IEEE spectrum

features

Spectral lines

Things looked worse in the thirties. Some of us now have a pessimistic view of the future of electrical engineering, believing that we are too many for the foreseeable demand. There was a very similar feeling in the 1930s, based on a scarcity of employment over nearly a decade and higher levels of unemployment than we now have. During the '30s none of us even daydreamed the order-of-magnitude expansion of our numbers and importance that was to come. We had powerful methods of analysis and synthesis that would later be applied to make our vacuum-tube technology produce radar, control systems, and computers, but we only perceived enough to hope to find a job doing FM equipment design, if FM became a commercial success. The historical parallel suggests that it may be worth considering whether our arts and methods of thinking are primarily of value for weapons and space systems or will be increasingly needed in more fundamental and less volatile fields.

The contemporary competences of electrical engineers can be classified by application in the three fields of power, information, and materials science. The acquisition, transmission, processing, storage, and display of information are our most generalized ability and the principal area of expansion of the past 30 years. Electric power is an essential service that will probably grow in importance as petroleum is replaced as a basic fuel. Many electrical engineers have become skilled in materials science at a sophisticated level because it was required to fabricate semiconductor, magnetic, optical, and acoustic devices, and they were as well qualified to learn these new techniques as were conventionally educated metallurgists and chemists.

Looking forward, there are reasons to believe that the field of information will expand. One reason is its very low consumption of energy and raw materials. The human race may be required to conserve fuel, food, raw materials, air, and water, but we can offer it relatively unlimited access to, and interaction with, information. A more fundamental reason for the expansion of information as a field of work is that the reception, processing, and generation of information are the primary occupations of the race. Most repetitive toil is being tooled out of existence. Even where it is retained, the toilers are now supported by doctrine and entertainment. We are developing the technology that will make a wide extension of information services feasible economically. The computer as we know it may only be an evolutionary form, but something like it, perhaps more easily coupled to people, can soon be a part of every child's education and life tools. We will be able to provide universal communication and information systems offering the opportunity for study, library research, recreational research, and convenient communication with people of like interests so that we can all enjoy the characteristics of a university throughout our lives. New forms of entertainment and art will develop as we make them practical. We will find input languages and techniques for sound and visual pattern synthesizers that will make them widely used instruments because they open artistic creativity to people who do not have all the talents required by conventional methods. Thus the applications of our information arts will surely expand where they are now applied but, as we continue to lower their cost, they can become an intimate part of every life.

Electric power cannot expand exponentially because of ecological constraints, but its application will become more sophisticated and it is likely to expand because it will be the practical way to replace combustion power. Power semiconductors are coming of age, and will be widely applied to both motors and static loads. They also increase the market for information processing because of their good response to electric control. As we come to regard electric power as a scarce resource, precious beyond its economic cost of production, we will devote more effort to optimizing its generation, distribution, and application. Finally, it is very likely that there will be a resurgence of electric transportation in the United States. For all of these reasons, it seems likely that engineering opportunities in power will continue to increase.

We may never regain the employment that was provided by weapons and space systems in the '60s but the civil applications of those arts in air traffic control, navigation, and communications will continue to grow. It is worth speculating about the possible civilian market for radar-like devices given the new, potentially inexpensive semiconductor microwave components and very inexpensive IC signal processing. Communications satellites are only at the beginning of their application.

The ideas and tools of electrical engineering are now used extensively in the life and health sciences. Much of the instrumentation is electric, and our ability to process and store electrical data is being used increasingly. Industrial processes and controls will continue to add tools with a high electrical engineering content, such as electron microscopes, ion beams, film-deposition equipment, and electric analytical and sensing equipment.

The opportunities for electrical engineers that we have mentioned are the visible ones. They correspond to the hopes we had in the '30s for FM and television. Perhaps, as in those days, the greatest opportunities are the ones not yet perceived.

David DeWitt, Editor

The promise of controlled fusion

Although controlled fusion for electric generation has presented extraordinarily difficult technological problems, recent experimental results have created a new atmosphere of optimism for the coming decade

Robert G. Mills **Princeton University**

One of the outstanding reasons for the active development of controlled fusion nuclear reactors as a source of electric power, aside from economic advantages, is that they are extremely attractive from an environmental standpoint. The basic principles and some of the most pressing problems are presented in this article. Among the problems considered are plasma confinement, heating, fuel injection, and exhaust collection.

It has been almost 20 years since the scientific community undertook the most difficult task in technology so far conceived. These years have seen periods of optimism alternating with periods of disappointment. In the past, optimism was generally associated with fresh new ideas that seemed to have much promise. In contrast, the high optimism that now pervades the fusion power fraternity is based on concrete experimental results that bode well for the future. Although practical reactors that will burn fusionable fuel to manufacture electricity have yet to be proved possible, the expectation is that scientific feasibility will be demonstrated during the seventies. If this crucial step is successfully taken, commercial power may make its first appearance in the nineties. Such a development would provide an inexhaustible supply of fuel at negligible cost, with worldwide distribution.

The cost of fusion power will be dominated by the investment cost of the plant. Economic feasibility cannot be conclusively demonstrated until after scientific feasibility has been shown, and an accurate catalog of plant components and their specifications can be compiled. Current estimates, however, show these systems to be potentially very attractive from an economic standpoint.

Today environmental questions loom larger and larger in all aspects of society's activities. From this point of view the advantages of fusion power seem especially attractive. Clean, quiet, and hazard-free, a nuclear fusion power plant would be a good neighbor.

This work was performed under the auspices of the U.S. Atomic Energy Commission.

Fundamental requirements for the production of fusion power

The fusion process. Of the dozens of known nuclear fusion reactions, very few seem to be of interest for potential use for power generation. Since the positive electronic charge on nuclei produces large repellent forces tending to prevent the nuclear collisions that are necessary to produce fusion reactions, only those with the largest nuclear fusion cross sections merit consideration. Table I lists the only reactions for which the cross section is larger than a millibarn (10^{-27} cm^2) at energies below 50 keV. Since the nuclear scattering cross sections are much larger, any system for producing net power from nuclear fusion must provide an environment in which a fuel nucleus undergoes many collisions with other nuclei before it is lost from the system. Consequently, if incoming fuel possesses organized motion, it will be randomized by subsequent collisions, and the

I. Nuclear fusion reactions

D + D	→	T (1.01 MeV) + p (3.03 MeV)
D + D	→	He^3 (0.82 MeV) + n (2.45 MeV)
D + T	→	He^4 (3.52 MeV) + n (14.06 MeV)
D + He^3	→	He^4 (3.67 MeV) + p (14.67 MeV)
T + T	→	He^4 + n + n

reactants will develop a kinetic equilibrium describable by a temperature; that is, they will possess a Maxwellian distribution, and the fuel will be in the form of a hot gas. Since the energies required to produce a reasonable reaction rate are in the kiloelectronvolt region, well above the ionization energies of the light elements that are of interest, this gas will be a fully ionized plasma.

Under these conditions, the reaction rate will be proportional to the square of the ion density and to the reactivity, a quantity defined as the average over the relative velocity distribution of the product of the nuclear fusion cross section and the relative velocity of collision. Figure 1 shows the energy dependence of the fusion cross section and the temperature dependence of the reactivity for the reactions of greatest promise. A plasma containing a mixture of deuterium and tritium ions will have a D-T reaction rate density of

$$n_D n_T \langle \sigma v \rangle_{DT} \quad (1)$$

where n_D and n_T are the deuteron and triton densities, respectively, and the term in brackets represents the reactivity. In addition there will be D-D fusion reactions at the rate

$$\frac{1}{2} n_D^2 \langle \sigma v \rangle_{DD} \quad (2)$$

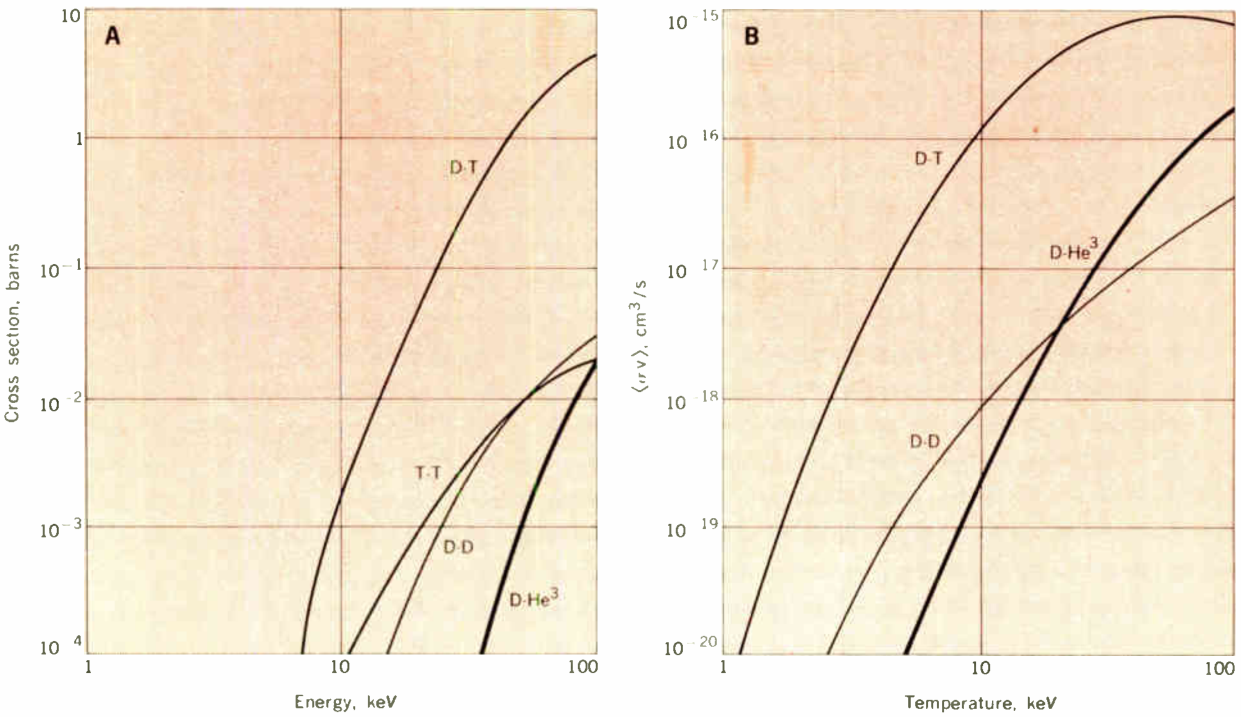
and T-T reactions at the rate

$$\frac{1}{2} n_T^2 \langle \sigma v \rangle_{TT} \quad (3)$$

In hydrogenic plasma, for every ion there is an electron. The strength of electrostatic forces is sufficient to keep all plasmas very close to overall neutrality—i.e., equal numbers of ionic charges and electrons.

The collisions between electrons and ions produce bremsstrahlung, or X-radiation, that is lost from the plasma, carrying away energy that can only be recovered in a heat cycle by cooling the reaction chamber's walls where the bremsstrahlung radiation is deposited. Since this process, like the fusion process, is a result of binary collisions, it too is proportional to the square of the particle density. It is also proportional to the square root of the electron temperature. Thus the bremsstrahlung

FIGURE 1. Energy dependence of cross section and reactivity.



power loss density, p_B, is given by

$$p_B = b n_e^2 T_e^{1/2} \quad (4)$$

where b is a constant, n_e is the electron density, and T_e is the electron temperature. In systems of relatively long confinement time the ions and electrons will be at essentially the same temperature, but in systems of short confinement time appreciable temperature differences can develop between the ion and electron seas.

Relative power output. The total energy U released as a result of a single fusion event can be considered the sum of three terms: the kinetic energy carried by the neutrons produced in the reaction, the kinetic energy carried by the charged particles produced in the reaction (for example, He^4, He^3, or p), and finally the energy that is ultimately released as the result of the absorption of the neutron in the material surrounding the reaction chamber. The last depends on the details of the machine, but will usually be in the range of 4 to 5 MeV. It consists of the energy released as a result of the radioactive decay of the unstable nuclei formed by neutron absorption. Most of it is in the form of gamma radiation; some, in the form of charged particles such as beta decay products. The sum of the first two, the total energy release of the fusion reaction per se, is listed as E in Table II. The energy released to charged particles, E^*, is important since this is the portion available in the plasma to provide the energy lost by radiation from the electron sea.

The effective total power density released from a thermonuclear reacting plasma will be obtained by multiplying the reaction rates given by Eqs. (1), (2), and (3) by the total energy U released as a result of each reaction. As an example, consider a plasma of ion density n_i, made up of 50 percent deuterons and 50 percent tritons. Then, by (1), the effective power density p_F produced by D-T fusion reactions is given by

$$p_F = \frac{1}{4} n_i^2 \langle \sigma v \rangle_{DT} U_{DT} \quad (5)$$

Neglecting the considerably smaller energy release by D-D and T-T reactions, this quantity multiplied by the plasma volume would give the total power generated by the reaction.

The power density released to charged particles within the plasma is given by

$$p_F^* = \frac{1}{4} n_i^2 \langle \sigma v \rangle_{DT} E_{DT}^* \quad (6)$$

which is available to supply radiation losses.

The cross section and reactivity curves shown in Fig. 1 could be misleading to someone seeking the optimum operating temperature for a reactor plasma. He might conclude that the output power would rise with temperature throughout the temperature range depicted. This misinterpretation would result from ignoring the pressure limitation that any practical system will have. The true temperature dependence of the fusion power density of a system follows by introducing into Eq. (5) the fact that the operating density of a pressure-limited system will vary inversely with the temperature. Since

$$n_i \approx \frac{1}{T} \quad (7)$$

the fusion power density will be given by the following relation:

$$p_F \approx \frac{\langle \sigma v \rangle U}{T^2} \quad (8)$$

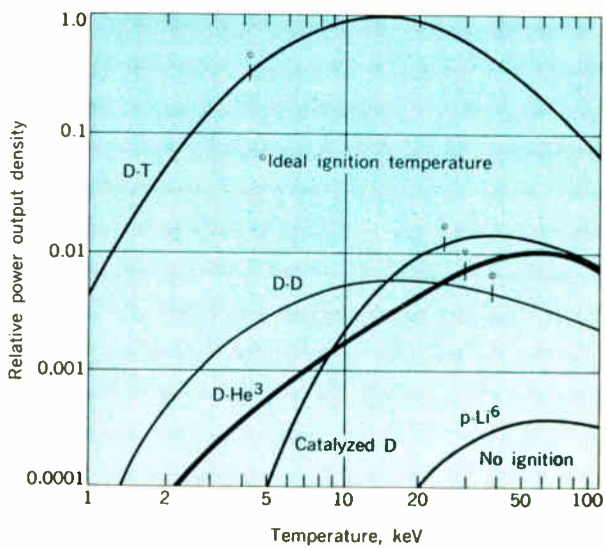
This function is plotted in Fig. 2 (normalized to one for the most favorable case, the D-T reactor at 15 keV) for a number of possible fuel systems.

The power released to charged particles rises more rapidly with temperature than does the bremsstrahlung radiation loss. Therefore an ignition temperature exists for the reactions that have sufficiently large cross sections. There are remarkably few of these.

The Lawson criterion. An analysis of the minimum necessary conditions for net power production in a fusion reactor was first published by J. D. Lawson in 1957.[1] His almost model-free system assumes the perfect confinement for a time τ of a reacting hydrogenic gas of ion density n_i. The total energy release of the system is assumed to be recovered at an efficiency η, and is used to heat the fuel charge to the operating temperature. He found that the product, $n_i\tau$, was the significant parameter to evaluate the feasibility of a system, and that it is a function of η and T_i.

For short confinement of a D-T system, his condition is expressed by

FIGURE 2. Relative output power densities for several typical fuel mixtures.



II. Energy yields

Reaction	Total Energy Yield E, MeV	Yield to Charged Particles E*, MeV
D + D	3.6	2.4
D + T	17.6	3.52
D + He^3	18.34	18.34

$$\left(\frac{1}{4} n_i^2\langle\sigma v\rangle_{DT} U_{DT}\tau + \frac{3}{2}(n_i + n_e)kT + bn_e^2 T^{1/2}\tau\right)\eta$$
$$= \frac{3}{2}(n_i + n_e)kT + bn_e^2 T^{1/2}\tau \quad (9)$$

The total energy release, assumed to pass through a thermal recovery cycle of efficiency η, is shown on the left and consists of three terms: the first is the power release from Eq. (5) multiplied by the confinement time τ; the second is the thermal-energy content of the ions and electrons (of particle densities n_i and n_e, respectively); and the third is the energy radiated by bremsstrahlung during the time τ and deposited on the chamber wall. The minimum condition for possibility is when this recovered energy is used to supply the thermal energy content of a fuel charge and provide the radiation losses during the confinement time τ—that is, the two terms on the right in (9).

Since we are dealing with a hydrogenic plasma, the electron density n_e will be equal to the ion density n_i, and Eq. (9) can be solved to yield the minimum $n_i\tau$ product required for a break-even condition. A successful reactor, of course, must be capable of operation at a higher $n_i\tau$ product than the Lawson criterion. The result is as follows:

$$n_i\tau = \frac{3kT[(1/\eta) - 1]}{\frac{1}{4}\langle\sigma v\rangle_{DT} U_{DT} + bT^{1/2}[(1/\eta) - 1]} \quad (10)$$

As an example, at an efficiency of 33.3 percent, the most favorable reactor mixture (50 percent D, 50 percent T) requires an $n_i\tau$ product of 10^{14} s/cm^3 at an ion temperature of 10 keV, the most widely quoted "Lawson criterion." His work can be generalized to yield the curves shown in Fig. 3 for the deuterium–tritium fuel system.[2]

Of the various possible fuels, only deuterium occurs in nature. If tritium or He^3 is to be used, it must be manufactured in the system. Since the D-T reactor requires by far the least stringent conditions, it has received the most attention. In the reaction

$$D + T \rightarrow He^4 + n$$

only one neutron is released when the tritium nucleus is consumed. The tritium requires breeding from lithium by the reactions

$$n + Li^6 \rightarrow T + He^4$$

and

$$n + Li^7 \rightarrow T + He^4 + n \quad (11)$$

and great care must be taken with the neutron economy to provide neutron multiplication and to avoid losses. Otherwise, less tritium will be bred than consumed. The region surrounding the vacuum tube, called the blanket, where the tritium will be bred, will be discussed later. It tends to be expensive.

There are abundant resources of lithium. However, there would be a decided advantage in avoiding the necessity for breeding fuel, and the ultimate goal of the fusion power program is to find a method for burning only deuterium, so abundant in nature that the potential energy content of a single liter of ordinary water is equivalent to that of about 300 liters of gasoline. As is seen in Fig. 2, the D-D reaction has a disadvantage

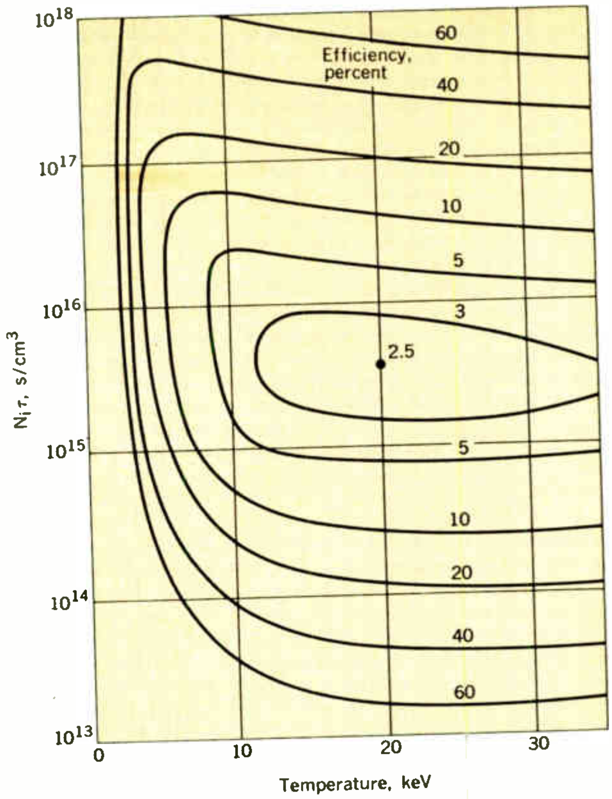


FIGURE 3. Lawson contours for the D-T reactor.

III. Catalyzing the reaction 3D → He^4 + p + n

Reaction	E	E*	n
D + D → T + p	4.04	4.04	0
D + D → He^3 + n	3.27	0.82	1
D + T → He^4 + n	17.6	3.52	1
D + He^3 → He^4 + p	18.34	18.34	0
6D → 2He^4 + 2 p + 2 n	43.2	26.7	2
or			
3D → He^4 + p + n	21.6	13.3	1

of about a factor of 200 in fusion power density relative to D-T. This requires so much larger a machine for a given power output that hypothetical models that are economic have not yet been suggested.

A way out of the difficulty may be to catalyze the reaction

$$3D \rightarrow He^4 + p + n \quad (12)$$

by the method illustrated in Table III. Basically, in a D-D reactor system, sufficient tritium (about 0.5 percent) and He^3 (about 15 percent) are added to make all four reactions proceed at the same rate, thereby maintaining a constant inventory of tritium and He^3. The factor-of-3 improvement in power density over the simple D-D reactor should be sufficient to make such systems attractive, since a blanket structure that need not carefully preserve neutrons will be much less expensive than one suitable for breeding.

Many advantages would accrue to a system that made no neutrons and delivered all of its energy to charged particles. The reaction of this type of largest cross section is

$$p + Li^6 \rightarrow He^3 + He^4 \quad (13)$$

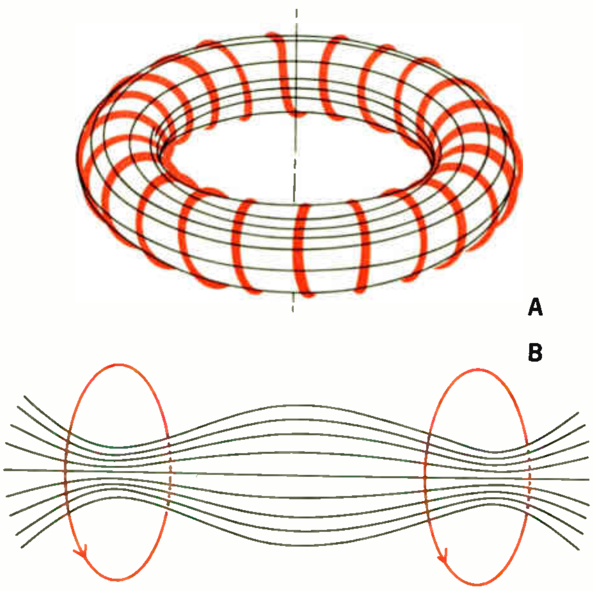
Note that

$$D + He^3 \rightarrow p + He^4 \quad (14)$$

does not qualify because in a plasma of deuterium and helium-3, this process would be accompanied by D-D reactions, half of which produce neutrons. Unfortunately, the reaction rate for Eq. (13) is insufficient to provide the bremsstrahlung radiation loss. Thus a reactor for utilizing this fuel would be extraordinarily difficult to achieve, although not impossible in principle as energy may be provided to the plasma from an external source.

Closed and open systems. The fundamental problem of the first phase of fusion power development, which lies in the realm of plasma physics research, is to heat a plasma to the temperatures required and (the real *sine qua non*) to confine it for the length of time implied by the Lawson criterion. It is this task that has occupied the attention of hundreds of scientists and engineers in a dozen countries for about 20 years. Since the confinement must be at long range, a gravitational or electromagnetic field is required. Failing this, only inertially confined systems (explosions) can be utilized. The stars are the only steady-state fusion reactors in existence, and they are as large as they are because of the extraordinary weakness of the gravitational field. The magnetic field is the appropriate tool for terrestrial applications. All magnetic containers for hot plasma fall into two general categories: those that are closed and those that are open. In their simplest possible forms, two archetypes are illustrated in Fig. 4. Neither of these is suitable for stable plasma confinement. An open system can be made suitable by the addition of "Ioffe bars," named after the Russian physicist, Ioffe, who announced this striking advance in 1961.

FIGURE 4. Confinement systems for (A) closed and (B) open magnetic fields.

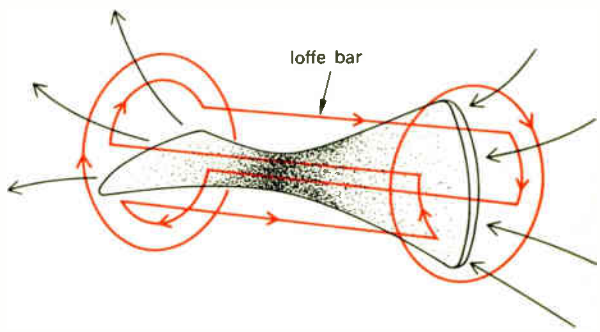


The basic idea behind Ioffe bars is to provide a magnetic field configuration in which there exists a region where the intensity of the magnetic field increases in every direction outward. The resulting field is illustrated in Fig. 5. In such an environment a typical plasma is stably confined, since escape for most particles requires an increase in energy. Unfortunately, the properties of magnetic fields require that any system with this property will have field lines that leave the volume, and although most plasma particles will be confined, those whose velocity vector is almost parallel to a field line leaving the volume will escape directly along the line. This "hole in velocity space" is replenished by the randomizing collisions of the plasma, and the gas is lost at a relatively high rate. In other words, open systems (called magnetic mirrors) leak badly through "end loss."

To avoid end losses, one should omit the ends by closing the field into a toroidal configuration. This produces a transverse field gradient that disrupts confinement completely. To make the toroidal field suitable for confinement one must add a poloidal field, which is orthogonal to a toroidal field. The magnetic field produced by a current in a closed circular ring is the simplest poloidal field. Three techniques for creating the field are illustrated in Fig. 6. In the first case an external set of helical-shaped windings is applied to the reactor tube; this is called a stellarator. In the second a current is induced to flow around the closed plasma ring; this is called a Tokamak. In the third case a ring conductor replaces the Tokamak's plasma current, and the plasma forms a closed hollow toroidal tube surrounding the conductor. Such machines have been called spherator, levitron, or floating multipole. The word "floating" is significant because the only way to avoid the fatal disturbance of the plasma by a penetration of material supports is to make the internal conductor a closed superconductor and levitate the insulated ring magnetically. Experimental machines of all these types have operated successfully. The basic confinement mechanism at work in all these devices is the effect the strong magnetic field has on the motion of a charged particle. The particle is confined to a helix about a given flux tube. Although the particles are free to move along the field direction, motion across the field is severely inhibited.

A dimensionless number, β, is used as an index of the effectiveness of a magnetic confinement scheme; it represents the ratio of the pressure of a plasma to the magnetic-energy density of the externally applied magnetic field, colloquially the ratio of "plasma pressure to

FIGURE 5. Magnetic mirror with well.



magnetic field pressure," as follows:

$$\beta = \frac{(n_i + n_e)kT}{B^2/8\pi} \quad (15)$$

The maximum possible value for the number is unity, which represents complete exclusion of the magnetic field by the diamagnetic plasma. Mirrors exhibit the ability of producing high-β confinement, but have the disadvantage of short confinement times. Toroidal geometries are capable of providing long confinement times but they have the disadvantage of being limited to low values of β.

The scaling law. The power density of a plasma varies as the square of the ion density. From (15), the density at a given temperature is proportional to βB^2. Therefore the total output power of a given type of reactor model will vary as

$$P_F \sim r^3\beta^2B^4 \quad (16)$$

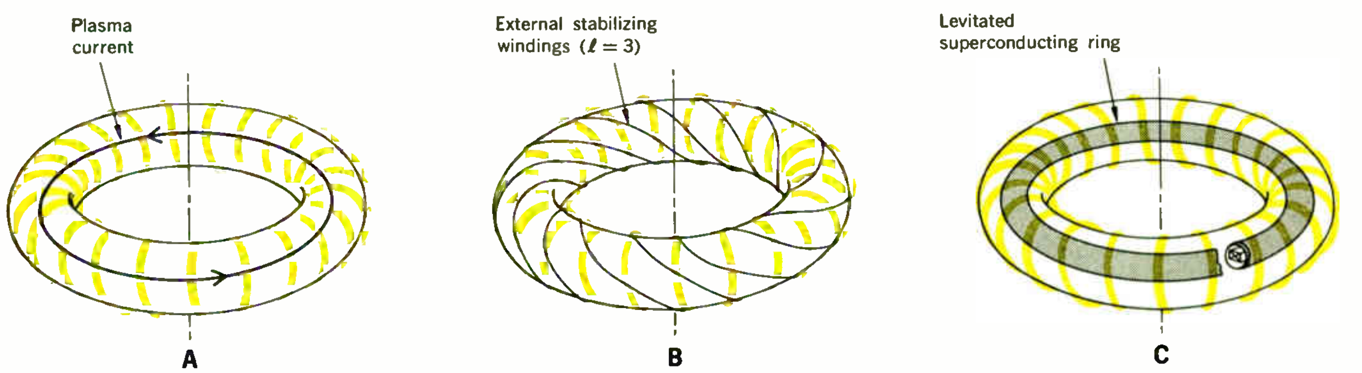
where r is any characteristic length of the machine, such as the plasma radius. This expression illustrates the importance of high magnetic fields to the fusion power program and the central role that superconducting technology will undoubtedly play in future power production.

Since scientific feasibility is yet to be demonstrated, it would be premature to select any particular model as most likely to lead to a reactor. However, a few comparisons can be made with respect to certain advantages and disadvantages of each. The open and closed systems lead to fundamentally different modes of operation. The closed-system advocates visualize a steady-state, or quasi-steady-state, system in which cold incoming fuel is heated to reactor temperatures by the charged reaction products trapped in the plasma. Injection is carried out in equilibrium with the diffusive plasma loss to give a density constant in time. Thus a nuclear "boiler" results, steadily burning the nuclear fuel with continuing cold fuel injection and hot exhaust collection in a closed fuel cycle. Since open systems are known to be incapable of producing sufficiently long confinement time to allow the ignition of cold fuel by the plasma, mirror reactors assume energetic injection of high-energy fuel in equilibrium with the stream of plasma escaping through a mirror. The short confinement time allows plasma conditions further removed from equilibrium than in a closed system, and the temperatures of the ion and electron seas can be quite different. This is important because the ion temperature needs to be kept high to lower the collision rate and minimize the end loss, whereas the electron temperature will sink under the loading of radiation, which becomes increasingly severe at the higher temperatures appropriate for mirrors. An open system, therefore, may have a large difference between the ion temperature and electron temperature and hence will not be governed by Fig. 2, for which the two temperatures have been taken as equal. Since the fuel is injected energetically, in principle the radiation power losses can be supplied also from energy carried by the incoming fuel, and the concept of an ignition temperature, which was described earlier, becomes irrelevant.

A mirror system must have a large amount of circulating power in the injector and mirror-stream-collection system, and there is a real premium for accomplishing both processes at the highest possible efficiency in order to make a mirror system possible. This need has led to the development at the Lawrence Radiation Laboratory, Livermore, Calif., of direct conversion devices of high efficiency in which the stream of plasma leaving the mirror is separated into ion and electron portions and electrostatically decelerated to convert its kinetic energy into dc electricity. Efficiency exceeding 90 percent may be achievable in this power loop. Note, however, that it is for only a portion of the total power of the system. The power in the neutrons and the power in the electromagnetic radiation must go through a thermal cycle, and the metallurgical limitation on the temperature of the vacuum wall will limit the efficiency of that portion of the plant to values similar to those for other heat engines. It is unlikely that the overall efficiency of a fusion power plant can exceed 60 percent. This figure, of course, would be a very significant improvement over present-day technology.

The proposed mirror machines have the relative advantage of high-β confinement, but suffer from the necessity of a very large circulating power. The toroidal machines have the advantage of being capable of as much confinement time as can be used, but suffer from the limitation of low β. Other hypothetical devices propose to combine the advantages of closed systems and high β by using pulsed theta pinches, devices such as the large equipment at Los Alamos, N.Mex., in

FIGURE 6. Toroidal configurations with poloidal shear. A—Tokamak. B—Stellarator. C—Spherator.



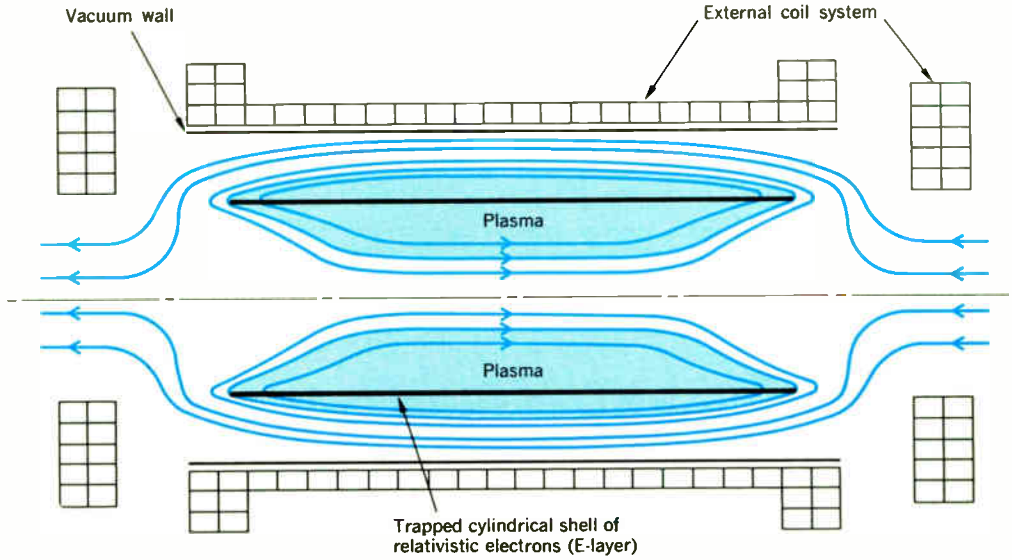


FIGURE 7. The Astron concept.

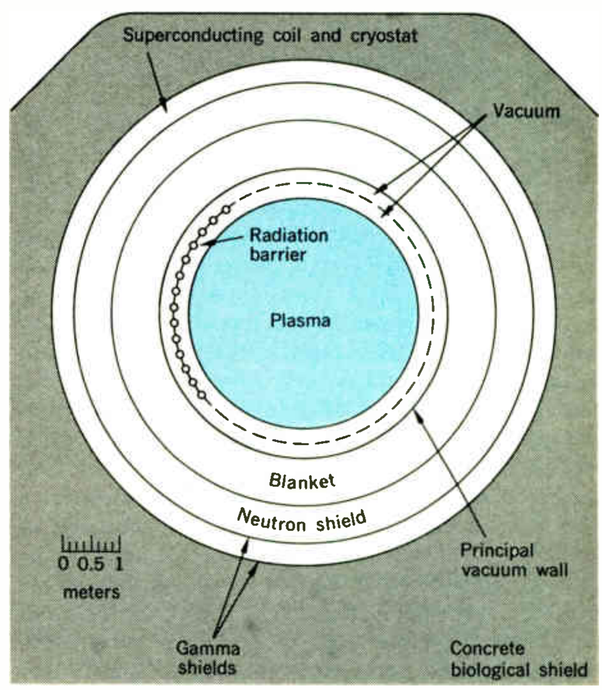
which a plasma is compressed by a rapidly rising, externally applied magnetic field. But this potential advantage tends to be nullified by the low duty cycle of the system. Finally, the condition that absolute magnetic wells imply mirrors (and their leaks) can be avoided if the current producing the field flows within the plasma. The Astron under development in California proposes to replace physical conductors by an energetic electron shell within the plasma volume itself. This ingenious but extremely complicated idea is shown in oversimplified form in Fig. 7. The trapped relativistic electrons (which make up the E-layer) are formed originally by generation in a large electron accelerator followed by injection and trapping within the mirror. In the next section we describe a system based on a closed geometry, but many of the features apply equally well to any reactor model.

