so will defeat the basic purpose of rapid communication of ideas. In recognition of this, the Institute, in partnership with other engineering societies, has embarked on a study of techniques for insuring a unified and effective approach to the management of the retrieval of information.

Earlier, in commenting on the quotation of Conant, mention was made of the problem presented by specialized literature. As a consequence of the explosion of technology in the last few decades, engineering societies have been subdivided to provide for special interests. The IEEE now exists as a closely knit federation of subsocieties, with separate publications serving the particular interests of professional groups. A single journal covering the contents of all of these publications and available to all the membership would be prohibitively expensive and would present most of us with the problem of storing an immense amount of unused literature. There is, however, a serious danger that the fragmentation of our literature in specialized journals will reduce the cross-fertilization between fields and ultimately work against the vitality of our technology. A society whose declared purpose is to maintain the intellectual vigor and breadth of its members needs to provide opportunities for easy communication between specialists and between specialists and generalists. Thus there is need for a journal edited with this objective in mind which is sufficiently free of the jargon of the specialist to be understandable to the nonspecialist.

Cross-fertilization can also be stimulated through appropriate organization of technical meetings. Within the Institute, professional Groups and Group Chapters through their technical meetings provide opportunities for communication between specialists in particular areas. With properly organized programs, the general meeting of the Institute and the various regional and Section meetings could have as their purpose cross communication between specialists and the serving of the needs of the generalist. Many of our Sections have expressed interest in this concept, and the establishment of a speakers' bureau was intended to encourage and support this interest. The Institute plans also to sponsor distinguished Institute lecturers commissioned to develop and present talks on new topics having broad implications.

The need for continuing education

Continuing education is a topic that has been much discussed in recent times. It is not a new idea, but the need for it to prevent intellectual obsolescence is now much more urgent in view of the rate at which our technology is expanding. These needs can be met relatively easily if access to educational institutions is readily available; if not, both the engineer and his employer are at a disadvantage. The provision of opportunities for continuing education by professional societies is entirely consistent with their objective of facilitating an interchange of information between their members. Such opportunities have been provided by local Sections of IEEE and its predecessor societies. A more concerted and coordinated effort by the Institute as a whole could make these opportunities more widely available and enable the exploitation of modern techniques whose use otherwise would be beyond the means of individual Sections. The recent action of the Board authorizing the initiation of a program of continuing education and providing for its support centrally is a step that will significantly increase the value of the Institute to its membership.

A professional society can serve as a vehicle for interaction between engineers in various countries. This is a readily accepted statement, but it remains true that the effectiveness of this interaction needs to be improved. A society whose affairs are dominated by the membership of a single country and oriented primarily to the needs of the membership in that country will be unattractive to membership elsewhere. For the development of international understanding, this interchange is sufficiently important to merit a major effort on the part of a professional society. The Institute has embarked on a program aimed at equalizing the service to members in locations remote from those in which there are major concentrations of the membership, so that they will feel more equal partners in a common undertaking.

The examples of the uses of a professional society that have been cited are not intended to be exhaustive but rather to illustrate the philosophy with which I believe we ought to approach the problem of making the Institute useful to its membership and the needs of society. My main message is that as members of an engineering society we should keep sight of the mission of engineering and recognize that its accomplishment requires a broad spectrum of talents and interests. We should not let the effectiveness of our activities as an engineering society be constrained by traditions inherited from societies formed to serve different objectives. If we keep in mind our own objectives we will be better able to structure the Institute so that it is a maximally effective instrument in the exploitation of the forces of nature for the use and convenience of man.

Chameleon in the sun: photochromic glass

Photochromic compounds have been known for a century but it is only recently that they have attracted serious attention. Their future promises to encompass many areas, including data storage and display, photography, and the protection of light-sensitive materials

Gail P. Smith Corning Glass Works

The properties characterized as photochromic—the ability to react reversibly to light—are found in various substances, both organic and inorganic. The two classes of materials are compared and the mechanism of the process is examined in each case. Of primary interest here are the inorganics, particularly the silver halide reversible photochromic glasses. Properties and ranges of properties of these are reviewed and some possible applications described.

Nearly a hundred years have passed since the appearance of the first published reference to the phenomenon that is now called photochromism. In a comprehensive review article, Brown and Shaw[1] note several early papers that describe color changes resulting from exciting radiation—but these identified references do not mention a still earlier art. It is recorded that Alexander the Great discovered a substance, whose composition has been lost in the obscurity of antiquity, that would darken when sunlight shone upon it. He dipped a narrow strip torn from the edge of his tunic into a solution of the material and wore this strip wrapped about his left wrist. Many of his soldiers did the same. By observing the changes of color during the day, they could tell the approximate hour. This became known as Alexander's rag time-band. (I am sorry that I cannot identify, and hence cannot give proper credit to, the author of this delightful footnote to history.)

In another, more recent review of photochromism, Schwab and Bertelson[2] establish a distinction between phototropism and photochromism. They define the more general reaction as "phototropy," or "phototropism," a spontaneously reversible change of a single chemical species between two states having different absorption spectra, with the change induced in at least one direction by electromagnetic radiation:

$$A_{\lambda 1} \underset{h\nu}{\rightleftharpoons} B_{\lambda 2} \qquad (1)$$

States A and B are usually ground electronic states, and are quantum mechanically stable, although one state in some cases may be a relatively long-lived excited electronic state that does not emit significant amounts of radiation. The restrictive case, in which at least one of the states absorbs visible light, is called "photochromism." This definition excludes flucrescent and phosphorescent materials, which re-emit light after irradiation. The reversibility of the change of state is a necessary criterion in distinguishing phototropic from ordinary and essentially irreversible photochemical processes.

Photochromic materials

Although photochromic substances have been known for nearly a century (both the Brown and Shaw and the Schwab and Bertelson articles carry extensive bibliographies), only during the last decade have these compounds become the subject of increased and serious attention. Much of this research has been supported by government agencies in view of the potential strategic importance of devices that react reversibly to light.[2, 3]

Organics. Schwab and Bertelson divide the reactions in different organic photochromic materials into some half-dozen categories; Windsor[4] reduces these to the following three main classes of general interest, based on how they work.

1. *Stereoisomers.* Absorption of light breaks one of the chemical bonds in a ring molecule, thus allowing the molecule to unwind and form a different geometrical arrangement. The reverse process is a re-forming of the bond. Examples of this class are the spiropyrans and the anils.
2. *Dyes.* A triphenyl methane dye, for example, is oxidized by energetic light; the absorption characteristics of the positive ion so formed are different from those of the original electrically neutral benzene rings.
3. *Triplet states.* In the class of polynuclear aromatic hydrocarbons, ground-state molecules are excited first to a singlet state by irradiation, and then go, via the lowest triplet state, to an excited triplet state. Visible light is absorbed in the triplet–triplet transition.

Inorganics. To the major classes of organic photochromics must be added several kinds of inorganics, also listed and described by Brown and Shaw[1]:

1. *Alkaline earth sulfides.* Traces of a metal such as manganese or bismuth appear to be necessary for photochromism.
2. *Zinc sulfide.* Lithopone, observed as early as 1881, is a compound of zinc sulfide and barium sulfate. The zinc sulfide appears responsible for the compound's sensitivity to light.
3. *Titania and alkaline earth titanates.* In the titanates, a contaminant, such as iron or any of several other metals, also appears necessary for darkening to occur.
4. *Mercury compounds.* Many of the mercury compounds, particularly those containing a halogen, have been observed to be photochromic.

In all of these materials, the photochromic response will depend on the intensity and spectral character of the incident light, on environmental parameters such as temperature, supporting matrix, or solvent, and, in most cases, on previous history. Most of the systems so far reported are only partially, or with difficulty, reversible, or are subject to fatigue—a change in behavior either with use or with time in storage. If the photochromic reaction is to be truly reversible, the quantum yield generally will be equal to or less than unity. When we compare this with a yield several orders of magnitude higher (in extreme cases, as high as 10^8) for ordinary silver halide photography, in which energy is added to the system chemically during development, we realize that photochromic processes are very "slow" in the photographic sense.[5] However, these light-sensitive materials are unique in that the image is formed directly and chemical processing to develop a latent image formed during exposure is unnecessary. In general, these inorganic materials are both reversible and reusable.

Glasses

Three general classes of photochromic glasses have been reported in the literature, and it is to these that I wish to direct primary attention.

Hackmanite types. Hackmanite is a naturally occurring mineral of the soda alumina silicate-sodalite group; it has the stoichiometric composition 18 ($Na_2O \cdot Al_2O_3 \cdot 2\,SiO_2$) $\cdot$ 3 NaCl $\cdot$ Na_2SO_4 and is supposedly a cubic crystal. It is usually opaque, white or blue, but can be melted to a glassy state, translucent to reasonably transparent, if a flux such as B_2O_3 is added. The minimum reported haze is 30 percent. With proper amounts of flux, the material darkens with exposure to ultraviolet light and can be bleached with longer-wavelength (visible) light.[6] Addition of other halides such as bromide and iodide can shift the absorption spectrum (color) of the resultant glassy material when it is darkened.

Cerium or europium. Cohen and Smith[7] report that in suitably purified base glasses, either of pure silica or soda-silica, the addition of small amounts of cerium or europium, typically 100 parts per million, has produced photochromic materials. Ultraviolet irradiation is absorbed by bands of cerium III or europium II centered in the ultraviolet and it transfers photoelectrons to nearby traps that absorb in the visible region, producing an amethyst color.

Decay times are typically a few seconds. Although the coloring and fading processes may be cycled repeatedly, the absorption band in the visible region (which produces the color) decreases in intensity with usage. This is believed to result from the photooxidation of the europium II to europium III. However, the band may be reduced, and the glass therefore resensitized, by exposure to short-wavelength ultraviolet light. These glasses, and

I. Composition of some typical photochromic silver halide glasses

Con-	Glass*						
stituent	1	2	3	4	5	6	7
SiO_2	60.1	62.8	59.2	59.2	60.1	52.4	51.0
Na_2O	10.0	10.0	10.9	14.9	10.0	1.8	1.7
Al_2O_3	9.5	10.0	9.4	9.4	9.5	6.9	6.8
B_2O_3	20.0	15.9	20.0	16.0	20.0	20.0	19.5
Li_2O	...	...	...	...	...	2.6	2.5
PbO	...	...	...	...	...	4.8	4.7
BaO	...	...	...	...	...	8.2	8.0
ZrO_2	...	...	...	...	...	2.1	4.6
Ag	0.40	0.38	0.50	1.50	0.40	0.31	0.30
Br	0.17	...	...	0.60	0.17	0.23	0.11
Cl	0.10	1.7	0.39	...	0.10	0.66	0.69
F	0.84	2.5	1.45	1.45	0.84	...	...
CuO	...	0.016	0.016	0.015	0.016	0.016	0.016

* Compositions are in weight percent; halogens are given as weight percent additions to that of the base glass.

their fatigue after exposure, have been studied in detail by Swarts and Pressau.[8]

Silver halide. Photochromic glasses that are truly reversible and do not show the effects of fatigue described have been reported by Armistead and Stookey[9] of Corning Glass Works. A wide range of base glasses has been found to be suitable; of these, alkali metal borosilicates are perhaps best from the standpoint of both general glass qualities (clarity, durability, ease of melting, and forming) and photochromic behavior. As for other photochromics, the composition and thermal history of the glasses play a large part in determining their resultant photochromic properties.

Silver halide glasses—composition and structure

Some typical compositions for silver halide photochromic glasses are given in Table I.[10] Most of the glasses investigated thus far are transparent in the unexposed state, darkening to a gray or reddish gray when illuminated. Glasses 6 and 7 in the table have had heavy metals added to increase their index of refraction to that required for ophthalmic use. At high concentrations of the silver and halogens, the glasses are photochromic—and either translucent or opaque. The upper limit of silver for the transparent glasses is usually about 0.7 percent by weight. The addition of other metals in the form of polyvalent oxides, including arsenic, antimony, tin, lead, and copper, increases the glasses' sensitivity and photochromic absorbance.

From observations of the glass, and by analogy with the known properties of silver halides, we can adduce several compelling reasons for asserting that the photochromic behavior of these glasses results from the silver halide crystals within them:

1. Heavy metals and halides are essential for photochromic behavior in glasses; silver is commonly used. As in bulk silver halides,[11] a small amount of copper oxide is an effective sensitizer for the glasses.
2. Phase separation is necessary in order that the glasses be photochromic. Crystalline silver chloride has been identified by X-ray diffraction in typical glasses.
3. The melting point of the separated phase is less than the annealing–strain point range of the glass (approximately 470–500°C).
4. The sensitizing optical absorption bands lie in the same wavelength region for the glasses and for crystalline silver halides.
5. The shapes of the optical absorption curves are similar for the two materials.
6. The rate of formation of color centers decreases with decreasing temperature.
7. Both types can be optically bleached, and the rate decreases rapidly with decreasing temperature.

The crystals are formed by precipitation from the homogeneous glassy matrix during initial controlled cooling or during a subsequent heat treatment (usually desirable for glasses in the lower ranges of silver halide concentration) at a temperature typically between the strain point and the softening point of the glass, long enough for the crystals to grow to their optimum size.

Electron micrographs of fractured surfaces of the photochromic glasses show small dense particles not seen in glasses that, either because of composition or heat treatment, are not photochromic. In Fig. 1, which is a photograph of a carbon replication of a typical fractured surface, the crystals are spheroidal because, as the glass cooled, they remained molten droplets until the glass had become rigid. Their average size and number can be determined, within limitations, by counting from such photographs, and, with more precision, by small-angle X-ray scattering. In general, glasses with particles that are less than about 50 Å in diameter are not photochromic. As the time or temperature of heat treatment for any one glass is increased, the average number of particles is reduced and

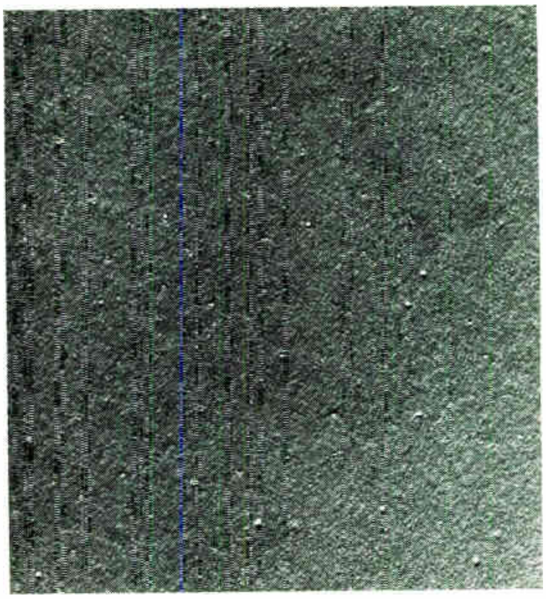
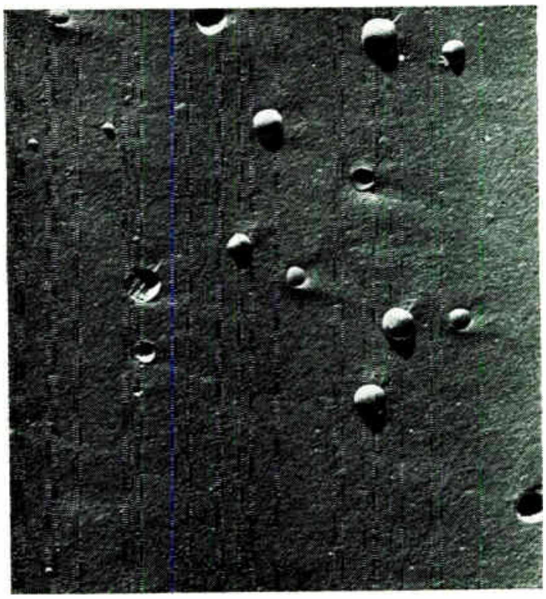


Fig. 1. Fractured surface of typical transparent photochromic glass, by carbon replication. Scale: one inch = 1.2 micrometers.

Fig. 2. Fractured surface of translucent photochromic glass, by carbon replication. Scale: one inch = 1.2 micrometers.



their size is increased, as would be expected from classical nucleation theory. Above about 300 Å in diameter, the particles scatter light and the resultant glass is opal. In the glass in Fig. 2 the crystallites were deliberately grown to relatively large sizes in order to show their spheroidal character more clearly. This glass, in which the particles were as large as 0.2 micrometer in diameter, was translucent.

For particles of average diameter (100 Å) present in a concentration of, say, 0.2 percent in the glass, there will be about 4×10^{15} particles/cm^3, with an average spacing of 600 Å between them.

In the photolysis of silver halide crystals, as in conventional photographic film, a latent image particle is formed from which elemental silver is developed in the traditional chemical processing, and the halogen diffuses away from the original crystal site. The photographic process may then be represented by the interaction of the incident photons with the silver halide crystal:

$$n\mathrm{AgCl} \xrightarrow{h\nu} n\mathrm{Ag}^0 + n\mathrm{Cl}^0 \nearrow \tag{2}$$

$$\swarrow$$

$$(\mathrm{Ag}^0)n, \text{ the latent image particle}$$

In the silver halide glasses, the halogen is held within the surrounding glass matrix and is available for recombination with the silver, which permits recovery of the glass to its original state after the light is removed. There are two independent recombination processes: (1) a natural thermal recovery, and (2) interaction with light of longer wavelength (lower energy) than that which darkens the glass, an optical bleaching. The unique behavior of these glasses results from the existence of these reverse processes.

In the photochromic process the reaction takes place as follows:

$$\mathrm{AgCl} \underset{\Delta,h\nu_2}{\overset{h\nu}{\rightleftarrows}} \mathrm{Ag}^0 + \mathrm{Cl}^0 \tag{3}$$

When copper is added to the glass in small amounts under reducing conditions it acts as a hole trap in the following reaction:

$$\mathrm{Ag}^+ + \mathrm{Cu}^+ \underset{\Delta,h\nu_2}{\overset{h\nu_1}{\rightleftarrows}} \mathrm{Ag}^0 + \mathrm{Cu}^{++} \tag{4}$$

to increase the amount of neutral silver atoms. (See Moser *et al.*[11] for a discussion of the role of copper as a sensitizer of silver halides.)

Photochromic properties of silver halide glasses

General behavior. The large possible ranges and variations in composition, coupled with variations in temperature and time interrelations of any subsequent heat treatment, give rise to wide latitude in photochromic properties, to greatly different rates of darkening and of recovery, and to a wide range of dependence of reaction rates and equilibrium states on temperature. The glasses are darkened by absorption of high-energy photons in the near ultraviolet or shorter wavelength visible region of the spectrum. The long-wavelength limit of the spectral sensitivity for darkening is higher for glasses containing heavier halogens. The spectrum of the light that induces darkening, as well as that which is most effective in optical bleaching, is continuous; there appear to be no sharp

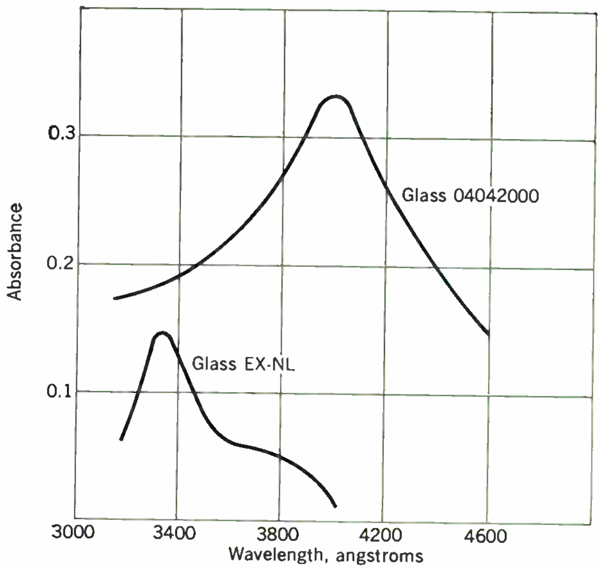


Fig. 3. Spectral sensitivity for two selected glasses. The glasses are activated by relatively long-wavelength ultraviolet or, in some compositions, by short-wavelength visible light.

Fig. 4. The approach to equilibrium absorbance at different levels of incident energy. Wavelength of the activating light was 4000 Å.

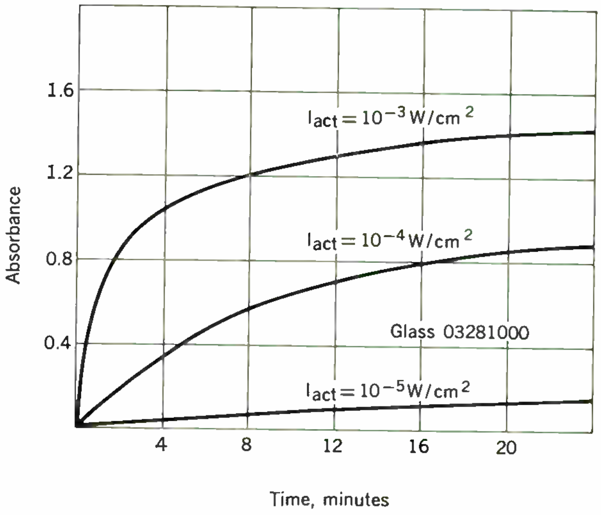
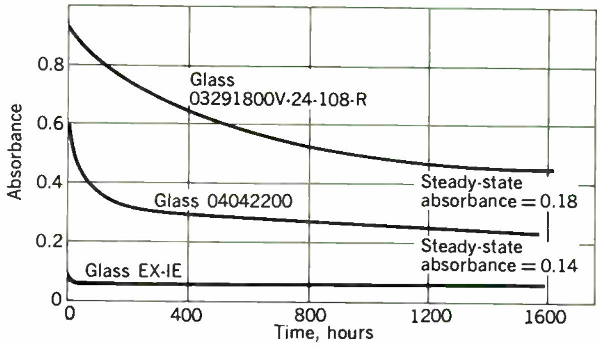


Fig. 5. Recovery of three glasses of different fading rates, after activation. The glass labeled EX-IE has a relatively very high thermal fading constant.



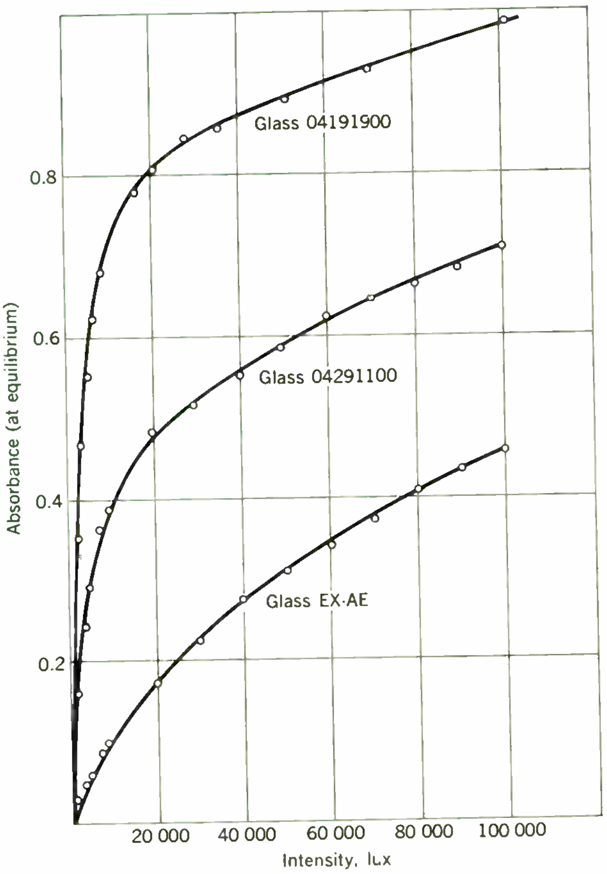
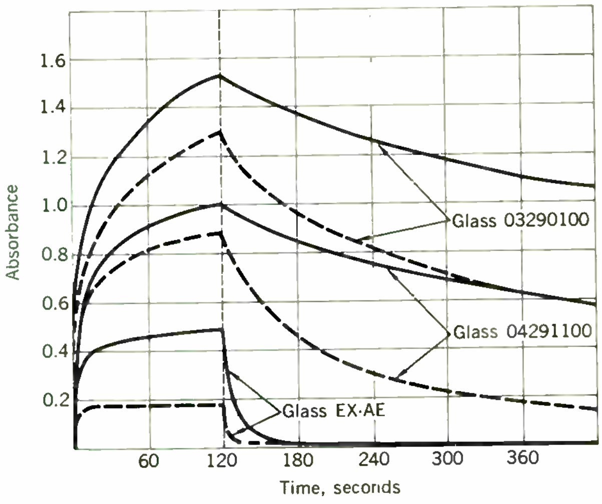


Fig. 6. Steady-state optical density versus light intensity, at constant temperature. The increase of optical density with increasing light intensity, and the equilibrium density, vary with the glass. The thermal fading rate constant of glass EX-AE is greater than that of glass 04291100. The glass labeled 04191900 is relatively a very slowly fading glass.

Fig. 7. Solid lines indicate the photochromic darkening and fading in three representative glasses at room temperature (23°C), with light of constant intensity. Dashed curves show photochromic darkening and fading in three representative glasses at 46°C. Length of exposure in both cases was 120 seconds. Comparison of the latter curves with those of Fig. 5 shows the interpendence of equilibrium absorbances, rates of approach to equilibrium, and temperature.



absorption lines or edges for either process.[5,12] Figure 3 shows the spectral sensitivity for activation of two different but representative glasses. At normal temperatures, the rate of darkening depends primarily on the intensity of the light—in the proper spectral region. Figure 4 shows the approach to equilibrium absorbance for a glass illuminated at three different intensities. The rate of recovery is determined mainly by glass composition and heat treatment. Recovery after irradiation is shown in Fig. 5 for three different glasses. Recovery (to half-maximum absorbance) in the dark at room temperature is measured in times that range from seconds to hundreds of hours.

Darkening and fading phenomena. For silver halide crystals, although the simple assumption of a single species and a single process is almost certainly not true, it does permit generalizations about the behavior of the glasses.[13] Under illumination, the change of concentration of absorbing color centers for this presumed mechanism will be given by

$$\frac{dc}{dt} = k_d I_d A - (k_f I_f + k_t)c \qquad (5)$$

where c is the concentration of color centers; k_d, k_f, and k_t are rate constants for darkening, for optical bleaching, and for thermal fading; I_d and I_f are the integrated intensities of the light, darkening and fading, over the respective wavelength ranges to which the glass is sensitive; and A is the number of sensitizable sites in the glass. When equilibrium is attained, $dc/dt = 0$, and the equilibrium concentration of absorbing centers will be as shown in the following:

$$c_s = \frac{k_d I_d A}{k_f I_f + k_t} \qquad (6)$$

Thus, the photochromic behavior of any glass will be determined by the relative magnitude of the rate constants describing it. These constants in turn are determined by the composition of the glass and by the state of the crystals produced within it; i.e., by its thermal history. If k_t is vanishingly small, then c_s is independent of the light intensity (assuming constancy of the ratio of darkening to bleaching light intensities); if k_t is large and becomes the determining rate constant, then c_s is proportional to the intensity.

The photochromic darkening for three glasses selected to have a wide range of darkening and fading rates is depicted in Fig. 6. The light source for these measurements was a high-pressure xenon arc; light intensities were measured with a photovoltaic meter. Glass thickness was about 6 mm for these samples. The relative linearity of absorbance with intensity is seen to be much greater for glass EX-AE, with a high thermal fading rate constant, than for glass 04291100 of intermediate fading rate, and for slowly clearing glass 04191900. The short time approach to equilibrium of three selected glasses under constant illumination at room temperature (23°C) is seen in Fig. 7(A); Fig. 7(B) shows the behavior of these glasses at 46°C, with the same light source. The shutter of the xenon arc was opened at time zero and closed after 120 seconds. The glasses were chosen to show differences in fading rate and equilibrium absorbance. The samples were maintained at the stated fixed temperature, so that the glass temperature was not appreciably raised by the energy

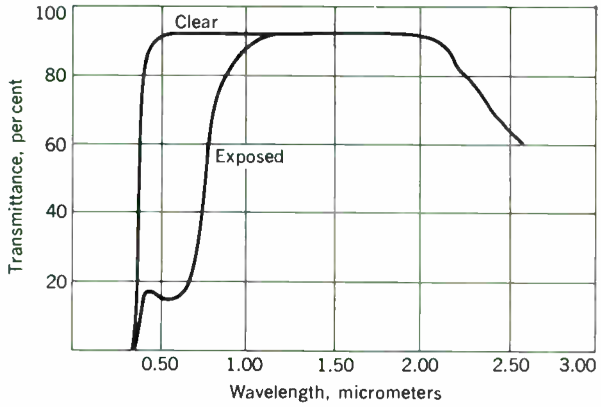


Fig. 8. Spectral transmittance of a typical photochromic glass (05171200). The glass, after exposure, has high absorbance in the visible region of the spectrum, with reduced amount in the infrared.

Fig. 9. Spectral distribution of energy from the sun (air mass 2) transmitted by a photochromic glass. Most of the reduction of energy is below 1-μm wavelength.

