IEEE spectrum

features

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

departments

Forum

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

Miami Beach revisited

I noted with considerable interest the letter from Prof. Balabanian in the April issue, in which he wonders why a recent Symposium on Circuit Theory was held in "pompously extravagant" Miami Beach. Although his question was admirably answered in the same issue by Prof. Wing, the asking of the question brings up several other points.

First of all, Miami Beach is primarily a middle-class resort, and it is this same middle class that pays the greater share of the individual income taxes from which the Federal Government derives most of its revenue. Since as Prof. Balabanian noted, most of the participants at the symposium have been receiving government grants, it seems to me that his criticism is more than a little gratuitous.

Second, in his list of socially necessary programs that are being shortchanged, Prof. Balabanian includes his own little niche, basic research. There are many who would question this classification, in view of the fact that circuit theory, for example, plays a negligible part in solving any of the major social problems that most of us agree exist.

This leads to my third point, namely, that engineering educators, as such, can do little in any direct way to attack these problems. What they can do, however, is concentrate on the basic missions of teaching, particularly at the undergraduate level, and encouraging the enrollment of qualified students from minority groups in scientific and engineering curricula. Although this might mean forgoing some of the more prestigious activities, such as symposia and the prolific publishing of papers, it would indeed be an instance of fulfilling the objective set forth in the last paragraph of Prof. Balabanian's letter, that objective being an awareness on the part of circuit theorists and engineers in general of their social responsibilities. It would also be more constructive than introducing the clichés of politics and journalism to the columns of a publication presumably written by and for engineers.

C. H. Wexler
Claymont, Dela.

Concerned scientists

This is to protest in the strongest possible terms against the statement published in the April 1969 issue of IEEE SPECTRUM (page 8) under, of all things, "Concerned Scientists." The IEEE is a nonpolitical organization and there is no room in its publications for open anti-American propaganda. While the register of the outrages these "concerned" people did not pay the slightest attention to would fill pages and pages, I refrain from listing a sample of them because that would be politics too.

I hope that such offensive articles will not appear again in SPECTRUM. If they do, I personally will not be able to maintain my membership.

L. F. Thomay
Montreal, Que., Canada

The biggest issue discussed at the M.I.T. gathering was the deployment of ABM. The biggest enterprise that may confront electrical engineers in the U.S. in the next decade is the ABM. The White House has specifically described the ABM decision as a political one. It is therefore unprofitable to pretend that electrical engineering is isolated from politics.

The Editor

The unsigned article headed "Concerned Scientists" in SPECTRUM for April constitutes an improper intrusion of politics into the pages of a professional journal. We should not permit IEEE and its publications to become still another base of operations for zealots, anxious for politico-economic controversy that requires an understanding of Communism. The charitable view is that the learned gentlemen from M.I.T. apparently see no threat in a revolutionary entity that has as its purpose self-aggrandizement to the point of destruction of this country's institutions.

I am sure that many of us are "concerned scientists and engineers" from the standpoint that we are convinced survival of the United States is at stake. We do not take the adolescent view that if ABM does not work perfectly it should not be built. Some defense, however flawed, is better than complete vulnerability. As for diversion of resources, it could be argued that the people most interested in environmental and social conditions are precisely those who seek to prevent destruction and carnage as a result of enemy action. But, of course, such an argument is lost upon anyone who denies the existence of the threat.

Technologists are free to decline work on any project. The "concerned" men of M.I.T. should exercise that privilege. But, if they do not believe there is an external threat to this nation, they should at least refrain from using SPECTRUM to organize "effective political action." Certainly the directors and editors of IEEE should prevent use of its journals for such a purpose.

M. R. Heembrock
Sunnyvale, Calif.

The M.I.T. faculty group is far from being the only one that is endeavoring to alter the present apportioning of research and development effort between military and nonmilitary projects. It is very unlikely that these efforts will have no effect on the social matrix in which electrical engineering is embedded, and on the professional opportunities open to and desired by electrical engineers. For IEEE to close its eyes and its pages to what is happening would be a disservice to its members.

The Editor

Transients and trends

International telephone, television, and telegraph charges should drop dramatically when the projected Intelsat IV communications satellite starts operating in 1971. This is the opinion of John A. Johnson, international vice president of the Communications Satellite Corporation (COMSAT). "It will be one of the few things in our economy in which we can foresee cost reductions," Mr. Johnson said in a statement issued at the recent Paris Air Show.

Mr. Johnson said that the basis for his statement was the much larger communications capacity of the Intelsat IV satellite. It will have the capability of handling 6000 telephone conversations at one time or 12 television channels compared with the current Intelsat III satellite that can handle 1200 telephone circuits or four television channels.

"The economics are quite simple," Mr. Johnson said. "When Early Bird went into commercial operation, the direct charge for land and earth station connections was about $32 000 a year. We expect the satellite costs for a telephone circuit in Intelsat IV to be about $3000 a year."

A description and drawing of the satellite appears on page 10.

Generating capability of the total electric utility industry in the contiguous United States is expected to reach 464.7 × 10^6 kW by 1974 according to the 45th Semi-Annual Electric Power Survey just published by the Edison Electric Institute. The survey presents data on the nation's electric utility industry as of this April 1.

Assuming median water flow conditions, the forecast represents an increase of 147.8 × 10^6 kW during the five-year span from December 1969 for an average annual increase of 7.9 percent. A 7.2 percent increase would approximate a doubling every ten years.

At year's end, the capability of the total electric utility industry is expected to be 317 × 10^6 kW, an increase of 9.8 percent above the actual December 1968 capability of 288.8 × 10^6 kW.

Output of electric energy by the total electric utility industry in 1974 is projected by the survey to be 2.075 × 10^{12} kWh. This figure exceeds the 1969 output by 644.5 × 10^9 kWh and represents a five-year average annual increase of 7.7 percent. These figures include utility generation plus purchases from outside sources, including net imports.

Total U.S. consumer electronic sales for the first quarter of 1969 show dramatic increases over 1968 figures. A report released by the Electronic Industries Association gives total television sales in the first quarter of 1969 as 3 363 695 sets versus 2 796 074 in 1968. Color television's share reached 1 604 962 sets. Some 2.6 million sets of first-quarter sales were produced in the United States. U.S. manufacturers also imported some 360 000 sets for merchandising under their own labels and non-U.S. label sets amounted to 340 000 units or about 10 percent of the total television market.

The total home radio market in the first quarter (excluding television and phonograph combinations and auto radios) amounted to 8.3 million units compared with 6.2 million units in the same period in 1968. The U.S. manufacturers produced 1.4 million units and imported 1.3 million units for sale under their own labels. Non-U.S. labels, with 5.6 million home radios imported, accounted for 67 percent of sales versus 57 percent in the first quarter of 1968.

Total U.S. sales of auto radios in the first quarter, at 3.3 million units, exceeded the 2.9 million units sold in the same 1968 period. The FM share of radio sales continued to increase. Some 3.7 million home radio sets and 440 000 auto radio sets were sold in the first quarter compared with 2.5 million home and 305 000 auto radio sets in 1968. About 13.4 percent of auto radios were FM compared with 10.4 percent in the first quarter of 1968. Home radio FM sales share (excluding radio–television–phonograph combinations) reached 44 percent, compared with about 40 percent in the same period in 1968.

Total U.S. phonograph sales in the first quarter totaled 1.5 million units compared with 1.4 million in 1968.

Magnetic tape continued to be the fastest growing consumer electronic product in the first quarter. Tape recorders, including reel-to-reel and cassette equipment, totaled 1.2 million units compared with 929 000 in 1968. Tape player imports totaled 801 000 units compared with 391 000 units in 1968. U.S. product figures in this category are not available.

The first commercial nuclear power station for the Netherlands will be built at Borssele near Middelburg. The electric utility company of the Dutch province of Zeeland, the Provinciale Zeeuwse Elektriciteitsmaatschappij, has awarded the Kraftwerk Union Aktiengesellschaft an order for the construction with an option on a planned extension for a second unit. The 400 MW (e) nuclear power station will have a light water reactor and will use enriched uranium for fuel. Construction is to begin at once and the station is scheduled to supply power as early as the spring of 1973 chiefly to the power-intensive industrial plants around Sloehafen on the Westerschelde.

The salary of the hypothetical 'median' engineer, as of late 1968, had risen to $14 800 per year. This figure is given in the latest Engineering Manpower Commission salary survey that shows that professional pay scales have continued their upward trend. The new "median" pay figure represents the sharpest rise since the survey was begun but, to some extent, reflects an increase in the median years of experience to 13.9 because of the growing maturity of the engineering profession. Salaries in most age brackets rose between 4 and 7.5 percent per year since 1966.

The salary study incorporated returns from 979 employers covering 191 000 engineers, making it the most comprehensive survey of engineers' salaries ever conducted. Results are available in a 90-page report, "Professional Income of Engineers 1968–69," which may be obtained for $5 per copy prepaid, and in the more detailed "Engineers' Salaries—Special Industry Report," designed primarily for personnel administrators, at $35 per copy. For either report write to Dept. P, Engineers Joint Council, 345 East 47 St., New York, N.Y. 10017.

A three-year study of the metric system has been given the go ahead by the U.S. Senate. Bill S. 1287 to authorize appropriation to the Department of Commerce of a total of $2.5 million for the study was passed by the Senate on May 14. Last year, the Secretary of Commerce was authorized by P.L. 90-472 to appraise the desirability and practicability of increasing the use of metric weights and measures in the United States.

nology taken together, the number of papers published annually may total two million. They appear in some 30 000 different specialized journals. And the communication process includes not only publication of original findings, but also reviews, surveys, abstracts, indexes, bibliographies, library services, preprints, meetings, and personal contact.

The first SATCOM recommendation calls for the establishment of a Joint Commission on Scientific and Technical Communication, which would report directly to the Councils of NAS and NAE. Its role would be to stimulate greater coordination among private groups and to bring them into closer touch with government agencies. The Academies, with their established prestige, nongovernment status, and close working ties with scientists, private organizations, and federal agencies, are the best base for the Commission, according to SATCOM.

The second recommendation outlines a philosophy of shared responsibility among government and private organizations for the effective communication of scientific and technical data. It points out that those who support research and development work must accept the responsibility to recognize the preparation and dissemination of information as an integral part of such work.

Intrinsic to many of SATCOM's recommendations is the conviction that the scientific and technical societies have a crucial role to play, and SATCOM challenges them to accept greater responsibilities. These include improving the quality, timeliness, and techniques of producing and distributing primary literature; assuring adequate basic abstracting and indexing of the primary information; stimulating the reprocessing and repackaging of information for special user groups; and conducting "exploratory and innovative studies," using qualified scientists, engineers, and practitioners, to evaluate the performance of their information services.

The societies, for instance, should take it upon themselves to encourage the preparation of more critical review articles and data compilations, an effort "which often requires great intellectual creativity." There is great need for such works, which organize and evaluate what is known about a subject and present it in language that can be understood, SATCOM says.

The Committee on Scientific and Technical Communication, a 25-member group under the chairmanship of Robert W. Cairns of Hercules Inc., was established in February 1966 at the request of the National Science Foundation.

Further meetings held by color TV study group

Previous reports covered the November 20, 1968, meeting of the *Ad Hoc* Color Television Study Committee of the Joint Committee on Inter-Society Coordination (JCIC) and field tests in Chicago on December 18 and 19 (see IEEE SPECTRUM for February, page 7, and May, page 10). Reported in the following are the meetings held on January 15 and March 11.

The question of how to interpret the observations made during the Chicago field tests was a major subject of discussion during the January meeting. (Copies of a brief report on the Chicago laboratory tests are available through A. E. Alden, SMPTE, 9 East 41 Street, New York, N.Y.) There was agreement that the results of the portion of the test in which the color signals were produced locally by a laboratory-type transmitter were meaningful. These results should be considered by the appropriate committees of the EIA and of IEEE. It was agreed that the results of the tests including broadcast transmitters could not be interpreted readily, as there was a strong likelihood that the propagation paths from each of the three participating transmitters to the input terminals of the receivers were different in unknown ways, and there were no data on the performance of the equipment between the common signal source and the three radiated signals.

Following discussion of this latter difficulty, W. C. Morrison, representative of the IEEE Group on Broadcasting, was appointed chairman of a task force on transmitters. This task force was charged with the design and conduct of experiments to determine the causes of variations of color observed on receivers arising in the portions of the broadcasting system between the output of "master control" in the studio and the input to the receivers. At the March meeting of the *ad hoc* committee, Chairman Morrison reported that development of a test program was proceeding satisfactorily and that tests would be conducted in Chicago in mid-April.

E. P. Bertero, chairman of the task force on the colorimetry of television camera systems, reported in January

plans for calculating the colorimetric performance of color television systems under different conditions, using the International Commission on Illumination 1964 uniform-color-space coordinate system, U^*, V^*, W^*. Results from such calculations for several camera characteristics, matrixing conditions, and picture-tube phosphors were presented at the March meeting. The results were given as the lengths of vectors in color space, between the color presented to the camera and its reproduction on the picture tube. The calculated results indicate the desirability for changing the matrixing when the picture-tube phosphors are changed from those assumed by the NTSC to those used in modern tubes. The matter will be considered by the EIA Receivers Committee.

C. E. Anderson, chairman of the task force on video tape recording questions, reported that the SMPTE Engineering Committee on Video Tape Recording had undertaken a study of the causes of the variations of color introduced by recorders.

D. M. Zwick, chairman of the task force on motion-picture films, reported that he was setting up a "Standard Review Room" to facilitate studies of motion picture films. He showed the *ad hoc* committee samples of two special test slides he intends to use in his studies. Mr. Zwick reported also that the European Broadcasting Union has issued recommendations for the minimum and the maximum density of color films intended for broadcasting. These EBU recommendations are quite similar to those under consideration in SMPTE engineering committees. He reminded the *ad hoc* committee that there are regional preferences in the reproduction of flesh color. The color preferred on the West Coast of the U.S. is warmer that that preferred on the East Coast, and the taste in Europe is for a still colder reproduction.

Chairman K. B. Benson of the *ad hoc* committee reported that the U.S. CCIR National Organization has formulated a proposal that line 17 of the television frame be reserved for international test signals (vertical interval test signals—VITS) for facilitating the exchange of 525 line programs among nations. This topic is the subject of FCC Notice of Inquiry, Docket no. 18 505, which was adopted by the Commission on March 21.

As a result of the studies of the *ad hoc* committee, several additional sources of variability of color as seen on home re-

Spectral lines

Professionalism in electrical engineering. Electrical engineers consider themselves professionals, and wish to be treated as such. Their path to this kind of recognition is sprinkled with booby traps, of which not the least vexatious is the drive by janitors toward professional status, to be achieved by changing the name "custodian" to "maintenance engineer."

Few other nouns in English are as vague as "professional." In its broadest meaning, it denotes one who takes money for that at which he or she is most accomplished. In its most restricted sense, a profession has lately been an occupation (law, medicine, the ministry, and others) for which one is trained in a graduate school after taking a bachelor's degree in the liberal arts.

Most U.S. engineers who speak of professional status for themselves probably do not have in mind either of these extremes. It is a curious thing that in the 19th century, when the U.S. was an underdeveloped nation, a liberal-arts education was highly prized, both for its own sake and as preparation for the professions. Now that the country is highly developed, we hear that a liberal-arts education is a luxury that today's students cannot afford.

Going back to origins of words is not always useful, but in this case I think it is. The great *Oxford English Dictionary* tells us that originally (in the 13th century) a "profession" was a declaration, promise, or vow made by one entering a religious order. As such, it connoted discipline, devotion to an ideal, and complete dedication of self. Its first extension (appropriately enough, in view of its significance) was to the military, "the profession of armes." Then came the "learned professions" as a term applied most notably to law, divinity, and medicine. We now have a spectrum of professions, with these three at one end, professors and city planners in the middle, and engineering, alas, somewhere near the other end.*

It seems to me that after all these centuries, the marks of a profession are still discipline, devotion to an ideal, and complete dedication of self. To the extent that he can claim these attributes, an actor can fairly call his work a profession. So can a clergyman. So can an engineer.

A different criterion is that professionals are those who win status through the esteem of their fellow practitioners, rather than through the approval of those whom they serve. For salaried men, "those whom they serve" presumably includes company management. Here we have one reason—perhaps the major reason—why engineering is not universally regarded as a profession. Professionals at the high end of the spectrum possess more freedom of action than a salaried man can easily get. Another characteristic of the top of the scale—incidental, perhaps, rather than essential—is that high-prestige professionals deal with the affairs of people rather than with the design and fabrication of things.

When all this has been said, we are still convinced that electrical engineering is to be classified not as a trade, but as a profession. How can we get nonengineers to see the light?

One move that receives frequent mention is to unionize. I see no common ground between unionism and professionalism. A profession exalts individual merit; a union exalts the welfare of the hive. In a profession, advancement comes through personal struggle for recognition, whereas in unions the respectable route to advancement is seniority. To opt for unionism is to renounce all claim to being a professional.

Nevertheless, this point must not be dismissed summarily. The engineers who think of unionism are very likely the men—many of them in the U.S. aerospace industry—who are hired by the hundred when a contract comes in, and laid off the same way when the contract terminates. It seems to me that these men have a real and clamant grievance. True, when jobs fold in one area, other jobs have opened up elsewhere. But who wants to spend four years or more in a university in order to become a migrant? Who wants to keep on buying a house in a new community when the demand is high, and selling it a few years later when the demand is zero? As a first step in eliminating such hardship and indignity, the IEEE should set up a committee to make an extended study of this problem and to report on ways of alleviating it.

There is evidence that few electrical engineers in the U.S. yearn to be unionized, but that many would like to see a strong drive—spearheaded, perhaps, by IEEE—toward organizing the profession along lines similar to those of the American Medical Association. This complex possibility merits attention here in a coming month.

J. J. G. McCue

* *Daedalus*, Fall 1963, was an issue on "The Professions"; it does not treat engineering as a profession at all.

Electronically expanding the citizen's world

Representatives of government, business, and the academic world got together at the IEEE 1969 International Convention's Highlight Session to discuss technology and its relationship to the man in the street

James D. O'Connell *Special Assistant to the President for Telecommunications*

Eugene G. Fubini *Consultant*

Kenneth G. McKay *Vice President Engineering, American Telephone and Telegraph Company*

James Hillier *Executive Vice President—Research and Engineering, RCA Corporation*

J. Herbert Hollomon *President, University of Oklahoma*

Is electronics extending itself to increase citizen opportunities for education, entertainment, and a better life? During World War II, electronics expanded to meet U.S. defense needs. Korea accelerated the advance of the technology even more. In all the unsettled years since, electronics technology has been facing new challenges. The Highlight Session at this year's IEEE International Convention and Exhibition served as a platform where some men with very broad horizons could speak, and others could listen and learn. Three of the speakers agreed that we are just on the threshold of the age of technical developments. The fourth maintained that the future of electronics and its relationship with society will hardly be determined by electronics at all.

Hon. James D. O'Connell

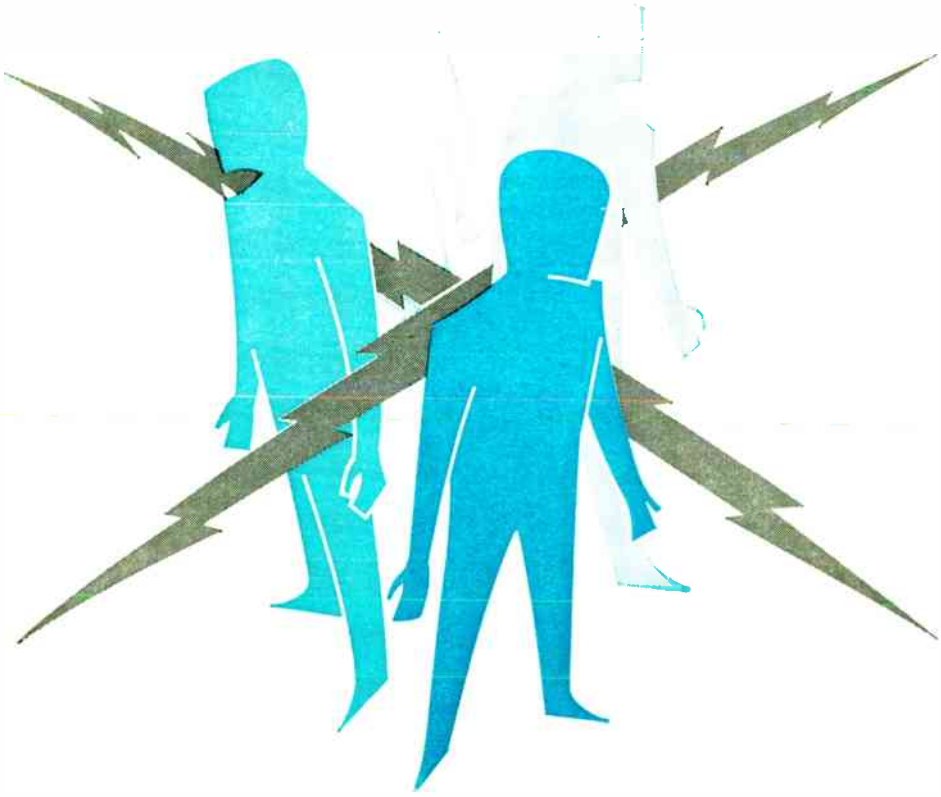
Ladies and gentlemen, I would like to start by asking Dr. Fubini to present his ideas on this subject.

Dr. Eugene G. Fubini

My purpose today is to talk to you about the effect computers will have on our lives, and I intend to begin with a declaration of faith in this future, I will however put emphasis on the limitations of the predictions that are current on the ability of the computer to penetrate our lives, rather than add predictions of my own to those that you read in the newspapers and popular literature. But this means I will have less fun, and I'll leave most of the fun to the people who will talk after me. I do feel that, from time to time, we need to take stock of the things we read and determine how valid they are.

The first question is: "Will the computer become an intimate part of our lives?" I think the answer is "yes." But this is not yet so today: Many of my friends have heard me make an analysis of the various stages of development of major technologies and I would like to briefly repeat it here. Major technologies have three stages: stage one, when you do what you used to do only you do it better; stage two, when you do new things you never did before; and stage three, and this is the important one, when you change your lives to match the new capability that the technology gives you.

Now, with the internal combustion engine we are clearly in the third stage. The supermarket is a necessary consequence of the invention of the engine, and so are the suburbs, the ghettos, and the social ills. I'll give you a very simple method to determine whether a technology is in the third stage. Take a magic wand and stop all the internal combustion engines in the United States and see what happens: life would not continue in the form we have in the U.S. today. But the computer is not in the



same stage yet. Stop all the computers in the United States and see what happens: The airplanes will be a little bit later than they should be. Your salaries will be paid with a certain delay. Your bank statements may be a little bit late also, but otherwise, I don't think your lives would change very much. You haven't modified yet your lives to match the availability of computers. We are not in the third stage with the computers. And let's see—let us try to ask ourselves—what does it take to get there?

Well, before the computer truly enlarges the citizen's world in the way we are talking about today, I think that first of all we have to change the way we do software. We must remove the barrier that separates the untrained man from the computer. We must make it possible for everyone to communicate with a computer, and we must increase computer literacy in our schools. We must bring down the cost of developing the software. Software is the only material good that I know in which the unit cost goes up with the size of production—and up quite a bit, as a matter of fact. We don't yet have a scientific approach to software generation and we need to find it. We must go from the art of software to the science of software and we have a long way to go.

We must also learn to measure the performance of computers; we really don't do it well at all. Computers are sold on size of memory and speed rather than on the ability to do a certain type of work. We have not yet found a good way of measuring "throughput." We must also increase their reliability and think of computers not as machines that are in the basement, but as devices that provide a service to a community and cannot fail. Let me assure you that this is an important change in the frame of mind of today's designers and manufacturers. We must achieve reliabilities that are unknown to computers today, but are feasible. We must make our terminals foolproof but inexpensive. And by the way, since I sit near Ken McKay, I should say that I think the videophone that he may be talking about may become such a terminal, but I would like to add to it a keyboard entry and a hard-copy output to make it a terminal that I would like to buy.

I'm fully convinced that all of these things will come true, and that we will soon be immersed in that third stage of the computer in our lives. But I'm also convinced that the two most common examples that are reported by all the newspapers and the magazines about this immersion need to be considered more carefully, and I'm talking about the "checkless" society and the computers in education.

Technological forecasting has become a very common form of entertainment these days, one which I enjoy also. I note that Wall Street is doing the job of technical forecasting with gusto. But one can hardly open a newspaper today without reading the promise of a new computerized

world. I agree, as you heard, but I'd like to make some critical remarks regarding these predictions.

When I was an Assistant Secretary of Defense I discovered a very important law. It's called "The American Syndrome." The American Syndrome when applied to defense problems was stated as follows: "If you can do it, do it." It was a jocular reproach to those people who felt that if a device could be built, one should build it, even if no clear operational or cost advantage would ensue from the construction of the device. I regret that I see the same general syndrome in some of our technological forecasts. And it is clear at the end of this 20th century that technologies can be found to do just about anything. The problem is not "how" to do something, but "what" to do. Our forecasters forget that not all possible events will occur. A computer in fact can eliminate checks and cash by recording all the financial transactions, and as a consequence we immediately see in the press the promise of a society that has neither cash nor checks. If technical feasibility were the only ingredient needed, these forecasts would be right. Unfortunately or fortunately, many other ingredients are necessary, and if you own stock in a factory that prints checks, I suggest you keep that stock. Also, I don't think you should burn all the cash.

There is a second pitfall that our forecasters tend to fall into. They tend to extrapolate linearly and neglect the fact that technologies saturate with time. We have improved the speed of computer elements by a factor of five every ten years; but we changed technologies four or five times. We went to vacuum tubes—they saturated; we went to transistors—they saturated; we are now going with integrated circuits and they'll saturate; and we are seeing large-scale integration on the horizon but we already know saturation will occur. I had a computation made that convinces me, and maybe it would convince you, that our present semiconductor technologies will saturate at a factor of ten—only ten—better than the best that's available in the laboratory, and a factor of about 100 from the best available in production.

To summarize then, forecasting will be wrong if one assumes that things will happen just because they can happen; if one neglects the fact that technologies saturate; and that for this reason short-range predictions tend to be optimistic; we must also remember, conversely, that the long-range predictions tend to be pessimistic, because brand-new needs and brand-new technologies are difficult to forecast. For example, take the "checkless" society.

To make it come true, we need a centralized computer with a number of terminals, but the terminals must be justified costwise. The saving on the handling of the checks doesn't save enough money to justify the plan. Furthermore, the customer must be willing to forgo the advantage of charge accounts in returning merchandise, and must forgo the advantage of canceled checks as receipts. I'm not convinced that all these will be acceptable in the society of the future. So I really invite you, before you predict the "checkless" society, to consider what use you make of checks, and not to assume that technical feasibility is sufficient for realization.

Let me talk about education. We have all heard of the tremendous future of data processing in education. And in fact, computers have been used to solve problems, to make drill-and-exercise painless, to simulate physical experiments, to teach students, to manage the movement of students in the different steps of the educational process. Yet the fact that the computer can be used for all these things is no guarantee that this will occur in large scale.

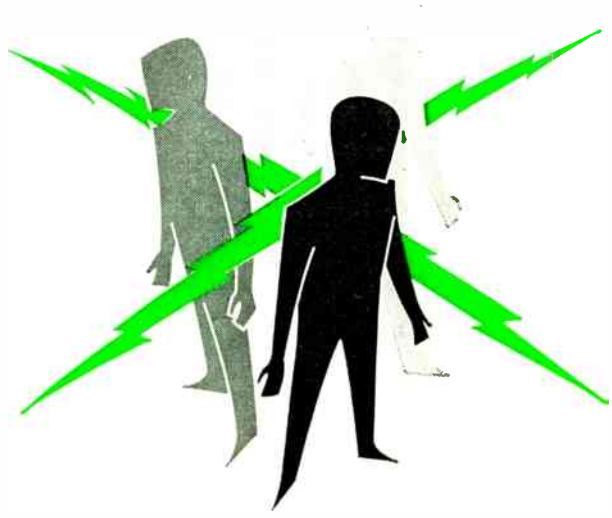
I would like to call your attention to the numerous mergers between companies in the electronics business and book publishers, and I note that the success of these mergers has not been outstanding. I don't intend to discuss this problem here but I urge you to keep separate, when you deal with computers in education, the different elements of the problem. Let us separate secondary education from industrial education, for instance, and let us separate problem solving from tutorial interaction.

In problem solving, the computer has changed our world. It has changed the meaning of the word "solving." In order to solve a problem we only need to write the rule that makes the solution possible; then we can let the machine do the job that we used to do. That has changed the character and the level of what we are doing.

In the tutorial interacting computerized system of instruction we have a cost problem. The computer costs between $1 and $3 per student contact hour, and in the typical secondary school there is not enough money to pay for this. We may note that only very rich secondary schools spend more than $5 per student-year on books, audiovisual aids, and technological devices. The cost of tutorial interaction is acceptable then only in industrial education and in military education. The different elements of the computer in education will interact differently in our society of the future.

I predict that we will see two things coming to the fore: problem solving and simulation, and that computer-assisted instruction will drag.

Before I conclude, I would like to give an example of a computer simulation that has had a great impact on me and explains why I think this is one of the most powerful elements in education. This is the experiment: You take a computer and send a simulated stream of particles from the left into a cathode-ray tube, and a stream of particles from above, into the same cathode-ray tube. All these particles are the same size. You then tell the computer to



say that these particles attract each other with Newton's law, and the ordinary laws of mechanics apply. And then you let the experiment run. It's breathtaking to see what happens. I can't forget it. Right in front of you the universe is forming. You see spiral galaxies coming out; you see all the types of galaxies that are present in the world today—they are being created in front of you through this simple mechanism. I never really understood how a galaxy was formed until I saw that simulation. And it is with experiments of this kind that I think we can do a type of education that we have never been able to do before. And I'm thoroughly convinced of the impact that these things have.

In conclusion, let me reaffirm my convictions. We are going to see the computer penetrate our lives just like television or the internal combustion engine or the jet aircraft. We have not yet solved all the problems that stand in the way of this achievement. But the solution is both possible and near. We are going to see the electronic data-handling devices appear in force; they are neither obvious nor expected today. There is no doubt that we will see computers force us to modify our lives.

Hon. James D. O'Connell

Thank you very much, Dr. Fubini. Next I'd like to ask Dr. McKay if he would give us his views on this evening's subject.

Dr. Kenneth G. McKay

Ladies and gentlemen: Tonight we're indulging in futurology. This activity puts a communications engineer in somewhat of a peculiar position because in a sense virtually everything we have thought of can be done, and you have to stretch for exceptions. I mean, we're not quite ready to put on three-dimensional television but holography may fix that up for us, and you can see that the extreme exceptions tend to emphasize the extraordinary range of audio, visual, and other information that can be delivered anywhere on demand. So this question of what can be done is essentially trivial. The real question is, "What probably will be done?" This isn't just a technical problem; it hinges rather heavily on a blend of technology and economics and human desire. For example, it's now technically possible for NASA to put every person in New York City into orbit around the moon. I suggest that both economics and human desire negate this achievement, at least for the near future. This is my response to Dr. Fubini's American Syndrome.

Let's see where we are today, before we peer into the future. Three principal modes of electrical communication stand out: there's interactive communications—personal, private, connected, two-way; there's spectator communications or entertainment, which is primarily one-way, single-source, and multisink. And there's instructive communication, which can be a combination of these. Obviously there's some overlap between the modes. For example, a multiparty telephone line might be thought of as combining the interactive and the entertainment modes.

Let's look at interactive communications first. There we have already achieved one of the highest states of development that Dr. Fubini was referring to. We do have over 100 million telephones in the United States—one for every two persons. Of course, the user sees only the end instrument. What really makes it useful are the transmission facilities and the switching centers that form the network. We have extended the use of these facilities to computers, and their monotonous but accurate vocabularies do enter the network through more than about 85 000 data sets; they range in speed up to some 2000 bits per second on the switched voice network. In a few months we'll introduce something higher—about 3600 bits per second on the switched network.

What can the citizen expect from interactive communications in the future? My most obvious prediction is a continued increase in the number and the accessibility of telephones, in a very wide variety of forms. I think they'll appear in some unaccustomed places. Right now we're a nation that's distinguished for its mobility, and yet we're rather restricted in our ability to communicate while in transit. Plans are currently under way that, if they come to fruition, will provide for a major increase in the use of mobile telephones—in automobiles and private boats, as well as on commercial aircraft and on trains. This hinges upon the effective exploitation of the frequencies above 900 megahertz. You know, one of the attractive features of the high-speed train between New York and Washington is on-board dial telephone service, and the traveling public apparently really likes it.

I think generally you can see ways in which we can gain greater access to the network, and I'm going to couple this with the fact that we're going to be putting in more and more in the way of high cross-section coaxial cable, and ultimately waveguide, producing a predictable decrease in the cost of long-distance transmission. This simply means that we'll have a deeper permeation of our affairs with telephone communications.

Now let's switch to data—primarily used now by government and buisness concerns. If computers are really going to be useful to the citizen, they have to be accessible to him where he is—in his home, in his business, wherever he happens to be. This really calls for an extension of the combination of computers and communications. Right now, you can call up a computer and you can ask it a question, using Touch-Tone buttons, and you can receive an answer by computer-controlled voice; many business firms are doing this today. This suggests that if you want to keep your household accounts this way—this will be a way you can do it. Of course you may want to carry on a more sophisticated dialogue with your friendly neighborhood computer and make a record of what he says, and at this point we move into some slightly more sophisticated keyboard instruments. For example, teletypewriters. Although old hat for business, they're still somewhat costly for the household; but the costs have been going down, and I think this will continue as household markets develop.

So I think that accessible computers can not only cope with your income tax, but provide you also with immediate information on a variety of items that are of immediate interest—like ski reports, or vacation reservations, together with the probability of what the weather's going to be at that time. Actually the likelihood of realization of this kind of information and these information services hinges largely on the cost of collecting the data itself.

We have yet to add a most important aspect to interactive communications—the visual mode. Really, today's telephone leaves us living in a world of the blind, and we hope to change this with visual telephone or Picture-

phone, which restores vital face-to-face dimensions to communications.

Right now we're involved in some trials with Westinghouse that appear quite promising. We've improved the Picturephone set, which you may have seen some years ago at the New York World's Fair or in cartoons in the *New Yorker* magazine. The new set uses double the bandwidth—about one megahertz—and this does permit good resolution for face-to-face communication. It's not as good as a well-adjusted television set, but it's adequate. Now of course the Picturephone set is like a miniature television studio, but that's just the set itself. To provide service we require a complete switched network for signals of one-megahertz bandwidth. Over short distances we propose to send signals in analog form; over all long-distance transmission, the signals will be in digital form. Obviously we use a great deal of equipment that we already have in the present network, but we need a good deal more of it, and this means that the cost of Picturephone service will be correspondingly higher. So we expect initially it will be used primarily for business purposes—just as the telephone was when it was first introduced. Then later on it will begin to appear in the home when the president of the company decides he'd like to see what's going on there too.

Incidentally, some of you ladies in the audience may reject Picturephone lest it peer at you when you don't want to be seen. These fears are groundless. There's a handy button on the set that permits the viewer to see who's calling while the caller sees nothing. So milady can always choose whether she can be seen or not.

Just as telephones can interact with computers, so can Picturephones, and this is now echoing a point that Dr. Fubini brought up—that computers can be programmed to answer in graphical language just as well as in spoken language. The resolution we will get on Picturephone, of course, is not as good as on a TV screen or on the CRT displays that are commonly used for output devices with computers, but simple graphs can be adequately displayed and we are investigating the whole range of possible applications. When you call Weather Information you might also be able to see a weather map.

Now in your point about hard copy, Dr. Fubini, I think this is an interesting one. I can suggest a half-way measure; I think it's easy to connect on a video recorder and you have a semihard copy. You can play it back and see what it was you were seeing.

Dr. Fubini. I was trying to save money, and the video recorder is really too expensive for my hard-copy reproducer.

Dr. McKay. That's something we can bring up later.

An important point here is that the same transmission and switching facilities that will carry Picturephone signals can also be used for other purposes, such as high-speed data and facsimile. Here is an asset whose potential is yet to be explored. When we realize the existence of a geographically dispersed switched network capable of carrying signals of one-megahertz bandwidth, or digital information of the order of one megabit per second, it should open up a very wide range of applications. And these really have not been properly considered up till now—I think primarily because of the relatively high cost of the terminal equipment.

Now I put this to you: this is a challenge to the entire electronics industry—to provide equipment for broadband application that the householder can afford.

Let me shift briefly to the field of spectator communications. I think the most probable advance will be the availability of more television channels to the user through cable television. This does satisfy the desire for a broader choice of programs, but doesn't accomplish the further goal of providing a given program to a specific user at the time of his choosing. There are different ways you can do this. One would be to reserve one of the channels in the cable television system for your exclusive use for a period of time, and call up a central tape file and get your particular program played just for you. Or conversely, you could tape the program automatically on a video recorder in your home, when it's being presented, put up your own tape file, and then play it back at your convenience. Just for what it's worth, I think the home recorder will be more economical in the long run.

Much has been written about the potential benefits of domestic satellites to provide nationwide distribution of TV. My present view is that the costs of transmission are completely overwhelmed by the cost of providing the programs. And so for the user himself, this distinction between transmission modes really loses relevance.

Now, I'd like to talk briefly about the matter of instructive communications. Everything that can be spoken or drawn or printed or seen can be communicated electrically, and these are also the modes that are used for knowledge to be transmitted from one generation to another. I feel that this application of communications—to education—has possibly the greatest unrealized potential impact, as far as the citizen is concerned. Of course one reason we apply communications rather gingerly in the instructive field is that to be successful we really need a rethinking of the basic purposes and the philosophy of education. After all, what facets of the human mind and body are to be subject to the educational process?

I think the social aspects of education, for example, seem to require some direct interaction among individuals. Nobody's going to qualify for the football team via Picturephone. So let's just concentrate for a moment on the acquisition of knowledge and the intellectual interplay between teacher and student.

The vast lecture hall seems rather impersonal. This is a far cry from the tutor-pupil relationship, and now the question is, can we restore this through the use of a knowledge machine—a computer? We can store a large amount of knowledge here, and we can produce it in accord with the individual student's ability to absorb it. One suspects that this judicious blending of computers and communications can enable a student to learn at the time and place of his own choosing. I'm talking about the possible, but what is probable? As you know, there are experiments being carried out in many parts of the country to try to evaluate the usefulness of computer-aided education. This is, I think, going to be a very lengthy task because it's complicated by a wide spectrum of environmental and sociological conditions. But we do have, at least here and there, some promising results in terms of at least a deeper understanding on the part of the students. We need these evaluations of effectiveness, and then we must return to the actual cost. I think it's still too early to prejudge these because they do involve the overall cost of education, but we can see some clues right now. Obviously a single computer is large enough and expensive enough that you must be able to share

it over a large number of students; you really must be in the time-shared mode with communications. You can see some other things such as the fact that when you're using communications, bandwidth still costs money and so you don't use television bandwidths when data bandwidths on voice channels would be adequate.

So you can build up a sort of scenario of little hints as to how to go about this, but I think that it's going to be a substantial period of time before we can focus in on all of the various modes of communication and how computers and communications can be most effectively used.

Finally, I think we ought to avoid one pitfall, and that is we must not think of the communications technology as being mature and without future surprises. Integrated circuits are complicated, but their evolution is progressing and at least a part of every new piece of communications equipment that's under development does involve integrated circuits or LSI. The obvious goals are lower costs and smaller sizes, but the most significant goal is increased reliability, because this will enable effective operation of much more complex systems than we now contemplate.

I'm going to combine that with computer simulation of these systems in the hope that we can actually design them so they will work, and if I put these things together I think we will design systems that are highly complex and yet be able to produce them economically.

These are some of the patterns to be woven by communications into the lives of our people. They'll permit the citizen greater freedom of choice in where he lives, where he works, what he sees, and what he learns. Ready access to computers will reduce complexities to simplicities. They're all possible; to make them probable we must choose to have them.

Hon. James D. O'Connell

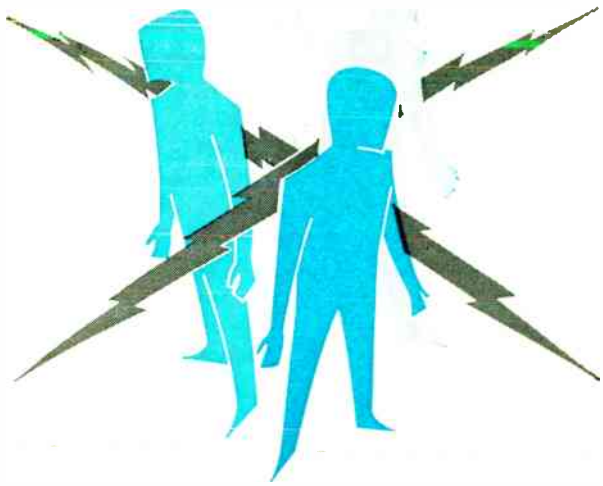
Thank you very much Dr. McKay. I'm going to resist the temptation to get some discussion started on the panel and give you an opportunity to hear the other distinguished gentlemen that we have here before opening it up for further discussion. I'd like to introduce Dr. Hillier.