One hypothetical reactor

In principle, once scientific feasibility is established, a power-producing reactor can be built by combining the following features: (1) a vacuum-tight reaction chamber, in which a low-density pure gas of fusible isotopes, such as an equal-volume mixture of deuterium and tritium at a density of about 5×10^{14} cm^{-3}, can be heated to reacting temperature and confined such that the average residence time at high temperature (10 keV or 10^8 K) is about 1 second; (2) a structure consisting of blanket and shield surrounding the reaction chamber in which neutrons will be thermalized and tritium will be generated; (3) a thermal cycle in which the heat produced in the blanket is converted to electricity; and (4) a magnetic coil system for generating the confining magnetic field. The translation of these broad categories of systems and the numerous subsystems they imply from conception to fully engineered designs represents a technological challenge of the highest order, the detailed consideration of which has just begun. Consider the cross section of a toroidal reactor and a few of the problems that immediately come to mind.

Functions of the elements. Figure 8 presents an elemental cross section of a hypothetical machine. Proceeding from the center out, the various principal elements meet these requirements. The low-density plasma consists of an equal mixture of deuterons and tritons (plus their associated electrons), at an operating temperature of 10^8 K. This plasma is fed continuously with cold fuel, possibly in the form of D-T droplets or pellets, and is pumped out at the same average rate from the reaction chamber. The pressure of this low-density but high-temperature plasma is a few atmospheres, and is supported by a magnetic field threading the entire region approximately 10 teslas (100 000 gauss) in intensity. Between the plasma surface and the wall is a vacuum region providing thermal insulation of the hot gas from the relatively low-temperature wall. The first

FIGURE 8. Cross section of hypothetical fusion reactor.



wall is a thin, cooled structure for intercepting the electromagnetic radiation load and alleviating the cooling problem of the vacuum wall. The radiation barrier has vacuum on both sides and therefore need not have high strength. The main vacuum wall isolates the reacting region from the surroundings to maintain the purity of the low-density plasma.

Beyond the wall lies the blanket, which performs two functions: (1) the energy of the 14-MeV neutrons is converted to heat and (2) tritium is bred by neutron absorption in lithium. The shield absorbs almost all of the neutrons to prevent their absorption in the magnet structure outside.

The external magnetic field coils will be superconducting and therefore immersed in a cryogenic medium, presumably liquid or gaseous helium at a temperature in the range of 4–10 K.

Problems. A great many problems are associated with each of the functions just described. The plasma presents four: confinement, heating, fuel injection, and exhaust collection. The first two have received by far the most substantial attention since the beginning of the fusion research programs. This has been very appropriate, since the primary question is the possibility of such machines—a question that is yet to be firmly answered. It will be indicated later that there is good reason to anticipate such a proof within a few years. Therefore, a significant number of people are now turning their attention to the problems of fuel injection and exhaust collection.

Fuel injection is not a simple matter. One would hope to adopt as simple a system as possible, such as a high-velocity droplet injector. Such equipment exists, but no experiments have been made to determine whether or not droplets (or pellets) can penetrate deeply into a high-temperature, high-density plasma. Calculations based on the simple energetics of a droplet encountering a plasma indicate that so much energy would be intercepted that the droplet would be evaporated, ionized, and therefore immobilized by the magnetic field, essentially at the surface of the plasma region. In a different model, a dielectric pellet at high velocity becomes polarized in the transverse magnetic field and thereby generates a shielding magnetic field that prevents interaction with the hot plasma. This might predict complete penetration, with the droplet leaving the far side—an even less satisfactory result. Perhaps experiment will show the truth to lie somewhere between, but today it is an open question. Failure of neutral macroparticle injection would require the application of techniques of energetic neutral beams (a technology developing well in laboratories studying mirror technology) or cyclic operation with injection of a fuel charge to initiate the cycle.

Pumping the throughput of fuel is not straightforward. In equilibrium at optimum condition, a D-T reactor burns only 2 percent of the fuel charge in one cycle through the plasma, and a new type of pump is required. Such a device is called a divertor and is one of the most important features of a fusion reactor. This has been tested in experimental devices with low-energy plasmas, but, of course, not yet with reactor-grade plasmas. Figure 9 shows a magnetic field plot of a type of divertor that has been operated on a research machine. The outer layer of the plasma is stripped off and delivered to an external vacuum chamber of higher base pressure where vacuum pumps of relatively conventional design may be used to concentrate the exhaust.

The method by which this is done is easily understood from the figure. The coil in the transverse plane of symmetry of the divertor has a current direction opposite to that of the toroidal field coils. The magnetic field resulting has a zero at the point indicated. Therefore, that part of the magnetic flux (such as typical field line number 1) lying beyond this is removed from the main chamber into the divertor chamber. Field lines within (such as number 3) are not diverted. The separatrix (line 2) is the boundary between the diverted and undiverted flux. In this experimental divertor the diverted flux carried its confined plasma to a small region of the divertor chamber's wall. In a reactor, the exiting plasma would have a large thermal energy content, and field shaping would be done to spread the particle flux over a larger area to allow adequate cooling.

Second only to the basic problem of plasma confinement, the problem of the vacuum wall seems ominous. This may be the technologically most limiting feature of a reactor. The wall must be structurally sound, of long life, and, for D-T operations, not an absorber of neutrons. In fact, a neutron multiplier would be useful. It will see vacuum on one side and a high-pressure coolant on the other. The heat load from neutrons and gamma radiation must be conducted to the outer surface. The resultant temperature gradient produces high stresses that must be supported. This structural element will be expected to withstand a lifetime flux of some 10^{23} neutrons/cm^2, equivalent to the highest encountered in fission reactors, but of a spectrum of considerably higher energy. During a week of typical operation, every atom of the vacuum wall will have been displaced at least once as a result of neutron collisions. An economically reasonable lifetime is not yet assured, and research on this subject has been undertaken. Several materials have been suggested, including molybdenum, niobium, and vanadium. Molybdenum seems to be subject to excessive rates of void formation due to (n, α) reactions; niobium becomes neutron-activated, leading to an important after-heat level; vanadium is the latest choice for D-T models; all are expensive. If deuterium reactors become possible, stainless steel may be suitable, but it, too, may have excessive void formation rates.

The blanket and neutron shield have probably received more design attention than any other nonplasma fusion engineering problem. It is not surprising, therefore, that it seems much more tractable now than it did in the early days. Nevertheless, the D-T reactor does represent a very sophisticated neutron engineering problem. For every triton burned only one neutron is generated, and it takes one neutron to breed a triton. Losses are, of course, unavoidable; so neutron multipliers (such as beryllium) are required in the blanket, and absorbers should be avoided. The neutron economy, however, does not appear crucial, and fuel doubling times as short as several weeks seem possible—much shorter than would be essential. Another way to describe this advantage is in terms of the breeding ratio required to provide a useful fuel doubling time of, say, eight years. For the D-T system only 1.01 or 1.02 is needed for the net breeding ratio.

The heat-transfer problems within the blanket and with the associated heat exchangers for steam generation for the turbines have been studied to some extent,

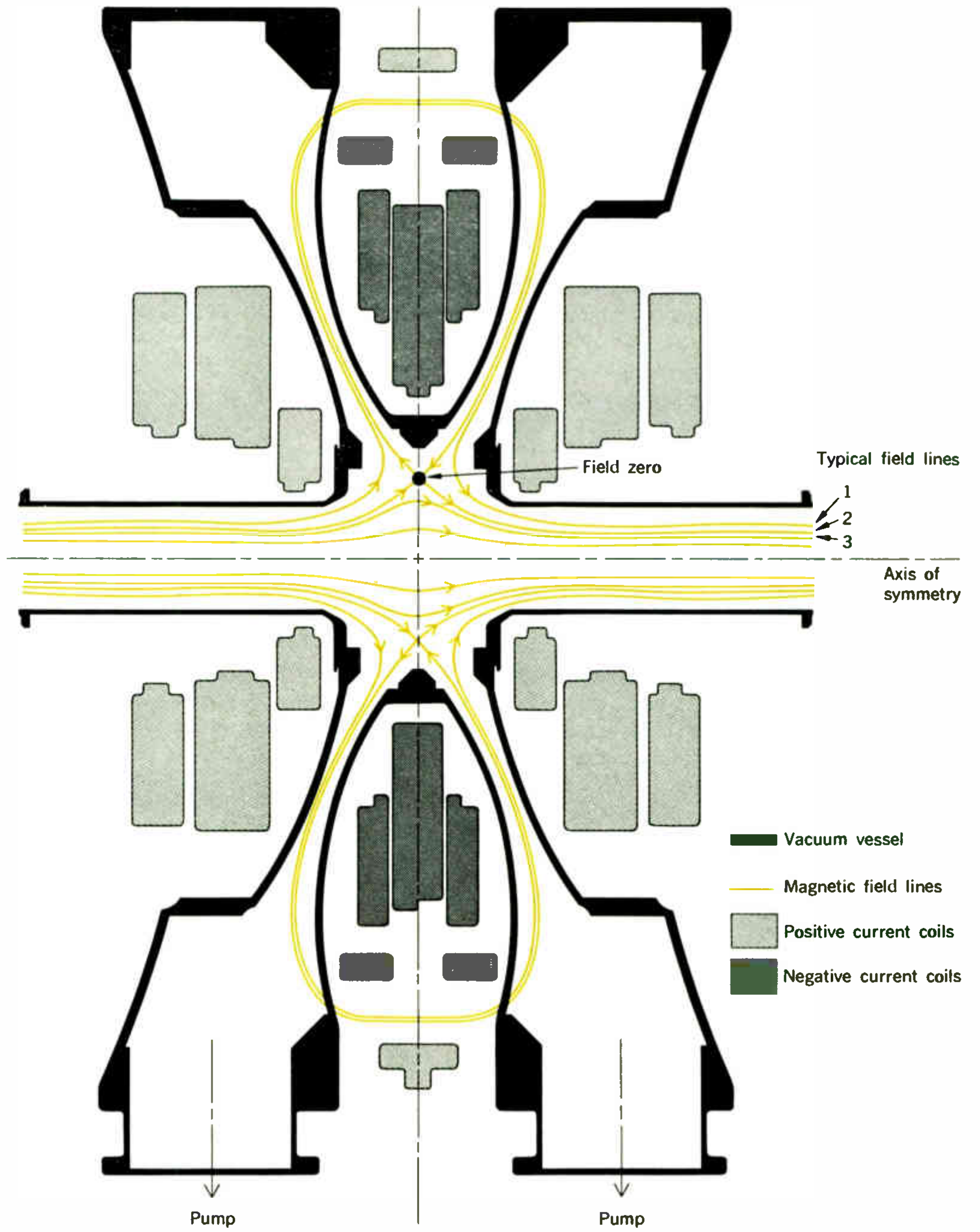


FIGURE 9. Divertor field of Model C stellarator.

and further work is desirable. Candidates for use within the blanket are molten salts such as flibe (lithium beryllium fluoride), lithium nitrates, molten metals (lithium or potassium), helium, and steam. The electrical conductivity of the liquid metals is a disadvantage, since these metals are flowing in a magnetic field. Thus, either large pumping power, electrical insulation, or both must be provided. The combination of very-high temperature, alkali metal, and intense neutron and gamma radiation poses an almost impossible task for the current state of insulation technology. In fact, the corrosion problem is a challenge throughout the blanket. Fortunately the liquid-metal fast breeder reactor program is learning much that will be needed if and when fusion blankets and shields become a reality.

Outside the shield lie the magnet coil and structure, charged with the task of generating magnetic fields of average values close to 10 teslas. To do this in a toroidal geometry, the maximum field strength adjacent to the conductors near the axis of the machine may equal or exceed 15 teslas, the highest practical level for Nb_3Sn, the most popular superconductor now available for high field magnets. New materials are becoming available (for example, V_3Ga) that are reported to be usable beyond 20 teslas, but no appreciable experience has yet been developed with these materials. Regardless of the electrical stability of the superconductors, the force problem will be a challenge to designers. A 16-tesla magnetic field behaves in many respects as if it were a gas at a pressure of 15 000 lbf/in^2 (about 10^8 N/m^2), and clever design must be done to support the coils reliably without massive supports penetrating the Dewar, which might cause unacceptably large heat leaks into the cryogenic region.

The major heat leak into the Dewar may come from the leakage of neutrons from the shield. When they are captured, each liberates a few million electronvolts of energy, and when this occurs within the cryogenic region, a serious refrigeration problem results. A challenging optimization problem occurs near the axis where the field is at its maximum. This leads the designer to the lowest possible conductor temperature and the thinnest possible shield, but both of these lead to higher refrigeration power. He will also be plagued by some magnificent problems of differential thermal expansion, since he will be dealing with a rather large structure, in which the outside is at room temperature and the next layers are near absolute zero. Moreover, beyond the shield, into the blanket, inner layers rise rapidly to the maximum temperature allowed by the metallurgical limits of the vacuum wall—to say nothing of the plasma within, which operates at a temperature much higher than that of the center of the sun!

Advantages and disadvantages. The promise of controlled fusion is for a widely distributed, inexhaustible primary fuel source of negligible cost. The subsidiary advantages include high thermal efficiency, the absence of dangerous by-products, and immunity to clandestine production of weapons-grade nuclear material. The ash of the nuclear fusion reaction is ordinary, inert helium, a useful product that may be in short supply within a few decades.

The question of safety is frequently misunderstood. We do not have choices between "safe" systems and "unsafe" systems. Every human activity is beset with a number of hazards. An important function of technology is to reduce hazards to a practical minimum that is small, but never zero. The ethics of the engineering profession require that a product or service offered to or affecting the public have a hazard potential so small that few knowledgeable and rational people feel concern over its use. The quantitative problem facing the engineer is to optimize the cost of the measures necessary to make a system safe.

The future needs for electric power for mankind for all but the immediate future must be met by either breeder fission reactors or nuclear fusion reactors. In all probability both systems will prove useful, and both will be safe; but there may be a cost advantage for the fusion system because it will be inherently so docile. Whereas a fission plant contains sufficient fuel in its core for many months of operation, the tenuous plasma of a fusion reactor contains enough fuel for only a few seconds of operation at most, and there is no chance for a nuclear excursion with damage to the plant. Accidental failure of magnetic confinement will dump the plasma on the wall, cooling the plasma immediately, quenching the reaction, and causing a wall temperature rise of less than 100°C.

The fusion machine itself will be a strong source of radiation, requiring careful shielding, but there will be no radioactive output until the day the plant becomes obsolete or the vacuum tube must be replaced. For good economy this should be not more often than every 20 years, but when the time comes, the inner parts must be removed, compressed, and safely stored until the induced radioactivity dies away.

The plant will contain an inventory of tritium, a hazardous radioactive gas, but engineering studies show that it will occur at high concentration only in a very small region of the plant that will be easily protected at a relatively low cost and thereby avoid hazard to plant operating personnel.

Local problems of thermal pollution will remain with man as long as he burns fuel. There is a small but significant advantage to fusion reactors in that a portion of their output is amenable to direct conversion to electricity. However, a substantial quantity of the energy (that portion carried by neutrons, bremsstrahlung, and synchrotron radiation) must go through a thermal cycle, and this portion is subject to the metallurgical limitations of available materials just as is any heat engine. Overall plant thermal efficiency should lie between 45 and 60 percent.

The principal disadvantage of fusion power is its tendency to come in large packages. It is indeed unfortunate that no one has yet come forward with a scheme to generate fusion power without simultaneously generating high-energy neutrons. It is the range of neutrons in matter that forces one to adopt blanket/shield thicknesses of 1 or 2 meters. It would be economically cavalier to use such a large, expensive blanket on a small-diameter, low-power plasma. The plasma dimensions tend to approach or exceed the blanket dimensions, and the net electric power output becomes at least 1000 MW.

Present status of controlled thermonuclear research

Plasma diffusion in a magnetic field. The concept of a controlled thermonuclear reactor involves a magnetically confined, high-temperature, reacting plasma with its density at a maximum at the center but falling to essentially zero at the boundary where the plasma encounters material of low temperature. This is not an equilibrium situation, and a steady state can result only if there is a source of material at the center (produced by fuel injection) and a continual removal at the edge. A reactor can be successful only if transport of the material from the center to the edge is sufficiently slow. Since there is a density gradient, this transport process should be governed by the law of diffusion, Fick's law, which states that the current density of particles, $j(cm^{-2}\cdot s^{-1})$, is given by

$$j = -D\nabla n \qquad (17)$$

where ∇n is the gradient of the particle density and D is the diffusion coefficient $(cm^2\cdot s^{-1})$.

In ordinary neutral gases, the diffusion coefficient is given by

$$D = \frac{1}{3}\lambda v \qquad (18)$$

where λ is the mean free path between collisions and v is the average velocity of the particles. For a gas of charged particles in a magnetic field, however, the situation is quite different.[3] Diffusion in a neutral gas may be viewed as a random walk process, where collisions produce random new directions of motion of step length λ

between each collision. In the presence of a magnetic field, however, the charged particles do *not* travel a step length λ in a straight line between collisions. Magnetic forces confine them to a helical motion, with a gyroradius r_g, about a magnetic field line. In a homogeneous, uniform magnetic field, the cross-field diffusion coefficient, $D_\perp$, may be sharply reduced from the ordinary diffusion coefficient if the field strength is high enough to make $r_g \ll \lambda$:

$$D_\perp = \frac{D}{1 + (\lambda^2/r_g^2)} \tag{19}$$

It is this effect that makes magnetic confinement possible, and the classical diffusion coefficient given in Eq. (19) predicts extremely slow diffusion rates, orders of magnitude slower than are necessary for a reactor.

The real world is somewhat more complicated than the idealized situation of a uniform magnetic field, and (19) usually does not apply. Instead of the gyroradius of a particle about a field line, we must consider the entire path of a particle throughout the confinement device. If this device is toroidal, there is a field gradient that produces drifts of the orbit across the field, and the poloidal field, described earlier, produces an elongated and relatively large orbit before the path again closes on itself. These paths (often called "bananas" because of the shape of the projection of the orbit on a cross section) lead to a substantial increase in the rate of diffusion.[4] In effect, the larger width of the banana tends to replace r_g in (19). This results in faster diffusion.

The fundamental research problem in plasma physics studies oriented toward fusion power is to learn the laws governing the diffusion of a magnetically confined plasma in a real situation, a joint theoretical and experimental task. This study has its theoretical roots in astrophysics since almost all of the matter in the universe is in the plasma state and penetrated by the weak interstellar magnetic field. The experimental roots lie in work done in gas discharges such as those in mercury-arc rectifiers similar to those Irving Langmuir was studying in the twenties when he gave the name plasma to this interesting form of matter, and in radar-switching devices that the physicist Bohm was studying in the forties when he proposed the semiempirical "Bohm diffusion," predicting rates of diffusion independent of density, proportional to electron temperature, and inversely proportional to the magnetic field. This result was in striking contrast to theory, which predicted that the rates should increase with the density, should vary inversely as the square of the magnetic field, and, most important, should vary inversely as the square root of the temperature rather than rising linearly with it, as predicted by Bohm. Classical diffusion was far slower than needed for a practical reactor. Bohm diffusion was so fast that a reactor big enough to break even energetically would produce many thousands of megawatts. Since this would be an impractically large block of power, a crucial question was whether or not Bohm diffusion was a law of nature. During the fifties and sixties various simple magnetic geometries demonstrated confinement times well in excess of Bohm values, but all toroidal systems seemed disappointingly close to Bohm performance, roughly two orders of magnitude short of that needed for practical reactors.

The Russian Tokamak. In 1968–69 dramatic results were achieved when the Russians announced kiloelectronvolt plasmas of substantial density ($\sim 10^{13}$ cm^{-3}) and, most important of all, confinement times in excess of 10 ms, some hundred times "greater than Bohm." These results were received with some skepticism in the West largely on the basis that the reported observations, explained by a high-temperature Maxwellian plasma, could equally well be explained by a two-component plasma consisting of a low-temperature main body in which was imbedded a small group of "runaway" or high-energy electrons accelerated by the electric field used for ohmically heating the plasma. To choose between these interpretations, one can determine the electron velocity distribution in the plasma directly by measuring the Doppler-shifted frequency spectrum of light scattered by the electrons. This procedure has become practical with the development of lasers. In 1969 such instrumentation was not available in the Moscow laboratory, but the English had suitable equipment and offered to take it to Moscow to measure the electron velocity distribution. The Russians welcomed this assistance, and in September the dramatic results were announced. The plasma temperature was indeed high, in fact somewhat higher than had been originally claimed.

This stimulated great interest throughout the world in Tokamak devices, and the fabrication of several was undertaken during the winter of 1969–70. The first Western Tokamak went into operation in May 1970 at Princeton, that laboratory having modified its Model C stellarator into a Tokamak by removing the straight section and moving the two ends together. The resulting device is very close in dimensions to the T-3 Tokamak in Moscow. Experiments with it gave results very similar to those in T-3, and subsequent measurements have carried the detailed knowledge of the plasma characteristics considerably further.

Significance of results. The excitement over these latest toroidal results stems not merely from a large increase in the observed plasma confinement time but, more important, from the fact that conditions are now improving with temperature rather than deteriorating. It appears that a diffusion process is operating. In these circumstances the confinement time τ will vary as

$$\tau \approx \frac{a^2}{D} \tag{20}$$

where a is the plasma radius and D is the diffusion coefficient. In present toroidal plasmas, the electron temperature, ion temperature, density, and confinement time are, respectively,

$$T_e = 1500 \text{ eV}$$
$$T_i = 500 \text{ eV}$$
$$n_i = 5 \times 10^{13} \text{ to } 10^{14} \text{ cm}^{-3}$$
$$\tau = 0.01 \text{ second}$$

These are observed in small laboratory devices with the plasma radius between 10 and 20 cm. The diffusion coefficients measured are already small enough to be useful for reactors, systems in which the plasma radius a must be an order of magnitude higher.

Possible pitfalls. The $n_i\tau$ product of existing experi-

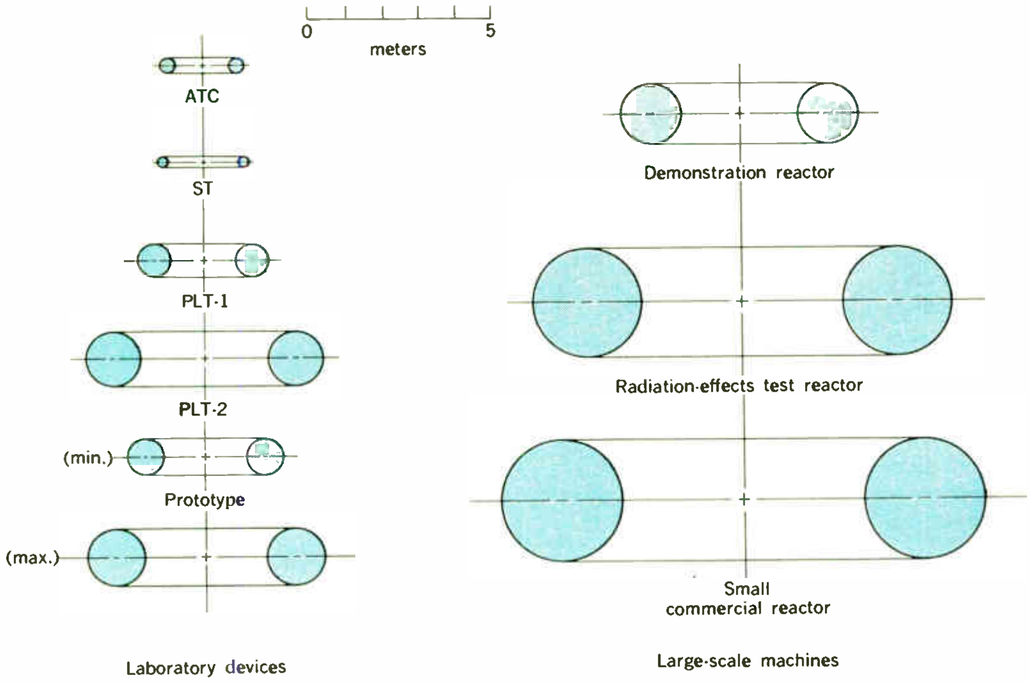


FIGURE 10. Comparative plasma cross-section sizes in various closed geometries.

ments is about 10^{12} s/cm³, whereas the Lawson criterion requires 10^{14} s/cm³ to prove reactors possible. The previously presented scaling argument gives reason for optimism that a radius ten times larger would provide this, but to prove scientific feasibility these conditions must be achieved at reactor ion temperature, another factor of ten above existing conditions. No one can give assurance that the threshold of some new unanticipated instability may not be crossed as temperatures go up. The boundary of the plasma must lie within the vacuum wall; therefore, the radial gradients of both temperature and density will increase as reactor conditions are approached, and these are known sources of drift waves that could interfere with confinement. Experience will tell.

Keeping up with this rapidly developing field is difficult for the nonspecialist, since the current literature, especially in plasma physics, is highly technical. However the engineering literature oriented toward reactor problems is growing, and those interested in learning more of this fascinating field will find much to their liking. A good starting point would be the Proceedings of the Culham Conference, which was sponsored by the British Nuclear Energy Society in 1969.[5]

Future events

Scientific feasibility and the Lawson criterion. An oversimplified explanation for the excellent confinement being observed now is that it is a natural result of operating larger research machines. As the machine size grows, the confinement time grows, and as this time grows, higher temperatures are achieved from heating methods of limited power. When this fact is linked to the observation that plasma behavior is more favorable at higher temperatures, the importance of adequately sized machines becomes clear.

Larger machines, of course, cost more, and it is significant that several conceptual designs for a variety of devices aimed at demonstrating scientific feasibility of fusion power all lead to estimated costs in the $25–30-million range, an amount equal to the entire national annual budget distributed to all laboratories investigating this important question.

What is an adequate demonstration of scientific feasibility? To most people working in the field it means a 10-keV plasma at a density of 10^{14} ions/cm³ confined in any device for a period of 1 second. To others, ignition conditions seem adequate: a temperature of about 5 keV and an $n_i\tau$ product approaching 10^{14} sec/cm³. (Remember that the Lawson criterion depends on the efficiency of the associated power plant, and there is no absolute lower limit on $n_i\tau$.) Still others insist that, to be absolutely convincing, the device demonstrating the foregoing properties must be fueled with a D-T fuel mixture in order to release nuclear energy at a rate that is actually faster than the power required to operate the machine.

Plasma physicists expect to demonstrate scientific feasibility in more than one device during the seventies. Having accomplished that, what will be the subsequent intermediate steps before commercial power will be a reality?

A possible course of events. From the point of view of toroidal confinement, Fig. 10 compares plasma sizes of laboratory devices with those of the larger machines required to reach the commercial power level. The ST plasma is in the existing Princeton Tokamak. The ATC is a new machine in fabrication now to go into operation in the spring of 1972. The PLT-1 machine has a preliminary design available, but is not yet authorized. It is designed for the largest plasma that can be provided by the existing 200 000 kW of motor–generator capacity at the Princeton laboratory, shown in Fig. 11, and thus would be the quickest and most economical path to a large toroidal plasma.

It is conceivable, but unlikely, that scientific feasibility could be shown with this machine. It is the PLT-2 machine that is planned to settle definitely the question of scientific feasibility in closed geometries. Its design criterion is to make it as large as would be necessary under the most pessimistic extrapolation of current knowledge to demonstrate feasibility.

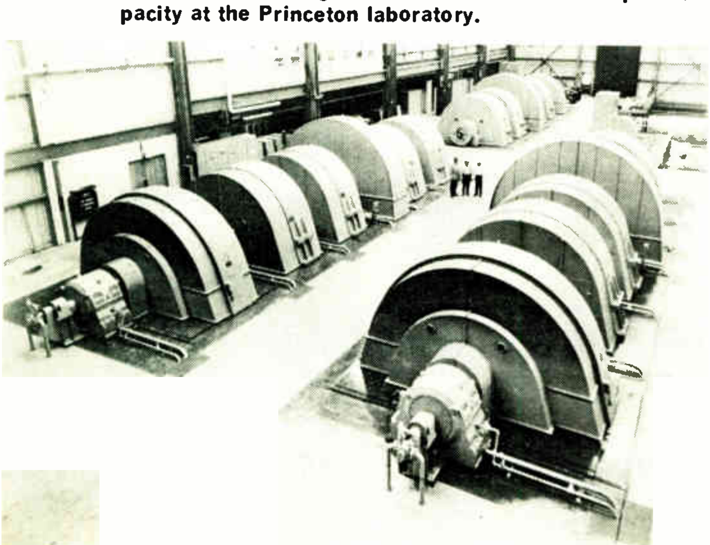
The course of events in the laboratory beyond PLT-2 cannot be predicted accurately now. The PLT could serve as a large test bed for further plasma physics experiments and development to complement power development work at larger sites. Conceivably, it could be converted to a prototype reactor. Or a prototype reactor could be built as a separate device, as shown in Fig. 10, its size depending on the experimental results of the preceding machines. This device would burn deuterium, tritium, and He^3 fuel combinations but it would not breed the necessary isotopes, which are available from fission plants.

Moving on to large-size devices, a demonstration plant would generate a small amount of power, say 100 MW, and would be self-sufficient, breeding its own fuel. However, this type of plant would be too small to be economic and also would be intermittent and developmental in its operations.

The important questions of radiation damage and blanket and shield design cannot be fully answered until the potent neutron fluxes of the D-T reactor become available. A radiation test facility should be designed and fabricated as soon as it has been proved possible, and confidence has been developed that it can be operated. Five full years of successful high-flux operation should be accumulated before the large sums needed for a commercial reactor should be committed.

The schedule for such a program can only be guessed at, but with a broad brush: feasibility in the seventies, demonstration in the eighties, and commercial power in the nineties. The cost? An average of $100 million a year until we are ready for the new industry.

FIGURE 11. Motor–generator sets of 200 000-kW peak capacity at the Princeton laboratory.



If magnetic confinement fails. Lest we forget, it has not yet been proved that a controlled thermonuclear reactor is possible. If closed geometries fail, mirrors may succeed. If mirrors fail, too, perhaps pulsed devices or the Astron will be possible. If all magnetic confinement fails, laser-ignited microbombs may carry the day, or even mini-bombs in underground cavities. If none of these schemes is economically feasible, then fission breeder reactors will have the full responsibility for fueling the future of mankind.

Closing on this cautionary note, however, should not mask the fact that today, in contrast to the situation a few years ago, a majority of scientists and engineers knowledgeable in the field of controlled thermonuclear research believe that fusion power will be possible and will become practical in this century.

REFERENCES

1. Lawson, J. D., "Some criteria for a power producing thermonuclear reactor," *Proc. Phys. Soc. (London)*, vol. B70, pp. 6–10, 1957.
2. Mills, R. G., "Lawson criteria," *IEEE Trans. Nuclear Science*, vol. NS-18, pp. 205–207, Aug. 1971.
3. Glasstone, S., and Lovberg, R. H., *Controlled Thermonuclear Reactions*. Princeton, N.J.: Van Nostrand, 1960, pp. 451 et seq.
4. Kadomtsev, B. B., and Pogutse, O. P., "Trapped particles in toroidal magnetic systems," *Nucl. Fusion*, vol. 11, pp. 67–92, 1971.
5. *Proc. Nucl. Fusion Reactor Conf.*, British Nuclear Energy Society, London, 1969, CLM-MFR.

Reprints of this article (No. X71-111) are available to readers. Please use the order form on page 8, which gives information and prices.

Robert G. Mills (SM) is senior staff member and head of the Engineering and Development Division of Princeton University's Plasma Physics Laboratory. He has been associated with the controlled thermonuclear program at Princeton since 1954 and is now in charge of the newly organized Reactor Studies Program, which is undertaking a comprehensive analysis of the technology of toroidal fusion power reactors.

After obtaining the bachelor's degree in electrical engineering at Princeton in 1944, he served in the U.S. Navy, during which time he was assigned to the Naval Research Laboratory, where he assisted in the development of specialized communications equipment. He received the M.A. degree in mathematics from the University of Michigan in 1947 and the Ph.D. degree in nuclear physics in 1952 from the University of California, Berkeley, where he served as a research associate at the Lawrence Radiation Laboratory. A National Research Fellow (1952), Dr. Mills taught and conducted research at the University of Zurich from 1952 to 1954. He is the author of numerous papers on the engineering aspects of thermonuclear research. He is a member of the American Physical Society, the American Nuclear Society, Phi Beta Kappa, and Sigma Xi. He is chairman of the IEEE Technical Activities Board Finance Committee and a past chairman of the IEEE Nuclear Science Group.

The present status of engineering employment

A benchmark survey of engineering employment shows the picture to be brighter than supposed—but still sobering. Results confirm many past conclusions concerning the geographic nature of the problem and highlight the degree of severity existing in mid-1971. Overall unemployment for U.S. engineering professionals who are members of engineering societies is 3.0 percent. One significant fact to emerge is that electronics engineers have an unemployment rate equivalent to that of engineers in the aerospace field; both are the highest among all engineering disciplines —more than 5 percent.

The specter of joblessness facing many engineers in the U.S. during the current economic decline can now be given some form and substance as the results of the long-awaited first nationwide survey of engineering employment are made available and can be evaluated. The survey, conducted by the Engineers Joint Council for the National Science Foundation* and completed in August, sampled some 59 000 engineers—all of whom were members of engineering societies. This figure represents a 60 percent response to a query sent to 20 percent of the 500 000 individual members of 23 major engineering societies, representing all fields of engineering employment. A follow-up query sent to the 40 percent nonrespondents produced a return with the same results obtained from the original responders.

The most notable and the first general result obtained by NSF showed that the unemployment rate for engineers was running at 3.0 percent. The immediate tendency is to conclude that the national unemployment situation for engineers is not as severe as had been surmised. This conclusion undoubtedly has some validity, and considering that the general population has an unemployment rate over 6 percent, one could argue that engineers, as a group, are faring better than the general population. However, this conclusion may be misleading, because if we take into consideration the fact that in the recent past the engineering unemployment rate has been less than 1 percent, it can be argued that, with the same incremental rise in unemployment for engineers as a group as that for the general population, the engineers are faring as poorly as the general population; indeed, as some of the tables show, in some geographical areas they are much worse off.† Moreover, the EJC believes that the NSF study is erroneous in that engineers working part time or doing nonengineering work should be classified as completely unemployed rather than employed. On the basis of this reasoning, EJC has produced a second set of figures that reflect this rearrangement in categories. Table I and others in this article give both the results as reported by NSF and those developed by EJC.

In addition, data have been extracted and compiled showing the rates as a function of field of specialization, geographical area, age, and degree. From these we can arrive at some general conclusions that should be of help in planning for the future and in concentrating remedial efforts in those areas and disciplines of greatest need.

* Reported in NSF Bulletin 71-33, *Science Resources Studies Highlights.*

† The basis for the assertion that employment of engineers is at least equivalent to that of the general population lies in the fact that in previous EJC surveys (1964–69 period) on behalf of the National Engineers Register less than 1 percent of the respondents reported themselves out of work. If one accepts the economic dictum that a "full employment" economy is one which maintains a 3 percent overall unemployment rate, then the incremental change for engineers in general (between 1 and 3 percent) roughly equals that for the general population (3 to 6 percent). Thus, engineers are faring at least as poorly as the general population.