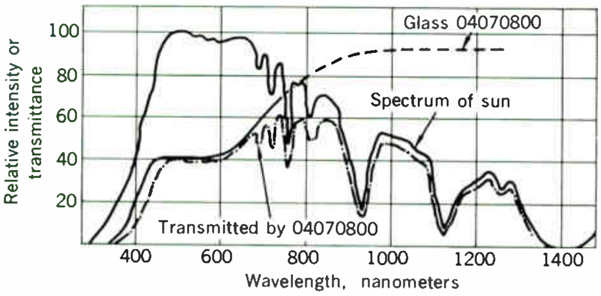
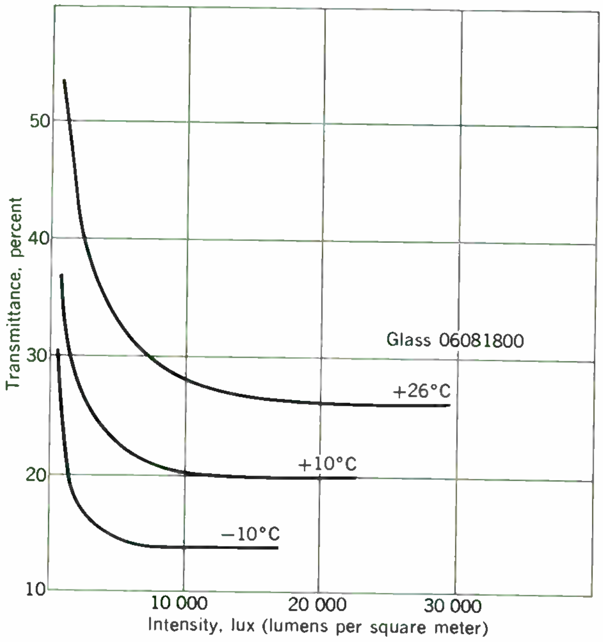


Fig. 10. Transmittance versus increasing intensity, in sunlight at three temperatures, for a glass with relatively moderate thermal fading rate.



absorbed. The increased effect of change in ambient temperature on the thermal fading rate, and therefore on the equilibrium absorbance, of the faster clearing glass, EX-AE, is seen from these curves.

Response to sunlight. The measurements were made with artificial and controlled sources that were relatively rich in ultraviolet light. Sunlight, at the surface of the earth, represents an uncontrolled light source with the amount of ultraviolet light, and its ratio to visible light, constantly changing throughout the day and from day to day with the change in the weather and seasons. The response of glasses to sunlight is important in many applications, such as in eyeglasses and in windows in buildings, automobiles, and aircraft. Curves of transmittance versus wavelength for one glass (6 mm thick) are shown in Fig. 8. From the transmittance at equilibrium and the spectral distribution of energy of the sun at the surface of the earth,[14, 15] the amount of energy absorbed by the darkened glass can be determined.[16] Figure 9 presents a trace of the spectral distribution of the energy of the sun at the surface of the earth (air mass 2), together with the spectrum of the amount of that energy transmitted by a typical darkened photochromic glass.

For most uses in sunlight, we would, by following Eq. (6), choose glasses in which the thermal fading rate constant k_t has been made relatively large so that (1) the equilibrium absorbance increases with the intensity to as large values of intensity as possible, and (2) the absorbance decreases as rapidly as possible when the illumination goes to very low values, i.e., at night. However, since k_t would be expected to be temperature-dependent, this implies a glass composition whose transmittance is determined to a large extent by the temperature of the glass. This dependence on temperature is noted in the curves of Fig. 7.

A typical glass (06081800), which has a favorable balance of several photochromic parameters for use in buildings, shows a reduction in transmittance with increasing solar intensity for three mornings of different ambient temperature as in Fig. 10. These curves exhibit the expected increase in equilibrium transmittance with temperature, and also the increase in intensity of illumination at which equilibrium is approached at the higher temperature.

When the transmittances of the photochromic glasses are recorded over an entire day, they show a characteristic pattern: their transmittance begins to decrease at dawn (actually before sunrise because of the ultraviolet light scattered to the glasses by the atmosphere), continues to decrease until saturation is achieved, remains at approximately that transmittance throughout the day, begins to increase before sunset, and continues to clear at a constantly reducing rate until the next morning, when the pattern is repeated. Figure 11 shows smoothed traces of several selected glasses (6 mm thick). Temperature here was nct externally controlled; it was the equilibrium temperature for these glasses mounted in a vertical, south-facing panel.

The transmittance versus wavelength curves for all of these glasses are similar to that of the 05171200 glass shown in Fig. 8, so that the glass labeled 03281800, with lowest equilibrium transmittance, would probably have been warmest. Note that the characteristic shape of all these traces is the same: any one of them can, by translation of the entire curve along the transmittance axis or by stretching or compressing the transmittance scale, be reasonably well superposed on any of the others. These curves appear smooth partly as the result of the compensatory dependence of transmittance on intensity and on temperature. That is, as the sunlight intensity increases, the absorbance of the glass increases. At solar intensities high enough to produce an appreciable temperature increase in the glass, however, the rate of change of absorbance with intensity is small (compare Fig. 10), and this is offset by the decrease in equilibrium absorbance resulting from the increase in the thermal fading rate of the glass.

It follows, then, that changes in intensity produce changes in transmittance at high-intensity values when the temperature of the glass is controlled. The transmittance of a selected photochromic glass mounted in a temperature-controlled enclosure, and double-glazed with commercial soda-lime plate glass versus the intensity of the sun for a typical day, is shown in Fig. 12(A.). On this summer day the trace of the incident light, here on an arbitrary scale, shows a reasonably bright morning, heavy cloud around 1000 hours, variable cloud and sunshine in early afternoon, and a relatively clear evening. The corresponding trace of transmittance shows darkening of the glass from dawn to about 0700 hours, with clearing after about 1700 hours. The glass clears from about 40 to 52 percent transmittance with the heavy cloud at 1000 hours that reduced the illumination, measured normal to the glass, from about 35 000 to about 6000 lux (lumens per square meter). Also shown is the response to incident light intensity change resulting from a cloud as seen just before 1300 hours.

A similar glass, similarly glazed, is seen in Fig. 12(B) to increase in transmittance from 30 to 57 percent during a very severe storm, when the external illumination decreased from 28 000 lux to about 500 lux. Therefore the light transmitted, when it became relatively very dark outside, was about twice that which a window of fixed transmittance of 30 percent, such as a gray light-absorbing window, would have admitted.

Optical behavior and applications. Another, potentially important application of photochromic materials is in the display of information. Data can be recorded in photochromic glass in two ways: by darkening the glass with short-wavelength light in a desired pattern; or by uniformly darkening the glass and bleaching it, in the desired pattern, with longer wavelength light.

This application is demonstrated in Fig. 13.[5] The 1-mm-diameter spots were produced by activating clear glass, and by bleaching, at a longer wavelength, previously exposed glass. To achieve both maximum change of absorbance and persistence of the stored information, a glass of relatively low thermal fading rate would be used. Figure 14 shows some typical bleaching curves for different activation levels. The bleaching light was at a wavelength of 6000 Å, with 5×10^{-3} W/cm^2 applied to the sample.[5]

As pointed out previously, in general, glasses that become the darkest are the slowest to clear, which implies a cycling time too slow for display applications involving normal information rates. However, the rate of clearing can be accelerated by external heating or by overall exposure to wavelengths longer than those that are used for darkening.

Figure 15 compares the recovery times of a glass sample exposed to ambient temperatures of 200°F (94°C) and 300°F (149°C) with that of a control sample that was allowed to recover at room temperature.[17] The data indicate that heating the sample in this way will increase its fading rate.

More rapid heating can be obtained, as seen in Fig. 16, by coating the sample with a transparent, electrically con-

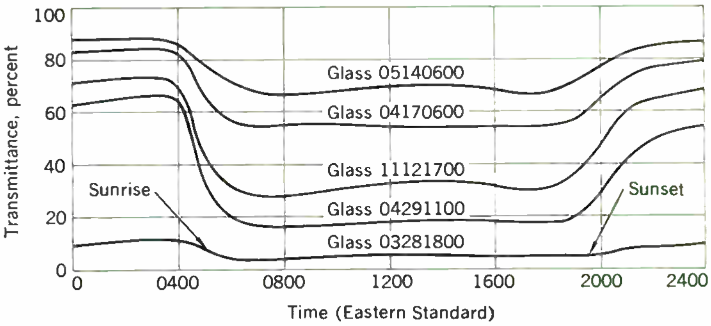


Fig. 11. Transmittance of five representative photochromic glasses throughout a midsummer day (June 26) in Corning. The night of June 26 was cooler than that of June 25.

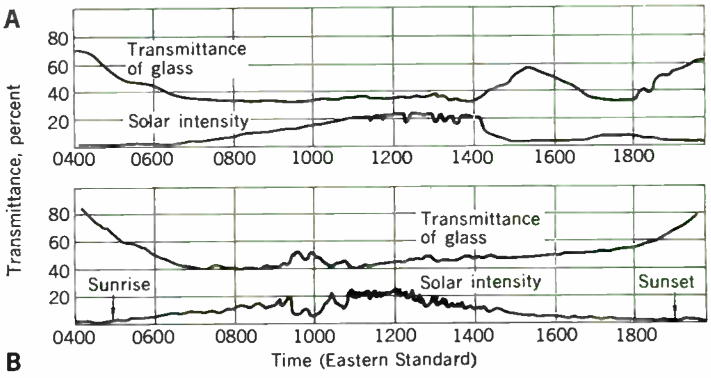
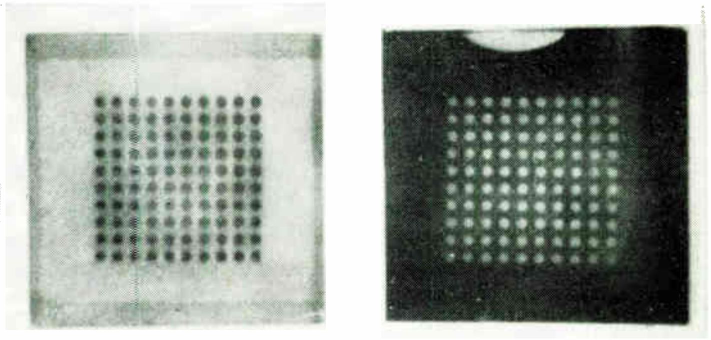


Fig. 12. Transmittance of selected photochromic glass, double-glazed with soda-lime glass (A) throughout a midsummer day (June 9) in Corning; and (B) on a day (July 2) with a severe thunderstorm.

Fig. 13. Digital information as dark spots on previously clear glass, or as bleached spots on previously activated glass. Diameter of spot is 1 millimeter.



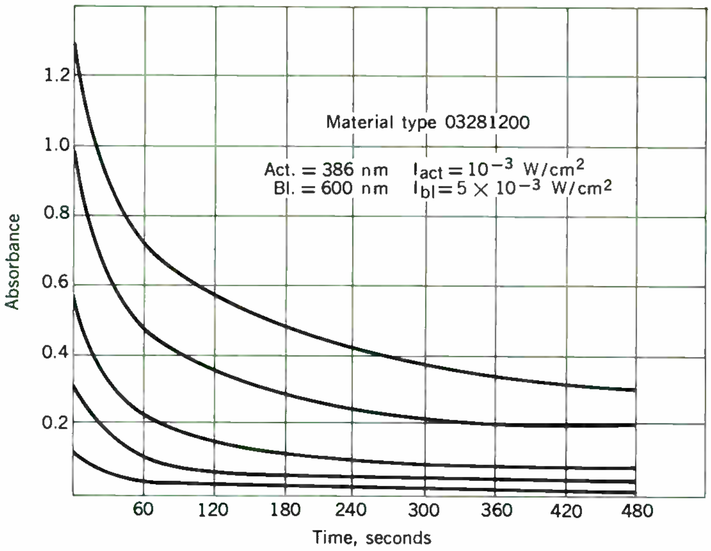
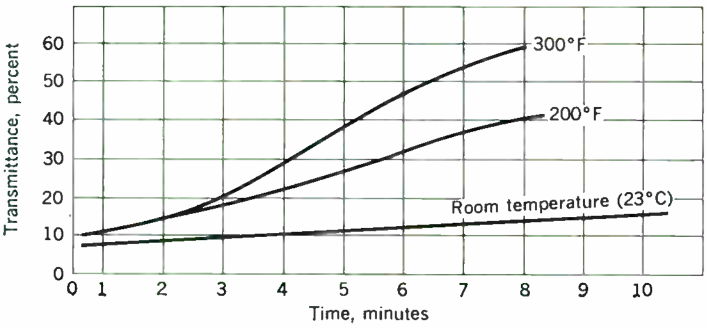


Fig. 14. Decrease of absorbance with time by bleaching from different activation levels. The energy of the bleaching light was 5×10^{-3} W/cm² at 6000 Å.

Fig. 15. Accelerated fading of photochromic glass (type ED-LD) by heated ambient air.



ducting tin oxide coating. By reducing the thickness of the glass, and hence increasing the heating rate, the glass was made to recover essentially completely in less than ten seconds.

The ultimate resolution obtainable in these glasses may be very high. The crystallite sizes are at most a few hundred angstroms; the crystallite separation is an order of magnitude larger—which is small compared with that in other photographic materials. High-speed photographic materials of low resolution contain crystal sizes of 20 000 Å and low-speed materials of very high resolution have sizes of approximately 1000 Å.

An additional illustration of resolution, and of possible applications of photochromic glasses to photography, is seen in Fig. 17. Figure 17(A) is a photographic positive, made in the conventional manner from a film-based negative. Photochromic glass, 0.15 cm thick, was exposed to ultraviolet light through this negative, and the resulting positive was photographed; Fig. 17(B) is the positive print made from the resulting negative. It can be seen from the photograph that a gradient in photochromic density under uniform illumination can be produced within this glass.

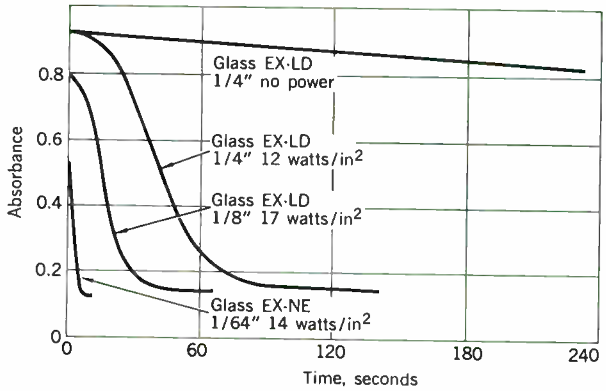


Fig. 16. Accelerated fading of photochromic glass by heating with conductive coatings on the glass. The glasses reached temperatures of the order of 200°C, for rapid recovery.

Reversibility. No significant changes in photochromic behavior have resulted from cycling glass samples with a commercial "black light" source (3600 Å) up to 30 000 cycles. There were also no apparent solarization effects causing changes in darkening or fading rates after accelerated ultraviolet exposure equivalent to 20 000 hours of noonday sunshine. A sample has also been cycled, one cycle per minute, using a constant output source with switching filters transmitting at wavelengths of 4000 Å and 6200 Å; i.e., "writing" and erasing. The activation energy of 1 mW/cm² and the bleaching energy of 13.3 mW/cm² produced a cyclic 0.3 change in optical density. To date, there is no apparent fatiguing after more than 300 000 cycles[5]; the tests are being continued in order to accumulate more conclusive data.

Conclusion

What does the future hold for photochromic materials? We have seen possibilities for glazing, for information storage and display, and for photography. In a combination of the last two, a 1245-page Bible was reproduced on a plastic strip only 5 cm square, a reduction of the order of 50 000 to 1 in area.[18] Sunglasses made with organic photochromics are now on the market and prescription lenses of photochromic glass have been made commercially available.[19]

With the development of materials that react quickly enough, and to a high enough optical density, eye protection against nuclear bursts may be possible. The characteristics required for nuclear flash protection have been set down in some detail[20]:

1. Flash detection and triggering within 10 μs.
2. Shutter closure within 50 μs.
3. Visible light transmission of 75 percent open and 0.01 percent closed, essential over a 20° field and desirable over a 60° field.
4. Minimum transmission outside the 0.4- to 0.7-μm region.
5. Resolution of at least 5 seconds per centimeter of aperture.
6. Clear aperture compatible with optical instruments.
7. Unlimited reuse.

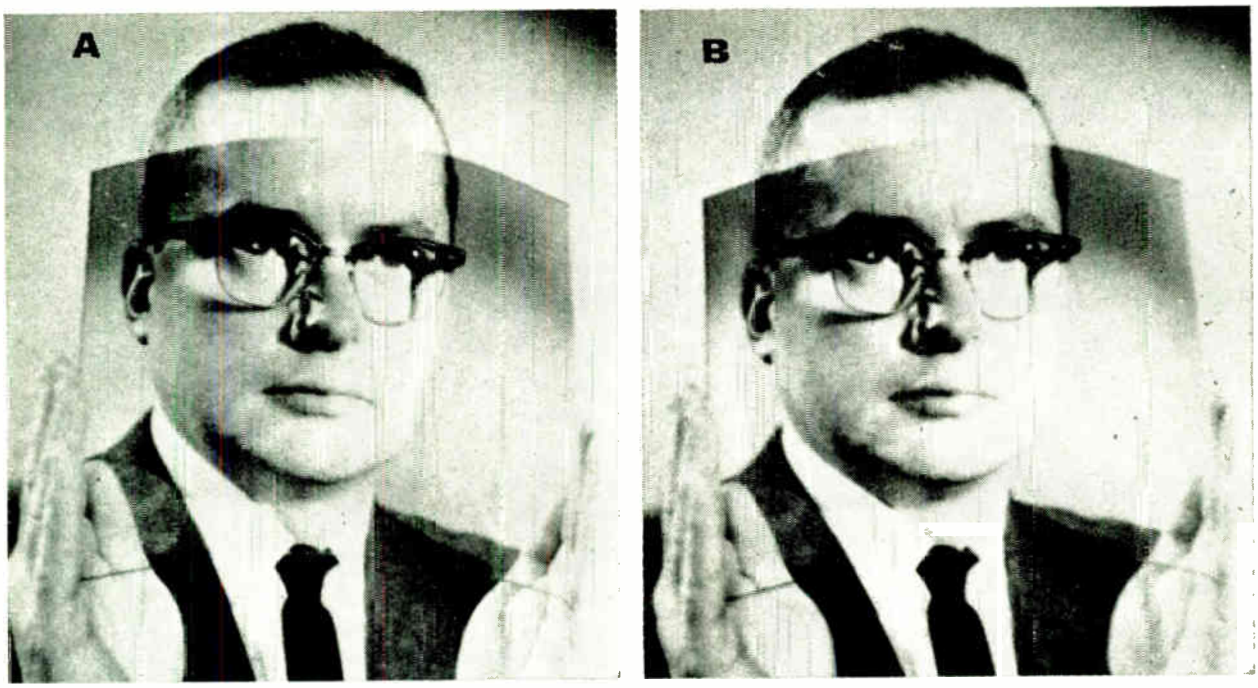


Fig. 17. The use of photochromic glass as a temporary positive in photography. (A) Conventional photographic positive. (B) Positive made from foregoing film with photochromic glass used as an intermediate positive.

8. Permissible operating temperature from −40°C to 52°C and storage temperature from −62°C to 74°C.
9. Resistance to nuclear and thermal radiation of 10 kt at 500 meters.
10. Resistance to shock and vibration of armored vehicles.
11. Power requirements: man pack or vehicle.
12. Minimized production and logistic problems.

These requirements have been amplified and revised for some specific development projects, but even so they present a formidable set of specifications for meeting a critical need. Kropp *et al.*[21] report that a combination of several dye-enzyme systems can meet some of the most difficult of the current requirements and that, with further work, all requirements for an operational system show promise of being met.

Other possible uses for varied photochromic materials are as dyes, paints, or coatings for radiation control; for bottles and containers for light-sensitive drugs and foods —and, of course, beer. And they can even be used to make a sun-tanning doll for the children.

REFERENCES

1. Brown, G. H., and Shaw, W. G., "Phototropism (photochromism)," *Rev. Pure Appl. Chem.*, vol. 11, no. 2, pp. 2–32, 1961.

2. Schwab, H., and Bertelson, R. C., "Photochromism—state-of-the-art review," presented at 1966 Symp. on Unconventional Photographic Systems, Soc. of Photographic Scientists and Engrs., Washington, D.C., pp. 94–106.

3. "Photochromism and phototropism," U.S. Dept. of Commerce, Nat'l., Bur. of Stds., Clearinghouse for Federal Scientific and Technical Information. Springfield, Va., Nov. 1964.

4. Windsor, M. W., "Photochromism," in *Encyclopedia of Chemistry*, G. L. Clark, ed. New York: Reinhold, 1966, pp. 816–818.

5. Megla, G. K., "Optical properties and applications of photochromic glass," *Appl. Opt.*, vol. 5, pp. 945–960, June 1966.

6. Radler, R., and Chenot, D., "Synthesis of inorganic phototropic materials for high density computer memory applications," Tech. Doc. Rept. AL TDR 64-170, Oct. 1964.

7. Cohen, A. J., and Smith, H. L., "Variable transmission glasses sensitive to sunlight," *Science*, vol. 137, p. 981, Sept. 21, 1962; *Ceram. Abstr.*, p. 41i, Feb. 1963.

8. Swarts, E. L., and Pressau, J. P., "Phototropy of reduced silicate glasses containing the 570 mu color center," *J. Am. Ceram. Soc.*, vol. 48, pp. 333–338, July 1965.

9. Armistead, W. H., and Stookey, S. D., "Photochromic silicate glasses sensitized by silver halides," *Science*, vol. 144, pp. 15–154, Apr. 10, 1964; U.S. Patent 3 208 860, Sept. 8, 1965.

10. Smith, G. P., "Photochromic silver halide glasses," Paper 108, 7th Intern. Congr. Glass, Brussels, Belgium, July 1965. New York: Gordon and Breach, 1966.

11. Moser, F., Nail, N. R., and Urbach, F., "Optical absorption studies of the volume photolysis of large silver chloride crystals," *Phys. Chem. Solids*, vol. 9, pp. 217–234, 1959.

12. King, C. B., and Plummer, W. A., Internal Repts., Corning Glass Works.

13. Stookey, S. D., Internal Rept., Corning Glass Works.

14. Moon, P., "Proposed standard solar-radiation curves for engineering use," *J. Franklin Inst.*, vol. 23, pp. 583 617, Nov. 1940.

15. Condit, H. R., and Grim, F., "Spectral energy distribution of daylight," *J. Opt. Soc. Am.*, vol. 54, pp. 937 944, July 1964.

16. Smith, G., and Justice, B., "Photochromic glass: light and heat control with variable-transmittance glazing," presented at ASME 1964 Winter Meeting.

17. Justice, B., and Leibold, F. B., Jr., "Photochromic glass—a new tool for the display system designer," *Inform. Display*, pp. 23 28, Nov.–Dec. 1965.

18. "Photochromism," *Mod. Plastics*, Aug. 1965.

19. "BESTLITE, photochromic glass for ophthalmic lenses," Corning Glass Works Bull. OPD1, Aug. 1965.

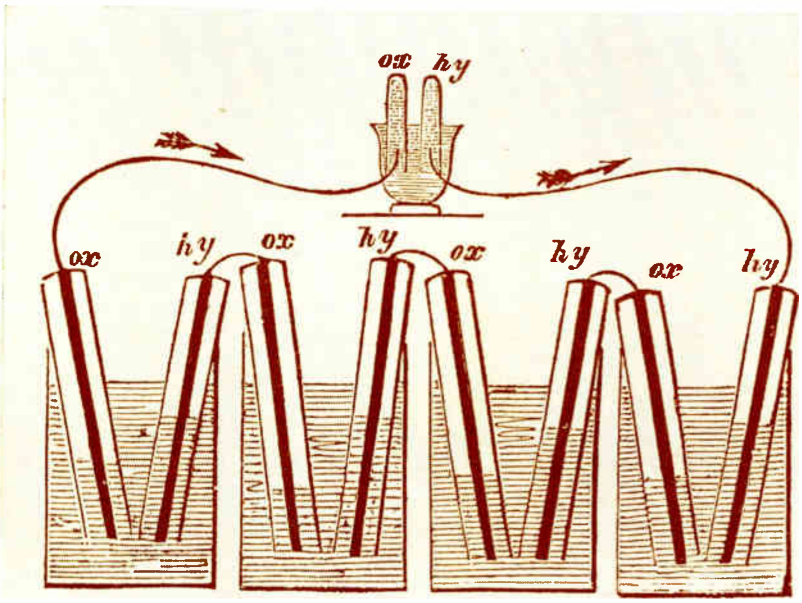
20. Britten, A. J., "Eye-protective devices," *Ordnance*, Nov.–Dec. 1964.

21. Kropp, J. L., Windsor, M. W., Brake, J. M., and Moore, R. S., "Research on photochromic dye-enzyme systems for flash-blindness protection," Air Force Materials Lab. Tech. Rept AFML-TR-65-423, Mar. 1966.

Fuel cells and fuel batteries—an engineering view

Although many formidable obstacles must be overcome before fuel cells can be economically competitive with established commercial energy sources, these devices are finding increasing use in special applications in which convenience is the primary criterion

H. A. Liebhafsky *General Electric Company*



(Left). The first fuel-battery system (1842). Several years before, Grove had the dream of "effecting the decomposition of water by means of its composition"—in our language, of using a hydrogen-oxygen fuel battery (four cells of which are shown connected in series) as the power source for an electrolysis cell. Electrolyte was sulfuric acid and electrodes were platinum.

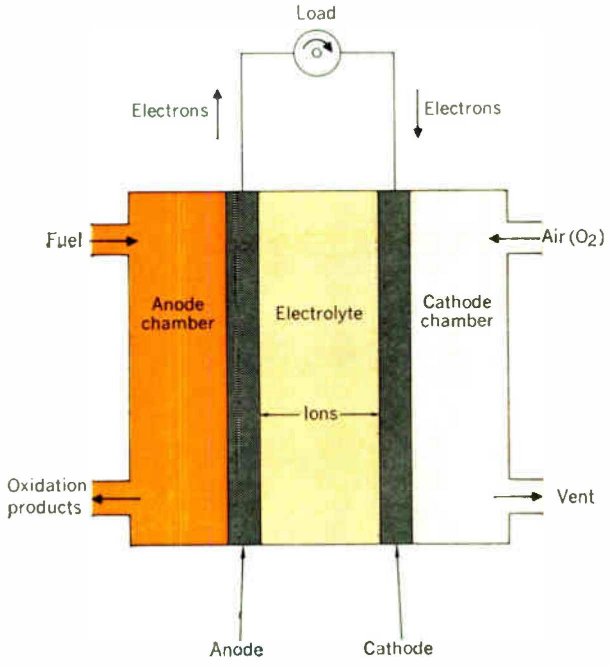
Fig. 1 (right). The tendency of fuels to give up electrons and oxygen to capture them leads to an electron transfer from fuel to oxygen during combustion. In the fuel cell the process proceeds at two electrodes in a more orderly way. Electrons given up by fuel at anode flow through external circuit, where they can do work, to be captured by ion flow through the electrolyte, which is virtually impervious to electrons

If fuel cells and fuel batteries attain their expected usefulness in the future, electrical and electronics engineers may some day be joined by a new kind of engineer—the electrochemical engineer. The fuel-cell problems this new engineer will have to face include those involving various types of fuels and their properties, efficiency, reliability, life, and operating temperatures. All of these properties are tied in with the important consideration of economics.

In any definition of fuel cells and fuel batteries, *fuel* in its classic sense should be the key word. Fuel cells should react *conventional* fuels (by which we mean the fossil fuels and substances readily derived therefrom) electrochemically with oxygen, preferably from air. A fuel cell, therefore, is an electrochemical cell in which energy from such a reaction is converted directly and usefully into low-voltage dc energy. Fuel cells electrically connected (in series, parallel, or series-parallel) make fuel batteries. The most common type of fuel cell is shown in Fig. 1.

These definitions are more important than they seem. Though restrictive, to make the subject manageable, they still include devices of great diversity. *Fuel* as used here excludes important substances often included, such as atomic "fuels" (e.g., uranium) and metals such as zinc or sodium. (The National Electrical Manufacturers Association[1] is much less restrictive. By their definitions, an electrochemical cell in which cesium and fluorine combine continuously would have to be called a fuel cell.)

The words *directly* and *usefully* imply that the device has an anode at which fuel is oxidized and a cathode at which oxygen is reduced, and that the conversion proceeds at voltages not greatly below the maximum possible and at reasonably high current densities. *Low voltage* and *direct current* are important to the electrical engineer,



who knows of course that electric energy of this kind, though different from that ordinarily generated and transmitted, is of greatest importance to the electrochemical industry.

The reaction between conventional fuels and oxygen liberates only enough energy to give us about one volt per cell under ideal conditions. As Fig. 1 shows, electrochemical reactions normally generate direct current. Schemes have been proposed to produce what has been loosely called alternating current from fuel cells, but the electrical engineer need not concern himself with such cells for some time to come.

The conventional fuels

The important conventional fuels may be listed, in order of decreasing reactivity, as hydrogen (in a class by itself), compromise fuels, and hydrocarbons.

Hydrogen belongs by itself because it is simple and highly reactive, the first characteristic probably being responsible for the second. When hydrogen reacts at an anode, it loses only one electron per atom and forms simple products. This probably explains why hydrogen can give us high current densities (mA/cm^2 of geometric electrode surface) with minimum loss of voltage from the theoretical. Current density and rate of electrochemical reaction are proportional.