Dr. James Hillier

My assignment in this program tonight is to extend our discussion of electronically expanding the citizen's world to the consumer, and more specifically to the consumer in the home.

As the title implies, I am supposed to polish a little smog off my crystal ball and give you my version of what the electronic future might hold for us ordinary citizens. However, before we do that, it might be a good idea to take a quick look at where we are.

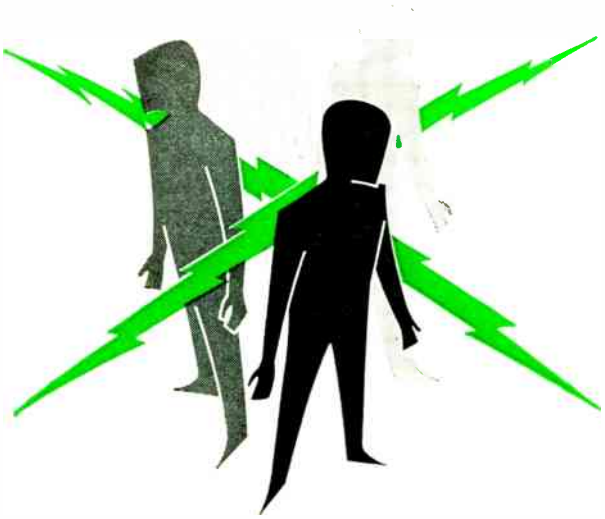
I suspect, if you throw the concept of electronics in the home to the man on the street, his immediate reaction will be to think of television, and then he'll add the radio and the record player, and if he thinks a little more, he will probably add the telephone. While the telephone goes back to nonelectronic antiquity, he will realize that the steady increase in service that has occurred during his lifetime has been largely based on electronics. If a man has a somewhat technical inclination, a bit more thought might add a host of home appliances that also had their origins before electronics, but which in recent years have become more sophisticated through the gradual addition of electronic devices. Finally, if our man is really aware of



the world around him, he will realize that he is immersed in an invisible but pervasive electronics environment with which he interacts continuously. Every action he takes is dependent on or sets in motion a stream of electronic impulses that controls the manufacturer of the products he buys, controls and guides the transportation he uses, keeps tabs on his financial transactions, provides much of the communications service through which he learns of the options available to him, expresses his wishes, and maintains control of his own destiny.

I characterized this part of our man's electronic environment as "invisible" because it is something he can appreciate intellectually but which does not reach strongly into his awareness as he lives his life. A large part of that electronic environment does no more than let the world operate about the same as it always has, by compensating for the greater numbers of people and the much greater complexity of our lives. As a result of this part of electronics, our man sees only slow change. Our man also has seen relatively little new that electronics has offered him for his home during the past 20 years. With the possible addition of color television he has the same boxes and the same services that he had shortly after World War II. It is true that radios and tape players have become very small, inexpensive, and very portable, and that television is following in the same course. This is a result of the solid-state revolution in electronics technology. But the services are still the same.

The most significant point of awareness of the growth of electronics for our man has probably been when the television screen has turned away from "show biz" to become a window on the world. Then he cannot really miss the fact that now he is seeing events in Paris or Tokyo as they happen. A very dramatic demonstration of this point indeed was the flight of Col. Borman and his crew around the moon last Christmas. While these very men were being transported physically through space to the vicinity of the moon, some 230 000 miles distant, all the world was accompanying them—vicariously, but nevertheless, accompanying them—through the agency of electronic communications.



Thus, my view of electronics and the consumer in his home contains some peculiar contrasts and contradictions. I see a rapidly changing and advancing electronics industry, contrasted with an almost static picture of electronics in the home, or at best, a very slowly evolving one. I see a large part of our advances in electronics, exclusive to a certain extent of defense electronics, being used to compensate for the ever-increasing load being thrown on our commercial, manufacturing, and financial institutions by our increasing population and by the compounding complexity generated by that population and by the increasing options that individuals have available to them.

Parenthetically, I suspect that if we pursue this line of thinking a little bit further, one conclusion we might reach is that the major impact of electronics on the man in the street has been an increase in his paycheck; that's a real way of "electronically expanding the citizen's world."

But let me return to my assigned theme. At this point I have to ask myself some questions. Why does the consumer in the home seem to be the forgotten man? Why haven't we offered him some different electronic boxes in the past 20 years? And another good question is, perhaps, what kind of a consumer is he today?

One thing is clear: many of us have been busy colorizing and transistorizing, while the rest of us have been preoccupied with the developing technologies and with the development of large systems for industry and for defense. And if you don't believe that, a quick look at the roster of the IEEE Groups is one of the best confirmations I can give you.

However, our preoccupation with other matters is not the whole story. The task of introducing a new box into the home is an exceedingly complex one. Only two options seem to be available to us. One is by the radical change of, or addition to, the existing boxes. The other is by the introduction of a new service. The basic purpose of most electronics systems is to provide a service—a service in which the largest ingredient is information. This is true whether we are discussing television or the navigation of a jet plane by satellite.

When we sell or rent our consumer one of our present boxes, we are really providing him with the means of access to a service that he desires or needs, a service that is already in existence and for which the cost—as he appraises it—is reasonable. This is certainly true for television, radio, phonograph, and the telephone group. When we talk about additional electronic boxes in the home, we must talk about extensions of the present services or the addition of completely new services. The new ones must satisfy the same conditions of desire, cost, and availability. There is another characteristic of these services that is worth noting. That is, that the services we now provide electronically to the consumer at home were not completely new when they were introduced. They were all convenient extensions and improvements of services already available and being used. The messenger, the mails, the telegraph—preceded the telephone. The lecture hall, the concert hall, the theater, and the movies all preceded radio and television. I see no signs that we're at the point of violating this trend in the future, even though I also see no basic reason why we could not.

These thoughts begin to reveal the rationale of my view of the future development of electronics in the home. I believe we should think more in terms of new services rather than new boxes. I believe that the new services will be extensions of existing services; that the electronic versions will be more current, more immediate, more convenient, ultimately lower in cost, and available to more people. Note that I'm changing, for some of you, perhaps, the rules of the game. Instead of saying "Let's invent new boxes," I'm saying "Let's entrepreneur new services." Obviously I'm turning up the contrast to make a point. I recognize that it is possible to invent a new box that will catalyze the entrepreneuring of a new service. Also the entrepreneuring of a new service will have a high technical content and will probably require the invention, or at least the development, of new boxes as it proceeds.

I hardly need to add that entrepreneuring a new service is orders of magnitude more difficult than inventing a new box. It is not surprising that our consumer has not been offered one recently, but we must offer him one soon if the consumer electronics business is to break out of its pattern of maturity.

We can find some more clues about the future if we take a brief look at our consumer. We'll recognize he has more leisure time and more money to spend, but his extra money has also given him more options as to how he spends his leisure time. This one point has significance with regard to our consumer's use of the existing services and the opportunities to provide new ones. Note, however, that we may find electronics either participating in his options or competing with them.

Another significant point is that while our consumer's information supply has been increasing very rapidly, so also have his needs. Since he is an integral part of his information system and has a very limited bandwidth, he is going to have to exercise greater selectivity as to the information he seeks and accepts. In this I believe he will put a premium on ease and speed of access. This signifies an opportunity for electronics.

I believe you will generally agree that with color television we reached the end of the major quantum jumps of bringing entertainment into the home electronically. Some further refinements such as high-fidelity television or 3-D television may evolve if some of the

standards and spectrum problems and the economic problems can be solved. However, even if these do come, they will not represent a major change in the concept of the television service.

We do see some possibilities of other dimensions that fit our picture of the consumer. For instance, there are gadgets that let you show color slides or home movies on your color television screen. There's an element of the "do-it-yourself" concept in these approaches. These will certainly appeal to the man who has everything. They also could be quite successful on a much broader base if our consumer decides that the additional convenience is worth the extra cost.

Note that in these devices we have added possibilities for electronics in the home by utilizing only a part of the television system; namely, the display. Carrying this further, many people are thinking of providing packaged video programs that can be displayed on the television tube. This is the television counterpart of the phonograph record that we tend to call the video record system.

When a video record is developed it will fit our pattern of an additional service quite well. While it is true that anything that could be put on a video record could also be put on regular television, there are good reasons for believing that the programs available on video records will cover a wider range of entertainment, edification, and education. The video record will give the consumer the opportunity to select more precisely the program he wants, and give him complete flexibility as to the timing of his viewing. You will have noted that I included edification and education in the repertoire of video records. I have a personal theory that says that people are much more selective, as to timing as well as to subject matter, when they are seeking edification or education-type programs. If valid, this would explain at least some of the difficulties educational television channels are having. With very few channels in any one location, the probability of a hit—that is, an encounter of the right program at the right time for a specific individual—is much too small. With video records the probability of such a hit will be determined only by the size of catalogs of available programs, which can continually expand. In fact, it is not difficult to go further and visualize the video record as the replacement of the library book. But we might ask the question as to when.

Technically the video record has been here for some time. Many companies have been aiming at this market with video tape players suitable for the home. They've had some success in the educational and particularly the training environment, but so far their success in home use has been rather limited.

More recently we have had demonstrations of a new video record system known as the EVR system. While this seems to be aiming for the home market eventually, the immediate target is also the educational market. I believe the real problem to be solved in the video-record system concerns the cost of the record to the consumer. The consumer has developed a pattern of spending for entertainment which it seems to me is giving us rather good guidance if we'd listen to it. When our consumer plays his television set, which he does at some length, he is paying a few pennies an hour for his entertainment. When he goes to the movies, the ball game, the hockey game, the price goes to a little over a dollar an hour, and so on. It's always in that range. I don't believe the consumer is ready to pay $25 or $40 an hour for entertainment, and certainly not for blank magnetic tape.

Obviously I believe the cost of the video record has to come down sharply before we will have a product. Even when we accomplish that, we still have to entrepreneur the total service—all the way from the recording of the program through to the consumer's player. It's not going to be easy, but here is a large area of opportunity.

Let me conclude my remarks by turning for a couple of minutes to one other very large area of opportunity. And I'm thinking of a whole complex of possible home or business information services. For instance, I believe that it is inevitable that you and I will ultimately replace the card punch operator who presently exists at several points in the chain of events initiated by the transactions we make. The gasoline pump operator has already reached that stage. He is only one of the early warning signs. But let's look on this positively and indicate the entrepreneurial opportunities to provide our consumer with a host of desirable information services.

Now, I'm an independent inventor—I'm going to say "Let us dream of a black box full of electronics." I shall impose one condition on my dream: everything in that box is available today, or is so close to being available that we can assume that it is. I will not spoil my dream by mentioning costs.

The box looks familiar. It has an alphanumeric keyboard and a kinescope display. It has a couple of additional features; there's a lens system that looks out from a small television camera, and a slot that will eject sheets of paper with printing on them. It has all the presently available inputs—a television antenna, a cable television connection, a Picturephone line, and one or more standard telephone lines. Now, just think of all the fun you can have with this gadget. Right now you could use it as a time-sharing terminal; or you could use it for television or telephone facsimile; and very shortly, you could use it as a Picturephone.

Then if you could develop the services, you could use it for broadcast facsimile, either on a television signal or on a cable television channel. And you would have an enormous selection of classes of information there; electronic delivery of the newspaper is the way the news people usually talk of this, but I'm thinking of a much broader service. It could provide electronic mail service eventually. It would also give you access to all forms of information banks as they are developed. In fact, this one box could provide essentially all the information services ever envisioned by the science fiction writers. Yet I have stipulated that everything in the box is technically available today.

Why don't we have it, or parts of it? Two reasons: first, we have to entrepreneur the services so that they are available and in a form that the consumer wants and can afford; second, we have to solve along with that the chicken-and-egg problem between entrepreneuring the service and the reduction of cost of the equipment. This is a very large area of opportunity.

It's obvious from my brief remarks that I believe there are many opportunities to expand the citizen's world by electronics—many more than I've mentioned here tonight. They are not exclusively technical challenges, but they will require substantial entrepreneurial skills and very creative ones. Finally I've emphasized that in think-

James D. O'Connell (F) attended the University of Chicago and received the B.S. degree from the United States Military Academy in 1922. He did graduate work at Northwestern University and received the M.S. degree in communications engineering from Yale University in 1930. Since May 1964, he has been Special Assistant to the President for Telecommunications/Director of Telecommunications Management. He retired from the U.S. Army with the rank of lieutenant general. During World War II, he served in North Africa and in Europe; and in Japan following the cessation of hostilities. From 1951 to 1959, he was Deputy Chief Signal Officer and then Chief Signal Officer of the Army. As such he was responsible for tactical and strategic military communications operations and research and development of U.S. military communications equipment. Prior to assuming his present position, Mr. O'Connell was a consultant in communications to Stanford Research Institute, Page Communications Engineers, Northrop Corporation, Granger Associates, Data Dynamics, Inc., and Fred W. Morris, Jr. & Associates. From 1959 to 1964 he was a member and chairman of the Joint Technical Advisory Committee, IEEE. He is also a member of the Armed Forces Communications and Electronics Association. Mr. O'Connell is the author of numerous papers on military and satellite communications. He was awarded the Army's Distinguished Service Medal for meritorious action.

Eugene G. Fubini (F), until recently IBM vice president and group executive responsible for the company's Research Division and the Advanced Systems Development Division, is now consultant to IBM and other elements of the electronics industry. Prior to joining IBM, Dr. Fubini served as Assistant Secretary of Defense Research and Engineering for the U.S. government. Born in Turin, Dr. Fubini was educated in Italy. He attended the Turin Technical Institute and received the doctorate in physics from the University of Rome in 1933. During World War II, Dr. Fubini served as a scientific consultant and technical observer to the U.S. Army and Navy in the European Theater of Operations. After the war, Dr. Fubini joined Airborne Instruments Laboratory, Melville, N.Y., as an engineer in 1945, and in 1960 was appointed vice president of the Research and Systems Engineering Division of the AIL Division of Cutler-Hammer Corporation. In March 1961, Dr. Fubini joined the Office of Defense Research and Engineering of the Office of the Secretary of Defense. In June 1963, while serving as Deputy Director of Defense Research and Engineering, he was nominated Assistant Secretary of Defense. Dr. Fubini is a member of the Scientific Advisory Board of the Air Force, the Defense Science Board, the National Academy of Engineering, and chairman of the Scientific Advisory committee of the Defense Intelligence Agency. He has received the Presidential Certificate of Merit and the Defense Medal for Distinguished Public Service.

Kenneth G. McKay (F) has been vice president-engineering of the American Telephone and Telegraph Company since December 1966. Dr. McKay joined Bell Telephone Laboratories, Inc., in 1946 where he did research into the physics of solids, including studies of secondary electron emission and electron bombardment conductivity in insulators and semiconductors. He was associated with Bell Laboratories in various capacities for over 20 years. He was named director of development of solid State devices in 1957; excutive director of components and solid State devices in 1958, and vice president in charge of systems engineering in 1959. He was elected executive vice president of the Laboratories in 1962. Dr. McKay was graduated from McGill University, where he received the B.S. and M.S. degrees in 1938 and 1939. He received the Ph.D. degree from the Massachusetts Institute of Technology in 1941 and, for the next five years, worked with the National Research Council in Canada. Dr. McKay has written extensively on solid-state physics for scientific publications. He is a Fellow of the American Physical Society. He is also a member of the Research Society of America and the National Academy of Engineering.

James Hillier (F), executive vice president, Research and Engineering, RCA Corporation, first came into prominence for his contribution to the development of the electron microscope and for his role in encouraging the growth of electron microscopy as a research technique in biology, medicine, chemistry, and other sciences. Dr. Hillier is well known in the field of research management. Dr. Hillier studied at the University of Toronto, where he received the degrees of B.A. in mathematics and physics in 1937, M.A. in physics in 1938, and Ph.D. in physics in 1941. In 1953 he was appointed director of the Research Department of Melpar, Inc. A year later he was named administrative engineer, research and engineering, at RCA. In November 1955 he was appointed chief engineer, RCA Industrial Electronic Products. In 1957 he became general manager of RCA Laboratories and a year later vice president. Dr. Hillier is a Fellow of the American Physical Society, the American Association for the Advancement of Science, and an Eminent Member of Eta Kappa Nu, as well as past president of the Electron Microscope Society of America. He is a member of the Commerce Department's Technical Advisory Board and chairman of the Advisory Council of the Electrical Engineering Department of Princeton University.

J. Herbert Hollomon became the eighth president of the University of Oklahoma on July 1, 1968, following the retirement of Dr. George Lynn Cross. As president-designate, he initiated and directed a comprehensive study of the university and its future by some 600 faculty and staff members students, and community leaders in Oklahoma and other states. Dr. Hollomon came to the university from Washington, D.C., where since May 1962 he had been Assistant Secretary of Commerce for Science and Technology. Dr. Hollomon was graduated from Augusta Military Academy in Fort Defiance, Va., in 1936, and received from the Massachusetts Institute of Technology the B.S. in science in 1940 and the Ph.D. in metallurgy in 1946. In 1946 he became a research associate in the General Electric Company Research Laboratory, later rising to assistant to the manager of the Metallurgy Research Department, and in 1952 to manager. While with General Electric, Dr. Hollomon was an adjunct professor of metallurgy and served on the science development council of Rensselaer Polytechnic Institute. He has also held advisory posts at Harvard University, Cornell University, George Washington University, and M.I.T. Included among his many honors is the Rosenhain Medal of the British Institute of Metals. He was the first U.S. recipient of this award. He is also a member of the Federal Council for Science and Technology, a consultant to the President's Science and Advisory Committee and a member of the National Advisory Heart Council. On January 13, 1969, he was elected president of the Mid-America State Universities Association.

A guided tour of the fast Fourier transform

The fast Fourier transform algorithm can reduce the time involved in finding a discrete Fourier transform from several minutes to less than a second, and also can lower the cost from several dollars to several cents

G. D. Bergland *Bell Telephone Laboratories, Inc.*

For some time the Fourier transform has served as a bridge between the time domain and the frequency domain. It is now possible to go back and forth between waveform and spectrum with enough speed and economy to create a whole new range of applications for this classic mathematical device. This article is intended as a primer on the fast Fourier transform, which has revolutionized the digital processing of waveforms. The reader's attention is especially directed to the IEEE Transactions on Audio and Electroacoustics for June 1969, a special issue devoted to the fast Fourier transform.

This article is written as an introduction to the fast Fourier transform. The need for an FFT primer is apparent from the barrage of questions asked by each new person entering the field. Eventually, most of these questions are answered when the person gains an understanding of some relatively simple concept that is taken for granted by all but the uninitiated. Here the basic concepts will be introduced by the use of specific examples. The discussion is centered around these questions:

1. What is the fast Fourier transform?
2. What can it do?
3. What are the pitfalls in using it?
4. How has it been implemented?

Representative references are cited for each topic covered so that the reader can conveniently interrupt this fast guided tour for a more detailed study.

What is the fast Fourier transform?

The Fourier transform has long been used for characterizing linear systems and for identifying the frequency components making up a continuous waveform. However, when the waveform is sampled, or the system is to be analyzed on a digital computer, it is the finite, discrete version of the Fourier transform (DFT) that must be understood and used. Although most of the properties of the continuous Fourier transform (CFT) are retained, several differences result from the constraint that the DFT must operate on sampled waveforms defined over finite intervals.

The fast Fourier transform (FFT) is simply an efficient method for computing the DFT. The FFT can be used in place of the continuous Fourier transform only to the extent that the DFT could before, but with a substantial reduction in computer time. Since most of the problems associated with the use of the fast Fourier transform actually stem from an incomplete or incorrect understanding of the DFT, a brief review of the DFT will first be given. The degree to which the DFT approximates the continuous Fourier transform will be discussed in more detail in the section on "pitfalls."

The discrete Fourier transform. The Fourier transform pair for continuous signals can be written in the form

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft}\,dt \tag{1}$$

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{i2\pi ft}\,df \tag{2}$$

for $-\infty < f < \infty$, $-\infty < t < \infty$, and $i = \sqrt{-1}$. The uppercase $X(f)$ represents the frequency-domain function; the lowercase $x(t)$ is the time-domain function.

The analogous discrete Fourier transform pair that applies to sampled versions of these functions can be written in the form

$$X(j) = \frac{1}{N}\sum_{k=0}^{N-1} x(k)e^{-i2\pi jk/N} \tag{3}$$

$$x(k) = \sum_{j=0}^{N-1} X(j)e^{i2\pi jk/N} \tag{4}$$

for $j = 0, 1, \cdots, N-1$; $k = 0, 1, \cdots, N-1$. Both

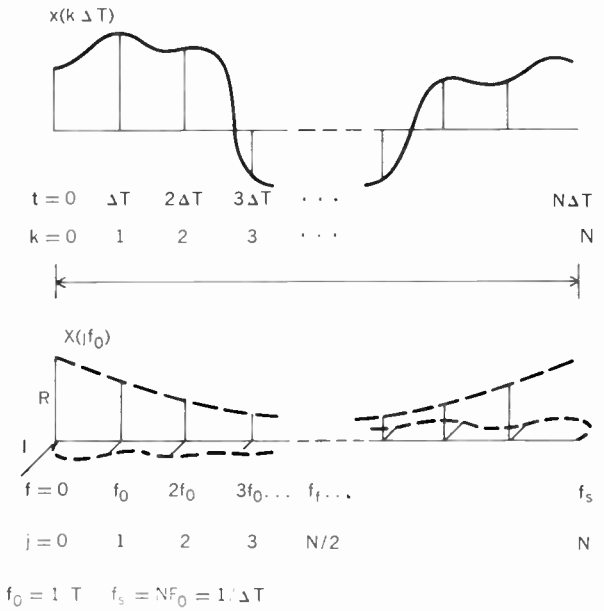
$X(j)$ and $x(k)$ are, in general, complex series. A derivation of the discrete Fourier transform from the continuous Fourier transform can be found in Refs. 12 and 23.

When the expression $e^{2\pi i/N}$ is replaced by the term W_N, the DFT transform pair takes the form

$$X(j) = \frac{1}{N}\sum_{k=0}^{N-1} x(k)W_N^{-jk} \tag{5}$$

$$x(k) = \sum_{j=0}^{N-1} X(j)W_N^{+jk} \tag{6}$$

FIGURE 1. A real signal and its complex discrete Fourier transform displayed in the FFT algorithm format.



An example of a real-valued time series and its associated DFT is shown in Fig. 1. The time series $x(k\Delta T)$ is assumed to be periodic in the time domain of period T seconds, and the set of Fourier coefficients $X(jf_0)$ is assumed to be periodic over the sample frequency f_s. Only one complete period of each function is shown.

The fundamental frequency f_0 and the sample period ΔT do not appear explicitly in Eqs. (5) and (6), but each j should still be interpreted as a harmonic number and each k still refers to a sample period number. That is, the true frequency is the product of j and f_0 and the true time is the product of k and ΔT.

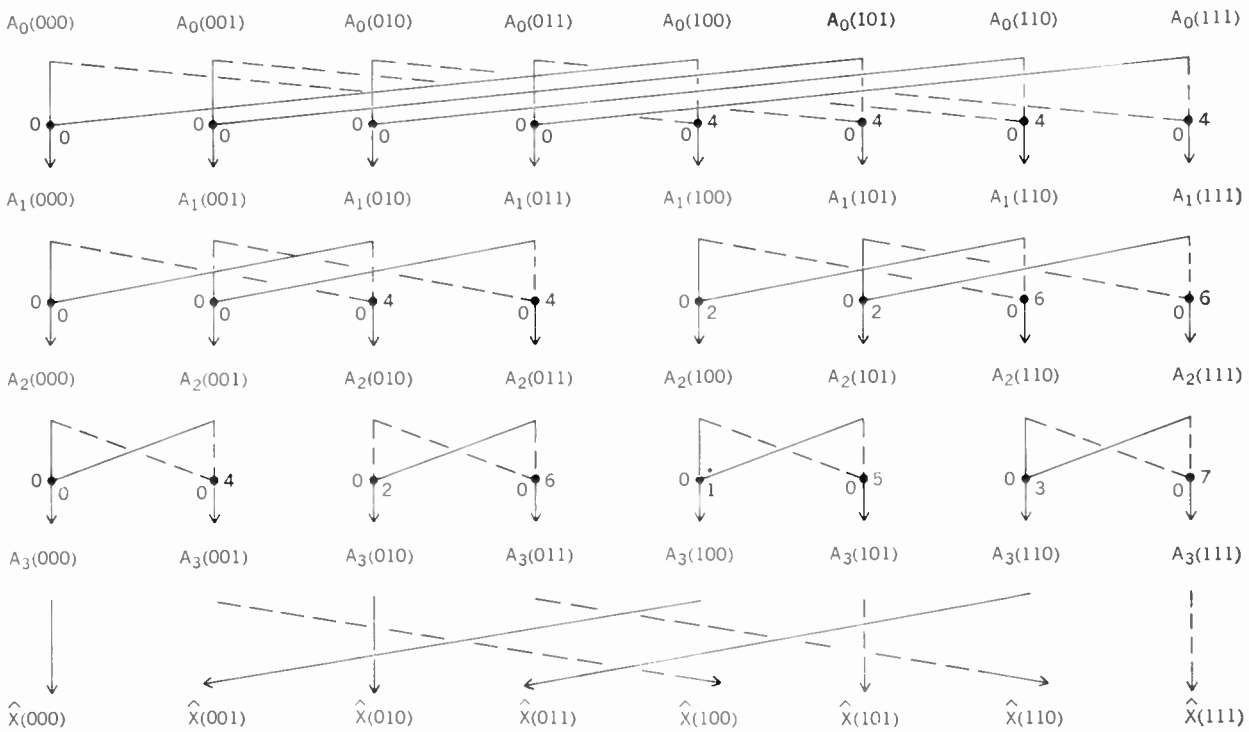
When the $x(k)$ series is real, the real part of $X(j)$ is symmetric about the folding frequency f_f (where $f_f = f_s/2$) and the imaginary part is antisymmetric. Since $X(j)$ has been interpreted as being periodic, these symmetries are equivalent to saying that the real part of $X(j)$ is an even function, and that the imaginary part of $X(j)$ is an odd function. This also means that the Fourier coefficients between $N/2$ and $N-1$ can be viewed as the "negative frequency" harmonics between $-N/2$ and -1. Likewise, the last half of the time series can be interpreted as negative time (that is, as occurring before $t = 0$).

Derivation of the Cooley–Tukey FFT algorithm. A derivation of the Cooley–Tukey FFT algorithm[8] for evaluating Eq. (6) is given in this section for the example of $N = 8$. This derivation is also appropriate to the forward transform, since Eq. (5) can be rewritten in the form[9]

$$X(j) = \frac{1}{N}\left[\sum_{k=0}^{N-1} x(k)^* W_N^{jk}\right]^* \tag{7}$$

where the asterisk refers to the complex conjugate operation. Alternatively, the FFT algorithm used for comput-

FIGURE 2. A flow diagram of the Cooley–Tukey FFT algorithm for performing an eight-point transform.



ing Eq. (6) can be altered by redefining W_N to be exp $(-2\pi i/N)$ and dividing each result by N.

Using Cooley's notation,[8] the FFT algorithm involves evaluating the expression

$$\hat{X}(j) = \sum_{k=0}^{N-1} A(k)W^{jk} \tag{8}$$

where $j = 0, 1, \ldots, N-1$, and $W = \exp(2\pi i/N)$. Note that $\hat{X}$ and A can be interpreted as X^* and x^*/N, respectively, if the forward transform is being computed, and can be interpreted as x and X, respectively, if the inverse transform is being computed.

When N is equal to 8, it is convenient to represent both j and k as binary numbers; that is, for

$$j = 0, 1, \cdots, 7 \qquad k = 0, 1, \cdots, 7$$

we can write

$$j = j_2 4 + j_1 2 + j_0 \qquad k = k_2 4 + k_1 2 + k_0 \tag{9}$$

where j_0, j_1, j_2, k_0, k_1, and k_2 can take on values of 0 and 1 only. Using this representation of j and k, Eq. (8) becomes

$$\hat{X}(j_2,j_1,j_0) = \sum_{k_0=0}^{1}\sum_{k_1=0}^{1}\sum_{k_2=0}^{1} A(k_2,k_1,k_0)W^{(j_2 4+j_1 2+j_0)(k_2 4+k_1 2+k_0)} \tag{10}$$

Noting that $W^{m+n} = W^m \cdot W^n$, we have

$$W^{(j_2 4+j_1 2+j_0)(k_2 4+k_1 2+k_0)} = W^{(j_2 4+j_1 2+j_0)k_2 4}W^{(j_2 4+j_1 2+j_0)k_1 2}W^{(j_2 4+j_1 2+j_0)k_0} \tag{11}$$

If we look at these terms individually, it is clear that they can be written in the form

$$W^{(j_2 4+j_1 2+j_0)k_2 4} = [W^{8(j_2 2+j_1)k_2}]W^{j_0 k_2 4} \tag{12}$$

$$W^{(j_2 4+j_1 2+j_0)k_1 2} = [W^{8j_2 k_1}]W^{(j_1 2+j_0)k_1 2} \tag{13}$$

$$W^{(j_2 4+j_1 2+j_0)k_0} = W^{(j_2 4+j_1 2+j_0)k_0} \tag{14}$$

Note, however, that

$$W^8 = [e^{2\pi i/8}]^8 = e^{2\pi i} = 1 \tag{15}$$

Thus, the bracketed portions of Eqs. (12) and (13) can be replaced by a one. This means that Eq. (10) can be written in the form

$$\hat{X}(j_2,j_1,j_0) = \sum_{k_0=0}^{1}\sum_{k_1=0}^{1}\sum_{k_2=0}^{1} A(k_2,k_1,k_0)W^{j_0 k_2 4}W^{(j_1 2+j_0)k_1 2}W^{(j_2 4+j_1 2+j_0)k_0} \tag{16}$$

$A_1(j_0,k_1,k_0)$

$A_2(j_0,j_1,k_0)$

$A_3(j_2,j_1,j_0)$

In this form it is convenient to perform each of the summations separately and to label the intermediate results. Note that each set consists of only eight terms and that only the latest set needs to be saved. Thus the equations can be rewritten in the form

$$A_1(j_0,k_1,k_0) = \sum_{k_2=0}^{1} A(k_2,k_1,k_0)W^{j_0 k_2 4} \tag{17}$$

$$A_2(j_0,j_1,k_0) = \sum_{k_1=0}^{1} A_1(j_0,k_1,k_0)W^{(j_1 2+j_0)k_1 2} \tag{18}$$

$$A_3(j_0,j_1,j_2) = \sum_{k_0=0}^{1} A_2(j_0,j_1,k_0)W^{(j_2 4+j_1 2+j_0)k_0} \tag{19}$$

$$\hat{X}(j_2,j_1,j_0) = A_3(j_0,j_1,j_2) \tag{20}$$

The terms contributing to each sum are shown in Fig. 2. Each small number refers to a power of W applied along the adjacent path. The last operation shown in Fig. 2 is the reordering. This is due to the bit reversal in the arguments of Eq. (20).

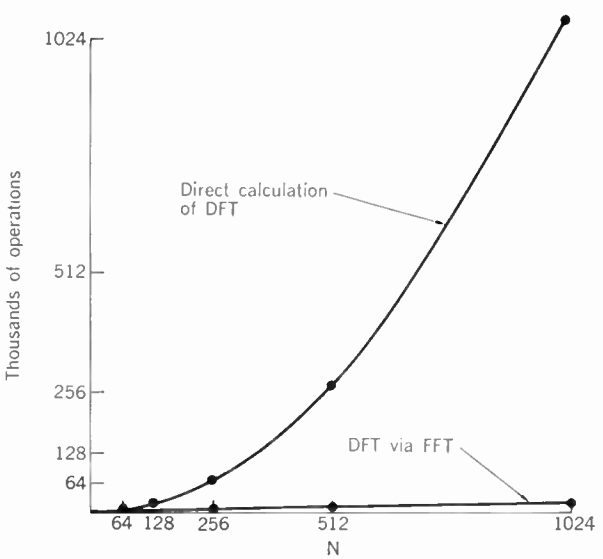
This set of recursive equations represents the original Cooley–Tukey formulation of the fast Fourier transform algorithm for $N = 8$. Although a direct evaluation of Eq. (8) for $N = 8$ would require nearly 64 complex multiply-and-add operations, the FFT equations show 48 operations. By noting that the first multiplication in each summation is actually a multiplication by $+1$, this number becomes only 24. By further noting that $W^0 = -W^4$, $W^1 = -W^5$, etc., the number of multiplications can be reduced to 12. These reductions carry on to the more general case of $N = 2^m$, reducing the computation from nearly N^2 operations to $(N/2)\log_2 N$ complex multiplications, $(N/2)\log_2 N$ complex additions, and $(N/2)\log_2 N$ subtractions. For $N = 1024$, this represents a computational reduction of more than 200 to 1. This difference is represented graphically in Fig. 3.

What can it do?

The operations usually associated with the FFT are: (1) computing a spectrogram (a display of the short-term power spectrum as a function of time); (2) the convolution of two time series to perform digital filtering; and (3) the correlation of two time series. Although all of these operations can be performed without the FFT, its computational savings have significantly increased the interest in performing these operations digitally.

Spectrograms. The diagram in Fig. 4 represents a method of obtaining estimates of the power spectrum of a time signal through the use of the fast Fourier transform.

FIGURE 3. The number of operations required for computing the discrete Fourier transform using the FFT algorithm compared with the number of operations required for direct calculation of the discrete Fourier transform.



In this case the square of the magnitude of the set of complex Fourier coefficients (that is, the periodogram) is used to estimate the power spectrum of the original signal.

A snapshot of the spectrum of the signal can always be computed from the last T seconds of data. By taking a series of these snapshots, estimates of the power spectrum can be displayed as a function of time as shown in Fig. 5. When the audio range is displayed, this is usually called a sound spectrogram or sonogram.

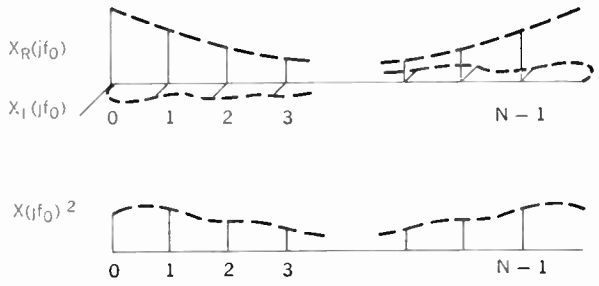


FIGURE 4. The power spectrum of a real function computed by taking the sum of the squares of the real and imaginary components of the discrete Fourier transform Fourier coefficients.

FIGURE 5. A spectrogram made by using the fast Fourier transform algorithm.

In some cases it is of interest to go one step further. When the spectrum of a signal contains a periodic component, this spectrum can be compressed by taking the logarithm, and then the fast Fourier transform can be taken again. The result is called a cepstrum (pronounced "kepstrum").[13,19] An example of using the cepstrum to determine the pitch period of a speaker is described in Ref. 14. For a more complete discussion of short-term spectrum and cepstrum analysis see Refs. 10–21.[13-17]

Digital filtering. In a linear system, one is frequently confronted with the problem of either (1) determining the output, given the input and the impulse response, or (2) finding the impulse response, given the input and the output. Both of these problems can be approached rather easily in the frequency domain.

In Fig. 6, the output $c(t)$ is formed by convolving the input $g(t)$ with the impulse response of the system $h(\tau)$. For sampled functions this convolution takes the form

$$c(k) = \frac{1}{N}\sum_{\tau=0}^{N-1} g(\tau)h(k-\tau) \tag{21}$$

This equation represents the linear system in Fig. 6, as long as $h(\tau)$ is assumed to be zero for $\tau < 0$. If both $g(k)$ and $h(k)$ consist of N consecutive nonzero samples, the series $c(k)$ can consist of $2N-1$ nonzero terms.

Since the FFT algorithm gives us a fast way of getting to the frequency domain, it is interesting to consider

$$C(j) = G(j)\cdot H(j) \tag{22}$$

where $C(j)$ is the DFT of $c(k)$, $G(j)$ is the DFT of $g(k)$, and $H(j)$ is the DFT of $h(k)$. By Eq. (6), this is equivalent to

$$c(k) = \sum_{j=0}^{N-1} [G(j)\cdot H(j)]W^{jk} \tag{23}$$

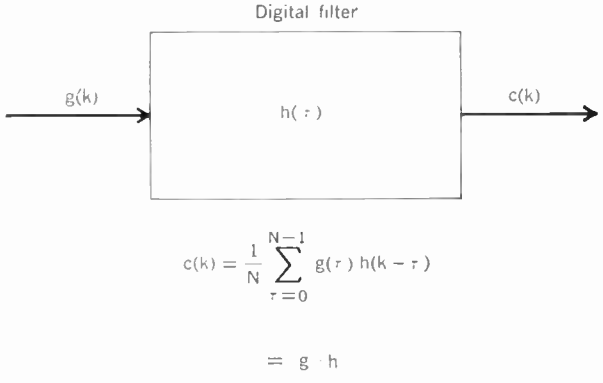
By Eq. (5), this can be written as

$$c(k) = \sum_{j=0}^{N-1}\left(\frac{1}{N}\sum_{\tau=0}^{N-1} g(\tau)W^{-j\tau}\right)\left(\frac{1}{N}\sum_{\hat{\tau}=0}^{N-1} h(\hat{\tau})W^{-j\hat{\tau}}\right)W^{jk} \tag{24}$$

Since all of the sums are finite, this can be rewritten as

$$c(k) = \frac{1}{N}\sum_{\tau=0}^{N-1}\sum_{\hat{\tau}=0}^{N-1} g(\tau)h(\hat{\tau})\left[\frac{1}{N}\sum_{j=0}^{N-1} W^{+j(k-\tau-\hat{\tau})}\right] \tag{25}$$

FIGURE 6. The response of a linear system to a driving function g(k) expressed as the convolution of the input signal with the impulse response of the system.



This can be simplified through the use of the orthogonality relationship

$$\sum_{j=0}^{N-1} W_N^{nj} W_N^{-mj} = N \qquad \text{if } n = m \bmod N$$

$$= 0 \qquad \text{otherwise} \tag{26}$$

Thus Eq. (25) is equal to zero unless $\hat{\tau} = k - \tau$, for which we have

$$c(k) = \frac{1}{N} \sum_{\tau=0}^{N-1} g(\tau) h(k - \tau) \tag{27}$$

This equation is identical to the desired Eq. (21) but the requirement that $h(\tau)$ be zero for $\tau < 0$ is not met. In representing $g(k)$ and $h(k)$ by their Fourier coefficients the assumption is made that they are periodic functions. A method of sidestepping this problem is described later.

Correlation. By considering the equation

$$C(j) = G(j) \cdot H^*(j) \tag{28}$$

and following a development similar to that of the preceding section, we obtain the result

$$c(k) = \frac{1}{N} \sum_{\tau=0}^{N-1} g(\tau) h(\tau - k) \tag{29}$$

This is the form of the cross-correlation function of $g(k)$ and $h(k)$. When $h(k) = g(k)$ we obtain the autocorrelation function. The problem again is that both $g(k)$ and $h(k)$ were assumed to be periodic in finding their Fourier coefficients. This problem is discussed further in the next section.

From Eqs. (29) and (27), it is clear that convolution is simply the process of correlating one time series with another time series that has been reversed in time.

What are the pitfalls?

The three problems most often encountered in using the discrete Fourier transform appear to be aliasing, leakage, and the picket-fence effect. Also of interest are the problems associated with the blind use of Eqs. (22) and (28) to perform convolutions and correlations. An ever-present problem, which is not discussed here, concerns finding the statistical reliability of an individual power spectral estimate when the signal being analyzed is noise-like. A discussion of this last problem can be found in Refs. 11, 16, 17, 19, and 21.

Relating the DFT to the CFT. Most of the time, engineers are interested in the discrete Fourier transform only because it approximates the continuous Fourier transform. Most of the problems in using the DFT are caused by a misunderstanding of what this approximation involves.

In Fig. 7, the results of the DFT are treated as a corrupted estimate of the CFT. The example considers a cosine-wave input, but the results can be extended to any function expressed as a sum of sine and cosine waves.

Line (a) of Fig. 7 represents the input signal $s(t)$ and its continuous Fourier transform $S(f)$. Since $s(t)$ is shown to be a real cosine waveform, the continuous Fourier transform consists of two impulse functions that are symmetric about zero frequency.