I. Engineering employment and unemployment rates as computed by the NSF and EJC

Employment Status	Computed by NSF: Number	Computed by NSF: Percent	Computed by EJC: Number	Computed by EJC: Percent
Total survey respondents	59 200		59 200	
Not employed; not seeking employment	3 500		3 500	
In labor force	55 800	100.0		
Employed in engineering work	50 400	90.3		
Employed in nonengineering work	3 700	6.7		
Unemployed and seeking employment	1 700	3.0		
Employed in nonengineering work by choice			3 200	
Employed part time in engineering by choice			500	
Total in or seeking full-time engineering work			52 000	100.0
Employed full time in engineering work			49 500	95.3
Not employed full time in engineering work			2 500	4.7
Employed part time in engineering, seeking full-time employment			300	0.6
Employed in nonengineering work, engineering not available			500	0.9
Unemployed and seeking employment			1 700	3.2

II. Unemployment rates by field of specialization

Field of Specialization	Number in Labor Force	NSF Unemployment Rate	EJC Employment Problem Rate
Aerospace engineering	3861	5.3	7.6
Chemical engineering	2072	1.9	2.8
Civil engineering	5626	1.2	1.9
Communications*	1398	2.9	4.6
Computer/mathematics*	1293	3.7	6.5
Electrical engineering*	4769	2.2	3.6
Electronics engineering*	4262	5.3	7.7
Environmental/sanitary engineering	1089	1.6	2.5
Industrial engineering	1972	2.8	5.2
Management/business administration	3091	3.0	4.8
Manufacturing engineering	2751	4.5	7.0
Mechanical engineering	5232	2.8	3.9
Metallurgical engineering	1797	2.8	4.5
Petroleum engineering	1149	0.7	2.0
Plant/facilities engineering	1406	2.3	3.2
Product engineering	1343	3.1	4.9
Systems engineering*	1610	4.1	6.3
Nonengineering	1167	4.5	20.7
Specialty not reported	1617	4.9	13.4

* Major disciplines covered by IEEE.

Cable television has had a phenomenal growth record both in the United States and in Canada. It has been predicted[1] that by 1975 a minimum of 25 percent of all U.S. homes will be wired with 20 or more channels and by the late 1970s that figure may well rise to at least 60 percent.

The rapid growth of the cable television industry to date has taken place for two principal reasons: people living in rural communities have been able to get good television reception via cable that they could not get without it, and people in urban areas have been able to get better-quality pictures with cable than they were able to get by direct reception of broadcast signals.

Now, as the '70s move ahead, there is a new impetus to further growth in cable television systems. This impetus stems from the demonstrated engineering capability to bring many more channels and new communication services to subscribers while, at the same time, giving them the facility to "talk back" to the system and, in some cases, to all other subscribers on the system. This two-way or bidirectional capability opens up a whole new realm of application possibilities. The implications of two-way cable facilities are so important to the overall communications policies of the United States that government agencies and other policy-proposing groups are taking a close look at this emerging new technology.

Many responsible people, realizing that cable television with 20, 30, 40, or more channels and two-way facilities for visual, audio, and data communications is feasible right now, are concerned with putting this tremendous capability to the best possible use to improve the life style of all citizens. Suddenly there has evolved a need for putting into proper perspective the relative importance to the nation of its existing telephone networks, its broadcasting networks, and its cable television systems. It is a task fraught with social, economic, and political overtones, most of which are beyond the scope of this report.

Dr. Peter C. Goldmark, who retired last December as president of the Columbia Broadcasting System Laboratories, is quoted in a recent *New York Times* interview as being firmly convinced that the entire structure of broadcasting and communications is headed for a king-size shakeup. Dr. Goldmark believes, according to the *New York Times* report, in the ultimate marriage of domestic satellites and cable television. The union, he says, will outmode the monopoly of existing networks and stations and provide for a profusion of coast-to-coast and local services that cut across entertainment, culture, information, instruction, education, business, and population dispersal. He feels that just two satellites could easily provide as many as 40 or more nationwide channels.

EIA recommendations

In October 1969, the Industrial Electronics Division of the Electronic Industries Association submitted a document[2] to the Federal Communications Commission. This well-conceived report took the stand that the services to be provided by broadband communications networks in the late '70s and early '80s were of "landmark importance," of "national resource dimensions," and that development of these resources should be a national goal. The EIA suggested to the FCC that it provide a regulatory environment allowing the development of two types of broadband communication networks in the United States:

1. A video telephone system similar to the "Picturephone" system of AT&T with the ability to act as a video output terminal with limited keyboard access to computers and transmission and reception of high-speed facsimile information.

2. A broadband communications network that would be a minimum 300-MHz-bandwidth pipe to provide many information services for home, business, and government, including broadcast video, first-class mail, and educational material, with limited return bandwidth for receiving and tabulating specific requests and responses by individual users of the cables.

The EIA report stated that "the public interest will be best served for the immediate future by permitting tests of different systems or services by different entities in various cities to afford some basis in experience before any particular system or service becomes established on a widespread basis."

NAE recommendations

In June of this year, the Committee on Telecommunications of the National Academy of Engineering submitted a report to the U.S. Department of Housing and Urban Development. This comprehensive report[3] is a thought-provoking document that echoes some of the EIA suggestions but adds new ones as well. The Committee on Telecommunications believes that modern communications technology, thoughtfully applied, can help in relieving many problems besetting the cities and can upgrade the level of city life. The NAE report talks about four basic communications networks for the future:

1. The telephone network for transmitting pictures, voice, and written material between two points.

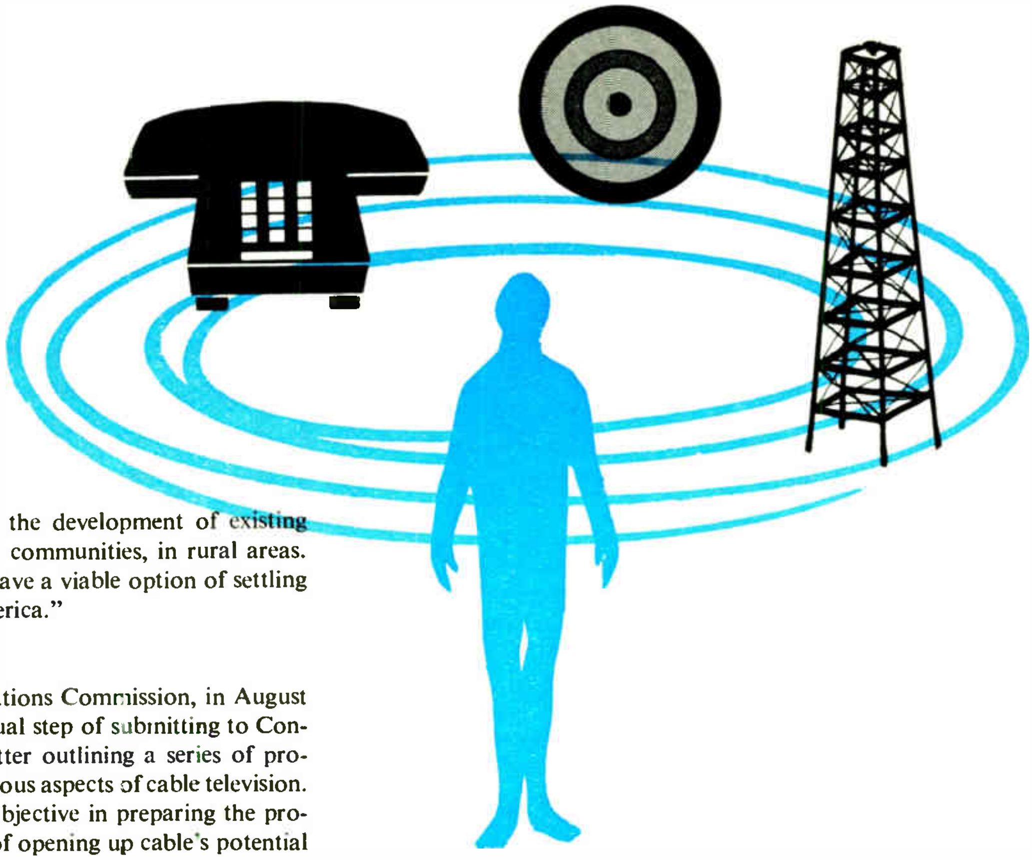
2. A network based on existing cable television systems, which can distribute information from central facilities to offices and homes, with a capacity of as many as 30 television channels and a limited call-back capacity for polling or making requests.

3. A broadband communications network carrying up to 30 equivalent television channels in both directions interconnecting major public institutions and large commercial enterprises.

4. A multipurpose city sensing network to collect data on such items as weather, pollution, traffic, vehicle location, and power status.

The NAE report concludes: "...many of the cities' problems are caused by high density living conditions in an era of increasingly rapid change. Communications technology, imaginatively applied, could offset the trend in which the vast majority of Americans today, and more in the future, live on a small percentage of the available land.

"We suggest an exploratory program to examine how broadband communications technology could be applied to business, government, education, health care, and



entertainment to stimulate the development of existing small communities, or new communities, in rural areas. As a result, people would have a viable option of settling in either urban or rural America."

FCC position

The Federal Communications Commission, in August of this year, took the unusual step of submitting to Congress a 55-page detailed letter outlining a series of proposals for regulation of various aspects of cable television. The Commission's stated objective in preparing the proposals was "to find a way of opening up cable's potential to serve the public without at the same time undermining the foundation of the existing over-the-air broadcast structure."

Citing the potential of cable television as a medium for community expression, the Commission said it would direct its efforts to "ensure the development of sufficient channel availability on all new cable systems to serve specific recognized functions." To achieve these goals, the Commission proposed that cable systems provide a channel for nonbroadcast use for each broadcast signal carried and that new systems in the so-called top 100 markets have a potential capacity for at least 20 channels. It would require, in addition, a public access channel to be available for five years at no cost for state and local government use, and a similar channel for local educational groups. "Excess" channels would have to be available for lease to all potential users, systems would be required to ensure that channels were available for use in response to demand, and two-way capacity would be required.

Williams bill

Senator Harrison A. Williams, Jr., on August 4, 1971, introduced a bill to the U.S. Senate (S.2427) to amend the Communications Act of 1934 to provide for the regulation of community antenna television systems. The Declaration of Policy in the proposed bill reads:

"Section 1. The Congress declares that community antenna television systems provide an essential service, and that access to a full offering of diversified television communications and supplemental electronics services via such systems is not a luxury, but is an essential condition for the survival of our society; that few communities will need, or be able to support, two such systems and that a full offering of such service must be uniformly and universally available for public needs and public use."

OTP wired-city experiments

The Office of Telecommunications Policy has accepted a proposal from Malarkey, Taylor and Associates in Washington, D.C., for tasks to be performed in connection with evaluating experiments for the "wired city." The proposal, when submitted later this year, will list and evaluate candidate experiments, evaluate locations and methods of operation, examine various funding arrangements, and prepare implementation schedules for experiments that are selected for further consideration.

The Malarkey, Taylor proposal emphasizes that FCC regulatory restrictions make the private entrepreneur reluctant to invest his money in the development of cable television in the major markets. The problem, therefore, according to the proposal, "is to determine whether there may not be some set of commercial services that, in the aggregate and when added to the allowable off-the-air television service, will make the coaxial cable wiring of the large cities an attractive enterprise for private investors. The wired-city potential can then be developed and the overriding goal of alleviating urgent urban problems and improving the life-style in urban areas via the use of modern telecommunications concepts can be pursued through normal profit motivated commercial process."

It is encouraging to realize that responsible individuals and organizations are so deeply concerned about the future of cable television at a time when the general public is making few, if any, demands for additional services beyond reception of commercial television programs. It may just be that, by sheer coincidence, the market will not develop until a well-defined national policy has evolved.

Basic two-way systems

When it began to appear desirable to provide more than the standard 12 VHF channels—either for additional one-way programming or for additional programming plus return-signal capability—various techniques were proposed. All are variations of a few basic schemes: use of UHF channels, frequency-division multiplexing of a single cable, dual cables, and switched multiple-cable networks either with or without multiplexing. Each of these possibilities will be described but first some discussion of distortion problems and frequency allocations is appropriate since both become so much more troublesome in multichannel two-way systems.

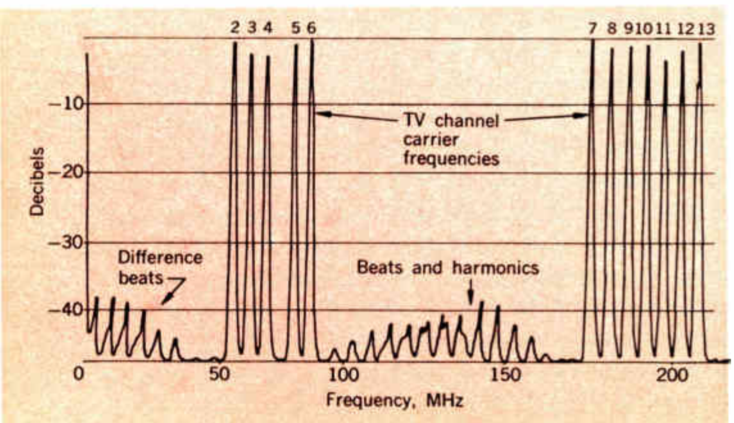
Noise and distortion

Noise and various forms of distortion are critical factors in any CATV system. Thermal noise in the system determines the lowest signal levels that can be allowed without producing snowy pictures. Distortion products determine the maximum signal levels that can be tolerated without producing "windshield wiper" effects and "herringbone" patterns in the picture. The optimum CATV system operating level is usually set at the midpoint between the two[4] even though the individual amplifier gain resulting in the longest usable system is less (8.69 dB).[5]

The present state of the art in device technology limits the maximum number of amplifiers in cascade in a cable television system. In the United States this maximum is 30 to 35 amplifiers. Increasing the number of amplifiers lowers the quality of the picture on a subscriber's television receiver beyond acceptable limits. Theoretically, an improvement in transistors that would raise their output capabilities by only 3 dB or reduce the distortion products they produce by only 6 dB would permit twice as many amplifiers to be used. Feedback techniques are not effective for reducing the more troublesome distortion products.

Distortion in CATV amplifiers is caused by nonlinearities in the transfer characteristics of the transistors used in the amplifiers. Distortion takes the form of cross modulation, where one frequency is modulated by the modulation of another frequency; of harmonics, where unwanted signals are generated at frequencies that are multiples of the desired frequencies; and of beats, where two or more desired signals combine to create an interfering signal.

FIGURE 1. Frequency spectrum obtained when 12 CW signals on carrier frequencies for channels 2–13 are introduced into a CATV amplifier operating at levels higher than normal.



Second-order distortion products that are either sums, differences, or second harmonics of the television channel carrier frequencies are not a problem with 12-channel systems. The standard carrier-frequency assignments established by the FCC took these distortion products into account. As shown in Fig. 1 (see Ref. 4, p. 27), the sums or differences do not fall within the frequency bands of any of the 12 VHF channels. The second harmonics of all the low-band carriers (channels 2–6) fall between the carrier frequencies of channels 6 and 7. When more than 12 channels are being considered, however, these same distortion products become troublesome. The severity of the problem depends on the particular frequency-assignment scheme being used.

The spurious signals caused by third-order distortion can give trouble in any multichannel system since they can fall within some of the channels. In 12-channel systems, cross modulation has usually been the determining factor for maximum output levels of the amplifiers. But in systems with more than 12 channels, third-order intermodulation products can limit the system sooner than cross modulation.

Samuel Colodny of AEL Communications Corporation has done some interesting research into the question of third-order products in systems with more than 12 channels. He has shown, for example, that there are 13 ways of getting the carrier frequency of channel 2 through third-order products in a standard 12-channel system. This figure rises to 71 for 26-channel systems, using the popular frequency-assignment scheme of Table I.[6] Channels G, H, and I (Table I) for this same 26-channel system have their carrier frequencies produced in 165 ways by third-order products. If seven midband channels (Table I) are converted as a block to channels 7–13, and if inverted superband channels are converted as an alternate block to channels 7–13, then the total of the third-order products that duplicate the carrier frequency of channel 2 is 35 and the worst channels have 93 such products.

Frequency assignments

From the previous example it is easy to see why frequency assignments in more-than-12-channel systems are so important—so important, in fact, that they are being looked at carefully by the Frequency Allocation Subcommittee of the IEEE's Coordinating Committee for Cable Communication Systems, which will issue a report later this year. Dr. Robert S. Powers, chairman of the subcommittee, has planned the report to cover the following engineering considerations:

1. Systems containing the 12 standard VHF channels and up to 14 additional channels arranged in several

alternate ways. Additional channels will be converted to "standard" frequencies by use of converters, which may or may not invert each of the channels in the conversion process. The study of standard 41–47-MHz IF and higher IFs to avoid oscillator leakage and image problems will be included.

2. A system in which carriers starting at approximately 50 MHz up to 350–400 MHz will be phase-locked to each other. Carriers will be 6 MHz apart in most cases but the effect of making an exception in order to match the standard 12-channel assignments will be explored. With this technique, the effect of third-order distortion products will be to raise or lower the carrier levels of other carriers and increase cross-modulation effects slightly—a situation hopefully less serious than coherent beats.

3. An assignment technique in which all the additional channels are "right side up" and inversion techniques are not used. Cases of both contiguous assignments and offset assignments will be discussed. A study of 45.75-MHz and higher IFs to avoid oscillator leakage and image problems will be included.

4. Dual-cable systems with an A to B switch at the subscriber's receiver rather than a set-top converter.

Dr. Powers states further that the final report will "recognize the broadest aspects and consequences of the frequency assignment problem, including two-way systems and the implications of standardization or lack of it on the consumer, on manufacturers of the equipment, and on cable operators."

The IEEE Subcommittee Report will be a valuable contribution to the industry and should help clarify some of the confusion in regard to distortion and frequency assignments in more-than-12-channel systems. It should also pave the way for setting standards for a multichannel cable television receiver.

The set-top converter

Cable television got its start because it was profitable to bring good-quality commercial television signals via cable to persons living in rural areas with poor broadcast television reception. It later evolved that cable television also had a role to play in cities where high-rise buildings and other high-density living arrangements either precluded direct line-of-sight paths from television transmitting antennas to home receivers or caused reflections of the broadcast signals to appear on receiver screens as delayed ghosts. The cable solved this particular ghosting problem but, in the process, created one of its own. If the broadcast signal were strong enough it could be picked up directly by a television receiver even though no receiving antenna was connected to the receiver. The same signal, delayed slightly by the cable system, reached the receiver later than the broadcast signal. The result was that both signals appeared on the receiver screen, resulting in a "leading ghost."

One technique to counteract the direct-signal-pickup problem uses a device known as a converter. It is a self-contained unit placed in a subscriber's home on top of or near the television receiver. The cable is connected to the converter rather than to the receiver. The output of the converter feeds the receiver at some unused VHF channel position. In New York City, for example, channel 12 is used. The receiver remains tuned to channel 12 at all times and the converter is used to tune in the

I. One proposal for CATV channel assignments

Channel	Frequency Range, MHz	Channel	Frequency Range, MHz
2	54–60	7	174–180
3	60–66	8	180–186
4	66–72	9	186–192
5	76–82	10	192–198
6	82–88	11	198–204
A	120–126	12	204–210
B	126–132	13	210–216
C	132–138	J	216–222
D	138–144	K	222–228
E	144–150	L	228–234
F	150–156	M	234–240
G	156–162	N	240–246
H	162–168	O	246–252
I	168–174	P	252–258

desired VHF channels. There is no direct pickup by the receiver of the unused channel because there is no signal being broadcast on that channel and the converter is shielded carefully.

The direct-pickup problem and the resultant need for a converter could both be avoided if commercial receivers had better built-in shielding, including 75-ohm coaxial-cable input terminals. But so far the majority of receivers —at least those sold in the United States—do not have such features.

With the more-than-12-channel systems being introduced, a converter is often necessary in the subscriber's home so that the extra channels can be viewed on the subscriber's existing television receiver. Since the new converters, in most cases, operate on the same principle as a 12-channel unit, some discussion of its operation is pertinent.

Many early 12-channel converters contained a VHF tuner that accepted the cable signals, down-converted them to 45.75 MHz, and then up-converted them to the desired output channel (channel 12 in New York City in the previous example). This technique, rather than some other approach, was used because this is the traditional operating principle of a standard television tuner in a television receiver.

When the cable was multiplexed in some fashion to distribute more than 12 channels, some new problems arose. Gene Walding, writing in the July 1971 issue of *TV Communications*,[7] states that one problem is that the local oscillator of the more-than-12-channel converter radiates energy into the mid-, high-, and superband spectrum with the result that converters can "talk" to each other. Figure 2, from his paper, illustrates what can happen. Converter no. 1, tuned to channel 5, has its local oscillator operating at 123 MHz. Converter no. 2 (which may be located in the apartment next door) is tuned to channel A (Table I) which covers the range from 120 to 126 MHz. The local oscillator of converter no. 1 interferes with reception of channel A, causing serious distortion or herringbone image on the screen of receiver no. 2.

Mr. Walding's company, Oak Electro/netics, has introduced a new 26-channel converter that is said to eliminate this particular problem. As shown in the block diagram of Fig. 3, it does not down-convert to 45.75 MHz. Instead it up-converts to an IF of 330 MHz and then down-converts to the desired output channel. By so doing, the local oscillator of the converter, in the 386- to 572-MHz range, is well out of the CATV signal band.

Some other manufacturers are also using the up-conversion technique in their converters.

Use of UHF channels

At the beginning of this section on two-way-system approaches, the basic schemes for obtaining additional channels were mentioned. One technique was use of UHF channels. At first consideration, this seems to be a likely choice. Most receivers in homes in the United States have built-in capacity for reception of UHF signals because of an FCC rule requiring such capability in sets manufactured since 1966. This choice has not proved to be a popular one, however, because of a combination of increased cable losses at the higher frequencies and the unavailability of UHF tuners of good enough quality for use with cable television systems.

Single-cable techniques

Frequency multiplexing a single cable to obtain additional VHF channels has received much attention. The reason is obvious. With few exceptions, existing installations are single-cable installations. Why not add to, or retrofit, these systems to gain more channel capacity? With the added channel capacity, one or more channels can be used for return or "upstream" signals from subscribers, making possible many two-way or bidirectional applications. The single-cable principle has been used in many experimental systems, some of which will be described later.

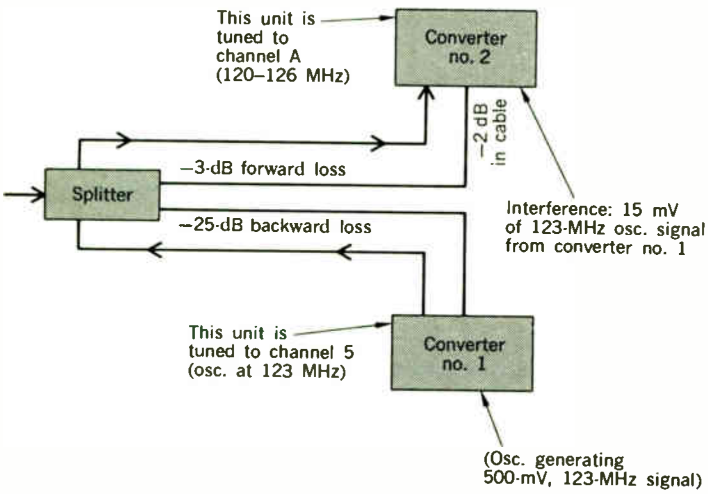
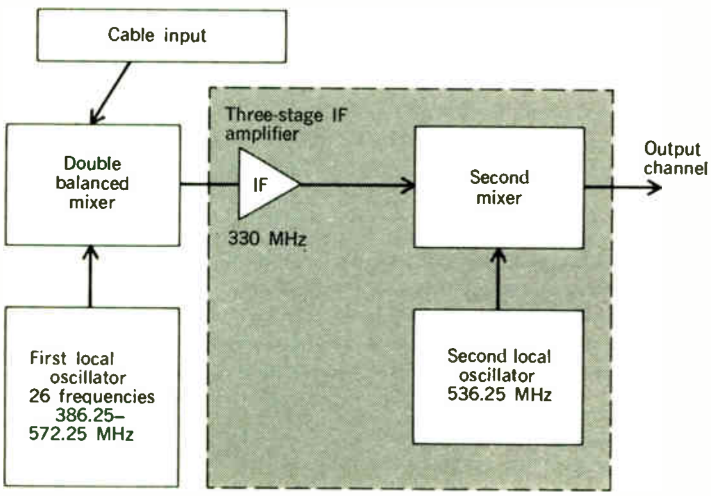


FIGURE 2. Converter-to-converter interference resulting from proximity of two converters.

FIGURE 3. Block diagram of Oak 26-channel converter.



An elementary single-cable bidirectional system would be one employing a single cable trunk and a single feeder. A trunk is the main coaxial cable in a CATV system between the head end—the equipment at the start of the system that receives, amplifies, and otherwise processes the broadcast television signals for distribution on the system—and the community or portion of it served by the system. A feeder is the coaxial cable running between bridger amplifiers or bridgers located on the trunk, the line extenders or distribution amplifiers in the feeder system, and the taps used to obtain signal voltages from the cable for subscribers.

One configuration of an elementary system is shown in Fig. 4.[8] It is called a single-cable, subsplit system because the frequency spectrum is split below channel 2, with the return information for the head end occupying the frequency band from 5 to 30 MHz, providing enough return bandwidth for one television channel (6-MHz bandwidth) and 19 MHz of data (or some other split of the bandwidth). The outgoing or "downstream" signals at channel 2 and at higher frequencies use the conventional frequency assignments. In an experimental system of this configuration, the outgoing portion used push–pull amplifiers designed for 27-channel operation: 12 standard VHF channels, nine channels in the midband region (channels A through I of Table I), and six superband channels above channel 13 (channels J through O, Table I). The feeder return and the trunk return amplifiers are also push–pull. The spectrum split is obtained with high–low filters or diplexers.

Filter requirements can present challenging design problems.[9] Passband flatness, insertion loss, return loss, and differential delay distortion are some of the parameters that must be considered. It is beyond the scope of this report to discuss filter design for CATV systems but the interested reader is referred to the published literature.[9–11]

Another version of the configuration in Fig. 4 employs a frequency split in the midband (108–174 MHz). Fourteen television channels are available for outgoing distribution (channels 7–13 and J–P, Table I) and 14 channels plus 19 MHz of data are available in the return path since it now covers the frequencies from 5 through 108 MHz. This version of the system would be a special-purpose one since only the seven high VHF channels (channels 7–13) are available for distribution to a standard television set without a converter.

Multiple-cable techniques

The simplest of the multiple-cable proposals is the addition of a second cable to make an additional 12 VHF channels available. The receiver is equipped with a simple A-B switch. Under ideal conditions, the subscriber receives 12 standard channels in position A (assuming the channels are all being used) and 12 additional channels in position B for a total of 24. In practice, the total may

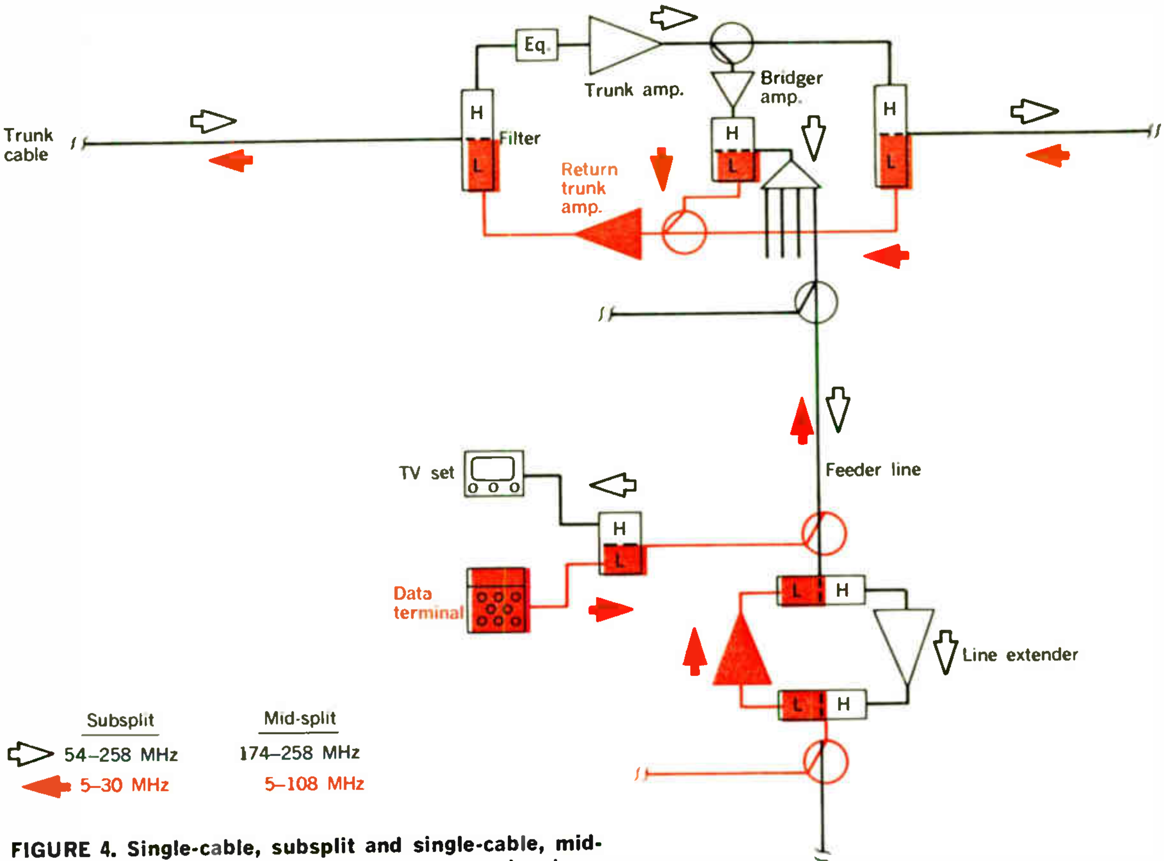


FIGURE 4. Single-cable, subsplit and single-cable, mid-split systems with upstream and downstream signals.

be less than 24.[12] Interference problems between cable A and cable B are readily controlled.

Variations of the dual-cable approach are being field-tested for two-way systems. In fact, of all approaches to increasing channel capacity with two-way feasibility, the dual-cable approach seems to be winning favor with the majority of cable system operators. Dual-cable systems may be more expensive initially than single-cable systems but they can either eliminate or reduce the amount of multiplexing necessary with the result that there are fewer complex amplifiers, filters, distortion problems, etc., to worry about. Hubert J. Schlafly, president of TelePrompTer Corporation, for example, has directed that all future installations of that company will be dual-cable installations. And TelePrompTer has already experimented with single-cable two-way installations.

Figure 5 shows an elementary dual-trunk, single-feeder system.[8] Trunk cable A is one-way only and uses the outgoing frequency spectrum from 54 through 258 MHz for carrying 28 channels (channels 2–13 and A–P). The A cable distribution is two-way with the 54–258-MHz spectrum from the A trunk outgoing, and 5–30 MHz returning from the subscriber locations to the B trunk station. The B trunk cable is the mid-split, two-way system that was shown in Fig. 4. The 5–30-MHz portion of the B return spectrum is used by the A feeder return signals, which are coupled over to the B station from a high–low split filter in the A station. The 30–108-MHz portion of the B return and the B outgoing have limited access and return.

The dual-cable system in Fig. 5 has the capability for full distribution of 28 channels outgoing and one channel plus 19-MHz return. For limited-access private service, it offers 14 channels outgoing and 13 channels return.

Another configuration is a dual-trunk, dual-feeder system.[8] The A cable is a conventional one-way CATV system. The B cable is a full two-way system with a trunk return of 5–108 MHz, a feeder return of 5–30 MHz, and outgoing trunk and feeder from 174 to 258 MHz. This configuration has all the advantages of that in Fig. 5 with the additional capacity of the B outgoing distribution available to the subscriber through use of an A-B switch. In many cases, enough channels would be available by means of the switch that converters would not be needed.

If a subscriber had an A-B switch and no converter, for the dual-trunk, dual-feeder system just described, he would be able to receive 19 channels (channels 2–13 on cable A plus channels 7–13 on cable B). If the subscriber had a multichannel converter, he would be able to receive 42 channels (channels 2–13 and A–P on cable A plus channels 7–13 and J–P on cable B). One channel plus 19-MHz return from 5 to 30 MHz would be available for public service and 13 channels from 30 to 108 MHz for private service.

The dual-cable concept can be carried further, obviously, by adding a third cable, a fourth, etc. But economic considerations tend to rule out this approach in all except highly specialized applications or in urban areas where installation labor dominates cost and the incremental cost of multiple cables is not as significant.

There are other adaptations of the multiple-cable

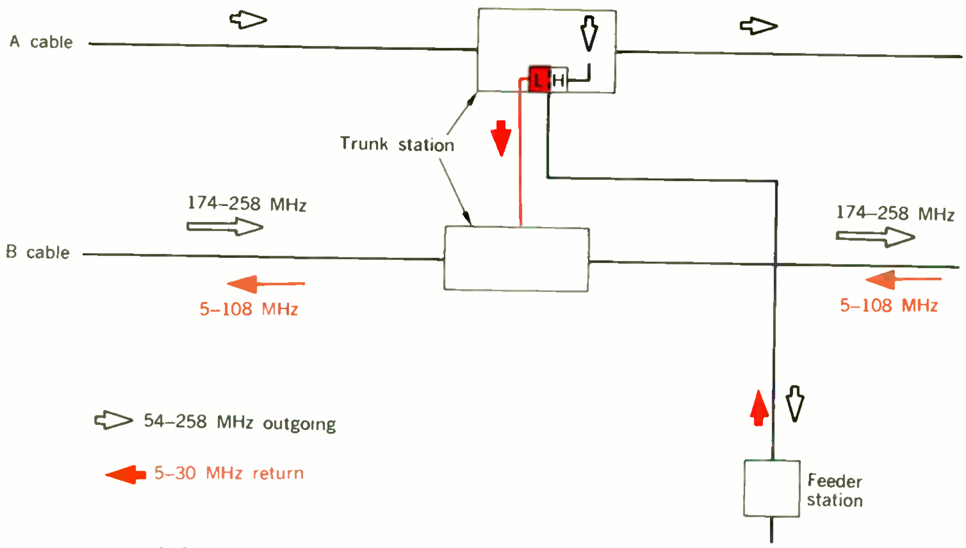


FIGURE 5. Dual-trunk, single-feeder system showing upstream and downstream signals.

approach that will be described later. One such system offers an ideal arrangement for two-way operation. It is the Rediffusion system, which uses multipaired HF cables rather than broadband, coaxial cables. Each subscriber's receiver is wired with two pairs—a control pair and a signal pair. The subscriber dials a coded number for any one of 36 channels he wishes to receive. An electromechanical rotary device with 36 reed switches in a local distribution exchange then connects the subscriber to the desired channel. In a somewhat comparable system by Ameco, known as DISCADE, coaxial cable is used together with solid-state switching. And GTE Laboratories has just proposed a system that uses the Rediffusion mechanical switching setup with coaxial cables.

Actual two-way systems

Even though no real market now exists for two-way CATV systems, many enterprising cable system operators and equipment suppliers are actively engaged in attempts to create that market through experimentation with two-way systems in the field. The remainder of this report describes some of these working systems as well as others that have been proposed but not yet implemented. The few systems described have been selected carefully to show the variety of engineering designs being used to make single-, dual-, and multiple-cable CATV systems feasible as well as to show the imaginative applications of these systems that are being demonstrated. Space limitations preclude descriptions of all two-way systems proposed or in existence.