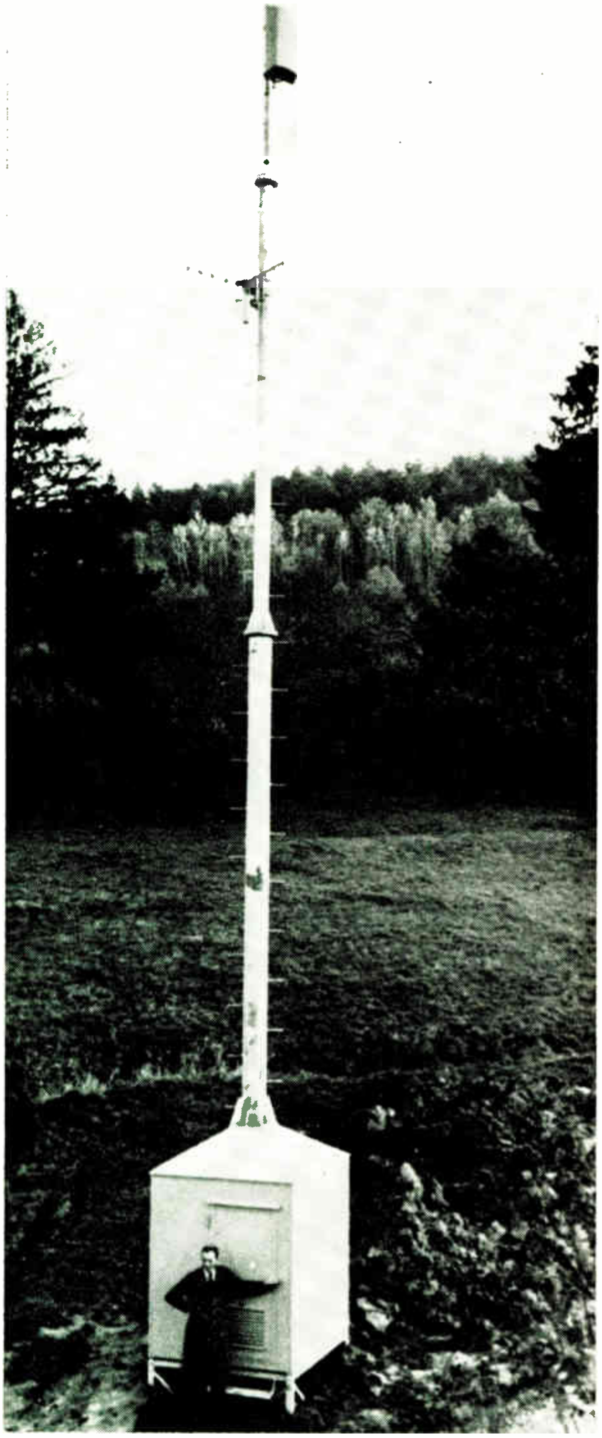
Hydrogen has always had a dominant position in the fuel-cell field (see title illustration) and hydrogen fuel cells and fuel batteries will be emphasized in this article. Hydrogen has serious disadvantages, among which only high cost and difficulties in handling and storage need be mentioned here. Because of these disadvantages, we must look to other fuels for the future.

The hydrocarbons, especially the liquid hydrocarbons, are among the most important and desirable of all fuels.[2] Unfortunately they are low in anodic reactivity, and their reactions are complex and can lead to many products. They are strong where hydrogen is weak, and weak where hydrogen is strong. The direct hydrocarbon fuel cell is a most difficult research assignment, but its successful accomplishment entails rewards that would outweigh the difficulties.

As their name implies, compromise fuels are of reasonable reactivity, cost, availability, and energy content, and are not too difficult to handle or store. Methyl alcohol and ammonia are prime examples. Hydrazine would be suitable for specialized applications were its price to drop tenfold or more. The compromise fuels are likely to be the earliest successors to hydrogen in direct fuel batteries; hydrazine qualifies now for special military applications in which fuel cost is unimportant and the toxicity of hydrazine can be tolerated.

So far we have not mentioned the commonest fossil fuel, coal. At the beginning of the century, scientists and engineers began to wonder whether the dream "electricity direct from coal" could be realized, whereupon the fuel cell, which had been almost dormant since the work of Sir William Grove, suddenly became popular. In 1900, the overall efficiency of steam plants was only about 10 percent. At this efficiency, they would have offered much less serious competition to a fuel-battery central station than today, with an efficiency four times as great. In

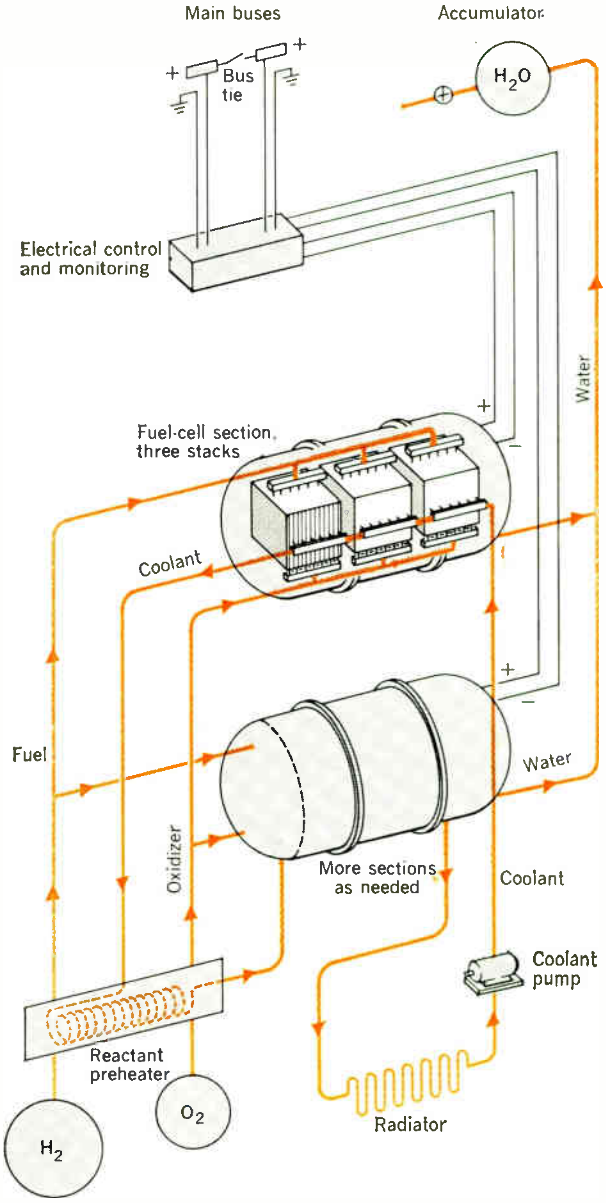
Fig. 2. Fuel-storage-battery-powered television relay station, Sudwestfunk Baden-Baden, Germany.[3] A similar station, 2300 meters away, is the power source for television for the Zermatt Valley, Switzerland. Both batteries were installed toward the end of 1965 and operated satisfactorily over the winter. The battery is rated at 24 watts, and 28 ± 2 volts is maintained by a special dc/dc converter. The fuel is methanol dissolved in strong potassium hydroxide for ordinary service. For service at very low temperatures (to −25°C), a formate is added to the solution. The fuel-electrolyte solution is consumed during operation and must be replaced after 5000 to 6000 hours. The long period between replacements necessitates a large battery. (Photo courtesy Brown, Boveri and Co., Baden, Switzerland)



1896, W. W. Jacques developed a carbon/air battery that delivered 16 amperes at 90 volts. The electrolyte was molten potassium hydroxide (costly), which was changed to potassium carbonate (much cheaper) as the battery operated. This alone makes the battery expensive to operate. In addition, its efficiency—erroneously placed at 82 percent of the theoretical—was grossly overestimated. Moreover, the inventor did not come to grips with the difficulties that would have arisen from impurities (ash, sulfur) had he used coal instead of a much purer carbon.

Inert fuels, such as hydrocarbons, can be used today, but only indirectly; that is, they must be changed to substances, mainly hydrogen, that are more reactive at fuel-cell anodes. Examples of such changes are the reaction of carbonaceous fuels with steam, which is being

Fig. 3. Fuel-battery system for Project Gemini. Water transport to the accumulator is accomplished without moving parts by means of wicks and a pressure differential across a porous member. (Courtesy Direct Energy Conversion Operation, General Electric Co., Lynn, Mass.)



widely investigated, and the decomposition ("cracking") of ammonia, which will be used to provide hydrogen for a fuel-battery-powered submarine in Sweden. Indirect systems thus combine a chemical plant with a fuel battery, and the combination brings problems not present with the fuel battery alone. Ideally, the chemical process should be carried out in the anode chamber to facilitate heat and mass transfer. Indirect systems will be of great interim value.

Oxygen or air?

For applications, such as space missions, in which the nitrogen of the air cannot be tolerated, oxygen must be used at the fuel-cell cathode. For terrestrial applications, in which oxygen is too expensive or cannot be carried because of weight or volume restrictions, ambient air must be used. But the use of air has important drawbacks that concern the engineer.

Most fuel-cell electrodes are highly porous and therefore their true surface is many times the geometric; this is one road to high geometric current density, since this current density increases with true surface area. At high current densities, cathode pores can become filled with nitrogen, thus creating a mass-distribution barrier for oxygen and injuring cell performance. One remedy is to make the cathodes very thin (0.25 mm or so thick) and the pores large, but this practice introduces problems of its own. Especially in a fuel battery, where passages must be narrow to conserve space, forced convection of the air will usually be needed for acceptable current densities (say, 100 mA/cm^2). As nitrogen leaves a battery containing an aqueous electrolyte, this gas may carry with it enough water vapor to interfere with cell operation. Particularly at high current densities, the carbon dioxide (about 0.03 percent) present in the air may give trouble with alkaline electrolytes either by precipitating solids in the electrodes or by reacting with the bulk electrolyte; therefore, scrubbing the air to remove carbon dioxide or frequent changes of electrolyte may be necessary. Clearly, the expression "free as air" needs qualification when we are referring to the fuel battery.

The problems of air operation are important also because various air batteries that are not fuel batteries (for example, zinc/air batteries) might be attractive for applications, such as vehicular, in which high current densities are needed. One desirable by-product of fuel-cell research is an air cathode that can serve other power sources as well.

Fuel batteries vs. storage batteries

Storage batteries do not use conventional fuels. Instead, they contain the chemical energy they convert, and hence must be recharged when this energy is depleted. Ideally, the fuel battery can be an invariant converter that delivers energy so long as fuel and oxygen are supplied.

These two kinds of power sources are complementary more often than they are competitive. Storage batteries are favored for high power over short times (starting an automobile or short space missions); fuel batteries are favored when the load profile calls for moderate power over longer times (space missions longer than several days). The trade-offs that must be made usually are not simple, and they must be made on the basis of the complete energy system for the load profile in question. In the case of the fuel battery, for example, one must consider energy source plus fuel plus oxygen plus peripheral equipment, with proper debits or credits for the reaction products. To handle high peak loads, storage batteries may be used and kept charged by fuel batteries in continuous operation.

Metal/air batteries, such as the zinc/air battery mentioned previously, are hybrid devices; with respect to the anodes, they are storage batteries; with respect to the cathode, they are fuel batteries. A different hybrid device is the fuel-storage battery of Fig. 2, in which the fuel (methyl alcohol) is stored in the electrolyte (potassium hydroxide), which changes to carbonate as the battery operates. The solution must be replaced when the fuel is exhausted, and the cost of potassium hydroxide, unfortunately not negligible, enhances the energy cost. The cathode operates on air.

Why do we want fuel batteries?

Fuel batteries are considered attractive because they are convenient and because they promise eventually to be low-cost sources of electric energy. Cost must be judged relative to convenience: because of the convenience it offers, a fuel battery may prove successful in an application (for example, a space mission) even though the cost of the energy it produces is prohibitive by central-station standards. Theconcept of "convenience" embraces such qualities as

1. High power rating for unit volume
2. High power rating for unit weight
3. Quietness
4. Cleanliness
5. Operation on air
6. Continuous unattended operation over long periods
7. Production of useful water

In assessing the first two qualities, the complete system (see Fig. 3) must be considered. In reference to quality 4, the ultimate is a hydrogen/oxygen battery in dead-ended operation; in other cases, complete oxidation of the fuel is the most desirable way of achieving a harmless battery exhaust. Quality 6 involves reliability, maintenance, and life. Quality 7 is important particularly on space missions.

Factors contributing to low-cost electric energy include

1. High efficiency
2. Low-cost fuel and oxygen
3. Low maintenance cost
4. Long life
5. Low capital investment

Not all of the factors determining cost have been listed. Research and development costs have been omitted because they are incapable of general assessment; for example, terrestrial fuel batteries benefit from the knowledge gained in developing fuel batteries for space missions. Research and development costs are both high in an absolute sense, with development costs very high relative to research costs. Adequate life tests are expensive.

Efficiency

The immunity of the fuel cell to the Carnot-cycle restriction (see Fig. 1) was for a long time its greatest attraction. In a modern central station, the Carnot-cycle efficiency could be near 65 percent, and the overall efficiency near 40 percent. The overall efficiency of smaller energy sources that the fuel cell hopes to displace is considerably less than 40 percent. Statements by reputable authorities often mention efficiencies greater than 65 percent for the fuel cell, and the popular press is sometimes even more optimistic.

There are efficiencies of various kinds. We shall proceed conservatively, and define an overall efficiency (called the comparative thermal efficiency) for the fuel battery and for the fuel-battery system. These efficiencies are comparable with the 40 percent just mentioned for central stations. The two definitions are analogous.

For the battery (or the system):

$$\text{Eff}_{\text{CT}} = \frac{\text{Net useful work}}{\Delta H \text{ of fuel consumed}}$$

In the denominator, ΔH is the *higher* heat of combustion of the fuel actually consumed in making available the net useful work in the numerator; some of this fuel may be consumed by peripheral equipment. In both cases, the electric energy required by the peripheral equipment (such as pumps) with its demand for parasitic power must be subtracted from the gross electrical work

$$\int Ei\,dt$$

available at the fuel-battery terminals; the system efficiency consequently may be considerably lower than the battery efficiency.

The efficiency of a single fuel cell will usually exceed considerably the two efficiencies given previously. Detailed discussion would take us too far afield. We shall simply say that under most conditions this cell efficiency is determined principally by the *voltage efficiency* under operating conditions. This efficiency is E/E_{max}, where E is the actual cell voltage and E_{max} is the maximum value of E, which can be calculated from thermodynamic data and could be realized only under completely reversible conditions.

The most important single characteristic of a fuel cell is its current-density–voltage curve (Fig. 4), which is an index of cell performance and, therefore, corresponds to an upper limit for the performance of battery and of system. The current *density* (not current) is chosen as abscissa, not only because current density is proportional to the rate of electrochemical reaction, but also because it helps determine watts per square centimeter, a ratio that helps establish the size and weight of a power source of given rating.

From the point of view of efficiency, the vital feature of current-density–voltage curves is that cell voltage always decreases with increasing current density throughout the useful operating range. To realize maximum efficiencies, the cell would have to be operated at current densities too low for doing finite work: microamperes from a large power source are seldom useful.

Overall efficiencies are often thought of primarily in their relation to fuel cost. We hope the time will soon come when such thinking is justified for fuel batteries. In this early stage of their development, however, overall efficiencies are important primarily because they determine unit capital cost (dollars per kilowatt) and in special applications (space missions, portable power sources) because these efficiencies fix the weight and volume of reactants that must be carried for doing a given amount of work.

Reliability and working life

The crucial questions of reliability and expected life cannot be answered firmly until there has been much more experience with fuel batteries. The answers will differ with the type of battery and with the duty cycle for a given type. "Reliability" and "life" are concepts difficult of exact or general definition. In space applications, where the fuel batteries are isolated and cannot be attended, life may be taken as synonymous with mean time to failure, including failure of peripheral equipment. In Project Gemini, it will be remembered, all the difficulties to the time of writing were chargeable to the peripheral equipment—none to the fuel cells themselves. In terrestrial applications, where opportunities exist for adjustment, repair, and replacement, a battery or a system will have a useful life far exceeding mean time to failure under the drastic conditions in space. Reliability and maintenance costs therefore cannot yet be assessed.

The life of single cells under steady load in the laboratory is thousands of hours: *uniformity* is the key to long life. When cells are assembled to make batteries, uniformity is more difficult to achieve (see below), with the result that the life of a single cell may be shortened below what it would have been were it operated alone. Further, when cells are connected in series, and the life of an entire stack is that of the cell which is the weakest link, statistical considerations lead to a stack life reduced considerably below the average life of a single cell operated alone. For terrestrial applications, it should be possible to choose conditions so that the life of the battery limits the life of the system.

This analysis is not meant to be discouraging. If individual cells show long life, as they do, electrochemical engineers should be able to design and develop batteries and systems of adequate life.

Unit capital costs

It is impossible to translate unit capital costs (dollars per kilowatt) into energy costs so long as life is unknown. What unit capital cost is reasonable depends upon the premium that the convenience of the fuel battery can command. In space missions for which the weight of other power sources is prohibitive, that premium is high. The premium is at a minimum in the usual large central stations. For a given terrestrial application, such as power sources for communication equipment, the premium is likely to be much higher for fuel batteries in military (as opposed to commercial) use.

A simple calculation will show the importance of unit capital costs in commercial applications. Fuel batteries are often suggested for utilizing waste hydrogen. With dc electric energy at one cent per kilowatthour, and with hydrogen and air at no cost, a hydrogen/air battery, at $300 per kilowatt installed and operating continuously and requiring no service, would produce just about enough electricity to recover the capital investment in three years. There are no fuel batteries now on sale at anywhere near $300 per kilowatt that will operate for three years under the conditions stated.

Tentative estimates of tolerable unit capital costs for fuel batteries intended for commercial use will be given. These are opinions not based upon detailed information. For small (10- to 100-watt) power sources, this cost is more than $1000 per kilowatt; such power sources will serve best where they can benefit from transistorized circuitry. For central stations, this cost is $100 per kilowatt. For first use in electric vehicles, it is $200 per kilowatt; for passenger automobiles, the ultimate dream, very much less. Building a reliable battery of adequate life for, say, $50 per kilowatt will not be easy, no matter what the fuel.

In the early stage in the development of fuel batteries, considerations of unit capital cost warrant the prediction that the terrestrial use of these devices will occur first in small sizes and in military applications.

Operating temperatures

The properties of the electrolyte are perhaps the most important determinant of fuel-cell operating temperatures. Of these properties, we shall mention only electrical conductivity. One function of the electrolyte is to complete the electric circuit (see Fig. 1) by the transport of ions, and it is desirable to keep the resulting I^2R losses low by close spacing of the electrodes and by choosing a temperature at which there is adequate conductivity. The following are typical examples (temperature ranges approximate):

Ion exchange membranes now available, below 100°C
Aqueous acid electrolytes, up to 200°C
Aqueous alkaline electrolytes, up to 300°C
Molten carbonate electrolytes, 500–600°C
Doped zirconia (solid) electrolytes, 900–1200°C

Where does research stand?

It is convenient, though imprecise, to say that the fuel cell belongs to research, and that the steps from cell to battery and from battery to system are engineering assignments.

Although research is never finished, one can say that enough is known about hydrogen/oxygen and hydrogen/air cells to make the designing and building of good batteries feasible.

Most research problems relating to energy conversion can be formulated as materials problems because the drive for high performance strains materials to their limits. We shall not concern ourselves with the usual types of materials problems, which arise in connection with sealing, corrosion, aging, decomposition, or evaporation.

Electrocatalysis is the main research problem with fuels other than hydrogen. For present purposes, we may (imprecisely) regard electrocatalysis as the process that raises *IR*-free performance curves like those in Fig. 4—that is, the process by which electrode reactions at constant temperature, pressure, and electrolyte are accelerated to give a higher current density at a given cell voltage. A good electrocatalyst must be inert toward the electrolyte, have large specific surface and active morphology, be or resemble a transition metal (see the periodic table of the elements), and (if necessary) double as a catalyst for chemical reactions that accompany the electrochemical reactions. Platinum is the best single electrocatalyst for fuel-cell electrode reactions as a group, although it is not the best for every reaction. But platinum is costly, needed for other purposes, and limited in supply. Science has not yet given us an understanding of platinum's unique position in electrocatalysis, and therefore we have no firm theoretical guidelines for attacking the electrocatalysis problem.

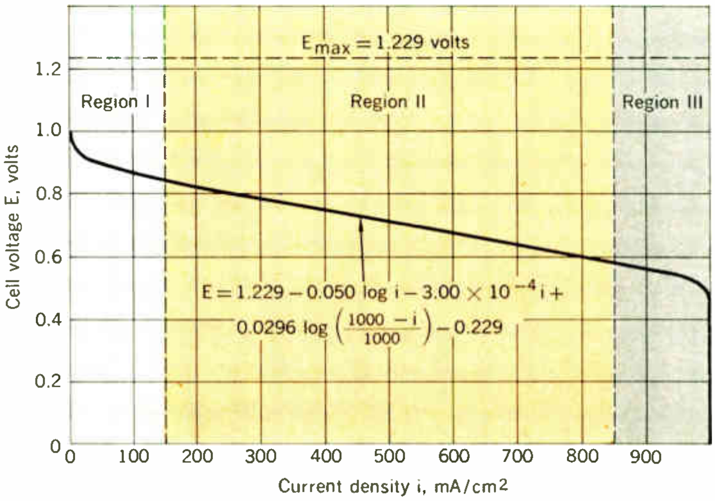


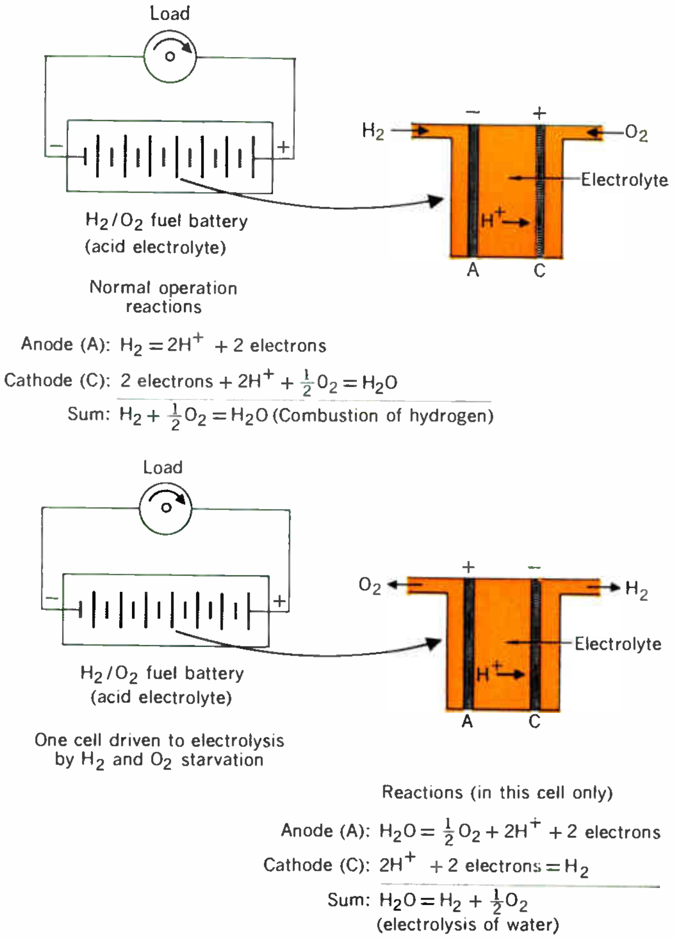
Fig. 4. Idealized performance curve for a fuel cell. The value $E_{max} = 1.229$ volts at 25°C is the maximum permitted by thermodynamics for an H_2/O_2 cell at standard conditions. In Region I, loss of voltage occurs principally at the electrodes. In Region II, this loss is increased by the internal resistance of the cell. In Region III, perpendicular increase in voltage results from a limitation in mass transport. The equation for the curve is explained in a forthcoming book, "Fuel Cells and Fuel Batteries," by H. A. Liebhafsky and E. J. Cairns (New York: Wiley).

The rates of chemical reactions increase with temperature. Although electrochemical reactions have complexities that enter into the temperature dependence of their rates, one is justified in assuming that higher temperatures bring higher rates, and that the electrocatalysis problems should be less serious at higher temperatures. This advantage will be at least partially offset by the increasing seriousness of the several types of materials problems, as stated previously. To illustrate, an oxide-ion electrolyte resembling doped zirconia, but of greater conductivity, would permit reduced operating temperatures for cells with these solid electrolytes and make them more attractive.

What of engineering?

The importance of uniformity in a battery was mentioned earlier. Only if conditions are uniform in a battery can the performance of the battery approach that realized for individual cells on a laboratory bench. The attainment of this uniformity is an engineering assignment because it depends upon the control of transport processes. A fuel battery consumes reactants and generates products and heat and electricity. The processes that transport mass, momentum, heat, and electricity must proceed at rates that maintain conditions uniform within the battery. Nonuniformity can result in many ways and have many undesirable consequences, one of the more serious of which is illustrated in Fig. 5.

In addition to ensuring uniformity in the battery, the engineers must also choose construction materials, regulate the electrical output of the battery, and make the step from battery to system. These engineering assignments have proved to be more formidable than many had anticipated, and the engineer today carries the principal burden in making hydrogen batteries successful.



Anode (A): $H_2 = 2H^+ + 2$ electrons

Cathode (C): 2 electrons $+ 2H^+ + \frac{1}{2}O_2 = H_2O$

Sum: $H_2 + \frac{1}{2}O_2 = H_2O$ (Combustion of hydrogen)

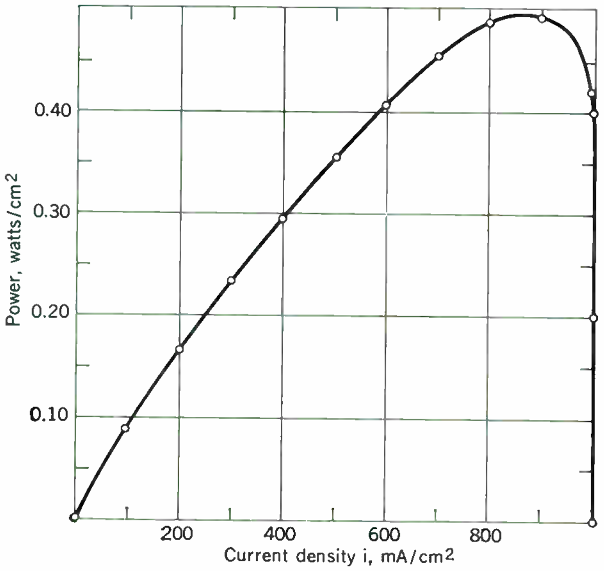
Anode (A): $H_2O = \frac{1}{2}O_2 + 2H^+ + 2$ electrons

Cathode (C): $2H^+ + 2$ electrons $= H_2$

Sum: $H_2O = H_2 + \frac{1}{2}O_2$
(electrolysis of water)

Fig. 5. Interrupting supply of oxygen and hydrogen to one cell in a series-connected battery causes the other cells to "drive" the afflicted cell. Undesired electrode reactions in the afflicted cell result.

Fig. 6. Idealized curve based on Fig. 4, showing how power generated by a fuel cell varies with current density. Comparison with Fig. 4 shows that the cell voltage at the current density for maximum power has fallen to about 0.6, which means reduced efficiency.



Electrical problems of the fuel battery

The electrical problems of the fuel battery are inherent in the performance curve of the fuel cell (see Fig. 4). Two favorable features stand out: (1) At open circuit, voltage is maintained without measurable consumption of fuel, there being no net electrochemical reaction at zero current density. (2) Voltage efficiency, and hence overall efficiency in the usual case, is higher the lower the current density. These features make the direct fuel battery desirable for equipment that must stand ready to perform during long idling periods, or that operates most of the time at low load. These advantages may be reduced in an indirect fuel-battery system owing to the energy required to keep converter or re-former ready for operation when load increases.

The low voltage of the single fuel cell leads to electrical problems, which are generally less serious with hydrogen as fuel because it gives higher cell voltages, at the same current density, than do most others. Hydrogen might yield an E of 0.7 volt at a current density for which hydrocarbons give an E of 0.3 volt. The obvious way to obtain needed high voltages from fuel cells is to connect them electrically in series.

As mentioned earlier, the greater the number of cells in series, the greater the chance one cell in the stack will fail; this will most often be a failure of the *least* reliable cell. The result could be simply an open-circuited stack or it could be something more serious. If the failure resulted from an interruption of the hydrogen or oxygen supply, the other cells in the stack could "drive" the one affected and cause unwanted reactions to occur at the electrodes. This is the serious lack of uniformity mentioned previously. As Fig. 5 shows, this type of failure could lead to the generation of oxygen in the hydrogen (anode) chamber and to the generation of hydrogen in the oxygen (cathode) chamber.

For cells connected in parallel, complete failure will usually not occur until the *most* reliable cell has failed, although there will have been a decrease in current prior to complete failure. From the standpoint of reliability, it is desirable to minimize series connections and maximize parallel connections.

There is a limit to how far one can go. Maximizing parallel connections implies the handling of large currents and the incurring of high I^2R losses, and there is the added difficulty that most electric equipment operates at voltages considerably above that of a single cell. Solid-state dc/dc converters are currently available at ratings from 20 watts to a few kilowatts, but they are inefficient at low input voltages. They are nevertheless valuable because they make it possible to reduce the number of cells connected in series, the extent of maximum reduction being set by the conversion inefficiency considered tolerable and by the probability of failure of a cell in the stack.

For small loads, dc/ac inversion can also be accomplished, but only with heavier and more costly equipment than dc/dc conversion requires. At present, we do not believe that inversion of fuel-battery power on a central-station scale need be considered; if such power can compete on this scale at all, it will have to compete for dc applications, notably in the electrochemical industry. The industry provides a large market: perhaps 5 percent of the 200 GW total U.S. generating capacity serves this market, about half of which produces aluminum.

