RCA REVIEW

a technical journal

RADIO AND ELECTRONICS
RESEARCH • ENGINEERING

Published quarterly by
RCA LABORATORIES
in cooperation with all subsidiaries and divisions of
RADIO CORPORATION OF AMERICA

VOLUME XXIV SEPTEMBER 1963 NUMBER 3

CONTENTS

PAGE

System Organization for General Communication Via Medium Altitude Satellites 293
D. G. C. LUCK

Superconductive Associative Memories 325
R. W. AHRONS

High-Speed Transistor–Tunnel-Diode Sequential Circuits 355
J. J. AMODEI AND J. R. BURNS

Parameter Optimization of an FM/FM Multichannel Telemetry System 381
D. H. SAPP

On a Problem in Single-Sideband Communications 391
J. DUTKA

Discussion and Applications of Electrostatic Signal Recording 406
I. M. KRITTMAN AND J. A. INSLEE

Communication-Satellite-System Handover Requirement and Associated Design Problems 421
H. J. WEISS

RCA TECHNICAL PAPERS 459

AUTHORS 462

RCA REVIEW is regularly abstracted and indexed by *Abstracts of Photographic Science and Engineering Literature, Applied Science and Technology Index, Bulletin Signalétique des Télécommunications, Chemical Abstracts, Electronic and Radio Engineer, Mathematical Reviews, and Science Abstracts* (I.E.E.-Brit.).

RCA REVIEW

BOARD OF EDITORS

SYSTEM ORGANIZATION FOR GENERAL COMMUNICATION VIA MEDIUM ALTITUDE SATELLITES

By

D. G. C. Luck

RCA Advanced Military Systems
Princeton, N. J.

Summary *— World-wide, general-purpose communication service via medium-altitude satellites poses problems different from those of long-range, heavy-traffic trunk service. Solutions to the general problem are generated from three principles — ready access of each user to the total system via any satellite, post-office ground stations, and geometric symmetry. Post offices generate no traffic, but serve for its collection, sorting, packaging, forwarding, and distribution. Separated operations of collection and distribution permit simple, flexible operation of complex systems. Contacting directly only its post office, not other users, each user central station serves a surrounding region via surface facilities. Symmetry equalizes tasks, avoiding local over-design and promoting system economy. Two alternative communications systems evolved from these principles are discussed:*

1. *Nine active satellites keeping relative stations at 8260-nautical-mile altitude, with just one post office near the north pole for very wide service, or 5 post offices for the entire world.*
2. *Five post-office stations in northern latitudes, with 250 random satellites at 1570-nautical-mile altitude (may be passive). Use of 8 post offices would provide a stronger net and add fringe-area service.*

Introduction

IT APPEARS that satellites will first be used in practical communication operations as a limited supplement to existing ground facilities. They will then serve only a moderate number of heavy-traffic centers separated by such natural barriers as to render contact by older means very difficult and expensive. They will simply interconnect these centers in pairs. The needs of this very important but highly specialized service have constrained a good deal of the more detailed system thinking that has been done to date regarding communication via satellites.[1,2]

[1] J. D. Rinehart and M. F. Robbins, "Characteristics of the Service Provided by Communications Satellites in Uncontrolled Orbits," *Bell Syst. Tech. Jour.*, Vol. 41, p. 1621, Sept. 1962.

[2] W. H. Meckling, "Economic Potential of Communication Satellites," *Science*, Vol. 133, June 16, 1961.

At a somewhat later time, it may be expected that the potential of satellites for rendering communication service will be much more fully exploited. They may then be used in a more general way, permitting any user, anywhere, to insert communication traffic directly into a system capable of delivering it to any other user.[3] When this situation comes about, satellite service may take over much of the function performed today by medium to long-haul ground facilities, rather than being just a limited adjunct to pre-existing facilities. Such a course of events is the normal consequence of an evolving new technology, and communication via satellites seems unlikely to depart from this norm. System thinking addressed directly to providing this more general sort of service, flexibly interconnecting many users, can profitably include approaches that are quite different conceptually from those already found appropriate for interconnecting in pairs a few users. Even in an era of general-utility service, of course, trunk-line interconnection in pairs of a few major heavy-traffic centers can remain a very important special service in its own right, and the general-use system should not crowd out this capability.

This paper explores certain basic concepts of organization that appear useful for configuring systems for general-purpose communication via satellites. The task postulated is that of enabling any user station, anywhere in the inhabited portions of the earth, to enter readily at any time into two-way communication with any other user stations it may select, wherever they may be. It is just this sort of operating characteristic that has made earth-synchronous satellite systems so attractive for general use. What will be done here is to show how the same properties may be provided in systems using satellites at medium altitudes. These systems will not employ direct satellite-to-satellite communication; only ground-to-satellite and satellite-to-ground communication paths will be used here. Examples will be given of system configurations that have the stated capabilities.