Vicom system

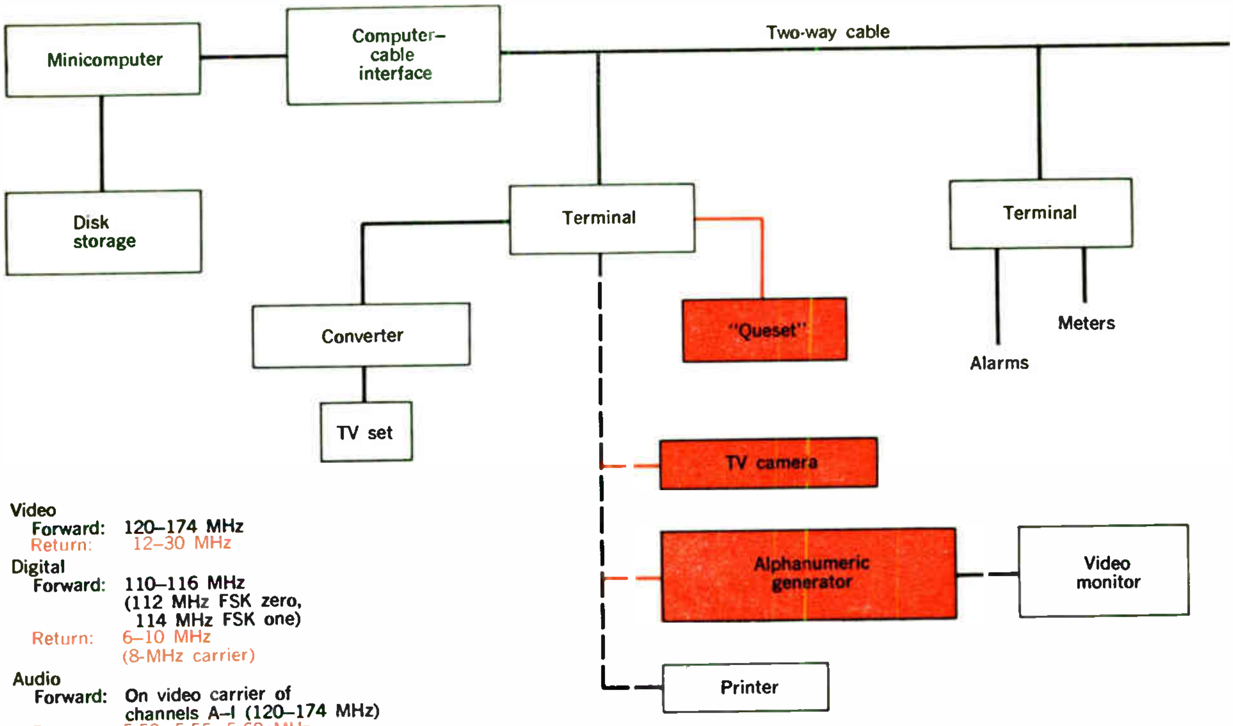
One of the first organizations to develop and demonstrate a two-way cable communications system was the Vicom Manufacturing Company in Dexter, Mich. It has two systems in operation—one at Overland Park, Kans., on Telecable Corporation's CATV network and the other in a major automotive plant.

The Vicom system is designed to accommodate both interactive and monitored services. The return capability includes audio, video, and digital functions under the control of a central computer. The system block diagram and frequency assignments are shown in Fig. 6. The interactive equipment at the subscriber location consists of the following units: a 26-channel converter; a terminal containing the digital and RF circuitry; and the "Queset," Fig. 7, which contains the microphone, 12 alphanumeric keys, and a push-to-talk switch. Several peripheral devices can be connected to the terminal on a temporary or permanent basis as shown by the dotted lines in Fig. 6. The alarms and meter-reading outputs could be connected to the same terminal but, for simplicity of wiring and security reasons, they are interconnected to a separately addressed digital terminal unassociated with the television set.

With the Vicom system, the following two-way environments can be established: remote program origination, audio–video interaction, and digital interaction.

A portion of the return-channel capacity of the system is used to originate programs from any location that contains a terminal. The program is retransmitted from the head end on a forward midband channel. Viewing of this program can be restricted to a particular group of subscribers through the digital interrogation link, which can selectively turn off the converter of those not authorized to view the program. The subgroup of subscribers who have been assigned this channel could be students, professionals, political organizations, club members, or any other assembly of people who desire to use the television medium for their own purposes.

The program originator is also supplied with an alphanumeric generator and video monitor. As each member of



Video
Forward: 120–174 MHz
Return: 12–30 MHz
Digital
Forward: 110–116 MHz (112 MHz FSK zero, 114 MHz FSK one)
Return: 6–10 MHz (8-MHz carrier)
Audio
Forward: On video carrier of channels A–I (120–174 MHz)
Return: 5.50, 5.55, 5.60 MHz

FIGURE 6 (above). Block diagram of the Vicom system and its various frequency assignments.

FIGURE 7 (right). "Queset" terminal for Vicom system.

the subgroup selects the assigned channel, he also uses the keyboard to enter his name. This information appears on the video monitor shown in Fig. 8A (taken at the Overland Park installation). If any member of the group wishes to talk, he enters the request on his keyboard and the information is immediately displayed on the control monitor as shown in Fig. 8B. The program originator inserts the displayed code into his keyboard. The computer then turns on the audio unit within the requesting terminal and a dialogue can ensue. The computer also presents a new code next to the subscriber's name, which can be used by the program originator to turn off the remote audio transmitter as shown in Fig. 8C.

A similar video request procedure can be followed if one of the participants wishes to have the video originate from his terminal. The cablecast request codes are displayed on the monitor for action by the program originator. When the remote camera is turned on, the originating camera is turned off under computer control.

The alphanumeric keyboard on the Queset can also be used to send any general message or data. For example, in the Overland Park equipment, a student may enter his response to a question and the result appears on the teacher's monitor, as shown in Fig. 8D.

Another example occurs in the merchandising test from a Sears, Roebuck store conducted on the same system. A customer has entered his order and it appears next to his identification number as seen in Fig. 8E. However, the purchase order is not accepted until a confirmation code is entered by the customer. The confirmed order appears with an asterisk next to the order number, as in Fig. 8F.

For the merchandising application the monitor information is seen by all of the subscribers. A more comprehensive technique is to provide the message information to each individual. The terminal contains sufficient storage so that a single line of alphanumeric data can be addressed and displayed on the subscriber's television set. In Fig. 8G, an instruction has been sent from the computer to the subscriber, and the subscriber's response is reflexed by the computer as shown in Fig. 8H. This capability provides access by each subscriber to low-cost data bases for retrieval of personal information of widely available data. For most cases it eliminates the need for hard-copy printouts, since the data are always accessible from disk storage.

A more extensive use of the megabit data rate for data entry and retrieval is shown in Fig. 9. The equipment is installed on an in-plant two-way cable within an automotive assembly line. The card reader is used to identify a specific automobile. The Queset is used to enter defects as found by inspectors and also to retrieve various descriptions of the defect data that had been stored on disk.

Initial use of the Vicom system at Overland Park has been for an experimental two-way educational program for a handicapped student confined to his home, and for the Sears experiment mentioned earlier. A variety of other subscriber services has been proposed for future implementation. They would either be available at no cost or at nominal fees, and would include: entertainment, including live local events and prerecorded shows; a wide range of musical selections available upon request that would be received on the subscriber's FM receiver; several different types of many-player, few-player, and one-player games; social group communications with the opportunity for these groups to originate cable-carried group meetings; library services with response to customer inquiries for information as to which visual response is desirable; retrieval of data in groups of 16 characters, displayed one after another by the internal one-line alphanumeric display of the terminal; family budget expenditure records with up to 32 categories, with security code, available for viewing on request; arithmetic calculation service for such elementary operations as tabulating a checking account; ordering of goods with digital keyboard with automatic billing and recording of the transaction; reporting of channel usage; reaction to programming; utilities telemetering with sensors

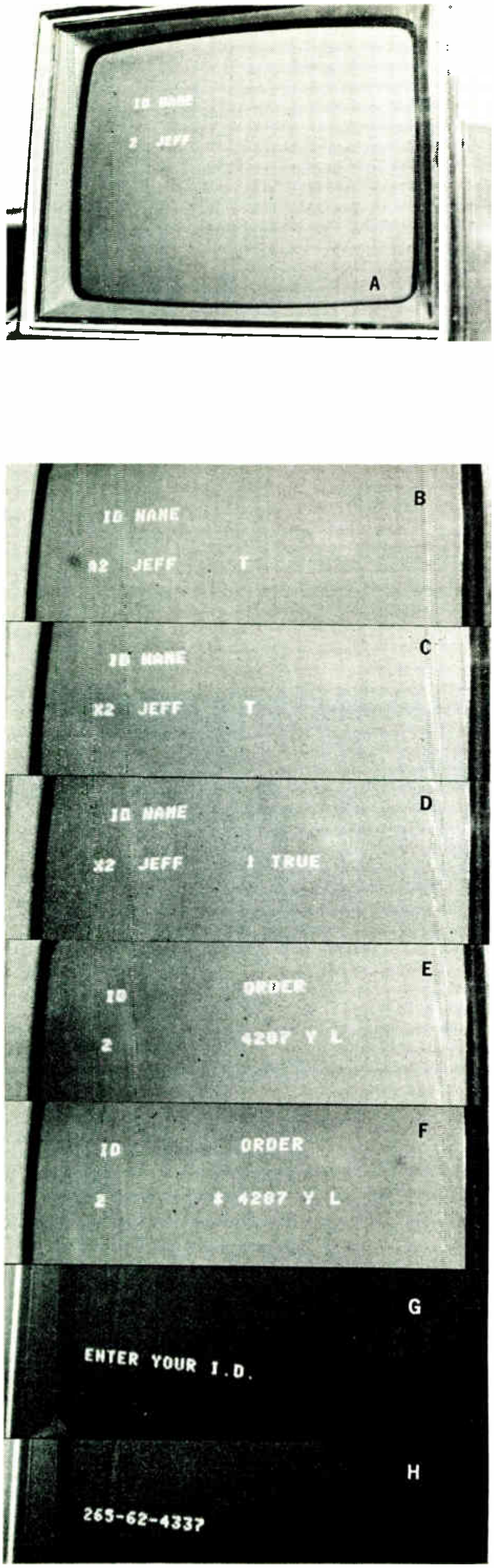
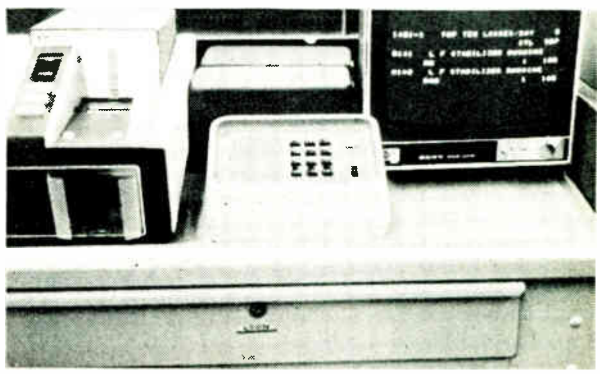


FIGURE 8 (left). Display monitor for program originator in Vicom system. Student enters his name via keyboard and it appears with identification number (A); student's request to talk is indicated by the letter T next to his name and code A2 is used to activate the audio circuit within student's "Queset" terminal (B); code X2 appears to allow student to turn off the audio previously activated (C); student has entered his response to a question and his answer is displayed to the teacher (D); customer has entered a catalog number and his identification number is displayed with his order (E); asterisk entered by customer confirms his order (F); 16 characters can be displayed on the monitor (G); subscriber's response to query in (F) is sent back to the terminal via the computer (H).

FIGURE 9 (below). Application of the Vicom system in an automotive plant. Display gives real-time data from product inspection.



modified for connection to the subscriber terminals; fire-alarm and burglar-alarm services; vote-tabulation service; vocational education; business communications for data exchange and education of employees; sale calculation and inventory records for small businesses; and signal-level diagnosis through automatic monitoring at various locations throughout the cable network.

TOCOM system

A system called TOCOM (TOtal COMmunications) has been developed by CAS Manufacturing Company of Dallas, Tex. In addition to regular television service, the system has many other possible applications, including home protection systems, pay television, surveys for television rating services, meter reading, cable amplifier monitoring, and subscriber response polls.

The single-cable, bidirectional system provides 26 channels of forward transmission (channels 2–13 and A–N) and return transmission in the 6–30-MHz range. Each subscriber on the system is provided with a remote transmitter–receiver (RTR) unit that includes a 26-channel converter; an RF receiving section; a crystal-controlled digital transmitter for return signals; and a digital control system. A complete system can be made up of as many trunks as required with each of the trunks accommodating up to 30 groups, each with 999 subscribers. Each trunk is assigned a specific return frequency so that all 999 RTRs on any one trunk operate on the same return frequency. Return-frequency assignments for trunks are spaced 0.75 MHz apart but may be spaced only 0.25 MHz apart if necessary.

Figure 10 shows a block diagram of an RTR.[13] Signals in the 54-MHz region and up are routed to the converter by means of a filter (not shown). The RF section receives interrogation signals at 50 MHz. The transmitter sends return signals on its assigned frequency somewhere between 6 and 30 MHz.

Figure 11[13] shows the format of the signal used to interrogate the RTRs. A central data terminal (CDT) transmits interrogation information to the RTRs with selected frequency coding in the 50-MHz range. First, a master reset signal is transmitted. It causes all of the RTRs throughout the system to come to a reset state. Next, an ID enable signal is transmitted that enables all the remote RTRs in the entire system to receive an ID code ten bits long. Each RTR decodes the ID signal and, if the code is applicable to any particular RTR, it will respond to the CDT to say, in effect, that it is ready to receive an interrogation. The CDT then transmits an additional signal, which identifies the particular work of the selected RTR that is to be interrogated. That particular RTR then transmits back to the CDT a 16-bit signal corresponding to the desired information, such as a meter reading. An RTR can be sampled in 30 milliseconds and the entire system in 30 seconds.

No TOCOM systems have been installed as yet (other than in the CAS corporate facility) but, with quantity production, the manufacturer hopes to set the price of an RTR at about $100. The CDT cost can range upward from $80 000, depending on the complexity of the particular system being installed.

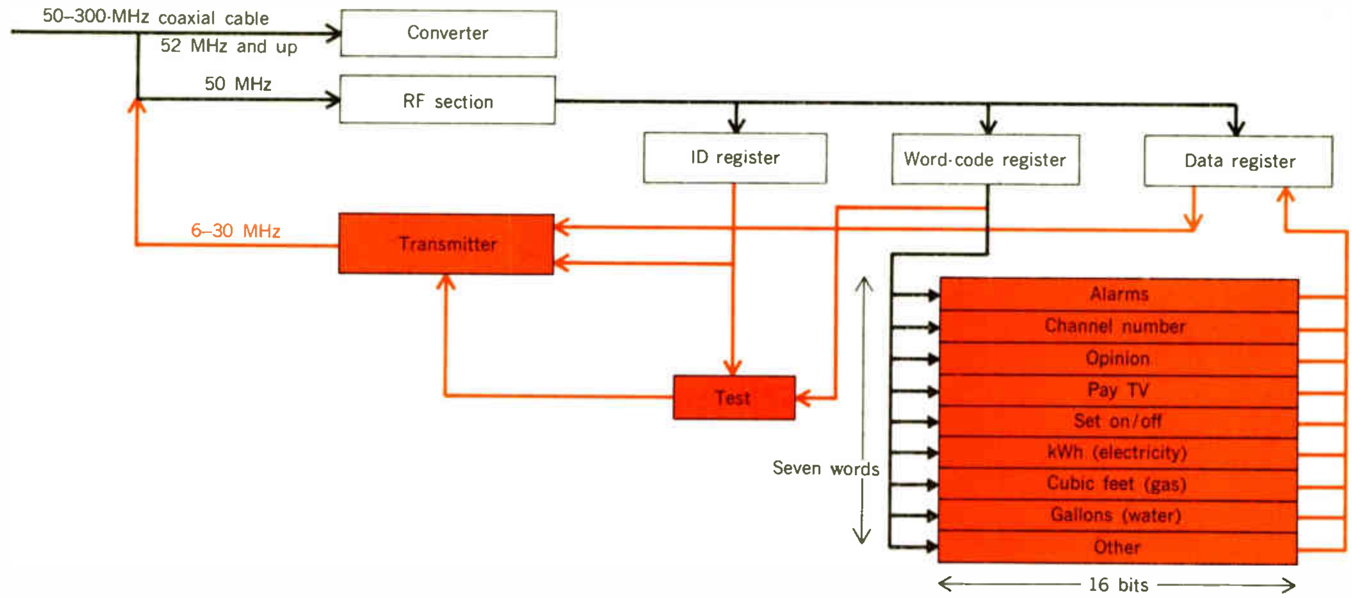
Subscriber Response System

Hughes Aircraft Company in Culver City, Calif., has developed a two-way system known as the Subscriber Response System (SRS). It provides video and digital downstream signals and digital upstream signals. It was developed primarily as a single-cable system but is readily adaptable to dual-cable installations. The system has been under field test for some time by TelePrompTer Corporation in Los Gatos, Calif., in a single-cable installation.

Figure 12 is a functional block diagram of the Subscriber Response System.[14] Two-way communications take place between subscriber terminals and a computer complex known as a Local Processing Center (LPC). At the head end, the SRS digital signals (on a 110-MHz carrier) are frequency multiplexed with the standard video signals and sent downstream through the cable network. At a subscriber's location, the composite signal is routed to a modem unit that converts the 26-channel television spectrum to a fixed frequency signal (usually channel 8 or 12) for reception by the television receiver. The modem performs all of the RF modulation and demodulation and most of the digital signal processing required at the subscriber terminals, and also provides the interfaces for all accessories in the system. All operating controls for the terminals are contained in the subscribers' consoles that are connected to the modems. A console contains a television channel selector switch, a keyboard, and a small strip printer.

The downstream signals occupy the band from 54 to 270 MHz. The downstream SRS signals use pulse code modulation at a one-megabit-per-second rate. The digital SRS signals, which are used to frequency-shift-key a 110-MHz carrier, occupy a 4-MHz bandwidth from 108 to 112 MHz. The upstream signal also uses a 4-MHz bandwidth, from 21 to 25 MHz. Upstream signals are also digital PCM at a data rate of one megabit per second but the 23-MHz carrier is phase-shift-keyed rather than frequency-shift-keyed.

Figure 13 shows a typical communications sequence with the SRS system.[14] The LPC initiates all communications by sending an interrogation message to each subscriber in sequence at a periodic rate. Figure 13 shows typical responses from a subscriber's terminal. If there is no response from the subscriber, either a break in the cable or a faulty subscriber terminal is indicated. If no service is requested, a terminal status report is automatically sent back to the LPC indicating the state of the terminal with regard to proper functioning of the terminal circuitry, the condition of accessory devices, and other diagnostic information. In the example shown in Fig. 13, another possible reply is that the subscriber requests restricted channel 24, which has special programming. The LPC checks the request for eligibility. If the subscriber is ineligible, no further action is taken, but, if the subscriber is eligible, his coded address is remembered by the LPC. After a group of 1000 subscribers has been processed, the LPC services the subscribers' requests. Referring again to the example shown in Fig. 13, the LPC will send a message to the subscriber enabling his television video reception for the restricted channel requested. Simultaneously, the LPC will prepare a billing record on magnetic tape (if there is a charge involved). When 1000 subscribers have been interrogated, the process is repeated for another 1000 subscribers, until the system limit of about 65 000 subscribers per LPC is reached. For a system of 10 000 active subscribers, the time required for any one subscriber to receive a response



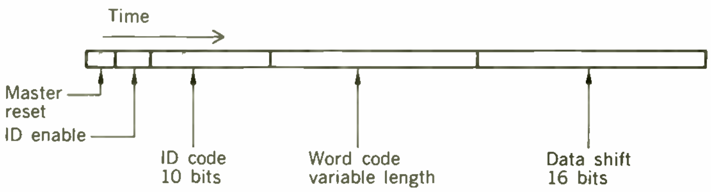
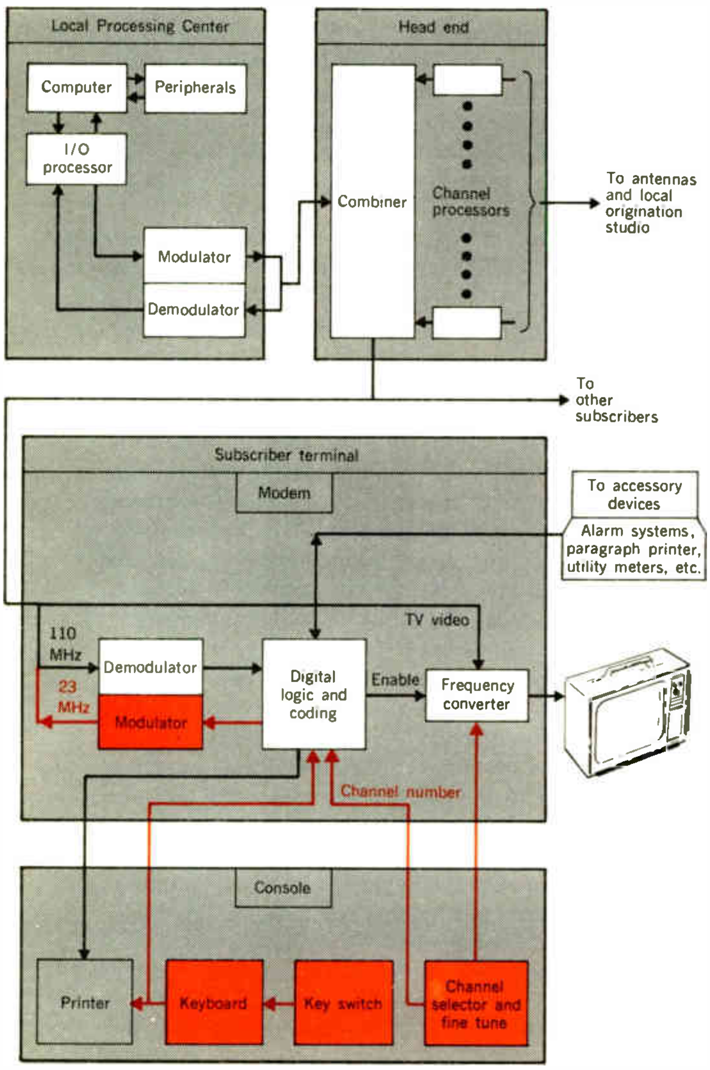


FIGURE 10 (above). Components of remote transmitter-receiver unit in TOCOM system.

FIGURE 11 (left). Format of the signal used to interrogate the remote transmitter-receivers in the TOCOM system.

FIGURE 12 (lower left). Functional block diagram of the Subscriber Response System.



to a request is less than 2 seconds.

Typical services that might be provided by the system include remote channel selection, premium television, restricted television, channel polling, opinion polling, emergency alarms, meter reading, accessory power control/timing, two-way messages, system diagnostics, and system controls.

Two-way message capability can be initiated by a subscriber with the small numeric keyboard that is included as part of his console. Messages may be entered in groups of up to 20 characters at one time. As the subscriber enters data into the keyboard each character is printed on a paper strip so that he can check for errors and also obtain a permanent hard copy record of purchases or other financial transactions. The two-way message capability has many applications, such as shopping at home, educational instruction, reservation services, stock market transactions and reports, quiz shows, mail and advertising, and data bank access.

TICCIT system

An experimental system called TICCIT (Time-Shared, Interactive, Computer-Controlled, Information Television) is being demonstrated currently in Reston, Va., by The MITRE Corporation. The system is innovative in the techniques used to provide a variety of two-way services but actual applications might be restricted somewhat by the fact that one television channel is required for every 600 interactive terminals.

TICCIT is a system for providing computer-generated information that can be received and displayed selectively by standard television receivers. Using one television channel on a wide-band cable, 600 separate television receivers can receive separate information provided by the computer at a rate of once every 10 seconds. A signal "refresh" device provided to the subscriber incorporates a video cassette recorder that allows the television screen to display the information sent by the computer at the standard television rate of 30 frames per second. By using standard telephone lines and a 12-button push-button phone, each subscriber can call for any kind of information the system is designed to provide, independently of all other subscribers.

In the experimental system at Reston, the CATV network of The Reston Transmission Company is used. It is a dual-cable system that provides each subscriber with 12 channels on each cable.

The computer for the Reston demonstration is located in McLean, Va. Voice and still-picture signals generated by the computer are transmitted to the head end of the Reston CATV system by microwave link. At the head end the signals are put onto channel 13 of cable A.

At the computer facility, standard telephone lines are terminated in modems that convert subscribers' information requests into a digital format. A character generator at the computer facility provides a standard television signal picture containing up to 800 characters (16 lines of 50 characters each). The character set consists of 96 alphanumerics.

Because of the characteristics of the home refresh device and the desire to maximize the use of each television channel, the demonstration system does not use the interlace feature of standard television. Instead, the transmitted picture frames contain only every other line of a standard picture, which permits sending up to 60 different pictures every second rather than 30. Address data, supplied by the character generator, are encoded near the last line of each previously transmitted picture frame as a sequence of 16 black or white elements, each measuring $^1/_{16}$ of the line length. Using 16 elements allows addressing of up to 32 000 different home terminals.

The home terminal is shown in the block diagram of Fig. 14. It uses a standard television receiver, a video tape recorder, and a coupler/decoder. The video tape recorder serves as the refresh memory. The type of recorder used records a $^1/_{60}$-second picture frame and plays it back continuously if the take-up and supply reel mechanism for the recorder is disabled. The recorder is driven by a composite video signal coming from the coupler/decoder, which also provides some control signals for the recorder mechanism. The coupler/decoder has a television tuner, an IF strip, and a detector so that the recorder can record material from any television channel. The decoder and control portion of the terminal examines lines 480 and 481 of each received picture frame and compares the black-and-white pattern with the address pattern set in the decoder. If the transmitted picture's address matches the decoder's address, the recorder is placed in the record mode for the entire picture frame. At the completion of the frame, the recorder automatically will be placed back in the playback mode. During all operations in the computer display mode, the recorder tape-motion hardware is inhibited.

When a user requests information from the system, he simply dials the phone number for the computer. It signals automatically with a beep when it is ready for interactive communication. The computer then displays on the television screen a directory of services and tells the subscriber to depress the number on his push-button phone corresponding to the service in which he is interested. In an operational system, alternatives to push-button phones would be electric typewriters or other devices operating on part of the two-way cable system.

Possible two-way applications for TICCIT are limited only by one's imagination. At the present time, software programs are being designed under the sponsorship of the National Science Foundation to provide individualized educational courses for home or school use. The system also provides capabilities for selective distribution of materials during "off hours." A movie or book or newspaper, for example, could be sent to a subscriber's video recorder to be stored for later display on his television receiver.

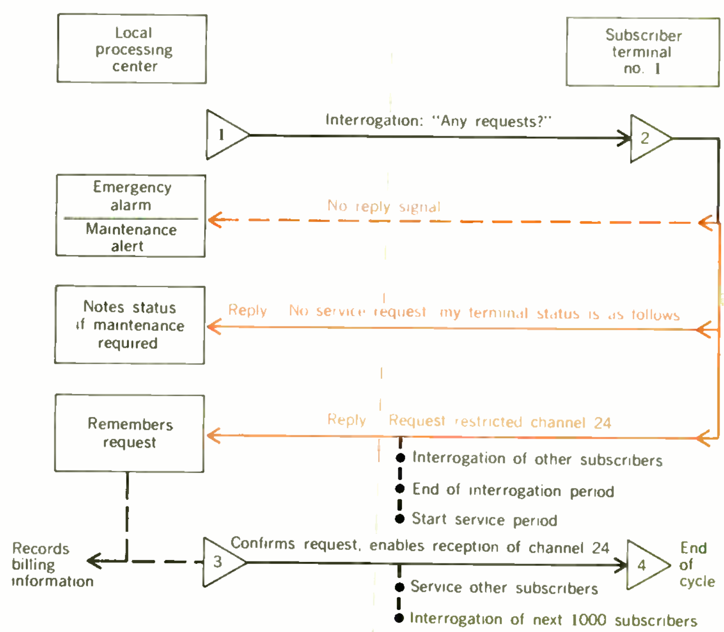
Dial-a-Program system

Since October 1970, a small-scale field trial of a unique approach to two-way communications has been under way at Dennis Port, Cape Cod, Mass. Installed by Rediffusion International Limited, the system is basically different from those already discussed but offers some very decided advantages. In fact, it may well be the ideal type of cable system for two-way communications.

The Rediffusion system is a multiple-cable approach but not in the sense of multiple trunks with various multiplexing schemes. It is, instead, a switched system whereby program material is supplied to each subscriber independently of the other subscribers on the system. The cables used are not coaxial. They are HF twisted pairs. And any subscriber, given the necessary equipment, can be a program source for all other subscribers on the system.

The heart of the basic system is a central exchange to which each subscriber is connected by two pairs of cables. One carries program signals in either direction; the other carries control signals. At the exchange, a 36-position

FIGURE 13. Typical communications sequence with the Subscriber Response System.



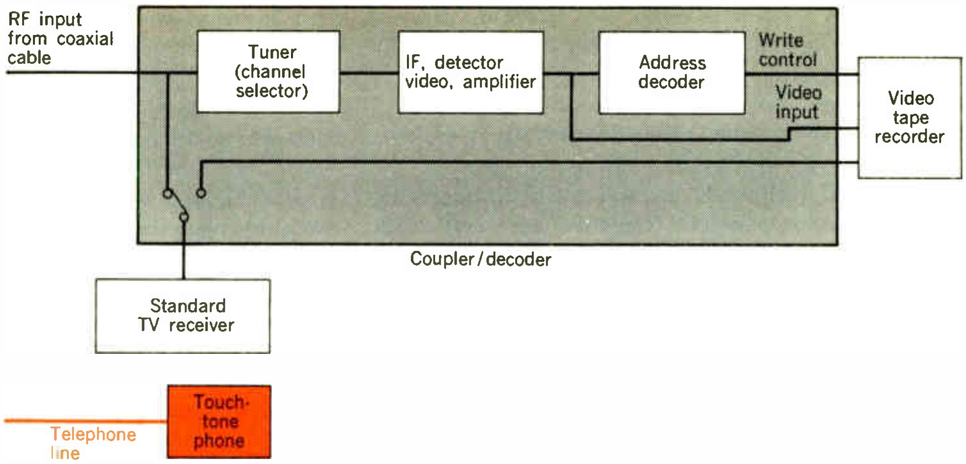


FIGURE 14. Home terminal for the TICCIT system.

reed selector switch is dedicated to each subscriber. Each of the 36 positions on the switch can be connected to a program source. The subscriber is provided with a telephone-type dial selector switch that generates pulses of direct current to operate the 36-position switch in the exchange. When a subscriber wishes to receive a particular program on his standard television receiver, he merely dials the code number corresponding to that program material. (Channel 0 of the exchange switch gives a list of programs available and their dialing codes.) The dc pulses are sent to the exchange on one of the two control cables. The selector arm on the exchange switch moves through the 36 positions in sequence until the one selected by the subscriber is reached, and then stops in that position, making the connection between the subscriber and the program material at the head end in the exchange. The program material is sent to the subscriber on the remaining pair of cables. When the subscriber wishes to switch to another program, he presses a reset button on the dial selector mechanism. This action causes a pulse to be sent on the second of the two control cables to the exchange where the switch dedicated to that subscriber is caused to return to the zero position. The subscriber then dials the code for the new program desired and the action takes place as before.

With this system it is a simple matter for any subscriber to be a program source for all other subscribers. For example, at the Dennis Port installation, one of the subscribers is a delicatessen. It advertises daily specials to the other subscribers through use of a television camera and a simple, handwritten sign. The camera signals are sent back to the exchange on the delicatessen's programming pair of wires. The advertising message then becomes available to all other subscribers who dial the proper code number. Upstream programming signals are in the 9–15-MHz band; downstream signals in the 3–9-MHz band.

At the present time, 11 off-the-air television channels, and one weather information channel, are on the Dennis Port system. In addition to the delicatessen programming, the administrative offices for the system, which are on the network, also originate programs, on occasion, for demonstration purposes. These programs are available to all subscribers. To demonstrate the full 36-channel capability of the system, programs are sometimes duplicated or triplicated and test patterns are used.

The extreme flexibility of the Rediffusion system is apparent from the preceding description. Even though the trial installation has a maximum of 36 channels available, it is a simple matter to dedicate a second exchange switch to each subscriber for a total of 72 channels, or even more if desired. An important advantage of the system is that channel capacity can be increased dramatically without making any changes whatsoever in the distribution network or the terminal equipment in any subscriber's home.

Because of the simple frequency division used on the programming cables, distortion is not a problem. Another advantage of the system is that, for certain applications, the need for expensive terminal equipment in a subscriber's home is eliminated. For example, in the TICCIT system discussed earlier, a video tape recorder or some other type of "frame snatcher" is needed by each subscriber. With the Rediffusion system, one or more frame snatchers could be located at the exchange and time-shared by subscribers.

As with any system, there are disadvantages as well as advantages and the Rediffusion system is no exception. Possibly the major drawbacks, depending on the particular installation, are the need for one central exchange for every 336 connections and, in the trial system, a limited distance from subscriber to exchange of about 550 meters. The 336-connection limit is partly determined by a restriction on the maximum number of 36-position switches in any exchange and partly by the complexity created by the numbers of cables to be connected between the exchange and the subscribers. Main distribution cables from the exchange contain 12 pairs, enough cabling for just six subscribers. Thus it is easy to see how quickly the cabling can become a problem. The 550-meter limitation on the trial system comes about partly because of the limited power output of available HF repeaters and partly as a result of video signal cross-coupling problems. Both problems can be alleviated somewhat[15] by higher-powered repeaters now under development, by use of individual subscriber amplifiers in the exchange serving those who are most distant, and by cable selection procedures whereby the cables with the least cross coupling are used for the most distant subscribers.

GTE/St. Charles proposal

GTE Laboratories, Incorporated, has submitted a proposal[16] for a two-way cable system for St. Charles Communities, Maryland, a new community being developed 48 km southeast of Washington, D.C. The proposal is for a switched system similar in network configuration to the Dial-a-Program system. In fact, Rediffusion's 36-position reed switches are to be used in the St. Charles system. The distribution network differs from the Dial-a-Program system in that coaxial cables are used in place of HF twisted pairs. The St. Charles system has the versatility of the Dial-a-Program system supplemented by the availability of a bandwidth of at least 45 MHz. This provides the capability of sending each subscriber two different television channels simultaneously for use with two different television receivers.

Local Distribution Centers (LDCs) are the heart of the St. Charles network. All program sources feed into an LDC. Every subscriber is connected to an LDC by a private coaxial cable that carries signals in both directions. Up to 24 private cables are bundled together in one large cable for distribution lines from an LDC to individual subscribers. Because subscriber connections are limited to about 1200 meters in length, a neighborhood may be served by more than one LDC.

The private coaxial cable is terminated inside a subscriber's home at a small service entrance box that contains coaxial cable connections, a simple network that permits transmission of security alarm signals over the cable, and connectors for the fire and burglar alarms.

At the LDC each subscriber is connected to a 36-contact selector switch—the Rediffusion switch—that provides him individual access to 36 program busses. The switch is activated by the subscriber by using a touch-call keyboard to select the desired program.

Each subscriber's line can carry three independent television signals, two from the LDC to the subscriber and one in the reverse direction, as well as a 20-MHz bandwidth of FM audio channels. In the return direction, audio talkback, preference responses, dial signals, and fire and intrusion alarms are carried to the LDC to be processed and delivered to the desired origination or monitoring facility. Since all subscriber lines can carry a video channel plus miscellaneous narrow-band signals back to the LDC, local programming is particularly easy. Any residence or building connected to the system can originate programs with a television camera and some auxiliary electronic modules.

Figure 15 shows the frequency-assignment scheme used for the St. Charles system. The designations A and B refer to the two separate television channels available to any subscriber who wishes them for reception on two different television receivers. For such service, the subscriber would also need a second touch-call keyboard to select the desired program for the second receiver. The second keyboard is tied in with a second 36-position switch at the LDC. Initially, the same set of 36 programs will be available on all switches so that the second receiver would not be able to receive any different programming. But, in the future, a second set of different programs could be made available on the second switch.

The television channels received by the subscriber, Fig. 15, are not on standard frequency carriers but are up-converted to channel 2 (54–60 MHz) and channel 4 (66–72 MHz). The FM sound channels that are received in the 45–65-MHz range are up-converted to 88–108 MHz for use by standard FM receivers.

Programming possibilities for the system are unlimited but suggested channel assignments for the initial system, in addition to 12 television entertainment channels and six FM channels, include: weather service; stock market information; shopping; scheduled educational programs; access to educational programs on tape; adult educational programs with talkback; adult educational programs with tape access; scheduled community programs with and without talkback; youth programs; senior citizen programs; and private two-way video channels for home-bound students and schools, patients at home and doctors, and nursing homes and hospitals.

The St. Charles system, as is true with the Dial-a-Program system, has the advantage of not requiring each subscriber to have a frame snatcher for educational applications. The frame snatcher can be located at the LDC and the expense to any one subscriber thus minimized.

DISCADE system

The final system to be described is one developed by AMECO, Incorporated, of Phoenix, Ariz. Installation of the DISCADE system is now under way in the Disney World complex in Orlando, Fla., where it will supply about 2000 hotel rooms, and another installation has just been completed in Daly City, Calif.