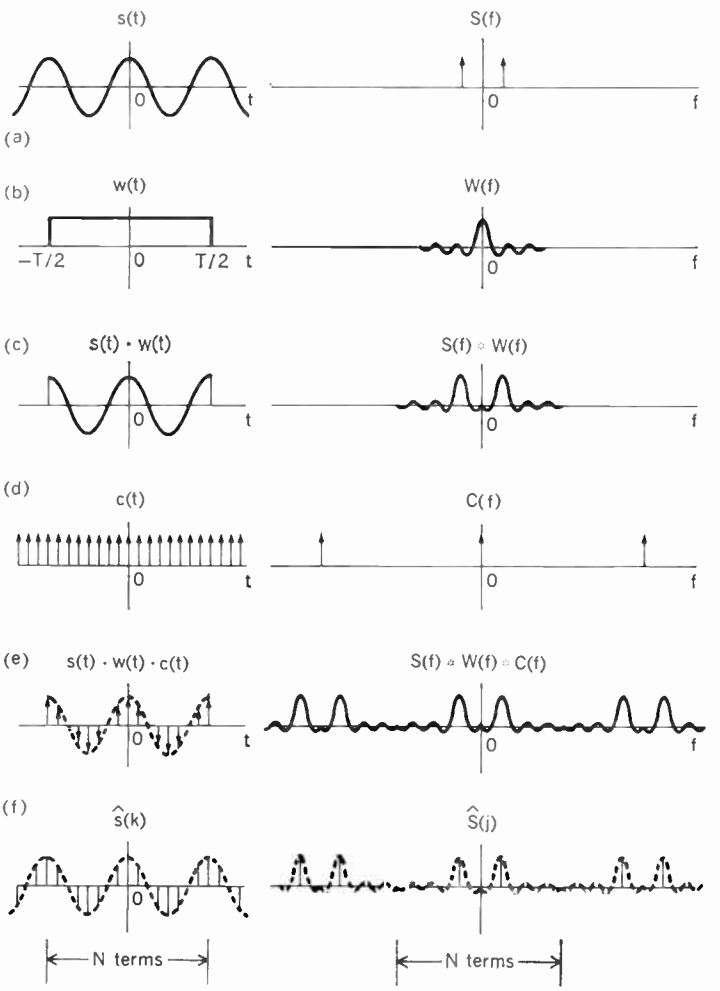
The finite portion of $s(t)$ to be analyzed is viewed through the unity amplitude data window $w(t)$. This rectangular data window has a continuous Fourier transform, which is in the form of a $(\sin x)/x$ function. [When this function takes the form $(\sin \pi x)/\pi x$, it is referred to as the sinc x function.[2]] The portion of $s(t)$ that will be analyzed is represented as the product of $s(t)$ and $w(t)$ in line (c) of Fig. 7. The corresponding convolution in the frequency domain results in a blurring of $S(f)$ into two $(\sin x)/x$-shaped pulses. Thus our estimate of $S(f)$ is already corrupted considerably.

The sampling of $s(t)$ is performed by multiplying by $c(t)$. (In Ref. 12 this infinite train of impulses is called a Dirac comb.) The resulting frequency-domain function is shown in line (e).

The continuous frequency-domain function shown in line (e) can also be made discrete if the time function is treated as one period of a periodic function. This assumption forces both the time-domain and frequency-domain functions to be infinite in extent, periodic, and discrete, as shown in line (f).

The discrete Fourier transform is simply a reversible mapping of N terms of $\hat{s}(k)$ into N terms of $\hat{S}(j)$. In this example, the N terms of $\hat{s}(k)$ and $\hat{S}(j)$ approximate $s(t)$ and $S(f)$ extremely well. This is an unusual case, however, in that the frequency-domain function of line (e) of Fig. 7 was sampled at exactly the peaks and zeros. The problems

FIGURE 7. The Fourier coefficients of the discrete Fourier transform viewed as a corrupted estimate of the continuous Fourier transform.



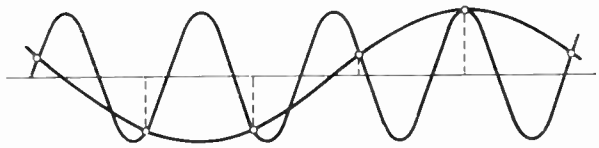


FIGURE 8. An example of a high frequency "impersonating" a low frequency.

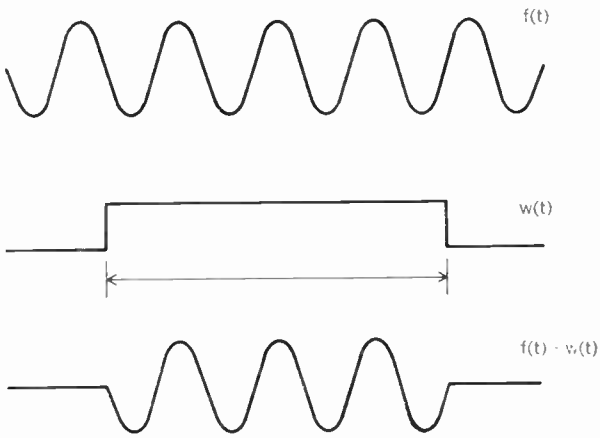
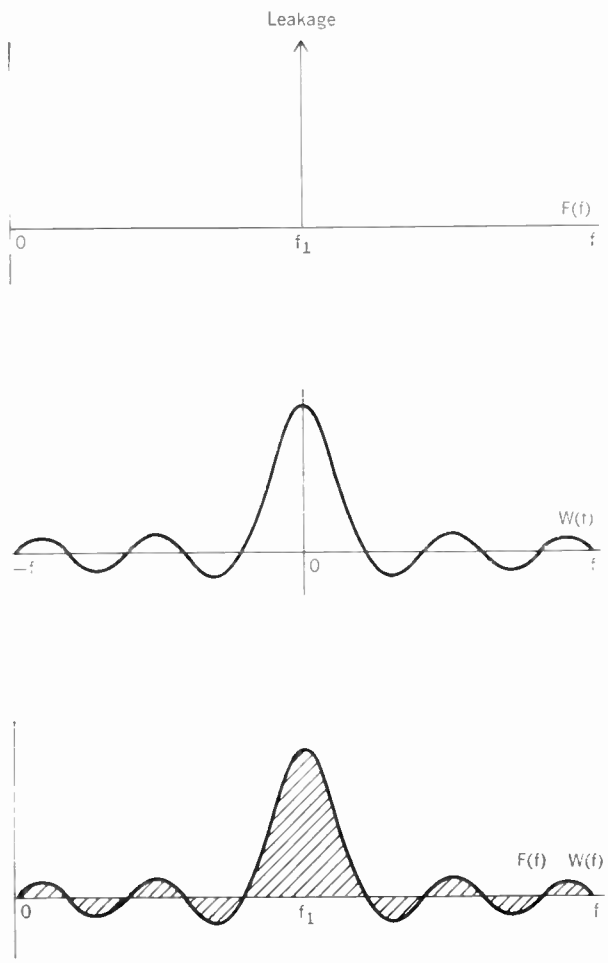


FIGURE 9. The rectangular data window implied when a finite record of data is analyzed.

FIGURE 10. The leakage of energy from one discrete frequency into adjacent frequencies resulting from the analysis of a finite record.



of leakage, aliasing, and picket-fence effects are associated with variations from this ideal condition.

Since both $\hat{s}(k)$ and $\hat{S}(j)$ are periodic, with a period N, any set of N adjacent terms can be used to characterize either set. The magnitude of $\hat{S}(j)$ is the same for samples of $\hat{s}(k)$ that are symmetric about the origin as it is for the DFT of a set of samples starting at the origin. The change in the phase of $\hat{S}(j)$ for different time origins can be determined by application of the DFT shifting theorem.[9]

Aliasing. The term "aliasing" refers to the fact that high-frequency components of a time function can impersonate low frequencies if the sampling rate is too low. This is demonstrated in Fig. 8 by showing a relatively high frequency and a relatively low frequency that share identical sample points. This uncertainty can be removed by demanding that the sampling rate be high enough for the highest frequency present to be sampled at least twice during each cycle.

In the diagram of Fig. 1, the sampling frequency f_s is shown to be $1/\Delta T$ samples per second. The folding frequency (or Nyquist frequency) f_f is shown to be equal to $f_s/2$. In this example, an input signal $x(t)$ will be represented correctly if its highest frequency component is less than the folding frequency. A frequency component 5 Hz higher than the folding frequency will impersonate a frequency 5 Hz lower than the folding frequency. Thus any components in $x(t)$ that are higher than the folding frequency are aliased (or folded) into the frequency interval below the folding frequency. In Fig. 7, aliasing occurs when $S(f)$ extends over a wider range of frequency than one period of $\hat{S}(j)$.

The cure to this problem involves sampling the signal at a rate at least twice as high as the highest frequency present. If the signal is known to be restricted to a certain band, the sampling rate can be picked accordingly. If the signal has been passed through a low-pass filter, a sampling rate can be chosen so that the components above the Nyquist frequency are negligible. Aliasing is discussed in more detail in Refs. 1–3, 10, and 12.

Leakage. The problem of leakage is inherent in the Fourier analysis of any finite record of data. The record has been formed by looking at the actual signal for a period of T seconds and by neglecting everything that happened before or after this period. As shown in Figs. 7 and 9, this is equivalent to multiplying the signal by a rectangular data window.

Had the continuous Fourier transform of the pure cosine wave in Fig. 9 been found, its contribution would have been limited to only one point on the frequency axis. (This is represented by the impulse at frequency f_1 in Fig. 10.) The multiplication by the data window in the time domain, however, is equivalent to performing a convolution in the frequency domain. Thus this impulse function is convolved with the Fourier transform of the square data window, resulting in a function with an amplitude of the $(\sin x)/x$ form centered about f_1.

This function is not localized on the frequency axis and in fact has a series of spurious peaks called sidelobes. The objective is usually to localize the contribution of a given frequency by reducing the amount of "leakage" through these sidelobes.

The usual approach consists of applying a data window to the time series, which has lower sidelobes in the frequency domain than the rectangular data window. This is

analogous to weighing the outputs of a linear antenna array to reduce the sidelobe levels of the antenna pattern.

A host of different data windows have been discussed in the literature; see Refs. 11, 12, 19, and 21. An example of Tukey's "interim" data window is shown in Fig. 11. In this window, a raised cosine wave is applied to the first and last 10 percent of the data and a weight of unity is applied in between. This window was suggested with reservations in Refs. 11 and 19. Since only 20 percent of the terms in the series are given a weight other than unity, the computation required to apply this window in the time domain is relatively small. Another window that can be applied conveniently is the Hanning window, described by Blackman and Tukey.[12] This window is a cosine bell on a pedestal, but it can be applied by convolving the DFT coefficients with the weights $-\frac{1}{4}$, $\frac{1}{2}$, and $-\frac{1}{4}$.[11]

When the computation required is not the overriding consideration, one can find a number of windows that give rise to more rapidly decreasing sidelobes than the cosine tapers described previously. An application of the Parzen window (which has a shape of $1 - t$, for $-1 \leq t \leq 1$) is described in Ref. 21, and an application of the Dolph–Chebyshev function is described in Ref. 31. The spectral window arising from an application of the Dolph–Chebyshev function has a principal lobe width that is as small as possible for a given sidelobe ratio and a given number of terms.[31]

Whatever the form of the data window, it should be applied only to the actual data and not to any artificial data generated by filling out a record with zeros.

Picket-fence effect. In Fig. 12, an analogy is drawn between the output of the FFT algorithm and a bank of bandpass filters. Although each Fourier coefficient would ideally act as a filter having a rectangular response in the frequency domain, the amplitude response is in fact of the form shown in Fig. 10 because of the multiplication of the input time series by the T-second data window. In Fig. 12, the main lobes of the resulting set of spectral windows have been plotted to represent the output of the FFT. The sidelobes are not shown here. The width of each main lobe is inversely proportional to the original record length T.

At the frequencies computed, these main lobes appear to be N independent filters; that is, a unity-amplitude complex exponential (of the form $e^{i\omega t}$) with a frequency that is an integral multiple of $1/T$, would result in a response of unity at the appropriate harmonic frequency and zero at all of the other harmonics.

The picket-fence effect becomes evident when the signal being analyzed is not one of these discrete orthogonal frequencies. A signal between the third and fourth harmonics, for example, would be seen by both the third- and fourth-harmonic spectral windows but at a value lower than one. In the worst case (exactly halfway between the computed harmonics) the amplitude of the signal is reduced to 0.637 in both of the spectral windows. When this result is squared, the apparent peak power of the signal is only 0.406. Thus the power spectrum seen by this "set of filters" has a ripple that varies by a factor of 2.5 to 1. One seems to be viewing the true spectrum through a picket fence.

A cure to this problem involves performing complex interpolation on the complex Fourier coefficients. This can be accomplished by the use of an interpolation function[29] or through modification of the DFT. The latter approach will be described here.

By extending the record analyzed with a set of samples that are identically zero, one can make the redundant FFT algorithm compute a set of Fourier coefficients with terms lying between the original harmonics. Since the width of the spectral window associated with each coefficient is related solely to the reciprocal of the true record length (T), the width of these new spectral windows remains unchanged. As shown in Fig. 13, this means that the spectral windows associated with this new set of Fourier coefficients overlap considerably.

If the original time series is represented by $g(k)$ for $k = 0, 1, \cdots, N - 1$, then the series analyzed in this example can be represented by $\hat{g}(k)$, where

$$\hat{g}(k) = g(k) \qquad \text{for } 0 \leq k < N$$
$$\hat{g}(k) = 0 \qquad \text{for } N \leq k < 2N \tag{30}$$

The additional Fourier coefficients that result are interleaved with the original set.

As shown in Fig. 13, the ripple in the power spectrum has been reduced from approximately 60 percent to 20 percent. The ripple can be made larger or smaller than 20 percent by the use of less or more than N additional zeros.

In practice the picket-fence problem is not as great as this discussion implies. In many cases the signal being processed will not be a pure sinusoid but will be broad enough to fill several of the original spectral windows. Moreover, the use of any data window other than the rectangular (or boxcar) data window discussed here,

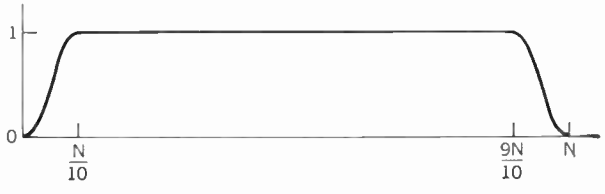
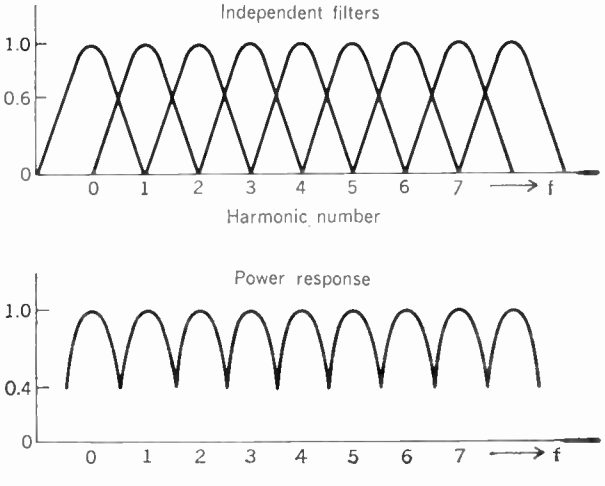


FIGURE 11. An extended cosine-bell data window.

FIGURE 12. The response of the discrete Fourier transform Fourier coefficients viewed as a set of bandpass filters (picket-fence effect).



usually tends to reduce the picket-fence effect by widening the main lobe of each spectral window.

Convolution and correlation. The blind use of Eqs. (22) and (28) to perform correlation and convolution is often inconvenient and usually incorrect. An example of incorrect usage is discussed first.

Given a function $g(t)$, one is frequently interested in convolving this function with another function $h(t)$. As shown in Fig. 14, this involves reversing (or flipping) $h(t)$, and sliding it by $g(t)$. The convolution of these functions is formed by computing and integrating the product $g(\tau) \cdot h(t - \tau)$ as a function of the relative displacement t. As indicated by Fig. 14, both functions are considered to be identically zero outside of their domain of definition. Where $g(t)$ and $h(t)$ are finite and sampled, this corresponds to the sliding, multiplying, and summing operations of Eq. (21).

When convolutions are computed with the aid of Eq. (22) and the fast Fourier transform, both of the functions are treated as being periodic. The corresponding forms of $g(\tau)$ and $h(t - \tau)$ are shown in Fig. 15. The result is a cyclical convolution, which is entirely different from the noncyclical convolution described previously.

Fortunately, this problem is easily sidestepped by simply defining and convolving $\hat{g}(\tau)$ and $\hat{h}(t - \tau)$ as shown in Fig. 16; see Refs. 7, 10, 23, 24, 28, and 31. In this case

$$\hat{g}(k) = g(k) \qquad 0 \leq k < N$$
$$\hat{g}(k) = 0 \qquad N \leq k < 2N \tag{31}$$

Thus the appropriate procedure is to

(1) Form $\hat{g}(k)$ and $\hat{h}(k)$.
(2) Find $\hat{G}(j)$ and $\hat{H}(j)$ via the FFT.
(3) Compute $\hat{C}(j) = \hat{G}(j) \cdot \hat{H}(j)$.
(4) Find $\hat{c}(k) = F^{-1}[\hat{C}(j)]$ via the FFT.

The series $\hat{c}(k)$ represents the $2N$-term series resulting from the convolution of the two N-term series $g(k)$ and $h(k)$. This technique is expressed in general terms in the section on "select-saving" in Ref. 24, and it is appropriate to correlation as well as convolution.

When one of the series convolved or correlated is much longer than the other, transforming the entire record of both functions is inconvenient and unnecessary.

If $h(k)$ is an N-term series and $g(k)$ is much longer, the procedure shown in Fig. 17 can be used. The $h(k)$ series can be thought of as the impulse response of a filter that is acting on the series $g(k)$. As shown in Fig. 14, the "present" output of the digital filter is merely a weighted sum of the last N samples it has seen. In the example of Fig. 17, this means that the $2N$-term series $h(k)$ convolved with one of the $2N$-term segments of $g(k)$ would result in N lags (or displacements), for which the correct N-term history is not available, and N lags for which the correct N-term history is available. Thus the first N lags computed are meaningless and the last N lags are correct.

In Fig. 17, overlapping sets of $2N$-term segments of $g(k)$ are shown. When convolved with $h(k)$, each of these segments contributes N correct terms to the final convolution. When all of these sets of N terms are pieced together, the result is the desired convolution of $g(k)$ and $h(k)$. Thus the length of the required FFTs is determined by the length of the short series rather than the long one. The length of the long series doesn't need to be specified.

Having exactly half of the values of $\hat{h}(\tau)$ be zero is a specific example of the "select-saving" method[24] and is made a requirement only to limit the discussion.

The segmenting technique used in convolving a short series with a long series can also be applied to the problem

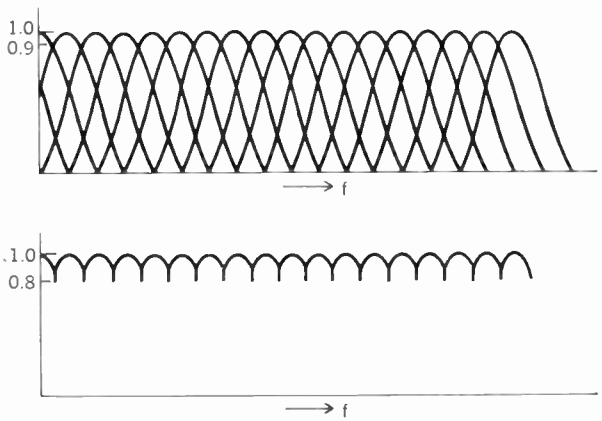


FIGURE 13. The reduction of the picket-fence effect, brought about by computing redundant overlapping sets of Fourier coefficients.

FIGURE 14. The procedure for performing noncyclical convolution on two finite signals.

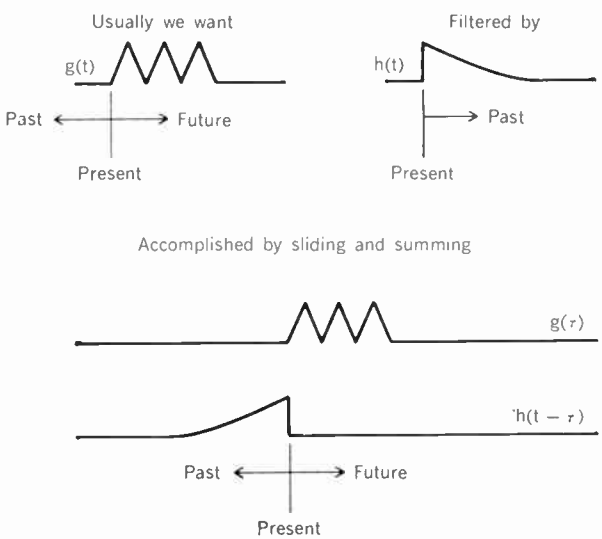
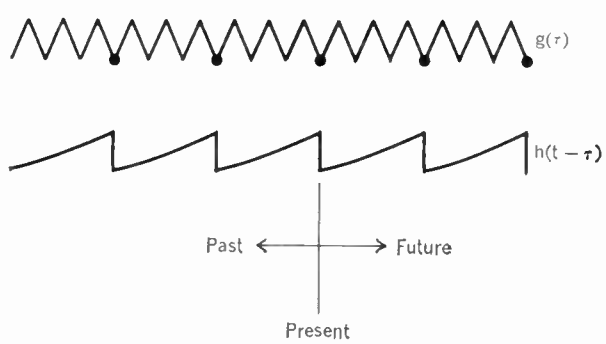


FIGURE 15. The cyclical convolution of two finite signals analogous to that performed by the FFT algorithm.



of computing N lags of the autocorrelation function of an M-term series when $N \ll M$. Figure 18 shows a function $g(t)$, which can be sampled to form an M-term time series. To compute N lags of the autocorrelation function of $g(t)$, this series is broken into overlapping segments of at least $2N$ terms. For each segment, the functions $\hat{g}(t)$ and $\hat{\hat{g}}(t)$ can be formed and sampled such that

$$\hat{g}(k) = g(k) \qquad 0 \leq k < 2N \tag{32}$$

$$\hat{\hat{g}}(k) = g(k) \qquad 0 \leq k < N$$

$$= 0 \qquad N \leq k < 2N \tag{33}$$

An appropriate procedure is to

(1) Form $\hat{G}(j)$ and $\hat{\hat{G}}(j)$ via the FFT.

(2) Compute $\hat{C}(j) = \hat{G}(j) \cdot \hat{\hat{G}}^*(j)$.

(3) Find $\hat{c}(k) = F^{-1}[\hat{C}(j)]$ via the FFT.

The first half of the $c(k)$ series represents the contribution of the first half of the $2N$-term segment to the first N lags of the autocorrelation function of $g(k)$. When the contributions from all of the segments are added together, the result is the first N lags of the autocorrelation function of all M terms of $g(k)$. This technique, called "overlap-adding," is described in detail by Helms.[24]

The choice of $2N$ zeros is made to limit the present discussion and can be changed. The optimum choice with respect to computational effort is discussed in Refs. 22 and 24. It is also a simple matter to compute any set of N lags and to apply this same technique to computing cross-correlation functions.

Frequency-domain filter design. Given the capability for convenient and efficient digital filtering, one possesses considerable degree of freedom in specifying the form of the filter. It becomes very tempting to make use of the ideal low-pass, high-pass, and bandpass filters that are so difficult to come by in the real world. This should not be done, however, without exercising great caution.

When a digital filter is specified in the frequency domain, this is equivalent to multiplying the Fourier coefficients by a window. This multiplication in the frequency domain is equivalent to performing a convolution in the time domain. Thus, the constraints discussed previously still apply but are not as readily apparent. For this discussion, the constraint is that at least half of the impulse response implied by the frequency domain filter be identically zero.

When an ideal filter, such as $H(f)$ in Fig. 19, is specified in the frequency domain, this implies a $(\sin x)/x$ impulse response in the time domain. In this example of an ideal low-pass filter, $h(t)$ does not go to zero, meaning that the tails of the $(\sin x)/x$ impulse response fold back into the T-second region. The result is aliasing in the time domain caused by undersampling the filter function in the frequency domain. This is directly analogous to the aliasing in the frequency domain caused by undersampling in the time domain.

The cure to this problem is always to choose a filter with an impulse response that dies out or can be truncated so that at least half of its terms are essentially zero. Two procedures based on truncating the impulse response are described in detail by Helms[30,31]; see also Refs. 32 and 33. In specifying a filter in the frequency domain, remember that a filter with a real impulse response implies a set of Fourier coefficients whose real part is an even function and whose imaginary part is an odd function. In the fast Fourier transform format, this implies that the real part be symmetric about the folding frequency [the (N/2)th harmonic] and the imaginary part be

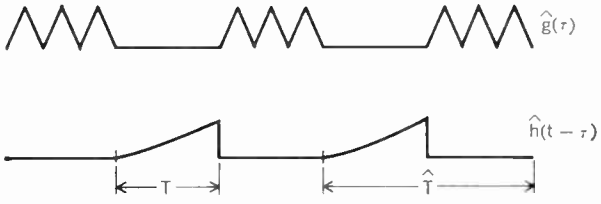


FIGURE 16. Noncyclical convolution of two finite signals analogous to that performed by the FFT algorithm.

FIGURE 17. A method for convolving a finite impulse response with an infinite time function by performing a series of fast Fourier transforms.

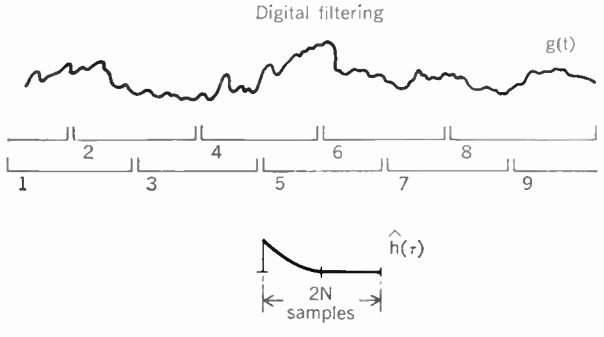
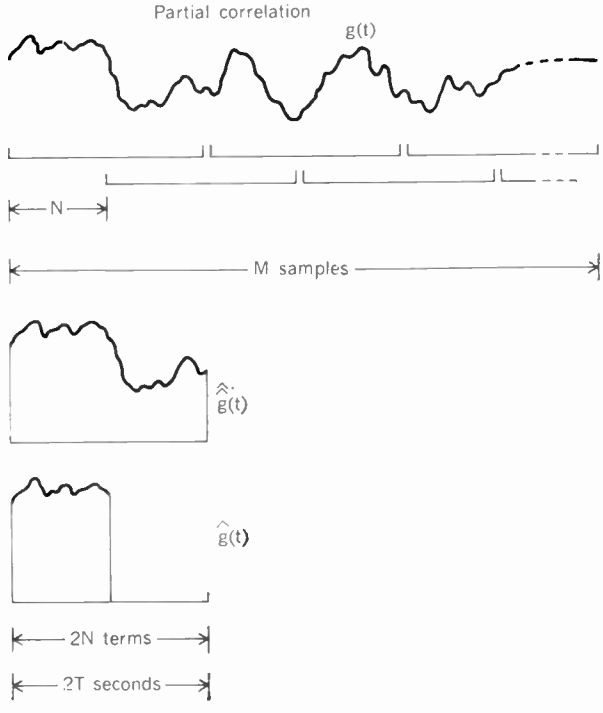


FIGURE 18. A method of using the fast Fourier transform algorithm to compute N lags of the autocorrelation function of an M-term series.



antisymmetric, as shown in the example of Fig. 1.

Applications. An example of a radar signal processing system is shown in the top half of Fig. 20. A chirped radar pulse, $s(t)$, is shown entering a matched dechirping filter followed by a Taylor weighting network.[62] In practice the input signal is corrupted by noise, the atmosphere, the equipment, etc., and the problem is to determine the effect this has on system performance.[61] In Fig. 20 this degradation is expressed in the form of a weighting function applied to $s(t)$ before it enters the matched filter. In this example the weighting is of the form $1 - x(t)$.

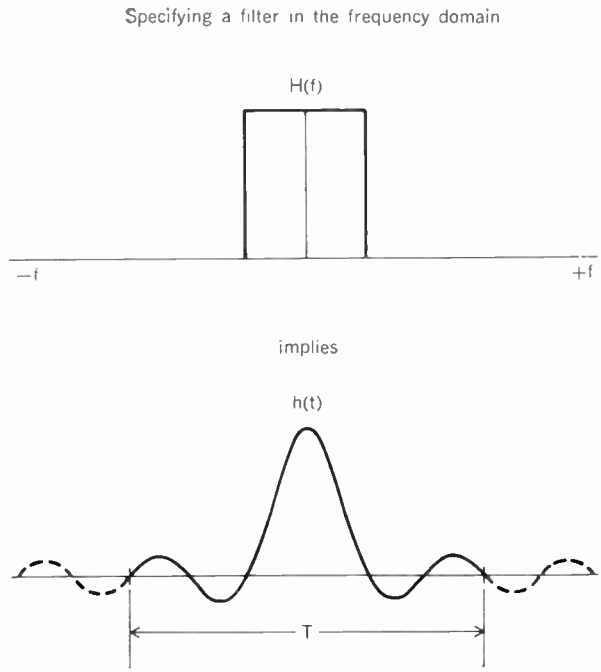
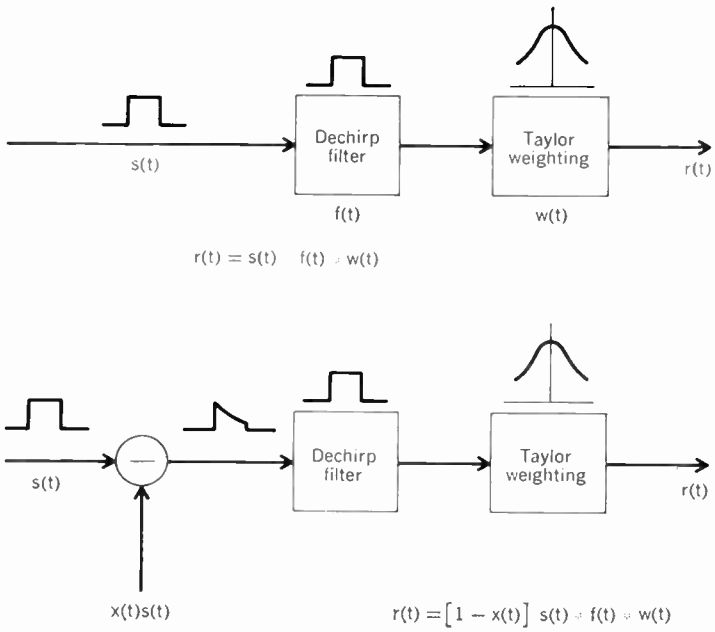


FIGURE 19. The (sin x)/x impulse response of a filter implied when a rectangular frequency response is specified.

FIGURE 20. Block diagrams illustrating the use of the FFT for simulating a radar signal processing system.



For our discussion, the matched filter can be viewed as finding the autocorrelation function of the chirped signal, and the Taylor weighting can be viewed as applying a weighting function in the frequency domain. The autocorrelation can be thought of as convolving the waveform $s(t)$ with itself reversed in time—that is, with $f(t)$. The multiplication by the Taylor weighting coefficients in the frequency domain corresponds to convolving the result of the first convolution with the impulse response of the Taylor weighting network $w(t)$. Thus the response of the system, $r(t)$, is actually the result of performing two convolutions. If $s(t)$, $f(t)$, and $w(t)$ can all be represented by N-term series, they each should be augmented with $3N$ zeros to assure that the final convolution is noncyclical.

The application of the fast Fourier transform to radar and sonar ranging systems can be viewed in terms of performing a correlation. Since correlation can be viewed as one function searching to find itself in another, it is clear how echo-ranging systems using a chirped or even a random-noise signal can operate. Since these correlations can now be done by means of the FFT, many of these operations can be performed digitally in real time.

How has it been implemented?

Software. The number of variations of the FFT algorithm appears to be directly proportional to the number of people using it.[34–53] Most of these algorithms are based on either the Cooley–Tukey or the Sande–Tukey algorithm,[11] but are formulated to exploit properties of the series being analyzed or properties of the computer being used.

Equations (17) through (20) represent the Cooley–Tukey formulation of the FFT algorithm. By separating the components of j instead of the components of k in Eq. (11), the Sande–Tukey equations would have resulted. In either case, the recursive equations can be written in the form of a two-point transform followed by a rereferencing (or twiddling[9]) operation. The resulting counterparts of Eqs. (17)–(20) are as follows:

$$\hat{A}_1(j_0,k_1,k_0) = \left\{\sum_{k_2=0}^{1} A(k_2,k_1,k_0)W_2^{j_0k_2}\right\} W_4^{j_0k_1} \quad (34)$$

$$\hat{A}_2(j_0,j_1,k_0) = \left\{\sum_{k_1=0}^{1} \hat{A}_1(j_0,k_1,k_0)W_2^{j_1k_1}\right\} W_8^{(j_12+j_0)k_0} \quad (35)$$

$$\hat{A}_3(j_0,j_1,j_2) = \left\{\sum_{k_0=0}^{1} \hat{A}_2(j_0,j_1,k_0)W_2^{j_2k_0}\right\} \quad (36)$$

$$\hat{X}(j_2,j_1,j_0) = \hat{A}_3(j_0,j_1,j_2) \quad (37)$$

The extension of the FFT algorithm from radix-two algorithms to arbitrary-radix algorithms is accomplished by representing the j and k variables of Eq. (8) in a mixed radix number system.[34,51] For the example of $N = r_1r_2r_3$, they take the form

$$\begin{aligned} j &= j_2(r_1r_2) + j_1(r_1) + j_0 \\ k &= k_2(r_2r_3) + k_1(r_3) + k_0 \end{aligned} \quad (38)$$

where $j_0, k_2 = 0, 1, \cdots, r_1 - 1$; $j_1, k_1 = 0, 1, \cdots, r_2 - 1$; and $j_2, k_0 = 0, 1, \cdots, r_3 - 1$. The resulting recursive equations can functionally be separated into r_1 point transforms, r_2 point transforms, r_3 point transforms, and rereferencing operations.

Since four-point transforms can be performed with only additions and subtractions, an algorithm with a large number of factors that are four requires less computation than a radix-two algorithm.[44] In much the same manner, a further reduction can be made by forcing the algorithm into a form where a large number of eight-point transforms can be done very efficiently.[35] The number of multiplications required by the resulting algorithms is reduced by 30 percent and 40 percent, respectively, compared with radix-two algorithms.

Several variants of the FFT algorithm have been motivated by characteristics of the series being transformed. When the series is real, the expected two-to-one reduction in computation and storage can be obtained by either putting the N-term real record in the form of an artificial $N/2$-term complex record,[7] or by restructuring the FFT algorithm.[36]

An algorithm that allows the value of N to take on any value (including a prime number) has recently been proposed.[25,38] For any value of N, the discrete Fourier transform can be written

$$X(j) = \frac{1}{N}\sum_{k=0}^{N-1} x(k)W^{-jk} \tag{39}$$

$$X(j) = \frac{1}{N}\sum_{k=0}^{N-1} x(k)W^{-jk+[(k^2-k)^2+j^2-j)^2]/2} \tag{40}$$

$$X(j) = \frac{W^{-j^2/2}}{N}\left\{\sum_{k=0}^{N-1}[W^{-k^2/2}x(k)]W^{(j-k)^2/2}\right\} \tag{41}$$

Note that Eq. (41) is in the form of a convolution; i.e.,

$$X(j) = \frac{W^{-j^2/2}}{N}\sum_{k=0}^{N-1} g(k)h(j-k) \tag{42}$$

where $g(k) = W^{-k^2/2}x(k)$, and $h(j-k) = W^{(j-k)^2/2}$. In performing this convolution, both of the series can be augmented with zeros until they contain a highly factorable number of terms $\hat{N}$. These extended series can be transformed by a conventional $\hat{N}$-term FFT algorithm to obtain the noncyclical convolution required by Eq. (42). The Fourier coefficients corresponding to the original N-term DFT of Eq. (39) are obtained by multiplying the results of this convolution by the unit amplitude complex exponential $W^{-j^2/2}$, where $W = \exp(2\pi i/N)$.

This algorithm should be very useful when employed with a hardware FFT processor that otherwise would be restricted to values of N that are powers of two.

A variety of different FFT programs for performing one-dimensional and multidimensional fast Fourier transforms have been made available by Cooley,[41,42] Sande,[48] Singleton,[49-52] and Brenner,[39] who have programmed most of the options described here plus several others as well.

Hardware. The advent of the FFT algorithm reduced the time required for performing a 2^{10}-point discrete Fourier transform from several minutes to less than a second. The advent of special-purpose digital hardware further reduced this time to tens of milliseconds.[54-60]

The cost of finding a 2^{10}-point transform used to be measured in dollars. The FFT algorithm reduced this cost to a few cents. Special-purpose hardware has reduced it further to hundredths of a cent.

Thus in the past five years, both the cost and execution time of computing a discrete Fourier transform have been reduced by nearly four orders of magnitude. Therefore, we can now build a relatively inexpensive special-purpose processor that is able to meet the real-time constraints of a wide variety of signal-processing problems.

A survey of FFT processors and their characteristics is reported in Ref. 55. Most of these processors can be classified in one of the four families of FFT processor machine organizations discussed in Ref. 56. These four families represent varying degrees of parallelism, performance, and cost.

Most of the hardware implementations built to date have relied on either a radix-two or a radix-four algorithm. The regularity of these algorithms and the resulting simplification in control have tended to discourage the use of the more general arbitrary-radix algorithms. In addition, most users of real-time FFT hardware seem to have adapted quickly to thinking in powers of two.

The author would like to thank D. E. Wilson for the loan of the title; R. A. Kaenel and W. W. Lang for their encouragement; and B. P. Bogert, W. T. Hartwell, H. D. Helms, C. M. Rader, and P. T. Rux for suggesting several improvements, which were incorporated in this article.

REFERENCES

Introduction to Fourier analysis

1. Arsac, J., *Fourier Transforms*. Englewood Cliffs, N.J.: Prentice-Hall, 1966.
2. Bracewell, R., *The Fourier Transform and Its Applications*. New York: McGraw-Hill, 1965.
3. Papoulis, A., *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

Historical development of the fast Fourier transform

4. Cooley, J. W., Lewis, P. A. W., and Welch, P. D., "Historical notes on the fast Fourier transform," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 76–79, June 1967.

Introduction to the fast Fourier transform

5. Brigham, E. O., and Morrow, R. E., "The fast Fourier transform," *IEEE Spectrum*, vol. 4, pp. 63–70, Dec. 1967.
6. Cochran, W. T., *et al.*, "What is the fast Fourier transform?" *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 45–55, June 1967.
7. Cooley, J. W., Lewis, P. A. W., and Welch, P. D., "The fast Fourier transform algorithm and its applications," IBM Research Paper RC-1743, Feb. 1967.
8. Cooley, J. W., and Tukey, J. W., "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, Apr. 1965.
9. Gentleman, W. M., and Sande, G., "Fast Fourier transforms — for fun and profit," *1966 Fall Joint Computer Conf., AFIPS Proc.*, vol. 29. Washington, D.C.: Spartan Books.
10. Gold, B., and Rader, C. M., *Digital Processing of Signals*. New York: McGraw-Hill, 1969.

Spectrum and cepstrum analysis

11. Bingham, C., Godfrey, M. D., and Tukey, J. W., "Modern techniques of power spectrum estimation," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 56–66, June 1967.
12. Blackman, R. B., and Tukey, J. W., *The Measurement of Power Spectra*. New York: Dover, 1958.
13. Bogert, B. P., Healy, M. J. and Tukey, J. W., "The frequency analysis of time series for echoes: cepstrum, pseudoautocovariance, cross-cepstrum and saphe-cracking," *Time Series Analysis*, Murray Rosenblatt, ed. New York: Wiley, 1963, pp. 201–243.
14. Noll, A. Michael, "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection," *J. Acoust. Soc. Am.*, vol. 36, pp. 296–302, 1964.
15. Oppenheim, A. V., Schafer, R. W., and Stockham, T. G., Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264–1291, Aug. 1968. (Reprinted in *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 437–465, Sept. 1968.)
16. Parzen, E., "Statistical spectral analysis (single channel case) in 1968," Tech. Rep. 11, ONR Contract Nonr-225(80)(NR-042-234), Stanford University, Dept. of Statistics, Stanford, Calif., June 10, 1968.
17. Richards, P. I., "Computing reliable power spectra," *IEEE Spectrum*, vol. 4, pp. 83–90, Jan. 1967.

18. Tukey, J. W., "An introduction to the measurement of spectra," in *Probability and Statistics*, Ulf Grenander, ed. New York: Wiley, 1959, pp. 300–330.

19. Tukey, J. W., "An introduction to the calculations of numerical spectrum analysis," in *Spectral Analysis of Time Series*, Bernard Harris, ed. New York: Wiley, 1967, pp. 25–46.