"Users" will here be considered to mean major public entities, such as fairly large cities, small countries, or isolated but well-populated islands. Each such major entity should be able to support one ground station of considerable capability and to provide a fairly unbroken flow of its own traffic. The system problems of the still later time era when high-powered satellites might render similar service directly to individual persons as true end users are not attacked in this paper. In the era considered, the major user entities are assumed to be con-

[3] E. A. Laport and S. Metzger, "Concept for an Intercontinental Satellite Communication System," *RCA Review*, Vol. XXII, p. 555, Sept. 1961.

nected to their local masses of small end users by all the familiar sorts of local-communication facilities. The only massive demand for communication capability that can be foreseen with assurance arises from the public telephone service. Total capacity required for all other known uses is very modest by comparison. Many new sorts of communication service have been imagined, but public demand for them cannot now be predicted with confidence.

Technique Aspects

Problems of apparatus design receive only minimal consideration here. What is done is to concentrate attention on such matters as mode of system operation, traffic-routing concept, satellite-orbit patterns, and location patterns of central ground stations. It is assumed, in order to pay full attention to these matters, that the characteristics of the satellite-borne repeaters are such that simultaneous access to the satellite service is available to many users, and that the communication capability of single satellites is adequate to avoid serious constraints on system design due to limited traffic capacity aloft.

These assumed capabilities are known to be fully feasible today technically, in that they are freely available through use of passive-reflector satellites. Passive reflectors as constructed today, however, are so heavy that rocket-vehicle costs to place conveniently large ones in orbit become formidable. The assumption made, then, is that active-repeater technology will reasonably soon approach the traffic capability and general availability of access exhibited by passive reflectors, without sacrificing the active-system advantage of light weight in orbit.

While the choice betweeen active-repeater and passive-reflector satellites is largely irrelevant to the objectives of this paper (so long as adequate flexibility of access and traffic capacity are provided), one more point should be noted. This is that simple, fully passive reflector satellites are severely limited in the use that they can make of directivity, with the result that higher working altitudes require markedly heavier reflectors. A consequence of this is that the vehicle costs to launch them exhibit a rather strong, sharp minimum for a particular choice of satellite altitude. This represents the interplay of number of satellites needed for coverage and size of satellite needed for good signals. Passive-system design is strongly constrained by the pronounced nature of this economically optimum altitude, and by its relatively low value. Because it seems more reasonable to stabilize the attitude of active than of passive satellites, and because the directivity of stabilized active-satellite antennas can be designed to match the

working altitude, active-satellite systems show a much flatter launch-cost minimum. This active-system optimum occurs at markedly higher altitude, and because of its flatness imposes much less constraint on the choice of system geometry.

Minimum attention is also given to questions of type of signal and type of modulation. So long as the criterion of multiple access to satellites is always met, interaction between choice of modulation and choice of system configuration remains weak. Economies of power or spectrum that are attainable with certain choices of signal character and modulation method give comparable advantages with any reasonable system configuration.

Trainable ground antennas are necessary for all low and medium-altitude systems, and the systems here discussed are no exception. Satellite altitudes desirable in communication systems are high enough to require only low angle-tracking rates and accelerations, which pose no severe antenna-driving problems. If ground-antenna power gains are to be high, tracking must be quite accurate. Data and computation for acquisition and tracking of satellites are essential, but this seems to have been greatly over-rated as a source of problems in routine operations. Data needed can be largely reduced to an up-to-date table of clock settings at selected standard times, and can be kept highly reliable and accurate if attempts at long-term extrapolation are avoided. Computation is in the nature of angle-coordinate transformation between two polar-coordinate systems rotating with respect to one another. This can be done by a basically simple analog assembly of two clock-driven cranks, connected by a radius rod, through a two-axis gimbal with angle sensors, or it can be done digitally if desired. Once suitable means are provided and smooth operating routines are learned, tracking becomes a continuing but minor task.

State of the art takes an especially heavy toll in the area of cost of launch vehicles and their operation. The situation today is such that it literally makes little difference to overall cost per pound in orbit, even for very low orbits, whether the payload is fabricated out of common dirt or pure gold. Fuel consumption for launching and orbital injection is very heavy, but fuel costs remain less than 10 per cent of the total cost of a launched vehicle. Any avenue that will lead to significant reduction of cost in orbit is important. Expectation of some early gain in economy through development of multiple-payload launch capability is to be noted. Overwhelmingly important as this cost constraint is, it distinguishes among system configurations only by emphasizing the importance of holding down total weight in orbit for any

system. Since launched cost per pound-in-orbit increases with orbital altitude, severity of launching costs strengthens those system-tradeoff factors that favor lower altitudes.

Availability of the technology for maintaining relative positions (phases) of several satellites along one orbit is assumed in much of what follows. Such technology is generally considered feasible, but has not yet been tried in orbit.* It can lead to highly significant economy of satellites in a carefully configured system. In the alternative case of a randomly phased system, the statistical problem of determining and describing fully the system capability is an elaborate one. No clearly best general solution to this problem of performance description has yet emerged. This limits the confidence with which statements can be made regarding the performance and cost economy attainable in randomly phased systems.