The DISCADE system[17] is similar to both the Dial-a-Program and GTE systems in that it is a remotely switched cable distribution system. It differs from both in that each television or other communications channel, or two of these channels in another version of the system, is distributed to a subscriber on a separate coaxial cable. Another difference is that all switching and control functions are accomplished with solid-state circuits.

The original Daly City system consisted of a 20-channel head end and a space-division-multiplexed subfrequency (7–13-MHz) transmission system with each of the 7–13-MHz channels connected to a separate coaxial cable. Four Area Distribution Centers, each with a capacity of 48 switching modules, were installed to provide service to 192 subscribers. Problems arose in that the equipment used, particularly repeater amplifiers and Area Distribution Centers, was physically large and difficult to install under existing pole attachment practices. The system cost appeared to be uncomfortably high.

In a newer version of the system, known as DISCADE II, two channels are carried on each cable. One half of the VHF channels are converted to channel A, with a 35.25-MHz visual carrier frequency and a 39.75-MHz aural carrier frequency, and one half to channel B, with a 23.25-MHz visual carrier frequency and a 27.75-MHz aural carrier frequnecy. One channel A and one channel B signal are connected to each coaxial cable transmission line. The number of transmission lines used is determined by the total number of channels. In this case, for a 20-channel system, a total of ten transmission cables are used. An additional cable carries the FM band throughout the system. The FM signals are converted at the head end to the 20–40-MHz band and then block-converted upward to 88–108 MHz for reception by standard FM receivers.

The channel selector in the subscriber's home is de-

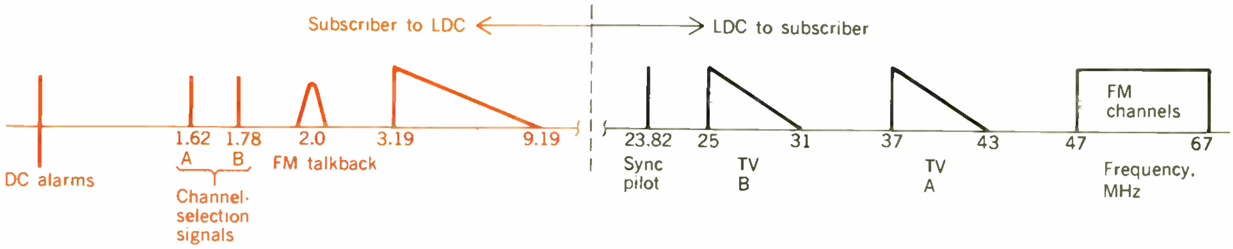


FIGURE 15. Frequency-assignment scheme for the GTE/St. Charles system.

signed to switch from one line to the next on alternate channel selector switch positions. Channel A and channel B frequency-conversion oscillators are alternately switched on and off with each channel selection position.

In the Disney World development, the frequency of the television signal, carried one channel per cable, matches the intermediate frequency of the television receiver being used and thereby eliminates the need for any tuning or conversion at the receiver.

Current development programs for the DISCADE system include design of equipment for automatic monitoring of the subscriber's service connection and design of equipment for providing subscriber response and home-security service.

A new DISCADE System is also planned for Salt Lake City, Utah.

Conclusion

From this necessarily brief discussion of two-way systems it should be apparent that there are a great many techniques being proposed or actively under field test for achieving two-way communications on CATV systems. It is perhaps ironic that with so many different systems evolving there are so few customers demanding two-way services. But, as with all new developments, demand comes only when the public has been exposed to the advantages of such developments. That exposure will surely take place as enterprising government agencies, cable system operators, and equipment manufacturers take the lead in putting systems into the field. The public will benefit in the long run and one can only hope that an overall communications program will emerge that will permit CATV systems, the telephone network, and commercial television broadcasting to complement each other rather than compete.

REFERENCES

1. Singer, A. L., Jr., "Issues for sudy in cable communications," Alfred P. Sloan Foundation, New York, N.Y., Sept. 1970, p. 4.

2. "The future of broadband communications," IED/EIA response to FCC Docket 18397, Part V, Oct. 29, 1969.

3. "Communications technology for urban improvement," report to the Dept. of Housing and Urban Development, Committee on Telecommunications, National Academy of Engineering, Washington, D.C., June 1971.

4. Simons, Ken, *Technical Handbook for CATV Systems* (3rd ed.). Philadelphia, Pa.: Jerrold Electronics Corp., Mar. 1968, p. 47.

5. Simons, Ken, "The optimum gain for a CATV line amplifier," *Proc. IEEE*, vol. 58, pp. 1050–1056, July 1970.

6. Rheinfelder, W. A., *CATV System Engineering* (3rd ed.). Summit, Pa.: TAB Books, Feb. 1970, p. 227.

7. Walding, G., "Spectrum pollution and the set top converter," *TV Commun.*, vol. 8, pp. 142–148, July 1971.

8. Barnhart, A. W., "Two-way systems performance," paper presented at 1971 NCTA Convention, Washington, D.C.

9. Rogeness, G. G., "Two-way repeater station utilizing hybrid thin-film amplifiers as building block," *Proc. 1970 NCTA Convention* Chicago, Ill.

10. Avins, J., Harris, B., and Horvath, J. S., "Improving the transient response of television receiver," *Proc. IRE*, pp. 274–284, Jan. 1954.

11. Rogeness, G. G., "Two-way transmission on the CATV cable," paper presented at 1968 NCTA Convention, Boston, Mass.

12. Taylor, A. S., "A proposal to the FCC for increasing TV channels," *TV Commun.*, vol. 8, pp. 51–55, July 1971.

13. Osborn, W. F., "TOCOM system: bidirectional cable television information and control transmission system," paper presented at 1971 NCTA Convention.

14. Callais, R. T., and Durfee, E. W., "The subscriber response system," paper presented at the 1971 NCTA Convention.

15. Gabriel, R. P., "Experience with the Dial-a-Program System," paper presented at the 1971 Northeast Research and Electronics Meeting, Boston, Mass.

16. "Proposal for an experiment in communications-electronics services for St. Charles Communities, Maryland," vols. I and II, GTE Laboratories, Inc., Waltham, Mass., Sept. 1971.

17. Hickman, J. E., and Kleykamp, G. C., "Multicable solution to communications systems problems," paper presented at the 1971 IEEE Internat'l Convention, New York, N.Y.

Reprints of this article (No. X71-112) are available to readers. Please use the order form on page 8, which gives information and prices.

Ronald K. Jurgen (SM), managing editor of IEEE Spectrum, has been in the technical editing field for 21 years and has been managing editor of IEEE Spectrum since its birth in January of 1964. He was graduated from Rensselaer Polytechnic Institute in 1950 with the B.S. in E.E. degree. He received his basic editorial training as an assistant editor on the staff of Electrical Engineering, and then with Electronics under the editorship of Donald G. Fink, now IEEE's general manager.

In 1953, Mr. Jurgen, in the capacity of editor, started Sutton Publishing Company's monthly journal, Electronic Equipment, now EDN/EEE. After five years in that position, he returned to Electronics to be feature editor. His next position was manager of scientific information, on the corporate level, for IBM. He rejoined Sutton to become editor of Industrial Electronics and, in 1961, returned to Electrical Engineering to become editor in 1962.

Placing atmospheric CO_2 in perspective

Contrary to popular belief, increases in atmospheric CO_2 have not drastically altered the world climate; in fact, over the last three decades, rising CO_2 levels have helped to slow down the effects of a steady decrease in world temperature

Arthur D. Watt Westinghouse Georesearch Laboratory

Atmospheric carbon dioxide, which amounts to 320 parts per million (ppm) by volume, rather than being a pollutant, is essentially a thread of life woven through the globe on which we live. In the past century alone, the amount of CO_2 in the atmosphere has increased by 40 ppm, with levels increasing at a current rate of about 0.75×10^{10} tonnes per year. Fortunately, man can tolerate CO_2 levels many times present concentrations, and plant life actually grows better at increased CO_2 levels. What does cause concern is the effect that atmospheric CO_2 has on the earth's climate. It appears that the 40-ppm increase over the last century may have contributed to a global temperature increase of the order of 0.2 K. Since 1940, however, the global atmospheric temperature has been decreasing—an indication that other factors (such as atmospheric dust) are of much greater importance in determining the overall heat balance of the world.

Each year, approximately one tenth of the carbon dioxide in the atmosphere is exchanged at ocean and land surfaces, with the biosphere playing a decidedly important role. At present, man is injecting into the atmosphere approximately 2×10^{10} tonnes (metric tons) of CO_2 per year as a result of fossil fuel burning. This is about 10 percent of the annual exchange and 1 percent of the atmospheric total. By far the largest input of CO_2 to the atmosphere is derived from the tropical oceans, amounting to an estimated 15×10^{10} tonnes per year.

Rate of exchange is primarily governed by temperature, *p*H, and wave action, which are the most significant factors in maintaining the level of atmospheric CO_2. Also of great importance is the continual removal of CO_2 by plant growth, believed to consume 7×10^{10} tonnes per year. Despite the delicate balances that contribute to the overall rate of exchange, however, atmospheric CO_2 levels are annually increasing by about 0.75×10^{10} tonnes, or over a third of the total amount introduced to the atmosphere by man in burning fossil fuels.

Approximately 99 percent of the earth's atmospheric materials (5.3×10^{15} tonnes) are contained in the first 30 km of this life-supporting layer surrounding the earth. The major constituents of the atmosphere are gases and water vapor. The gases consist (by volume) of 78 percent nitrogen, 20 percent oxygen, 0.9 percent argon, and 0.03 percent CO_2. In addition, there are numerous other gases present, including ozone with a concentration of the order of 0.000 005 percent. Water vapor amounts to approximately 0.2 percent (by weight), which is equivalent to 2–3 cm of precipital moisture distributed over the whole earth. There is also about 0.008 percent of water droplets and ice crystals, and a small amount of dust with diameters predominant in the region from 0.1 to 1 μm with an effective weight of approximately 0.000 000 1 percent. Atmospheric dust loading varies over a range of several factors, very often resulting from volcanic action.

Gases are rather uniformly distributed up to a height of some 90 km, with the exception of ozone, which is concentrated in a layer 20–60 km above sea level. Water vapor is contained in the first several kilometers above the earth's surface. Fine dust particles (0.1 μm in radius) are found near the earth's surface in the lower troposphere, with increased concentration occurring during windy days in dry areas and near urban centers. Coarser particles (near 1 μm radius) are found in the stratosphere, with a maximum near 18 km above the earth's surface and a concentration that increases after large volcanic eruptions.

There has been considerable conjecture for a number of years as to the effects of the growing amount of atmospheric carbon dioxide on the world's climates. These claims range all the way from rather minor increases in world temperature to several degrees of change with a resultant melting of the ice caps and the inundation of large coastal cities.

Carbon dioxide in the air is one of the essential keys to all life. All primary production of land-based food is dependent upon atmospheric CO_2. In fact, 93 percent of the basic materials used by land plants in the photosynthetic production on carbohydrates come from the carbon dioxide of the atmosphere. The remaining 7 percent is supplied from groundwater. From this we can surmise that, rather than being a contaminant in the air we breathe,

Based on material presented at the First Westinghouse School for Environmental Management, Fort Collins, Colo., June 15–July 10, 1970.

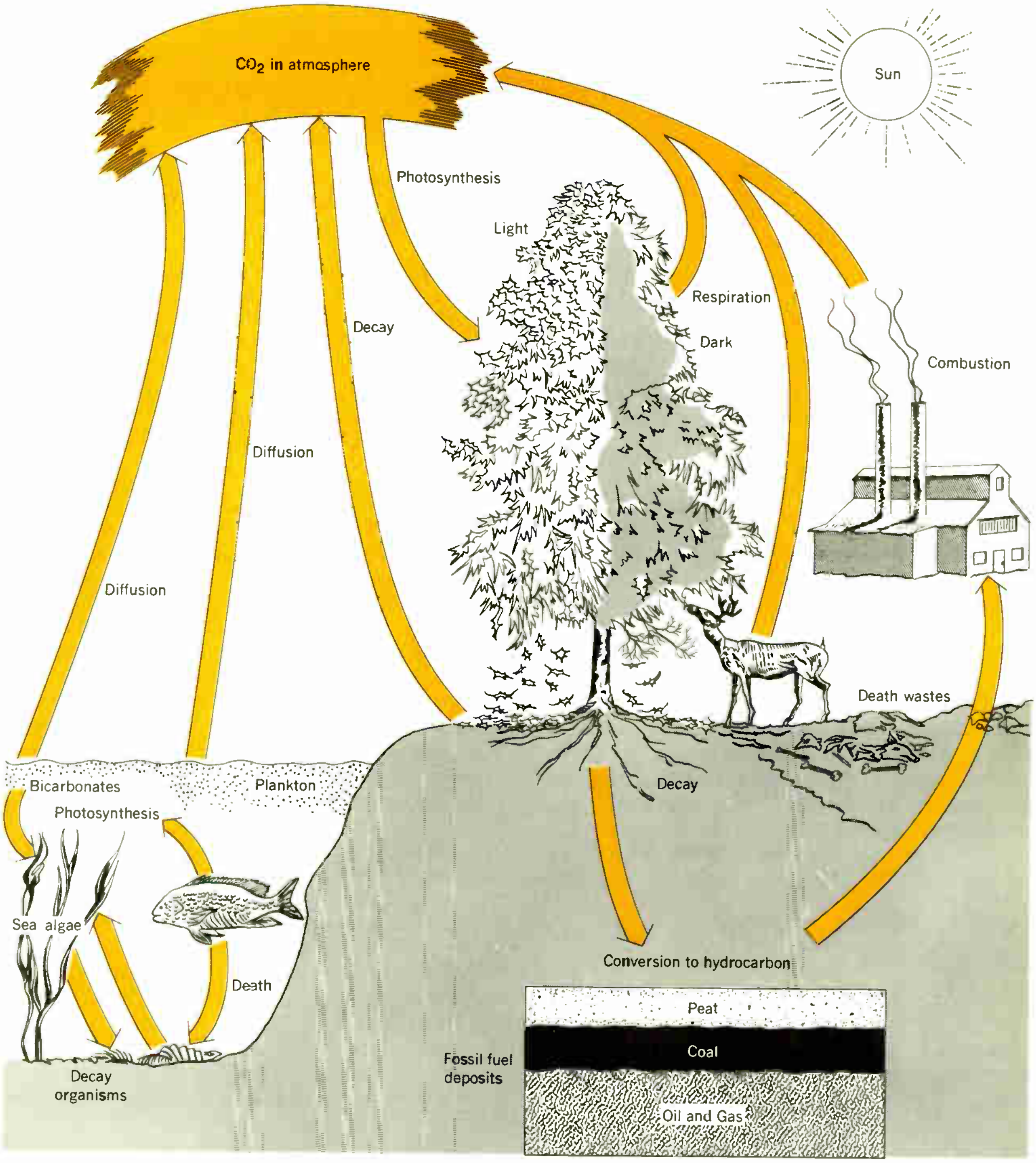


FIGURE 1. The carbon cycle in the ecosystem. (Adapted from Smith,[2] p. 49)

CO_2 is absolutely essential to life on earth.

It is instructive to observe that even if all the world's economic hydrocarbon fuel resources were burned in one year and no greater absorption occurred, the atmospheric CO_2 would only increase to three to six times the present level, i.e., to about 1000 to 2000 parts per million (ppm). By comparison, the human safety limit for prolonged exposure to CO_2 is reportedly[1] about 5000 ppm or 15 times the present level. The upper limit for plant response has not been established, but it is known that plants grow more rapidly with increased levels of CO_2 ranging up to 100 times the present atmospheric level.

In view of these considerations, it appears that the only significant remaining area for possible concern over increasing atmospheric CO_2 lies with its possible effects on global climate. Since existing literature on this subject is controversial, we will first examine the basic CO_2 cycle along with observed spatial and temporal variations of atmospheric CO_2. The probable effects of increasing CO_2 levels will then be examined in the light of experimental and theoretical evidences.

The basic CO_2 cycle

In man's environment, there are a number of highly significant cycles by which both dominant and trace elements are cycled through the combined physical and

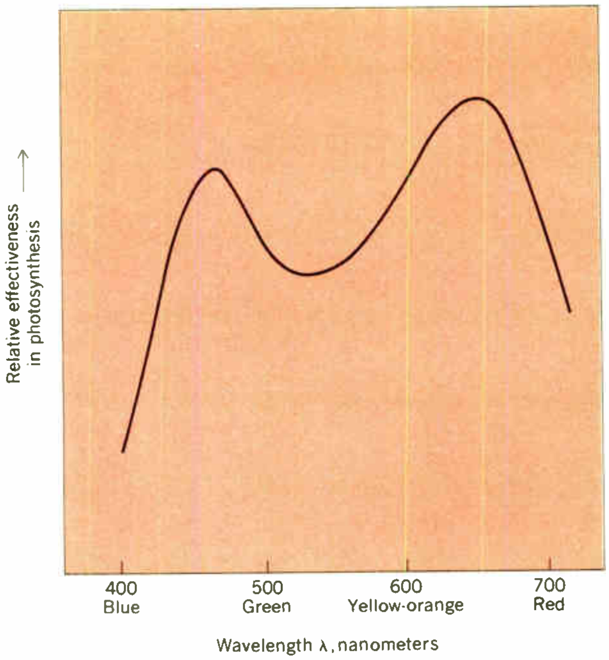


FIGURE 2. Action spectrum of the photosynthesis process. (From Keeton,[3] p. 118)

biological systems. The characteristic cycling of carbon dioxide through a local ecosystem is shown in Fig. 1. Solar energy is a basic requirement for the photosynthesis process, which is expressed in simplified form as

Photosynthesis

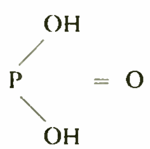
		(from the air) ↓		solar radiation ↓↓↓				(to the air) ↑	
H_2O	+	CO_2	+	$\xrightarrow{\text{light energy}}$	$\frac{1}{6}\{C_6H_{12}O_6\}$	+		O_2	(1)
18 grams		44 grams		112 kcal	30 grams			32 grams	
water	+	carbon dioxide	+	light energy	→	glucose	+	oxygen	

The regions of the electromagnetic spectrum that are most effective in photosynthesis are the orange-red portion with a secondary peak in the green region (see Fig. 2).

The actual photosynthesis and respiration process involves various other essential elements and catalysts. Keeton[3] pointed out that "the reduction of carbon dioxide to form glucose proceeds by many steps, each catalyzed by an enzyme." Two key nitrogen- and phosphate-containing organic compounds are involved: ADP (adenosine diphosphate) and ATP (adenosine triphosphate). Energy is added in the reaction

$$\text{ADP} + Ⓟ + \text{energy} \xrightarrow{\text{enzyme}} \text{ATP}$$

where Ⓟ designates the addition of a phosphate group



Energy is removed in the reaction

$$\text{ATP} \xrightarrow{\text{enzyme}} \text{ADP} + Ⓟ + \text{energy}$$

From these considerations, it is apparent that nitrogen and phosphorus availability as well as trace elements for enzymes are important factors in plant growth.

Based on these energy relationships, it can be shown that the first-level production of 1 kg of glucose requires 3.7×10^3 kcal = 4.35 kWh of energy. In this production, approximately 1.47 kg of CO_2 is consumed along with 0.6 kg of H_2O. Tracer studies have revealed that the oxygen in the glucose all comes from the oxygen in the carbon dioxide and, as a result, about 93 percent of the glucose produced comes from the CO_2 in the air and only 7 percent (i.e., the hydrogen) comes from water. The oxygen given off, on the other hand, comes from both water and CO_2.

In plants, there is continual respiration going on day and night, during which time glucose is combined with oxygen to yield carbon dioxide, water, and energy. This reaction is shown in the following simplified form:

Respiration

	(from the air) ↓		(to the air) ↑				
$\frac{1}{6}\{C_6H_{12}O_6\}$ +	O_2	⟶	CO_2	+ H_2O	+ energy		(2)
30 grams	32 grams		44 grams	18 grams			

Typical plants use up approximately 40 percent of the first-level $C_6H_{12}O_6$ in the production of energy to be used in plant-life housekeeping processes, including the transport of water through the plant and the conversion of first-level glucose to the final carbohydrate forms such as insoluble cellulose ($C_6H_{10}O_5$) within the plant. From this we conclude that the maximum theoretical efficiency expected from plants is 60 percent, and the production of 1 kg of carbohydrate materials in the plant requires 6.2×10^3 kcal = 7.2 kWh* of energy from a spectrally matched source. Since 180 grams of glucose yield 162 grams of cellulose plus 18 grams of water, the production of each kilogram of cellulose removes 1.6 kg of CO_2 from the atmosphere.

The photosynthesis/respiration process has a theoretical maximum efficiency of 60 percent, which requires the assumption that the light spectrum matches the photosynthesis activity curve. Solar energy contains a consider-

* By way of comparison, the heat content of a typical wood is 5 kWh/kg.

able amount of energy outside the action spectrum for photosynthesis and, as a result, the overall carbohydrate production must be modified by a spectral utilization factor that amounts to 10–20 percent for terrestrial solar energy. The result is that we would anticipate an upper bound for photosynthesis production of carbohydrates in the region of 6–12 percent of the total solar energy received. In actual conditions, all the solar energy incident to the earth's surface does not reach plant leaves; for normal growing conditions and spacing between plants, approximate efficiencies of food production from solar energy are 4 percent for algae, 2 percent for sugar beets, 0.2 percent for wheat, and 0.1–1 percent for forest trees.

Returning to the pictorial carbon cycle of Fig. 1, the carbohydrates, once produced by plants, are seen to be eaten by herbivores, or else to decay, in which case most of the carbon dioxide originally consumed is returned to the atmosphere through bacterial action. If appropriate heat and pressure are applied to the carbohydrates, a process of conversion to hydrocarbons takes place and the familiar coal deposits result. In this conversion to hydrocarbons, oxygen is driven off and the heat content per unit mass of the material increases. The net result is carbon withdrawal from the atmosphere and its storage in fuel deposits.

Withdrawal from the short-term cycle can also occur in the marine environment. Phytoplankton take the CO_2 that has diffused from the atmosphere into the upper layers of water and convert it into carbohydrates. This in turn serves as food for marine life, which in some instances produces coral atolls. Such production effectively withdraws carbon from the short-term CO_2 cycle but is available for reintroduction, under proper conditions, to the atmosphere–ocean CO_2 exchange process. Before attempting the formulation of an overall CO_2 cycle, we will examine global carbon deposits and observed levels of atmospheric and oceanic CO_2.

Global carbon deposits and observed CO_2 levels

About 99.9 percent of the earth's carbon is contained in the land. Estimates place the earth's crust deposits at 2.7 $\times 10^{16}$ tonnes, largely as carbonates and with only a small amount in the form of hydrocarbons and a much smaller amount as carbohydrates (Fig. 3).* The world's oceans contain about 0.1 percent of the total, and the atmospheric carbon content in the form of CO_2 is 0.0026 percent or approximately 2.6 percent of that contained in the oceans. The carbonate deposits on land are widely distributed in a nonuniform manner, as are ocean deposits.

At present, the amount of carbon dioxide observed in air has an annual average value of approximately 320 ppm by volume. Since air has a mass of 28 kg per unit volume as compared with 44 kg of CO_2 per unit volume, this is equivalent to 505 ppm of CO_2 by weight. With an atmospheric pressure of 10^4 kg/m² distributed over 5.1 $\times 10^{14}$ m² of global surface, 505 ppm of CO_2 yields 2.6 $\times 10^{12}$ tonnes of CO_2 in the earth's atmosphere. The equivalent amount of carbon is 7 $\times 10^{11}$ tonnes. Observations over the past century have shown that atmospheric CO_2 is increasing according to the trend plotted in Fig. 4. More detailed observations in the last decade have shown

* Much of the data given here for CO_2 deposits and exchange rates were obtained from Lieth,[4] p. 3895.

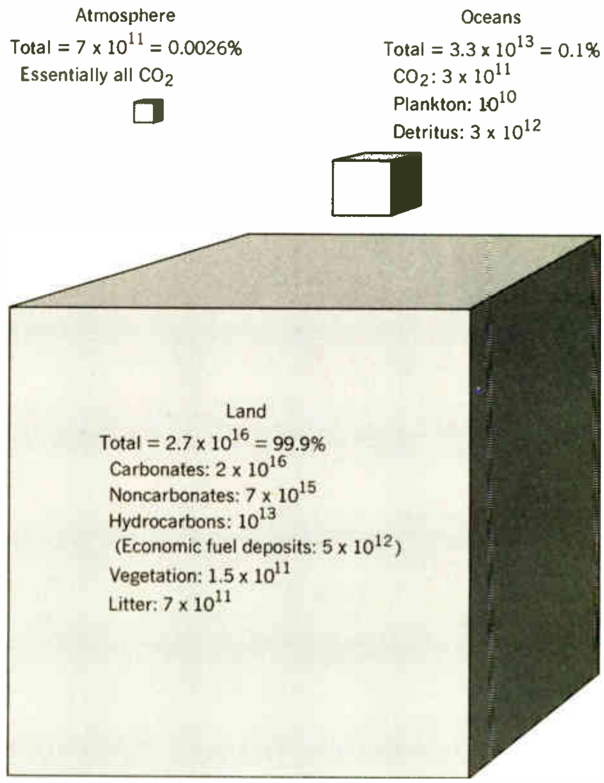


FIGURE 3. Global carbon distribution. Values are given in tonnes of carbon; to obtain equivalent values of CO_2, multiply by 3.7.

FIGURE 4. Trend of atmospheric CO_2 (mean annual data from the northern temperate troposphere).

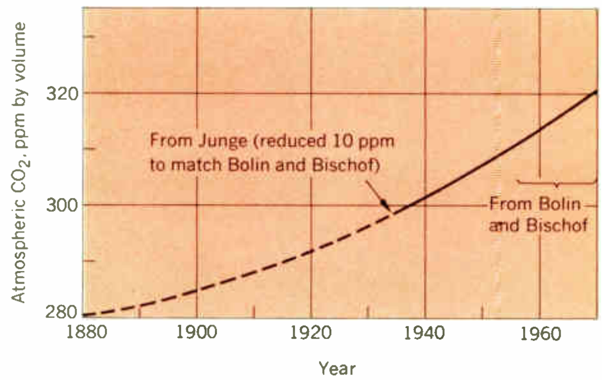
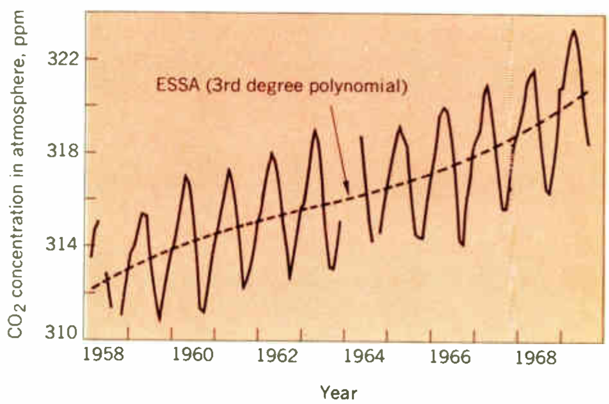


FIGURE 5. Mean monthly atmospheric carbon dioxide concentration at Mauna Loa, Hawaii. (1958–1963 data from Pales and Keeling, J. Geophys. Res., vol. 70, 1965; 1963–1969 data, Bainbridge [Scripps] private communication)



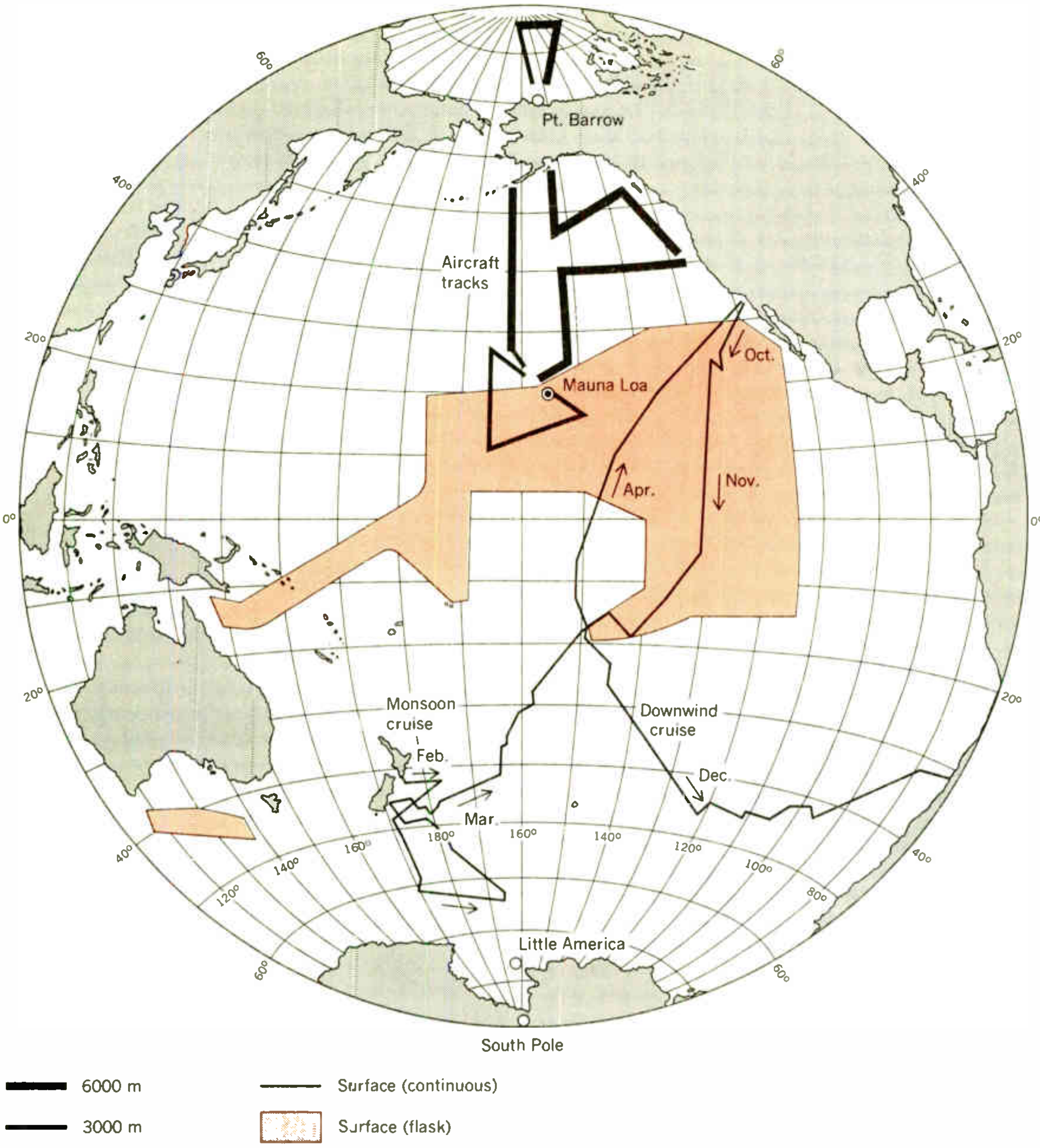
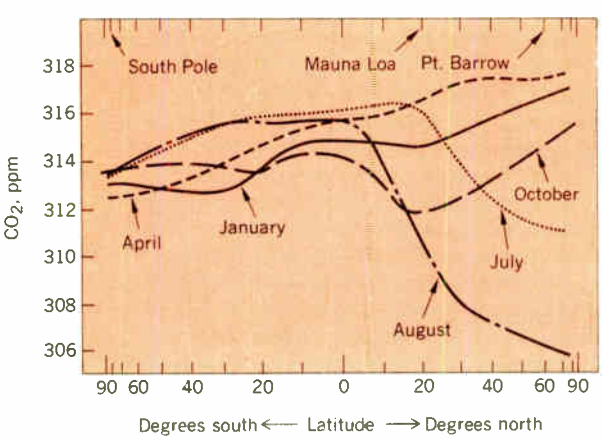


FIGURE 6. Location of stations and tracks for sampling atmospheric CO_2. (From Bolin and Keeling[5])

FIGURE 7. The concentration of atmospheric CO_2 near the earth's surface as a function of latitude in a region near 150°W longitude showing the large changes with season. (From Bolin and Keeling[5])



a significant annual variation in the atmospheric CO_2 levels as noted in Fig. 5. Before attempting to explain these temporal variations, it is well to examine spatial variations (with latitude) for the global area (near 150°W longitude) shown in Fig. 6 as described in detail by Bolin and Keeling.[5]

Figure 7 indicates a trend in January from near 313 ppm at the South Pole to nearly 318 ppm at the North Pole. As the season progresses, the trend smooths out and covers a slightly greater range in April. By July, we see a remarkable change taking place—the concentration from 20°N to the North Pole drops drastically to a low of 311 ppm at Point Barrow. This trend, which begins in June, occurs rapidly and indicates the presence of an active scrubbing mechanism operating during the northern-hemispheric summer. An even greater drop is reached in August when large areas of arctic tundra are active and polar ice cover is low. At this time the Point Barrow values reach a low of 306 ppm. By October, the point of lowest concentration has moved down to 20°N latitude with an appreciable recovery of CO_2 levels at Point Barrow. The trend shifts slowly back to the northern-hemisphere winter curve for January.

These rather large spatial variations emphasize the need for careful measurements when attempting to arrive at temporal variations. It can also be seen from these illustrations that temporal variations will differ greatly at various geographic locations.

Atmospheric injection and scrubbing—the global CO_2 cycle

The exchange of CO_2 between the atmosphere and land biosphere has been described, and it was seen that 1.6 kg of CO_2 is removed from the atmosphere for each kilogram of cellulose produced. Correspondingly, the decay or burning of a kilogram of cellulose returns 1.6 kg of CO_2 to the atmosphere. Moreover, the burning of a fossil-fuel hydrocarbon such as coal, which is nearly 80 percent carbon, returns 3 kg of CO_2 to the atmosphere for each kilogram of coal consumed.

The largest amount of atmospheric CO_2 exchange involves the oceans. Data reported by Sverdrup, Johnson, and Fleming[6] (Fig. 8) show the conditions necessary for exchange of CO_2 between the surface waters of the ocean and the atmosphere. For example, with a typical ocean *p*H of 8.1 and temperatures above 20°C, the flow of CO_2 will be to the atmosphere at atmospheric concentrations of CO_2 near our present 320 ppm by volume. For colder water temperatures, the flow is from the atmosphere to the sea. The ease with which this exchange process takes place has been considered in detail by Kanwisher.[7] He shows, for example, that the net CO_2 flux density across the air–sea interface increases with surface roughness. Combining temperature, wind, and area factors, it appears that a large amount of CO_2 removal occurs in the southern oceans.

A conceptual view of the estimated* global annual exchange of CO_2 between the atmosphere, hydrosphere, and biosphere is drawn in Fig. 9. Primary injection is seen to be from the tropical oceans with a large relatively steady scrubbing by the southern oceans. The tundra and Arctic Ocean scrubbing components—each shown as 0.5×10^{10} and 1.6×10^{10} tonnes of carbon per year; i.e., 7.8×10^{10} tonnes of CO_2—are primarily active in the late summer months since in the winter and early spring much of the far north is covered with ice or snow, as shown in Fig. 10.

The large annual fluctuation of CO_2 levels that we have described at high northern latitudes are now seen to be the result of the varying activity of arctic tundra. The low atmospheric CO_2 levels correspond to the time when the snow and ice cover is gone from arctic lands and when the polar ice cover is at a minimum.

The rather persistent bulge in observed CO_2 at mid-latitudes (from 20°S to 20°N) results from warm ocean waters giving off CO_2 to the atmosphere. The persistent low point in the 40°–50°S latitudes corresponds to a rather effective CO_2 scrubbing by rough cold seas and the

* Preliminary values based on data from Lieth[4] and others and calculations by the author.