The performance curve in Fig. 4 also permits conclusions about operation at various power levels. As Fig. 6 makes clear, operation at maximum power density is possible only at reduced efficiency, and this reduction becomes prohibitive at current densities above that for maximum power density.

The fuel battery as a chemical plant

In space, the water generated by H_2/O_2 batteries will be drunk or used in other ways. The fuel battery will then be not only a dc generating plant but a chemical factory as well. Is this appealing concept likely to prove widely useful on earth? We think not.

It is true that many important chemicals are produced by oxidation, and that such oxidation can often be done advantageously at an anode. Although we do not exclude the possibility that electricity may be a useful by-product in special cases, such as the oxidation of sodium amalgam in the preparation of caustic, we do not think the combination of chemical factory and fuel cell will prove generally useful for these reasons[4]:

1. The amount of electric energy produced annually by the power industry is so large that the by-product electricity we are considering will appear very small beside it. For example, a rough estimate shows that in the United States the electric energy produced in one month (5.6×10^9 kWh) is equivalent to that of all the sulfuric acid—29 million tonnes (32 million tons)—made in two years. (SO_2 is assumed as the starting material.) Sulfuric acid was chosen because it is a high-tonnage chemical, not because it is adapted to manufacture in a fuel cell. It follows that any chemical made in amounts below about 1 million tonnes annually could not produce by-product electricity in significant amounts.

2. The value of such by-product electricity is low relative to that of the chemicals produced. This is true even in the case of sulfuric acid: less than 1 cent for a kilowatthour that is equivalent to 5½ kg of acid, worth about 12 cents. This twelvefold ratio will be much greater with most other chemicals.

3. An electrochemical device must usually meet different requirements for the optimum generation of electricity and for the optimum production of a chemical. Conditions for the latter process can be more closely controlled if a voltage is imposed on the cell—that is, if electricity is consumed instead of generated. An improved yield or a chemical of better quality should usually justify this approach.

Fuel batteries for energy storage

Fuel batteries are feasible for energy storage in space applications, provided that solar-energy converters are available. The scheme here is to convert an excess of solar energy into electric energy during the orbital day, use this excess to electrolyze a working substance (for example H_2O), and recombine the products of electrolysis (H_2 and O_2) in a fuel battery to produce electric energy during the orbital night. The electrolyzer and fuel battery here constitute a regenerative system; the two may be the same device. Such energy storage sounds attractive, but there are problems with both the solar converter and the electrochemical system.

A recent article on pumped storage[5] shows that this method of storing energy on a large scale is so economical as to make competition by electrochemical regenerative systems appear hopeless. The efficiency of these systems, being the product of the efficiencies of fuel battery and electrolyzer, is much lower than that of fuel battery alone.

The present outlook for fuel batteries

Anyone called upon to predict the outlook for fuel batteries is entitled to quote Mr. Justice Holmes[6]: "Every year if not every day we have to wager our salvation upon some prophecy based upon imperfect knowledge."

The following prediction[7] was made before 1960: "The current increase in fuel cell activity, if maintained, makes it likely that fuel cells will serve as power sources in special applications within the next 5 years. Successful, practical model cells are already with us. The future of central-station fuel cells cannot be predicted today." Figures 2 and 7 show that the first sentence of this prediction was not rashly optimistic.

Next, the reader should examine a recent, authoritative, and more detailed prediction by Lord Rothschild,[8] speaking for Shell Research Ltd., where important fuel-cell work is being done. This is a conservative prediction, reconcilable with the statements outlined below.

The predictions that follow are made within these boundary conditions: (1) They are based on the published material we know. (2) They include applications, such as space and military, in which the fuel battery commands a premium for convenience. (3) They assume that air, when available, will be used at the cathode. Air is considered unavailable in space and under water. These predictions will not be documented, and only a few examples will be cited.

Space. The fuel battery has established itself for space missions (General Electric; Fig. 7). Future missions are scheduled to use fuel batteries, by Pratt and Whitney, based on the distinguished work begun over 30 years ago by F. T. Bacon.

Portable (carriable by one or two men). Successful application within three years seems certain, with power sources for military communication equipment in a preferred position.

Transportable (carriable by vehicle with batteries for propulsion excluded). Already installed by Brown, Boveri (Fig. 2). Successful military applications of other types at higher ratings seem likely.

Propulsion. Successful applications will come first on military vehicles. Golf carts with hydrazine batteries have been demonstrated by Allis-Chalmers. Military fork-lift trucks should operate on fuel batteries within ten years. The Swedish submarine effort was mentioned previously.

The passenger automobile seemingly affords the fuel battery a great opportunity, but unit capital cost is such a formidable hurdle now that other problems are scarcely worth discussing. For the present, effort should be concentrated on automobiles that use storage batteries,[9] perhaps new types not yet in use, which might be replaced or augmented by satisfactory fuel batteries. Opinion in Great Britain holds that the locomotive or the fuel-battery-powered railroad car is a more promising application than the passenger automobile. Hydrogen/oxygen batteries to operate all three could be built at a high price today; it will be remembered that Allis-Chalmers used such a battery to power a tractor in 1959.

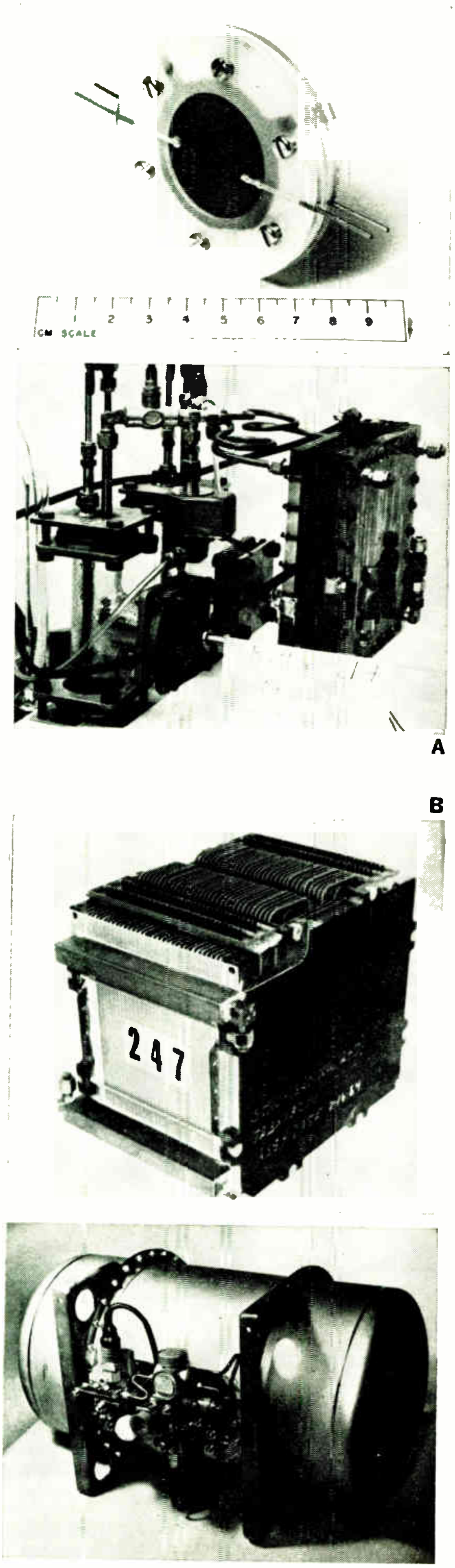


Fig. 7. Development of the ion exchange membrane fuel battery by the General Electric Company. A—Physical Chemistry Section, Schenectady, N.Y., 1955-1960; top: Grubb-Niedrach cell, 0.02 watt; bottom: Cairns-Douglas battery, 15 watts. B—Direct Energy Conversion Operation, Lynn, Mass., 1959-1965; top: DECO stack, 300 watts; bottom: DECO Gemini battery, 1000 watts.

The home. Steady progress (Broers, TNO, Holland; Institute for Gas Technology, Chicago) being made on methane/air batteries with molten-carbonate electrolytes leads one to expect experimental home installations exceeding 20 percent in comparative thermal efficiency within five years; such fuel batteries would be connected to banks of storage batteries as energy reserve for peak loads.

Central stations. The earlier prediction[7] stands, with a few added remarks. The central station ranks with the passenger automobile in difficulty as an application for the fuel battery. It differs from the automobile in that unit capital cost is a less serious hurdle here than overall energy cost. The future of the fuel battery in the large-scale generation of electricity seems linked to the future of natural gas. The growth of atomic energy installations and the effect this will have on the coal industry both enter the picture because one must look perhaps a decade ahead for the earliest time that a central-station fuel battery might begin to be used. But there is hope for the fuel battery in smaller central stations that serve a single community—stations in which the use of heat and of electric energy will be efficiently combined and distribution costs will be reduced.

The reader wishing to reconcile these predictions with reference 8 will note that we have stressed fuel batteries of low ratings, and have included space and military applications.

A logical position at present seems to be that fuel-cell research should continue so long as significant progress is made, and that the engineering development of H_2/O_2 and of H_2/air batteries for favorable applications should be emphasized. Fuel batteries will prove themselves indispensable in some applications and useful in many others.

The author gratefully acknowledges the assistance of Dr. E. J. Cairns in the preparation of this article.

REFERENCES

1. "Fuel cell definitions," Standards Pub. No. CV 1-1964, National Electrical Manufacturers Association, New York, N.Y.

2. Peattie, C. G., "Hydrocarbon-air fuel cell systems," *IEEE Spectrum*, vol. 3, pp. 69-76, June 1966.

3. Plust, H. G., Private communication, Brown, Boveri and Company, Mar. 21, 1966.

4. Liebhafsky, H. A., and Cairns, E. J., "The fuel cell and the power industry," Rept. 60-RL-2382C, General Electric Co., Mar. 1960.

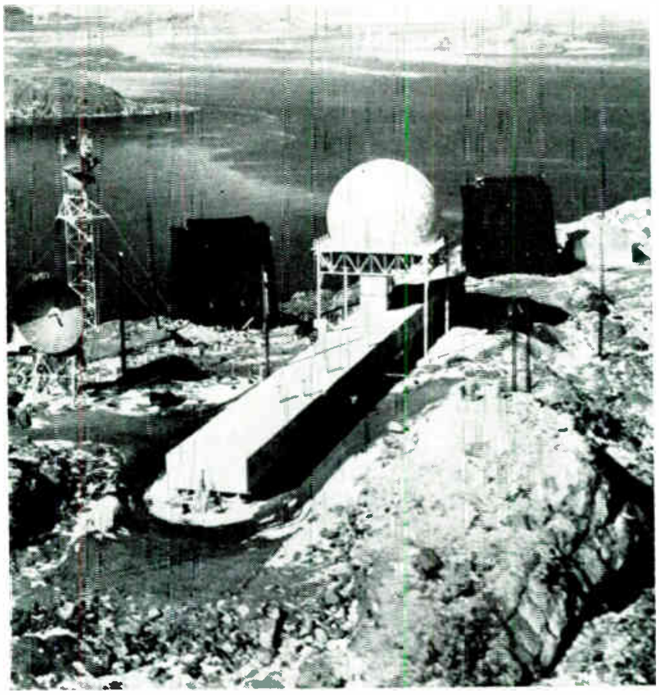
5. Friedlander, G. D., "Pumped storage—an answer to peaking power," *IEEE Spectrum*, vol. 1, pp. 58-75, Oct. 1964.

6. *Abrams v. U.S.*, 250 U.S. 616 624 (1919), Oliver Wendell Holmes dissenting.

7. Liebhafsky, H. A., and Douglas, D. L., "Fuel cells as electrochemical devices," *Ind. Eng. Chem.*, vol. 52, pp. 293-294, Apr. 1960.

8. Lord Rothschild, "Fuel cells," *Sci. J.*, vol. 1, p. 82, 1965.

9. See, for example, Maxwell Boyd, "Electric 60 mph 'mini' out soon," *London Sunday Times*, Feb. 27, 1966.



Ten years of DEWLine

The U.S. Air Force's DEWLine after ten years of operation has undergone a gradual major overhaul. It has become, as well as a surveillance system, a highly reliable communications network, and seems on its way to being a permanent fixture in the Arctic

Nilo Lindgren *Staff Writer*

Constructed and maintained under cruel Arctic conditions, the DEWLine was established principally to act as an early warning system against the approach of manned bombers; but with the development of ballistic missiles, with the construction of BMEWS further north, and with the increasing military and other communications traffic across it, the emphasis on DEWLine functions has been shifting. This article reviews some of the highlights of DEWLine's evolution and its development, after a three-year upgrading program, into a sturdy, high-capacity communications network.

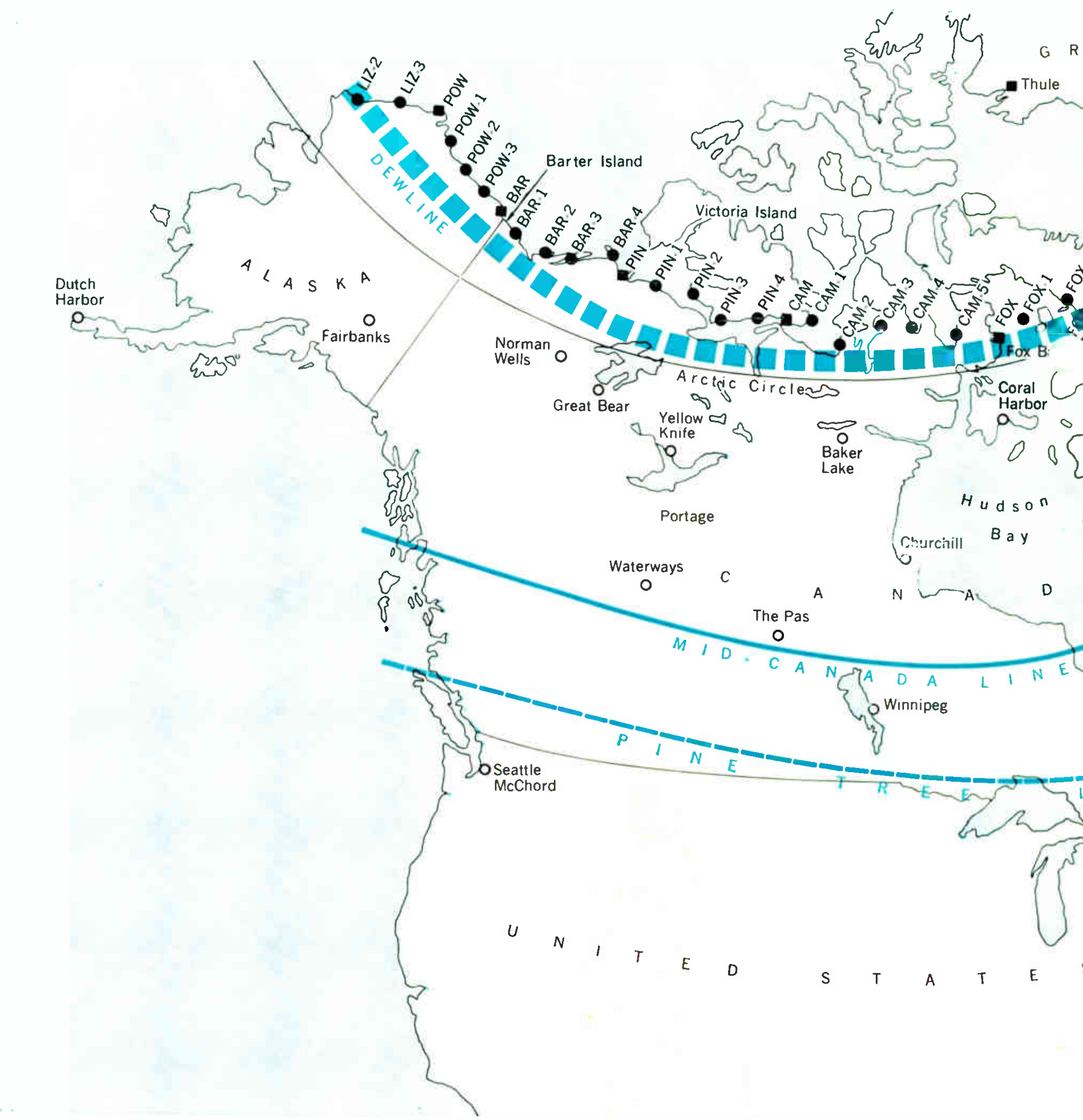
The DEWLine has long been in and out of the news. As an operating system, the DEWLine is ten years old, and today it is in the news again, not only because it has just passed its tenth anniversary, but because the system has been upgraded to take on an important new function, that of a highly reliable communications network.

Original DEWLine functions

For those who at this late date need a recapitulation, DEWLine stands for "Distant Early-Warning Line." The system consists of a series of early warning stations, located about 160 km apart, and stretching across the northerly region where explorers of an earlier day sought the Northwest Passage, a region extending from the Aleutian Islands in the west to the east coast of Greenland in the east. It also includes a larger water jump over to the west coast of Iceland. Along most of its length of 9600 km, it skirts north of the Arctic Circle, and in all comprises six main stations (DYE, FOX, CAM, PIN, BAR, and POW) (see Fig. 1) and 27 auxiliary stations. Several stations provide rearward (southward) communication links. The eastern extension consists of four early warning stations in Greenland and a support facility and terminal station at the head of the long Sondrestrom fjord. The most distinctive features of the main stations, aside from the rugged and desolate terrain of rock and snow in which most have been planted, are the huge clifflike tropo antennas, which stand in brooding pairs, and the golf ball radomes housing the surveillance radars (see title illustration).

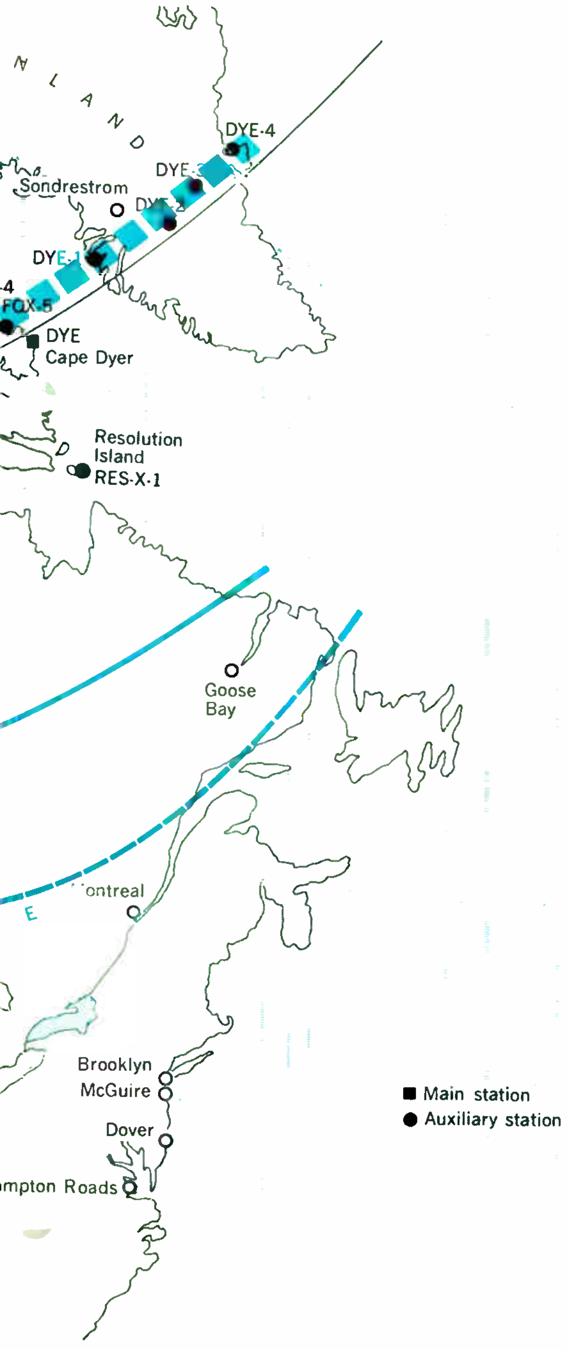
The principal original function of the DEWLine system was, of course, the detection and reporting of airborne objects intruding upon or operating within the Distant Early Warning Identification Zone. Through its system of lateral communications along the length of the line, and in many places through the overlapping of its radar coverage in the same volume of space, the line forms in effect a big three-dimensional electronic fence looking northward (they "forward tell it") out to 250 km (with the more powerful radars out to 320 km), and from grazing up to 20 000 meters across the entire line. Small radar gaps over the water at the eastern end of the line are covered by planes equipped with early warning radar, which are flown on a secret random schedule. Anything that passes through this electronic fence will be detected,

Fig. 1. The DEWLine, the Mid-Canada Line, and the Pine Tree Line originally formed three tiers of a complex electronic surveillance and warning system.



monitored, and eventually identified as friendly or hostile. Major rearward communication links were set up down the east and west coasts of the North American continent and down through central Canada, and in some cases linked up with Canadian commercial communication outfits that were steadily pressing their systems northward. DEWLine engineers have taken advantage of these existing commercial systems wherever they could.

In the event of an intrusion in the DEWLine airspace, the typical scenario of events reads briefly as follows: If an unidentified object is detected on the radarscope of a DEWLine station (this part of the surveillance has from the beginning been by eye—there are no automatic systems), the radar technicians, who are assigned on an around-the-clock watch, report their findings immediately to the data center of the sector's main station. There, U.S.A.F. and R.C.A.F. personnel register the course of the object and alert NORAD (North American Air Defense Command) headquarters at Colorado Springs in the United States, which in turn alerts SAC headquarters. If the airborne objects penetrate further southward and are positively identified as hostile, they come under the surveillance of the Pinetree Line (see Fig. 1). During this time, their potential target areas are analyzed, and the heads of the Canadian and U.S. Governments are alerted, as are many defense agencies. In theory, then, should a hostile intrusion proceed this far, the populations would have taken cover, the retaliatory jet bombers would be on their way north, and the intercept jet fighters and missiles would be beginning their business. It is a grim scenario, and one which luckily we have not seen acted out on the real world stage.



Theoretically, the DEW system was to give the populations to the south a two-hour warning of the approach of manned bombers. But barely was the system operating at full effectiveness than it was threatened with obsolescence by the development of long-range ballistic missiles with the capability of overpassing or even bypassing the DEW system entirely.

Nevertheless, as long as even the dim threat of a manned bomber attack lingers, the DEWLine serves its original surveillance function. Furthermore, from the beginning the DEW system had served, as well as surveillance, a number of subsidiary functions. Its communications and electronics activities included long-haul communications, navigational aids, and short-haul communications and support systems. In this respect, with the passage of time, the top of the world had become more busy rather than less, and the various subscribers to the DEW communications facilities were queueing up for services. Work on the BMEWS (Ballistic Missile Early Warning System) further north began in 1961, and the DEW system lay waiting as a natural backup. By 1963, the pressure had grown on the DEW system to serve as a data link for BMEWS traffic. However, the problem was that the DEWLine had not been originally engineered for the reliability called for by BMEWS. So, once again, the DEW system entered a new stage of evolution.

Conception and evolution

The evolution of the DEWLine is, as much as anything, the story of how a series of unique engineering problems were resolved and systems operated under the most arduous conditions imaginable.

DEWLine was conceived in 1952 by a group of scientists called the Summer Study Group at the M.I.T. Lincoln Laboratory. Within a year, the first prototype of a DEWLine station was installed in Streator, Ill., and by 1954, the first trial sites had been set up near Barter Island ("Kaktovik" in Eskimo), Alaska. Barter Island, in the twenties, was the site of extensive fur trading between the white man and the Eskimo, a trade that had gradually diminished. The Eskimos of the fifties, perhaps to the despair of some anthropologists, have been employed both on the construction and the maintenance of the DEW system facilities.

Between 1954 and 1957, for 2½ years, upwards of 23 000 U.S. and Canadian engineers and construction workers labored to survey and lay out the station sites, traversing more than a 1½ million km in the process. They constructed the buildings and installed and checked out the radar and communication equipments under the cruelest of environmental conditions. The saga has been well documented.

Because of the haste with which the line was installed, there was no time to go through the usual stages of engineering design and development, so that often equipment was constructed directly from breadboards. This led to many design problems during the initial operation and maintenance periods. In effect, then, it was necessary for the system to continue evolving as these design problems were resolved.

Initial responsibility for the DEWLine construction was given by the U.S. Air Force to the American Telephone and Telegraph Company. The system went into operation, actually before all initial construction was complete, on October 24, 1956, ten years ago. The contract for its operation and maintenance is held by Federal Electric Corporation, worldwide service subsidiary of the International Telephone and Telegraph Corporation, which, among other things, has trained the personnel at the Air Force's DEWLine school in Streator, Ill., and kept the line manned by more than 1000 technicians.

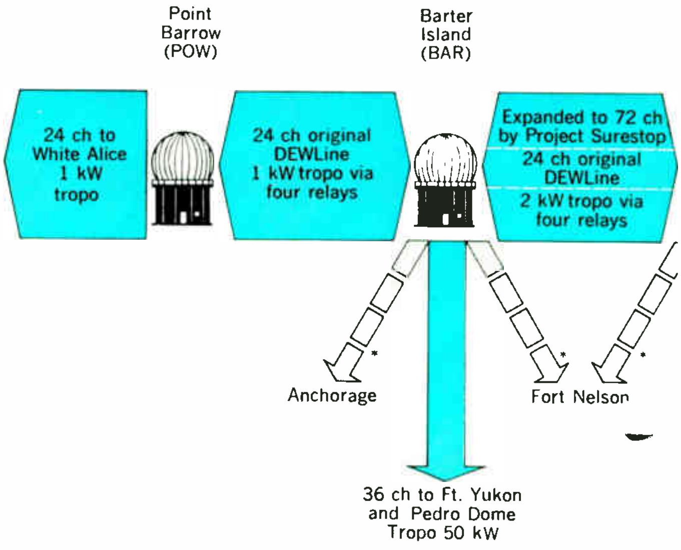
One of the primary technical problems that had to be resolved in the early days was the design of suitable lateral communications along the Arctic chain and reliable rearward links. Systems then available were unsuited for the awful Arctic weather and atmospheric phenomena. For short hauls, tropospheric scatter schemes, then in their early stages of development, seemed the best suited for the lateral hops, but at that time there were very little data on over-the-horizon links with tropo (a survey of tropo systems and their capabilities appears in "Tropospheric scatter communications—Past, present, and future," by Frank A. Gunther, in the September 1966 issue of SPECTRUM). After the successful demonstration of tropospheric scatter with a one-kW klystron, this was the mode decided on. The combination of radar siting criteria and the klystron power capabilities of those days led to the choice of setting the station sites at about 160-km intervals. This multichannel UHF tropospheric propagation radio system provided duplex voice and teletypewriter communications laterally. The original concept for the long-haul rearward communications links involved a dual-diversity IS-101 VHF ionospheric scatter communications system. However, this system remained rather limited in terms of reliability and communications capacity. It had only four teletypewriter channels available and, although it was possible to span over 1400 km between stations, the initial cost and subsequent maintenance costs were high in relation to the number of channels. However, between 1959 and 1963, the hardware people made such progress with tropospheric systems that it became feasible to deactivate all ionospheric systems and rely totally on new tropo systems for rearward communications. This is not to say that there were no rearward tropo systems in the beginning. At Cape Dyer, an AN/FRC-45(V) triple-diversity tropo scatter system with 10-kW output transmitted information rearward to Brevoort and Resolution Island (see Fig. 1). At POW Main, at the western end of DEWLine, the same kind of equipment transmitted information rearward down through the Air Force "White Alice" communication system.

Meanwhile, the klystron art was being driven hard, and power outputs were pushed upwards, so that by 1958, when highly reliable communication links were required between Thule (the original trading post named "Ultima Thule" meaning "The Utmost End") southwards to Cape Dyer (DYE Main) (see Fig. 1), this was bridged by SSB high-power tropo. The Lincoln Laboratory, once again acting out a pioneer role, working with Eitel-McCullough, developed a 50-kW klystron, which, through brute force power, provided sure tropo links over the big southward water gaps. As things turned out, this proved to be a fortunate development, since the underwater cable subsequently installed, owing to the severe iceberg conditions, has not been as reliable as had been hoped. These two systems, after the further development of the high-power klystrons, were augmented by the installation of a 100-kW FM tropo system between Thule and FOX during 1964.

On the radar end of things, the stations generally were equipped with AN/FPS-19 equipment—a dual-beam, dual-antenna early warning radar, operating in the L band, with both beams operating simultaneously on a back-to-back basis and the tilt of each antenna controlled at the console by the operator. These radars, in the Alaskan and Canadian sectors, were supplemented in the early years by a "gap-filler" Doppler radar, the AN/FPS-23. The Doppler radar, though not as discriminating in its target-location capability as the search radar, covered from 80 meters upward to overlap significantly the search radar coverage. However, as air traffic grew in the north, which was in part a consequence of increased DEWLine acti\ities, the Doppler system was showing an alarm condition much of the time. Furthermore, with experience, it became evident that coverage by the FPS-19 was adequate, and in 1963, therefore, Doppler radar was phased out. As Sam Bennie, manager of DEWLine operations, puts it in his fine Scottish burr, "We finally had the courage to switch the Doppler off."

In the Greenland sector, at the four DYE stations, surveillance is maintained with AN/FPS-30 radar systems.