20. Welch, P. D., "A direct digital method of power spectrum estimation," *IBM J. Res. Develop.*, vol. 5, pp. 141–156, 1961.

21. Welch, P. D., "The use of the fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 70–73, June 1967.

Use of the fast Fourier transform in convolution, correlation, digital filtering, etc.

22. Cooley, J. W., "Applications of the fast Fourier transform method," *Proc. IBM Scientific Computing Symp. on Digital Simulation of Continuous Systems*, Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1966.

23. Cooley, J. W., Lewis, P. A. W., and Welch, P. D., "Application of the fast Fourier transform to computation of Fourier integrals, Fourier series, and convolution integrals," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 79–84, June 1967.

24. Helms, H. D., "Fast Fourier transform method of computing difference equations and simulating filters," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 85–90, June 1967.

25. Rabiner, L. R., Schafer, R. W., and Rader, C. M., "The chirp *Z*-transform algorithm and its applications," *Bell System Tech. J.*, vol. 48, pp. 1249–1292, May–June 1969.

26. Sande, G., "On an alternative method of calculating covariance functions," unpublished technical note, Princeton University, Princeton, N.J., 1965.

27. Singleton, R. C., "Algorithm 345, an Algol convolution procedure based on the fast Fourier transform," *Commun. Assoc. Comput. Mach.*, vol. 12, Mar. 1969.

28. Stockham, T. G., "High-speed convolution and correlation," *1966 Spring Joint Computer Conf., AFIPS Proc.*, vol. 28. Washington, D.C.: Spartan Books, pp. 229–233.

The picket-fence effect

29. Hartwell, W. T., "An alternate approach to the use of discrete Fourier transforms," to be published.

Frequency-domain filter design

30. Helms, H. D., "Fast Fourier transform method for computing difference equations and simulating filters," *IEEE Trans. Audio and Electroacoustics*, vol. AU-15, pp. 85–90, June 1967.

31. Helms, H. D., "Nonrecursive digital filters: design methods for achieving specifications on frequency response," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 336–342, Sept. 1968.

32. Kaiser, J. F., "Digital filters," in *System Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, eds. New York: Wiley, 1966.

33. Otnes, R. K., "An elementary design procedure for digital filters," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 336–342, Sept. 1968.

Algorithms and software

34. Bergland, G. D., "The fast Fourier transform recursive equations for arbitrary length records," *Math. Comput.*, vol. 21, pp. 236–238, Apr. 1967.

35. Bergland, G. D., "A fast Fourier transform algorithm using base 8 iterations," *Math. Comput.*, vol. 22, pp. 275–279, Apr. 1968.

36. Bergland, G. D., "A fast Fourier transform algorithm for real-valued series," *Commun. Assoc. Comput. Mach.*, vol. 11, pp. 703–710, Oct. 1968.

37. Bergland, G. D., and Wilson, D. E., "An FFT algorithm for a global, highly parallel processor," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, June 1969.

38. Bluestein, L. I., "A linear filtering approach to the computation of the discrete Fourier transform," *1968 NEREM Record*, pp. 218–219.

39. Brenner, N. M., "Three Fortran programs that perform the Cooley–Tukey Fourier transform," Tech. Note 1967-2, Lincoln Laboratory, M.I.T., Lexington, Mass., July 1967.

40. Cooley, J. W., and Tukey, J. W., "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, pp. 297–301, Apr. 1965.

41. Cooley, J. W., "Harmonic analysis complex Fourier series," SHARE Doc. 3425, Feb. 7, 1966.

42. Cooley, J. W., "Complex finite Fourier transform subroutine," SHARE Doc. 3465, Sept. 8, 1966.

43. Danielson, G. C., and Lanczos, C., "Some improvements in practical Fourier analysis and their application to X-ray scattering from liquids," *J. Franklin Inst.*, vol. 233, pp. 365–380, 435–452.

44. Gentleman, W. M., and Sande, G., "Fast Fourier transforms—for fun and profit," *1966 Fall Joint Computer Conf., AFIPS Proc.*, vol. 29. Washington, D.C.: Spartan Books.

45. Good, I. J., "The interaction algorithm and practical Fourier series," *J. Roy. Statist. Soc.*, vol. 20, series B, pp. 361–372, 1958; addendum, vol. 22, pp. 372–375, 1960.

46. Pease, M. C., "An adaption of the fast Fourier transform for parallel processing," *J. Assoc. Comput. Mach.*, vol. 15, pp. 252–264, Apr. 1968.

47. Rader, C. M., "Discrete Fourier transforms when the number of data samples is prime," *Proc. IEEE*, vol. 56, pp. 1107–1108, June 1968.

48. Sande, G., "Arbitrary radix one-dimensional fast Fourier transform subroutines," University of Chicago, Ill., 1968.

49. Singleton, R. C., "On computing the fast Fourier transform," *Commun. Assoc. Comput. Mach.*, vol. 10, pp. 647–654, Oct. 1967.

50. Singleton, R. C., "Algorithm 338, Algol procedures for the fast Fourier transform," *Commun. Assoc. Comput. Mach.*, vol. 11, pp. 647–654, Nov. 1968.

51. Singleton, R. C., "Algorithm 339, an Algol procedure for the fast Fourier transform with arbitrary factors," *Commun. Assoc. Comput. Mach.*, vol. 11, pp. 776–779, Nov. 1968.

52. Singleton, R. C., "Algorithm 345, an Algol convolution procedure based on the fast Fourier transform," *Commun. Assoc. Comput. Mach.*, vol. 12, Mar. 1969.

53. Yavne, R., "An economical method for calculating the discrete Fourier transform," *1968 Fall Joint Computer Conf., IFIPS Proc.*, vol. 33. Washington, D.C.: Spartan Books, pp. 115–125.

Fast Fourier transform hardware

54. Bergland, G. D., and Hale, H. W., "Digital real-time spectral analysis," *IEEE Trans. Electronic Computers*, vol. EC-16, pp. 180–185, Apr. 1967.

55. Bergland, G. D., "Fast Fourier transform hardware implementations—a survey," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, June 1969.

56. Bergland, G. D., "Fast Fourier transform hardware implementations—an overview," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, June 1969.

57. McCullough, R. B., "A real-time digital spectrum analyzer," Stanford Electronics Laboratories Sci. Rept. 23, Stanford University, Calif., Nov. 1967.

58. Pease, M. C., III, and Goldberg, J., "Feasibility study of a special-purpose digital computer for on-line Fourier analysis," Order No. 989, Advanced Research Projects Agency, Washington, D.C., May 1967.

59. Shively, R. R., "A digital processor to generate spectra in real time," *IEEE Trans. Computers*, vol. C-17, pp. 485–491, May 1968.

60. Smith, R. A., "A fast Fourier transform processor," Bell Telephone Laboratories, Inc., Whippany, N.J., 1967.

Other

61. Gilbert, S. M., Private communication, Bell Telephone Laboratories, Inc., Whippany, N.J.

62. Klauder, J. R., Price, A. C., Darlington, S., and Albersheim, W. J., "The theory and design of chirp radars," *Bell System Tech. J.*, vol. 39, pp. 745–808, July 1960.

G. D. Bergland (M) received the B.S., M.S., and Ph.D. degrees from Iowa State University in 1962, 1964, and 1966, respectively. While at Iowa State he was a teaching assistant in the Electrical Engineering Department and a research assistant in the Engineering Experiment Station. From 1964 to 1966 he held a National Science Foundation traineeship. In 1966 he joined the Digital Systems Laboratory at Bell Telephone Laboratories, Inc., Whippany, N.J., where he conducted research in the area of special-purpose computer organizations. Since 1968 he has been the supervisor of the Computer Systems Studies Group, which is involved in the study and evaluation of real-time applications of special-purpose data-processing techniques. He is a member of Sigma Xi, Phi Kappa Phi, Eta Kappa Nu, and Tau Beta Pi.

7/28/69

The Tiros decade

It is doubtful that there have been many projects, even in less complex and competitive fields, that measure up to the complete success of the Tiros/ESSA meteorological satellite program; still greater promise is expected from a second-generation series

Abraham Schnapf *RCA Corporation*

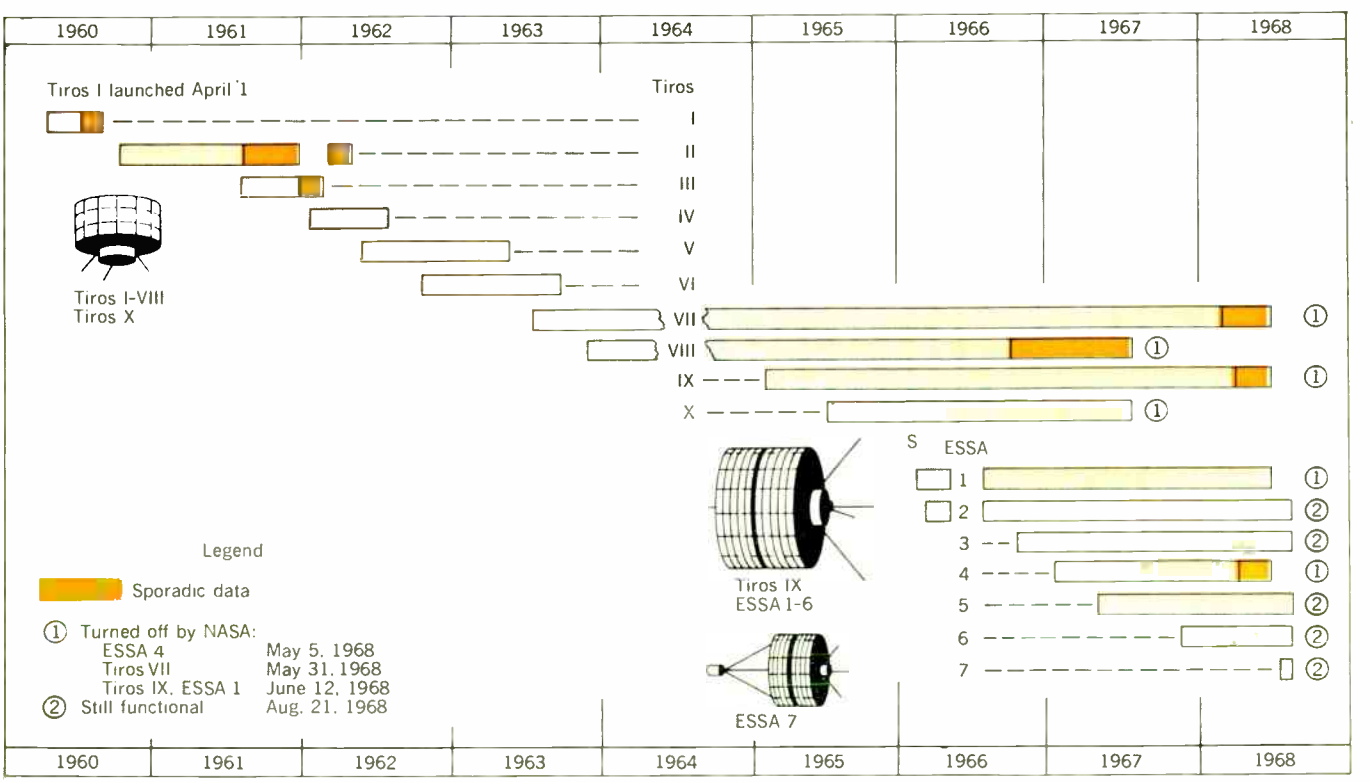
The space age—started by the launch of Sputnik I—is now ten years old. Men have orbited the earth and moon, and spacecraft have landed on the moon and traveled to distant planets. Communications satellites make possible live television among continents and, with the aid of meteorological satellites, giant strides are being made in our understanding of weather. RCA has played a key role in many of these areas, but none more important than its role in the meteorological satellite program—Tiros. Throughout the space decade, RCA has been involved in this program, which provided the first important peaceful and beneficial use of outer space by the nations of the world with the implementation of the Tiros Operational System (TOS) in 1966. This article reviews the performance of the ten Tiros and nine ESSA satellites (part of TOS) orbited during this decade, and gives some insight into the improved TOS satellites that will be launched in the near future.

RCA's participation in the Tiros (Television and Infrared Observation Satellite) meteorological satellite program started in 1958, with the first Tiros satellite being launched on April 1, 1960. Since then, a total of ten Tiros and nine ESSA satellites have been successfully orbited (Fig. 1), providing nearly continuous space observations of this planet's weather phenomena for more than nine years.

Originally, Tiros was an experimental system. However, in February 1966, the Tiros Operational System (TOS) was implemented with the successful orbiting of the ESSA 1 and 2, expanding the basis Tiros system and providing the world's first operational satellite system capable of observing the earth's cloud cover on a daily routine basis.

Now, second-generation TOS satellites are under development; these will further enhance the potential of the Tiros system for global operational weather observation and forecasting.

FIGURE 1. Summary of Tiros/TOS performance.



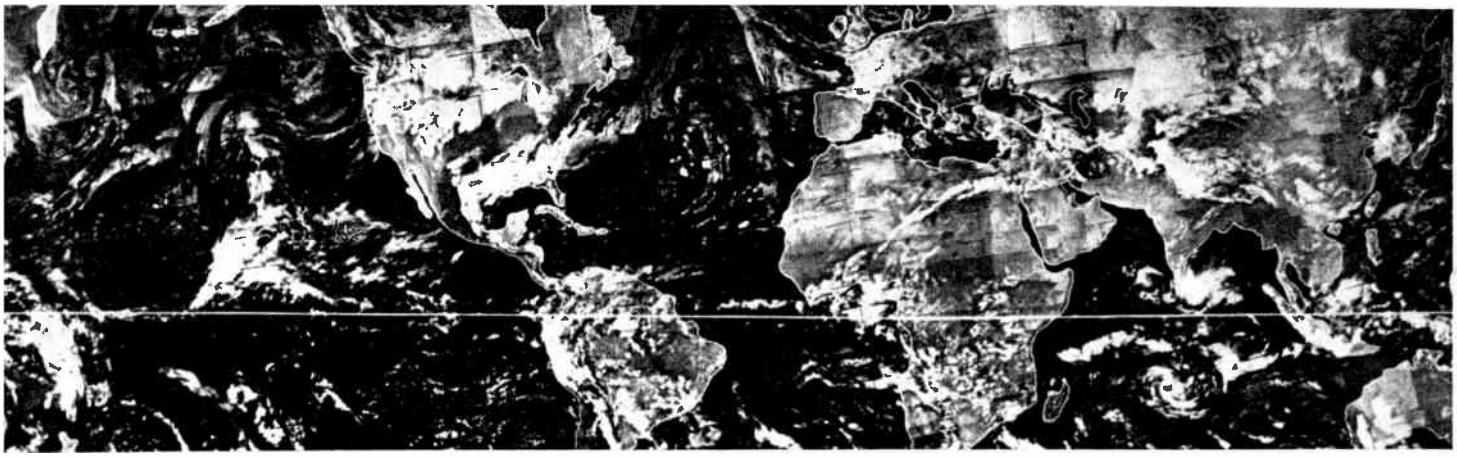


FIGURE 2. First complete view of the world's weather. This photomosaic is composed of 450 photographs taken by Tiros IX during its 12 orbits on February 12, 1965. (Provided through the courtesy of U.S. Environmental Science Services Administration.)

FIGURE 3 (right). Portion of an orbital sequence of ESSA 3 television photographs showing Saudi Arabia, Somaliland, and the Indian Ocean. (Gridding added by digital computer at the U.S. National Environmental Satellite Center.)

Meteorological benefits

Through the Tiros research and development programs, a reliable and useful meteorological tool has been developed; namely, the Tiros Operational System. This system has been in operation for more than three years, providing routinely and without interruption daily global weather observation to the National Environmental Satellite Center of the United States and local APT (Automatic Picture Transmission) weather photographs to more than 400 stations located throughout the world.

The important product of TOS is the observation of weather conditions in all parts of the world and the provision of the weather data rapidly and in useful form. Major weather systems, such as fronts, storm centers, tropical and extratropical flows, hurricanes, typhoons, and distinctive cloud patterns are viewed by the television cameras in the ESSA satellites. These data are then relayed to earth, to be applied to other, previously obtained, data for analysis and forecasting.

Several of the more than 1 180 000 television pictures returned by the Tiros and ESSA satellites are shown in Figs. 2, 3, and 4. Although reproduction processes result in the loss of detail, the high quality of the Tiros and TOS photographs is evident in these illustrations. The photographs were taken with television cameras equipped with wide-angle lenses.

Evolution of the global weather satellite system

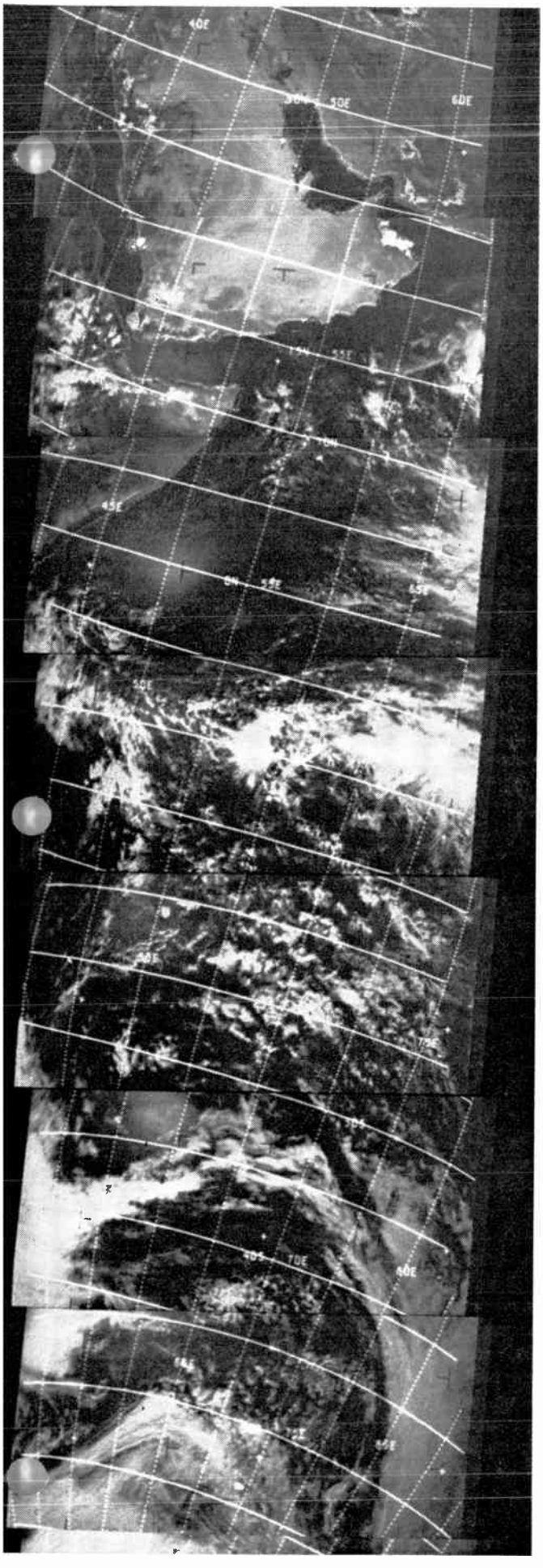
Tiros I, the world's first meteorological satellite, was launched April 1, 1960, with the primary objective of demonstrating the feasibility of observing the earth's cloud cover by means of slow-scan television cameras in an earth-orbiting, spin-stabilized satellite. This satellite included both a wide-angle and a narrow-angle television camera, and was placed in a circular orbit at an altitude of 400 nautical miles (740 km), with the orbit inclined 48 degrees to the equator.

The first historic television pictures from space were received on the very first orbit of Tiros I, immediately and clearly demonstrating the feasibility of the system. Figure 5 shows the first picture taken on the first orbit, a wide-angle picture showing for the first time the earth's cloud cover from space over the northeastern part of the United States and part of Canada. With the reception of pictures from Tiros I, a new and powerful tool for the meteorological community became a reality.

Tiros II was orbited on November 23, 1960, to demonstrate, in addition to the wide- and narrow-angle television cameras, an experimental, five-channel, scanning infrared radiometer, and a two-channel nonscanning infrared device. Both of these devices were developed by the NASA Goddard Space Flight Center. They measured the thermal energy of both the earth's surface and atmosphere in order to provide data on the planet's heat balance and add a new dimension to the understanding of weather. A magnetic torquing coil was added to Tiros II (and all Tiros satellites thereafter) so that a controlled magnetic field about the satellite would interact with the earth's field in space and, hence, provide control of the satellite's attitude. In this way, camera pointing, thermal control, and the use of available solar power were enhanced.

Tiros III, IV, V, VI, and VII were launched between July 1961 and June 1963 to provide continuous observation of the earth's cloud cover for limited operational use. With each of these satellites, particular emphasis was given to providing early warning of severe tropical storms, hurricanes, and typhoons. In addition to cloud-cover observation, the satellites were employed experimentally to detect sea ice and snow cover, and to support the Indian Ocean expedition, Ice Reconnaissance experiments, and other research programs. These spacecraft each contained two, slow-scan, ½-inch (1.27-cm) vidicon television camera systems as the primary sensors.

Tiros VIII, launched in December 1963, included both a 1.27-cm Tiros camera and a 1-inch (2.54-cm) automatic picture transmission (APT) camera. This marked the first in-space use of the APT system that had been developed for Nimbus I. The APT camera utilizes a very slow-scan vidicon, as compared to the 1.27-cm television camera. The latter requires 2 seconds to scan its 500-television-line image; the APT camera requires 200 seconds for readout of its 800-television-line image. By virtue of the 2-kHz bandwidth of the APT system, Tiros VIII was able to transmit direct, real-time television pictures to a series of 45 relatively inexpensive APT ground stations located around the world.



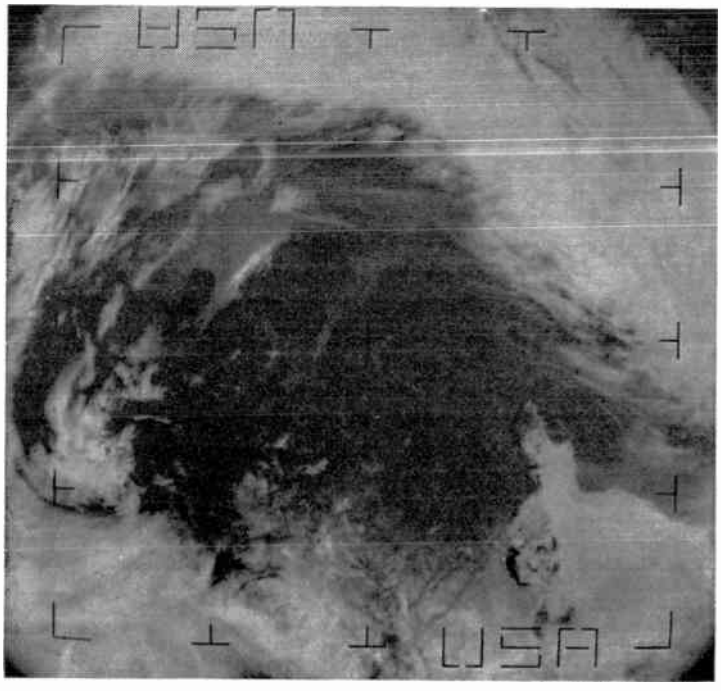


FIGURE 4. The one-millionth picture taken by the Tiros/ESSA series; picture shows the Hudson Bay area and northern New England on May 27, 1968.

Tiros IX, the first "wheel-mode" satellite, was launched in January 1965 with the objective of expanding the capability of the Tiros satellites to provide complete global weather observation on a daily basis. This represented an increase of four times the daily observation provided by the predecessor Tiros satellites.

With its new design, Tiros IX differed from its predecessors in many aspects, and was the forerunner of the satellites now used in the Tiros Operational System.

A primary difference was that in Tiros I through VIII the two television cameras were mounted on the baseplate of the satellite with the optical axes parallel to the inertially stabilized spin axis. Hence, the camera axes were parallel to the orbit plane and viewed the earth for about 25 percent of each orbit. In the Tiros IX configuration, the television cameras were mounted diametrically opposite one another and looked out through the sides rather than through the baseplate of the satellite. The satellite was injected into orbit with the spin axis in the orbital plane; however, the spin axis was then maneuvered by an improved magnetic-torquing system to an attitude normal to the orbit plane. Thus, the spinning satellite "rolled" along its orbital path and the field of view of each camera passed through the local vertical once during each spin or "roll." At the proper interval in the picture-taking sequence, the camera shutter was triggered to take a photo of the local scene when the camera was looking down at the earth. Hence, throughout the sunlit portion of the orbit, the earth below the satellite could be observed by means of a sequence of overlapping photos. By placing the wheel satellite in a near-polar, sun-synchronous orbit, the entire earth could be observed on a daily basis.

Tiros X, the last of the research and development series of standard Tiros satellites, was launched in July

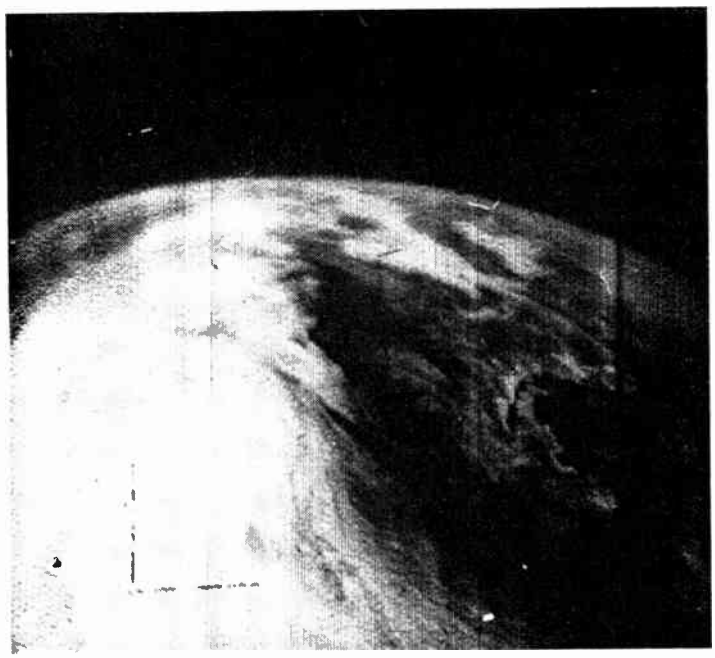


FIGURE 5. First television picture of earth's weather received from a satellite; picture was taken from a 640-km altitude, and shows the weather over the northeastern U.S. and part of Canada, including Nova Scotia and the St. Lawrence River, on April 1, 1960. A 1.27-cm television camera was used.

1965 to provide hurricane and tropical storm observations.

ESSA 1 was launched on February 3, 1966, into a 400-nautical-mile (740-km), near-polar, sun-synchronous orbit to become the first operational satellite providing global observations on a daily basis. This satellite (like its predecessor, Tiros IX) utilized two 1.27-cm vidicon camera systems, wherein a pair of pictures (one from each camera) produced a picture swath 2200 miles (3500 km) wide and 800 miles (1280 km) long along the orbit track. Contiguous coverage was provided by programming the cameras to take "picture pairs" every two minutes along the sun-illuminated portion of the orbital track. With the 14.5 orbits completed each day, a total of 450 television photos were available for transmission to the Tiros ground stations at Fairbanks, Alaska, or the station at Wallops Island, Va. From these stations, the satellite television and telemetry data were retransmitted to the Environmental Science Services Administration's (ESSA) National Environmental Satellite Center (NESC), for processing, analysis, and retransmission to major weather centers in the United States, as well as abroad.

The TOS satellites. With ESSA 1 on station and providing operational global observation for readout in the United States, the second ESSA satellite, ESSA 2, was successfully placed in orbit on February 28, 1966. ESSA 2 was actually the first of the TOS-design spacecraft. It was launched into a 750-nautical-mile (1400-km), sun-synchronous orbit to complement ESSA 1 in the Tiros Operational System by providing direct, real-time readout of APT pictures to the APT ground stations located throughout the world.

This pair of operating satellites fulfilled the commitment made by the United States to provide an operational meteorological satellite system in the first quarter of 1966.

Later in 1966, on October 2, the ESSA 3 satellite was launched. It replaced ESSA 1, in which a television sensor had ceased operating. ESSA 3 was launched into a 750-nautical-mile (1400-km), sun-synchronous orbit. In this satellite, the original Tiros wide-angle camera system was replaced with a modified Nimbus camera, the Advanced Vidicon Camera System (AVCS), which provided higher resolution and a larger picture area than the 1.27-cm Tiros cameras. [In general, the AVCS and APT camera systems both have similar resolution, which, at 790 nautical miles (1460 km) altitude, is two nautical miles (3.7 km) at the local vertical.]

To ensure uninterrupted daily global photocoverage, ESSA 4 (the second of the TOS APT satellites) was placed in orbit on January 26, 1967. This satellite replaced ESSA 2, which was providing limited global coverage due to a slow orbital drift that had taken place since launch. ESSA 5, the second TOS AVCS satellite, was placed in orbit on April 20, 1967, to provide back-up for ESSA 3. ESSA 6, carrying APT cameras, was successfully launched on November 10, 1967, to replace ESSA 4, whose one remaining operational camera did not provide satisfactory operational data. ESSA 7 was launched on August 16, 1968, to ensure an uninterrupted supply of AVCS photographs. There have since been two additional successful ESSA launches, ESSA 8 and ESSA 9; thus extending the perfect record of the Tiros/TOS programs to 19 successful spacecraft in 19 launches.

Description of the Tiros Operational System

The Tiros Operational System is sponsored by the U.S. Department of Commerce, and is managed and operated by Environmental Science Services Administration's National Environmental Satellite Center, under the technical direction of the National Aeronautics and Space Administration's Goddard Space Flight Center (NASA/GSFC). To meet the full operational objectives of the system, it is required that two TOS meteorological satellites be in orbit at all times; one carrying the APT subsystem for direct local readout to APT stations throughout the world, and the second carrying the AVCS, which is capable of storing global video data for readout to associated ground stations that immediately relay the data to the NESC for processing and analysis.

The TOS ground station network. The two primary TOS ground stations are the Command and Data Acquisition (CDA) stations located near Fairbanks, Alaska, and Wallops Island, Va. The locations of these stations permit direct communications with a TOS satellite on every orbit, except one, on a daily basis. The two stations are similar, and each is equipped for either manual or automatic transmission of commands to the satellite by means of a low-gain antenna. A separate, 85-foot (25-meter), steerable, parabolic antenna is used to receive satellite-transmitted telemetry data and the AVCS video transmissions; APT video need not be received by the CDA stations.

All telemetry data transmitted from the satellite are recorded on seven-channel tape recorders at the CDA station; when AVCS video data are received, a second tape recorder is employed. The telemetry data from the satellite's 137.77-MHz beacon also are recorded on paper-chart recorders and, in addition, are transmitted

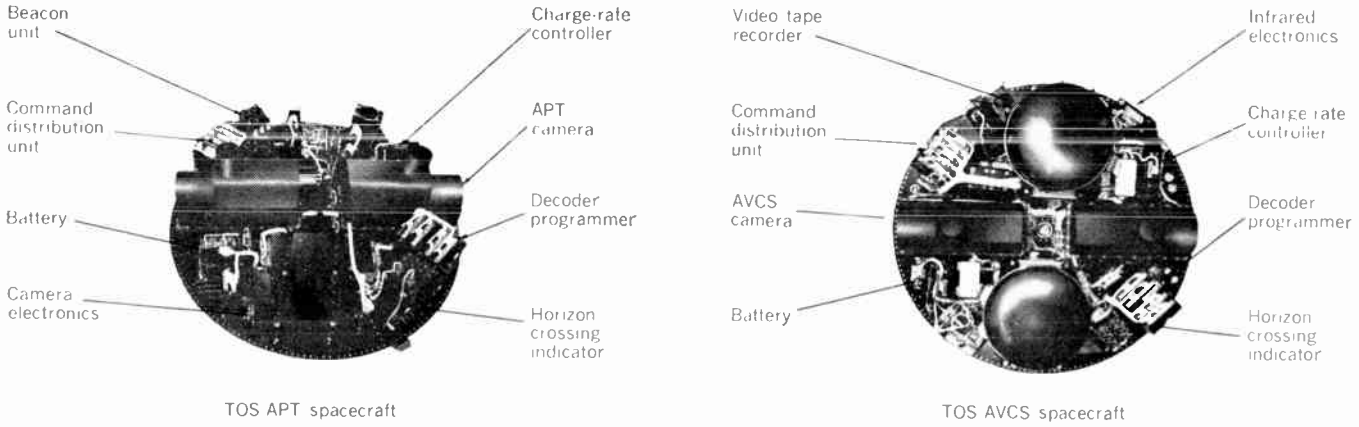


FIGURE 6. Comparison of the TOS APT and TOS AVCS baseplate layouts.

in real time to the TOS Operations Center (TOC) at the National Environmental Satellite Center, Suitland, Md., and to the TOS Evaluation Center at the Goddard Space Flight Center, Greenbelt, Md. The video data are played back at an 8-to-1 reduced rate over the 48-kHz broadband transmission line to TOC.

The TOS satellite ephemeris data are provided by NASA's Space Tracking and Data Acquisition Network (STADAN). These data are used for providing the CDA stations and APT stations with orbit tracking data, and also by TOC to permit picture gridding by computer and to facilitate programming for future events.

Under ESSA control, programming instructions and pertinent ephemeris and tracking data (derived from NASA STADAN station data) are forwarded from TOC to the primary CDA stations in advance of that station's contact with the satellite. TOC also monitors the performance of the satellite throughout its operational life, in order to budget the power supply and maintain the spin-axis attitude, the spin period, and the operating temperatures within the optimum design limits.

Physical description of the satellite. The TOS APT and TOS AVCS satellites are similar in their general external physical characteristics. The satellite structure is similar to that of previous Tiros satellites, consisting of an 18-sided right polyhedron, 22.5 inches (57 cm) high and 42 inches (106 cm) in diameter. A reinforced baseplate carries most of the subsystem components, and the cover assembly ("hat") provides mounting area for the solar cells on its outer top and side surfaces. The dynamics-control devices provided on the satellite consist of attitude- and spin-control magnetic coils, and mechanical and liquid precession dampers. These devices are mounted inside the hat structure. Openings in the hat provide viewing ports for various sensors mounted on the baseplate. A crossed-dipole antenna projects from the underside of the baseplate and a monopole, or whip, antenna extends vertically from the center of the hat. On ESSA 7 and 9, this antenna was modified to combine, in one structure, the receiving antenna and a biconical S-band transmitting antenna, which is being given an "in-space" test prior to its use on the improved TOS system described later. In addition, the AVCS satellite is equipped with terrestrial-radiation sensors developed by the University of Wisconsin for measurement of the earth's heat balance.

The APT satellite with redundant APT cameras weighs 285 pounds (130 kg); the AVCS satellite, with redundant AVCS cameras and video recorders, and the University of Wisconsin radiometers and associated equipment, weighs 325 pounds (147 kg), 345 pounds (156 kg) with the new S-band transmitting system. Except for data recorders and an infrared subsystem, the functional diagram for the TOS AVCS spacecraft is essentially the same as that for the APT spacecraft. Figure 6 permits a side-by-side comparison of the two types of TOS spacecraft.

Satellite dynamics control. The identical dynamics-control subsystems for the two types of TOS satellites include attitude- and spin-control coils, and nutation dampers. The primary technique used for controlling the spinning satellite's attitude is magnetic torquing. Since the mission requires a sun-synchronous orbit, in which the orbit plane precesses in synchronism with the earth's motion around the sun, the satellite's inertially stabilized spin axis must be precessed at the same rate in order to maintain the "wheel" attitude (in which the spin axis is normal to the orbit plane). To achieve this continuous slow drift in the proper direction, a magnetic bias coil (MBC) is used in conjunction with the quarter orbit magnetic attitude control (QOMAC) coil. The MBC is similar to the magnetic attitude-control devices employed on Tiros II through VIII; the QOMAC coil is similar to that used first on Tiros IX.

The current-carrying MBC generates an electromagnetic field of selectable strength, the dipole moment of which is colinear with the satellite's spin axis. The device consists of a coil of wire and a stepping switch to vary the amount and direction of current in the coil.

The QOMAC coil, operating in a low-torque mode, provides the fine control over spin axis motion required to keep the satellite in mission attitude. In its high-torque mode of operation, this coil provides the rapid attitude change required to initially achieve the wheel attitude. This attitude change, of approximately 90 degrees, is required because upon injection into orbit the satellite's spin axis lies in or near the orbit plane, rather than normal to it. In the high-torque mode, the spin axis precesses approximately 10 degrees per orbit, and in the low-torque mode, approximately 2 degrees per orbit.

The magnetic spin control (MASC) coil is used to maintain the satellite spin rate at the optimum level for

operation of the television cameras. Current through the MASC coil is commutated at one-half-spin intervals to provide a motor effect that can be used to increase or decrease the spin rate. Normally, the MASC coil is activated only near the earth's poles, where its operation is most efficient.

The second-generation TIROS Operational System

In 1966, design studies were initiated by NASA/GSFC, in consultation with ESSA, and are being implemented by RCA under the Tiros M and Improved TOS (ITOS) program. This program utilizes, whenever possible, the proven technology and satellite hardware developed by the Tiros and TOS satellites, and sensors developed for other NASA programs.

Since the TOS system is an operational system, the changes planned for Tiros M/ITOS were chosen to reflect an orderly transition. The phase-in process between TOS and Tiros M/ITOS reflects the requirements of common usage of existing ground-station and data-processing facilities to accommodate the simultaneous use of TOS and Tiros M/ITOS during the initial, developmental flights of the new series of satellites.

An underlying requirement for the development of operational systems is cost effectiveness. In keeping with this requirement, the second-generation satellite has been configured to the capabilities of the Delta launch vehicle, a reliable, low-cost launch vehicle. In addition, the system design calls for placing all sensors on a single satellite, rather than having separate APT and AVCS satellites. Thus, only one satellite and one launch, as opposed to two satellites and two launches in the present operational system, will be required to meet the mission requirements for local and global data readout.

Like TOS, Tiros M/ITOS offers the capability for providing daytime, direct, real-time APT readout to stations throughout the world, and readout of stored AVCS daytime observations to NESC. However, Tiros M/ITOS will also be configured with a scanning radiometer (SR) subsystem capable of daytime and nighttime cloud-cover observations for direct readout of local data and stored readout of global data. The SR has a visual channel of 0.5–7 μm, with a resolution at the local vertical of 2 nautical miles (3.7 km); its infrared channel is 10.5–12.5 μm, with resolution equal to 4 nautical miles (7.4 km). Use of the SR subsystem with the AVCS will enable photocoverage of the entire earth every 12 hours.

All of the primary sensors will be supplied in redundant sets on Tiros M/ITOS to provide the necessary backup in the event of failure or degradation of a device.

In addition to the primary sensors, the Tiros M/ITOS satellite will contain the following secondary sensors:

- A flat-plate radiometer, developed by the University of Wisconsin, to provide global measurements of the earth's thermal energy.
- A solar proton monitor, developed by the Applied Physics Laboratory of Johns Hopkins University, to measure the solar proton energy at the satellite's orbital altitude.

The capability for using all of these sensors on a single spacecraft results from the use of an advanced stabilization technique on the Tiros M/ITOS satellite. This type of stabilization allows the main body of the spacecraft to rotate at only one revolution per orbit. Thus, the sensor side of the satellite always faces toward earth. This is in contrast to the basic TOS satellites, in which the entire spacecraft spins at either 10.5 or 9.2 r/min, depending on the basic timing requirements of the sensors (APT or AVCS) employed.

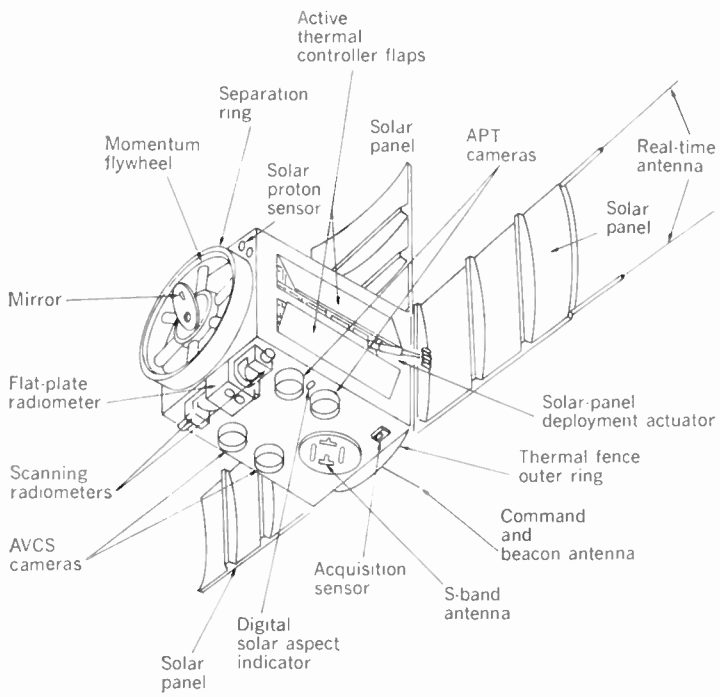
System design. The general physical configuration of the Tiros M/ITOS satellite is shown in Fig. 7. The satellite is a rectangular, box-shaped structure, with each of the sides measuring approximately 122 cm in length and 102 cm in width. On the bottom of the structure, a cylindrical transition section attaches to the 94-cm-diameter ring section of the second stage of the launch vehicle.