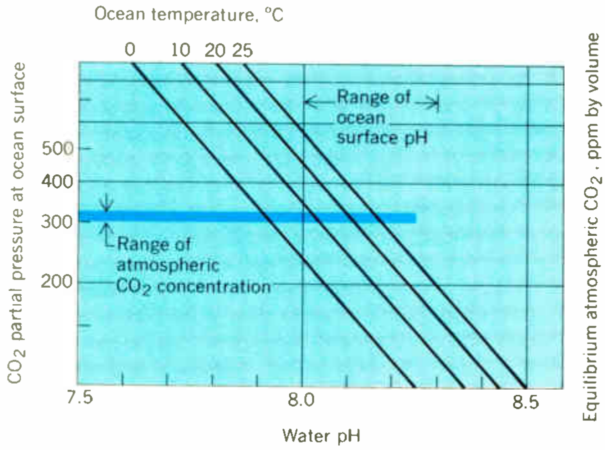
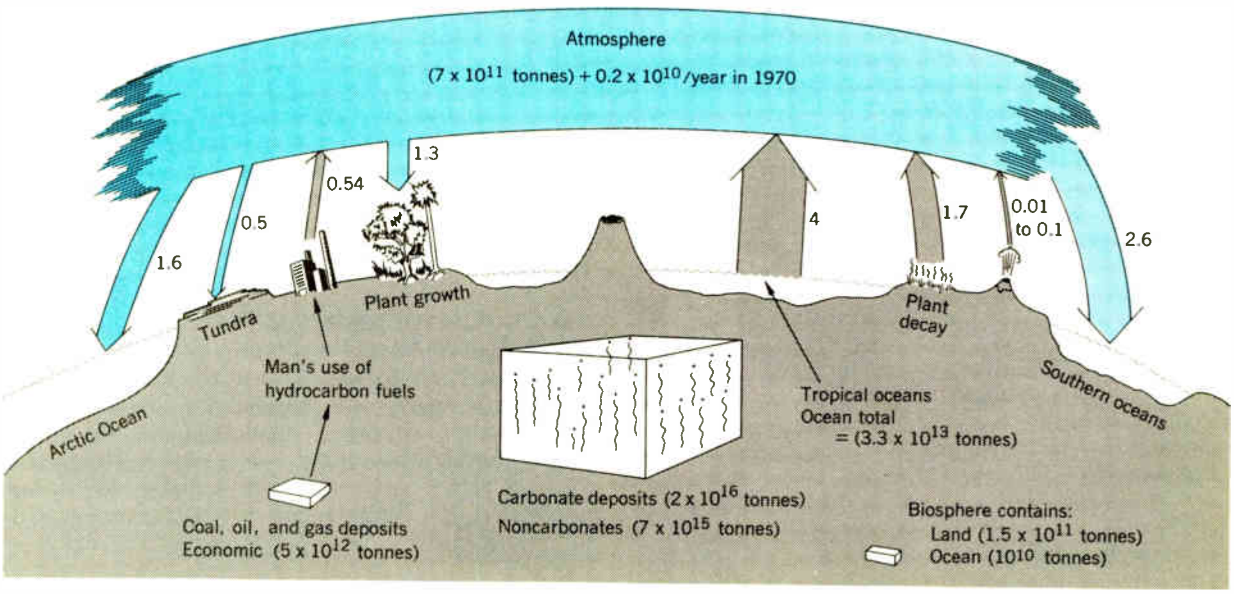
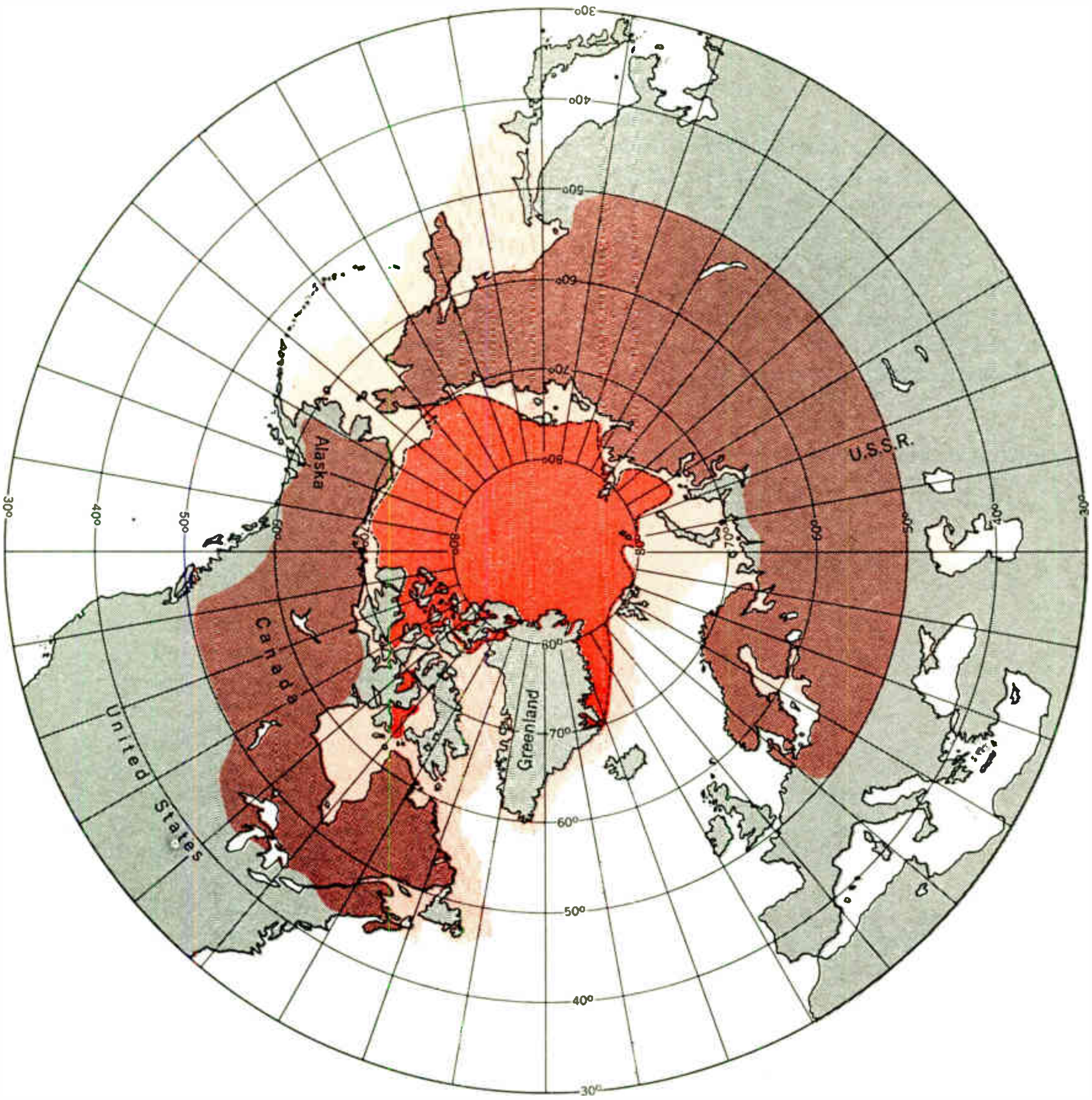


FIGURE 8. CO_2 exchange between ocean and atmosphere. (Based on data from Sverdrup et al.[6])

FIGURE 9. Estimated global deposits and annual exchange of CO_2 between the atmosphere, hydrosphere, and biosphere. Flow values are in units of 10^{10} tonnes/year of carbon; to convert to CO_2, multiply values shown by 3.7.

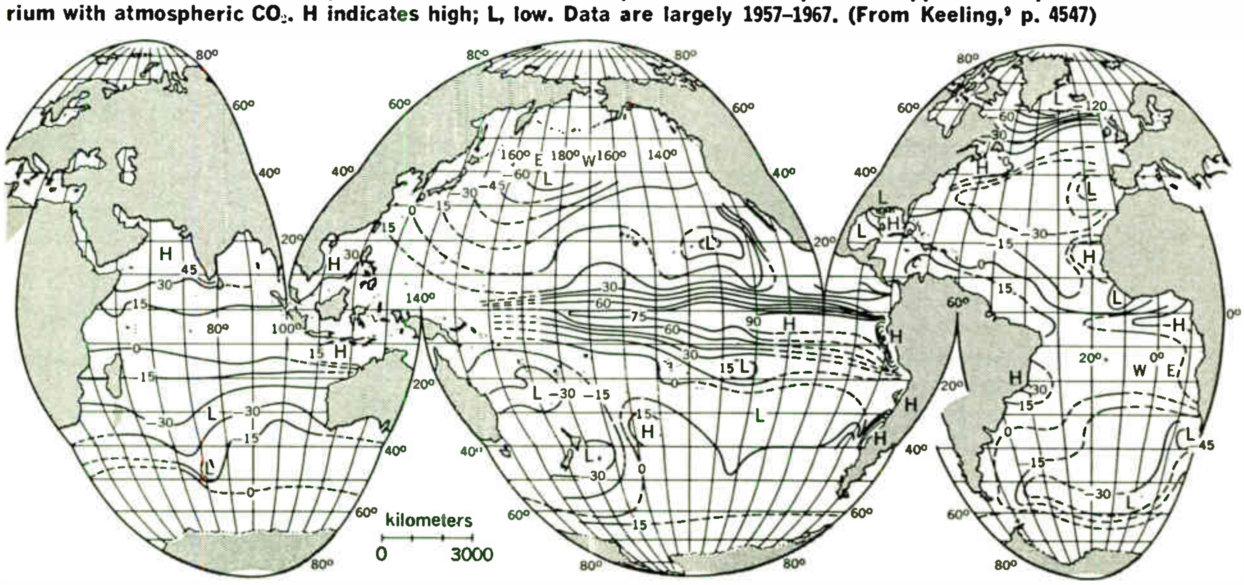




Snow-covered in winter | Usually ice-covered | Usually becomes ice-free for several months

FIGURE 10. Variations in arctic snow and ice cover. (From Sater[8])

FIGURE 11. The distribution of p_{CO_2} of the world's oceans expressed as the departure in ppm from equilibrium with atmospheric CO_2. H indicates high; L, low. Data are largely 1957–1967. (From Keeling,[9] p. 4547)



True distances on mid meridians and parallels 0° to 40°

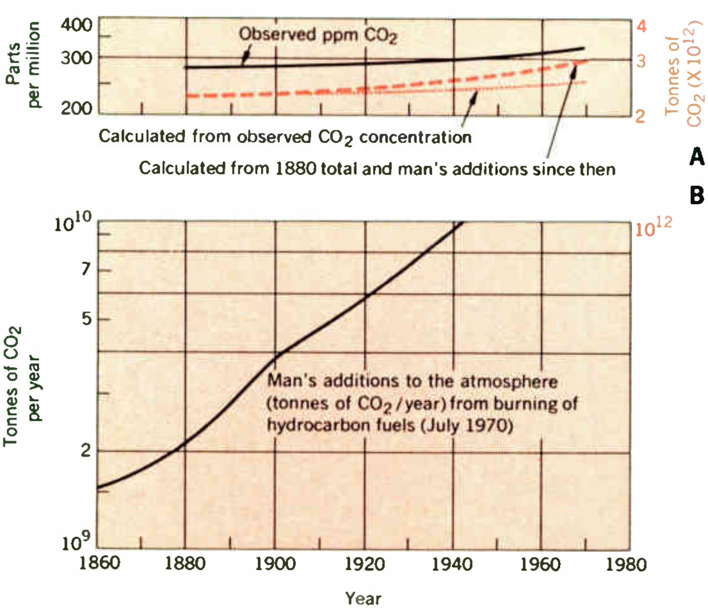
large amount of exposed water in this latitude range.

Some idea of the areas in the ocean that are giving off CO_2 to the atmosphere and those that are taking it from the atmosphere can be seen in Fig. 11. This graph shows the departure in parts per million from equilibrium with atmospheric CO_2. As expected, the high areas are found in the equatorial regions and the lows in the northern and southern oceans. Although there is a considerable amount of data available on CO_2 levels in the atmosphere, as well as partial pressures over surface ocean water, there still are not enough data (pointed out by Keeling[9]) to "clearly establish the oceanic cycle of carbon and the rate of exchange of CO_2 between the atmosphere and the world's oceans." In view of this, we have estimated global gains in the land biosphere and assigned the remainder as the net exchange between the oceans and the atmosphere.

I. Estimated atmospheric CO_2 balance (1970)

	Unit Exchange, 10^{10} tonnes/year	
	C	CO_2
Inputs to atmosphere from		
Tropical oceans	4.00	14.80
Plant decay	1.70	6.30
Man (fuel burning)	0.54	2.00
Volcanic action	0.01 (variable)	0.05
Total:	6.25	23.15
Outputs from atmosphere to		
Northern oceans	1.60	6.00
Southern oceans	2.60	9.60
Arctic tundra	0.50	1.80
Other plants	1.35	5.00
Total:	6.05	22.40
Distribution of man inputs to atmosphere		
Atmosphere gain	0.20	0.75
Biosphere (land) gain		
Lumber	0.02	0.07
Forests, soil, and peat deposits	0.10	0.38
Ocean gain	0.22	0.80
Total:	0.54	2.00

FIGURE 12. Observed atmospheric CO_2 increase compared with calculations of the potential increase due to man's use of hydrocarbon fuels.



We will now consider the overall balance of CO_2 and man's influence. Additions of CO_2 to the atmosphere calculated on the basis of the amount of hydrocarbon fuels consumed are shown in Fig. 12. In 1970, the level exceeded 2×10^{10} tonnes of CO_2 added to the atmosphere per year. This corresponds to the value of 5.4×10^9 tonnes of carbon indicated in Fig. 9. The observed concentration in parts per million by volume for the past 90 years is shown by Fig. 12A. The dashed curve shows the atmospheric level in 1880 plus man's total addition since then. This is given in tonnes of CO_2 in the atmosphere using the right-hand scale. According to these calculations, the present value would stand at 3×10^{12} tonnes if all man's additions remained in the atmosphere. The actual amount of CO_2 in the atmosphere calculated from observed CO_2 concentrations is shown by the dotted curve as being about 2.6×10^{12}.

Over the past decade, the CO_2 level has increased by approximately 1.4 ppm per year by mass. This amounts to a gain of 7.5×10^9 tonnes per year, which is about one third of man's additions. This means that approximately two thirds of man's input to the atmosphere—i.e., 1.25×10^{10} tonnes/yr—is being removed by some scrubbing mechanism. An estimated annual carbon dioxide budget on a global basis is shown in Table I. Direct data are not yet available to show whether the net oceanic or biosphere CO_2 flows are positive or negative. In this table, the best-known values are the atmospheric gain of 7.5×10^9 tonnes of CO_2 per year and the 2×10^{10} tonnes/yr for man's input of CO_2 to the atmosphere.

The world production of timber and forest products, excluding firewood, is estimated at 7×10^8 m^3/yr or near 4.5×10^8 tonnes/yr. Since there are 1.6 tonnes of CO_2 scrubbed per tonne of cellulose produced, this would amount to almost 7×10^8 tonnes of CO_2 removed from the atmosphere. Standing timber is now increasing in the United States and may be increasing on a global basis. Assuming a small amount for forest increase plus an allowance for increasing humus and peat deposits, a total of 4.5×10^9 tonnes of CO_2 per year could easily be accounted for by the land portion of the biosphere.

This would leave about 8×10^9 tonnes of CO_2 to be removed annually via the oceans. The details of the oceanic portion of the carbon cycle, spelled out by MacIntyre,[10] are quite complex. The near-surface balancing equilibrium could be produced by a slight decrease in pH. In actuality, some near-surface CO_2 is removed via plankton production, which may be transferred to calcium carbonate structures. The cold seas near ice packs also remove CO_2 from the atmosphere, and some is transferred to the tropical oceans via the ocean circulation system for use by phytoplankton and for release to the atmosphere. Some carbon in the form of HCO_3^- is transferred to great depths by the annual sinking of high-salinity waters that result from sea-ice production. This can be trapped in deep-ocean clay deposits, which have the ability to absorb vast amounts of bicarbonate ions.

Effects of atmospheric CO_2 levels and temperature on vegetation

In the preceding section, we saw the importance of plant growth in removing CO_2 from the atmosphere, and noted that the exchange process between the ocean and the atmosphere was dependent upon temperature at the atmosphere–ocean interface.

In a similar manner, we shall see that temperature affects the rate of CO_2 assimilation in plant growth. In Fig. 13, this rate for potato leaves in daylight is given as a function of temperature. Two atmospheric concentrations of CO_2 are included. The lower curve is representative of present-day atmospheric conditions and shows, for example, that for these plants a maximum assimilation rate of approximately 0.2 mg of CO_2 per square centimeter of leaf surface per hour occurs at a temperature of 20°C. It would seem from Fig. 13 that, if the atmospheric CO_2 level were increased to 40 times present levels, the plant productivity for this species would increase by four to one.

Figure 14 gives the relative growth of three different plant species as a function of the partial pressure of CO_2 in millibars. Since there are 505 ppm by weight of CO_2 in the atmosphere, the present partial pressure is 0.5 mbar on this illustration. From these results, it appears: (1) that, at present atmospheric levels, plant uptake increases almost directly with CO_2 concentration, and (2) for temperatures below 20°C, CO_2 uptake increases with temperature.

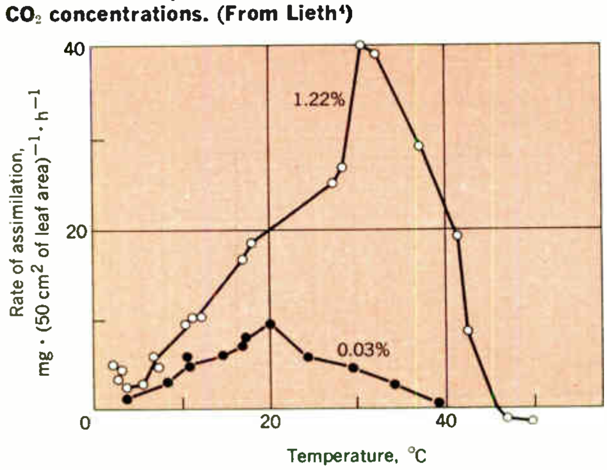
It can be surmised that if increasing levels of atmospheric CO_2 produce an increase in global temperature, more effective CO_2 assimilation by plants will have a self-regulating effect upon global temperature. Even if temperature effects are small, as will be indicated, increased plant activity as a function of CO_2 level should act as a regulator that will tend to stabilize the level of atmospheric CO_2.

The effect of clouds and rain on plant productivity is demonstrated in Fig. 15. The amount of CO_2 assimilated by plants on clear days is much greater than that assimilated during a rainy day. Another important factor is that, as the sun sets, plant respiration starts returning CO_2 to the atmosphere, as evidenced by the negative assimilation value shown on the lower-right-hand side of Fig. 15.

The effect of plant activity on the near-surface-level concentrations of CO_2 in the atmosphere is shown in Fig. 16. In the early afternoon, atmospheric CO_2 levels measured near the ground are much lower than those at night in vegetated areas. The rapid decrease of CO_2 at low heights during early morning hours is due to the onset of thermal convection of air. The large variations of from 300 ppm during the day to 400 ppm at night show that care must be taken in measuring procedures to allow for this effect when attempting to determine either long-term trends or spatial variations in CO_2 levels.

The actual productivity of land areas varies greatly

FIGURE 13. Rate of CO_2 assimilation in potato leaves as a function of temperature in full daylight at two different CO_2 concentrations. (From Lieth[4])



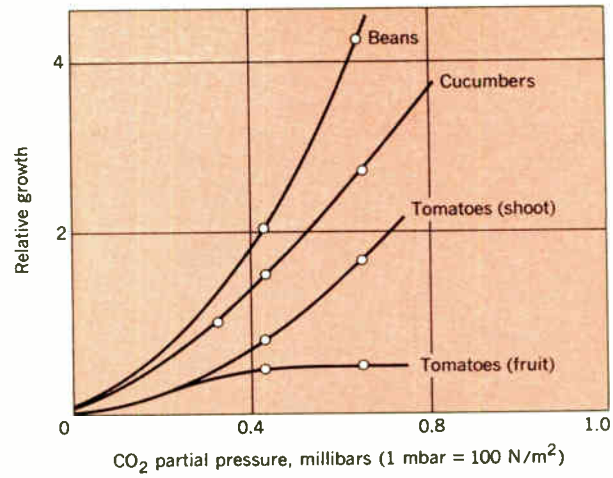


FIGURE 14. Rate of growth of various domestic plants as a function of the CO_2 concentration of the air. (From Lieth[4])

FIGURE 15. The variation of net CO_2 assimilation of corn in a field during both a clear and a rainy day, including soil respiration rate. (From Lieth[4])

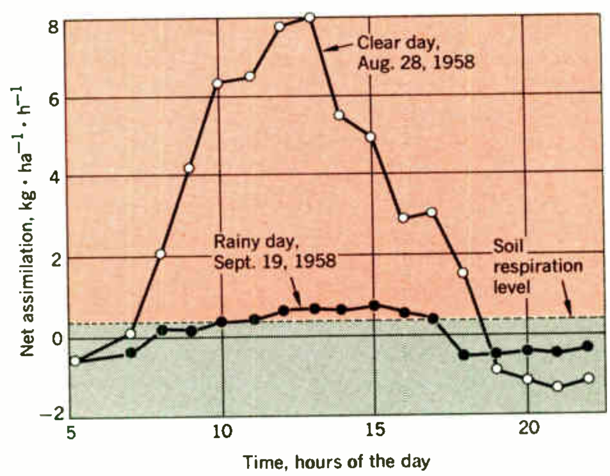
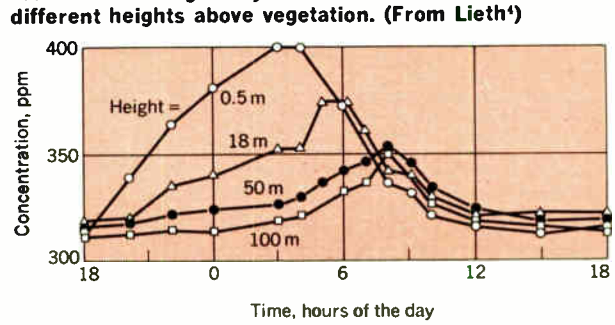


FIGURE 16. Average daily variation of CO_2 concentration at different heights above vegetation. (From Lieth[4])



II. Potential productivity of the earth and the population it could support*

North Latitude, degrees	Land Surface, ha × 10^8	Number Months Above 10°C	Carbohydrates, $Mg \cdot ha^{-1} \cdot yr^{-1}$	Square Meters per Man to Support Life: With No Allowance for Urban and Recreational Needs: m^2/man	No. Men × 10^9	With 750 m^2/man for Urban and Recreational Needs: m^2/man	No. Men × 10^9	Percentage Agricultural Land
Column 1	2	3	4	5	6	7	8	9
70	8	1	12	806	10	1556	5	52
60	14	2	21	469	30	1219	11	38
50	16	6	59	169	95	919	17	18
40	15	9	91	110	136	860	18	13
30	17	11	113	89	151	839	20	11
20	13	12	124	81	105	831	16	10
10	10	12	124	81	77	831	11	10
0	14	12	116	86	121	836	17	10
−10	7	12	117	85	87	835	9	10
−20	9	12	123	81	112	831	11	10
−30	7	12	121	83	88	833	9	10
−40	1	8	89	113	9	863	1	14
−50	1	1	12	833	1	1583	1	53
Total:	132				1022		146	

* From deWit,[11] p. 317.

with incident solar energy, soil moisture, and soil nutrients. Lieth[4] has shown, for example, that a typical European forest will yield a yearly production of 10 tonnes of dry material per hectare (1 ha = 10^4 m^2), which results in a scrubbing of 15 tonnes of CO_2 from the atmosphere. In a Malayan forest, yearly production is about 25 tonnes/ha, resulting in a scrubbing of 38 tonnes of CO_2.

DeWit[11] has calculated the potential food productivity of the earth assuming that the only limiting factor is solar energy. Table II shows the land surface available in each 10° latitude band, and the number of months when the temperature is about 10°C. The maximum potential productivity is listed in column 4 in tonnes of carbohydrates per hectare per year. It is interesting to note that maximum productivity is indicated in the 20° latitude regions. This effect in the real world is even more pronounced since in many tropical regions the soils have been heavily leached so that near-surface minerals are not available and as a result productivity, particularly in terms of agriculture, is often very low. The difference between the values quoted in column 4 by deWit[11] and those of Lieth[4] indicate that real-world productivity is perhaps one fourth to one fifth the idealized productivity based on solar energy alone. For many land areas the primary limitation is availability of water.

The world's oceans have 70 percent of the global area and could potentially support a large amount of biological activity and resulting CO_2 scrubbing. In actuality, as shown by Isaacs,[12] large areas of the oceans have low productivity. This results when the near-surface waters (through which the sun's rays can penetrate) are deficient in essential nutrients such as nitrogen and phosphorus. The areas of high productivity are in the upwelling regions shown in Fig. 17, where nutrients are brought to the surface to replace those consumed by the phytoplankton. The heavily shaded areas, which amount to approximately 10 percent of the ocean area, are believed to be the major contributors to the estimated total annual oceanic biological uptake of 1.5×10^{10} tonnes of carbon,[13] which is equivalent to 5.5×10^{10} tonnes of CO_2 per year. Ocean productivity in some northern-latitude regions has been shown to have major variation throughout the year.[14] Land productivity at high latitudes is, of course, extremely variable from winter to summer.

Probable climatic effects of atmospheric CO_2

There is still considerable disagreement concerning the effect of atmospheric CO_2 on global climates. There is a prevalent misconception that atmospheric CO_2 acts as a blanket that rather uniformly turns back the infrared energy radiated from the earth and thereby increases the earth's temperature. In actuality, CO_2 is effective only in absorbing energy in several very limited portions of the earth's radiated frequency spectrum.

The energy spectrums involved are displayed in Fig. 18. Most of the energy received from the sun is in the region from 0.5 to 5 μm, whereas the major portion of energy radiated from the earth to space is in the 5–30-μm region. Both of these curves are for unattenuated radiation. In the first case, it represents energy from the sun before it encounters the earth's atmosphere, and in the second case it represents electromagnetic energy from the earth at the earth's surface before it passes through the atmosphere on its way to space. Both of these spectral distributions are modified considerably in passing through the atmosphere.

There are numerous atmospheric constituents that influence the radiant energy on its journey through the atmosphere, but those of primary concern are water vapor, carbon dioxide, and ozone. These triatomic molecules, which are very effective in absorbing certain portions of the radiation spectrum, are found in varying degrees at different locations in the atmosphere. Water vapor is concentrated in the several lower kilometers, carbon dioxide is rather uniformly distributed through-

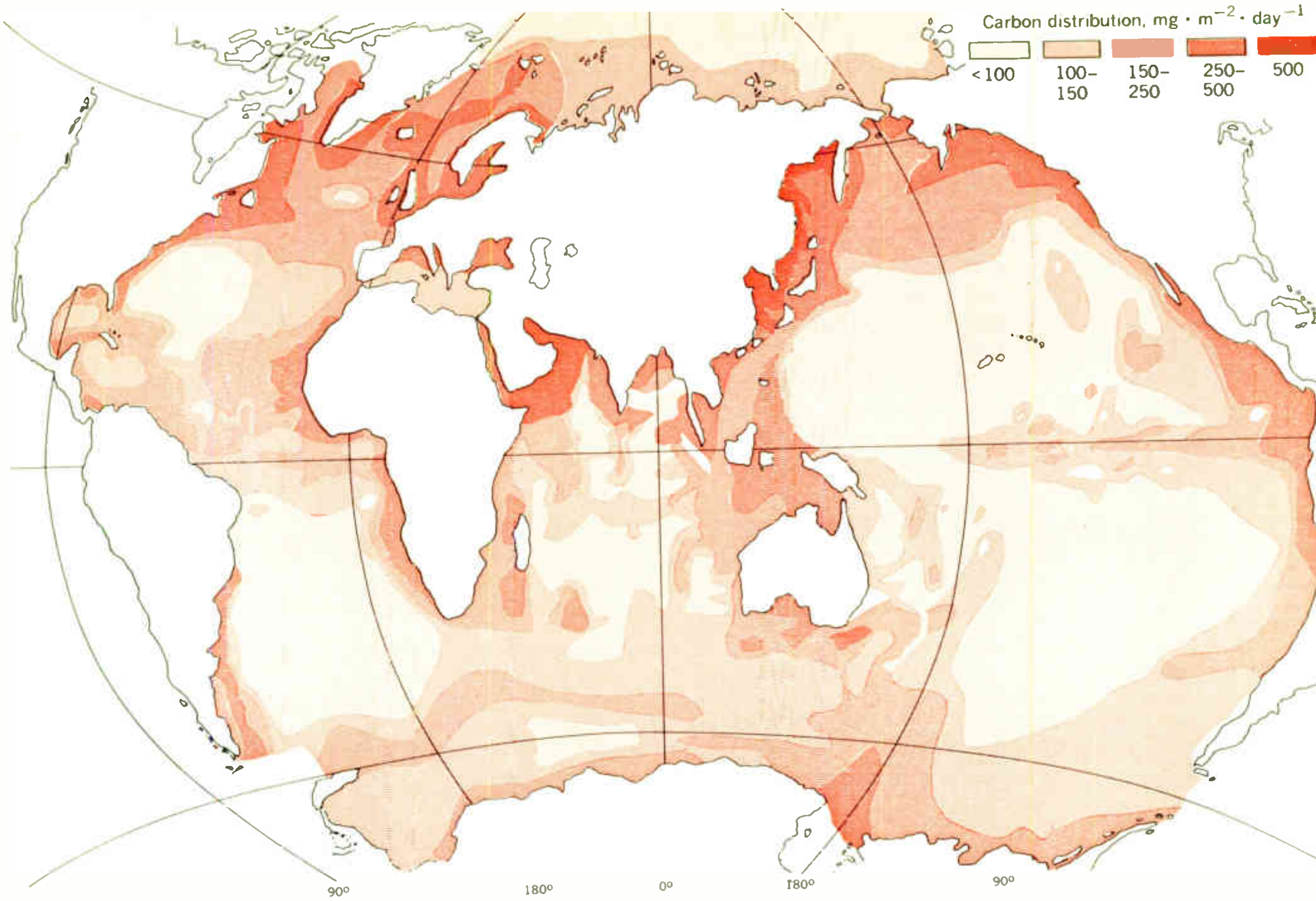
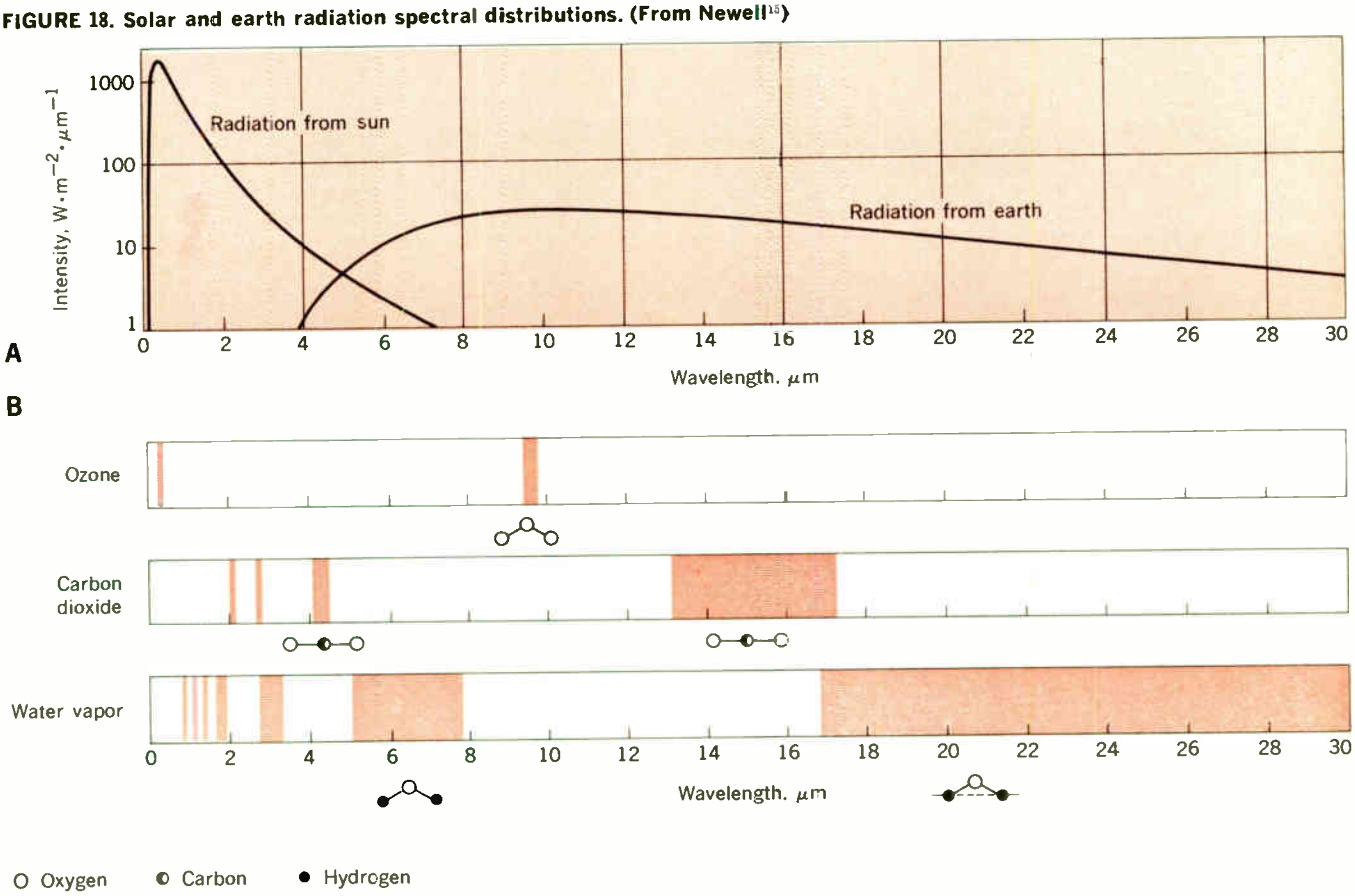


FIGURE 17. Distribution of primary biological production in the world's oceans. (From Koblentz-Mishke et al.[13])

FIGURE 18. Solar and earth radiation spectral distributions. (From Newell[15])



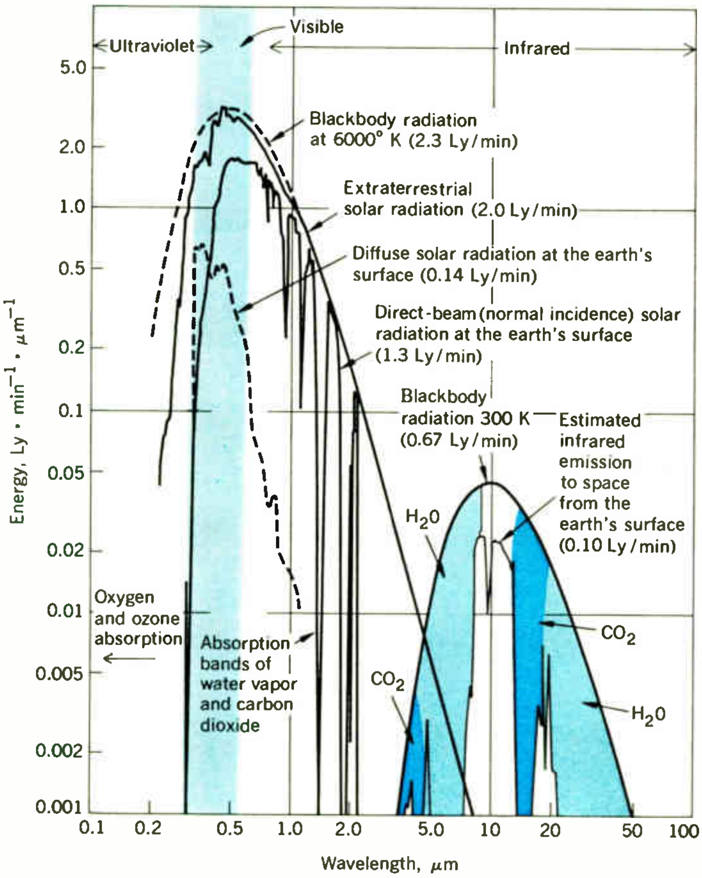
out most of the atmospheric layer, and ozone is largely concentrated in a layer from 20 to 60 km above the earth's surface.