The reader realizes, of course, that we are only skimming the barest highlights here. In addition to the equip-



ments mentioned, DEWLine stations have UHF and VHF air/ground transmitters and receivers, HF air/-ground receivers, disaster system monitors, emergency HF transmitters and receivers, VHF mobile communication equipment, and navigational aids such as AN/FRT-37 LF beacons, and in some cases TACAN and TVOR.

Hard times

DEWLine has had to live under the cruelest conditions from the very beginning, cruel to men and to equipment. These include ice storms, total darkness in winter, danger of sunburn and snow blindness when there is sun, and "whiteouts" on the Greenland glacier when planes must find their way down without quite knowing just when they will strike the ice landing field—and then have their skis stick when they try to take off again. Mountains of supplies must be rushed into the north by sea and air during the short summer months. The winter cold is so bitter that the momentary prop wash from a plane can cause serious injury to the bared skin, and a drop of spilled gasoline plummets the skin below freezing almost instantly. Such environmental conditions merely intensify the psychological problems for men living only with men in relative isolation.

On top of these problems, the equipment itself must be manned and maintained in a condition of minimum downtime. Even after the Herculean labors involved in its installation, the line presents curious and unusual problems of maintenance. To take just one rather spectacular example: Power klystrons, the heart of the long-haul vital tropo systems, cost up to $47 000. A single error in alignment and the replacement cost is $47 000 right on the spot. In fact, klystron failures have occurred at an unforeseen and costly rate (i.e., there were 68 failures reported between 1962 and 1965). Special studies had to be made in 1962 and again in 1965 by Federal Electric and Eitel-McCullough to resolve this problem alone.

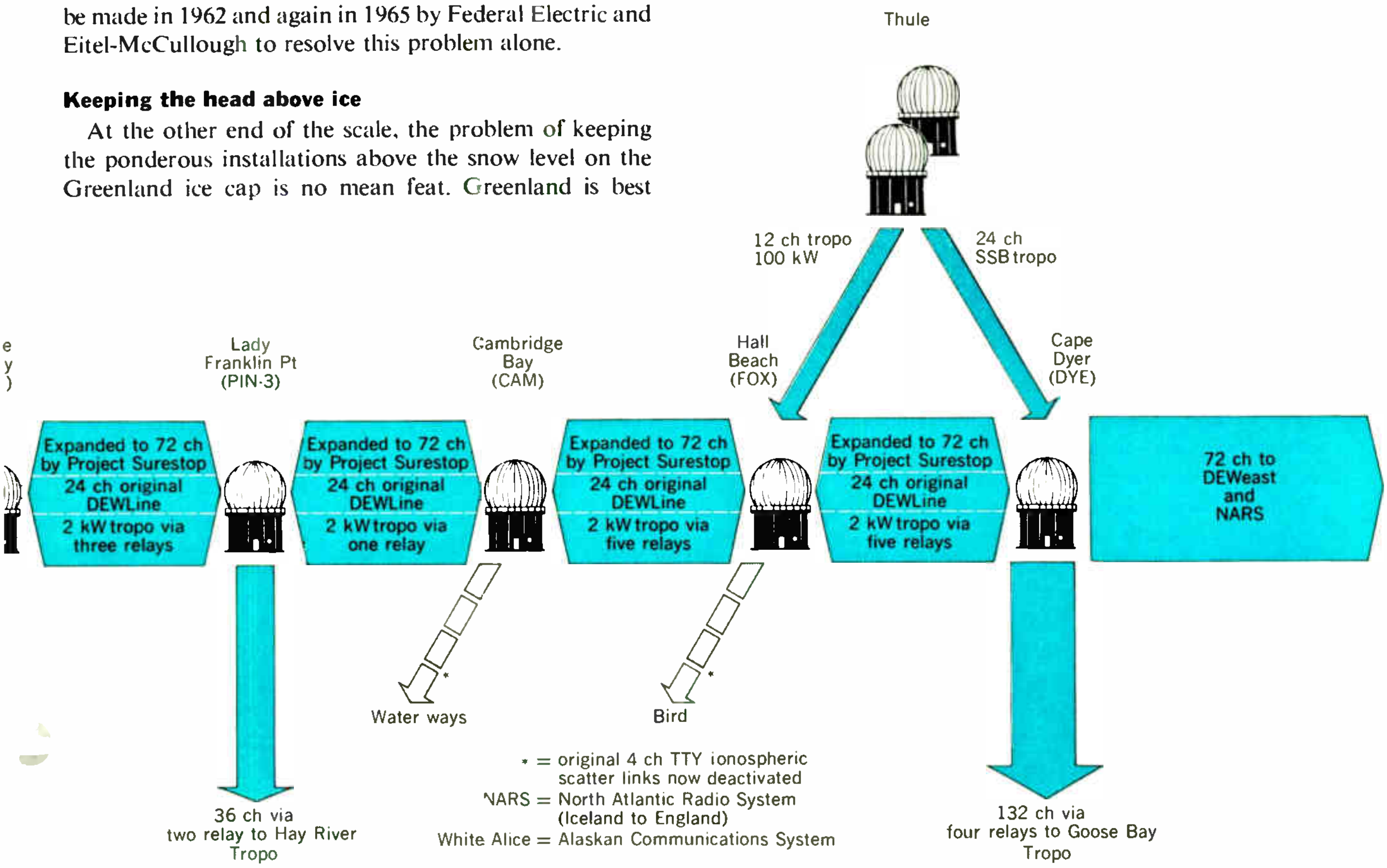
Keeping the head above ice

At the other end of the scale, the problem of keeping the ponderous installations above the snow level on the Greenland ice cap is no mean feat. Greenland is best described as an ice-filled bowl with the coastal mountains forming the walls of the bowl, the bowl being filled with accumulated snowfalls of centuries. The snow, under the pressure added by the weight of each successive snowfall, compacts to form solid ice. The icecap stations, DYE-2 and DYE-3, sit atop ice 3 km thick. These stations, weighing about 6000 tons each, are constantly bearing down on their iron feet, sinking into the ice about a half meter each year. As the annual snowfall (between one and two meters) does not melt, it becomes necessary to raise the buildings about four meters biennially to prevent their being buried. Each of the eight columns supporting the building incorporates hydraulic jacks for this purpose, and the columns themselves are periodically extended. Sewage problems and fire precautions, in a world where there is plenty of ice but no water, will be left to the reader's imagination.

Project Surestop

From the point of view of BMEWS needs and standards, the principal weaknesses of the DEW communications system were its unreliability and its noise. Nor did it possess anything like the channel capacity required by BMEWS. By DEW standards, the BMEWS people were lavish in their requirements in rearward communications. One grasps a feeling of the nature of the BMEWS reliability problem from one specification alone—no single BMEWS station can be down for more than 12 minutes

Fig. 2. Graphic summary of changes in the DEWLine communications network carried out under Project Surestop.



* = original 4 ch TTY ionospheric scatter links now deactivated
NARS = North Atlantic Radio System (Iceland to England)
White Alice = Alaskan Communications System

in an entire year. Furthermore, two routes, preferably through two different media, had to be set up for all BMEWS messages.

Some of the more significant deficiencies of the DEW system, from the point of view of BMEWS users, were as follows: The lateral communications system was basically a wired-in system, with minimum provision for interchanging equipments or for testing and monitoring functions. This prevented the adoption of a program of continuous testing and monitoring. The only patching and testing facilities provided were those at the master patching unit in the surveillance room at main stations and in the communications center at auxiliary stations. This master patch concept did not provide adequate flexibility for toll testing or rapid rerouting of traffic. Only two main switching centers were equipped as control centers for channel testing. Although DYE Main on Cape Dyer could test and control the North Atlantic Radio System (NARS), it did not have the capability of testing or controlling any of the 24 DEWLine channels. In addition to these major problems, a number of station links were operating with only marginal RF signal reliability owing to limited receiver sensitivity, low transmitter output power, and low antenna gain. Furthermore, the fact that the system operated on a dual-diversity rather than a quadruple-diversity scheme (see Gunther's discussion on diversity configurations) afforded neither acceptable signal nor system reliability. The system as designed provided for a maximum of only 24 voice channels, which could not handle the increased communications traffic. Furthermore, all lateral transmitters operated on a narrow bandwidth, resulting in channel capacity limitations and high intermodulation. Outages on the line would occur for various causes. For instance, with only one AN/FRC-45(V) transmitter operating at any time on a given terminal, a failure in this transmitter made the link inoperative during the time required to put the standby transmitter on the air. An especially crucial limitation was that in the power generating equipment. Primary power was supplied from only one bus. Failure of this bus resulted in a total communications outage and imposed severe limitations on circuit reliability.

Such were the types of deficiencies that had to be corrected. Upgrading of the DEWLine lateral communications and the rearward communications has been carried out under Project Surestop.

The DCA (Defense Communications Agency) consolidated and standardized the interfaces between the BMEWS and DEW systems, setting up the performance standards, and so on. Thus, the new requirements for the DEW system as a communications net reflect the general DCA standards. Very broadly, to become more reliable and to be suitable as a communications net, the DEWLine has over the past few years undergone several basic changes: the means of power generation at the stations had to be overhauled; channel capacity had to be expanded; station receivers were improved—in fact, the receive terminals were more or less replaced. In all, 23 stations in Canada and four stations in Greenland underwent modifications. Alaskan stations are unimproved as yet; however, these stations do not carry BMEWS traffic. Overall cost of the power improvements alone was about $20 million. Installation of the new communications equipment ran to approximately $15 million, of which some $10 million was spent on hardware.

The net result is that the DEW system has become a "marvelously noise-free communications system." The system is so set up that it can potentially handle 72 channels of lateral communication, of which 48 channels are fully installed now. If more capacity is needed in the future, all that will be required will be plug-in-type modules in the terminating equipment. In addition, there are now 72 channels of communication rearward out of Cape Dyer, 36 channels out of central Canada, 36 channels out of BAR, and 24 out of POW, Point Barrow, in the west. Also, 72 channels go out from DEW east to Iceland and Great Britain across NARS. Figure 2 summarizes the channelization plan as of August 1966. In all, there are now 5 700 000 circuit km in the DEW communication system.

Plans for the DEWLine include the transmission of autovon and autodin (automatic voice and data routing systems). Engineering work on these, and on high-speed data and error-correction systems, is under way.

The power upgrade program

The increased reliability of the current DEWLine facilities stems in part from the power upgrade program, which on the one hand increased power-generating capacity to meet the demands of new tropo equipment and to anticipate continually expanding needs, and on the other hand was designed to a new concept.

Specifically, under the program implemented for the Air Force by ITT, two additional 60-kW diesel generators were installed at the auxiliary stations on a special two-bus system, which permitted the use of both essential and nonessential loads on either bus. However, in the event of generator malfunction, load-shedding devices disconnect the nonessential load. With two generators on either bus operating at about 50 percent of rated load, the loss of one generating unit would insure that the total load of the affected bus will be maintained. Should the remaining generator be unable to support the entire load, the load-shedding equipment is reactivated. If an entire bus is lost, the essential load is automatically transferred to the remaining bus; if necessary, all nonessential loads will be thrown off.

At most of the main stations, completely new power buildings were constructed, two with 500-kW generating sets, the remaining with 150-kW sets; all have been equipped with two-technical-bus and automatic transfer equipment.

In terms of the tropo communication equipment, what this power upgrading system means is that if power-generating sets break down, traffic reliability is threatened, but the message does go through.

Does DEWLine have a future?

Now that the DEW system has been upgraded, one may wonder about its potential uses for other than military matters. Some observers feel that it has a vast potential for aerospace activities, as a vital east–west communication network for exchanging data with polar-orbit satellites, for experimental programs in aerospace communication and telemetry, and so on. Thus, it seems that one more Cold War expedient (a grandiose one to be sure), which one might have thought would have eventually faded away like the fur trade of that earlier era, is taking on the character of a permanent fixture.

The economics of desalination

In some instances desalination can provide a valuable source of fresh water; however, the high costs of integrating desalinated water into existing systems may result in a product that is not economically competitive with conventional supplies

S. Baron *Burns and Roe, Inc.*

The economic development of the process of desalination depends upon its ability to produce water that is competitive on a cost basis with water from conventional sources. Energy costs can be minimized by combining the water plants with power plants and thus allocating the energy charges. This article analyzes the thermodynamics of dual-purpose plants and shows how the economics is applied to obtain low energy costs for desalination. However, such plants also have certain cost limitations and penalties that must be included in any appraisal. The major factors to be considered are examined and discussed.

Desalination, or the production of fresh water from seawater (approximately 35 000-ppm solids) and from brackish water (over 2000-ppm solids), is a process that is highly dependent upon energy costs for its economic development. The energy costs for the rapidly developing evaporator processes for desalination of seawater are some 30 to 40 percent of the production costs in plants yielding 3800 cubic meters (1 000 000 gallons) per day, and 50–60 percent of production costs in proposed plants of 190 000 cubic meters per day and larger. The electrodialysis process, which is finding wide application for producing fresh water from brackish water, depends to a large measure on electric energy costs. Brackish water containing more than 5000-ppm solids generally is not treated by this method at present because the high energy requirements do not make the water costs attractive.

Recently developed processes, such as freezing, reverse osmosis, and ion exchange, though requiring less energy for water production, have compensating cost penalties. The freezing processes are more complex than evaporation, with greater capital costs per unit of water production. The reverse osmosis process, although low in energy, must bear the replacement cost of membranes on a six-month to two-year basis. Ion exchange, which theoretically has low energy requirements because solids are removed from saline water instead of water being removed from saline solutions as is generally done, has a high chemical cost because it is necessary to regenerate the ion exchange resins. Table I summarizes the 1963 and projected 1980 estimates of energy requirements for various desalination processes.[1] Interestingly, the accelerated pace of desalination developments since 1963 has indicated that the projected energy costs may be achieved much sooner than 1980, probably in the 1972–1975 period.

In 1965, the U.S. Congress authorized the Department of the Interior to allot $200 million to desalination developments for 1965–1971. The major part of this money (about 40 percent) will be spent on the evaporation process, in particular, on the promising multistage flash evaporation technique. The primary objective of the research, development, and construction of multistage flash evaporation units is to reduce or minimize energy requirements and costs. This can be achieved by developing low-cost, high-efficiency evaporation units and by combining water and power plants to permit economies in the energy charges for the water. Concurrently, the U.S. Atomic Energy Commission has undertaken the development of the heavy-water organic-cooled reactor, which potentially can produce very-low-cost steam in reactor sizes that will be required for the proposed large-scale desalination plants.

This article will analyze the thermodynamics of dual-purpose plants and will show how the economics can be

I. Energy requirements for various desalination processes

Process	Energy Requirement, kilogram-joules per cubic meter of product 1963	 1980
Evaporation	284 500	170 000
Freezing	170 000	100 400
Electrodialysis	142 000	86 000
Reverse osmosis	70 000	45 000

applied to obtain low-energy costs for desalination. However, the dual-purpose economic gains have certain cost limitations and penalties that must be appraised in order fully to evaluate the advantage to the total integrated power and water system. The major factors to be considered in an integrated system will be examined and discussed.

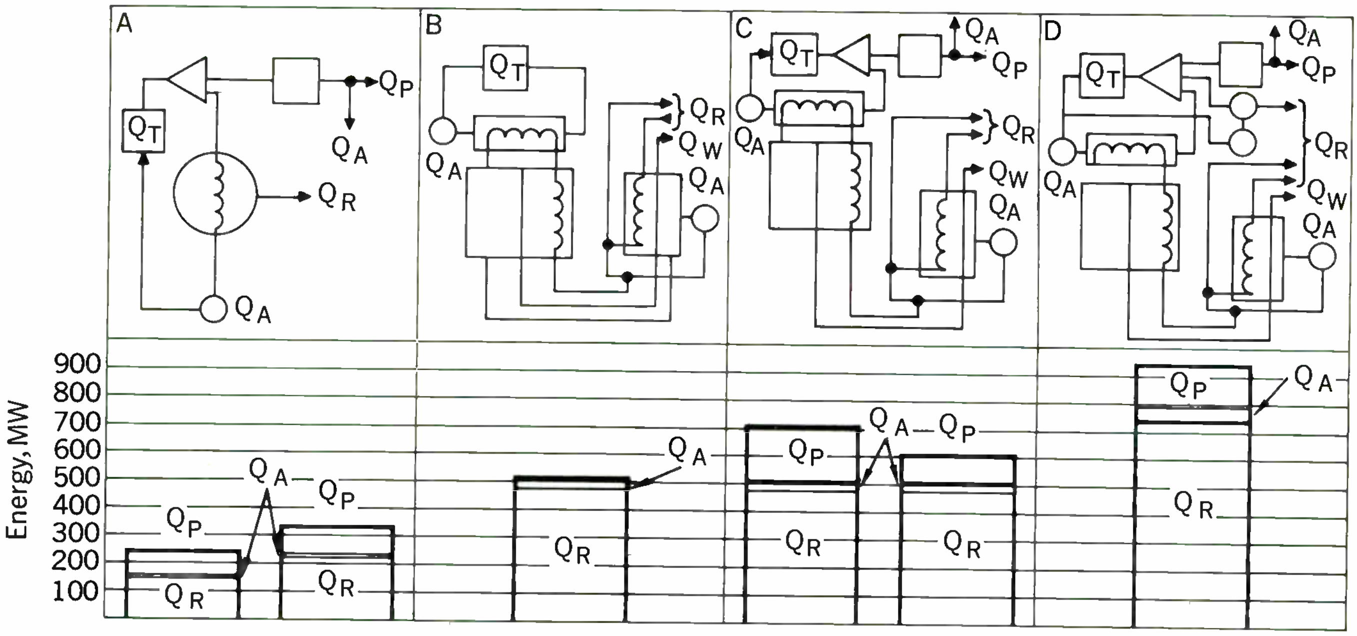
Thermodynamics

The dual-purpose power and water plant under consideration combines a steam turbine generator with a multistage flash evaporator that utilizes steam energy from the turbine for desalination purposes. By joining the power and water plants to a single heat source, certain obvious economies can be achieved in both capital and operating costs.

Obviously, the administrative, operating, and maintenance costs per unit of electric energy and water will be lower in combination plants than in separate single-purpose plants because of labor sharing. In addition, the use of a single heat source, even if it is equal in capacity to the sum of two separate heat sources, will have a lower unit cost of heat output because of the factor of size. The reduction of cost with size is especially pronounced in the case of nuclear reactors, which will be discussed later. However, the significant energy reduction for water production in dual-purpose plants is achieved by combining the plants and in the method of accounting.

Figure 1 presents the basic flow diagrams to be analyzed and the heat distribution among power, water, and auxiliaries, and in rejection. Excluding the heat loss of the boiler (10–15 percent for fossil fuels and less than one percent for reactors), a modern high-efficiency steam turbine–generator power plant will produce about 40 percent of salable power. In other words, as shown in Fig. 1(A), to sell 100 MWe of electric power (Q_P), the

Fig. 1. Thermodynamics of single- and dual-purpose plants.



Case Energy	A. Single-purpose power plant Fossil	 Nuclear	B. Single-purpose water plant Fossil or Nuclear	C. Noncondensing dual-purpose Fossil	 Nuclear	D. Condensing dual-purpose Nuclear
Thermal output, Mwt Q_T	250	330	480	710	610	940
Electric output, MWe Q_P	100	100	0	200	100	200
Water output, thousands of cubic meters per day Q_W	0	0	228	228	228	228
Auxiliary power, MWe Q_A						
Heat source, plant, MWe	5	5	5	10	10	15
Water plant, plant, MWe	0	0	20	20	20	20
Heat rejector, MWt Q_R	145	225	480	480	480	705

boiler plant will have a heat output of 250 MWt (Q_T), of which 5 MWe of power will be used for plant auxiliaries (Q_A) and 145 MWt of heat will be rejected to the condenser (Q_R). A nuclear power plant of the light-water type will be about 30 percent efficient; thus, for 100 MWe of power, the reactor will have an output of 330 MWt.

A high-efficiency multistage flash evaporator produces about 12.0 kg of water per kilogram of steam with heating steam at approximately 130°C. In a plant such as shown in Fig. 1(B) there is about 480 MWt of heat output from the boiler for 229 000 cubic meters per day of water (Q_W) with approximately 25 MWe of power purchased for plant auxiliaries. The total heat output of the boiler is rejected in the form of warm product water or in the heat rejection section of the evaporator, so that the boiler's 480-MWt heat output is rejected. Actually, the heat rejection is slightly higher because some of the purchased power is converted to rejected heat.

Most of the economic studies of dual-purpose plants[1-5] have examined the use of a topping turbine generator to the brine heater of the desalination plant. In a design such as shown in Fig. 1(C), the noncondensing dual-purpose plant has no heat rejection for power production and thus power is being produced at almost 100 percent efficiency. This arrangement is feasible operationally provided the plant is essentially base loaded, with the power and water being absorbed into an existing system. However, should demands for power and water fluctuate with time as well as with respect to each other, it may be necessary to provide for water storage capacity with a resultant increase in energy costs due to the partial loading of the unit and the fact that the steam bypasses the turbine.

When heat or thermal energy costs for water production in a noncondensing dual-purpose plant are calculated, it is general practice to charge only the difference between the total thermal energy of the dual-purpose plant and the thermal energy requirements of a single-purpose condensing power plant. In the nuclear case of Fig. 1(C), thermal energy charge for water would be (610 MWt − 330 MWt)/480 MWt, or 60 percent of the thermal energy for a single-purpose water plant. This is an appreciable reduction and is the major factor contributing to the low water costs with this design.

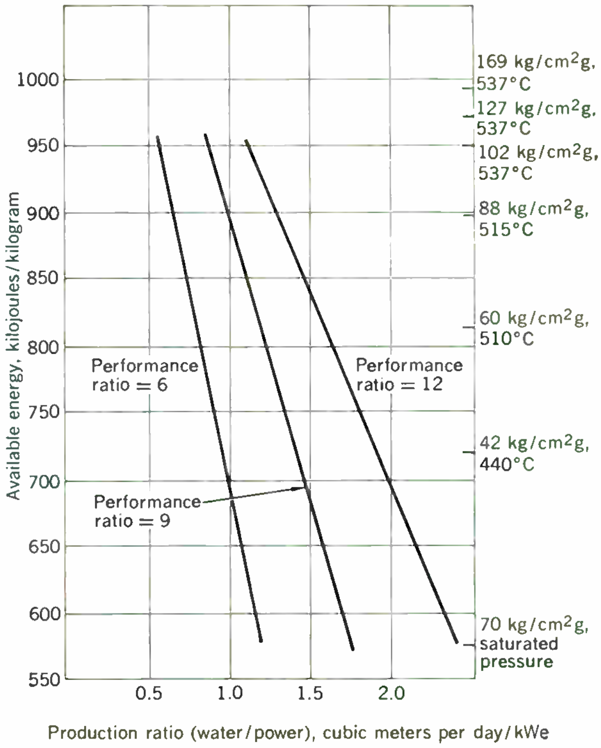
The condensing dual-purpose plant shown in Fig. 1(D) has the operational flexibility for varying the load of each product with time and with respect to each other but offers little or no thermal energy cost advantage over the separate single-purpose plants. The total heat of the combined plant is approximately the same as the sum of the heat requirements of the separate single-purpose plants so that the production cost of water of such a plant, base loaded, is higher than with the noncondensing design. If the demand upon the dual-purpose plant should fluctuate, the need for storage capacity is reduced; under some circumstances this may result in lower overall water costs than are possible with noncondensing plant.

Since the noncondensing turbine–generator desalination plant requires less heat than the condensing power plant it should be considered first for meeting the power and water requirements. Although this configuration has limited flexibility, particularly for part-load operation, there are still some parameters that can be varied in order to meet the different water-to-power production ratios. With a constant cooling water temperature, the design ratio of water to power can be modified by varying the steam conditions to the turbine, the backpressure of the turbine, and the performance ratio of the evaporator. This last is the ratio of the rate of water production per rate of thermal energy or steam flow required by the evaporator.

For the same steam turbine throttle conditions and evaporator performance ratio, the water-to-power ratio can be increased by increasing the backpressure of the turbine. Experience to date has limited the maximum brine temperature to 121°C (with steam around 130°C) to prevent scaling of the brine heater tubes with acid injection control and brine concentrations in the evaporator of some 70 000 ppm. The backpressure of the turbine would therefore be in the 2.1–3.2 kgf/cm^{2}a (30–45 psia) pressure range. Although lowering the backpressure below 2.1 kgf/cm^{2}a would decrease the water-to-power ratio, the evaporator costs would increase rapidly with decreasing brine temperatures at constant performance ratio—and thus the backpressure probably should not be varied if the evaporator performance ratio is held constant.

Figure 2 shows the approximate relationship of water/power production ratios with varying steam turbine throttle conditions and multistage flash evaporator performance ratios at a turbine backpressure of 2.1 kg/cm^{2}a (30 psia) at 121°C. Significantly, these curves show that by changing the turbine steam conditions from, say, 102 kg/cm^{2}g at 537°C, which is typical of fossil-fired plants, to 70 kg/cm^{2}g saturated steam conditions, which is typical of the water-reactor-type nuclear plants, we can double the water-to-power production ratios for a constant evaporator performance ratio. By varying the performance ratios of the evaporator we can achieve further significant changes in water-to-power production ratios. When we combine the variations in steam throttle conditions and the performance ratios (3 to 20), we see a design

Fig. 2. Available energy vs. production ratio (water power).



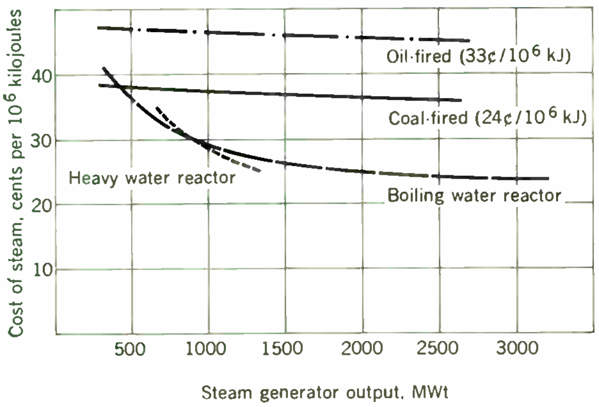


Fig. 3. Cost of steam vs. size of heat source.

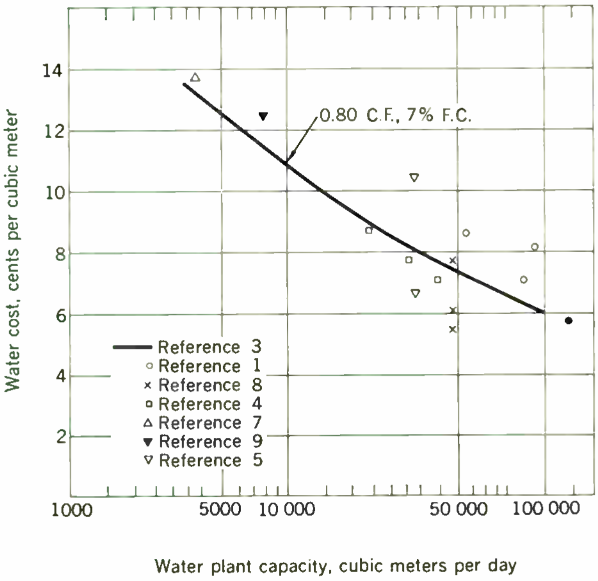
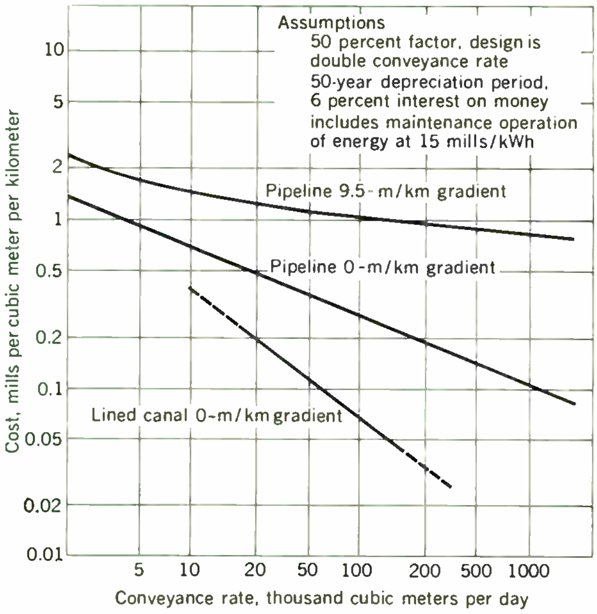


Fig. 4. Cost of water vs. size of water plant.

Fig. 5. Unit conveyance cost vs. average conveyance rate.



range in cubic meters per day/kWe of production ratios of water to power from 0.38 to 3.8.

Although a range of 10 in the water-to-power production ratio is achievable with the noncondensing turbine generator, once the ratio is fixed as the basis of the plant design, this plant cannot be operated at part load economically to meet varying power and water demands. If the turbine in this design is partly loaded, steam must bypass the turbine and be throttled to the evaporator to meet the water production load. Such an operation represents a loss of available energy and an increased energy cost for the power produced. Similarly, part loading the evaporator with full load on the turbine requires that turbine exhaust steam be condensed at 2.1 kg/cm^2a (30 psia), which represents a highly inefficient use of energy. To circumvent the high energy charges for part-load operations, the dual-purpose plant could be operated as a base load part of the time with the water storage sized to meet part-load water demand and the older or peaking power plants of the system meeting the part-load power demand. However, this alternative also represents cost penalties for product power and water because of the need for water storage and the operation of older power units.