The key to the control of the Tiros M/ITOS satellite is the momentum flywheel system. The spinning flywheel, coupled through a bearing to a despun platform, maintains effective stabilization. The flywheel axis is collinear with the pitch axis and contains a scanning mirror that will enable fixed infrared bolometers to detect sky–earth and earth–sky transitions in their field of view. The satellite's pitch-control system will regulate the speed of the motor-driven flywheel based upon position and rate signals derived from these infrared bolometers.

TOS-type QOMAC coils will be used to correct roll and yaw errors, as well as to perform the initial orientation maneuver. The MBC coils will be utilized to correct for the residual magnetic dipole and provide the basic one-degree-per-day precession rate to track the orbit regression of a sun-synchronous orbit. A magnetic spin-control coil will control momentum about the pitch axis, and liquid dampers will reduce satellite nutation.

As shown in Fig. 7, the three solar panels will be mounted along the main body of the satellite with their hinge lines at the top of the structure; during launch these panels will be held flat against the sides of the spacecraft in a stowed position. Once the satellite achieves mission mode, the panels will be deployed by actuators.

FIGURE 7. The Tiros M/ITOS satellite in mission mode.



In the deployed position, they will be normal to the sides and parallel to the top of the structure and will be approximately in the orbit plane.

Thermal control will be achieved by the application of passive and active thermal-control techniques. Most of the satellite is covered by multilayer insulation blankets, except for the primary sensor openings and the areas used for the active control device. Passive thermal control will be provided by a variable absorptivity device, designated the "thermal fence," mounted on the top of the satellite. As the sun angle varies with respect to the two vertical walls of the fence, the amount of solar absorption will also vary. The heights of the fence walls were selected to provide maximum heat input at a sun angle of approximately 60 degrees. This passive-control device will, by itself, maintain the satellite temperatures within design limits. However, an active-thermal-control system will also be utilized to augment the passive design, providing a narrower range of temperature variations throughout the satellite's mission life. This active device consists of thermal flaps, opened or closed to vary the effective emissivity of the spacecraft.

All of the satellite's electronic equipment will be mounted on three load-carrying sections within the structure or main body of the satellite. The equipment will be arranged on two side-wall members and on the base section. Sufficient volume will be provided for additional equipment, and the simple structural design will permit a variety of possible layouts. The two side walls will support the basic camera and recording subsystems, whereas the base structure will contain command, control, power supply, and communications equipment. The scanning radiometers will be mounted on the base section, at the lower edge of the earth-oriented side.

System operation. The Tiros M orbit-injection sequence will be initiated immediately after separation from the upper stage of the launch vehicle. The events from injection to the achievement of the mission-mode attitude, under nominal conditions, are expected to take up to 24 hours. Upon separation from the upper stage, the satellite will be spinning at approximately 3.5 r/min, with the spin axis approximately normal to the orbit plane. This spin rate will provide the required momentum value for controlling satellite attitude.

The momentum wheel in the pitch control system (stabilite) will then be spun-up to 115 r/min. At this point, the satellite will be completely stabilized about the spin axis of the flywheel. Then, the magnetic torquing coils will be employed to adjust the momentum vector (spin axis) from the initial injection attitude to mission mode, in which the spin axis will be normal to the orbit plane.

The pitch-axis control system will be commanded to achieve local orientation of the sensor platform by transferring most of the total momentum of the satellite to the flywheel. The spin rate of the flywheel will increase to 150 r/min, while the main body of the satellite will decrease to one revolution per orbit to keep the sensor side of the satellite facing earth throughout the orbit.

When the satellite approaches mission-mode attitude and desired spin rate, the solar panels will be deployed. In the deployed position, the panels will be in the orbit plane and fully illuminated by the sun. Under nominal orbital conditions, the sun vector will be at 45 degrees to the spin axis; however, complete mission operations can be realized with a sun angle within the range of 30 to 60 degrees with respect to the spin axis.

Each APT and AVCS camera on Tiros M/ITOS satellites will scan an area similar to that covered by the TOS satellite cameras, and each camera will be programmed for an 11-picture sequence on each orbit. The SR sensor will provide continuous coverage (i.e., will scan continuously) whenever it is turned on. Although this sensor will normally be used for nighttime cloud-cover observations, it will also be used for daytime observation when desired.

The secondary sensor data (from the flat-plate radiometer and the solar proton monitor) will be stored in serial, digital form on two tracks of an incremental recorder. A third track will record timing data.

The Tiros M/ITOS communications link will utilize the same beacon telemetry, APT transmission, and command links employed on TOS; however, the stored video data for the AVCS cameras, the SR sensors, and the secondary sensors will be transmitted at an S-band frequency, nominally 1.7 GHz.

The capacity of the Tiros M/ITOS command and control subsystem has been increased over that of the TOS subsystem because of the greater number of commands required for the multiple-sensor configuration used on Tiros M/ITOS and to provide for future growth.

Summary

The Tiros Operational System and the second-generation ITOS satellites represent the culmination of an orderly and progressive research and development program initiated over ten years ago. The routine practical application of meteorological satellites has been one of the most important products of this first decade in space, and today many countries of this planet are deriving direct, beneficial use from the Tiros Operational System.

Abraham Schnapf received the B.S.M.E. degree from the City College of New York in 1948, and the M.S.M.E. degree from Drexel Institute of Technology in 1953. He has been project manager of the Tiros and TOS programs at RCA Astro-Electronics Division, where the first Tiros program was started by NASA in 1960, and has had responsibility for the management of design and fabrication of ten Tiros and nine ESSA weather satellites—all of which were successful. In 1966, his responsibilities were increased to include the Tiros M/ITOS program. From 1950 to 1958, Mr. Schnapf was with the RCA Airborne Systems Department, where he managed the development and design of the Automatic MOD II Shoran Bomb Systems; he was also responsible for the design and development of an advanced airborne weapon system for Mach 2 fighter interceptors. From 1948 to 1950, he was responsible for the design of lighter-than-air car structures, jettisonable 1500-gallon fuel tanks, cockpit enclosures, and Radomes for the Goodyear Aircraft Corporation. A professional engineer of the State of New Jersey, Mr. Schnapf is also an associate fellow of the AIAA and a member of the New York Academy of Sciences, holds a number of patents, and has presented and published numerous professional papers. He has been cited in the "Aviation Week and Space Technology" Annual Laurels, is listed in "American Men of Science," and received the American Quality Control Society's 1968 award for Tiros/ESSA reliability.

The electronic highway

Significant work has been done in the field of highway automation, which represents one partial solution to some of the enormous traffic problems predicted for 1990 and thereafter

Robert E. Fenton, Karl W. Olson *The Ohio State University*

Although high-speed, electrically powered surface and subsurface transportation will help to alleviate future traffic congestion in some localities, a majority of the public will probably continue to prefer the mobility, privacy, and freedom afforded by the automobile. Therefore, some form of automation for individual vehicles appears to be desirable—in fact, necessary—if complete highway chaos is to be avoided. The approach described in this article involves the concept of a dual-mode system, whereby the vehicle (which must be specially equipped) is manually controlled on nonautomated roads and automatically controlled on automated ones.

An examination of traffic conditions today—congested roadways, a large number of accidents and fatalities, and extremely powerful automobiles—indicates the need for improvements in our highway system. Unfortunately, these conditions will be much worse in the next decade, for it is predicted that the total number of vehicles registered in the United States in 1980 will be 62 percent greater than in 1960, and 75 percent more vehicle miles will be traveled.[1] If one should look further ahead to the turn of the century, he would see vast sprawling supercities, with populations characterized by adequate incomes, longer life-spans, and increased amounts of leisure time. One predictable result is greatly increased travel. The resulting traffic situation could be chaotic, unless some radical changes are instituted beforehand.

It is obvious that the traffic problems cannot be solved simply by building more and larger highways, for the costs are too high, both in dollars and in the amount of required land. Many alternative solutions have been suggested: high-speed surface rail transportation; a high-speed, electrically powered, air-cushioned surface transportation system capable of speeds up to 350 miles (560 kilometers) per hour; high-speed travel in a deep rock tunnel; and so on. It is quite probable that either these or similar approaches will provide a partial solution. However, in the opinion of the writers, a majority of the public will not be satisfied with only city-to-city transit or even neighborhood-to-neighborhood transit via some form of public transportation. One needs only to witness the common use of private automobiles where such transit already exists. The role of a personal transportation unit is certainly justified by the mobility, privacy, and freedom afforded the occupants. It seems certain that this freedom, which dictates the spatial pattern of their lives, will not be relinquished.

In this light, one satisfactory solution would be the automation of individual vehicles. This approach has been examined by a number of researchers, for in addition to the retention of the individual transportation unit, it appears that considerable improvement in highway capacity and safety as well as a considerable reduction in driver effort can be achieved. However, there is an extremely large number of possible systems for achieving this goal—the writers have counted 1296—and great care must be exercised so that an optimum or near-optimum one is chosen.[2] In spite of this large figure, a number of researchers in the United States and in Great Britain have exhibited close agreement on the general structure of the system.

One system concept

The most frequently suggested system involves a roadway complex that consists of both automated and nonautomated roads. The main highways would be equipped for automation, but the various rural roads and urban streets would not be so equipped. The automated highways would be multiple-lane structures built for the carrying of both individual passenger and commercial vehicles. However, it might be desirable to set aside separate lanes for the latter.

An individual vehicle would enter the system at a special entrance point where—if it passed a rapid auto-

matic checkout—the driver would indicate his destination, and the vehicle would move to an entrance ramp from which it would be automatically merged into the traffic stream. The traffic-stream velocity would be fixed by a central traffic controller and would be dependent on such factors as weather, roadway conditions, and the state of the traffic stream. Once in the traffic stream, the vehicle would remain under automatic control until the driver's preselected exit was reached. Then the vehicle would be guided off the highway and onto the exit ramp, and control would be returned to the driver.

In the event of vehicle disability, the driver would signal that he had a problem, and the vehicle would be ejected from the main traffic stream. If it were operating, it would be routed to the nearest emergency exit. If it were not, the use of one lane would be lost until the vehicle could be moved off the highway. Hence, it would be necessary to direct the mainstream vehicles temporarily around the disabled one. Clearly, some provision would have to be made for clearing the roadway as quickly as possible.

Implicit in this approach is the concept of a dual-mode system—that is, one in which a vehicle is manually controlled by the driver on nonautomated roads and automatically controlled on automated ones. Thus, a driver would have essentially all of the mobility he has today. In contrast, several suggested single-mode systems involve the use of vehicles only on an automated guideway; thus, a user would probably be quite restricted as to destination.

It is expected that with the introduction and extended use of microcircuits, it will be possible to install all necessary equipment in the vehicle for several hundred dollars. The total investment in computers and highway-based sensors would probably average anywhere from $20 000 to $200 000 per lane mile (about $12 000 to $120 000 per lane kilometer), depending on the form of the chosen system and future technological advances.

One can expect two principal returns from such an investment: greatly increased lane capacity at high speeds and a reduction in the number of highway accidents. Estimates of the former range up to 800 percent and would depend, of course, on the chosen system design. The expectation of fewer accidents arises from the fact that an electronic system can provide a shorter reaction time and greater consistency than a driver can.

The evolutionary process

The evolutionary system concept is based on a highway complex of the future that will have evolved from the one of today. In contrast, a revolutionary system would not involve a natural outgrowth of our present highway system. The evolutionary approach was first discussed by Zworykin and Flory,[3] who pointed out the impossibility of converting many thousands of miles of highway and tens of millions of vehicles to automatic control overnight. It would be exceedingly difficult to build a controlled road and immediately require that all cars traveling on this road be equipped with the necessary controls, and herein lies the major disadvantage of a revolutionary approach. It is possible, of course, that this disadvantage could be offset by the advantages of a radically different system for optimum operation.

An evolutionary system would probably be implemented in three overlapping stages. First, one would have the installation and use of various driver aids so that the driver would be a more effective decision maker and improve the performance of his driver–vehicle system. The second stage might involve the gradual introduction of various subsystems for partial automatic control. The third would include the transition to complete automatic vehicle control. Each of these stages must be realized within the confines of one system so that the addition of each feature would contribute to the ultimate system.

The improvement of the decision-making capability of the driver has long been under study, and some promising methods have been reported. One of these is the Experimental Route Guidance System (ERGS), which is being developed at the General Motors Corporation under the sponsorship of the U.S. Bureau of Public Roads.[4] This method is an attempt to achieve dynamic route control by guiding drivers over nearly minimal time paths on a roadway network. A second is the Driver Aid, Information and Routing System (DAIR), which aids the driver by audio signing, a visual sign-minder (a method of displaying roadside signs within the car), code and voice emergency communication, and route guidance.[5] A third technique involves aiding the driver during the merging maneuver by means of specially designed entrance ramps with computer-controlled ramp-signaling devices.

A number of efforts have been made to improve driving performance, especially in the important driving task of car following. It has been found by several investigators that a tracking type of display, either visual or tactile, can be used to aid a driver and greatly improve his car-following performance.[6,7] Such driver aids would also be introduced during the first stage so that the performance of the driver–vehicle system would begin to approximate that of an automatic system—at least under equilibrium conditions.

Some of the subsystems that might be implemented during the second stage include automatic speed control, an automatic driver-alerting system, automatic steering, partial automatic control of the throttle and brakes, and emergency braking. One possible sequence of implementation might first include more widespread installation and use of automatic speed-control devices. A logical second step would involve the installation of a simple collision-warning device that would alert a driver when he came within a specified distance of a lead car. Since this distance would be a function of speed, it would be necessary to install a simple computer in the vehicle. This speed-control/warning-device combination

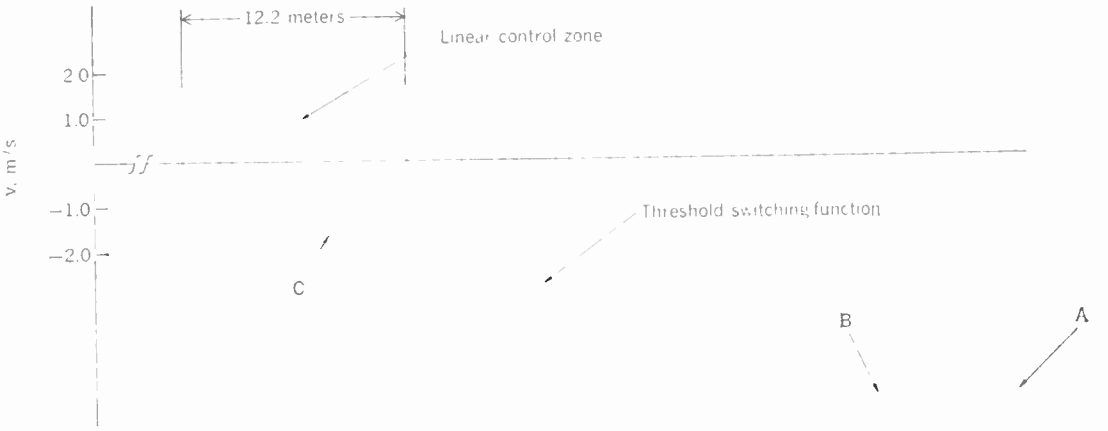


FIGURE 1. A typical overtaking situation as viewed on the v-h phase plane.

would presumably be used primarily for long trips under conditions of low-density traffic. A third possible step might be the installation of an automatic steering system, so that in situations of low-traffic density the driver need function only as supervisor—monitoring the system to ensure its satisfactory performance. (The problems of monotony and a lack of driver attentiveness would be important ones to be considered in these first steps.) In dense traffic the driver's task would be reduced to that of longitudinal control of his vehicle. However, an instantaneous override feature should be included on the automatic steering for emergency situations.

A first step in the third stage would be the installation of an automated longitudinal controller, which would naturally evolve from the relatively simple headway warning device previously used. Finally, the installation of an automatic merging and exiting system would result in a fully automatic highway.

This is clearly an oversimplified view of only one of many possible evolutionary sequences. It would require a number of years to implement, since it would be desirable—indeed, essential—to have each new feature tested and accepted by the driving public before proceeding to the next one. Further, if unnecessary chaos is to be avoided, each added feature must contribute toward the ultimate system and not increase the danger to unequipped cars.

The system components

It is not possible at present to specify completely all of the required system components, because the necessary knowledge is not available. However, the essential aspects can be discussed in general terms, and some of these in detail.

It would be highly desirable to have a limited number of vehicle types so as to simplify the interface between the roadway and the vehicle. In practice, this would probably mean that certain technical performance characteristics must be the same for all vehicles of a given type; however, one would expect to have the same wide range in vehicle appearance and comfort features that is available today.

It is probable that vehicles would be powered by the reciprocating internal combustion engine; however, a number of other prime movers—the dc motor, the gas turbine, the steam engine, and the linear induction motor—may be available and practical in the near future. The eventual choice will probably be dictated by such factors as air pollution and the continuing availability of cheap fossil fuel.

One attractive possibility involves the use of electrically powered cars, which would be self-powered via batteries on nonautomated roads and externally powered on automated ones. Here, power would be supplied through a pickup probe protruding from the vehicle, and control could be obtained by simply controlling the power flow.

An important question that is frequently raised is the advisability of allowing the driver to override the system. If he were able to do so, there would be a large measure of randomness in the system, which would be undesirable from a standpoint of both safety and system efficiency. Thus, while it would be necessary to allow the driver to be able to regain control during the evolutionary stage, it would probably be unwise to allow him to do so after the system was fully automated. In any case, it is clear that both during the evolutionary process and after it is completed, one must have a set of driver controls that would be compatible with the automatic system. This realization has led to research on various types of unusual vehicle control devices, including several different control-stick configurations.

A second important question concerns the location of the system's decision-making capability. In the system just described, this capability can be either contained within the roadway or split between the roadway and the vehicle. In the first case, one would have a centrally located computer, which would observe a given section of highway and communicate the necessary control information to each automobile. On the other hand, one could put some decision-making capability both into small digital computers located along the highway and into a small vehicle-based computer. One such possibility involves the use of a vehicle's decision-making capability in low-density traffic where the intersections are widely separated, and an external capability for use near intersections in congested areas.

A longitudinal control system for the automatic control of braking and acceleration would be required in each vehicle. The nature of this system will be dependent

FIGURE 2. Instrumented test vehicle.

on the chosen type of vehicle prime mover. If the power source were fossil fuel, an electrohydraulic system would be used; however, if an electric energy source were chosen, an all-electric control system would probably be employed.

The use of a longitudinal control system requires a means for measuring intervehicular spacing and relative velocity, which would normally be the inputs to this system. Various techniques for obtaining these quantities, including the use of radar, infrared communication, coherent-light techniques, and an electronic "block" system in the pavement, have been suggested. Note that some of these approaches would also require a means for communicating information to the vehicle. If, for example, a block system were used, a means for communicating between a block and a vehicle over that block must be available.

There is also a need for communication links to a central station so that there will be a complete "picture" of the traffic state at all times. If a centrally located computer is used for either partial or full stream control, it must also be able to communicate with the vehicle. It is not difficult to see that a satisfactory communication system would probably be very large and complex.

Each vehicle must contain a system for automatic lateral control or steering, with the basic system structure depending on the type of vehicle prime mover. If a linear induction motor were used as the drive unit, an effective guidance signal could be obtained from the motor field. On the other hand, if a conventional power plant were used, an external steering reference, such as an electromagnetic field, must be provided. In both cases, a controller—probably a simple position servomechanism—would be needed for turning the front wheels.

It is especially important that the number of vehicle breakdowns be minimized. This can be effectively accomplished by automatically checking out each vehicle as it enters the system. If the vehicle passed the test, it would be allowed to proceed onto the electronic highway; however, if it failed, it would be routed to a nearby service facility for the necessary repairs. The checkout would have to be extremely rapid so that it would not slow down the operation of the entire system.

Promising developments

A number of imaginative approaches to highway automation have been suggested and some experimental work has been done. The results obtained from some of these investigations appear quite promising.

Automatic longitudinal control. A method of longitudinally controlling individual vehicles is required for nearly every conceivable type of automated highway. Since no decision has been made as to which type is most desirable, any controller developed now should be sufficiently flexible that it could be used for most types.

This requirement can be satisfied by the use of a multimode control system such as the one being developed at The Ohio State University.[5-9] It may be conveniently described using a two-dimensional phase plane, where the relative velocity v between two vehicles is plotted versus the spacing or headway h between those vehicles. The relation between these variables is $v = dh/dt$. Time is not shown explicitly on this plane; rather, as time progresses, the point representing the instantaneous relative velocity and headway traces a "trajectory" on the plane.

Consider the phase plane shown in Fig. 1. It is divided into a number of regions—a certain mode of control

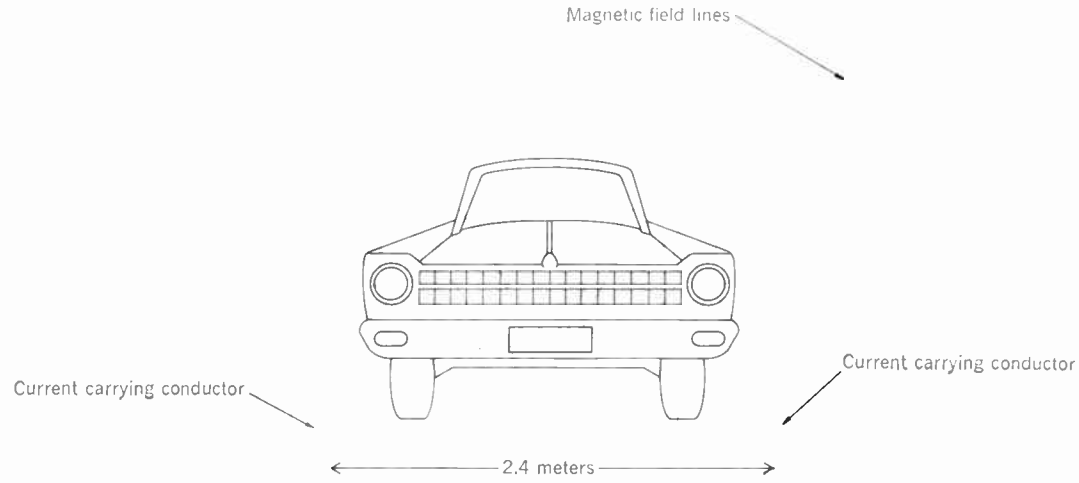
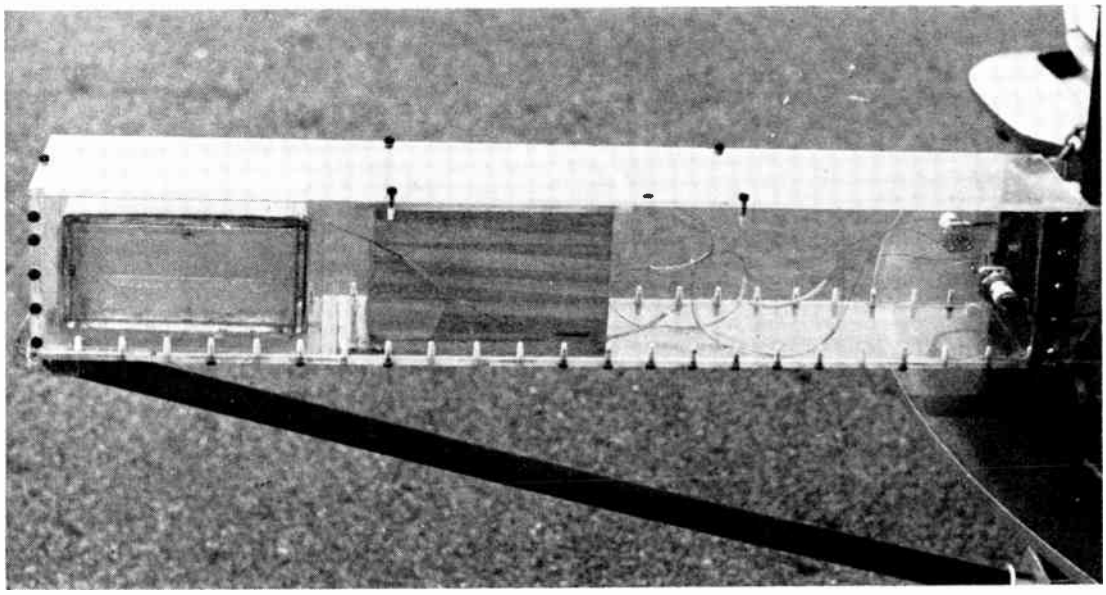


FIGURE 3. Magnetic field established by reference system.

FIGURE 4. Sensing coils for automatic steering system.



being associated with each region. The regions are separated by switching boundaries, and thus as a trajectory "moves" and crosses a boundary, the control mode changes. This change is made by a simple logic system that associates each point in the phase plane with a particular mode of control.

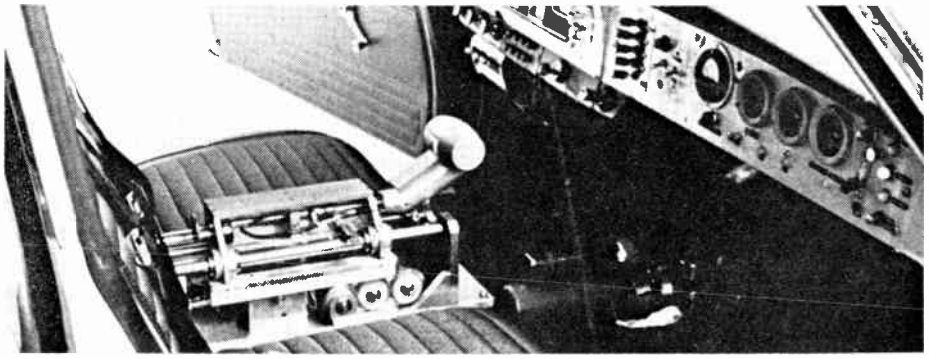
The operation of this system can easily be understood from an examination of the experimentally obtained trajectory shown in Fig. 1. Initially (point *A*), a lead and an overtaking controlled vehicle were both traveling at a constant speed, with the control system in the latter functioning as a speed regulator. At point *B*, the controlled vehicle entered the "zone of influence" of the lead car, and the control system responded by decelerating the vehicle at a constant rate. The trajectory progressed into the rectangular region at point *C*, which corresponds to a steady-state, car-following mode of control. This selection of control modes and switching boundaries is but one of a number of possible choices, and another might result in superior performance. However, it seems clear that a multimode system will be required because of the diverse nature of the number of required vehicle responses.

The vehicle shown in Fig. 2 was used in the testing of this system. The three main control functions—braking, acceleration, and steering—were accomplished by the use of electrohydraulic control systems. Excellent test results, of which Fig. 1 is typical, were obtained at speeds up to 65 mi/h (104 km/h).

It should be noted that the necessary headway and relative velocity information was obtained either via the use of a "phantom" lead car or a mechanical "yo-yo." In the former case the lead car was represented by a voltage fed into the on-board computer—the rationale being that the controlled vehicle cannot discern the difference between a signal from a phantom car and one from an actual lead car. In the latter case, a mechanical take-up reel was attached between the lead and follow-

FIGURE 5. Front-mounted control stick.

FIGURE 6. Control stick mounted between the driver and the passenger.



ing cars, and headway and relative velocity were obtained by measuring the reel displacement and angular velocity. It is obvious that a much more practical method must be developed before the automatic control of vehicles becomes a reality.

Automatic lateral control. The area of greatest experimental effort has been that of automatic lateral control or automatic steering. Experimental tests of such a system were reported as early as 1962,[10] and several organizations are now working on such systems.

The most widely tested external steering reference consists of a single cable buried in the center of the controlled lane. The cable is excited by an alternating current, which produces a magnetic field that induces a voltage in each of two tuned pickup coils mounted on either side of the vehicle's center line. The difference between these voltages is used to determine the location of the vehicle relative to the center line of the lane and to actuate the steering control unit. Experimental testing of such a system has been conducted by the General Motors Corporation in conjunction with RCA[10] and by the Road Research Laboratory.[11]

A conceptually similar system, now being developed, consists of two wires separated by about 2.4 meters and excited by a 7000-Hz signal to produce a magnetic guidance field (see Fig. 3).[8] This configuration was chosen because it yields a horizontal field component that varies linearly over the 2.4-meter spacing between the wires. In contrast, a single-wire system gives a linear characteristic that is no wider than the spacing of the two sensing coils on the front of the vehicle—approximately 0.6 meter. The vertical component of the field set up by the two-wire configuration is constant in a 1.22-meter-wide region in the center of the controlled lane. Thus, it seems possible that it can be used both as a phase reference and as a control variable in an automatic gain control.

The magnetic-field sensors for this system consist of

two coils: the reference coil with its axis in the vertical plane, and the other coil with its axis horizontal and parallel to the lateral axis of the vehicle. The coils are mounted in a plastic pod on the front of the vehicle, as shown in Fig. 4. It should be noted that no attempt has been made to achieve a minimum size configuration up to this time.

This system has been extensively tested up to speeds of 70 mi/h (112 km/h), and excellent performance was achieved when the tests were conducted on roads without buried steel reinforcing rods. When such rods were present, however, the magnetic-field strength fluctuated over a substantial range. In addition, the null point of the field varied laterally, thus causing the vehicle to track a rapidly shifting null and resulting in an uncomfortable ride. The investigators have made an initial unsuccessful attempt to overcome these difficulties by using a combination of passive filters and automatic gain control. It seems reasonably probable that satisfactory performance can be obtained by appropriate modification of this approach.

Compatible manual mode. The major automobile companies and others have experimented with unusual types of vehicle control devices, such as wrist-twist steering and the control stick developed for the General Motors Firebird III. One university research group has designed and tested similar devices specifically for use with an automated vehicle.[7] Their front-mounted control stick is shown in Fig. 5, and a side-mounted one in Fig. 6. In each case, the stick replaced the conventional controls —that is, the steering wheel, and the brake and accelerator pedals. The primary reason for using a control stick is the relative ease of obtaining compatibility between the automatic and manual modes with this device.

It was observed that the test drivers readily adapted to stick control and were quite confident of their driving abilities after approximately half an hour of practice. Their excellent driving performance in several car-following experiments also gave an indication of the merits of a control stick. However, one potentially serious problem is public acceptance of such a new and drastically different type of control device.

Conclusions

There seems little question that vehicle automation is technologically feasible; however, a tremendous amount of effort in both research and development will be required before a satisfactory automatic system is in operation. This effort must involve not only vehicle-control studies, but also an intensive investigation of the present driver–vehicle complex, since the knowledge gained will be necesssary for the proper specification and introduction of the control system components. Further, the need exists for intensive overall system studies so that optimum strategies can be chosen for headway spacing control, merging and lane changing, and the interfacing of automated highways with other modes of future transportation.

The authors were introduced to this subject by Dr. Robert L. Cosgriff, who has long advocated a broad systems approach to solving the problems of highway traffic, and directed early efforts at The Ohio State University toward this end. His efforts in their behalf were very much appreciated. The authors also wish to acknowledge the continuing support of their reseach efforts in highway automation by the Ohio Department of Highways and the U.S. Bureau of Public Roads.

REFERENCES

1. Wilbur Smith and Associates, "Further highways and urban growth," The Automobile Manufacturers Association, New Haven, Conn., Feb. 1961.

2. Fenton, R. E., "A strategic model for highway automation," in "A study of highway research needs and resources," Report EES299, Transportation Research Center, The Ohio State University, Columbus, Ohio, June 1968, Appendix E.

3. Zworykin, V. K., and Flory, L. E., "Electronic control of motor vehicles on the highway," *Proc. 37th Annual Meeting of the Highway Research Board*, pp. 436–451, 1958.

4. Carter, A. A., Jr., *et al.*, "Highway traffic surveillance and control research," *Proc. IEEE*, vol. 56, pp. 566–576, Apr. 1968.

5. Hanysz, E. A., *et al.*, "DAIR—a new concept in highway communications for added safety and driving convenience," *IEEE Trans. Vehicular Technology*, vol. VT-16, pp. 33–45, Oct. 1967.

6. Gantzer, D., and Rockwell, T. H., "The effects of discrete headway and relative velocity information on car following," presented at the 45th Annual Meeting of the Highway Research Board, Washington, D.C.

7. Fenton, R. E., and Montano, W. B., "An intervehicular spacing display for improved car-following performance," *IEEE Trans. Man-Machine Systems*, vol. MMS-9, pp. 29–35, June 1968.

8. Fenton, R. E., *et al.*, "One approach to highway automation," *Proc. IEEE*, vol. 56, pp. 556–566, Apr. 1968.

9. Bender, J. G., "Experimental studies in vehicle automatic longitudinal control," Rept. EES276A-5, Communication and Control Systems Laboratory, Department of Electrical Engineering, The Ohio State University, Columbus, Ohio, Aug. 1968.

10. Flory, L. E., *et al.*, "Electronic techniques in a system of highway vehicle control," *RCA Rev.*, vol. 23, pp. 293–310, Sept. 1962.

11. Giles, G. C., and Martin, J. A., "Cable installation for vehicle guidance investigations in the new research track at Crowthorne," Rept. RN/40 57/CGG, JAM, Road Research Laboratory, Crowthorne, England.

Robert E. Fenton (M) received the B.E.E., M.Sc., and Ph.D. degrees in electrical engineering in 1957, 1960, and 1965, respectively from The Ohio State University, Columbus. From 1957 to 1960 he served in the United States Air Force as an electronic intelligence officer. In September 1960 he joined the staff of the Electrical Engineering Department at Ohio State University. He was appointed associate professor in 1968. He is also an associate supervisor in the university's Communication and Control Systems Laboratory, where he is currently directing research work on highway automation. Dr. Fenton is a member of the Highway Research Board Committee on Highway Communication and is technical program chairman for the 20th Annual Conference of the IEEE Vehicular Technology Group. He has authored or coauthored some 20 papers dealing with the driver-vehicle system and the automatic longitudinal control problem.

Karl W. Olson (M) is an associate supervisor in the Communication and Control Systems Laboratory of the Electrical Engineering Department at The Ohio State University, Columbus. At the present time he is concerned with the instrumentation of automated vehicles and is directing research toward the development of an automatic vehicle-steering system. Prior to engaging in this program he was responsible for design and testing of a special-purpose computer for pattern recognition. He received three degrees in electrical engineering from The Ohio State University: the combined B.E.E. and M.Sc. degrees in 1959 and the Ph.D. degree in 1965. Dr. Olson is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi. He is the author or coauthor of a number of technical papers published in various journals.

Air pollution and electric power

Like education, pollution abatement is universally conceded to be a Good Thing—but there are tradeoffs to be considered. We must balance what we would like to have against what it will cost

Bruce C. Netschert *National Economic Research Associates, Inc.*

No one is in favor of air pollution. However, it is not possible today to eliminate it; the best we can do is to lessen it. We must balance what we would like against what it will cost. To the economist this suggests the cost–benefit approach by which the costs of a proposed decision are added up, a value is computed for the benefits, and if the latter exceeds the former there is economic justification. With this in mind the author discusses the various developments in the area of pollution abatement. For the energy technology of the future these developments point to two general paths, the first leading to an 'electric economy' and the second to a 'gas economy.'

"Just such disparity
As is 'twixt air and Angels' purity . . ."
John Donne

Pollution is an ugly word, and air pollution is something we instinctively abhor. None of us, however, is free from sin on this score, for we are polluters, one and all, even if we don't drive or smoke—as we are quickly reminded when a ventilating system fails. Nevertheless, no one is in favor of air pollution. It is not like water fluoridation or the protective tariff, on which there are clear divisions of opinion. Rather, pollution abatement is like education; there is universal agreement that it is a Good Thing. This being so, the observer from Mars might well wonder what all the commotion is about. Why don't we just pass a law forbidding air pollution?

This was, in fact, tried. In an attempt to eliminate the smoke problem, Edward I of England forbade the use of "sea coal" in London in the 14th century, with the death penalty for repeated offenses. The outcome of this attempt is indicated by the fact that Elizabeth I, almost 200 years later, also felt compelled to proclaim coal burning illegal. It is even less possible today to eliminate air pollution than it was for Edward and Elizabeth. The best we can do is lessen it.

Not only is some pollution inevitable; the economist concludes that in the absence of specific social action to the contrary, the mess we now find ourselves in was also inevitable. The "invisible hand" of classical economics, the market mechanism on which we still basically rely, is imperfect; here, as in numerous other instances, it is unable by itself to bring about the social optimum. The costs resulting from each individual act of air pollution are not incurred by the emitter, but by society as a whole. Thus, since the emission is costless to the emitter, there is no economic motive for him to do anything about it.

Tradeoffs

Pollution abatement cannot be done free; we must balance what we would like to have against what it will cost. Since it is neither possible nor desirable to attain Dr. Donne's "angels' purity," it is necessary first to define the ultimate goal or limit of air pollution abatement beyond which it would make no sense to pursue abatement further. In the last analysis, this goal is the threshold level, the demarcation between pollution in the meaningful sense and what is conveniently described by the oxymoron, "nonpolluting contamination." In this respect there is a similarity to another environmental pollutant, radioactivity—there is a natural background level from which we could, if we desired, isolate ourselves, but to do so would be inherently wasteful of our time and resources. There is further similarity in the matter of possible threshold exposure and the complications of instantaneous and cumulative doses. As you know, the question of the threshold in establishing radioactivity safety levels has generated a good deal of incandescent argument in health physics, and here, too, there is a parallel in air pollution. In many instances, especially the sulfur oxides, the threshold concentration that produces adverse physiological effects is still controversial.

In any event, within that limit, whatever it is, we must next establish a working goal toward which to aim our abatement efforts. There is only one way in which we can

make this decision: that is to compare the value of what we obtain with the cost of obtaining it. In the most general sense this is what is involved in any economic decision, made by anyone. To the economist it also suggests "cost–benefit analysis," the approach developed for similar decision making where social costs are involved, as in flood control, for example. The costs of the proposed decision are added up, a value is computed for the benefits, and if the latter exceeds the former there is, *ipso facto*, economic justification. In air pollution abatement, where there is a wide range of possibly justifiable actions, the cost–benefit approach offers the added advantage of determining the best action in terms of the optimal cost–benefit relationship.

Cost–benefit analysis has been developed into an elaborate technique in the justification of large public works such as dams, harbor improvements, and the like. One of the problems in applying it to air pollution decision making, however, is the far greater complexity of choices and the potentially wide ramifications of any particular choice. Consider, as an example, the reduction of sulfur oxide pollution from central power stations, one of the aspects of air pollution that has been emphasized in control programs to date.

One approach has been to set a standard for the maximum sulfur content of the fuel that is burned. If the fuel user desires to continue to burn the same type of fuel as before (coal or oil), he can meet this standard in a number of ways: by using fuel in which the naturally occurring sulfur content is low enough to meet the standard, by desulfurizing the fuel so that it meets the standard, by blending high-sulfur fuel with low-sulfur fuel, or by any combination of these. Now it happens that the supply of coal and oil with high sulfur content is higher than that with low sulfur content, so the price is lower, and this, of course, is why it was being used to begin with. Obviously, then, compliance with the air pollution standards means higher fuel costs, which are ultimately passed on to the consumer.

Cost analysis

It would appear to be a simple matter to take the difference between the high-sulfur fuel cost and low-sulfur fuel cost as the measure of the cost of the control program to society. There is much more to it than that, however. If the user decision is to go to foreign sources of naturally low-sulfur fuel oil, it means the use of African oil rather than the Venezuelan oil that previously constituted the chief supply. The result could be a net shift in the balance-of-payments position of the United States, since Venezuelan oil can be desulfurized either in the Caribbean or in the U.S. This balance-of-payments effect, as well as such noneconomic aspects as the implications for our international relations, must be included in the cost calculation. With desulfurization, the refiner also has a choice, over a wide range of proportions, between producing fuel oil or other, more valuable, petroleum products. The desulfurization cost itself may therefore be difficult to determine, since it may not be fully reflected in price.

The difficulties of costing this particular means of air pollution control are compounded, moreover, by the fact that the control decision is necessarily an *ex ante* one; it is before the fact. There is no way of forecasting accurately, beforehand, which alternative or combination of alternatives the fuel users will choose. It might appear to be a straightforward matter of comparing the costs of the various alternatives and assuming that the fuel users, being economically rational, will choose the one with lowest cost. This assumes, however, that the judgment of the fuel users and the pollution control authorities on the relative costs will be the same. In the case of the New York pollution control program the problem was even worse: no desulfurization capacity existed and it was impossible to guess how much would be installed as a result of the program—if, indeed, any would be!