The wavelengths at which each of these gases is effective in absorbing energy are shown in Fig. 18B as shaded bands. The width of these absorption bands is dependent upon the amount of material in the path traversed and also to a great degree upon atmospheric pressure. The so-called "pressure broadening" of spectral absorption bands can be explained as follows: First an absorption spectrum is measured with a particular amount of carbon dioxide. Then an additional amount of gas having no absorption in the region measured is added, which increases the total pressure. The absorption spectrum is again measured and seen to be appreciably broader. Details of such pressure broadening have been described by Plass.[16]

The incoming solar-energy spectrum as observed at the earth's surface and an estimated outgoing energy spectrum for the earth's radiation observed beyond the atmosphere are both shown in Fig. 19. The heavily tinted portions of the outgoing spectrum show where effective CO_2 absorption is occurring. The amount of energy absorbed by CO_2 is seen to be considerably smaller than the energy absorbed by water vapor. The CO_2 absorption bands at 4 and 15 μm are quite deep, and it will be shown later than even doubling the atmospheric CO_2 will only slightly increase the total portion of energy absorbed.

Relative to the heating of the atmosphere, it can be seen in Fig. 19 that a large portion of solar energy is absorbed by ozone and oxygen in the region below 0.3 μm. Theoretical calculations by Pressman[18] have shown that ultraviolet ozone absorption of solar radiation is the main radiative heat source of the atmosphere in the region from 30 to 70 km. Heating rates quoted are in the range of

FIGURE 19. Electromagnetic spectrums of solar and terrestrial radiation. Note that a langley (Ly) is a measure of solar energy density and equals 1 cal/cm² of irradiated surface. (Adapted from Sellers,[17] p. 20)



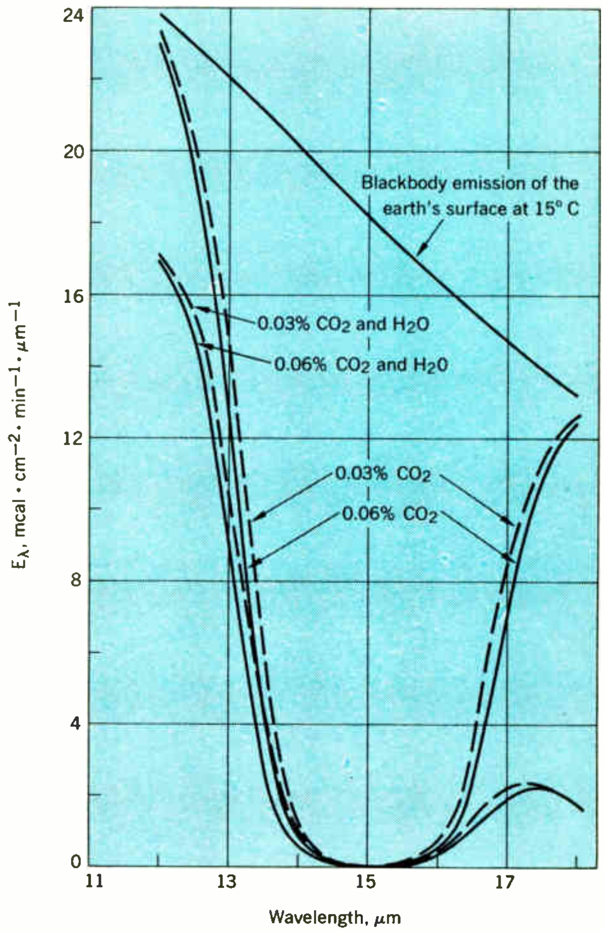
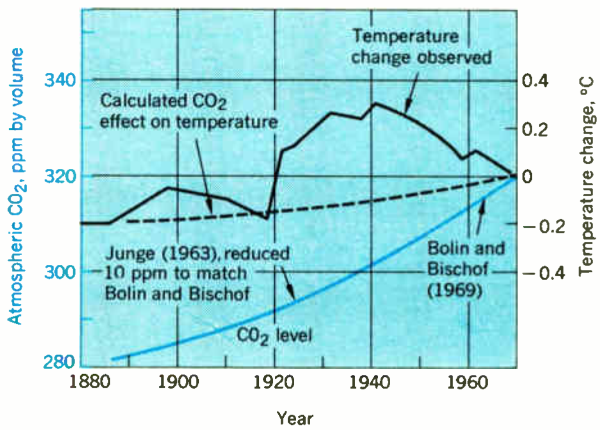


FIGURE 20. Outgoing effective radiation from the earth's surface for two CO_2 concentrations. (From Möller,[21] p. 3879)

FIGURE 21. Trend of atmospheric CO_2 (mean annual data taken from northern temperate troposphere).



1–10 K per day with a maximum heating usually at a height of 45 km. Below 30 km, the near-infrared bands of water vapor are the primary absorbers of solar radiation. The heating rates at these altitudes of the lower stratosphere may be as large as 4 K/day.

Plass[19,20] has shown that, in addition to the heating effects of CO_2 and O_3 in the atmosphere, the stratosphere is cooled by infrared radiation to space from CO_2 and O_3. The 15-μm CO_2 band is calculated to have a maximum cooling rate of 5 K/day at 45 km for typical conditions and, similarly, the 9.6-μm O_3 band has a maximum cooling rate of 2 K/day at 45 km. Under normal conditions, the sum of these cooling rates nearly balances the heating rates for solar radiation. This would indicate that, normally, the stratosphere is approximately in radiative equilibrium.

Some concept of the magnitude of change in net outgoing radiation from the earth's surface, with variations of CO_2 from 0.03 to 0.06 percent by volume (near present levels), can be seen in Fig 20. It is apparent that even doubling the atmospheric CO_2 levels has only a minor broadening effect on the primary absorption region at 15 μm. The net effect has been discussed by Möller,[21] and he comes to the conclusion that, "thus the theory that climatic variations are affected by variations in the CO_2 content becomes very questionable." At this point one may ask, "But what of the reported good correlation between increasing atmospheric CO_2 and increasing global temperatures?" Figure 21 shows average CO_2 concentrations with a superimposed world mean-temperature trend obtained from Mitchell.[22] It is clearly apparent that, although atmospheric CO_2 levels are still climbing, global temperatures have been dropping since 1941. Obviously, global temperatures do not go up directly with atmospheric CO_2. As a result, we must conclude that other factors in the global heat balance—such as atmospheric dust—are of greater importance and therefore have overridden the smaller, heat-capturing contribution of atmospheric CO_2.

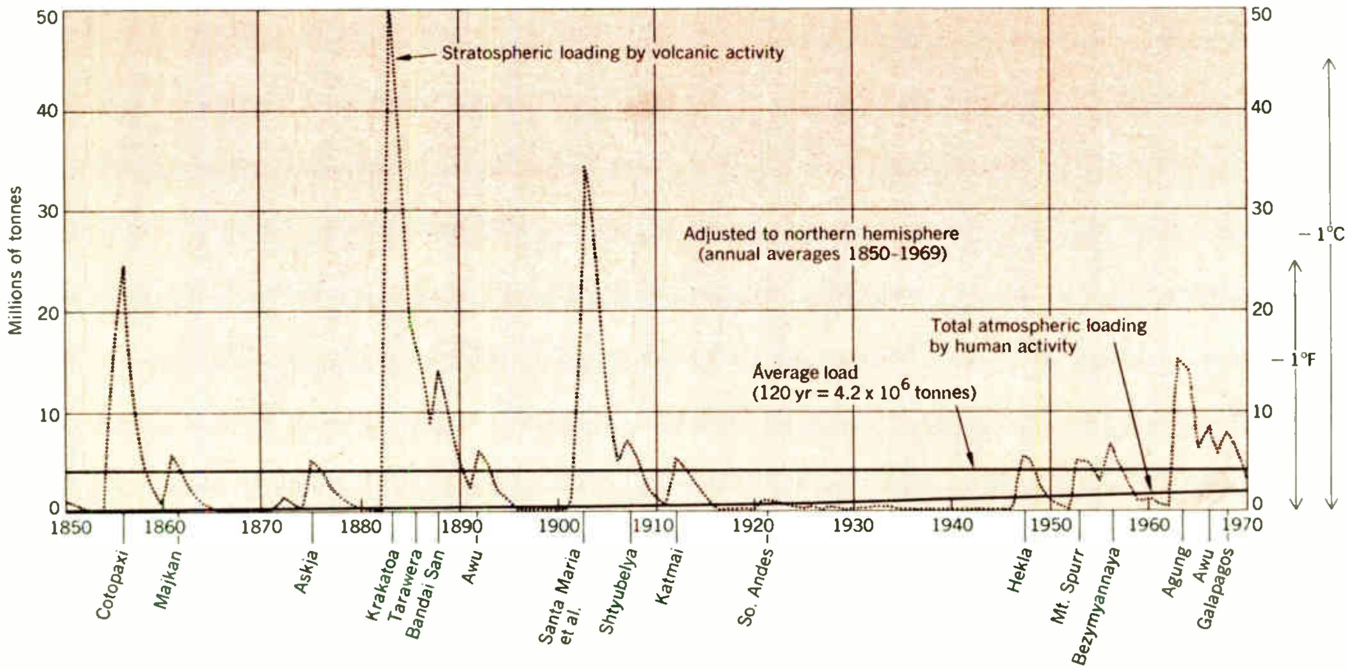
Calculations by Manabe and Wetherald[23] have indicated that a two-to-one increase in atmospheric CO_2 would produce a 1.5°C temperature increase at the earth's surface. With this, one can expect a stratospheric temperature decrease of 10°C at an altitude of 40 km. Using the observed 40-ppm increase in CO_2 levels from 1880 to 1970, an expected increase of 0.2°C is obtained by interpolation of Manabe and Wetherald's results; such a change is shown by the dashed curve of Fig. 21.

Atmospheric dust loading varies appreciably as indicated in Fig. 22. It is important to note that most of the long-residence atmospheric dust particles are in the range of 1 μm in diameter. Such small particles are effective in reflecting back to space the short-wavelength energy from the sun, but they are relatively transparent to the long-wave radiation from the earth. Some of the incoming solar energy is absorbed by the dust particles and some is reflected back to space. For a given tonnage, the smaller particles near 1 μm in diameter are more effective in intercepting solar radiation than are large particles. The height of the dust layer is also *very* important in determining its effect on surface temperature.[24]

Preliminary calculations by the author have indicated possible temperature changes derived from atmospheric dust to be somewhat larger than the reference −1°C shown by Mitchell on Fig. 22. Calculated global temperature changes are of the order of −0.03°C per 10^6 tonnes of high-altitude atmospheric dust in the 1-μm diameter range. Temperature records at numerous locations have shown actual drops in the annual temperature of 2–3°C following large volcanic eruptions.

The earth's temperature-regulating mechanism involves a number of beautifully interrelated mechanisms that are still not completely understood. It appears, however, that atmospheric CO_2 variations such as those

FIGURE 22. Estimated chronology of atmospheric dustload. (From Mitchell[22])



occurring at present are likely to have only a minor effect on the overall global temperature.

Need for a survey article of this type was first suggested to the author by Dr. J. H. Wright. Helpful discussions have been held with H. A. Gunther, F. S. Mathews, D. B. Large, and J. R. Portman. The preparation of the manuscript by Winifred Werth, illustrations by Gary Uridil, and a review of the text by Bob Hill are also gratefully acknowledged.

REFERENCES

1. Peterson, E. K., "Carbon dioxide affects global ecology," *Environ. Sci. Technol.*, vol. 3, pp. 1162–1169, Nov. 1969.
2. Smith, R. L., *Ecology and Field Biology*. New York: Harper & Row, 1966.
3. Keeton, W. T., *Biological Science*. New York: Norton, 1967.
4. Lieth, H., "The role of vegetation in the carbon dioxide content of the atmosphere," *J. Geophys. Res.*, vol. 68, July 1, 1963.
5. Bolin, B., and Keeling, C. D., "Large-scale atmospheric mixing as deduced from the seasonal and meridional variations of carbon dioxide," *J. Geophys. Res.*, vol. 68, July 1, 1963.
6. Sverdrup, H. V., Johnson, M. W., and Fleming, R. H., *The Oceans*. Englewood Cliffs, N.J.: Prentice-Hall, 1942.
7. Kanwisher, J., "The effect of wind on CO_2 exchange across the sea surface," *J. Geophys. Res.*, vol. 68, July 1, 1963.
8. Sater, J. E., (ed.) "The arctic basin," Arctic Institute of North America, Aug. 1969.
9. Keeling, C. D., "Carbon dioxide in surface ocean waters," *J. Geophys. Res.*, vol. 73, July 15, 1968.
10. MacIntyre, F., "Why the sea is salt," *Sci. Am.*, vol. 223, pp. 104–115, Nov. 1970.
11. deWit, C. T., "Photosynthesis: Its relationship to overpopulation," in *Harvesting the Sun*, A. San Pietro, ed. New York: Academic, 1967.
12. Isaacs, J. D., "The nature of oceanic life," in *The Ocean*. San Francisco: Freeman, 1969.
13. Koblentz-Mishke, O. J., Volkovinsky, V. V., and Kabanova, J. G., "Plankton, primary production of the world ocean," in *Scientific Exploration of the South Pacific*. Publication SBN 309-01755-6, Scientific Committee on Oceanic Research, National Academy of Sciences–National Research Council, Washington, D.C., 1970.
14. Parsons, T. R., and Anderson, G. C., "Large scale studies of primary production in the North Pacific Ocean," *Deep-Sea Res.*, vol. 17, pp. 765–776, Aug. 1970.
15. Newell, R. E., "The circulation of the upper atmosphere," *Sci. Am.*, vol. 210, Mar. 1964.
16. Plass, G. N., "Models for spectral band absorption," *J. Opt. Soc. Am.*, vol. 48, p. 690, 1958.
17. Sellers, W. D., *Physical Climatology*. Chicago: University of Chicago Press, 1965.
18. Pressman, J., "Seasonal and latitudinal temperature changes in the ozonosphere," *J. Meteorol.*, vol. 12, p. 87, 1955.
19. Plass, G. N., "The influence of the 9.6 micron ozone band on the atmospheric infra-red cooling rate," *Quart. J. Roy. Meteorol. Soc.*, vol. 82, p. 30–44, 1956.
20. Plass, G. N., "The infrared radiation flux in the atmosphere," *Proc. IRE*, vol. 47, pp. 1448–1451, Sept. 1959.
21. Möller, F., "On the influence of changes in the CO_2 concentration in air on the radiation balance of the earth's surface and on the climate," *J. Geophys. Res.*, vol. 68, July 1, 1963.
22. Mitchell, J. M., "The effect of man's activities on climate," presented at AGU Symp. on The Environmental Challenge, Washington, D.C., April 22, 1970; also "A preliminary evaluation of atmospheric pollution as a cause of the global temperature fluctuation of the past century," in *Global Effects of Environmental Pollution*, Singer, ed. New York: Springer-Verlag, 1970.
23. Manabe, S., and Wetherald, R. T., "Thermal equilibrium of the atmosphere with a given distribution of relative humidity," *J. Atmospheric Sci.*, vol. 24, pp. 241–259, May 1967.
24. Mitchell, J. M., "The effect of atmospheric aerosol on climate with special reference to surface temperature," *J. Appl. Meteorol.* (to be published).

BIBLIOGRAPHY

Bolin, B., and Bischof, W., "Variations of the carbon dioxide content of the atmosphere," Report AC-2, Institute of Meteorology/International Meteorological Institute, Stockholm, Sweden, Dec. 1969, p. 29.

Eriksson, E., "Possible fluctuations in atmospheric carbon dioxide due to changes in the properties of the sea," *J. Geophys. Res.*, vol. 68, July 1, 1963.

Eriksson, E., "The yearly circulation of sulfur in nature," *J. Geophys. Res.*, vol. 68, July 1, 1963.

Georgii, H.-W., "Oxides of nitrogen and ammonia in the atmosphere," *J. Geophys. Res.*, vol. 68, July 1, 1963.

Junge, C. E., *Air Chemistry and Radioactivity*. New York: Academic, 1963.

Kelley, J. J., Jr., "Carbon dioxide in the surface water of the North Atlantic ocean," *Limnol. Oceanog.*, vol. 15, pp. 80–87, Jan. 1970.

Matsushima, S., Discussion of "The vertical distribution of dust to 30 kilometers," by J. M. Rosen, *J. Geophys. Res. (Space Phys.)*, vol. 73, May 1, 1968.

Rosen, J. M., "The vertical distribution of dust to 30 kilometers," *J. Geophys. Res.*, vol. 69, pp. 4673–4676, Nov. 1, 1964.

Rosen, J. M., Reply to S. Matsushima, *J. Geophys. Res. (Space Phys.)*, vol. 73, pp. 3088–3089, May 1, 1968.

Arthur D. Watt (SM) received the B.S.E.E. degree from Purdue University in 1942, and has pursued further studies that include graduate studies at the University of Maryland (1945–1947), a special summer course on mathematical problems of communications theory (1953), and a special course on geochemical exploration (1967). As a project engineer with the U.S. Naval Research Laboratory from 1945 to 1951, he headed the Graphic Communications Group investigating communication design problems and theory, as well as methods of transmitting graphical material. While with the U.S. National Bureau of Standards, he was a project leader of modulation studies and VLF automatic communications projects, appointed chief of the Modulation Systems Research Section in 1956, and later promoted to assistant chief of the Radio Communications and Systems Division, Central Radio Propagation Laboratory, Boulder, Colo. As vice president of DECO Electronics, Inc., and director of the newly formed Boulder Division, he set up and directed basic and applied research in the fields of communications, electronics, and geophysics during the period 1960–1964. As chief scientist at DECO, and later as chief scientist of the Westinghouse Environmental Science and Technology Department, he directed work on many projects in the fields of communications, navigation, and geophysics. He has also directed long-range R&D on optimization of the Navy's VLF communication system, and has served as a member of the Omega Implementation Committee since 1963.

At present, Mr. Watt is a consulting scientist at the Westinghouse Georesearch Laboratory, Boulder, Colo., where he has been engaged in long-range planning of geophysical and earth science projects. Recently, he has been conducting studies on the influence of energy resources and environmental factors on electric power systems and other energy-using segments in the United States.

Mr. Watt is a member of the AGU, URSI, CCIR, RESA, the Omega Navigation System Committee, and Eta Kappa Nu. In addition, he has served on the IEEE Antennas and Propagation Committee, and is a professional engineer in Colorado.

Systems approach toward nationwide air-pollution control

II. The technical requirements

An effective system of pollution control must not only overcome technological problems—there are formidable political, legal, and economic obstacles as well

Robert J. Bibbero Honeywell Inc.

The general configuration for air-pollution control systems that was outlined in Part I will impose certain requirements upon system components—including hardware (sensors, data processors, communications), software (air quality laws, emission regulations, data processing, simulation programs), and of course the organization and the people who staff it. Most important of all, and surely the most costly, will be the gas-cleaning equipment, process controls, and operating changes that are required of industry, institutions, and private individuals. The means of implementing the systems functions described in Part I (Oct. IEEE Spectrum) will be considered here.

In operating an air-pollution control system, subsystem monitoring of current pollution levels must conform to temporal and spatial factors, which are established by the dose–effect criterions of pollution damage on the one hand, and by the requirements for area coverage and minimum measurement reliability levels on the other. The prediction subsystem, drawing on the same data supplemented by meteorological inputs, will be required to furnish both long-term future estimates and detailed short-term forecasts, but not necessarily with the same equipment. Forecasting will utilize up-to-date information on source emissions, applied to detailed mathematical models of the chemical and physical behavior of polluted air and the micrometeorology of cities. Unfortunately, much of the information needed to create these tools is incomplete, and questions may be raised as to our ability to handle the necessary simulation problems in a timely and economical manner.

The "business end" of pollution control consists of the codes and social organization that, in the long term, permit pollution-emitting sources to operate in the first place and police them after they are established. These can be termed the "prior" and "posterior" controls, depending on whether they are implemented before or after the source is created. Superimposed on these day-by-day regulations are the short-term emergency controls that are imposed only during dangerous episodes of abnormally intense pollution—largely a weather phenomenon. Since the lives of thousands of people may be endangered, these controls are of an entirely different magnitude of stringency and reaction time.

All of these considerations will establish the disciplines to which we must all conform and the price we must pay if we are to achieve and continue to enjoy the benefits of clean air.

Current level monitoring

The primary system function is to monitor the air quality in a specific locality. The first question is, "What pollutants will be sensed?" This will be determined by both national and local factors. Nationally, the state of the art with respect to the effects of pollutants on health and material objects is established by federal criterions. Locally, these criterions will be modified to suit conditions; all pollutants that are the subject of criterions will not be equal hazards everywhere. At present, criterions have been issued by the federal government through the Air Pollution Control Office (APCO) only for sulfur oxides, nitrogen oxides, particulates, photochemical oxidants, carbon monoxide, and hydrocarbons, but a total of 14 others have been scheduled over the next four years.

These criterions not only establish the type of pollutant, but the significant concentration level or, more correctly, the concentration and time duration. Current criterions contain a number of statements relating effects to time and concentration (see Table I; criterions derived from the 1970 Clean Air Act Amendments are described in the *Federal Register*, vol. 36, part 2, p. 8187, April 30, 1971, and on pages 27 and 28, Table III, in Part I of the present article), which can be expressed as points on a two-dimensional plot,[1] as in Fig. 1. A line fitted through these points crudely separates the space into "good" and "bad" air quality regions, representing one form of standard for air quality. A standard of this type is highly nonlinear as a consequence of its mixed origins, and is not very amenable to analysis.

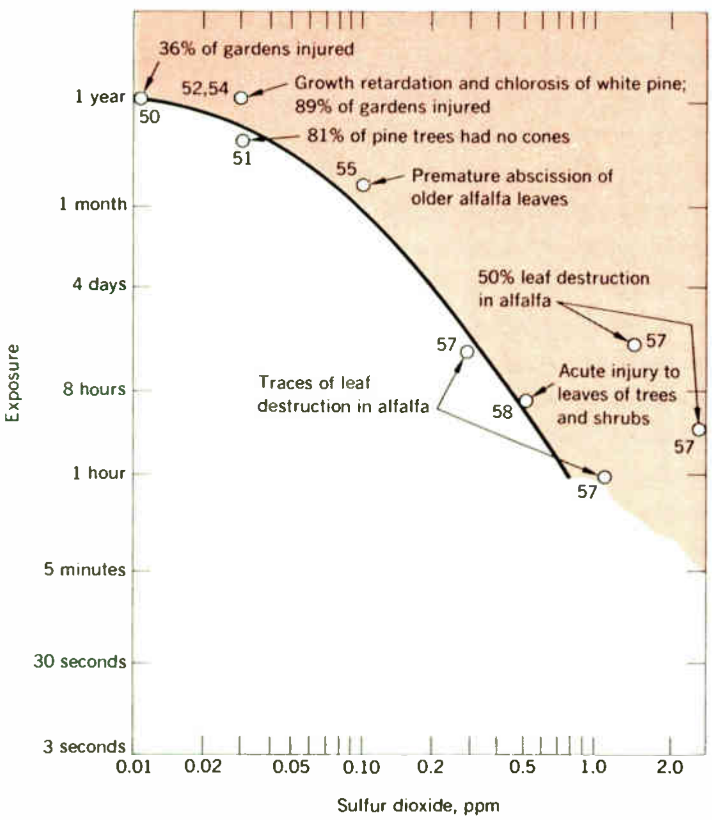
However, the introduction of time as a dimension of the damage criterion limits the permissible reaction time of the monitoring system instrumentation. Some traditional air-monitoring instruments are relatively insensitive and require a long operating period for each independent reading. One that is commonly used has a 20-minute time lag, whereas another requires an hour per reading. These instruments would be clearly unsuitable for monitoring a concentration level corresponding to damage within one hour or less, although they could be used satisfactorily at lower levels that are harmless unless sustained for a day or week. For timely warning of higher concentrations of pollutants such as carbon monoxide or oxidants, it is necessary to have rapid measuring instruments and any necessary averaging in the data-processing stage, a capability that must also be provided.

I. Minimum concentration–duration conditions defining pollution effects in current federal criterions*

Pollutant	Concentration	Averaging Time	Duration	Item Affected	Primary† Federal Standard
Sulfur dioxide	0.11 ppm	24 hours	3–4 days	health	0.14 ppm (24 hours)
	0.04 ppm	annual mean		health	0.03 ppm (ann. mean)
	0.03 ppm	annual mean		vegetation	
Particulates	300 μg/m³	24 hours	1 day	health‡	75 μg/m³
	80 μg/m³	annual mean		health‡	260 μg/m³ (24 hours)
	60 μg/m³	annual mean		materials‡	
Photochemical oxidants	0.03–0.3 ppm	1 hour	1 hour	health	0.08 ppm (1 hour)
	0.1 ppm	peak		health (eye irritation)	
	0.5 ppm	—	4 hours	vegetation	
Carbon Monoxide	10–15 ppm	—	8 hours	health	35 ppm (1 hour)
	8–14 ppm	week	week	health	9 ppm (8 hours)
Hydrocarbons (nonmethane)	0.15§	3 hours	6.00–9.00 a.m.	health	0.24 ppm (3 hours, 6–9 a.m.)

* Adapted from pertinent NAPCA "Air Quality Criteria," Publications AP 49, 50, 62–64.
† As per Clean Air Act Amendments (1970).
‡ With approximately 0.22 ppm SO_2.
§ Parts per million by volume of C (with photochemical oxidants).

FIGURE 1. Sulfur dioxide damage to vegetation.[1] (Courtesy of Academic Press)



Point and long-path sensors

In the design of systems to monitor current atmospheric pollution levels, one critical decision will be whether to employ "point" or "long-path" sensors. The former, as implied by the name, measure the concentration of the given pollutant at a single point in the area monitored and are the only type routinely employed today. A long-path (or open-path) instrument would measure the average concentration between two points separated by a considerable distance, perhaps one or more kilometers. None are yet commercially available except those types that merely monitor visibility. Lasers have made it possible to measure the concentration of trace quantities of pollutant gases over long paths, using spectroscopic techniques. Since a large number of point measurements are required to monitor an area in fine detail, the use of long-path devices might result in considerable savings in instrument costs, provided their average readings are representative of the true levels and do not mask dangerous pockets or persistent clouds of pollutants. This is a moot question, as evidenced by the statement of one APCO official: "It has never been shown that point-source instruments are optimum for measurement of air quality. Open-path instruments may possibly provide more representative results."[2]

As suggested, a primary consideration in designing a system utilizing point sensors is the number and location of sampling stations. Currently, the cost of sensors is the limiting factor. It is seldom that a sufficient number of stations are installed to produce timely data with the desired level of confidence.

West German federal practice, when locating monitoring sites to assess ambient SO_2 levels, calls for one station per square kilometer, located on a square grid. Other grid sampling systems have been designed with specific regard for the location and arrangement of existing sources, topography, prevailing meteorological conditions, and population distribution. This has led to some very economical designs, such as the 31 stations (approximately one per square mile or one per 2.6 km^2) in the locality of Rotterdam, Netherlands, and 1075 total sites for the national survey grid of Great Britain. The flexibility of such an approach is limited, however. In the United States, a posteriori analysis of results from random and selected sites has produced the finding that the statistical distribution of the pollutant concentration values from different points in an area fits a normal curve. A consequence of this distribution is a relationship between desired accuracy, averaging period, and the number of measurement stations required (see R. J. Bryan, in Ref. 1, vol. 2, p. 442). Applying this relationship to data from one study (Nashville), it was found that 245 stations, or four per 2.6 square kilometers, were required to estimate daily mean sulfur dioxide levels with a 95 percent assurance of ±20 percent accuracy. On the other hand, seasonal variations were estimated equally well with only one or two stations.

It is clear that these results are completely inconsistent with the estimate of 790 stations for the national network, derived in Part I from federal planning data. Even if only one station per 2.6 km^2 is assumed, representing a compromise accuracy of ±40 percent, then New York City, Philadelphia, and Chicago would require 1275 stations, and the Los Angeles–Long Beach area, by itself, 4000!

Other minor system factors influence the number and location of point sensors. To give representative measurements, sensors should be at least two "building heights" away from windward structures, far enough away from local pollution sources to reduce their contribution to 5 percent of the total contaminant (R. J. Bryan, in Ref. 1, vol. 2, p. 443), and generally be 3–4 meters above ground.

It appears reasonable that long-path sensors will avoid most of these problems if they are able to measure an average value of the pollutant concentration over a considerable distance. If the normal geographical distribution holds generally, path averaging should give good enough data to determine temporal fluctuations of a cloud of high concentration within an urban area. The location of the original source will be of importance in detecting violations, but long-path sensors may be able to perform this task by scanning and triangulation. Long-path sensors, however, must be more sensitive than point instruments if they are to detect the passage of pollutant "clouds" or plumes of smaller dimensions than their path. This may currently be a problem, as will be discussed.

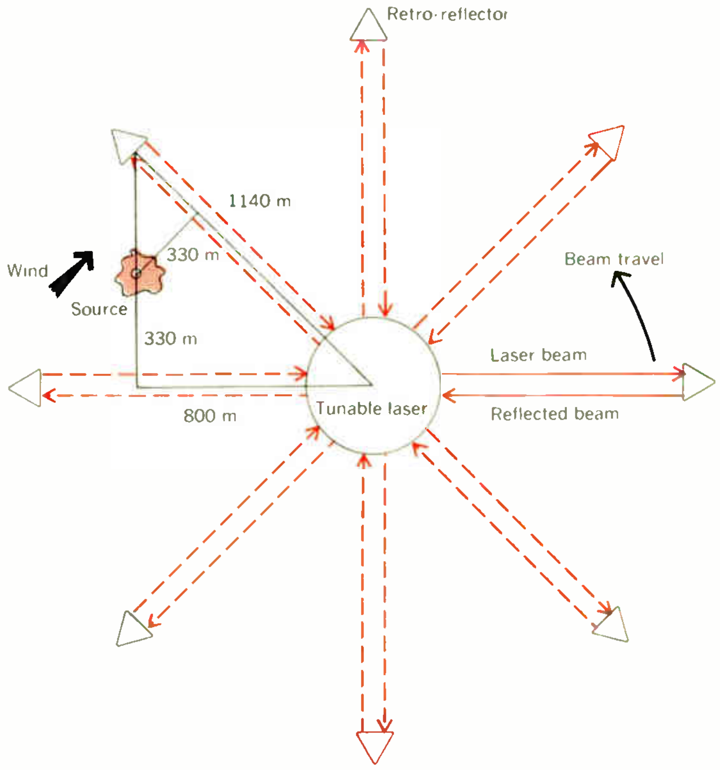
A brief review of the long-path sensors under current investigation reveals that the most highly developed is the correlation spectrometer of Barringer,[3] which projects the spectrum of a distant radiation source through an optical mask replicating the spectrum of a specific gaseous pollutant. Relative motion between the two images produces a modulated correlation signal, proportional to the amount of pollutant in the space between source and spectroscope. Both NO_2 and SO_2 have been detected from aircraft by utilizing the ultraviolet and visible regions (272–550 nm) and reflected sunlight as a source. Sensitivity to the average concentration of NO_2 across a 1000-meter path appears to be of the order of 0.01 ppm.

Hanst and Morreal,[4] followed by Hidalgo,[5] have constructed instruments following the principle of a two-beam differential infrared spectrometer, designed to operate over paths up to 3 km long, using a tunable CO_2 laser. The laser can be forced to oscillate on any of 20 lines between 9.2 and 10.8 μm by means of a diffraction grating in the cavity or by adding absorbing gas (propylene) to the cavity to shift wavelength. The use of two beams—one tuned to a prominent line of the absorption spectrum of the desired gas, the other to an adjacent, nonabsorbing region—eliminates the effects of atmospheric optical turbulence and scintillation, interference from molecules, and other scattering.

Hidalgo envisions an urban air-pollution-monitoring center consisting of a system of lasers emitting radial beams from a central location (see Fig. 2). Corner reflectors will return the radiation from distances of 1–2 km to receivers, which determine the differential absorption over the two-way path on both wavelength channels. If 5 percent differential absorption can be detected at 0.4862 μm, a sensitivity of 0.03 ppm of ozone should be achieved, with somewhat lower sensitivity for other pollutant gases.

Another technique frequently proposed for long-path measurements involves Raman scattering of a laser beam.[6,7] Raman spectrums result from inelastic collisions of monochromatic photons with molecular species during scattering, resulting in bands or lines of shifted wavelength. This wavelength shift is a function of the molecular vibrational modes, and permits unique identification of the molecule by comparison with "library" records. The Raman line intensities are a function of concentration, which normally leads to lack of sensitivity to trace components.[8,9] The combination of the intense monochromatic laser source and a long path tends to overcome this difficulty. The technique devised by Lederman and Widhopf utilized a Q-switched, 1-joule ruby laser, a high-gain photomultiplier, and a special absorption

FIGURE 2. Long-path laser sensor.



chamber, together with narrow-band filters and a spectrograph to isolate the Raman wavelengths and to show possibility of a kilometer-path analyzer. However, the ranges on plumes reported by Kobayasi and Inaba in field tests did not exceed 160 meters.

Microwave absorption spectroscopy has successfully detected formaldehyde, hydrogen cyanide, and a number of other molecules in interstellar space via radio astronomy in recent years[10] and has been proposed by one of its leading investigators, L. E. Snyder, as a detector of atmospheric pollutants. In this technique, receivers and transmitters operating in the gigahertz range would be located at opposite ends of a one-way propagation path. These would be tuned to the characteristic microwave absorption frequencies of common pollutants. Objections have been raised because of interference between molecules at absorption frequencies below that of ammonia.[11]

The effectiveness of long-path sensors in general may be evaluated by considering a simple mathematical model that retains only the essential common characteristics, such as highly directional geometry and sensitivity to the average quantity of a specific pollutant in the path. In this model, properties such as cost and reliability are disregarded. Hidalgo has done this for the laser absorption spectrometer previously referenced. The instrument is assumed capable of detecting an average concentration of 0.24 ppm of sulfur dioxide, corresponding to a sensitivity of 5 percent difference in laser energy absorption, over a two-way path of 2.2 km (2300 meters), which corresponds to complete area coverage with one sensor per 2.6 km^2 (see Fig. 2).

Since the long-path sensor monitors the average concentration along an extended line, and because topography and structures limit the line of sight at desirable monitoring heights (3–4 meters), it is not possible, in general, to scan the entire area. With the use of retro-reflectors, as suggested by Hidalgo, selected paths can be monitored. If there are gaps in the area coverage, it is legitimate to ask how long it will take to detect the rise in average concentration caused by a cloud of pollutant gas emitted from a source located near one of the monitored lines.