Thermal energy cost

As shown in Fig. 1, the thermal energy requirement of the dual-purpose plant is appreciably higher than that of either the single-purpose power or single-purpose water plant. Since the heat source of the dual-purpose plant is larger, its unit capital cost in dollars per kilogram of steam per hour would be less than the unit capital costs of the heat sources for the separate plants.

In the case of fossil fuels, the unit cost of fuel is fairly independent of the size of the boiler. However, unit nuclear fuel cost is dependent upon the reactor size; it decreases with the increasing size of the reactor. These reductions are achieved by economies in larger-batch fuel fabrication and reprocessing, and in shipping of larger batches. In addition, larger reactors give greater neutron economy, thereby resulting in increased plutonium production and lower enrichment requirements. Furthermore, the unit capital cost of nuclear heat source decreases more rapidly with increased size than is true of comparable size changes in fossil-fueled plants.

All these factors explain the fundamental reasons for considering nuclear reactors for large-scale dual-purpose plants. Figure 3 compares the approximate cost of steam from fossil-fuel and nuclear-fuel heat sources with the size of the heat source. For this comparison, the capital costs were written off annually at 10 percent fixed charges and the plant was assumed to operate at 90 percent of capacity. The unit nuclear steam costs for both light-water reactors and the natural uranium heavy-water reactors decrease more rapidly with increased output than do the fossil-fuel plant costs for comparable size changes.

The data for the light-water reactors were calculated from General Electric's price list for boiling-water reactors. The heavy-water reactor steam costs are from a study recently completed for International Atomic Energy Agency.[6] The fossil-fuel steam costs were based on oil costs of 33¢/10^6 kilojoules (35¢/million Btu), which is typical of the worldwide price of delivered oil (without duty), and 24¢/10^6 kJ coal, which represents a typical price for bulk purchases of coal in the United States.

The major fact demonstrated by these curves is the significant unit cost reduction of steam achievable with nuclear reactors as the size of the heat source increases. Since heat sources for dual-purpose plants exhibit appreciable size increases, it is apparent that nuclear steam costs are lower than fossil-fuel steam costs once the thermal rating of the heat source is over 500 MWt. When the heat source is appreciably greater than 1000 MWt, the cost of steam from a heavy-water reactor is lower than from a light-water reactor. Therefore, when desalination plants of 190 000 cubic meters per day and larger are considered, the water production costs should be lower for heavy-water reactor types. The United States is now developing a heavy-water-moderated organic-cooled reactor for application to large-scale desalination plants. The Canadian heavy-water reactor, which was the basis for the steam cost data in Fig. 3, shows similar economies for large-scale desalination application.

Cost of water by desalination

In recent years a number of studies have been conducted in the United States to determine the cost of desalinated water in dual-purpose plants.[1–5,7] These studies have examined production for multistage flash evaporator sizes from 38 000 to 950 000 cubic meters per day. The results of these studies have generally shown that desalinated water can be produced at costs approaching or competitive with costs of water supplied by traditional methods.

However, it is important to understand the basis for these results since they were predicated on capital cost extrapolations and economic assumptions that may or may not be valid. As explained earlier, capital and operating costs for water in a dual-purpose plant with non-condensing turbine generator are charged as an incremental cost over these costs for a single-purpose power plant of the same net electrical output. By this method there is a significant reduction in the thermal energy charge against water, and economies also result by charging labor and certain capital costs against two products.

The largest flash evaporators built to date and in operation are less than 7600 cubic meters per day in size and thus the calculations for these plants are based on exceedingly large extrapolations of both capital cost and evaporator performance. It is reasonable to expect that the unit capital cost of the evaporators will decrease with increasing size as has been amply demonstrated in other industries. Economies result from bulk purchases of materials, adaptation of mass production techniques, and greater mechanization in fabrication, as well as from the distribution of overhead and engineering costs over a larger base of direct costs. The performance of these large-scale evaporators has also been calculated based upon data developed from smaller operating units as well as pilot plant and laboratory data.

The Office of Saline Water (OSW) of the U.S. Department of the Interior has embarked upon a development program to close the technology and cost gap between the present-day units and the proposed large-scale units. The test site, which is now under construction near San Diego, Calif., will include evaporator test modules, large-scale pump testing facilities, material and corrosion testing facilities, and a flash evaporator plant in order to check out the design and cost data. The OSW schedules call for the initiation of construction by 1968 of a 190 000-cubic-meter-per-day plant to incorporate the test data developed at the test site and to demonstrate the economics and performance of large-scale desalination units. However, the validity of the evaporator cost extrapolations and the hydraulic and heat transfer data will not be verified until the 1970s when this plant has been completed and is in operation.

It would certainly be helpful to those considering desalination if the results of the various studies could be correlated. Unfortunately, this would be most difficult because of the different assumptions employed for capital cost estimating and the variation in site conditions, as well as different values applied for crediting power, use of fixed charges, plant capacity factors, evaporator performance ratios, and cost of energy. The only consistency has been in the charging of capital and operating costs for water in dual-purpose plants as an incremental cost above the costs of a single-purpose power plant with the same net electrical output.

Figure 4 shows the scatter of data from recent studies in the range of water plant sizes from 38 000 to 1.9 million cubic meters per day with nuclear energy as the source of heat. The solid line, relating water costs with size, was developed by the author[3] for 80 percent capacity factor plant with fixed charges at 7 percent and evaporator performance ratio of 12. It represents a fairly good trend curve and average for the scattered points of the many studies.

The Metropolitan Water District Study of Southern California,[8] based on a 570 000-cubic-meter-per-day water plant, used fixed charges at 5.66 percent, 90 percent capacity factor, with various sized electric plants from 120 MWe to 1600 MWe. The lowest cost water resulted from the largest power plant because of the low energy cost of the large-scale reactors associated with such a plant. The Israel study[5] was based on a 380 000-cubic-meter-per-day water plant and a 200-MWe power plant with fixed charges varying from 5 to 10 percent. The lowest cost water obviously resulted from the lowest fixed charges; the power credit was unchanged with the varying fixed charges so that water costs were very sensitive to the fixed charged rate. The parametric studies[1,4,9] correlate fairly well with the solid line since they used similar capacity factors and fixed charges.

The results of the various studies indicate that the costs of producing water by desalination should be about 13.2¢ per cubic meter in the 38 000-cubic-meter-per-day plant, around 7.9¢ per cubic meter at 380 000 cubic meters per day, and 5.8¢ per cubic meter for 1.9 million cubic meters. In areas with special water problems—such as the Florida Keys, Israel, and Southern California—where water costs are high, studies indicate that desalination can produce water at a cost that is competitive with that of conventional methods and thus can provide a source of fresh water to meet the growing needs of these communities. The significant fact pointed up by these studies is that although desalinated water can compete economically with industrial or domestic water, the method is still too costly to be considered for general agricultural applications except under certain special local conditions.

Conveyance and storage costs

Any complete analysis of desalination water costs must also take into consideration the costs of conveyance and storage. The output from a desalination plant must be conveyed to an existing water distribution system as well

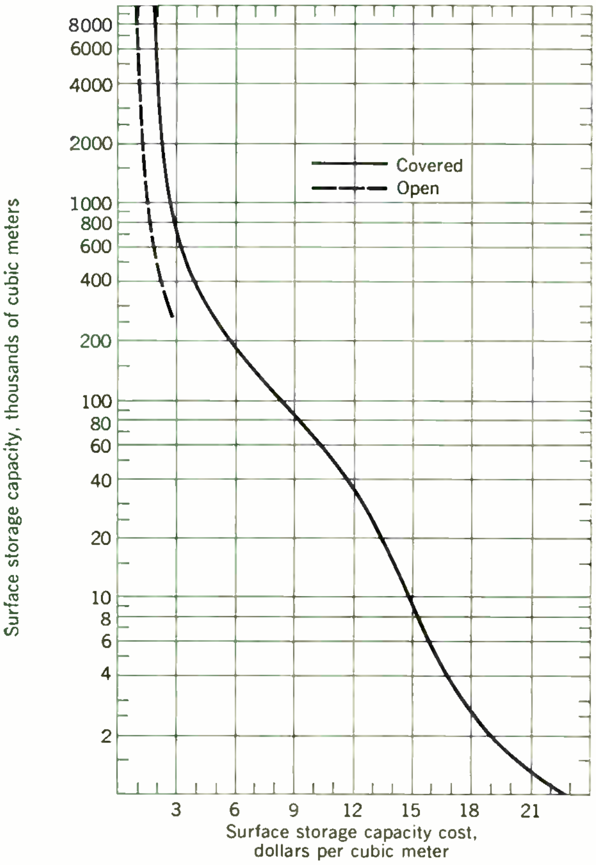


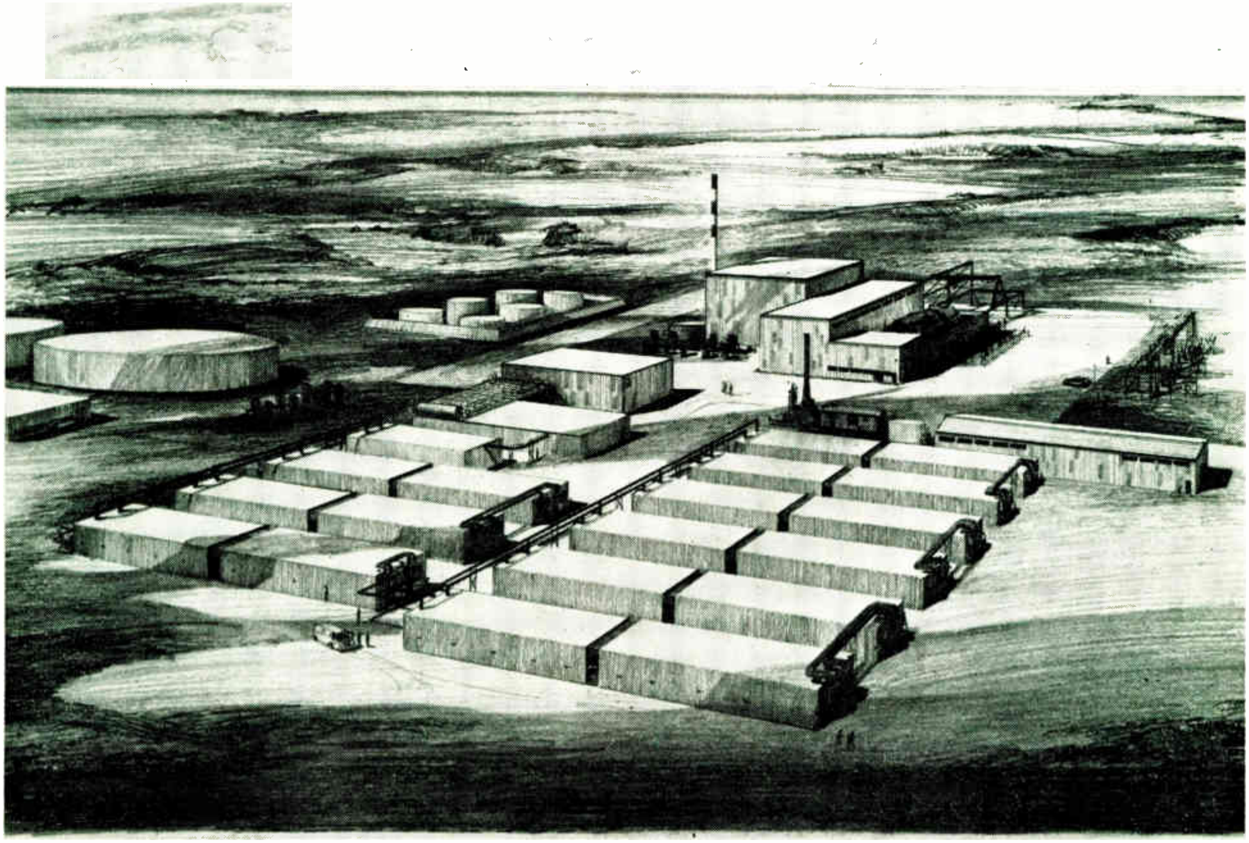
Fig. 6. Surface storage capacity vs. cost of surface storage.

as to the storage facilities that are necessary in order to maintain a balance between production and demand. The capacity of the conveyance system, which can be pipelines or canals, is defined by the output of the water plant. The storage facilities can be various types—steel or prestressed concrete for small storage capacities, lined excavated reservoirs for medium capacities, dams and reservoirs for large capacities.

Pipelines have many advantages over canals for water conveyance, although they are costlier in the same size category. Pipelines permit pressure pumping of water for long-distance conveyance as well as high velocities and therefore greater capacities for same line diameters; they can be mechanized for speedier installation and they can be operated under pressure to prevent contamination of water. Far less water is lost from pipelines by leakage and evaporation than from open canals. However, pipelines require greater maintenance and more frequent replacement because of internal and external corrosion, and are more expensive to install than earthen or concrete canals.

Figure 5 was developed from data in a paper presented at a 1965 United Nations seminar on desalination.[10] The curves compare the unit conveyance cost of water with average conveyance rate of the line for pipelines and open lined canals. These costs are based upon a 50 percent load factor so that the design rate is double the average conveyance rate. In addition, the depreciation period is assumed at 50 years with 6 percent interest on investment and pumping energy at 15 mills/kWh. Maintenance, operating, and repair costs are included in the unit conveyance cost.

Fig. 7. Dual-purpose plant for Florida Keys.



The cost of storage by reservoirs and dams is not only size dependent as in the case of storage tanks, but also depends upon the topography, soil conditions, and land value, and the related problems of flooding and stream control. It is, therefore, not possible to develop any meaningful correlation although the cost of large reservoirs and dams in the United States are reported to range from 4.5 to 9¢ per cubic meter for 1.15–2.3 billion cubic meters of storage capacity.[11]

The cost of surface storage capacity in the form of steel or prestressed concrete tanks and excavated lined reservoirs can be correlated with capacity because of their direct volume dependence. Such a correlation, shown in Fig. 6, was presented at the 1965 U.N. seminar; it was based on California construction costs. For capacities less than 10 000 cubic meters, costs were based on steel tanks; for sizes above 40 000 cubic meters, they were based on asphaltic concrete-lined excavated reservoirs. Between 10 000 and 40 000 cubic meters, both types are represented. The solid-line costs were calculated for roof construction whereas the broken line is without roof protection. By extrapolating the broken line to greater capacities we obtain costs of 9¢ per cubic meter, which approaches the upper limit of U.S. costs for large unprotected reservoirs.

Integrating desalinated water

Unfortunately, there is no best or preferred mode of operation for integrating the products from a dual-purpose plant into an existing power and water system. The most economical integration will depend upon the size of the existing system, the relative needs for power and for water, the quantities of additional power and water that will be required, and the specific characteristics of the system.

The method of integration and the plant design will vary from one extreme, in which a system is almost totally dependent upon desalination, to the other, in which desalinated water meets only a small portion of the water needs. In the former case, standby heat sources and additional storage and conveyance facilities may be necessary to assure a high plant availability, whereas in the latter it is unlikely that such factors will have to be considered.

Recent studies of applications in locations such as Southern California, Florida, Israel, Mexico, and New York City all fall between these extremes and probably represent typical future water problems. These are situations in which established communities have been obtaining their water from neighboring surface and groundwater supplies; however, with increasing industrial and population growth, new water supplies must be provided but these are either too costly to develop or just are not available. Under such conditions, desalinated water becomes an important part of the water supply. Figure 7 is a typical layout of a dual-purpose plant proposed for the Florida Keys, which has an electrical output of 50 MW and a water output of 38 000 cubic meters per day.

With this in mind, a number of major factors still must be considered before an economic solution can be reached. First, a site must be selected that is adjacent to a seawater or brackish water supply and that offers the most economic balance between power distribution and water conveyance to the demand centers. Since water conveyance costs appreciably more per unit distance than power distribution, the distance to storage or a central distribution system becomes a major cost consideration. On the other hand, if a nuclear reactor, which has the advantage of low-cost thermal energy in large sizes, is used, a compromise must be reached between a location convenient to populated water demand centers and one at distance from such areas for safety considerations.

Once a site has been decided upon, the capacity of the power and water plants and their mode of operation must be considered in relation to the cyclic demands for power and water. Generally, power and water demands are out of phase during the day and throughout the year so that water storage is necessary. On a daily basis, water demand is greatest during the daylight hours whereas power demand is greatest during the evening. Over the year, water demands are highest in the summer whereas the demand for power is greatest during the winter, although in some highly industrialized areas power demands are also high during the summer. Therefore, to meet these cyclic variations large amounts of water must be stored since power cannot be accumulated. Storage is required with reactors because they must be shut down for about a month each year for refueling. In this respect, the development of on-line refueling is certainly desirable for reducing water storage needs.

To illustrate the foregoing, consider a community requiring 200 MWe of electric power with water demand, as expressed in Fig. 8, consisting of 4½ months at 305 000 cubic meters per day and the remainder of the year at 153 000 cubic meters per day. A dual-purpose plant to meet these requirements can be either of the noncondensing or condensing turbine type combined with an evaporator as shown in Figs. 2(C) and 2(D).

In the noncondensing turbine cycle, the water plant would be base loaded at 230 000 cubic meters per day, with the excess over demand going to storage. Without on-line refueling, the nuclear reactor must be shut down for one month; therefore, as shown in Fig. 8, the excess water of Area A is stored in Area B to meet the demand for 305 000 cubic meters per day. In addition, the excess water production of Area C is stored as Area D to supply Area E's water during the shutdown. Such a design will require a storage reservoir of Area B capacity. This can be met by building a reservoir adjacent to the water plant or by pumping the output to an existing reservoir. In this analysis we will consider both alternatives: an existing reservoir assumed to be 42 km from the plant (50 ft/mile or 9.5 m/km gradient) and a new reservoir near the plant whose output must be conveyed 16.6 km to the main aqueduct or conveyance system.

To reduce the cost of a new reservoir, the dual-purpose plant can be designed with a condensing turbine so that the changes in demand can be met by variations in the extraction steam for the evaporators. In such an arrangement, the reactor will be larger and the power plant will have an additional circulating water system and condenser as well as a larger-capacity water plant. Storage capacity will be necessary for reactor shutdown but will be smaller, as represented by Area D. New storage capacities were oversized in all cases by 50 percent to take into account evaporation losses, leakages, and daily peaking variations.

For purposes of this comparison, water production costs taken from Fig. 5 were assumed to be 9.1¢ per cubic meter. Fixed charges were taken as 7 percent for the dual-

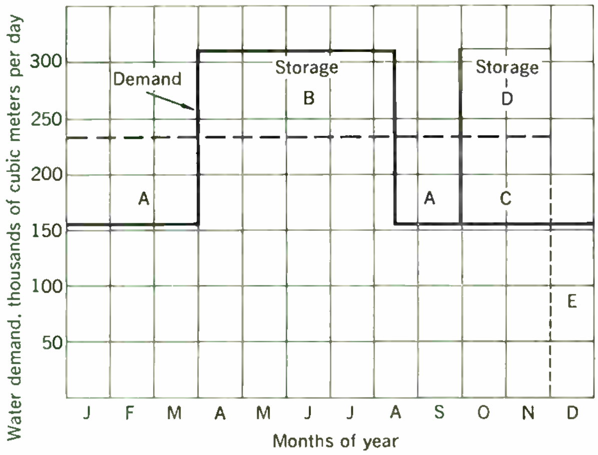


Fig. 8. Annual water demand curve.

II. Costs for integration of desalinated water output

	Cost per Cubic Meter		
Type of Charge	Case A*	Case B*	Case C*
Dual-purpose plant:			
Water production (Fig. 4)	$ 9.1	$ 9.1	$ 9.1
Increased energy (7 percent fixed charges)			1.47
Increased capital costs (7 percent fixed charges)			1.26
Storage (Fig. 6)	1.1		0.55
Conveyance (Fig. 5)	0.53	4.27	0.53
Total water costs	$10.73	$13.37	$12.91

* Note:

Case A: Base load noncondensing, storage 17-km conveyance system.

Case B: Base load noncondensing, 42-km conveyance system to existing storage.

Case C: Variable-load condensing, shutdown storage capacity, 17-km conveyance system.

purpose plant and as 6 percent for the conveyance and storage systems. Additional costs for integrating the water output would be as shown in Table II.

If the noncondensing turbine cycle were operated to vary with water demand, some of the exhaust steam would have to be condensed at 121 °C during the 153 000-cubic-meter-per-day operation in order to generate 200 MWe of power whereas during the 305 000-cubic-meter-per-day operation some throttle steam would bypass the turbine to the evaporator. Such an arrangement would be costly not only because of the increased reactor and evaporator size, but also because of the high energy costs, which are greater than those for Case C. This is the result of the turbine exhaust steam condensing at 131 °C and the throttle steam bypassing the turbine to the evaporator.

Conclusions

It is apparent from these costs, and under the conditions assumed, that the base-loaded dual-purpose plant with noncondensing turbine provides the most economic approach to desalination even though new storage capacity is required. Conveyance of water to an existing reservoir offers no economies because of the high cost of conveyance systems. Even variable operation of the water plant to reduce the storage capacity has no economic advantage because of the additional cost penalties for increased energy and capital.

Obviously then, when we analyze the economics of the situation, we must take into consideration more than just the basic production costs of the water. We also must study the cost effect of other factors such as conveyance and storage. These additional costs of integrating desalinated water can be very high indeed, and may result in a total delivered water cost that is not economically competitive with alternative surface or groundwater supplies that at first examination were considered too costly.

On the other hand, although too expensive to be considered for general agricultural applications, production of water by desalination can provide a promising source of fresh water for certain areas with special water problems where conventional supplies are inadequate to meet the needs of the growing population and increasing industrialization. However, we must recognize that the general emphasis in recent years on plant designs that give low water production costs have sometimes clouded the fact that these costs do not represent delivered water costs—a fact which should weigh heavily in any assessment of the feasibility of desalination for a particular application.

Revised text of a paper, "The Economics and System Integration of Dual-Purpose Power and Water Plants," presented at the World Power Conference, Tokyo, Japan, October 16–20, 1966.

REFERENCES

1. "An assessment of large nuclear powered sea water distillation plant," Office of Science and Technology, Executive Office of the President, Mar. 1964.

2. Hammond, R. P., "Large reactors may distill sea water economically," *Nucleonics*, vol. 20, pp. 45–59, Dec. 1962.

3. Baron, S., "Economics of reactors for power and desalination," *Nucleonics*, vol. 22, pp. 67–71, Apr. 1964.

4. Catalytic Construction Co. and Nuclear Utilities Services, "A study of desalting plants (15 to 150 MGD) and nuclear power plants (200 to 1500 MWt) for combined water and power production," Rept. NYO-3316-1, U.S. Dept. of the Interior and U.S. Atomic Energy Commission, Washington, D.C., Sept. 1964.

5. Kaiser Engineers and Catalytic Construction Co., "Engineering feasibility and economics study for dual purpose electric power–water desalting plant in Israel," to be published.

6. "Technical and economic data for nuclear and conventional power plants—preinvestment study on power including nuclear power in Luzon, Republic of the Philippines," Burns and Roe, Inc., New York, N.Y.

7. Burns and Roe, Inc., "Feasibility study of a dual purpose nuclear reactor power plant for the Florida Keys," Rept. NYO-10719, U.S. Dept. of Interior and U.S. Atomic Energy Commission, Washington, D.C., 1964.

8. "Engineering and economic feasibility study, phase I and II, for a combustion nuclear power–desalting plant (1965)," Bechtel Corp.

9. Burns and Roe, Inc., "Parametric cost studies pertaining to dual purpose power and water desalination plants," OSW Rept. 109, 1964.

10. Hansson, K. E., "Economics of conveyance of water," presented at 1965 Inter-Regional Seminar on the Economic Application of Water Desalination, U.N. Dept. of Economic and Social Affairs.

11. Hirshleifer, J., DeHaven, J. C., and Milliman, J. W., *Water Supply, Economics, Technology and Policy*. Chicago: The University of Chicago Press, 1960, pp. 182–183.

12. Golze, A. R., "Relationship between storage capacity and load factor of a desalination plant," presented at 1965 Inter-Regional Seminar on the Economic Application of Water Desalination, U.N. Dept. of Economic and Social Affairs.

Fundamentals of proportional navigation

A study of proportional navigation techniques, with emphasis on the time-varying behavior of the basic parameters; can provide valuable insight and information to the designer of antisatellite interceptor systems

Stephen A. Murtaugh, Harry E. Criel

Cornell Aeronautical Laboratory, Inc.

Proportional navigation has proved to be a useful guidance technique in several surface-to-air and air-to-air missile systems for interception of airborne targets. In this article, which is tutorial in nature, the basic theory of proportional navigation is presented and clarified. In addition, two variations on this guidance method are treated: one in which the commanded acceleration is biased by a small value of the measured rotational rate of the line of sight between the interceptor and its target, and one in which the line-of-sight rotational rate is reduced to a prescribed value (dead space) and then maintained at this rate until intercept. The analysis is directed, by example, to the case of the exoatmospheric interception of a satellite; however, the guidance theory presented is also applicable to the intercept of a nonmaneuvering airborne target.

Over the past two decades many papers and reports have treated various aspects of proportional navigation and specific applications of the technique, with emphasis on aerodynamically controlled interceptors having a speed advantage over a nonmaneuvering target. In recent years, proportional navigation has been analyzed for the potential role of guiding an interceptor to a target whose velocity is equal to, or even exceeds, that of the interceptor. In typical applications, intercept occurs outside the sensible atmosphere, so aerodynamic forces cannot be generated for vehicle control. Instead, the thrust of a rocket engine is used to provide the necessary maneuver forces, with engine swiveling or vehicle attitude control employed to direct the thrust in a direction normal to the line of sight between the interceptor and its target.

The application of proportional navigation and its several variations to exoatmospheric interceptors has been investigated by a number of researchers. However, most of the reports presenting the results of this work are either classified (because of the application) or else issued as company publications for internal distribution only, and are therefore not generally available. Notable among the companies that have contributed to this technology are Aerospace Corporation, Hughes Aircraft Company, Raytheon Company, and Space Technology Laboratories.

The analysis presented here is restricted to a two-dimensional end game, which is initiated when the interceptor seeker acquires a target satellite. It is assumed that prior to target acquisition the interceptor has been launched from the ground and guided toward a predicted intercept point on the basis of ephemeris data obtained from satellite tracking stations. Subsequent to target acquisition, the end game consists of nulling the projected miss distance produced by ephemeris uncertainties and midcourse guidance errors. This terminal homing phase will be accomplished by proportional navigation—that is, by generating control accelerations proportional to the measured rate of rotation of the interceptor-target line of sight and directing these forces so as to reduce or hold constant the rotational rate of the line of sight.

In addition to discussing the technique of proportional navigation, we shall attempt to demonstrate how linearized analysis, which can be performed without the need for extensive computer simulation, can provide much insight into various aspects of the preliminary design of a proportional navigation system and its effect on the interceptor configuration. As a result, certain effects are not considered, such as seeker gimbal friction, components of maneuver thrust acting out of the plane of the line-of-sight rotation, tracking noise, and time delays in providing the commanded maneuver forces. Although these effects can be important in the synthesis of a vehicle design, they are regarded as perturbations on the results presented here because of the difficulty in handling them in a linearized analysis. Tracker noise is frequently accounted for in an approximate manner in which the effect on a proportional navigation system is estimated based on experience and experiment. However, the two variations of proportional navigation treated are postulated as techniques that minimize the effects of tracker noise on the intercept accuracy, as discussed later.

Intercept geometry

The end-game geometry of the "ideal" collision course is depicted in Fig. 1 for the case of a constant-velocity, nonmaneuvering satellite target. The velocity of the target satellite V_S is dependent upon the orbital altitude. If the satellite altitudes of interest are considered to lie between 185 and 3800 km then the velocity range is relatively small: $V_S = 7800$ m/s at a 185-km altitude to $V_S = 6400$ m/s at a 3800-km altitude.

The velocity of the interceptor at the beginning of the end game, V_I, depends primarily upon the mass of the interceptor, the booster performance, and the intercept altitude. Typical interceptor velocities might fall in the range of 2500 to 7500 m/s.

The relative velocity V_R is the vector difference of V_I and V_S, as shown in Fig. 1, and is the velocity of the interceptor relative to the target. Based on the magnitudes of V_I and V_S, V_R ranges from almost zero to 15 000 m/s, depending upon the intercept geometry (head-on attack, tail chase, or lateral approach). However, practical considerations limit this range to relative velocities between 2400 and 12 000 m/s.

The line of sight $R_{(LOS)}$ is the line connecting the interceptor and the target. The length of this line at the time of acquisition is the acquisition range R_0. The value of R_0 depends upon the characteristics of both the terminal seeker and the target. Recent studies indicate that for radar or optical seekers and target cross sections of 0.2 to 0.5 square meter the acquisition range may be between 110 and 220 km. The same acquisition range appears to be reasonable for infrared seekers using cooled detectors.

The angle ϕ is the orientation of the line of sight with respect to inertial space; for convenience, the satellite track is used as the inertial reference. For the ideal collision course depicted in Fig. 1, the relative velocity V_R is aligned with the LOS; thus ϕ will remain constant throughout the end game. However, if V_R is not aligned with the LOS, the LOS will rotate at ever-increasing angular rates. Thus, the vehicles will not collide unless compensating maneuvers are performed by the interceptor.