But these difficulties pale besides those associated with the cost–benefit determination on coal. Sulfur exists in coal in both organic and inorganic form. Removal of most of the inorganic sulfur is technically feasible and is normally accomplished as part of the preparation of coal for the market (although not necessarily for the specific purpose of desulfurization). Removal of the organic sulfur, however, is economically infeasible at present and there is no foreseeable prospect for feasibility.

The ultimate standard established by New York City is a maximum of one percent sulfur content in the fuel burned. A Bureau of Mines survey of 1964 coal production showed that 62 percent of the total U.S. production had a sulfur content of greater than one percent. This production involved 70 000 miners, or 55 percent of the total mine labor force. In six producing states (Pennsylvania, Ohio, Illinois, Indiana, Missouri, and western Kentucky) *all* of the output was coal with more than one percent sulfur, and these states accounted for 37 percent of the total mine labor force. Low sulfur content is a premium quality in the metallurgical use of coal, and three quarters of the 1964 production of low-sulfur coal was for this use, both domestic and for export.[1,2]

It is clear from this that the cost of the New York City control program, if properly measured, would have to take into account the expected adverse effect on the economy of the coal districts supplying coal to New York City. There is also a balance-of-payments effect if low-sulfur coal is to be used in that city instead of being exported.

There is an alternative—the use of other fuels. Natural gas is sulfur-free. Like the cost of coal and oil, its cost to utilities is published information, so the difference in cost can be directly compared. Again, however, such direct comparison is insufficient. The imposition of a sudden, sizable increment of demand in the market for gas may well have its own effect on the price of gas. (The gas industry is, to be sure, regulated, but this merely introduces a time lag.) This effect will depend on the supply characteristics of the gas industry, and any price changes would, in turn, have effects on the demand for gas by various types of customers. It is necessary, therefore, to consider the elasticity of both gas demand and gas supply in considering the cost of a shift to gas. (There are difficulties here, too, but this is a matter economists prefer to discuss among themselves.)

The utility fuel user can also go nuclear, although this is possible only for new plants. In the opinion of some, this would not be a cost, but a benefit. I shall not attempt to argue the point except to note two things: (1) nuclear cost figures in recent months have exhibited a volatility more appropriate to the "high flyers" of Wall Street; and (2) one of the costs of avoiding air pollution via the nuclear route can be an increase in the thermal pollution of our water resources. In any event, the significance of

the nuclear option is that in determining the cost side of the cost–benefit equation, comparison of nuclear and fossil fuel must be made.

Similarly, the utility has the option in new plants of placing them outside the urban area and thereby (perhaps) escaping the sulfur limitation. It can be assumed that this also would definitely incur costs, since if it did not, it would have already been done. And these costs, too, must be taken into account.

The goal of reduced sulfur pollution can be sought by means other than limitations on the fuels. One alternative available to the control authority is desulfurization of the stack gases resulting from combustion. This has the advantage of requiring the investment to be made directly by the emitter rather than externally (as in fuel oil desulfurization). The cost of pollution abatement here is the addition to the unit cost of power being generated.

Once again, despite appearances, the costing is not a simple matter (at least, not yet). Stack gas desulfurization is still in the experimental stage, and it is not at all clear which of the many processes being tested will be best, and under what circumstances. Much has been made of the fact that by using fuel of sufficiently *high* sulfur content, stack gas desulfurization could be made costless or even profitable through the sale of the recovered sulfur or sulfuric acid. It has been pointed out, however, that if only one half of the sulfur emitted in the combustion of coal and oil in the U.S. were recovered in the form of elemental sulfur, it would increase domestic sulfur production by one third; if the sulfur were recovered as SO_2, it would provide a quantity of sulfuric acid almost equal to its total annual domestic consumption.[3]

A second alternative to limitations on the sulfur content of fuels is the high stack. The concept of the high stack as a means of SO_x pollution abatement is based on the distinction between the SO_x concentration in emissions and ambient air concentration. This proposition is valid, however, only if it is true that through the use of high stacks, high emission levels will not result in high ambient levels. This is another area in which controversy is sharp, mainly because there are simply not enough empirical data on which to base policy. High stacks are being built, nevertheless (the newest one is over 365 meters tall), and with sufficient experience it should become clear whether dispersion can be counted on to offset high emission levels.

The foregoing demonstrates that on the cost side of the cost–benefit calculation the problem is one of complexity plus an inadequate empirical basis. Small wonder, then, that an interdepartmental committee of the Federal government concluded, after investigating abatement costs at the national level, ". . .there are no acceptable national estimates of total investment or annual cost."[4]

Benefit analysis

Let us turn now to the benefit side of the equation. The problem here is to measure, in dollar terms, the benefits resulting from the abatement action. Unfortunately, there is no direct means of doing this: What is the value of clean air, as such? The best that can be done is to use the costs incurred because of pollution; in economic terms, the "opportunity cost" of pollution abatement. If, due to pollution, I spend X dollars, the value to me of an abatement program that eliminates that expense is those same X dollars, and this also holds true, of course, at the level of society as a whole. Pollution, however, is not a one-time thing but is continuous, so that there is not a single expense but a stream of expenses. The appropriate measure is therefore the total of those expenses over the appropriate time period, discounted at an appropriate rate.

As a theory this approach is impeccable, but as a practical means of measurement its usefulness is woefully limited. (For a discussion of the measurement problem see Ref. 5.) For example: a New York State air pollution official has estimated that residents of New York City would save $800 million per year, or $50 per capita, in cleaning costs for clothing, homes, and vehicles, if the State pollution criteria were attained.[6] On the other hand, figures presented by a staff member of the National Center for Air Pollution Control indicate that for the United States as a whole the cleaning cost attributable to air pollution is in the neighborhood of $2.9 billion.[7] We all know that New York City air is dirty, but it is unlikely that it is responsible for over one quarter of the total cleaning bills of the United States attributable to air pollution. Obviously, one or the other of these figures is implausible, and given the crudity of the basic statistics from which they were derived, both of them are of doubtful validity. I do not even have much confidence that they represent the correct orders of magnitude.

On another score, attempts have been made to correlate property values with air pollution. It is evident that, other things being equal, a lot immediately downwind from a rendering plant will have a lower value than a lot 8 km on the other side. But this is no help. Nor, I fear, is it much help to know, from the findings of a regression analysis of sulfur trioxide levels and property values in the St. Louis metropolitan area, that there is a decrease in property values of $245 per single-family residence for every increase of 0.5 milligram of sulfur trioxide per 100 square centimeters per day.[8] Such precision belies the fact that no multiple regression analysis can adequately deal with the complex relationships of the many variables involved, much less the psychological factors.

It is the latter, indeed, that stop measurement in its tracks. How is one to measure the health benefits of pollution abatement (assuming we can ever reach agreement on the health *effects* of pollution)? Or the esthetic values created by abatement? Granted, one could, after the fact, compare medical and hospital costs in a polluted area before and after abatement, but what are such values compared with the value to the individual of his improved health? We are in the realm of the wholly subjective, and until we possess some of the instruments of science fiction we must be content with frustration.

We have, then, two wholly different problems on the two sides of cost–benefit analysis. On the cost side we are led into a multiplicity of tradeoffs, each of which must be explored if we are to make our decisions with maximum objectivity, and which call for better data than have thus far been available. On the benefit side we are faced with formidable, if not wholly intractable, measurement difficulties. Does this mean it is fruitless to use the cost–benefit approach, even if it is the only tool we have for approximating rational decision making? I do not think so, but before commenting further on this point let me turn to a later stage in the process of air pollution abatement—the choice of regulatory criteria and the tools of regulation.

Establishment of pollution criteria

I believe it is important to recognize that the choice of criteria itself involves tradeoffs, which, although they may not be strictly economic, have economic implications. The criteria I have in mind are the ambient air standards that are the goal of the regulation and the maximum level of emissions permitted. Compare, for example, the ambient air SO_2 concentration criteria in Table I.

There is, as can be seen, a wide variation in levels, duration, and frequency in the various criteria, which is not surprising, since one would expect that the criteria would differ for different areas, each of which has its peculiar circumstances. The significance of the figures in the present context is their indication of the almost infinite number of combinations of level, duration, and frequency that is possible, and each combination has its own tradeoff, or cost–benefit relationship. I have no idea how any of the listed criteria were arrived at, but I suspect not much consideration was given to the tradeoff aspects of the choice. To the extent that this is true not only may our air pollution regulators be making implicit or unrecognized tradeoffs on the wrong terms, but also, the tradeoffs may be the wrong ones. The possibility becomes stronger, moreover, when the choice of criteria is widened to include not only ambient air levels but emission levels (without relating them to ambient air levels) and limitations on the sulfur content of fuels.

I am not suggesting that a marginal analysis in cost–benefit terms should be undertaken to determine which criteria constitute the optimum tradeoff of costs and benefits. (Indeed, I am in no position to make such a suggestion in view of the emphasis I have placed on the lack of data.) I do suggest, however, that at the very least the authorities should be aware that in the establishment of criteria, they are engaged in tradeoffs; and it is to be hoped that within the limits of the imperfect knowledge of both costs and benefits, their choice of criteria will represent balanced judgment rather than arbitrary discretion. It would be unfortunate too if the authorities were to become wedded to initially established criteria when there is still great uncertainty with respect to their validity as pollution indicators, not to mention their efficacy as regulatory standards. It is also to be hoped, therefore, that established policy does not become dogma, and that the policy makers will keep a continuing review under way and will be flexible enough to change policy if and when a need for change is indicated.

I. Comparison of ambient air SO_2 concentration criteria in three areas (all concentrations are by volume)

Area	Ambient-Air Standards for Concentration of SO_2
St. Louis metropolitan	0.25 ppm for 5 minutes, once in any 8 hours 0.10 ppm for 1 hour, once in any 4 days 0.05 ppm for over 24 hours, once in any 90 days
San Francisco Bay	1.5 ppm for 3 minutes during the daytime, 6 minutes during 24 hours ⋮ (14 intermediate categories by 0.1-ppm steps)
State of Colorado	0.5 ppm for 1 hour 0.1 ppm for 24 hours

But there is still more to the catalog. Beyond the choice of the standard or standards there is the option in the ways in which they may be applied. Should standards apply without discrimination, for example, or should there be some proportionality (or even disproportionality), depending on the degree of emission and/or its harmfulness? The interdepartmental committee to which I previously referred investigated this matter with simulation models. One model, of a hypothetical city of two million population, indicated that the cost of attaining an assumed pollution standard (60–75 percent reduction of human exposure to SO_x and particulates), by requiring abatement only by those emitting harmful wastes and able to achieve abatement of those wastes, would be three fifths of the cost of requiring all emitters to reduce their discharges by the same proportion. A second model, covering the central power stations in 20 cities with the worst SO_x problem, indicated that by taking into account the prevailing winds and the compass location of plants, and applying the standards on sulfur content of fuel used only to those plants in the prevailing wind quadrant, the additional fuel cost incurred by pollution abatement would be only one eighth of the cost if the standards were applied to all plants.[9]

Again it behooves me, in the light of what I have been saying, to observe that I put no faith in the actual numbers yielded by the models in view of the statistical deficiencies to which I have alluded and the compounding of assumptions within the models. Regardless of the actual levels, however, the numbers suggest that significant cost differences can result from different applications of a given standard. Since the choice exists with a given standard, the tradeoff in this instance is not between the costs and the abatement results. Instead, it is a balancing of the costs imposed by the different methods of application and the costs of administering those methods. Thus, in the case of the single-city model one would expect, a priori, that across-the-board administration of a given emission standard would be less costly than selective administration. But the simpler administrative scheme involves higher costs to the polluters, and hence to the public. The tradeoff, therefore, is all on the cost side, and the optimum solution can be described, somewhat tautologically, as the "least-cost cost combination."

Tax incentives

At this point you may be starting to read more about tradeoffs in air pollution abatement than you really care to know, but allow me one more example. I have emphasized that all abatement involves costs. As an alternative to the imposition of standards and the resulting ultimate imposition of costs on the public in general, there exists the possibility of using those costs to subsidize the polluters in their abatement efforts. Public funds could be given as outright subsidy, but I shall confine my discussion to the type of subsidy that has been most frequently proposed—the tax incentive. Since a tax incentive reduces government revenues, the incentive, properly computed and administered, could be made to result in a tax loss to the government equal to the cost of pollution control—i.e., the method of abatement I have been dis-

cussing heretofore. Assuming the government wishes to maintain its revenues at the level that would prevail in the absence of the tax incentive, other taxes would be increased. In the one instance the public pays through increased costs of the goods and services it purchases; in the other it pays through increased taxes.

Now it is unlikely that the incidence of the shifted tax burden on the public will be distributed in the same fashion as the direct costs of abatement. We have, therefore, another tradeoff in the decision between the two routes of obtaining abatement. Here the social advantages or disadvantages of the shifted tax incidence must be compared with those of the direct cost incidence. The use of numerical values is impossible; the most that can be done is to use what aid to judgment can be obtained from the theoreticians in fiscal and welfare economics. Once again, however, the important point in the present context is not the precision with which the decision (between control and incentive) can be made but the fact that if it is to be done at all rationally, the tradeoff must be recognized and in some fashion taken into account.

If the fiscal route is followed, there are further tradeoffs in the application of the incentive. Suppose, for example, the incentive is offered in the form of a tax credit or accelerated depreciation to be allowed for investment in abatement facilities. Abatement costs involve both capital and operating costs, but since such incentives would apply only to capital, they would tend to result in a disproportionate use of capital in abatement efforts. Thus, if a polluter had his choice between the use of low-sulfur fuel or stack gas treatment, he would be more inclined to opt for the latter, which would increase his cash flow, than he would in the absence of the incentive.

The same government study to which I have previously referred also investigated this subject and—noting among other things that fuel substitution (e.g., the use of low-sulfur fuel) was the indicated least-cost alternative in more than 60 percent of air pollution abatement—concluded that "tax writeoffs are not needed nor are they a desirable form for offering further assistance to industry."[10] Tax incentives, in other words, would not stimulate use of the least expensive abatement technique. This judgment was no doubt influenced by the recognition that the use of tax incentives for pollution abatement is politically undesirable because it increases the pressure for similar treatment for other worthwhile causes such as education and housing. At this point we are back to where we started: the tradeoffs we must consider because of limited means, and the demands on those means.

In laying stress on the difficulties of facts and measurements involved in the application of cost–benefit analysis, I am emphatically not suggesting that such analysis is useless and should therefore be avoided or bypassed. If pollution abatement is to be accomplished rationally, it is essential that costs and benefits, however defined, be compared. What I do urge is that, given the numbers, cost–benefit analysis be used in a pragmatic fashion, without being tied to a formula.

The future

Thus far, we have been considering the air pollution problem posed by present technology. Though it would be rash to speculate now on cost–benefit analysis of the resulting pollution problems, it is possible to make something more than guesses about the economics of energy use 30 years from now, as a guide to the directions in which the pollution problem may evolve.

The odds are very much against any truly fundamental surprises in energy generation. Even if such a surprise does come, the odds are even greater against its having any significant effect by the year 2000. Facilities now being built or planned must be written off before they can be supplanted. A new technology could become significant in a short time—say, a decade—only if there were a very large acceleration in the growth of the energy industry. We can thus rule out the likelihood that nuclear fusion will revolutionize the technology in the next 30 years.

Energy sources

Among possible energy sources are two old-timers, wind and sun. These can be dismissed for the same reason that has kept them from being significant for large-scale requirements to date: their intermittency and diffuseness require too great an investment per joule of output.

Tidal energy and geothermal energy can also be dismissed, because of their rare and localized occurrence. Here we must hedge, however. Surface phenomena indicating the existence of underground heat are indeed uncommon, but there are suggestions that buried intrusive masses still in the cooling stage may be more numerous.

Of seemingly greater potential are the oil shales of the mountain states and the tar sands of Alberta. Unless unexpected problems develop, the tar sands should contribute significantly to the North American supply of liquid fuels in coming decades. The extent to which they contribute to U.S. supply depends on U.S. import policy. With shale oil, the problem is economics; it may well be that if and when shale oil does constitute a supply of liquid fuels, it will be because of successful use of underground nuclear explosions to make possible *in situ* production from wells. At present it seems likely that synthetic fuel production will be based on coal, whose presence in the eastern U.S. gives it a geographical advantage. The real key to synthetic fuel from coal is availability of hydrogen at low cost; if this can be achieved, it could postpone the use of shale oil.

Transport of energy

In the field of energy transportation, there appear to be three significant avenues of possible development: the coal-slurry pipeline, liquefied natural or synthetic gas, and electrical conduction at low temperatures.

The coal-slurry pipeline has already demonstrated its commercial feasibility. The first venture was shut down only because its degree of success resulted in a lowering of railroad rates so as to put it out of business.

Liquefied natural gas is already being shipped by tanker from North Africa to Europe on a fairly large scale. The method is being projected for shipment from such places as Alaska, the East Indies, and Venezuela to Japan, Hawaii, and the east coast of the U.S.; it creates, in effect, a worldwide market for gas found anywhere within reach of tidewater. Furthermore, the volume reduction when natural gas is liquefied is a tempting offset to the high investment that would be required to build a pipeline for the overland movement of liquefied gas.

Large-scale energy transport using superconductors, or perhaps merely cooled conductors, is still in the earliest stages of investigation. A powerful stimulus toward the development of cryogenic transmission is concern for

visual pollution of the environment. Once the transmission line is underground, the problem is whether to use electrical insulation and conventional high-voltage transmission or thermal insulation and cryogenic transmission. Cryogenic transmission might not shave investment costs much, if at all, but would greatly reduce power losses, which now run from 2 to 5 percent of total generation.

Energy conversion

It is generally agreed that nuclear power generation cannot continue past the end of the century without exhaustion of the low-cost uranium deposits. Breeding is, therefore, essential. The U.S. Atomic Energy Commission is forecasting for 2010–2020 a nuclear power industry based wholly on plutonium. It is almost certain that with the Carnot cycle, breeder reactors will not much surpass the efficiencies—around 40 percent—now obtained in the best fossil-fueled central stations. Magnetohydrodynamics, with its potential efficiency of 50 to 55 percent, offers a means of hurdling this barrier.

Magnetohydrodynamics (used probably as a "topping cycle" in conjunction with the conventional cycle) is significant as a counter to the forces now tending to halt and reverse the long, uninterrupted downtrend in the cost of electric power. Although some of these forces stem from rising land values, most of them are a result of concern for pollution of the air, the water, or the scene. Within the next few years, costs of electricity—in real terms, aside from inflation—may start to rise. Such a rise could have profound results on the competition between the electric utility industry and fuels.

The gas industry is supporting intensive R & D on a fuel cell that would provide domestic electricity. It takes little imagination to contemplate the possible effects of success; the electric utility industry as we know it would be shaken to its foundations.

Technology of energy in A.D. 2000

Let us now consider in combination the developments that have been discussed separately. They point to two general paths. The first leads to "the electric economy," in which the breeder reactor and magnetohydrodynamics offset the costs stemming from concern for the environment. Electric space heating continues its invasion of the market; public pressure and favorable rates force the development of the electric automobile. The second path leads to "the gas economy," in which trucks and automobiles are fueled with liquefied natural gas and the home fuel cell provides domestic electricity. The gas economy is the same whether it is based on natural or synthetic gas from oil shale or coal, of which the supply is abundant.

The social forces that could bring the gas economy into being are impressive. Natural and synthetic gas are sulfur-free, so power generation by the home fuel cell would go a long way toward solving the sulfur problem; gas-fueled transportation would eliminate hydrocarbon smog. A high degree of success in developing low-cost hydrogen, which is necessary for large-scale production of synthetic gas, might result in hydrogen itself becoming the basic fuel. Then, with the possible exception of nitrogen oxides, there would be no atmospheric pollution at all. There would not even be carbon dioxide, and its elimination would remove the long-run worry about the greenhouse effect in the earth's atmosphere. Also, low-cost hydrogen would increase the prospects for general use of fuel cells, since this simplest fuel poses the fewest problems in fuel cell design.

Although I have proposed two scenarios for the development of energy technology through the remainder of the 20th century, I do not think that either of them is likely to occur as described. What is more likely is some in-between situation in which gas and electricity are dominant together, at the expense of the liquid fuels. The determining forces identifiable today are the cost of electricity and concern for the environment.

This article is based on two talks given by the author, "The Technology of Energy Use in the Year 2000" at the Dallas meeting of the American Association for the Advancement of Science, December 1968, and "Economic Trade-offs in Air Pollution Abatement" at the annual meeting of the American Physical Society, February 1969.

REFERENCES

1. "Report of the Working Committee on the Secondary Impact of Air Pollution Abatement," Coordinating Committee on the Economic Impact of Pollution Abatement, Dec. 15, 1967, p. 7.
2. Perry, H., and De Carlo, J. A., "The search for low-sulfur coal," presented at IEEE-ASME Joint Power Conf., Denver, Col., Sept. 18–21, 1966.
3. *Mech. Eng.*, Aug. 1965.
4. "Cost sharing with industry?" Summary Rept. (revised), Working Committee on Economic Incentives, Nov. 20, 1967, p. 12.
5. Kneese, A. V., *Economics and the Quality of the Environment—Some Empirical Experiences*, Reprint no. 71, Resources for the Future, Inc., Apr. 1968.
6. *Clean Air News*, pp. 2f, Feb. 13, 1968.
7. Gerhardt, P. H., "Air pollution research needs for improved economic analysis," presented at Annual Air Pollution Control Assoc. Conf., Cleveland, Ohio, June 12, 1967.
8. Nourse, H. O., "The effect of air pollution on house values," *Land Economics*, pp. 181–189, May 1967.
9. "Cost sharing with industry?" Summary Rept. (revised), Working Committee on Economic Incentives, Nov. 20, 1967, pp. 19–25.
10. *Ibid.*, pp. 27, 29.

Bruce C. Netschert received the B.A. degree in geology in 1941 and the Ph.D. degree in economics in 1949, both from Cornell University. He joined the University of Minnesota, Duluth, as assistant professor in 1949, and in 1951 became a commodity-industry analyst for the U.S. Bureau of Mines. Later, as a staff member of the President's Materials Policy Commission, he prepared supply forecasts to 1975 for copper, lead, zinc, tin, and rubber. From 1953 to 1954 he was a consultant in the materials area for the National Security Resources Board and its successor agency, the Office of Defense Mobilization. In 1954 he joined the Central Intelligence Agency, where he was branch chief responsible for nonferrous metals and nonmetallic minerals. As senior research associate for Resources for the Future, Inc., from 1955 to 1961 he conducted research on the future supply of fuels and energy sources, including nuclear and solar energy, and the future supply of the major metals. Since 1961 he has served as director of the Washington office of National Economic Research Associates, Inc., where he conducts and supervises research on a wide variety of subjects, with emphasis on fuels and energy resources. He is a member of Phi Beta Kappa, AIME, and the America Economic Association, and a fellow of the Geological Society of America.

Feedback controls on urban air pollution

Methods of monitoring and prediction, health considerations, smokestack regulation—these are some of the steps involved in the evolution of an effective system for controlling the quality of our air

E. S. Savas *Deputy City Administrator, The City of New York*

A conceptual frame of reference for viewing the various functions of an air pollution control agency in a metropolitan area is the conventional feedback-control diagram. A series of such diagrams is presented and discussed, to indicate a possible sequence of developmental stages through which an integrated system may evolve.

In many cities the present mode of operation of the local air pollution control agency is as indicated in Fig. 1. First, let us examine the inner loop on this illustration, which deals with the code standards for combustion equipment. Scheduled and unscheduled inspections and observations of emission sources by field personnel provide feedback and indicate code violations. Enforcement procedures are then applied to secure compliance. At the present time, the entire feedback process is strictly a manual one; this is indicated in Fig. 1 by the dashed line.

Air-quality control limits, shown in the outer loop, serve as the reference values to trigger control action when they have been violated, that is, when the measured air quality has deteriorated to a certain point. A representative set of control limits for New York City is shown in Table I. In many existing systems, air samples are taken and analyzed manually, and the results are transmitted by mail, messenger, or telephone to the decision-makers, who compare the observed air quality with the control limits and decide on what action, if any, to take. (The nonautomatic nature of the sampling, analysis, and feedback is indicated by the dashed line in Fig. 1.) Actually, before deciding on control action, i.e., declaring an alert the decision-makers take into consideration the weather forecast. In other words, control action is not predicated solely on a violation of the control limits, but also on a prognostication as to whether or not the condition will persist.

I. Control limits for air pollution conditions

Alert stage (example)
Stage is reached if for any consecutive six of the previous 12 hours:

1. Sulfur dioxide exposure (ppm-hrs) exceeds 2.0, and
2. Soiling index exposure (reflectance units of dirt shade—hrs) exceeds 25, and
3. Forecast predicts stagnation for at least 12 more hours.

Warning stage (example)
Stage is reached if for any consecutive six of the previous 12 hours:

1. Sulfur dioxide exposure (ppm-hrs) exceeds 3.0, and
2. Soiling index exposure (rud-hrs) exceeds 25, and
3. Forecast predicts stagnation for at least 12 more hours.

Emergency stage (example)
Stage is reached if in a 24-hour period:

1. Sulfur dioxide exposure (ppm-hrs) exceeds 15.0 and is rising, and
2. Soiling index exposure (rud-hrs) exceeds 200, and
3. Forecast predicts stagnation for at least 12 more hours.

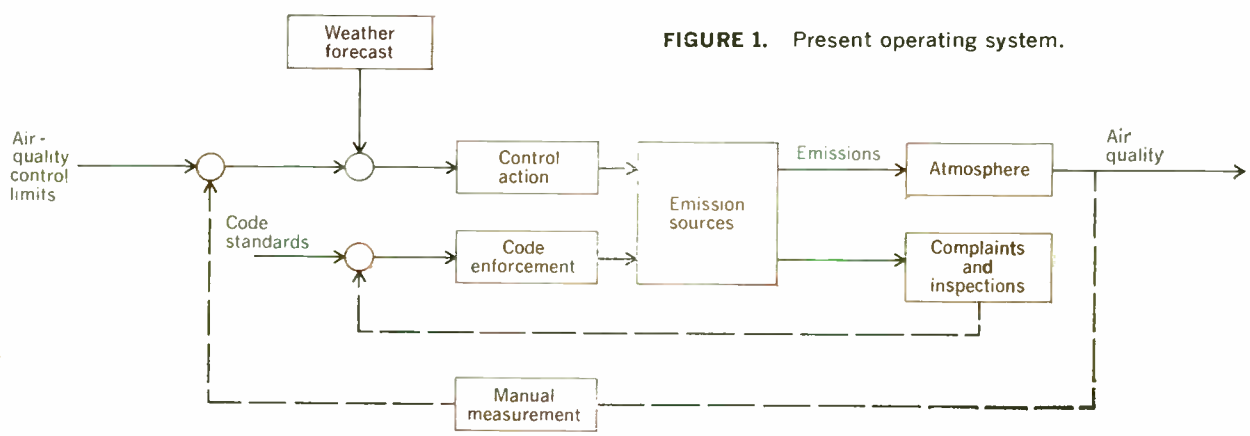


FIGURE 1. Present operating system.

Certain emergency control actions are usually available to the responsible public officials. These include such measures as prohibiting nonessential automobile travel, banning incineration, and requesting a switchover to different fuels for power generation. It should be noted that at the present time these are very coarse control actions, which impose uniform and heavy burdens throughout the city in terms of total cost and inconvenience; furthermore, it is difficult to determine the degree of compliance with these instructions, or their overall effectiveness.

The remaining blocks in Fig. 1 complete the picture by indicating that the emission sources release pollutants into the atmosphere and thereby determine the observed air quality.

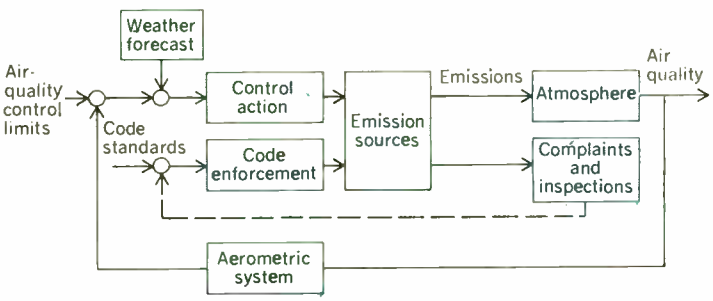


FIGURE 2. Mode of operation after aerometric system is functioning.

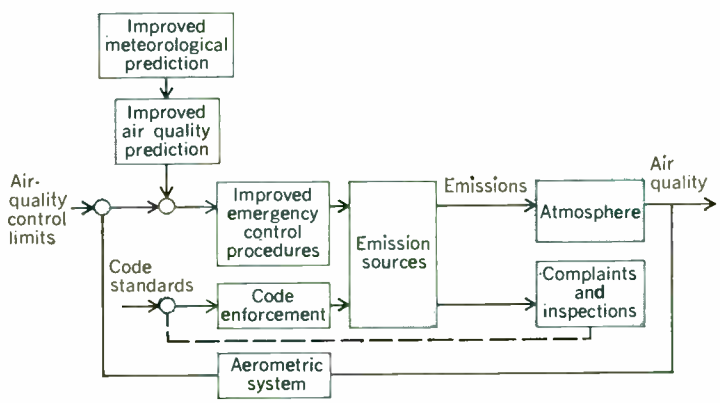
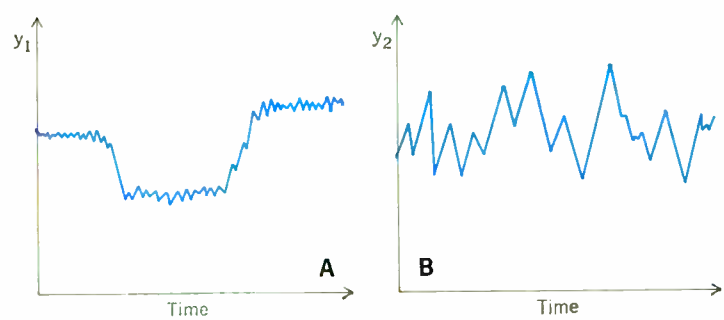


FIGURE 3. An effective early-warning system for air pollution control.

FIGURE 4. Two different pollution patterns that exhibit the same average annual pollution concentrations.



Air-quality monitoring system

The next step in the evolutionary sequence, the installation of an on-line aerometric system (an air-control monitoring system), has already been taken in several cities,[1] including New York. Information from remote, unattended sensor stations is sampled automatically by a computer to permit rapid assessment of the current air quality throughout the area. The on-line nature of this system is indicated by the solid line in the feedback loop of Fig. 2. This knowledge is useful in itself, for if serious pollution conditions develop they can be detected rapidly and control action can be taken to reduce emissions. Furthermore, the data that are acquired in machine-readable form can be analyzed to find the relation between pollution levels and particular emission sources. The sensor data usually include not only air-quality measurements, but also meteorological parameters such as wind direction and velocity.

Site selection. The sensor stations in an air-quality monitoring network serve four primary purposes: (1) to monitor air quality, to assure that dangerous conditions (whether predicted or not) do not go undetected; (2) to supply a record of air-quality data for future analysis of medical data; (3) to supply a record of air-quality data for testing of air-quality prediction models; (4) to provide input data for any real-time air-quality prediction technique that may be in use.

The location requirements for these four purposes differ considerably. For purpose 1 it is essential to locate the sensors in those spots known to suffer from relatively high pollution levels. Such spots could be pinpointed by means of mobile sensors and/or predictions based on a detailed diffusion model. For purpose 2 the sensors are best located in high-density residential areas, to indicate the air quality to which large population masses are exposed. For purpose 3 it would seem that locations having pollution levels close to the average of the area in question are desirable. Purpose 4 can be discussed only if one knows how the data are going to be used in the predictions as the problem is related to that of experimental design: What sensor location gives the maximum information for the purpose of the particular predictor in use?

To illustrate how requirements 2 and 4 may conflict, a sensor station located six floors above the ground in an area of high-rise apartments may be near the mean altitude of the population, but for an air-quality prediction model that requires some wind data it might be convenient to install the sensor station on the 30th floor of the building.

Similarly, by ignoring purposes 1 and 2 one could locate the sensors in such a way as to give a favorable but misleading picture of the city's air quality—misleading in the sense that neither the most exposed nor the most populous areas would be represented.

An effective early-warning system

An on-line air-quality monitoring system is a valuable and necessary first step, as it offers the prospect for considerably more effective measures to be taken. In particular, the prediction of future air quality (say, 6 to 48 hours in advance) would permit proper anticipatory action to reduce emissions before an "alert" condition is reached; this is manifestly superior to reacting merely after the fact, when air quality is already observed to be very poor. Figure 3 shows this next stage of development,

which incorporates improvements in meteorological prediction and air-quality prediction, and therefore makes possible improved emergency control measures.

Improved air-quality prediction may require use of an appropriate diffusion model. However, such models require predictions of meteorological variables, and these are unreliable at present in the crucial range of low wind velocities. Under these circumstances there is little incentive to use a sophisticated diffusion model for real-time prediction. It is doubtful whether the most detailed model would give any more accurate short-term predictions than would a crude model.

If such prediction is deemed unsatisfactory the initial improvement must come from better meteorological prediction. This may require an extended urban meso-meteorological monitoring network, as well as a wide-area weather watch. At the very least, better local weather prediction would lead to better prediction of major pollution episodes.

Scale of prediction. One question that comes to mind for improved air-quality prediction is: What is the required scale of prediction, both spatial and temporal: At one extreme, must one be able to predict the moment-by-moment concentration at any point in the city? At the other extreme, is it sufficient merely to be able to predict the gross 24-hour average for an entire metropolitan region? The answer must lie somewhere in between, but it is not clear precisely where.

The requisite time scale must depend on medical data, which should determine the minimum time of exposure that is likely to be harmful. Figure 4 illustrates two hypothetical situations in which the average annual pollution concentrations are equal. If both situations are hazardous to health, then it becomes important to predict the short-term peak heights of Fig. 4(B), clearly a more difficult task than to predict the extended pollution episodes of Fig. 4(A).

The requisite space scale must depend on the variability of pollution levels from place to place in the city. It is well known that sulfur dioxide concentrations can vary widely throughout a city at the same moment. However, it is commonly assumed that certain "pollution pockets" in the city have a rather consistently high pollution level; that is, although the absolute levels throughout the city will rise and fall with the weather, certain areas will be consistently worse relative to others.

In order to investigate this hypothesis, hourly data for an entire year from eight sulfur dioxide monitoring stations in New York City were examined. Twelve high-pollution days were selected. The readings at each of the stations were compared and are summarized in Fig. 5. The height of the bar indicates the number of days (of the 12) when the indicated station had the highest [Fig. 5(A)] and the lowest [Fig. 5(B)] readings in the city. Surprisingly enough, the same location, station H, had the lowest reading in the city on four of the 12 days and the highest on three of the 12. The rather even distribution of the high (or low) readings among the eight stations certainly suggests that there is considerable variability in the pollution pattern throughout the city—i.e., that the spatial pattern itself varies with time. Similar results appear in Fig. 6 for a selected sample of 31 days in 1966. Note that station G had the highest readings on five of the days, and the lowest on five other days.

The situation is quite different for carbon monoxide, whose concentration follows the same persistent spatial pattern as does vehicular traffic.

At the very least this brief investigation fails to support the hypothesis that if one knew that a fairly bad day were coming up, one could readily predict where in the city the high pollution levels would be found; it seems that small-scale air-quality prediction is necessary. This

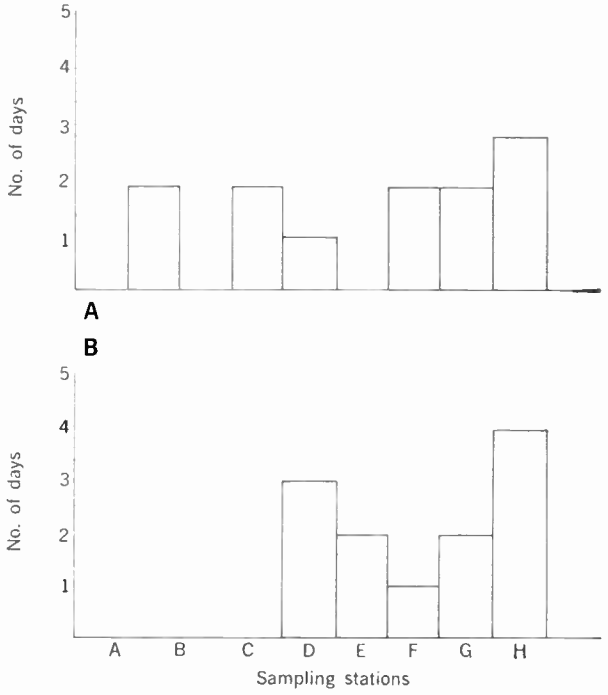
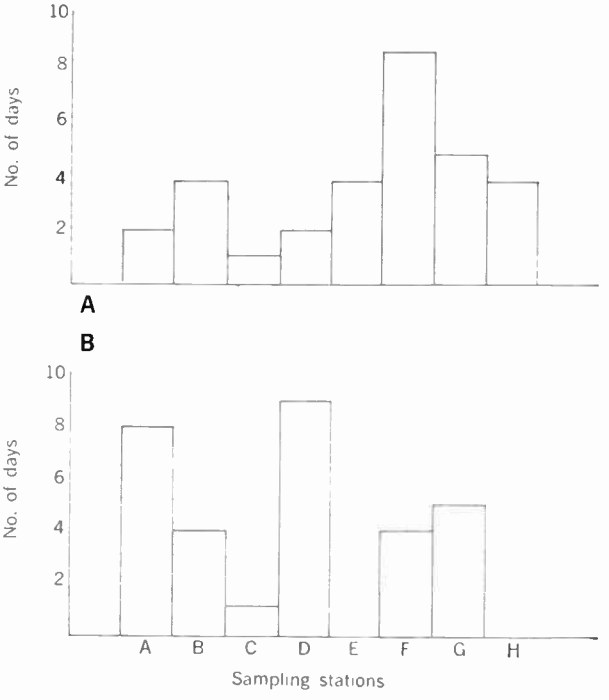


FIGURE 5. Distribution of highest (A) and lowest (B) sulfur dioxide hourly readings among stations for 12 days.

FIGURE 6. Distribution of highest (A) and lowest (B) sulfur dioxide daily average concentrations among stations for 31 selected days.



tentative finding, although dismaying on the surface, raises some interesting possibilities. On moderately bad days, it would appear that some parts of the city may be subject to hazardous pollution levels (if no control action is taken) whereas other areas may be comparatively safe. At present, emergency control measures are applied rather generally and indiscriminately throughout the city. To be able, however roughly, to predict localized air quality would afford the opportunity to apply control measures in a similarly localized manner, and thereby obviate the need for imposing economically drastic city-wide emergency measures. Furthermore, this capability would endow the control agency with the ability to act in marginal cases, when the situation does not warrant broad emergency measures but is still serious enough that limited action could properly be taken to alleviate a potential local hazard.

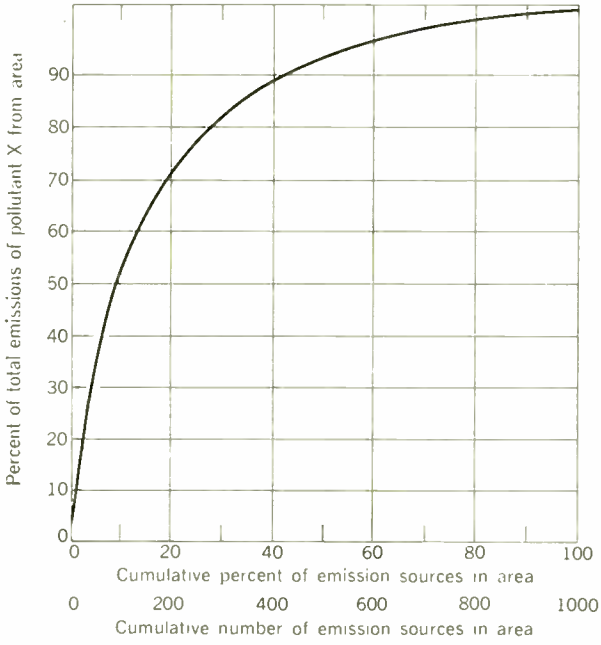
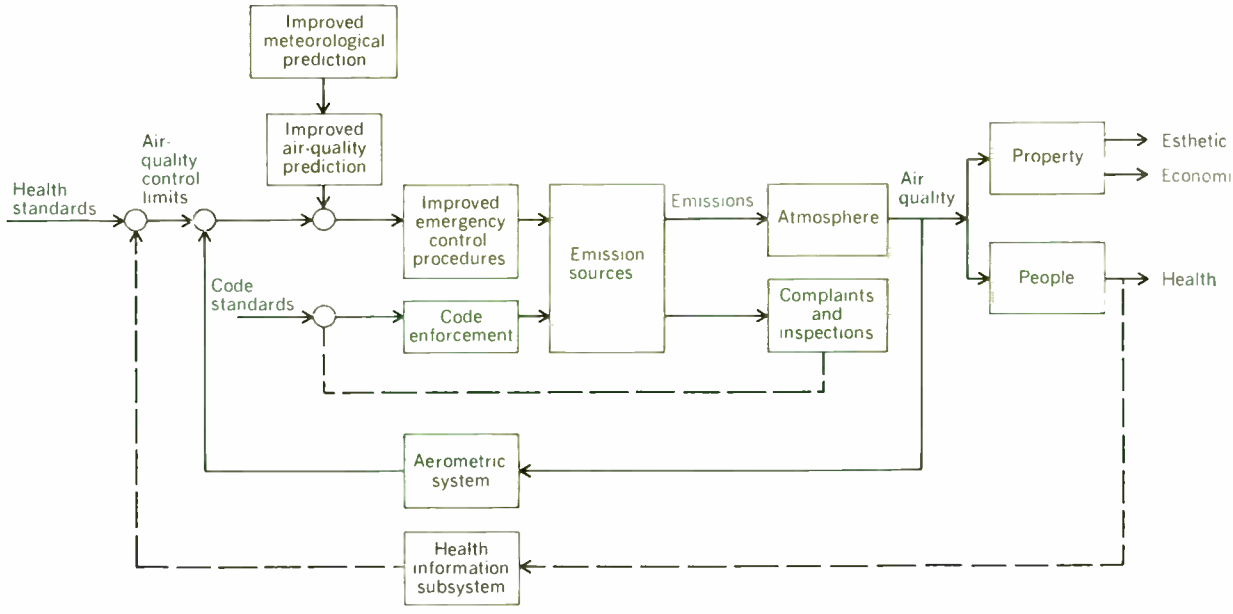


FIGURE 7. Cumulative emission characteristic of the emission sources in an area.