The strength of a moderate SO_2 source, say a 10 000-lb/h (4536-kg/h) steam boiler using 3 percent sulfur coal, is about 1 mole/s* (1 mole SO_2 equals 64 grams). In comparison, a large electric power plant (11 500 Btu/kWh or 12.1 kJ/Wh) using the same fuel will emit from 30 to 40 mole/s. Hidalgo's calculations show that emissions from either of these sources at a 1000-meter distance, propelled directly toward the monitored path by a moderate wind,† would require well over four hours to build up to a detectable concentration. Even at 100 meters, the smaller source would require between 15 and 30 minutes to be detected. Referring to Fig. 2, a sensor covering 2.6 km^2 with eight paths may have a source at a maximum range of 330 meters; this could emit 1 mole/s for more than 30 minutes before detection. Under the most unfavorable condition—a dead calm—a pollutant from the same source, transported solely by diffusion, would build up so slowly as to be undetectable within any reasonable time. It is apparent that a choice between an unreasonably large number of expensive point sensors and inadequate or slow long-path instruments is unpalatable. To focus more sharply on the problem, consider a specific example—the continuous automatic air-monitoring system in Allegheny County in Pennsylvania (the Pittsburgh area).[12] This system was planned to cover an area of more than 1800 km^2 with 18 stations, or double that number if cooperating industries constructed private monitors surrounding major sources. Similar to the Rijnmond (Rotterdam) system in the Netherlands, the philosophy behind the Allegheny system is to monitor ground concentrations near large sources with the hope that they will represent maximum exposures for any more distant, populated area, and that their magnitude during stagnant weather will serve as a warning in time to roll back emissions at the source. Experiments in New York City, however, indicating that the location of "pollution pockets" is not constant, expose this concept to some question, and suggest instead that small-scale measurements are indeed necessary.[13] Experience with the Allegheny system itself also tends to confirm the conclusion that pollution measured at any point is the sum of the contribution of multiple sources, and that spot location of ambient sensors cannot pinpoint the emissions from a single source.

However, if Allegheny County were to switch to area grid coverage at four stations per 2.6 km^2, nearly 3000 stations would be required. To monitor five gases at an average of $2000 for each sensor would total $30 million merely for the sensors. For the 100 cities whose total air clean-up costs were estimated by the National Air Pollution Control Administration (NAPCA) at less than $1 billion per year, this would entail spending more than three times as much for sensors as the annual cost of the control equipment. This is clearly untenable.

Several solutions—not mutually exclusive—exist to this dilemma. First, more versatile sensors than those currently used, capable of time-shared measurement of several different pollutant species, are available and permissible, since an interval of several minutes between readings is consistent with damage criterions. If only two sensors can measure all six pollutants at the same unit cost, the station total is reduced to one third or $10 million for the city. Although ten times the planned expenditure, this amount is certainly more reasonable.

Second, the concept of intelligent placement of sensors cannot be entirely wrong. Less-developed areas and flatter terrain will need fewer monitors, and better knowledge of local meteorology and source diffusion will suggest more economical sampling schemes than a uniform grid provides, even if results run short of Allegheny system's hopes. It is not unreasonable to expect an improvement factor of two to four in sampling by good design, reducing the sensor cost per city to within a $2.5 to $5 million range. Spread over several years, this should be a more equitable burden for the benefits received.

An ultimate solution, of course, would be the invention of inexpensive, maintenance-free, solid-state sensors (termed by APCO personnel as "third generation"[14]). Although a price of $1 per point sensor has not yet appeared on the horizon, such an achievement would reduce the cost of the monitoring system to essentially that of data transmission and handling.

* Coal emission factor: 38 × (%S)lb SO_2/ton [15.6 × (%S)kg SO_2/tonne], 87 percent efficiency, 12 000 Btu/lb (28 kJ/g).

† Transport coefficient equal to 10^5 cm/s (Ref. 5).

In sum, it appears most likely that the course of progress will be an evolutionary one. Many monitoring systems, like Allegheny County's, are already operational. As funds and better or cheaper sensors become available, these systems will extend their coverage, improving it with the knowledge they will gain from their own data collection. Ultimately, the state of perfection needed for detailed modeling and forecasting will be reached.

What will be the role of long-path sensors? Clearly, their sensitivity must be improved by several orders of magnitude if they are to be sufficiently responsive. Siting difficulties will remain and will require as much ingenuity as is needed for point sensors. To be competitive, the cost of a single long-path instrument, installed and with the geometry shown in Fig. 2, must not exceed the installed cost of four point-sensor stations, ($15 000–20 000). This assumes that the use of tunable lasers will permit simultaneous monitoring of six pollutants. If these targets can be met, there will certainly be at least a supporting role for these instruments in future monitoring systems.

Data processing

It is apparent from the foregoing that accurate and reliable reporting of air quality will require many hundreds or thousands of sensors in each metropolitan area. With large numbers of sensors reporting at frequent intervals, telemetry and data processing will assume increasing importance. Cost and efficiency suggest that digital data communication will be optimum, requiring digitizers and multiplexers at each location. Use of large-scale integrated circuits will keep the cost of providing these functions to a minimum. Communication must be reliable and bidirectional in order to provide for polling, calibration, and control commands from the central processor. It is probable that narrow-band wired communication will be preferred in most locations, because of its lower cost. Since "continuous" or "real-time" monitoring of air-pollution data will not normally require more than six data words per minute from each station, there is every opportunity to time-share communication channels for even greater cost savings.

Some storage for polling, but not more than a few bits per sensor, will be needed at each remote station. In the event that the sensor is complex and multicomponented, a local digital processor may be used at the remote station to manage it. If so, the processor can easily provide all the communication and storage functions mentioned.

Display and recording at each station should be minimal. For maintenance checks, a single display (digital) could be shared with all local sensors. If recording is needed, it could be provided economically by magnetic tape. These are not the current practices in today's monitoring stations, which are largely manual and rely on graphical records.

At a central location, the data are edited and validated, reduced, and displayed for interpretation. If the data are already in digital form, these functions can be performed automatically by the computer. Editing and validation consist of checking calibrations and recognizing and rejecting errors due to noise, component failures and drift, and statistical abnormalities. The program details will depend on the characteristics of the particular sensors, but should present no difficulty. In today's systems, some of these tasks are performed manually. They are particularly necessary because of the unreliability of present sensors. Data reduction includes the monitoring of instantaneous values and provisions for an alarm if preestablished limits are exceeded. In addition, data are averaged over various periods ranging from five minutes to a year, peaks and minima noted, and frequency distributions computed. It has been found, even for older networks with only tens of stations, that data reduction should be automated if polled more than four to six times per hour (R. J. Bryan, in Ref. 1, vol. 2, p. 455). With today's improved and low-cost computers, there will be no question of automating, even in smaller networks.

Data are displayed by means of conventional printouts and illuminated maps, alarms, etc. In addition, statistical evaluation and interpretation are conducted, all of which is facilitated by computer routines.

Trend prediction

The prediction of future air-pollution levels at any location within a monitoring subsystem's boundaries will be a fundamental function of the nationwide control system. We should be concerned with both short-term fluctuations and long-term trends of air pollution at several levels, varying from the fine scale within cities to that of the entire nation and the globe. Although digressing somewhat, we should be aware of the global problem, since local systems, feeding data into national archives, may detect long-term national or global background trends.

At the global level, we must be concerned with the ultimate disposal of all pollutants dumped into the atmosphere. Air parcels move from west to east, generally, making a complete circuit of the globe in about 12 days at 30 degrees latitude, but with some north–south meandering rather than along a straight line. Parcels containing pollutants, such as clouds from nuclear or volcanic explosions, have been identified and tracked over such paths for several global circuits.[15]

If the natural purging mechanisms of precipitation, absorption, chemical reaction, and biological processes do not remove all pollutants from the atmosphere in one global circuit, the concentration will ultimately build up to lethal proportions.[16] Evidence of man-made pollutant buildup over the oceans in the last 40–60 years has been recently reported.[17]

Such global, long-range predictions are of international interest, and beyond the scope of the systems envisioned here. The necessary measurements perhaps will be best made via satellite and, in fact, development of an instrument (a correlation spectrometer) for this purpose is under way. Nevertheless, these considerations illustrate the national and international nature of air-pollution problems, and encourage the study of local records over long periods of time to investigate possible trends or cycles.

On the national level, an air parcel traveling in the prevailing direction over the United States might pass over a number of cities, accumulating more pollution than it loses between them over some parts of its path and building up the local average or background levels for the more easterly cities. In one (purely conjectural) path, illustrated in Fig. 3, the air mass picks up a heavy concentration on the West Coast, drops nearly to background levels over the desert and Rocky Mountains, builds up to increasing peaks on its way to the East

II. Source types*

I. Fuel burning for heat and power
  A. Utilities
  B. Residences (private houses, apartments, etc.)
  C. Industrial
    1. Manufacturing (per SIC code)
    2. Commercial (stores, hospitals, hotels, etc.)
    3. Processing (e.g., laundries, service stations)

II. Incineration
  A. Municipal
  B. Industrial and commercial
  C. Residence
  D. Apartment house
  E. Open refuse burning

III. Transportation
  A. Motor vehicles (gasoline, diesel)
  B. Trucks and buses
  C. Railroad engines
  D. Ships
  E. Aircraft

IV. Industrial and commercial
  A. Manufacturing (e.g., chemical, petroleum, metals)
  B. Agriculture (spraying, dusting)
  C. Commercial (dry cleaning, spray painting)
  D. Miscellaneous (sewage treatment, demolition)

* From A. T. Rossano, Jr., in Ref. 1, vol. 2, pp. 615–618.

III. General emission factor[19] (tonnes/unit*)

Source	SO_x	NO_x	Particulate	CO	HC
Residual oil (1000 hp or more)†	0.203‡	0.052	0.004	negl.	0.0016
Residual oil (less than 1000 hp)	0.203‡	0.036	0.006	0.001	0.001
Distillate oil (1000 hp or more)	0.024‡	0.052	0.004	negl.	0.0016
Distillate oil (less than 1000 hp)	0.024‡	0.036	0.006	0.001	0.001
Anthracite coal (residential)	0.011‡	0.004	0.10[a]	0.025	0.005
Anthracite coal (commercial, governmental)	0.011‡	0.004	0.025[a]	0.025	0.005
Bituminous coal (residential, commercial, governmental)	0.019‡	0.004	0.018[a]	0.025	0.005
Bituminous coal (industrial)	0.019‡	0.010	0.018[a]	0.002	0.0005
Natural gas (residential, commercial, governmental)	0.0002	0.058	0.010	0.0002	negl.
Natural gas (industrial)	0.0002	0.107	0.009	0.0002	negl.
Gasoline	0.004	0.057	0.005	1.455	0.262
Diesel oil	0.020	0.111	0.055	0.030	0.090
Aircraft	(b)	(b)	(b)	(b)	(b)
Open-burning dump	0.0006	0.0003	0.024	0.043	0.040
Municipal incinerator	0.001	0.001	(c)	0.0004	0.0007
Residential, commercial, governmental, and industrial incineration	0.0002	0.0008	0.010	0.013	0.018
Backyard paper burning	0.0006	0.0003	0.002	—	0.073
Gasoline evaporation	—	—	—	—	0.060
Solvent losses (dry cleaning)	—	—	—	—	1.95

* Fuel oil, gasoline, and diesel fuel, 1000 gallons (3.8 m^3); coal, tonnes; natura gas, 10 ft^3 (0.28 m^3); refuse, tonnes; dry cleaning, 1000 people.
† 1 horsepower = 746 watts.
‡ Dependent on following sulfur content of fuel: residual oil, 2.55%; distillate oil, 0.3%; anthracite coal, 0.6%; bituminous coal, 1.0%.
a Dependent on ash content of coal (anthracite coal, 10.0%; bituminous coal, 7.0%), type of firing unit, and type of control.
b Dependent on type of aircraft.
c Dependent on type of control.

Coast, and then decreases over the Atlantic before building up again over Europe.[16] Such a buildup of average levels, even for a short time, would cause the recipient city to be more vulnerable to high-pollution episodes resulting from its local source and weather fluctuations.

It is obvious that air pollution is no respecter of state or regional boundaries, as any New Yorker who has complained about New Jersey factory emanations will testify. Even now, something like 7 percent of the carbon monoxide measured in downtown Washington, D.C., is alleged to come from Baltimore, representing a separate air quality control region (AQCR).[18] It is clear that a national solution to air-pollution problems must include an exchange of data to make air-pollution forecasts feasible.

Short-term forecasts of maximum concentrations that may cause acute pollution symptoms are a better-defined objective of the nationwide air-monitoring system and its local activities. In communities where topographical and weather factors favor pollution buildup, increasing industrialization and population make it an imperative function. The primary purpose of such information is to provide early warning for decision and implementation of control measures for existing sources or, in a larger time frame, to serve as the basis for air-resources management and land-use planning.

The predictive information system must provide a dynamic model of the local air resource so that its behavior can be simulated, given the following inputs:

1. Current concentration and distribution of pollutants.
2. Source strengths and distribution.
3. Present or predicted meteorological factors.
4. Geographical and topographical theater.

Clearly, the model must be a mathematical simulation capable of being programmed on a computer, so that large quantities of data and alternate measures can be evaluated in "real time."

Current levels of pollutant concentration

Unless there appears some unique causal relationship or statistical correlation between dependent variables, there is no reason to believe that prediction to a given accuracy will require any fewer input data than current level measurement. Therefore, everything said about current level monitoring will also apply to the prediction function. The necessary accuracy should be determined on a cost–benefit basis, balancing the cost of acquiring, monitoring, and recording data, the cost of unnecessary control measures, and the risk of failure to control.

Source factors

An inventory of the sources of emission in a locality is needed in order to construct an appropriate mathematical model. Models tested to date have all used some degree of simplifying assumptions. Source inventories are also required to design a sampling and analysis program, to serve as a basis for control or zoning laws and planning, and as a means for interpreting effects. A classification of source types has been suggested by A. T. Rossano, Jr. (in Ref. 1, vol. 2, pp. 615–618) to aid in taking a community inventory (Table II). The data should include, in addition to source type, the kinds of emissions, discharge rates, number of sources, location, and the process, raw materials, and control used. Sources should also be inventoried as to which ones emit quantities of oxides, CO, organics, and particulates, or specific toxic substances.

Information so obtained should be supplemented with a questionnaire soliciting data on plant size, manufacturing procedures, operating schedules, fuel and solvent usage, and refuse burning. Additional engineering inspection will yield information on process flow diagrams, raw materials, fuel types, emission rates, and gas-cleaning devices. These data can be updated through a computerized registration and permit system, as in Chicago and other major cities.

Emission factors, if not available directly from plant data, can be estimated from general tabulations similar to that in Table III or Ref. 19. It is recognized that such information will not be complete or accurate, since emission tables merely represent a statistical average of many installations and operating conditions. In most research to date, approximate data have proved good enough, but this will not hold true in the future.

Source data may be assembled on a gridded map and correlated with demographic and land-use information. The time element is introduced through traffic flow information, temperature (degree-day) data, and similar sources. It is then in form to be introduced into a computation scheme.

Some use has been made of aerial photography to identify sources. It is expected that use of infrared photography, radiometers, and special sensors for inventory purposes will increase, since they suggest an economical means to mechanize this task.

Much of the prediction activity that has been accomplished in recent years has been concerned with seasonal average, rather than episodic air quality conditions, in order to establish facts for setting air quality regional boundaries.[18,20–22] For this purpose, it is adequate to utilize averages and approximations for sources, meteorology, and topography. Local control systems, attempting to predict just these extreme conditions that are excluded by the averaging process, will require more refined data. In the case of source strength, it is desirable to know the current emission of each major source whenever there is any danger of pollution buildup. It can be envisioned that, ultimately, many sources will be instrumented and will telemeter information to the data center. In this way, direct feedback of information following control measures can be obtained as an additional benefit.

Meteorological data

The importance of global air movements in the transportation and dispersal of worldwide pollution has already been discussed. Local weather information and air-mass movement are even more vital (as much as source strength and distribution) in predicting pollution concentration maxima. The concept of the self-contained "air shed" was introduced within Part I of this article; its boundaries are so defined as to result in a minimum of pollution transport into or out of the area. In practice, the concept is compromised by demographic and political considerations when the AQCRs are designated (by EPA).[22] Even in theory, the air-shed concept has been challenged on the basis that episodic statistics may differ substantially from those used in AQCR designations.[23]

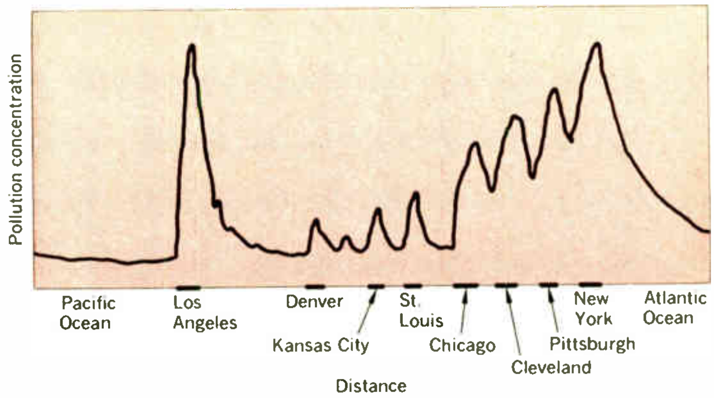
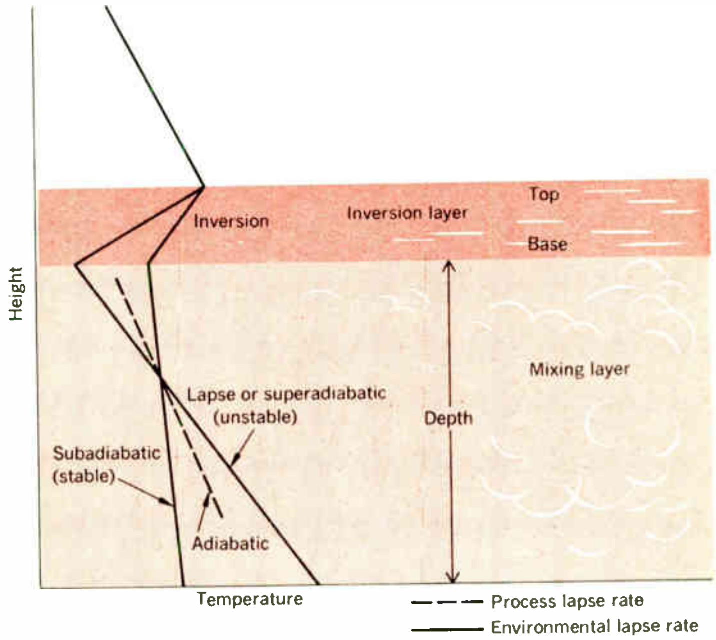


FIGURE 3. Hypothetical example of pollution distribution across the United States.[16]

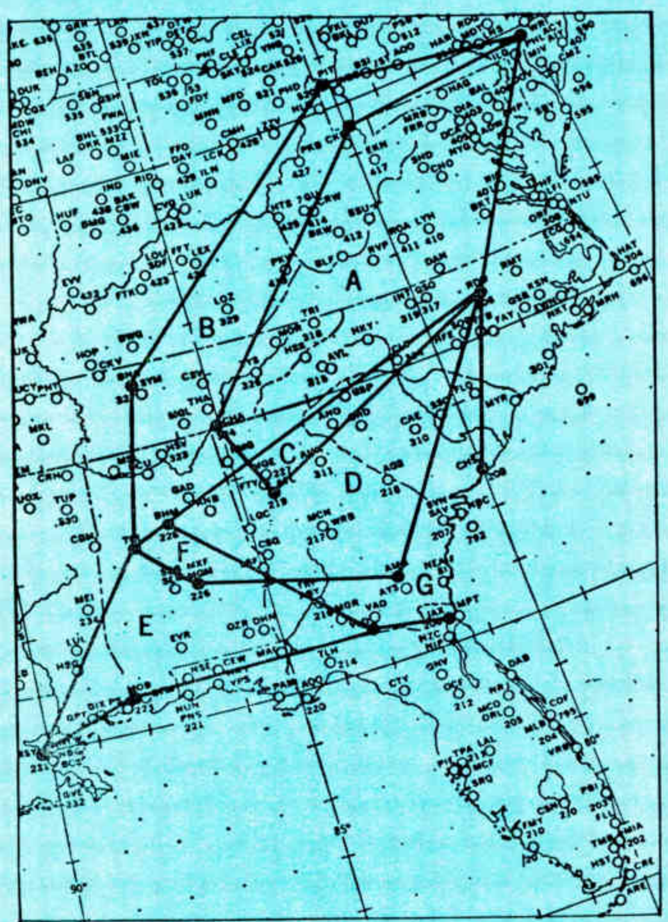
FIGURE 4. Temperature lapse rate and mixing layer.



But the prime influence of local meteorology, topography, and source geometry on the concentration and movement of polluted air masses has not been questioned.

Wind velocity determines the mass movement of pollutant clouds. Turbulence causes mixing and rapid diffusion. The lapse rate, or rate of temperature decrease with height, determines the stability or the degree of stagnation of the air mass in the vertical direction (see Fig. 4). A parcel of air expanding as it rises, without heat exchange, cools at the (process) adiabatic rate; if the environmental lapse rate is the same, the parcel is hydrostatically in neutral equilibrium. If the environmental lapse is greater (superadiabatic), the air parcel is accelerated upward and is unstable; if lower, the parcel is stable. If the environmental lapse rate is negative (temperature rising with altitude), the stable region is called an inversion. An inversion or stable air mass above a city acts like a cap, trapping the pollutants beneath it. The thickness of turbulent air under the cap is called the "mixing depth." The adiabatic lapse is, roughly, 10°C

FIGURE 5. Advisory of high air-pollution potential, National Meteorological Center (Ref. 25, p. 292). (The American Association for the Advancement of Science, 1970)



Advisory No. 122 (National Meteorological Center Advisory of High Air Pollution Potential, 28 July to 1 August 1970, 1300 E.S.T.).

28 July: begin areas *A, C*
29 July: continue areas *A, C*
begin areas *B, D, F*
30 July: continue areas *C, D, F*
begin areas *E, G,*
end areas *A, B*
31 July: continue areas *E, F,*
end areas *C, D, G*
1 August: end areas *E, F*

ESSA Meteorological Support Unit statements of local high air pollution potential, 27 July to 30 July 1970.

27 July: Washington, D.C. (and Baltimore, Md.), begin 1130 E.D.T.
28 July: Washington, D.C. (and Baltimore, Md.), continue;
Philadelphia, begin 1300 E.D.T.
29 July: Washington, D.C. (and Baltimore, Md.), continue;
Philadelphia, continue;
New York City, local statement on dispersion conditions (not local high air pollution potential) were issued at 1100 and 1500 E.D.T.
30 July: Washington, D.C. (and Baltimore, Md.), end 1700 E.D.T.;
Philadelphia, end 1200 E.D.T.

per kilometer for dry air. Its exact value is affected by temperature, height, and the degree of saturation of the air.

Surface wind direction can be obtained by means of vanes of various types, equipped with electric direction sensors; and the velocity of winds aloft measured by pilot balloons or "tetroons," and tracked by radar or theodolites. Smoke rockets and colored or fluorescent power tracers are also used. Wind speed is measured with a variety of anemometers: cup, propeller, or windmill, and hot-wire types (thermocouple or thermistor). Speed aloft can also be obtained from monitors placed on television towers.

Turbulence can be measured using rapid-response vanes and instrument computers that determine the standard deviation of the direction fluctuation, known as "sigma." A two-directional vane or bivane will measure sigma in the vertical as well as the horizontal plane. Other methods used include gust accelerometers, bladed wheels similar to restrained cup anemometers, and various airborne devices.

Lapse rate may be measured from thermocouple-instrumented towers, aircraft, wiresondes, and similar means. Near the ground, it may be measured by a thermal radiometer, a two-sided differential thermopile instrument that determines the net cooling or heating load at the surface. The vertical temperature structure can also be explored with microwave or infrared radiometer techniques. Laser radar (lidar) is capable of measuring the height of the mixing layer through reflection from particulates concentrated below the interface. This height is an indirect measure of lapse rate. Conversely, the mixing depth, or the height of the mixing layer (a significant parameter in current diffusion models), may be calculated by comparing the dry adiabatic lapse rate, measured from the surface temperature, with the morning environmental sounding and noting the altitude where the two temperatures are equal.

Solar radiation is important in photochemical reactions, which produce oxidants as secondary pollutants. A number of instruments (actinometers) have been designed to measure this parameter.

Hydrometers used for pollution meteorology are of standard design, as are precipitation instruments, rain gauges, and the like. Precipitation is the most important mode of purging the air of pollutants.

Recommended instrumentation for a moderate-size city may consist of several television towers with wind vanes and totalizing cup anemometers at 10, 20, 40, and 80 meters, topped by a gust accelerometer, bivane, or three-axis anemometer. Sigma computers, rain gauges, and actinometers are required. A wide horizontal deployment of sensors is recommended in order to monitor valley or shoreline winds. Investigations of low-level inversion prevalence using wiresonde temperature probes is also suggested (E. W. Hewson, in Ref. 1, pp. 329–387).*

The system installed by the city of Chicago, described by Cramer,[24] obtains meteorological data from eight remote stations, but has a capacity for 59, each handling five channels of information. Wind direction and speed are digitized at 5-second intervals, electronically stored, averaged over a 15-minute period, and recorded. The

* This is a complete survey of instrumentation requirements.

stored averages are polled at intervals and telemetered to the central data station for printout and display. In more recent installations, on-line analog and digital computers are being used for assessment of pollution conditions and diffusion analysis. An interesting example of direct computer control of a pollutant source through meteorological measurements is found in the Enrico Fermi atomic power plant at Monroe, Mich. A valve in the stack is operated by an analog computer, which accepts wind-speed and -direction data, computes the mean speed and the direction standard deviation (sigma), and compares them with preset dispersion criterions (R. C. Wanta, in Ref. 1, vol. 1, pp. 187–223). If they are not met, the stack gases are diverted to temporary storage until good dispersion is restored.

The influence of topography is very pervasive throughout the entire process of predicting pollution transport and dispersion. The effect of surface roughness on wind speed is a well-known instance. Sea–land and mountain–valley interfaces, as a result of differential heating, cause local winds that deviate from the prevailing direction. Valleys also cause channeling of wind direction, as does the rectangular "waffle" pattern of city streets. Passes and mountain ridges increase wind speed by constricting the air mass, whereas valleys have an opposite effect. The correlation of a bowl or valley topography with stagnation episodes (Meuse Valley, Columbia River, and Los Angeles basins) is often noted. However, a city on an open plain also favors formation of stagnant air pools because of its surface roughness, acting as a sink for the wind's kinetic energy (see Wanta, Ref. 1).

Supplementing the measure of pollutant dispersion that can be deduced from local meteorological data are those of a synoptic nature. The Environmental Science Services Administration (ESSA) in cooperation with APCO provides forecasts over large areas whenever persistent high-pressure air masses lead to stagnant conditions conducive to air-pollution episodes. Figure 5 provides an example of one of these "High Air Pollution Potential Advisories." The extent of this stagnant area is cited as a reason for "nationwide forecasting and coordinated control strategy."[25]

Conclusion

To enjoy the benefits of clean air at the least overall cost to our economy, we must invest in a flexible nationwide system that will enable each community, according to its pollution sources, topography, and meteorological characteristics, to engage in an optimum local control strategy, but one that is coordinated nationally in order to benefit from the economies of scale, a uniform burden on industry, and an exchange of data and techniques. Before reaching fruition, this system must leap many hurdles—political, legal, and economic—not the least of which are the technical and scientific problems concerned with safe use of the air resource to transport and dilute effluent wastes. Statistically representative monitoring of current levels of air pollution to insure levels consistent with human health will require more and better chemical and meteorological data than have hitherto been available to any locality on a long-term basis. This need will continue to spur exploration of exotic sensors for ambient monitoring, such as lasers, Raman spectrographs, and others, for which, perhaps, the basic phenomenons have not yet been identified. Less glamorous but equally essential is a detailed inventory of polluting sources, their locations, and means for monitoring their temporal variations.

In Part III we shall discuss the formidable problems of merging these data into a model describing the complexities of gas dynamics, air-mass transport, chemical reactivity, and urban topography; and thus be able to generate a detailed, short-term pollution forecast that will permit meaningful control actions to be taken on the time scale of sensitive human receptors.

REFERENCES

1. Stern, A. C., ed., *Air Pollution*, 2nd ed. New York: Academic Press, 1968, vol. 3, p. 607, Fig. 2; originally published in *Air Quality Criteria for Sulfur Dioxide*, U.S. Department of Health, Education, and Welfare.
2. Altshuller, A. P., "Instruments for monitoring air pollutants," presented at Instrument Society of America meeting, Philadelphia, Pa., May 19–22, 1968; also in *Environmental Pollution Instrumentation*, R. L. Chapman, ed. Pittsburgh: Instrument Society of America, 1969, pp. 1–6.
3. Barringer, A. R., in *Environmental Pollution Instrumentation. Ibid.*, pp. 49–67.
4. Hanst, P. L., and Morreal, J. A., *J. Air Pollution Contro Assoc.*, vol. 18, no. 11, pp. 754–759, 1968.
5. Hidalgo, J., "Study of the application of laser technology to atmospheric contamination measureation," Final Tech. Rept, NASA contract NAS 12-664, Radiation Laboratory, Tulane University, May 1, 1968–Apr. 30, 1969 (Clearinghouse, NBS, Springfield, Va., N69-36282).
6. Lederman, S., and Widhopf, "Specie concentration measurements utilizing Raman scattering of a laser beam," Paper 70-224, AIAA 8th Aerospace Sciences Meeting, New York City, Jan. 19–21, 1970.
7. Kobayasi, T., and Inaba, H., "Laser-Raman radar for air pollution probe," *Proc. IEEE*, vol. 58, pp. 1568–1571, Oct. 1970.
8. Rosenbaum, E. J., in *The Encyclopedia of Spectroscopy*, G. L. Clark, ed. New York: Reinhold, 1960, pp. 675–678.
9. Stewart, J. E., in *The Encyclopedia of Spectroscopy. Ibid.*, pp. 678–681.
10. Wick, G. L., "Interstellar molecules: Chemicals in the sky," *Science*, vol. 170, pp. 149–150, Oct. 9, 1970.
11. *Product Eng.*, Sept. 14, 1970.
12. Stockton, E. L., "Allegheny County's air monitoring program," Paper 69-207, 62nd Ann. Meeting of the Air Pollution Control Assoc., New York City, June 22–26, 1969.
13. Savas, E. S., "Feedback controls on urban air pollution," *IEEE Spectrum*, vol. 6, pp. 77–81, July 1969.
14. O'Keeffe, A. E., "Needs in electronic instrumentation for air-pollution analysis," *IEEE Trans. Geoscience Electronics*, vol. GE-8, pp. 145–148, July 1970.
15. Newell, R., "The global circulation of atmospheric pollutants," *Sci. Am.*, vol. 224, pp. 33–34, Jan. 1971.
16. Ludwig, C. B., *et al.*, "Study of air pollution detection by remote sensors," Rept. CR-1380, NASA, Washington, D.C., July 1965, p. 1-1–1-2.
17. *The New York Times*, Oct. 18, 1970.
18. "Report for consultation on the Washington, D.C., National Capitol Interstate Air Quality Control Region," U.S. Department of Health, Education, and Welfare, USPHS, July 1968.
19. Duprey, R. L., "Compilation of air pollutant emission factors," NAPCA, USPHS Publication 999-AP-42, Raleigh, N.C., 1965.
20. "Report for consultation on the Metropolitan Denver Air Quality Control Region," USPHS, Oct. 1968.
21. "Report for consultation on the Philadelphia Interstate Air Quality Control Region," USPHS, Oct. 1968.
22. "Report for consultation on the Los Angeles Air Quality Control Region," USPHS, Nov. 1968.
23. Stern, A. C., cited in *Chem. Eng. News*, vol. 48, June 1970.
24. Cramer, H. E., in *Environmental Pollution Instrumentation. Op. cit.*, p. 14.
25. Morgan, G. B., *et al.*, "Air pollution surveillance systems," *Science*, vol. 170, pp. 289–296, Oct. 16, 1970.

Reprints of this article (No. X71-114) are available to readers. Please use the order form on page 8, which gives information and prices.

New product applications

Network optimization for two-way CATV system design is programmed on computer

Complete layouts and designs for modern CATV distribution systems resulting from a package of computer programs is claimed as a unique service by Network Analysis Corporation. The resultant designs are said to be free of the approximations and inadvertent errors introduced by human designers. The CATV system designer is faced with a number of competing variables such as cross modulation, noise, bandwidth, temperature, alternate routes, and component characteristics.

Among the crucial decisions in the design are selection of head-end sites, location of messenger strand cable, selection of trunk distribution points, amplifier output levels and gains, assignment of automatic gain and slope control, and temperature compensation.

The design of the one-way system shown in Fig. 1A was performed by a human designer who followed the seemingly reasonable rule of thumb that the trunk should be kept as short as possible. The superior design produced by computer is given in Fig. 1B. This design has a longer trunk and even requires an extra trunk amplifier. The cost was less by $158.

The design of two-way systems is complicated by the fact that most present systems are one-way and many two-way systems will evolve as expansions of them. However, it is often true that efficient two-way systems cannot be obtained simply by moving amplifiers and changing output levels. As an example, an effective one-way design is shown in Fig. 2. If this section of feeder is near the end of the trunk, it may happen that with the added noise and cross modulation for two-way systems, a cascade of two extender amplifiers will be intolerable. No simple modification of the design is possible for two-way systems.

The computer, however, can obtain radically different alternate solutions with nearly equal costs, meeting additional constraints. The one-way design shown in Fig. 3 costs only $11 more than the design in Fig. 2 and contains the same number of extender amplifiers. It has no more than one extender in cascade, and thus can be expanded for two-way performance.

Write Network Analysis Corp., Beechwood, Old Tappan Road, Glen Cove, N.Y. 11542.

Circle No. 55 on Reader Service Card

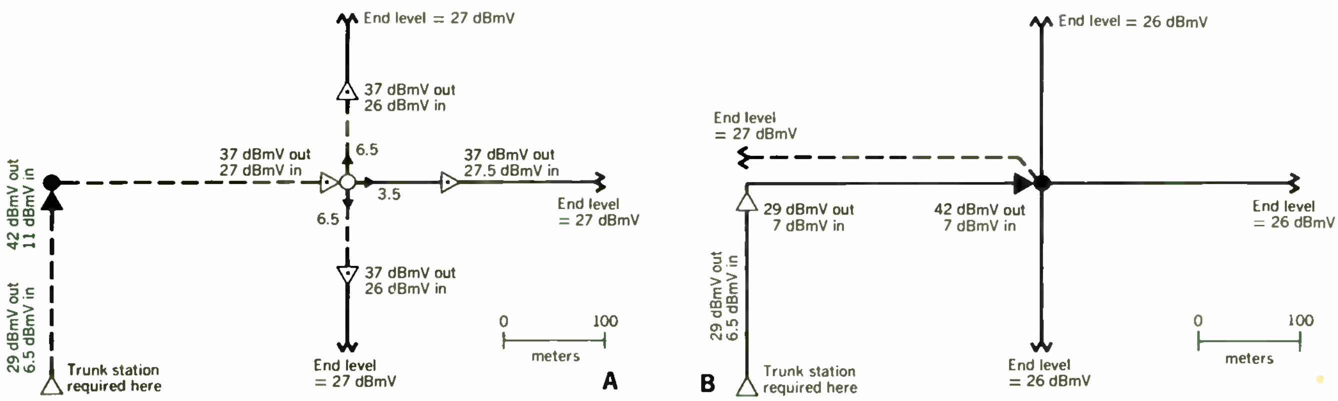


FIGURE 1. A—Manual design for system. B—Computer design costs less.

FIGURE 2. Design with two extender amplifiers in cascade.

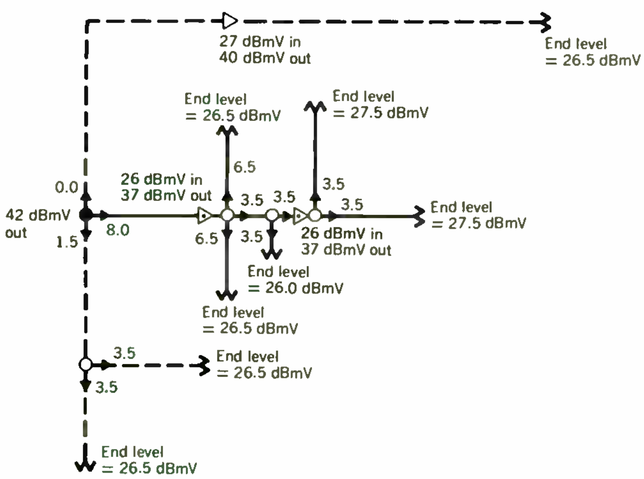


FIGURE 3. Design with no extender cascade of length two.