The geometry shown in Fig. 1 assumes that some form of midcourse guidance has placed the interceptor on a collision course with the target satellite. In reality, when terminal homing is initiated, the interceptor will not be on a collision course with the target because of uncertainties in the actual location of the satellite and because of interceptor midcourse guidance errors. Thus, if the interceptor is launched on the basis of the estimated target position, at the time of acquisition the target will be displaced from its predicted position. This condition

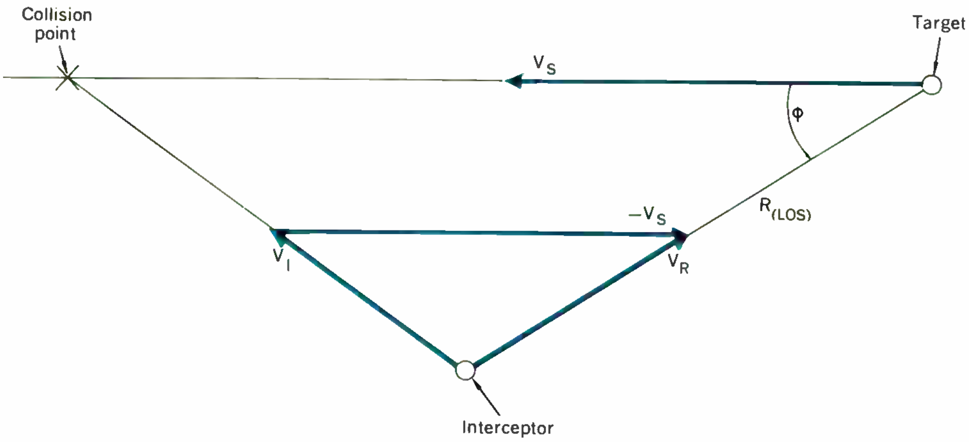


Fig. 1. Typical end-game geometry.

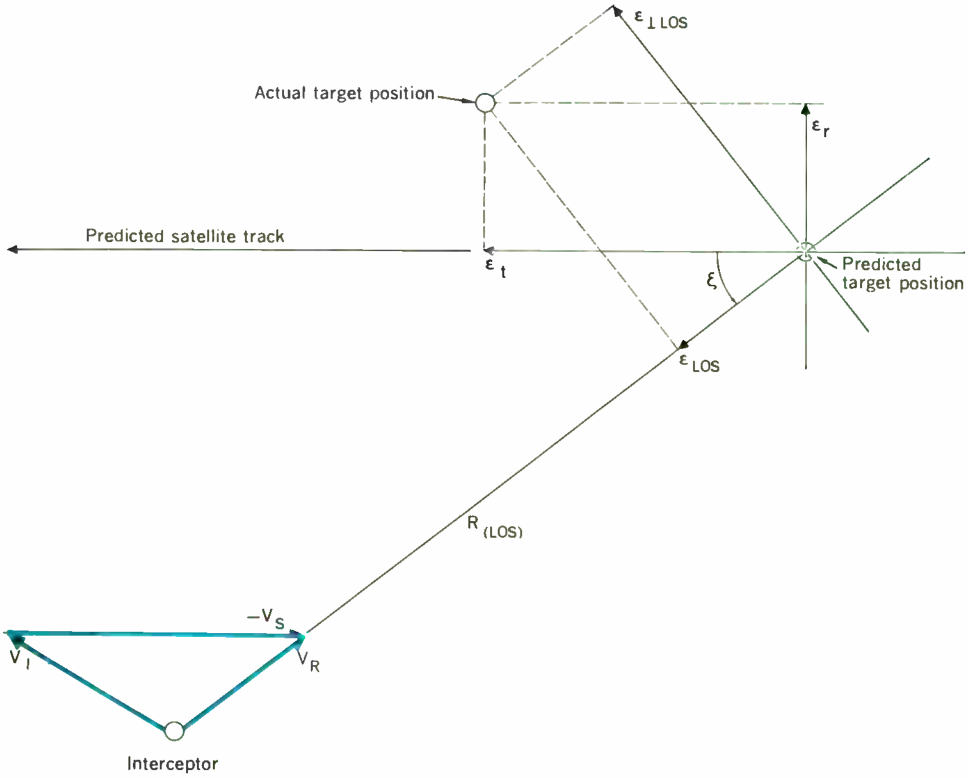


Fig. 2. Target uncertainty at acquisition.

is illustrated in Fig. 2 for the two-dimensional case. In this analysis it is assumed that midcourse guidance is perfect; however, midcourse errors could be regarded as additional uncertainties to the target position relative to the interceptor.

As shown, the error in the predicted target position has a cross-track component, ϵ_r, and an along-track component, ϵ_t. These components can be transformed to error components along and normal to tne nominal LOS ($\epsilon_{\rm LOS}$ and $\epsilon_{\perp\rm LOS}$).

Because the relative velocity V_R is aligned with the nominal LOS, any error $\epsilon_{\rm LOS}$ along the LOS alters the time of intercept but does not contribute to a miss distance. The miss distance that the interceptor must null is equal to the target uncertainty normal to the LOS:

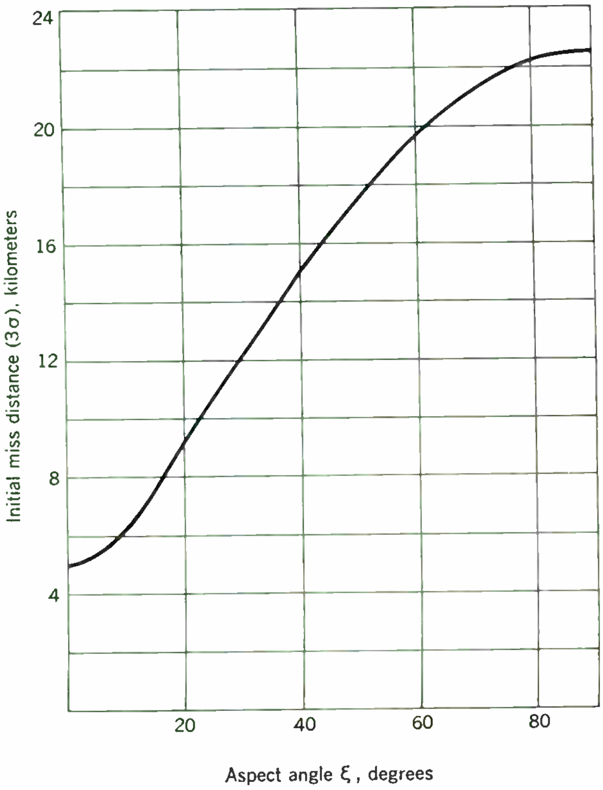
$$\epsilon_{\perp\rm LOS} = \epsilon_r \cos \xi + \epsilon_t \sin \xi$$

In terms of the error distribution normal to the line of sight, the standard deviation is given by the following expression under the assumption that ϵ_r and ϵ_t are independent:

$$\sigma_{\perp\rm LOS} = \sqrt{\sigma_r^2 \cos^2 \xi + \sigma_t^2 \sin^2 \xi}$$

The errors in predicting the position of the target correspond to uncertainties in determining the satellite ephemeris by ground tracking stations. For the sake of illustration, it is assumed that cross-track measurement errors are of the order of 2.0 km (circular error probability). Errors tangential to the orbit (that is, along the track of the satellite) are manifested as uncertainties in the time at which the satellite will pass over a given point. A typical timing uncertainty of 1 second (1σ) results in an error of about 7.5 km for a target speed of 7500 m/s.

Fig. 3. Effect of interceptor-target angle on initial miss distance.



In summary, the target position uncertainties just selected yield the following typical standard deviations in the target position:

$$\sigma_r = 1.7 \text{ km} \quad \sigma_t = 7.5 \text{ km}$$

Corresponding to these values of radial and tangential errors, the 3σ error normal to the nominal LOS is shown in Fig. 3 as a function of interceptor-target aspect angle ξ. The actual conditions at acquisition are depicted in Fig. 4. The interceptor, which is on an ideal collision course with the predicted target, moves up the nominal line of sight until the target comes within the acquisition range $R_{\rm (LOS)}$. At that time, the interceptor acquires the target, which is displaced from the closing flight path a distance M_0. This distance, whose 3σ values were presented in Fig. 3, is the miss distance that would result without terminal guidance and which must be nulled by the interceptor's guidance system.

Terminal guidance

In order to null the projected miss distance M_0, proportional navigation will be examined as a means of generating the appropriate guidance and control of the interceptor. The end-game geometry is depicted in Fig. 4 as a relative intercept; that is, the target is considered fixed in space and the interceptor is approaching the target with a relative velocity V_R.

The equations of motion of the interceptor are derived in vector form:

$$\mathbf{F} = m \frac{d^2\mathbf{R}}{dt^2}$$

As shown in Fig. 5, a unit vector triad (**i**, **j**, **k**) is oriented with the origin at the target, **j** aligned with LOS, and **k** normal to the plane of the vehicle velocity vectors.

The rotation of the **i**, **j**, **k** triad with respect to an inertial reference (taken as the satellite track) is

$$\mathbf{\Omega} = \dot{\phi}\mathbf{k}$$

Thus

$$\mathbf{R} = R\mathbf{j}$$

$$\frac{d\mathbf{R}}{dt} = \dot{R}\mathbf{j} + \mathbf{\Omega} \times \mathbf{R} = -R\dot{\phi}\mathbf{i} + \dot{R}\mathbf{j}$$

$$\frac{d^2\mathbf{R}}{dt^2} = -(\dot{R}\dot{\phi} + R\ddot{\phi})\mathbf{i} + \ddot{R}\mathbf{j} + \left[\mathbf{\Omega} \times \frac{d\mathbf{R}}{dt}\right]$$

$$= -(R\ddot{\phi} + 2\dot{R}\dot{\phi})\mathbf{i} + (\ddot{R} - R\dot{\phi}^2)\mathbf{j}$$

F, the applied force on the interceptor, is assumed to be the only control force applied to the interceptor normal to the line of sight. Differential gravitational effects have been neglected because of the short duration of the end game (of the order of 20 seconds), and the relatively large magnitude of the control forces. Thus, $\mathbf{F} = F_c\mathbf{i}$.

Equating the components of $\frac{\mathbf{F}}{m} = \frac{d^2\mathbf{R}}{dt^2}$:

$$\ddot{R} - R\dot{\phi}^2 = 0$$

$$R\ddot{\phi} + 2\dot{R}\dot{\phi} = -\frac{F_c}{m}$$

A few observations should be made before we proceed into the discussion of proportional navigation. First, note that if guidance were not applied (that is, $F_c/m = 0$), the solution to the differential equation in ϕ reveals that the LOS rotational rate would increase according to

$$\dot{\phi} = \left(\frac{R_0}{R}\right)^2 \dot{\phi}_0$$

where the subscript 0 refers to initial conditions. When guidance is applied, the relative velocity vector remains closely aligned with the LOS, and a very good approximation is $V_R \approx -\dot{R}$, where $\dot{R}$ remains constant since it is assumed that there is no acceleration component along the LOS. The equations of motion then become

$$\ddot{R} = 0$$

$$R\ddot{\phi} - 2V_R\dot{\phi} = -\frac{F_c}{m}$$

A virtual miss distance M is defined as the miss distance that would result if guidance were terminated at a particular value of time to go, T.

$$M = V_R T^2 \dot{\phi} = \frac{R^2\dot{\phi}}{V_R}$$

Note that

$$T = \frac{R}{V_R} = \left(\frac{R_0}{V_R} - t\right)$$

and, in the integration of the equations of motion, $dt = -dT$.

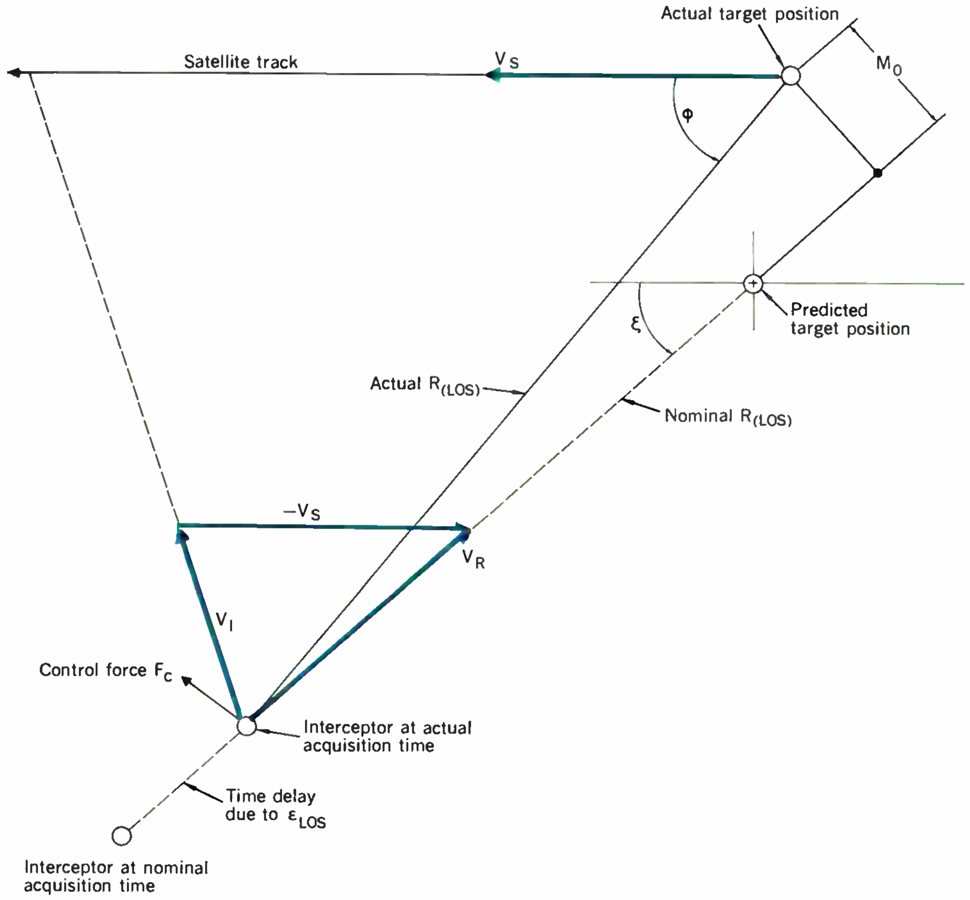

Actual target position
Satellite track
V_S
M_0
φ
Predicted target position
ξ
Actual $R_{(LOS)}$
Nominal $R_{(LOS)}$
$-V_S$
V_R
V_I
Control force F_C
Interceptor at actual acquisition time
Time delay due to ε_{LOS}
Interceptor at nominal acquisition time

Fig. 4. Acquisition geometry, actual conditions.

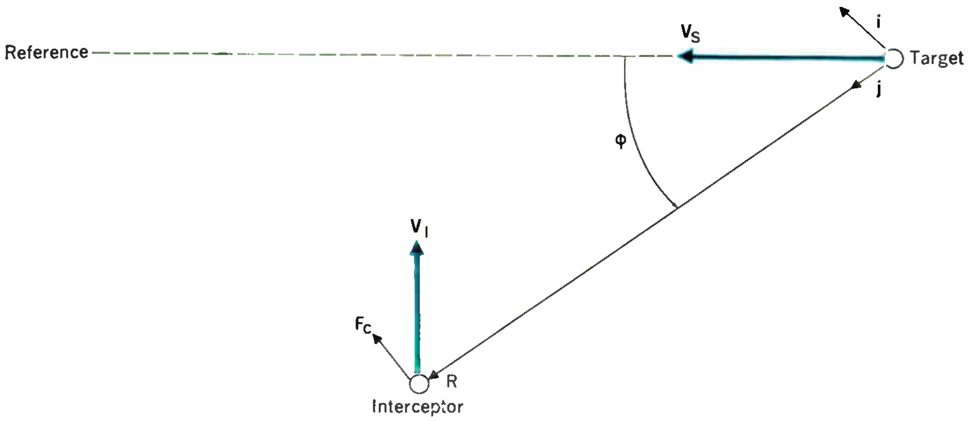


Fig. 5. Target-interceptor vector geometry.

True proportional navigation. In true proportional navigation, the interceptor acceleration commanded (F_c/m) is applied normal to the line of sight (see Fig. 4) and is proportional to the line-of-sight angular rate—that is, $F_c/m = k\dot{\phi}$. In order to facilitate the analysis, it is convenient to set the constant equal to λV_R, where V_R is the constant closing velocity and λ is defined as the navigation constant. For true proportional navigation, the equation of motion becomes

$$R\ddot{\phi} - 2V_R\dot{\phi} = -\lambda V_R\dot{\phi}$$

From this equation and the expressions for time to go and virtual miss distance, the following relationships for line-of-sight angular rate, LOS angular acceleration, and miss distance as a function of T and λ are obtained:

$$\dot{\phi} = \left(\frac{T}{T_0}\right)^{\lambda-2}\dot{\phi}_0$$

$$\ddot{\phi} = -(\lambda - 2)\dot{\phi}_0\frac{T^{\lambda-3}}{{T_0}^{\lambda-2}}$$

$$M = M_0\left(\frac{T}{T_0}\right)^{\lambda}$$

where

$$\dot{\phi}_0 = \frac{M_0V_R}{{R_0}^2} \quad \text{and} \quad T_0 = \frac{R_0}{V_R}$$

and where the subscript 0 denotes initial conditions at the beginning of the end game.

From these equations, it can be seen that if $\lambda < 2$, the LOS angular velocity and acceleration approach infinity as $T \rightarrow 0$; therefore, in the closing moments of the end game an infinite control force is required. A singular solution is represented by $\lambda = 2$. This represents the case of a constant control acceleration, acting through the end game, of just sufficient magnitude to null the initial miss distance. For the case of $2 < \lambda < 3$, the LOS angular acceleration approaches infinity as $T \rightarrow 0$; therefore, an intercept requires an infinite torque on the seeker in order to follow the line of sight. For $\lambda > 3$, the equations are well behaved and all accelerations remain finite throughout the end game. These phenomena are apparent from Fig. 6, which depicts LOS rate vs. time to go for various values of λ. As λ is assigned larger values, the LOS rate is reduced more rapidly, because the initial acceleration magnitude is proportional to λ. In the design of actual systems, it has been the practice to restrict λ to values between 3 and 6 in order to avoid instability in a system with time lags and noise. In this analysis, λ is assigned a value of 4 unless otherwise noted.

It can be seen from these expressions that for realistic values of λ, the LOS rate is reduced throughout the intercept, reaching a zero value (and a zero miss distance) as time to go becomes zero. However, there are variations of proportional navigation in which the LOS rate need not be controlled to a zero value, as will be discussed later. The cumulative velocity increment ΔV imparted to the interceptor during intercept is defined as

$$\Delta V = \int_0^T \left|\frac{F_c}{m}\right| dt$$

For true proportional navigation, it can be shown that the velocity increment required to intercept when an initial miss distance exists is

$$\Delta V = \frac{\lambda}{\lambda - 1}\frac{|M_0|}{T_0}$$

It is assumed here that the interceptor can provide the thrust called for; that is, the system is not thrust saturated by the initial acceleration requirement.

The initial acceleration required of the interceptor at the beginning of the end game in order to follow proportional navigation and to null the initial miss distance in the time available is

$$a_I = \lambda V_R\dot{\phi}_0 = \frac{\lambda M_0}{{T_0}^2}$$

This acceleration is a maximum at the beginning of the end game and decreases thereafter in direct proportion to the instantaneous LOS rate. From the foregoing equations, note that, for a given initial miss distance, as λ is increased the required velocity increment decreases and the required initial acceleration increases.

Plots of required velocity change and required initial acceleration as a function of initial miss distance for various closing velocities are presented for true proportional navigation in Figs. 7 and 8. In all cases, $\lambda = 4$ and the range at which guidance is initiated is 185 km. The dependency of a_I and ΔV on R_0 is shown in Fig. 9 for an initial miss distance of 15 km and closing velocity of 9000 m/s. Figures 7 and 9 are for ideal systems having no time lags, tracking noise, gyro drifts, or other random effects that may alter the ΔV requirements. In addition, it is assumed that the required acceleration can be provided; as will be shown later, if the interceptor is thrust limited and the time available is sufficient for the intercept to be accomplished, then the ΔV requirement will be still larger.

For the sake of illustration, consider that a 270-kg interceptor is being synthesized. A single-engine, variable-thrust, liquid-fueled interceptor configuration with a gimbaled seeker having an acquisition range of 185 km against a typical satellite target is selected. The orientation of the interceptor would be controlled so as to position the engine in the direction that thrust is desired. The seeker and its gimbal structure are designed to operate in a 20-*g* acceleration environment. Assume that launching the interceptor to achieve intercept at a selected point in space results in the interceptor approaching the target at a 40-degree aspect angle. This means that the intercept should be capable of nulling a 3σ target position error of 15 km (see Fig. 3). On the basis of booster performance and intercept geometry, an estimated closing velocity of 9000 m/s exists between the two vehicles, thus allowing 20.5 seconds to accomplish the intercept.

For $\lambda = 4$ and target acquisition at 185 km, an interceptor initial acceleration of 14.5 *g* is required. This means the initial maneuver force must be 38 600 newtons (8700 lbf) directed so as to reduce the angular rate of the LOS. The required maneuver force will diminish throughout the intercept to zero value at the end. In general, the throttling range of a rocket engine is limited and will not allow reduction from maximum thrust to very small values. Therefore, either the control force must be terminated at a small LOS rate and the resulting miss distance tolerated, or else a smaller throttleable engine might be added for the portion of intercept where small maneuver forces are required.

A velocity increment of 980 m/s is needed to accomplish this intercept. The mass of propellant required to provide this velocity increment can be calculated from the following expression, which relates the velocity gain due to thrusting (neglecting drag and gravity) to the specific impulse of the propellant and the vehicle and fuel mass.

$$\Delta V = I_{sp} g_c \ln \frac{W_0}{W_0 - W_f}$$

where

I_{sp} = propellant specific impulse, seconds
g_c = acceleration due to gravity at sea level, m/s^2
W_0 = vehicle initial total mass, kg
W_f = consumed propellant mass, kg

For a liquid fuel-oxidizer combination providing 300 seconds specific impulse in the near-vacuum of space at a combustion chamber pressure of 1200 lbf/in^2 (828 N/cm^2), the propellant required will weigh 77 kg.

Proportional navigation with a bias. As the LOS rate approaches zero, seeker tracking noise can cause the measured LOS rate to vacillate between positive and negative values. This requires the interceptor to accelerate and decelerate as the sign of the LOS rate changes, thereby imposing severe requirements on the interceptor control system. In order to alleviate these effects, variations of proportional navigation, in which maneuver forces are not commanded when the LOS rate has been reduced below a specific value, can be employed. In these schemes the LOS rate is reduced to a nonzero value; intercept is still achieved, but at the cost of additional propellant. Two of these variations of proportional navigation are discussed in this section and the next.

In a biased proportional navigation scheme, acceleration is commanded only when the magnitude of the LOS rotational rate $\dot{\phi}$ exceeds some positive bias value $\dot{\phi}_B$. The command acceleration is proportional to the difference between the magnitude of the actual LOS rate and the bias.

$$\begin{aligned} \frac{F_c}{m} &= \frac{\dot{\phi}}{|\dot{\phi}|} \lambda V_R(\dot{\phi} - \dot{\phi}_B) \qquad && \text{for } |\dot{\phi}| > \dot{\phi}_B \\ &= 0 && \text{for } |\dot{\phi}| \leq \dot{\phi}_B \end{aligned}$$

where $\dot{\phi}_B > 0$. In the analysis that follows, it is assumed that the initial miss distance is such that the initial LOS rate is positive and is greater than $\dot{\phi}_B$. The equation of motion then becomes

$$R\ddot{\phi} - 2V_R\dot{\phi} = -\lambda V_R(\dot{\phi} - \dot{\phi}_B) \qquad \dot{\phi}_0 > \dot{\phi}_B$$

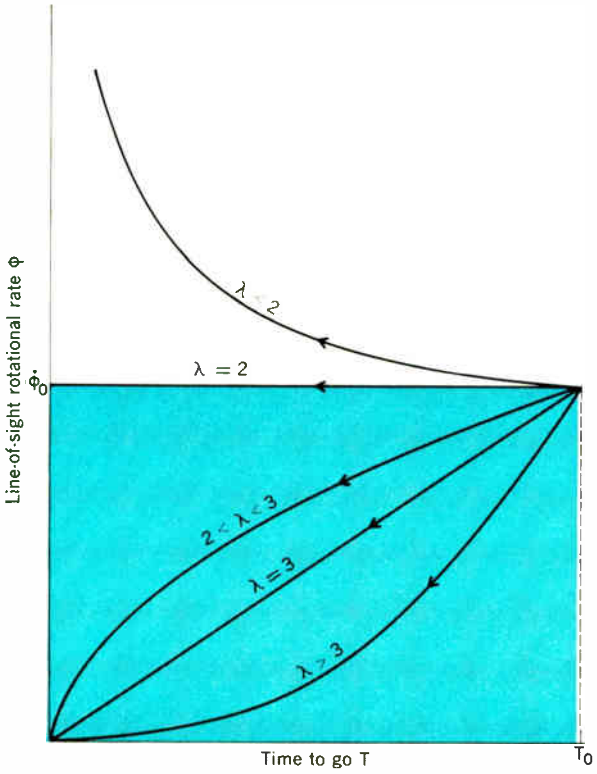
From this relationship, the following can be determined:

$$\dot{\phi} = \left(\frac{T}{T_0}\right)^{\lambda-2} \dot{\phi}_0 + \frac{\lambda}{\lambda - 2} \dot{\phi}_B \left[1 - \left(\frac{T}{T_0}\right)^{\lambda-2}\right]$$

$$\ddot{\phi} = \frac{T^{\lambda-3}}{T_0^{\lambda-2}} \left[\lambda\dot{\phi}_B - (\lambda - 2)\dot{\phi}_0\right]$$

$$M = M_0 \left(\frac{T}{T_0}\right)^{\lambda} + \frac{\lambda}{\lambda - 2} \dot{\phi}_B \left[V_R T^2 - \frac{M_0}{\dot{\phi}_0}\left(\frac{T}{T_0}\right)^{\lambda}\right] \qquad \dot{\phi}_0 > \dot{\phi}_B$$

Fig. 6. Line-of-sight rotational rate vs. time to go, for different navigational constants.



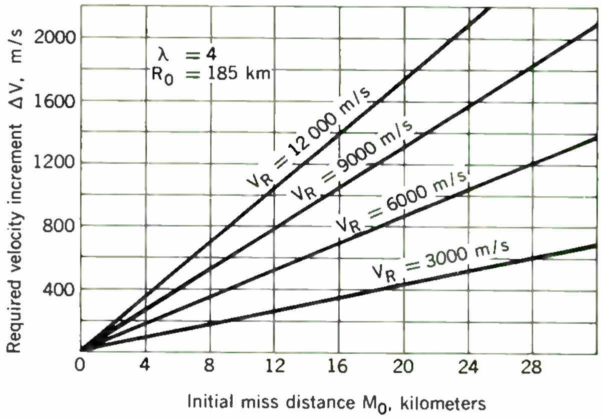
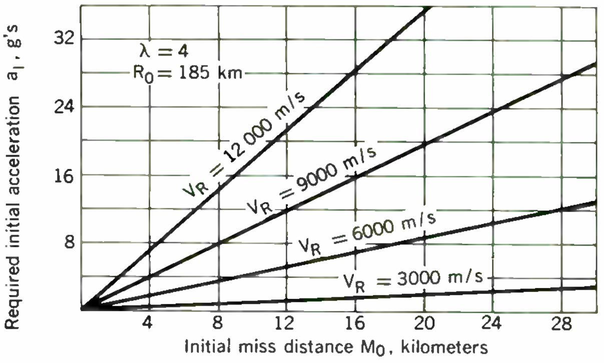


Fig. 7. Required velocity increment for true proportional navigation as a function of initial miss distance.

Fig. 8. Initial maneuver acceleration for true proportional navigation as a function of initial miss distance.



It is important to note that as time to go approaches zero, the miss distance approaches zero, but the LOS rate $\dot\phi$ approaches $(\lambda\dot\phi_B/\lambda - 2)$. For $\lambda = 4$, $\dot\phi$ approaches $2\dot\phi_B$.

The initial acceleration required is less than that for true proportional navigation by an amount equal to $\lambda V_R\dot\phi_B$. For the case of $\lambda = 4$ and $\dot\phi_B = 1$ milliradian per second, this reduction amounts to 1 to 5 g for closing velocities of interest (2400–12 000 m/s). In fact, for a given λ, a biased proportional navigation system always commands less acceleration at a given LOS rate than does true proportional navigation. Over the duration of the end game, this reduction in control results in an increased ΔV requirement for the biased system. Specifically,

$$\Delta V = \frac{\lambda}{\lambda - 1}\left[\frac{M_0}{T_0} + R_0\dot\phi_B\right]$$

The increased ΔV requirement of the biased system over that of true proportional navigation is given by

$$\Delta_B(\Delta V) = \frac{\lambda}{\lambda - 1}\, R_0\dot\phi_B$$

The increment of ΔV due to the bias is shown in Fig. 10 for $\dot\phi_B = 1$ mrad/s and $\lambda = 3$, 4, 5, and 6. The initial acceleration and required velocity increment for intercept using biased proportional navigation are shown in Fig. 11, which compares these parameters with the same parameters for true proportional navigation (and for dead-space navigation, discussed in the next section).

Note that when the initial miss distance is such that the initial LOS rate is less than the bias, $\dot\phi_B$, no control force will be commanded initially. In this case, the LOS rate will increase as the vehicles approach, and control forces will be commanded when the LOS rate exceeds the bias. However, because only small maneuver forces will be commanded, the LOS rate will continue to increase and, as before, it will approach $(\lambda\dot\phi_B/\lambda - 2)$ as T approaches zero.