FIGURE 8. Taking health aspects into consideration in an air pollution control system.



Emergency control actions. Given adequate meteorological and air-quality prediction methods that suffice to indicate the expected levels of pollutant concentrations in different areas of the city for a day or two in advance, the local control agency can issue directives for localized emergency control measures during the earliest stages of anticipated major and minor pollution episodes. When that capability is forthcoming, the agency can plan such localized control measures through a study involving:

1. Division of the city into suitable areas within which the control measures would be applied.
2. The listing, in rank order of pollution potential, of the principal emission sources in each of the areas for each type of pollutant; this is done from an emission inventory.
3. Estimation of the response time for each major emission class. The response time is the total time required to communicate a directive from the air pollution control center to the controlling party of the emission source, and for the latter to take the requested action and actually reduce the emission rate to the desired value.
4. Formulation of realistic emergency control measures, estimation of the degree of compliance with each measure, and use of a simulation system and economic evaluation procedure[2] to determine the most suitable and effective emergency measures. These could involve scheduling of power plants, selection of a minimum number of key emission sources (in each area) to be temporarily shut down, prohibition of traffic on certain arteries, etc.

Figure 7 illustrates, for hypothetical data, the fact that in some areas a very small fraction of the total emission sources are responsible for a major fraction of the emissions from that area. By preparing such a characteristic chart for each area of the city, and listing the specific sources, the local control agency is in a position to improve its emergency planning for each stage of an air pollution alert. It must be remembered, of course, that the chart for an area will differ according to seasonal and other time-dependent factors.[2]

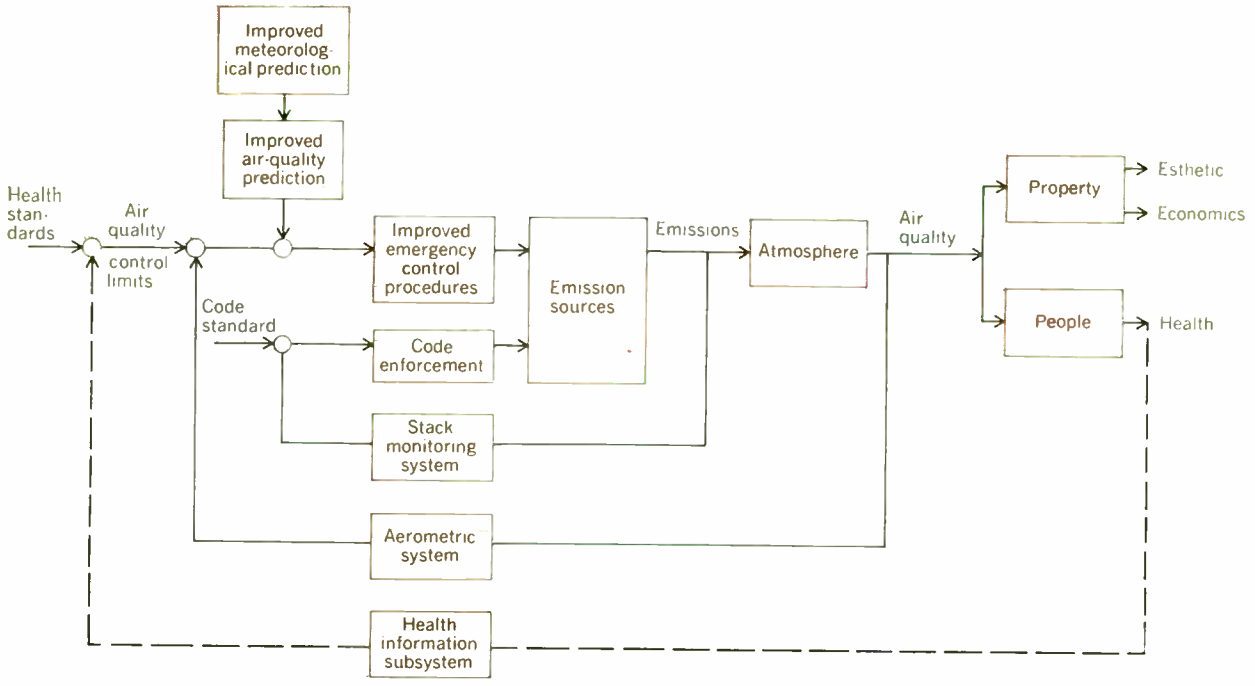


FIGURE 9. Stack monitoring in a total health-based air-pollution-control system.

Introducing health considerations

Present air-quality control limits are somewhat arbitrary, and are not rigorously founded on the medical effects of air pollution. As pointed out in the foregoing, it is not yet known what threshold effects there are, if any, from duration of exposure, dosage levels, etc. The next step in the progression, therefore, is to embed the air-quality control system within a larger and more meaningful system that explicitly considers the relationship between air quality and community health.

The block diagram of Fig. 8 shows the system of Fig. 3 nested within a medical feedback loop. It illustrates the fact that the air, and its pollutants, acts upon both people and property, and that it has an impact on the physiological, economic, and esthetic condition of the community. In such an ultimate system, health standards are established based on analysis of medical data, and these in turn are used to set the air-quality control limits. The latter may be dynamic; that is, it may be necessary to change them from time to time as feedback of medical information so dictates. As the dashed line of Fig. 8 indicates, this feedback need not be automatic, although conceivably it could be if it is a convenient by-product of a city-wide, computer-based health information system that is developed and implemented for a broader purpose.

Stack monitoring

A subsequent step that one can realistically envision for a municipal department of air pollution control in this progression of increasing effectiveness and increased mastery of the problem is the use of automatic sampling, analytical, and recording equipment to monitor the smokestacks of major pollution producers (see Fig. 9). This would tend to minimize inefficient "smoke-chasing" activities by field personnel and would increase effective control over these emission sources. Such monitoring might be accomplished by installing on-site devices on the stacks themselves, and either telemetering the data to a central point for recording and analysis, or recording it locally in sealed recording units that can be examined periodically by authorized inspectors. Another approach is to use remote, telescopic, infrared analyzers that can be focused, from a stationary platform or from a patrolling helicopter, on suspect smokestacks.[3] In either approach, computers would be used to analyze raw data and report on the findings. Conventional data processing would be applied to the licensing, inspection, and complaint-processing activities.

REFERENCES

1. Lynn, D. A., and McMullen, T. B., "Air pollution in six major U.S. cities as measured by the continuous air monitoring program," *J. Air Pollution Control Assoc.*, vol. 16, pp. 186–190, 1966.

2. Savas, E. S., "Computers in urban air pollution control systems," *J. Socio-Econ. Plan. Sci.*, vol. 1, pp. 157–187, 1967.

3. Barringer, A. R., "Molecular correlation spectrometer for remote sensing of gaseous pollutants," presented at 60th Annual Meeting, Air Pollution Control Association, Cleveland, Ohio, June 1967.

E. S. Savas, who was appointed Deputy City Administrator for the City of New York by Mayor John Lindsay in 1967, directs the Management Science Unit, which encompasses systems analysis, operations research, and computer systems planning. The recipient of B.A. and B.S. degrees from the University of Chicago in 1951 and 1953, respectively, and of M.A. and Ph.D. degrees in physics from Columbia University in 1956 and 1959, respectively, he joined the staff of IBM Corporation in 1959 as control systems consultant. In 1966 he became manager of urban systems for IBM, heading a newly created group to explore the application of management science methods and advanced computer technology to urban problems. Dr. Savas has lectured before the Japan Management Association as well as at various universities in the U.S. and is the author of a book, "Computer Control of Industrial Processes," and of numerous articles.

How to prototype hybrid-circuit patterns and screens at budget prices

An application of existing screening and photographic techniques enables design engineers to develop complex circuit patterns easily and quickly, and at low cost

Leon Jacobson *General Electric Company*

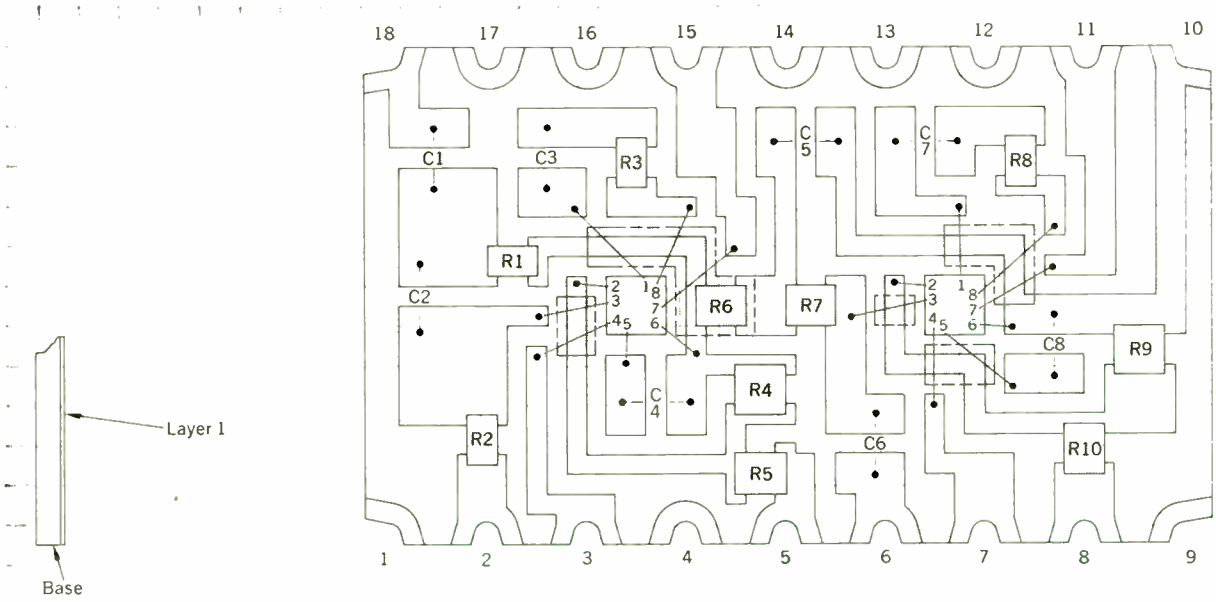
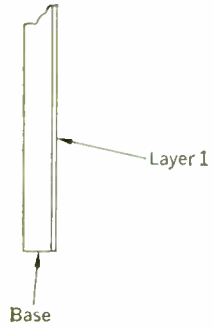


FIGURE 1. Ten times normal layout drawn on grid paper.

Printed circuit taping methods are used to make thick-film master patterns. Usually, the patterns are laid out at ten times normal size and photographically reduced to actual circuit size. The screen patterns are then made using commercially available indirect emulsion material. A home movie light and a photographic print frame process the emulsion instead of a carbon arc and a vacuum frame. These simple techniques result in quick turnaround time and minimum cost to produce development circuits.

Often, when building prototypes, the normal techniques of developing screen patterns are too costly, and production time too long. In addition, engineers need prototypes that have the capacity for quick and easy circuit and component changes. Here is a simple method by which master artwork and screen patterns can be made with a minimum of equipment in only a few minutes. The method is not revolutionary, but is an application of existing techniques used to produce thick films.

The popularity of the thick-film process as used in



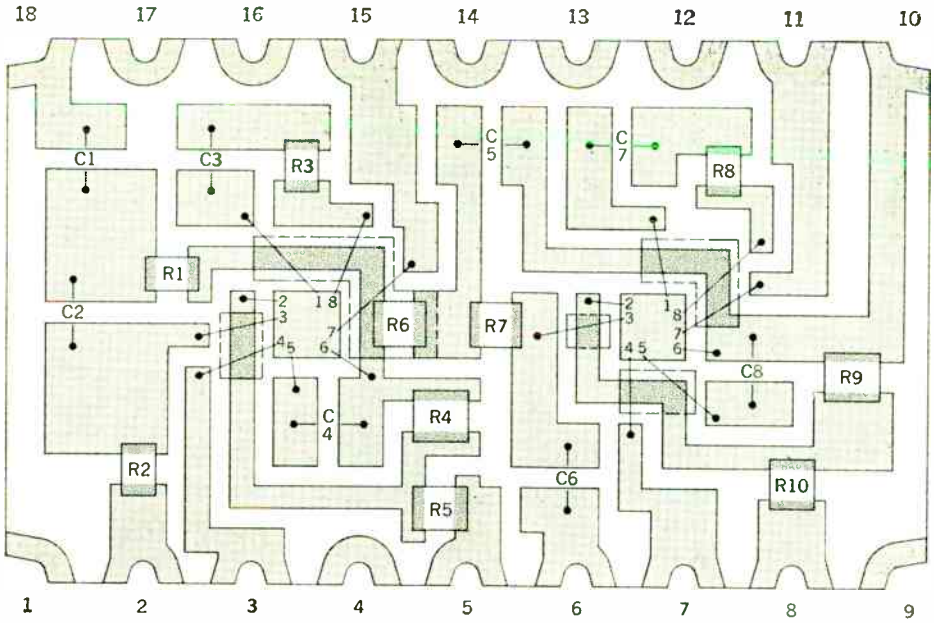
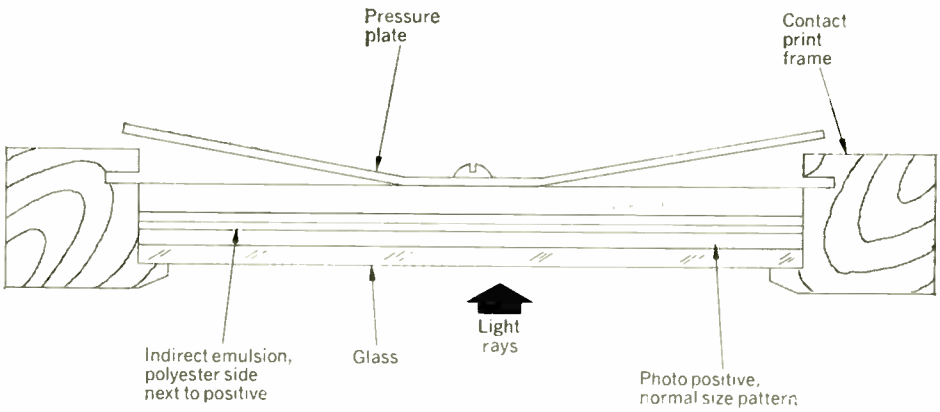


FIGURE 2. Taped conductor pattern.

FIGURE 3. Contact print frame retains polyester sandwich.



making passive components on ceramic substrates has been, in part, due to the ease with which the component patterns can be produced. And since silk-screen printing is the basis for making thick films, costs are usually lower than those of other microminiaturization processes.

Normally, however, artwork involves coordinatograph scribing and photographic reduction to make a film positive, which in turn is used to produce a screen pattern. Usually screens are fabricated by specialty houses. The technique described here cuts screen-making complexity and permits in-house production of intricate hybrid circuits by small engineering laboratories.

Master drawings

Once the schematic has been completed and the values of the various components calculated, a thick-film layout ten times normal size (Fig. 1) is made on conventional grid paper. In making the layout, the designer may find it useful to use color codes to distinguish resistors, crossovers and capacitor dielectrics one from the other. (It is also suggested that the grid lines be used, wherever possible, as the boundary lines that determine the component or conductor.)

Most engineers are familiar with the black adhesive-backed tapes used to produce printed circuit patterns. These tapes are commercially available in widths of 0.254 to 25.4 mm, and cut to an accuracy of ±0.05 mm. When they are used to make a master drawing, the reduction accuracy is quite good. For example, a 0.5-mm error in the taped master produces only a 0.05-mm error in the final reduction.

In the thick-film process, the need for a master is determined by each printed pattern. For each screening, therefore, a master pattern exists. Patterns are needed for conductor(s), crossovers, capacitor dielectric(s), resistor(s), and resistor protective glass encapsulant. By fastening the grid paper layout onto a light box with a sheet of transparent plastic over it, we have a surface on which we can make our first pattern, usually the conductor pattern (Fig. 2). At this time a registration pattern, such as corners or targets, is also applied. (A felt-tip pen is useful for marking an identifying code outside the

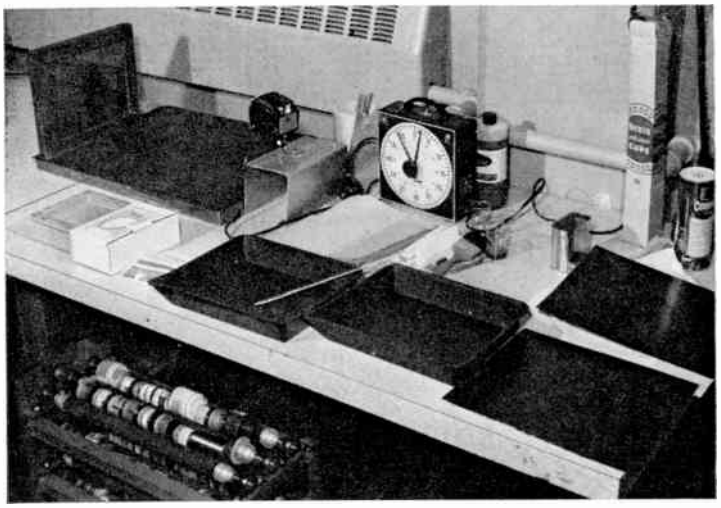


FIGURE 4. Setup for exposing indirect emulsion material.

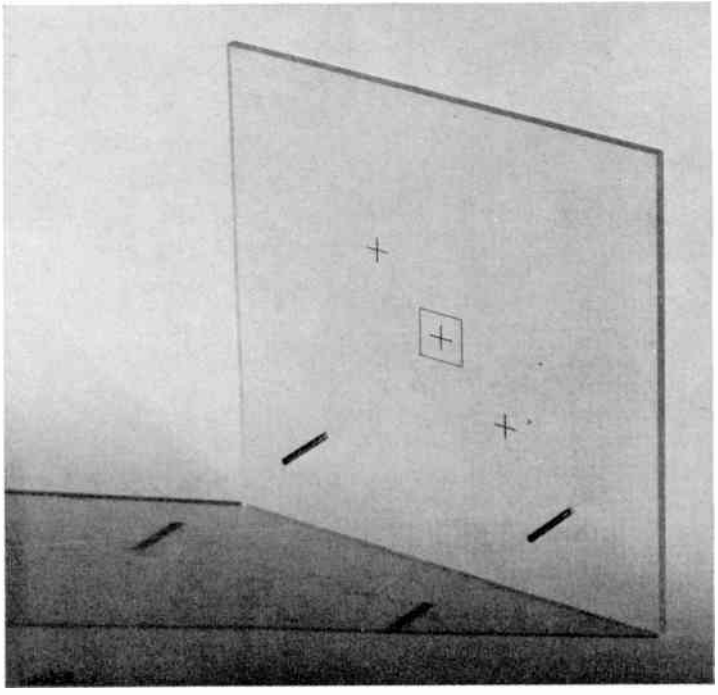
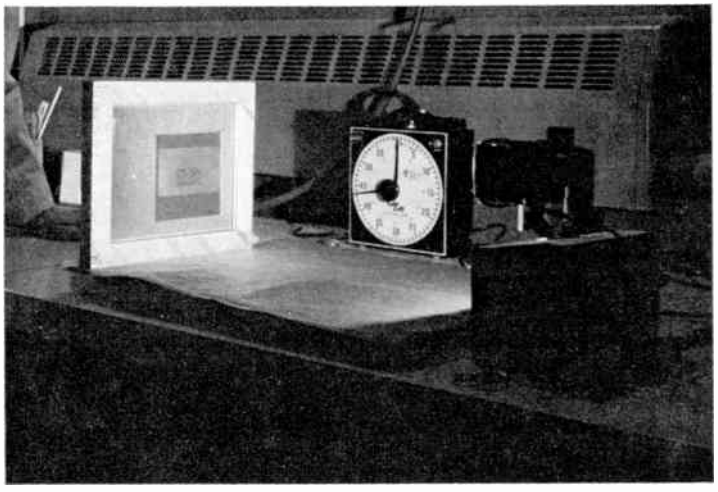


FIGURE 5. Lay-up board showing alignment target and pins.

FIGURE 6. Use of halogen lamp to expose hybrid pattern.



pattern area.) On top of the first taped master, another transparent sheet is added, and another taped delineation, such as the resistor pattern, made. This process is continued until all the needed master drawings are completed.[1]

The masters are photographically reduced and a normal-sized positive is made for each. Normally, an experienced technician can tape a master drawing in about one-half hour. Should changes be required, they can be made in a few minutes and the master rephotographed. If a Polaroid copy camera is used, the total process time can be reduced even more. Since the taping accuracy is usually considered about 0.500 or 0.760 mm, the accuracy of the finished photo (ten times reduced) becomes 0.050 to 0.076 mm.

Screens

The commercially available "silk" screens used for noncritical work, such as signs and posters, are generally made in one of two ways. The method most used by screen specialty houses utilizes a photosensitive direct emulsion painted on a raw (bare mesh) screen and allowed to dry. The pattern is placed in contact with the screen; then, using a high-intensity carbon arc, photographically exposed. Next, the screen is developed, and

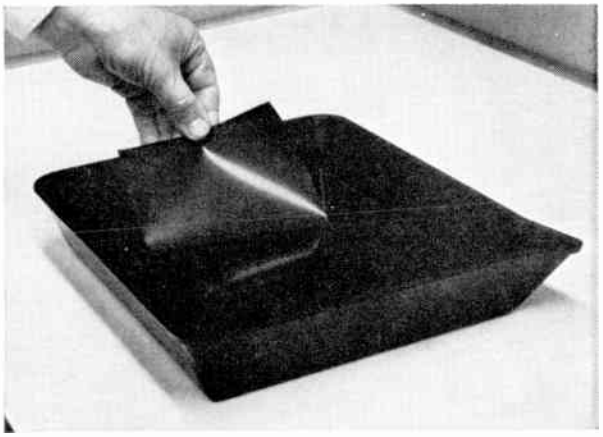
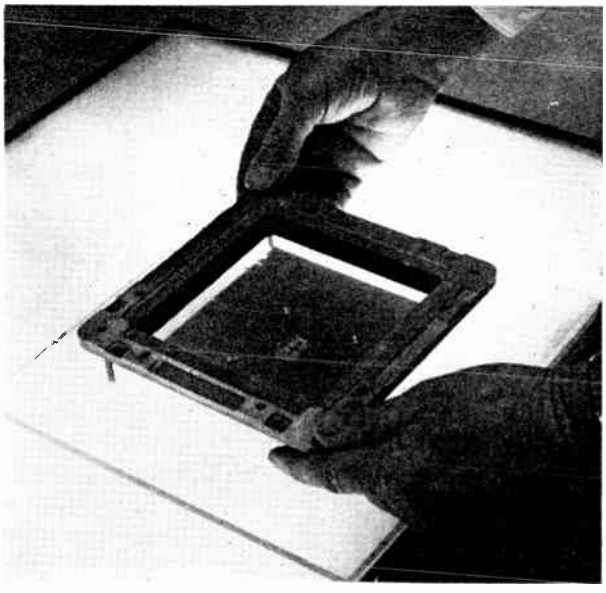
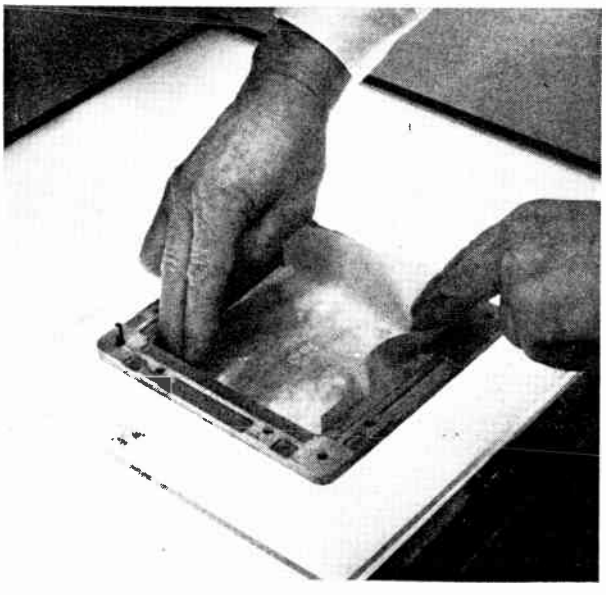


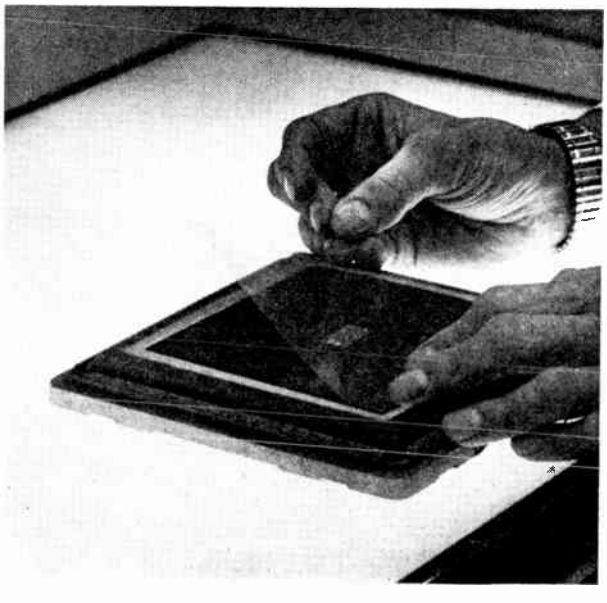
FIGURE 7. Material is developed under yellow safelight.

FIGURE 8. After being developed, material is washed in flowing warm water to clear printing areas of emulsion.

FIGURE 9. A (top)—Pattern is located on target and screen applied to emulsion. B (center)—Blotting paper is used to remove excess moisture. C (bottom)—After the emulsion has dried, the polyester backing is peeled away.







the open area through which the ink is to be squeezed "washed out" to remove any unexposed emulsion material.

The second method uses an indirect emulsion of light-sensitive gelatinous film, backed by a polyester sheet. Photographic exposure, development, and washing out is similar to the direct emulsion method, but work is performed on the gelatinous/polyester sandwich. When the developing process is complete, the wet film is transferred to a clean raw screen to which the gelatin adheres. Upon drying, the polyester backing is carefully stripped off and the screen is ready for use. Screens made in this way will print 100 to 200 patterns and can be cleaned in an organic solvent, such as triclorethylene, without damage. The raw materials for both processes are available from most commercial screen supply houses. Because of its simplicity,[2] this article will discuss the indirect emulsion screen process.

As was previously started, a carbon-arc vacuum frame printer is normally used for exposing the emulsion materials. But these printers are large and expensive. The exposure system described here uses components that are small, low in cost, and readily available at any photographic equipment supply store. It consists simply of a tungsten halogen lamp (bulb code DWY)[3] from a home movie set and an old-fashioned contact print frame (Figs. 3 and 4). At about 50 cm, the lamp exposes the indirect emulsion material in about four to five minutes. And the frame assures close contact of the film positive with the screen emulsion material. Total cost of these items is about $15.00.

The only other equipment necessary is a photographic print developing tray and a pinaligned lay-up board (Fig. 5). The board is made of a plastic plate with an alignment target scribed on it; locating pins are pressed in it to match the mounting holes of the raw screen frame.

After development and washout, the completed wet indirect emulsion is placed, gelatin side up, on the lay-up board and its pattern aligned with the plate's target. Next, the raw screen is placed on the alignment pins and dropped onto the gelatin until it adheres. During development, a yellow safelight should be used to prevent damage to the emulsion.

Most vendors who make completed screens for use in commercial thick film manufacturing will also supply "raw" screens. That is, they will sell the metal frame with the mounted screen mesh material stretched to the proper tension, but without an emulsion or pattern applied. These can be used as the basis of the process being described.

Indirect emulsion material can also be purchased from commercial screen supply houses. It is best to have it cut about 7.50 mm larger than the mesh area of the screen frame; this does away with the necessity of using a fill-in material on areas not covered by the emulsion and not part of the pattern. Measured packets of the developer chemical used for processing the screen material and bottles of water-soluble fill-in are available from the supply stores. The technician should have on hand a can of a household abrasive cleaner to preclean the raw screens. A bleaching agent, such as "Clorox," to remove the indirect emulsion from the screen when it is no longer needed should also be available. This, of course, allows the screen to be used again and again until its tension no longer meets minimum requirement.

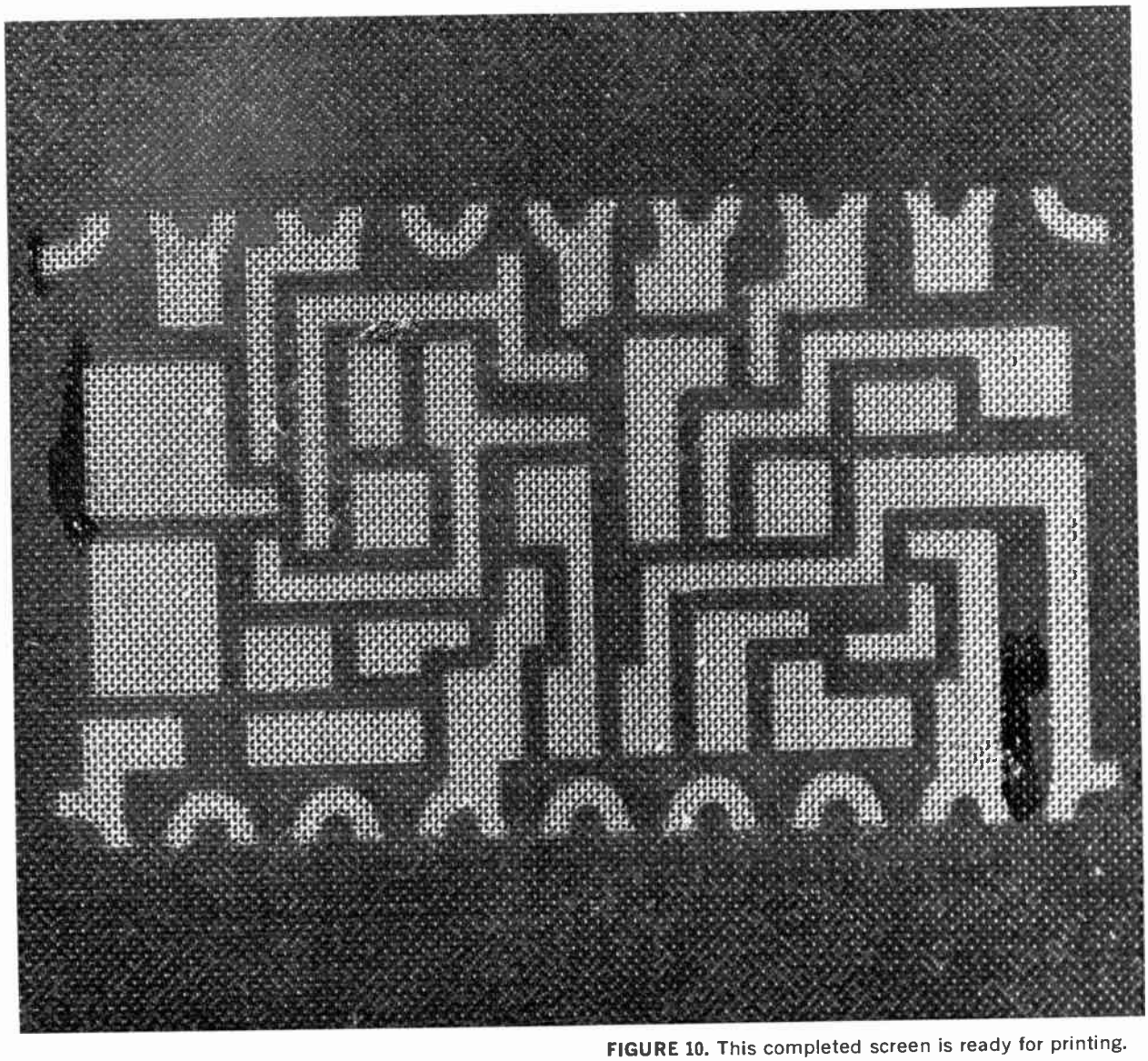


FIGURE 10. This completed screen is ready for printing.

Process

Once all the equipment and various materials have been gathered and checked, the technician can go ahead and make the screens. The process consists of six simple steps:

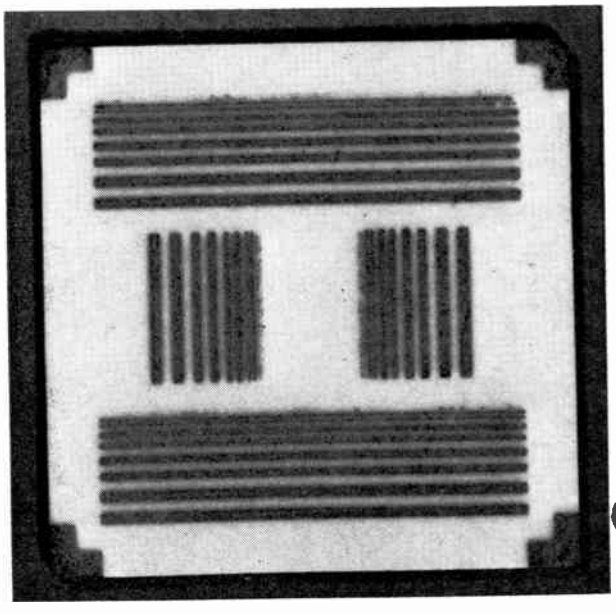
1. Check to see that the photographic positive is against the print-frame glass and the plastic backing of the emulsion is against the positive (Fig. 6); then expose the indirect emulsion. (Handle unexposed material in a yellow safelight environment.)

2. Develop the material, in accordance with the manufacturer's instructions as stated on the developer solution, under yellow safelight (Fig. 7).

3. Carefully hold material, gelatin side up, under gently flowing hot water (use temperature recommended by manufacturer) until all printing areas are clear (Fig. 8).

4. Place gelatin side up on lay-up board; locate pattern on target and gently apply screen to emulsion with enough pressure to adhere properly [Fig. 9(A)]. Using clean newsprint or paper toweling blot any excess moisture [Fig. 9(B)]. The indirect emulsion should now be affixed to the side of the screen that comes into contact with the substrate to be printed.

FIGURE 11. A typical sample of a printed substrate.



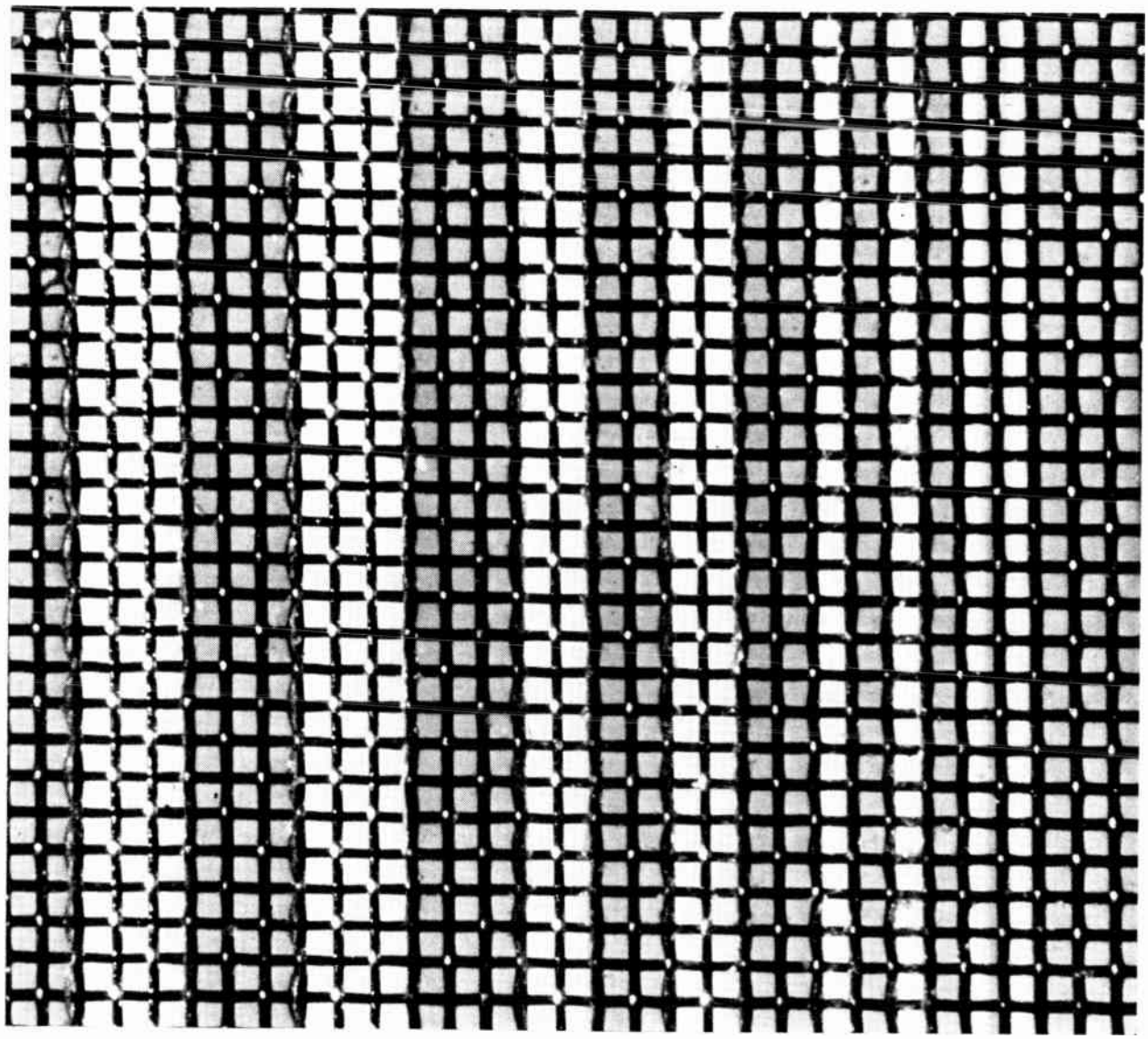


FIGURE 12. A typical screen pattern microphotograph.

5. Allow the emulsion to dry at room temperature in a clean, lint-free atmosphere.

6. Peel the polyester backing from the emulsion [Fig. 9(C)] and the screen is ready for use (Fig. 10).

Although this process has been developed for non-critical needs, it is unusually stable and good enough to produce prototype circuits. A test pattern was accurately made using pairs of lines at spacings of about 0.500, 0.380, 0.130, and 0.076 mm. This arrangement seemed to supply a worst-case condition because of the parallel lines and spacings. The patterns were printed using a platinum gold paste pressed through a 200 mesh (per inch) screen. A sample of a printed substrate is shown in Fig. 11. Figure 12 shows a microphotograph of the screen pattern. From left to right are a pair of 0.380-mm lines, 0.250-mm lines, 0.127-mm lines, and finally 0.076-mm lines. Note that the 0.076-mm lines have started to be irregular and filled in some areas and are not able to reproduce continuous line patterns. With special care and a high-resolution paste, continuous 0.127-mm lines were printed; however, these lines should be considered the limit of this process. Lines this narrow should not be used for prototype work. Table I gives the limits of the readings taken on ten printed samples of test patterns. The conclusion is that the 0.250-mm line is the smallest line width considered practical for prototype work.