The foregoing discussions and the solutions to the equation of motion are based on an initial LOS rate $\dot\phi$ that is positive. If the initial LOS rate is negative, the solutions are the negative of those shown.

Again, consider the interceptor example cited earlier, this time using biased proportional navigation. When a 1-mrad/s bias is used, the initial maneuver acceleration required at 9000-m/s closing velocity is reduced by 3.7 g to 10.8 g, resulting in a control force requirement of about 29 000 newtons (6500 lbf). However, from Fig. 10, it is seen that an additional 246-m/s velocity increment must be added to the interceptor, raising the velocity increment required from 980 to 1226 m/s and increasing the propellant mass by 15.5 kg to 92.5 kg.

Note that a control force will be commanded for all values of LOS rate in excess of the bias value, and as the time to go approaches zero, the LOS rate approaches $(\lambda\dot\phi_B/\lambda - 2)$ and commanded acceleration approaches $(2\lambda\dot\phi_B V_R/\lambda - 2)$. Thus, in this example, the LOS rate approaches 2 mrad/s at intercept. Therefore, the engine must be throttleable not to zero but to 3.7 g, or a thrust of about 6530 newtons (1470 lbf), allowing for a reduction in weight as fuel is consumed.

Proportional navigation with a dead space. In the case of proportional navigation with a dead space, acceleration is commanded only when the magnitude of the LOS rate $\dot\phi$ exceeds some given magnitude $\dot\phi_d$. The magnitude of this acceleration (when $\dot\phi > \dot\phi_d$) is proportional to the actual LOS rate $\dot\phi$. Thus, proportional navigation with a dead space is identical to true proportional navigation until $\dot\phi = \dot\phi_d$. When the LOS rate has been reduced by proportional navigation to the dead-space value,

$$T_d = T_0\left(\frac{\dot\phi_d}{\dot\phi_0}\right)^{\frac{1}{\lambda-2}}$$

At this time the commanded acceleration goes to zero. In the absence of control forces, the LOS rotational rate will tend to increase beyond the dead-space value, causing an acceleration to be commanded, which in turn drives the LOS rate back to the dead-space value. This cyclic control persists until intercept, causing $\ddot\phi$ to oscillate positively and negatively, producing an average $\ddot\phi$ of zero. Thus, as T approaches zero, the LOS rate is maintained at constant $\dot\phi$ and the miss distance approaches zero.

$$M = V_R T^2\dot\phi_d \qquad \text{for } T \leq T_d$$

The initial acceleration required, a_I, is the same as that for true proportional navigation. The ΔV is computed in two increments:

From $T = T_0$ to $T = T_d$, where $F_c/m = \lambda V_R\dot\phi$:

$$\Delta V_1 = \int_{T_0}^{T_d} \lambda V_R\dot\phi \; dt$$

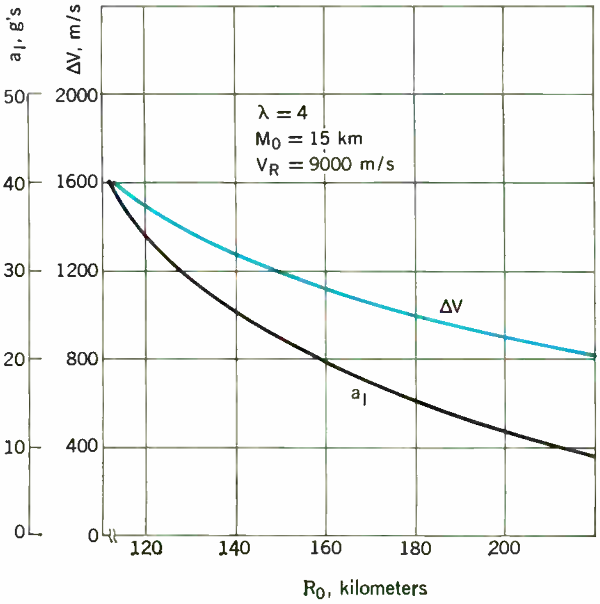
$$= \frac{\lambda}{\lambda - 1}\frac{M_0}{T_0}\left[1 - \left(\frac{V_R\dot\phi_d T_0{}^2}{M_0}\right)^{\frac{\lambda-1}{\lambda-2}}\right]$$

From $T = T_d$ to $T = 0$, $\dot\phi = \dot\phi_d$. It is apparent from the equation of motion that the acceleration required to maintain $\dot\phi$ constant (that is, $\dot\phi = \dot\phi_d$ and $\ddot\phi = 0$) is $2V_R\dot\phi_d$, irrespective of the value of λ. Therefore,

$$\Delta V_2 = \int_{T_d}^{0} 2V_R\dot\phi_d \; dt$$

$$= 2V_R\dot\phi_d T_d = \frac{2M_0}{T_0}\left(\frac{V_R\dot\phi_d T_0{}^2}{M_0}\right)^{\frac{\lambda-1}{\lambda-2}}$$

Fig. 9. Initial acceleration and cumulative velocity increment vs. acquisition range for true proportional navigation.



The total ΔV is then given by

$$\Delta V = \frac{M_0}{T_0}\left[\frac{\lambda}{\lambda - 1} + \frac{\lambda - 2}{\lambda - 1}\left(\frac{V_R\dot{\phi}_d T_0^2}{M_0}\right)^{\frac{\lambda-1}{\lambda-2}}\right]$$

This ΔV requirement represents an increase over that required for true proportional navigation (see Fig. 11 for a comparison with both true and biased proportional navigation). Specifically, the increment due to dead space is

$$\Delta_d(\Delta V) = \left(\frac{\lambda - 2}{\lambda - 1}\right)\frac{M_0}{T_0}\left(\frac{V_R\dot{\phi}_d T_0^2}{M_0}\right)^{\frac{\lambda-1}{\lambda-2}}$$

The increased ΔV increment of the proportional navigation system with dead space over that of true proportional navigation, $\Delta_d(\Delta V)$, is shown in Fig. 12 as a function of initial miss distance for $\dot{\phi}_d = 1$ mrad/s and for selected combinations of closing velocity and initial range. It is apparent from Fig. 12 that the larger the initial miss distance and the shorter the duration of the end game, the more closely the system with dead space approaches the true proportional navigation scheme.

In the region of small initial miss distances, the curves of Fig. 12 have been terminated at the point where $\dot{\phi}_0 = \dot{\phi}_d$. For smaller initial miss distances, $\dot{\phi}_0 < \dot{\phi}_d$ and thus no control acceleration is applied until such time as the LOS rate $|\dot{\phi}|$ builds up to the value of $\dot{\phi}_d$.

Again, consider the example interceptor illustration. It is seen from Fig. 12 that the use of proportional navigation with 1-mrad/s dead space results in a slight increase in the ΔV requirement for intercept, amounting to about 58 m/s over that for true proportional navigation. This is considerably less than that needed for the system using an equivalent amount of bias and requires only 4.1 kg more propellant than does true proportional navigation. In addition, the required throttling range of the engine is reduced because the minimum maneuver thrust commanded is established at a nonzero value determined by the values of navigation constant, closing velocity, and dead space selected. In this example, the throttling requirements will be eased inasmuch as the minimum control thrust will be about 6900 newtons for a 1-mrad/s dead space as opposed to zero for true proportional navigation. A disadvantage of this technique is that once the LOS rate has been reduced to $\dot{\phi}_d$, the engine must operate in an on–off mode, thereby requiring a multiple restart capability. One alternative to this last requirement is to change the navigation constant from whatever value is used to 2 at the time that $\dot{\phi}_d$ is reached, and then operate this engine at constant thrust as in true proportional navigation with $\lambda = 2$.

Considerations of nonideal conditions

The foregoing analysis of forms of proportional navigation was performed under the assumption of ideal conditions. These results can be extended to include certain additional considerations that depart from the

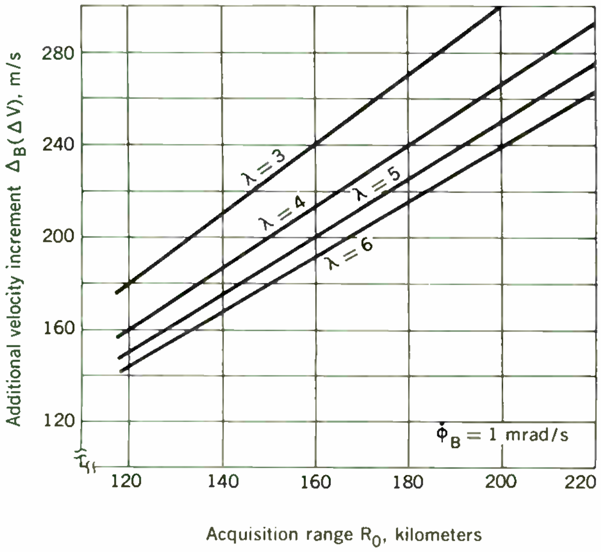
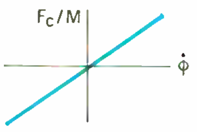
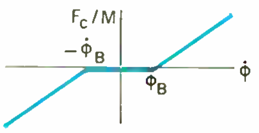
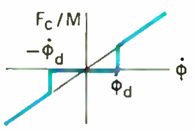


Fig. 10. Increased velocity increment for proportional navigation with bias.

Fig. 11. Comparison of biased and true proportional navigation with dead-space navigation.

Guidance type:	True	Bias	Dead Space
Commanded acceleration	 Slope = λV_R	 F_c/M $-\dot{\phi}_B$ $\dot{\phi}_B$ $\dot{\phi}$ Slope = λV_R	 F_c/M $-\dot{\phi}_d$ $\dot{\phi}_d$ $\dot{\phi}$ Slope = λV_R
Initial acceleration (maximum)	$\lambda V_R \dot{\phi}_0$	$\lambda V_R(\dot{\phi}_0 - \dot{\phi}_B)$	$\lambda V_R \dot{\phi}_0$
Required velocity increment $\Delta V = \int_T \left\lvert\frac{F_c}{m}\right\rvert dt$	$\frac{\lambda}{\lambda - 1} R_0 \dot{\phi}_0$	$\frac{\lambda}{\lambda - 1} R_0(\dot{\phi}_0 + \dot{\phi}_B)$	$\frac{\lambda}{\lambda - 1} R_0 \dot{\phi}_0 \left[1 + \frac{\lambda - 2}{\lambda - 1}\left(\frac{\dot{\phi}_d}{\dot{\phi}_0}\right)^{\frac{\lambda - 1}{\lambda - 2}}\right]$

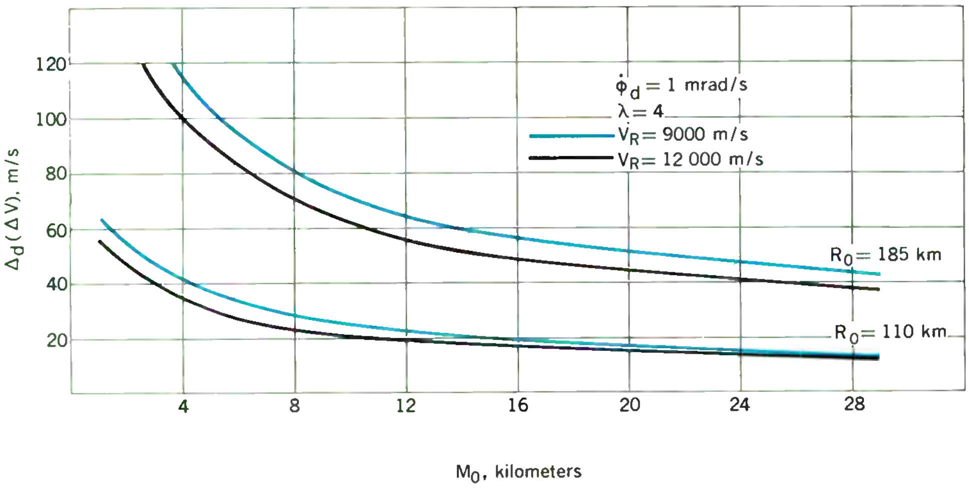


Fig. 12. Increment of ΔV for proportional navigation with dead space.

ideal case and, as a result, place an additional burden on the interceptor subsystems. Brief discussions of the effects of two such considerations—acceleration saturation and gyro drift—are presented in this section.

Acceleration saturation. The condition of acceleration saturation results when the commanded maneuver acceleration exceeds the acceleration that the interceptor propulsion system can provide. If acceleration saturation persists throughout the entire end game, a finite miss distance will result; that is, the available acceleration will not be sufficient to null the initial miss distance within the available time. However, for some intercept geometries the acceleration saturation will last only during the initial portion of the end game; that is, the maximum available thrust is applied until the line-of-sight angular rate is reduced to a point at which the acceleration commanded is an attainable value. Then a zero miss distance can be achieved, but the propellant required (ΔV) is greater than that necessary to intercept for an unsaturated mode.

For the example interceptor, consider that the vehicle has a maximum thrust of 53 300 newtons and an initial mass of 270 kg (corresponding to an initial maneuver capability of 20 *g*). Figure 13 presents the results of a digital computer simulation showing the effects of acceleration saturation during terminal guidance for proportional navigation with dead space ($\dot{\phi}_d$ = 1 mrad/s). The solid lines in the figure define the ranges of intercepts (initial miss distances) that can be accomplished even though the commanded acceleration cannot always be provided, for selected closing velocities and acquisition ranges with the 53 300-newton-thrust interceptor. The circles on each curve indicate the maximum miss distance that could be nulled in an unsaturated mode—that is, if the acceleration commands never exceeded the interceptor capability of 53 300 newtons thrust.

It is seen that the miss distance that can be nulled is increased considerably in theory, simply by operating the engine at maximum thrust as long as required within the time available, provided there is sufficient propellant. If 100 kg of the interceptor mass can be allocated to propellant (neglecting additional propellant required for time lags, noise, etc.), then for proportional navigation with 1-mrad/s dead space the capability exists for increasing the interceptor velocity to almost 1370 m/s. This provides the upper limit on miss distances that can be nulled in the acceleration-saturated mode, as indicated in Fig. 13.

Also shown are colored curves, diverging from the black curves, which show the intercept capability that could be provided if the interceptor was not thrust limited. For a given miss distance, the difference between the two curves indicates the increased velocity increment required to intercept in the acceleration-saturation mode instead of increasing engine thrust.

Gyro drift. It has been assumed that during the end game the target is continuously tracked by a gimbaled seeker in the interceptor. A gyro is then used to measure line-of-sight rate with respect to inertial space. Consequently, a drift rate in the gyro will manifest itself as an apparent LOS rate, thereby introducing an error in the system. The following two types of gyro drift can be investigated in light of the net effect on a true proportional navigation scheme: constant drift and *g*-dependent drift.

When a gyro drift is present, the seeker senses it as an additive term to the true LOS rate:

$$\dot{\phi}_m = \dot{\phi} \pm \dot{\phi}_G$$

where $\dot{\phi}_m$ is the measured LOS rate, $\dot{\phi}$ is the actual LOS rate, and $\dot{\phi}_G$ is the gyro drift.

Thus, the acceleration commanded of the control system is

$$\frac{F_c}{m} = \lambda V_R(\dot{\phi} \pm \dot{\phi}_G)$$

In the case of a constant drift, $\dot{\phi}_G$ can be considered as a constant bias for all LOS rates; thus, the acceleration command curve (F_c/m vs. $\dot{\phi}$) is shifted so that even when the true LOS rate is zero a control force is generated by the gyro drift. Therefore, a constant gyro drift can be examined in a manner similar to that used for proportional navigation with a bias. As was shown for the case of proportional navigation with a bias: as $T \rightarrow 0$, $\dot{\phi} \rightarrow \pm(\lambda\dot{\phi}_G/\lambda - 2)$. But unlike proportional navigation with a bias, if the drift is in the opposite direction of the initial LOS rate, a reversal in the control force will result when the magnitude of the true LOS rate equals the drift rate. Constant drift does not contribute to a miss distance, but it does require an increment in ΔV that is equal to

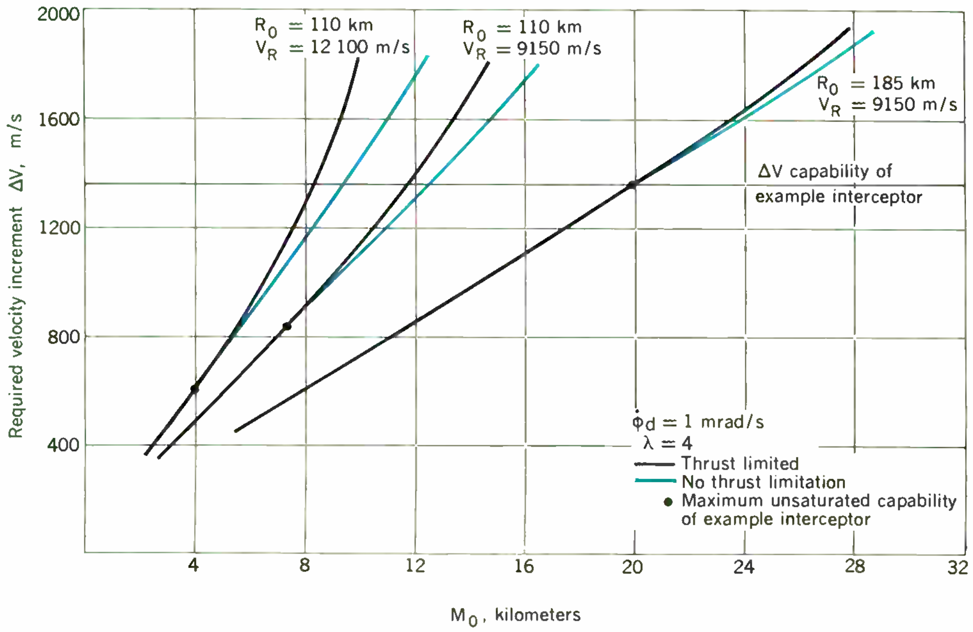
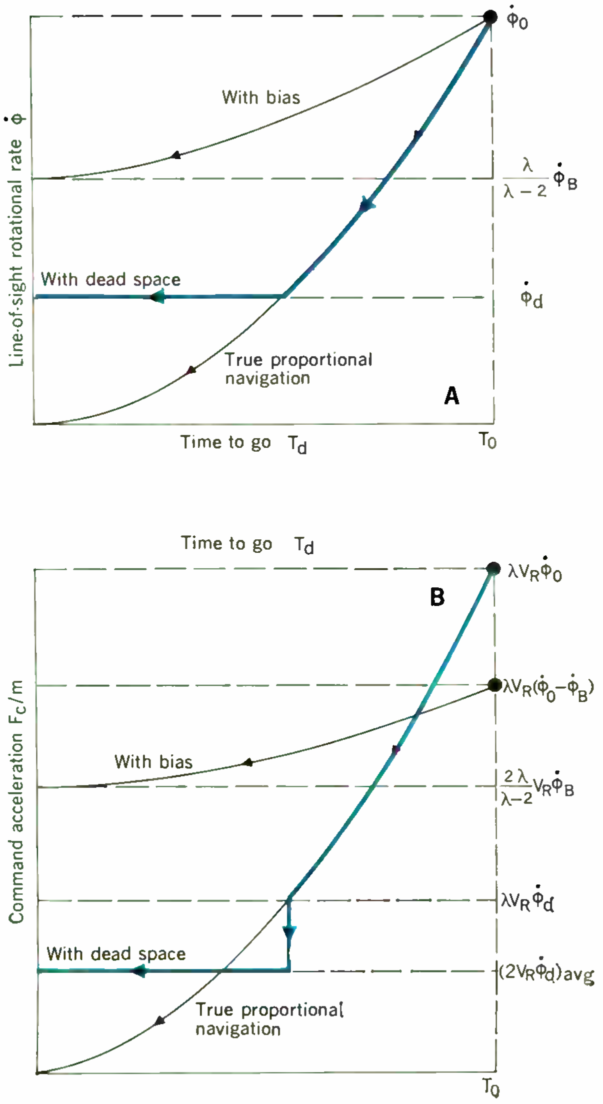


Fig. 13. Effect of acceleration saturation for proportional navigation with dead space.

Fig. 14. Line-of-sight rotational rates (A) and control accelerations (B) for various forms of proportional navigation.



$\pm(\lambda R_0\dot{\phi}_G/\lambda - 1)$. For gyro drifts of about 0.1 mrad/s, the increase in ΔV is only about 24 m/s for an R_0 of 185 km and $\lambda = 4$.

For the case of g-dependent gyro drift, the drift is proportional to the interceptor acceleration:

$$\dot{\phi}_G = \pm K\frac{F_c}{m}$$

where K is the drift proportionality factor. Thus, the acceleration commanded by the proportional navigation system is

$$\frac{F_c}{m} = \lambda V_R\left(\dot{\phi} \pm K\frac{F_c}{m}\right) \quad \text{or} \quad \frac{\lambda}{1 \mp \lambda V_R K} V_R\dot{\phi}$$

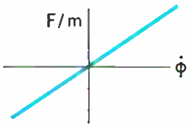
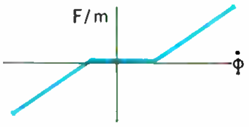
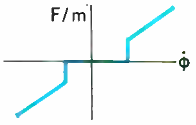
Therefore, the system behaves like a drift-free system with a navigation constant

$$\lambda' = \frac{\lambda}{1 \mp \lambda V_R K}$$

For example, when $\lambda = 4$ and for 9000-m/s closing velocity, if the drift proportionality factor is 0.1 mrad/s per g, then $\lambda' = 6.3$ (if the drift is in the direction of the measured LOS rate). This results in a 60 percent increase in initial acceleration and a 10 percent decrease in required velocity increment.

Concluding remarks

The preceding discussion has defined the characteristics of proportional navigation and several of its variations and also considered its application to a satellite interceptor. The differences among the various proportional navigation techniques have been described and the effects on the interceptor maneuver propulsion system illustrated. The differences noted in the propulsion requirements can have significant effect on the development of an interceptor configuration. For a given intercept condition, the variations in propellant required and maximum maneuver thrust for the different guidance techniques are not too critical. However, the need for a

Guidance type:	True Proportional Navigation  	Proportional Navigation plus Bias (1 mrad/s)  	Proportional Navigation plus Dead Space (1 mrad/s) 
Acceleration commanded Initial (gravity units) Final (gravity units)	14.5 g 0	10.8 g 3.7 g	14.5 g 3.7 g
Thrust commanded Maximum Minimum	38 600 N (8700 lbf) 0	28 900 N (6500 lbf) 6530 N (1470 lbf)	38 600 N (8700 lbf) 6860 N (1550 lbf); 0
Throttling ratio	>100:1	4.4:1	5.6:1
ΔV required (⊥ to LOS)	980 m/s	1226 m/s	1038 m/s
Propellant required*	77 kg	92.5 kg	81.1 kg
Problems	Wide throttling range Noise at low LOS rates	Maximum fuel requirement	On–off–on operation

$*\Delta V = I_{sp} g_c \ln \dfrac{W_0}{W_0 - W_f}$

Fig. 15. Comparison of guidance propulsion requirements for example interceptor.

wide throttling range and for on–off operation can impose severe design requirements.

The true proportional navigation system with $\lambda > 2$ requires a maneuver thrust capable of continuous modulation from maximum to zero. This is not available presently and would probably have to be approximated by an engine with a throttling range of 50 or 60 to 1. A further problem occurs at small LOS rates because of tracking noise which can cause the measured LOS rate to vary in sign, thereby calling for incorrect maneuver forces. This effect is reduced by the two variations of proportional navigation considered.

Control based on proportional navigation with bias reduces the requirement for extreme throttling range because the initial acceleration required is less than that for true proportional navigation by an amount equal to $\lambda V_R \dot{\phi}_B$ and the thrust required at intercept is $(2\lambda V_R \dot{\phi}_B/\lambda - 2)$ minimum. Required thrust variations may be no more than 10 to 1. Proportional navigation with dead space also reduces the range of engine throttling required; however, because the commanded acceleration is cyclic after the LOS rate is driven to the dead-space value, the engine requires a multiple restart capability. The advantages of reduced throttling range and minimization of tracking noise effects at small LOS rates make these two variations more attractive than true proportional navigation, with the variation using bias resulting in the simplest engine because it does not require a restart capability. In summary, the time history of the rotational rates of the line-of-sight and control accelerations for the three forms of proportional navigation considered in this article are depicted in Fig. 14. In addition, Fig. 15 presents a comparison of the guidance and propulsion requirements for these three forms of proportional navigation, based on the interceptor example used here.

A point of interest is the fact that the majority of the results presented have been obtained from closed-form solutions to the end-game kinematics. These solutions, based on simplifying but realistic assumptions and approximations, yield considerable insight into many of the important design parameters and constitute a powerful tool in the preliminary analysis of a sophisticated guidance technique.

Revised text of a paper presented at the IEEE Aerospace Systems Conference, Seattle, Wash., July 11–15, 1966.

BIBLIOGRAPHY

Adler, F. P., "Missile guidance by three-dimensional proportional navigation," *J. Appl. Phys.*, vol. 27, pp. 500–507, May 1956.

Bennett, R. R., and Mathews, W. E., "Analytical determination of miss distances for linear homing navigation systems," Tech. Memo 260, Hughes Aircraft Co., Mar. 31, 1962.

Booton, R. C., Jr., "Optimum design and miss distribution of homing missiles," Meteor Rept. 50, M.I.T., Mar. 31, 1950.

Gallagher, J. M., Jr., and Trembath, N. W., "Study of proportional navigation systems with linearized kinematics," Dynamic Analysis and Control Lab. Rept. 75 (Confidential), M.I.T., Apr. 15, 1953.

Janus, J. P., "Homing guidance," Rept. TOR-469 (9990)-1, Aerospace Corp., Dec. 10, 1964.

Jerger, J. J., *Systems Preliminary Design.* Princeton, N.J.: Van Nostrand, 1960.

Kishi, F. H., and Bettwy, T. S., "Optimal and suboptimal designs of proportional navigation systems," presented at the Symp. on Recent Advances in Optimization Techniques, Pittsburgh, Pa., Apr. 21–23, 1965.

Locke, A. S., *Guidance.* Princeton, N.J.: Van Nostrand, 1955.

Newell, H. E., Jr., "Guided missile kinematics," U.S. Naval Research Lab., May 22, 1945.

Schechter, H. B., "A brief survey of trajectory guidance, and propulsion aspects of orbital rendezvous," Memo. RM-3275-PR, Rand Corp., May 1963.

IEEE publications

scanning the issues
advance abstracts
translated journals
special publications

Scanning the issues

Transistor Failures. Mysterious failures in transistors under certain operating conditions were first reported in 1958. The failure phenomenon, now generally known as second breakdown, has since been the subject of intensive research. Transistor designers have been seeking to understand the phenomenon in order to eliminate it, and circuit designers have worked on special designs to circumvent it.

Despite the research, the phenomenon of second breakdown is still something of a mystery. The first organized discussion of the problem by the various investigators took place at an IEEE meeting in 1965. The papers from that meeting, and others that have been written since, are now being published in two special issues of the IEEE Transactions on Electron Devices. The guest editor, W. M. Portnoy, points out that it is generally accepted that second breakdown is thermal in origin. However, the conditions under which it occurs, the sites of its occurrence, and the nature of these sites constitute problems that are still under discussion, if not dispute. The difficulty of making sensible, nondestructive measurements is partially responsible for the mystery. Portnoy anticipates that the efforts to understand and describe second breakdown will continue. But he also expects that the greatest efforts will be expended where potential economic gains are greatest. The problem might be circumvented by some combination of device and circuit design before the phenomenon is completely understood.

A review of what is now known about second breakdown appears in the lead paper, "A Survey of Second Breakdown," by H. A. Schafft and J. C. French. The significance of the problem is measured out in one sentence of theirs: As the need for highly reliable electronics systems, and higher power and higher frequency transistors, has grown, so has the problem of second breakdown.

A graphical view of the second-breakdown phenomenon is reflected in Fig. 1. These swept V_{CE}–I_C characteristics represent three different constant base current drive conditions; the letters R, F, and O signify, respectively, reverse base drive operation, forward base drive operation, and operation with the base open-circuited. For sufficiently large collector currents, each of the three curves shows an abrupt decrease in V_{CE}. This drop in voltage is the most obvious indicator of the initiation of second breakdown.

Although it is not possible to trace Schafft's and French's review of the research that has been going on, their discussion of the meaning of second breakdown may be useful to many engineers who work with transistors. The authors say: What do we mean by second breakdown in diodes or in n^+nn^+ structures or perhaps in other semiconductor devices? We think that the way the term "second breakdown" has been generally applied suggests that it is meant to cover the range of thermally induced phenomena that are manifested by an apparently spontaneous decrease in voltage and a simultaneous constriction of current. If second breakdown is the result of thermal processes or the dependence of the parameters of the device

Fig. 1. Swept V_{CE}–I_C characteristics of transistor. Mysterious low-voltage mode of curves R, F, and O is called second breakdown.

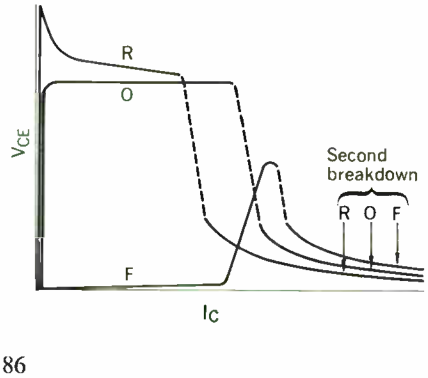


Fig. 2. Multiplate antenna test section.