Note that the photographic process of making screens produces line widths larger than the original artwork line widths. This is due to the dispersion of the light rays from the photo lamp. Increased line widths on the substrates are caused by flow characteristics of the paste. Again, Table I proves that artwork, precise enough to meet the needs of a prototype circuit, is produced when the indirect emulsion process and tapes are used to make screens. A completed circuit is shown in Fig. 13.

Conclusions

The use of printed-circuit taped artwork and commercially available indirect emulsion screen fabrication materials allows design engineers to obtain prototype samples quickly. With circuits of about 30 printed components, a layout can be made, the taped artworks produced, and the required screens made (a quantity of five) in about

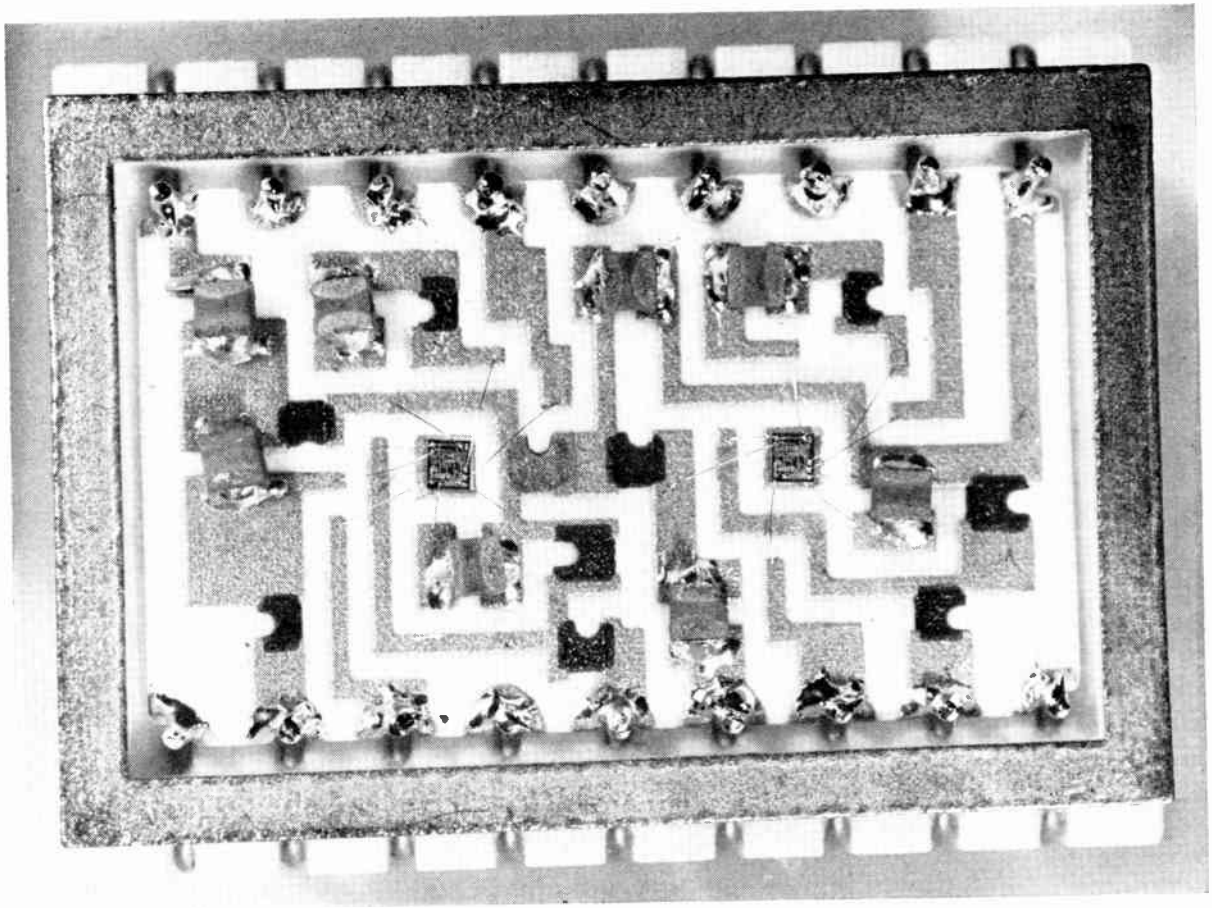


FIGURE 13. Completed circuit ready for use.

six man-days. Changes need minimum turnaround time. Simple changes can be made right on the film positive, the old emulsion washed off the screen mesh with Clorox, and the raw screen used again. More extensive changes can be made by moving or recutting the tape on the master and then rephotographing. Although these techniques are not designed for production units that require 0.100-mm line widths, they can assist the engineer in breadboarding his circuits. In many cases, the breadboard closely resembles the final product.

The assistance of L. Connelly and J. Cleary in the technical preparation of this material is greatly appreciated.

This article is based on the paper, "A Low Cost Master and Screen Making Technique," presented at the Electronic Components Conference (G-PMP, EIA), Washington, D.C., April 30–May 2, 1969.

I. Process limits for three line widths, mm

Theoretical Line Width	Film	Screen	Substrate
0.508	+0.0178 −0.0330	+0.0333 −0.0203	+0.1067 −0.0076
0.381	+0.0254 −0.0609	+0.0254 −0.0051	+0.0914 −0.0178
0.254	+0.0178 −0.0457	+0.0127 −0.0305	+0.1092 −0.0178

REFERENCES

1. Linden, A. E., *Printed Circuits in Space Technology*, Englewood Cliffs, N.J.: Prentice-Hall, 1962, pp. 18–36.
2. Kasloff, A., *Photographic Screen Process Printing*, Cincinnati, Ohio: Signs of the Times Publishing Co., 1962.
3. "General Electric photographic lamp and equipment guide," Photo Lamp Dept., General Electric Co., Cleveland, Ohio, p. 10.

Leon Jacobson received the bachelor of science degree in electrical engineering from Princeton University in 1947. After graduation he was employed by Gibbs and Hill in New York City on power plant, railroad electrification, and power-line design work. In 1949, he joined the Chase Aircraft Corporation in Trenton as supervisor of their electronics testing laboratory. He joined the General Electric Company, Syracuse, N.Y., in 1951 as a product design engineer, and was involved in printed circuit development. In 1957, Mr. Jacobson became a project engineer in advanced packaging techniques for radio-controlled missile guidance equipment. He was appointed consulting engineer in the area of packaging standards in 1961 and later manager of a reliability organization. He then became consultant in microelectronic packaging. His present assignment is technical manager of advanced components and standards.

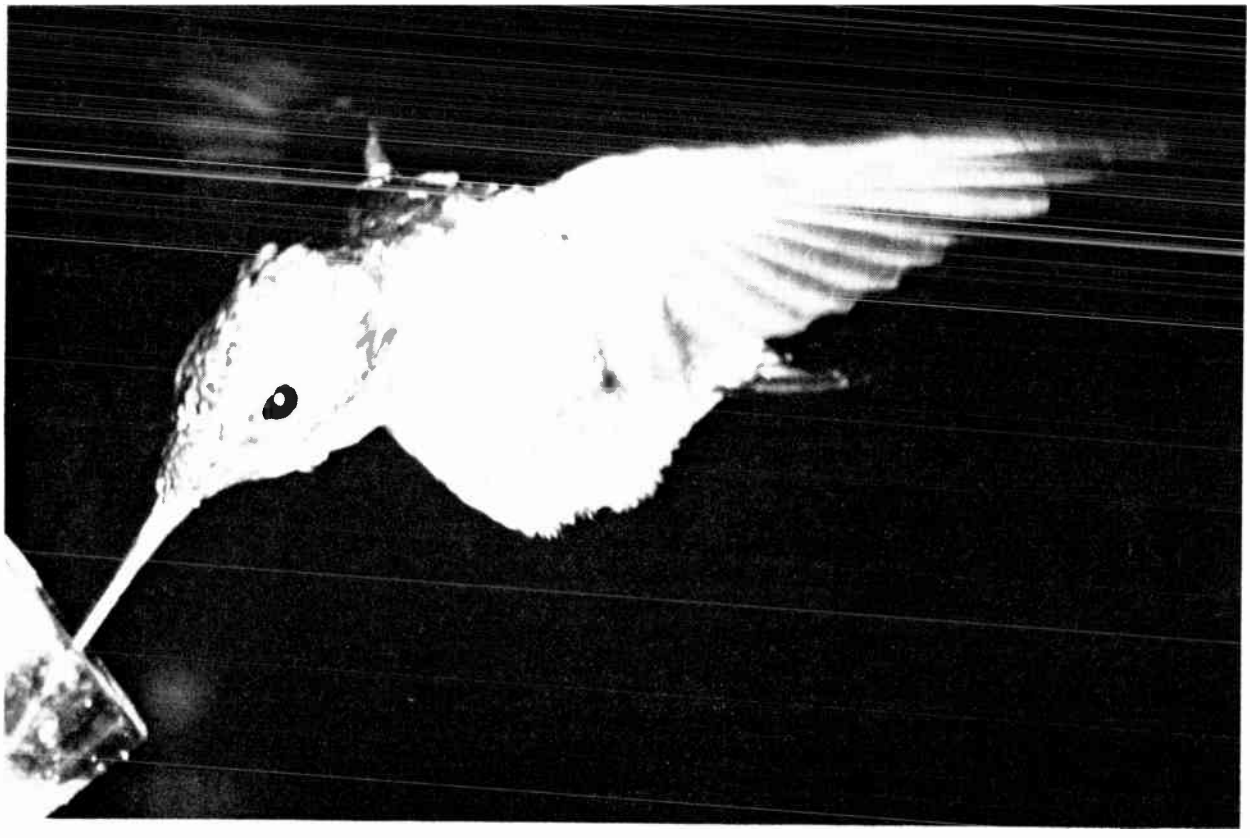


FIGURE 1. A feeding hummingbird; taken with an exposure time of 0.5 ms, this photograph demonstrates an example of undesirable wing blur.

Improvements in electronics for nature photography

To obtain exposure times of less than 100 μs and still maintain correct color balance, a group at M.I.T. has developed strobe units with fused-quartz flash lamps, special electrolytic capacitors, and regulated power supplies

Harold E. Edgerton, Vernon E. MacRoberts *Massachusetts Institute of Technology*
Manmohan Khanna *University of Toronto*

Nature photography was almost the first application of the electronic flash system of lighting when it became commonly available.[1] Although these first units used paper capacitors charged to high voltages and discharging at microsecond rates, later "strobes" employed lower-voltage capacitors and operated at longer flash durations. Hence, today's conventional strobe is not suitable for the rapidly moving subjects of nature photography. This article describes a development in strobe equipment that incorporates new and improved design to create flash units of not only lighter weight, but with exposure times of less than 100 microseconds.

Almost from inception, persons keen enough to grasp the impact of the electronic flash system of lighting on nature photography, such as Kane,[2] Mili,[3] Griffin,[4] Marden,[5] and others,[6–8] saw and exploited its unique advantages—short exposure, large output, and daylight color quality.

The first equipment of this type used paper capacitors charged to 3000 volts, which produced flashes less than 100 μs in duration. Later, improved flash lamps of higher efficiency employing smaller electrolytic capacitors (450 volts) were developed that greatly reduced the weight of the flash equipment; however, flash duration became longer than a millisecond for most units. Hence, the conventional modern-day electronic flash lamp ("strobe") commonly used for all sorts of routine photography has become unsuitable for the rapidly moving subjects of nature photography, such as the moving wings of birds, because of the extended exposure times.

It should be noted that the color cover photograph was taken by Crawford Greenewalt using the equipment described in this article. In particular, the wing tips of hummingbirds, if photographed with a flash of light of

conventional exposure time, will show an objectionable blur of the wings. Figure 1 was taken with a flash duration of 0.5 ms; an exposure time of less than 100 μs is required to satisfactorily expose an end-on view of a wing and some photographers even insist upon a still shorter flash.

Several portable special flash units for bird photography were made at M.I.T. in 1950 with support from the Research Committee of the National Geographic Society. This equipment, which was also furnished to Allen[9] and Van Riper,[10] had a flash duration of about 250 μs, slightly too long for photographing hummingbirds if the wings are in a position to show maximum velocity across the field of view. This unit used the GE FT-110 flash lamp excited at 900 volts through a 250-μF capacitance using a series-parallel connection of four 250 μF, 450-volt electrolytic capacitors.

There are other special-duty strobes, such as the General Radio Strobolume (30 μs) and the EG&G, Inc., type 549 Microflash, which have sufficiently short flash durations. The light from the latter will produce crystal-sharp pictures since the exposure time is less than half of a microsecond. However, because this equipment is somewhat bulky and requires an ac power connection, it is not convenient for field use. More important, the light output is not sufficient to take color photographs of birds at the aperture needed for the desired depth of field.

Greenewalt's[11,12] excellent hummingbird photographs were exposed with a special flash unit that uses three FT-220 lamps at 2000 volts, each excited by a 14-μF paper capacitor. The short (75-μs) flashes of light from this unit produce bird photographs that are sharp and clear without blur. The flash equipment used by Hosking and Newberry[13] and Davidson[14] are also of special design involving paper capacitors that are operated at relatively high voltage.

There continues to be a great interest in bird photography using electronic flash and color film. Awareness of the need for improvement in this field has been intense for years, as a result of pressure by Arthur A. Allen,[9] Walker Van Riper,[10] Crawford Greenewalt,[11,12] Eric Hosking,[13] Treat Davidson,[14] Walter Scheithauer,[15] and many others. The results reported here stem particularly from comments and interest expressed by Greenewalt. Also, Hosking has listed for us the desired qualities of portable flash equipment for bird photography. Foremost on this list is the demand for a lightweight apparatus for use in out-of-the-way places.

A strobe equipment is described here that incorporates improved new components leading to flash units of lighter weight and an exposure time of less than 100 μs. The improved circuit elements are

1. Fused-quartz flash lamps of short arc length and relatively large diameter, to produce an efficient and brief flash powered from an electrolytic capacitor. Quartz is used because a glass flash lamp of the same size might be crazed by the intense discharge.

2. A special electrolytic capacitor of low internal resistance. Such a capacitor is required for producing a short flash. The conventional electrolytic capacitor has an internal resistance that reduces the efficiency of light production and also prevents the flash from being of short duration. Paper capacitors have low internal resistance, but the electrolytic capacitor stores more energy per kilogram.

3. A power supply of small weight and size with voltage regulation for constant voltage. Several types of charging circuits are described here.

The light measurements, experiments, and designs reported are the joint effort of the Strobe Lab staff at M.I.T. together with students* of the Electrical Engineering Department who were working to meet the requirements of a course laboratory project during the spring term of 1968.

Circuit theory

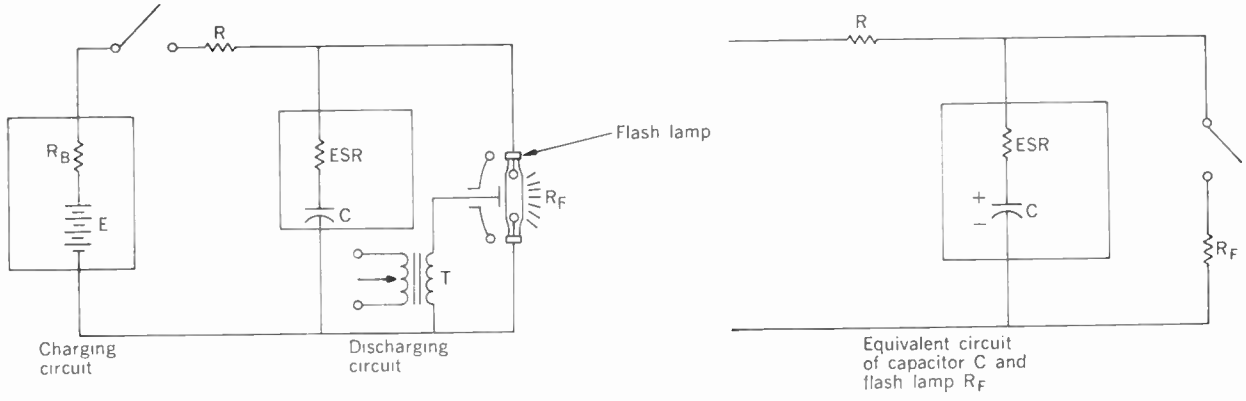
An approximate circuit representing the xenon flash lamp excited from an electrolytic capacitor is shown in Fig. 2. There are three nonlinear circuit elements that make the exact mathematical analysis of the circuit impractical at the present time:

1. The battery E, whose open-circuit voltage decreases with use and with time. The internal series resistance R_B becomes larger as the battery is discharged; it is also a function of temperature and time.

2. The electrolytic capacitor C, with its internal "effective series resistance" (ESR). This series resistance depends upon the capacitor construction. The Sprague Electric Company recommends the 250-volt type D

* The students who participated are D. Bovin, P. Chu, R. Dorman, A. Fillat, R. Latham, C. Owen, S. Poppe, and G. Varga.

FIGURE 2. Circuit of a xenon flash lamp showing the three nonlinear electric devices that are considered in the approximate analysis.



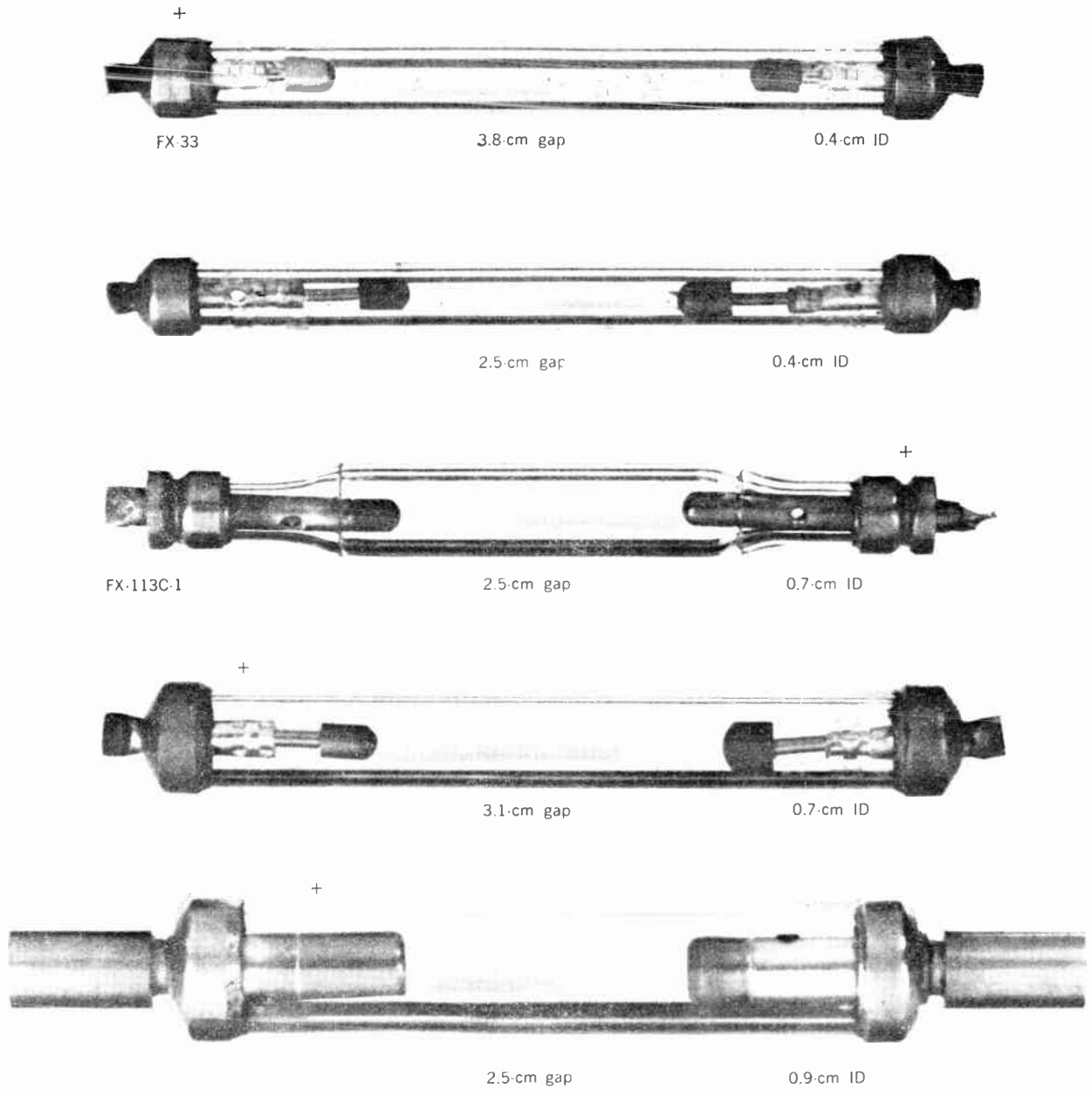


FIGURE 3. Quartz flash lamps containing xenon gas that are used in the design of a flash unit for high-speed nature photography.

45495 as having the lowest ESR per farad. Xenon flash lamps are not commonly available at 250 volts, so two capacitors in series were used in the final circuit of 500 volts. Two strings of series capacitors constitute a total connected capacitance of 250 μF for each flash lamp, with 31 joules of energy stored in the four capacitors when the combination is charged to 500 volts. The ESR of the four-capacitor combination is about 0.25 ohm.

3. The flash lamp, whose current increases from zero to a high value and then decreases again to zero after the discharge has stopped. The "resistance" R_F of a flash lamp can be defined in terms of peak current and initial voltage. Thus, $R_F = E/I$ if the energy exceeds about 30 joules per cubic centimeter, which is the condition for the arc completely to fill the arc chamber. If the arc does not fill the cross section, the tube resistance will be higher.[16]

If the circuit model (right-hand side of Fig. 2) is considered to be adequate, then the following conclusions can be drawn:

(a) The discharge time of the light to $1/e$ is $C(R_F + \text{ESR})/2$ seconds. This is a time constant for power, not current, so the 2 in the denominator is required. The one-third peak duration of an actual flash is usually about 50 percent longer due to the build-up time of the discharge, which is not considered in the linear model.

(b) The efficiency of transfer of energy into the flash from the capacitor is $R_F/(R_F + \text{ESR})$. Neither this statement nor the preceding is strictly valid, however, since both ESR and R_F are functions of voltage, time, age, temperature, current density, and perhaps certain other factors.

It is important to note that $R_F + \text{ESR}$ should be small

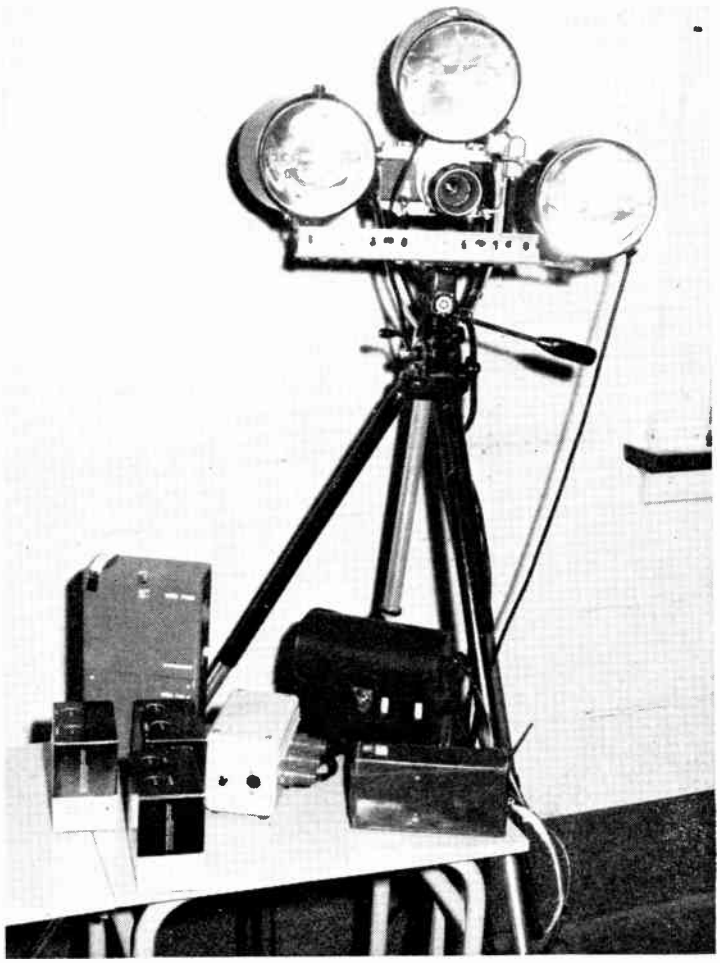
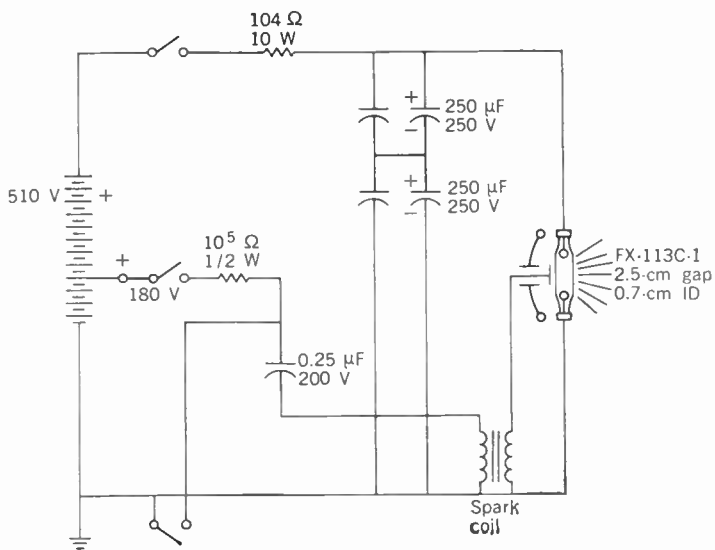


FIGURE 4. Three-lamp nature photography equipment with low-ESR capacitors and low-resistance flash lamps. Flash duration is less than 100 μs.

FIGURE 5. A battery-operated short-duration electronic flash unit used for bird photography.



I. Light output and flash duration (with low ESR capacitors)

Lamp	Gap, cm	Inside Diameter, cm	R_F,* ohms	Output, horizontal cd·s	Duration, μs
FX-33	3.8	0.4	0.50	124	120
Special	2.5	0.4	0.33	100	90
FX-113C-1†	2.5	0.7	0.11	95	60
Special	2.5	0.9	0.06	66	50

All data measured at 500 volts; capacitors are composed of four 250-μF, 250-volt capacitors (Sprague D45495) connected in a parallel-series combination. The stored energy at 500 volts in four capacitors is 31.6 joules. The ESR of one capacitor is 0.25 ohm.

* Lamp resistance referred to the EG&G, Inc., FX-1 flash lamp of 2.0 ohms at a 15.2-cm length and an inside diameter of 0.4 cm assuming the resistance is proportional to arc length and 1/area.

† This lamp was selected as being well suited for photographing hummingbird wings in action.

II. Results of tests with one battery and a 250-μF capacitor

Voltage		Number of	Recovery
Initial	Final	Flashes	Time
491	458	48	2 days
482	465	60	30 minutes
460	444	46	—

in order to realize a short-duration flash of light, and that R_F should be greater than ESR in order to achieve greater efficiency.

The flash lamps

Five quartz flash lamps of different internal dimensions were tried as light sources, each operating from a series-parallel system of four low-ESR capacitors (Sprague No. 45495). The lamps are shown in Fig. 3. Along with the dimensions of each arc, Table I gives approximate output as measured with a General Radio type 1501 light meter, which uses an RCA type 929 phototube. The flash duration was measured with a type 929 phototube, a 1000-ohm resistor, and an oscilloscope. It appears from Table I that any of the lamps except the FX-33 would be satisfactory for our purposes, since their flash duration is less than 100 μs.

The flash lamp with a 2.5-cm gap and 0.7-cm ID (FX-113C-1 in Table I) was selected for practical equipment, partly because of its low R_F. Three such lamps were mounted in reflectors of six-inch diameter; Fig. 4 is a photograph of the setup used for these three lamps in hummingbird photography.

The power supply

Several systems of supplying power to charge the capacitors have been considered for this nature flash unit. Factors of importance are

1. The number of flashes that can be taken with the battery.
2. The regulation of light output from the start of a freshly charged or new battery to the end point.
3. The time required to charge the capacitor for the next flash.
4. The weight.

The simplest power system is the standard commonly

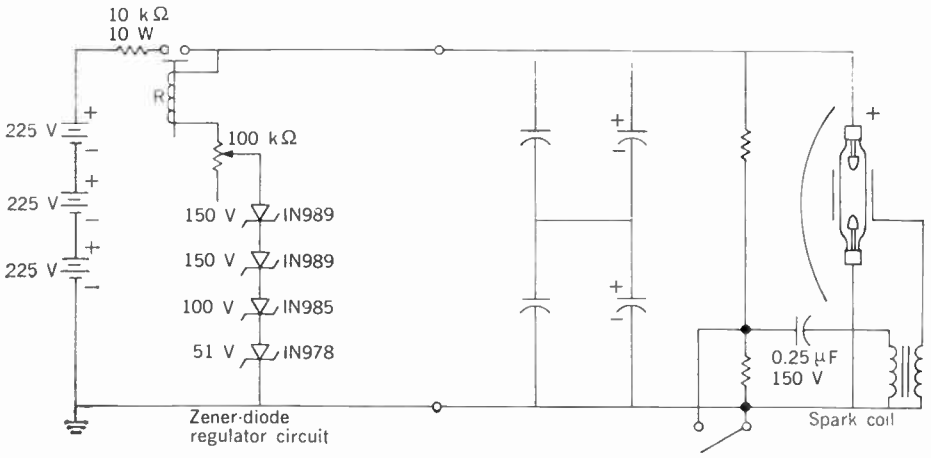


FIGURE 6. Regulated power supply with a string of Zener diodes used to open a relay when the desired voltage (500 volts) is reached. Specifications: R Sigma 4R-8000 S-SIL 2-A contacts; coil = 8 kΩ, 1.6 mA closed, 0.75 mA open.

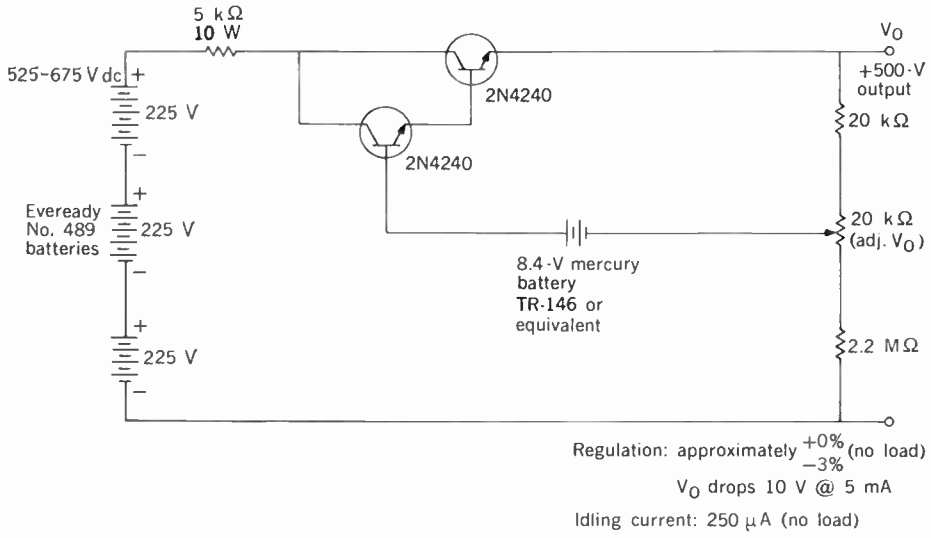


FIGURE 7. Regulating circuit for charging an electrolytic capacitor to 500 volts. With batteries at 650 volts, 750 μF (three capacitors similar to those of Fig. 8) will be charged to 480 volts in about 12 seconds.

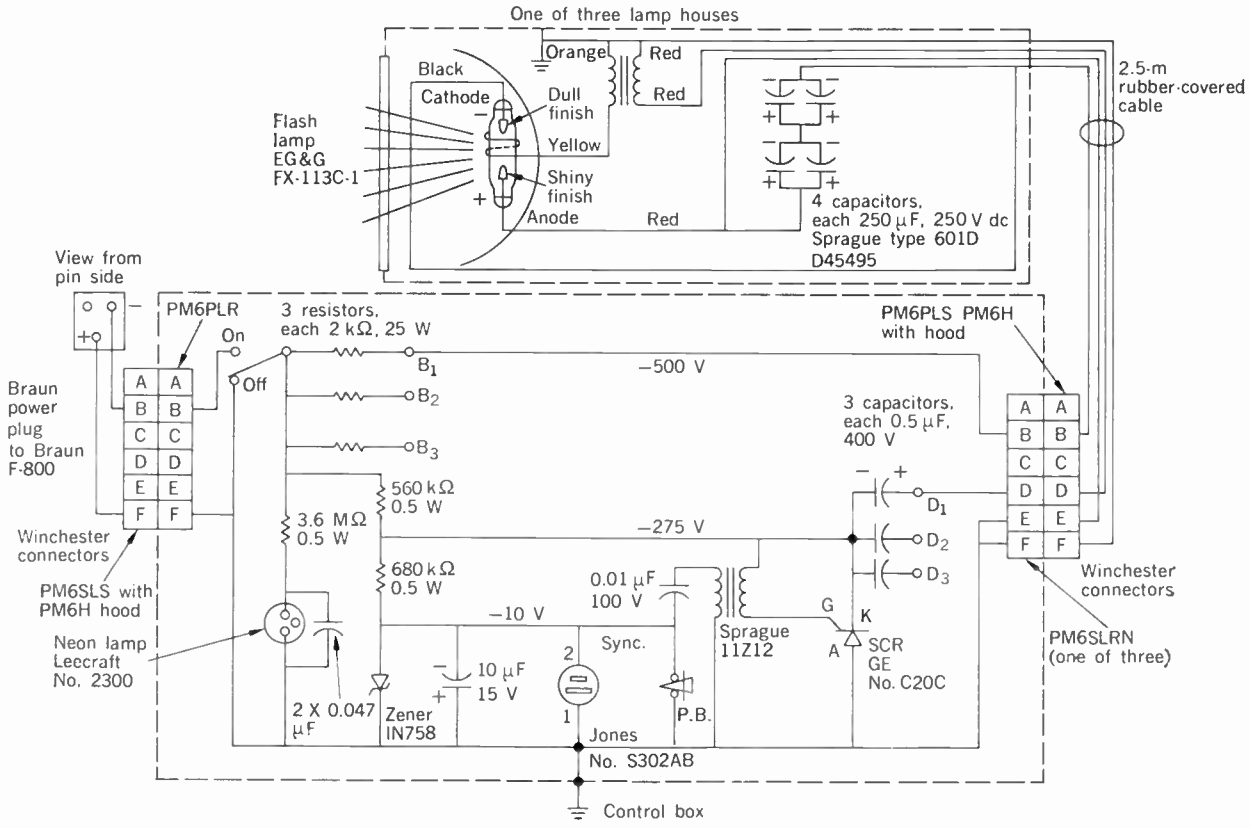


FIGURE 8. Complete circuit of a three-lamp electronic flash unit of less than 100 microseconds and powered by a Braun F800 unit.

available 510-volt dry battery, as used in many strobe flash units. The battery weighs 0.74 kg, and is available from Burgess (No. 497), Mallory (No. PF497), and others. A tap at 180 volts is available for use in the spark circuit, as indicated in Fig. 5. This battery has rather poor regulation and low voltage output when a large number of flashes are obtained in a short time. It would be worthwhile for a photographer to have a voltmeter with him to check the voltage and to make corrections for the optical output as the capacitor voltage is reduced during the life of the battery.

Tests of one battery (previous life unknown) with a 250-μF capacitor charged through a 10-kΩ resistor and a one-minute flash rate indicated the performance listed in Table II. Notice that the rated 510 volts is never achieved.

There are several commercial flash units with power supplies that can be used to charge to 500 volts the special equipment described here. Two examples are the Braun F800 and the Multiblitz Press. These units have regulators to hold the voltage constant. The user must remember that the metal prongs on the plug connecting to the special flash will be live even if removed, since the capacitors can retain their charge for many minutes. He will do well always to flash the lamps after turning off the switch and before pulling out the plug!

Two types of regulator charging circuits were tried, and are displayed in the schematics of Figs. 6–8. Figure 6 shows a Zener-diode regulator with a relay to open the charging circuit when the capacitors have been charged to 500 volts. Figure 7 shows a reference-voltage regulator system that uses an 8.4-volt mercury battery as a reference. The power supply should be selected to be commensurate with the photographer's particular needs in respect to such parameters as weight, regulation, number of flashes, and lifetime.

A complete electronic flash system has been assembled for photographing hummingbirds; the color picture on the cover of this issue was taken using this equipment. This system consists of three lamphouses connected by a small junction and control box to a Braun F800 flash-unit power supply. The circuit is shown in Fig. 8, and features include:

1. Stored energy of 31 joules per lamp.
2. Light output of 1500 beam candela seconds per lamphouse with a 28 degree spread.
3. Light duration of 65 μs.
4. Voltage-regulated power supply for constant light output.
5. Up to 200 flashes from one battery charge.
6. Light weight and portability:
   (a) Three lamphouses and junction box (3.37 kg).
   (b) Braun F800 power supply (1.95 kg).
   (c) Aluminum carrying case and accessories (3.2 kg).
   (d) Total weight of three-lamp flash system and accessories in carrying case (9 kg).

A safety switch is incorporated that interrupts the high-voltage circuit from the power supply and dissipates the energy stored in the flash capacitors so that the power plug may be disconnected safely.

REFERENCES

1. Edgerton, H. E., and Killian, J. R., Jr., *Flash, Seeing the Unseen*. Newton Center, Mass.: Branford, 1954 (1st ed. 1939).
2. Kane, H. B., Series of nature books for children, published by Alfred A. Knopf, New York.
3. Mill, J., "Cock fight," *Life*, Apr. 17, 1939.
4. Griffin, D. R., "Bats," *National Geographic*, July 1946.
5. Marden, L., "Color photographs of parakeets and parrots," *National Geographic* (unpublished).
6. Webster, F., *Internat'l Cong. on Technology and Blindness*, vol. 1, p. 49, 1963.
7. Cahlander, D. A., McCue, J. J. G., and Webster, F. A., *Nature*, Feb. 8, 1964.
8. McCue, J. J. G., *National Geographic*, Apr. 1961.
9. Allen, A. A., "Stalking birds with color camera," *National Geographic*, June 1948, May 1954.
10. Van Riper, W., Niedrach, R. J., and Bailey, A. M., "Nature photography with high speed flash," Denver Museum of Natural History pamphlet.
11. Greenewalt, C., *Hummingbirds*. New York: Doubleday, 1960.
12. Greenewalt, C., *National Geographic*, Nov. 1960, Jan. 1963, July 1966.
13. Hosking, E., and Newberry, C., *Birds in Action*. London: Collins, 1949.
14. Davidson, T., "Photoelectric nature photography," *PSA J.*, pp. 31–36, Mar. 1958.
15. Scheithauer, W., *Hummingbirds*. London: A. Baker, 1967.
16. LeCompte, W. W., and Edgerton, H. E., "Xenon arc transients, electrical and optical," *J. Appl. Phys.*, vol. 27, pp. 1427–30, Dec. 1956.

Harold E. Edgerton (F, L) received the B.S. degree from the University of Nebraska in 1925, and the M.S. and D.Sc. degrees from the Massachusetts Institute of Technology in 1927 and 1931, respectively. His pioneering research work in the field of stroboscopic photography was the foundation for the development of the present-day electronic speed flash. He has designed watertight cameras with electronic flash lamps, is a consultant on underwater flash photography and stroboscopy, and has been working with Capt. Jacques-Yves Cousteau in explorations of the floor of the Mediterranean Sea. He holds the position of institute professor at M.I.T. and also serves as honorary chairman of the board of EG&G, Inc. Dr. Edgerton currently is developing sonar devices for the positioning of equipment in the sea and for the exploration of the sub-bottom structure.

Vernon E. MacRoberts is currently a project technician with the Department of Electrical Engineering at M.I.T. in Cambridge, Mass. A 1954 graduate of the Lowell Institute School at M.I.T., Mr. MacRoberts has worked with Dr. Edgerton at M.I.T. for more than 25 years. He has designed and built much high-speed photographic and sonar exploration equipment, used for scientific investigation throughout the world.

Manmohan Khanna (S) received the B.S. degree in electrical engineering in 1965 and the M.S. degree in electrical engineering in 1968, both from the Massachusetts Institute of Technology. After working under an IBM cooperative program in Poughkeepsie, N.Y., from 1965 to 1966, he was employed as a teaching assistant in Dr. Edgerton's Stroboscopic Laboratory during 1967–1968. Mr. Khanna is a member of the honorary society Eta Kappa Nu, and is now studying toward a Ph.D. degree in stochastic control at the Department of Electrical Engineering, University of Toronto, Canada.