User Station Operations

Systems which achieve the general-service capability aimed at here have the characteristic that the entire complex of satellites appears to each user station as a single, sky-girdling entity. Any satellite that can be seen (above a selected minimum elevation angle) by a user station can provide to that station the full service of this central-system entity. It can deliver to the user station traffic addressed to it by any other user station. Switching and routing to assure expeditious delivery of traffic to the station to which it is addressed are accomplished entirely within the central entity.

To receive continuous service, each user station must keep one antenna continuously trained on one satellite, and this is all that it need do. It can then examine the satellite output for traffic addressed to itself, and can feed its own traffic into the multiple-access satellite as desired. This means, of course, that the system must provide at least one satellite visible to each user at all times, and must keep all satellites energized as live elements of the system at all times. If interruption of service is to be avoided while shifting to a new satellite when the one in use is leaving the field of view, two trainable antennas are required for each user. To hold loss of service to an acceptable level during major maintenance on one of the two working antennas, a third antenna may be required. Since there are many hundreds of potential user entities in the world, simplicity of equipment and operating tasks at user stations can be very important.

Continuous contact with the central system entity makes it easy to

* Note added in proof: Syncom II has just provided an encouraging first trial.

provide each user with up-to-date clock settings and any other ephemeris data needed for satellite acquisition and track programming. Central supervision of individual choices of satellites to be tracked can also be provided easily, should it be desired. Routing chosen within the world-wide central system is immaterial to the user stations, except as some types of transmission may be sensitive to the different time delays associated with different routes.

Large satellite speeds and long transmission distances result in appreciable Doppler shifts and time lags. To avoid confusion in system operation, the convention may be made that nominal values of frequency and of synchronous-signal timing (if such is needed) will exist at the satellites rather than at the ground station. Each ground station then becomes responsible for offsetting the radio frequencies and pulse timing (if any) of its own output signals, in just the degree necessary to provide the chosen nominal values on the satellite with which it is working. Since the ground station knows at all times the range to and relative line-of-sight speed of that satellite, these corrections can be determined readily.

User-station operation under the pair-interconnection system concept contrasts sharply in many respects with that described above. For each desired link, each of the pair of users interconnected must be able to see a satellite located in a particular region of its field of view, and usually must track it with a separate antenna. For each user having widespread correspondence, this can necessitate a considerable number of simultaneously working antennas. Even when two or more correspondents can be reached by a user through one satellite, the times of beginning and ending of mutual satellite visibility are likely to differ between them, making satellite changeover a complex, stepwise process. Each user station must concern itself piecemeal with the routing of all its traffic. Coordinated dissemination of fresh ephemeris information to all users and coordinated routing and dispatching of system traffic are probably possible in a pair-by-pair system, but are not a naturally inherent capability.

Character of Central System

Operation of all satellites as a unified system entity available as a whole to any user station that can see any one satellite could, in principle, be attained by direct intercommunication among satellites. However, no start appears to have been made on the extensive development program needed for this. Such an approach would greatly complicate the satellites, particularly through the necessity of providing

switching equipment within them. Suitable technology undoubtedly will develop and it is possible that in time direct satellite-to-satellite interconnection, which can have real geometric virtues, may become a desirable method of providing a unified global communication system. At the beginning of the era of general communication service by satellites, however, alternatives that use simpler, hence more reliable, satellites seem preferable, and only such alternatives are studied here.

Attention is concentrated in this paper on systems that are organized analogously to postal services. A very few special, extremely capable ground stations are provided, in carefully chosen locations, as the core of the system. These function as major post offices, acting to concentrate, sort, forward and distribute all communication traffic. The satellites, which are kept simple, serve as outlying postal stations for initial collection from user stations and final distribution to users of all traffic. The satellites also serve in a quite distinct second capacity, providing a trunk-line relay function between post offices as needed. It becomes the collective responsibility of the few central post-office stations to maintain contact with all satellites in the sky at all times. In this way the need is met for all satellites to be continuously live elements of the system, in order that every satellite may be available to any user in sight of it at any time. All problems of traffic switching and routing are concentrated in the few major central stations on the ground; these also provide system supervision and universally available ephemeris-data broadcasts. System organization and traffic flow under this concept are displayed in Figure 1.

It is basic to the suggested concept of communication systems that user stations and post-office stations exist for different purposes. The former originate and terminate all communication traffic for the satellite system, while the latter simply perform an intermediary service. In general-purpose operation of the system, no user station communicates directly (i.e., via one satellite-relay hop) with any other user. Each user communicates, instead, only with a post-office station. If it can do so, each post-office station passes traffic from each of its user stations directly (that is, via a second single satellite-relay hop) to those other user stations to which the traffic is addressed. When relative location of user stations prevents such simple two-hop handling of traffic from originator to addressee, the collecting post-office passes traffic on to the appropriate distributing office, via additional satellite-relay hops between offices as needed. In principle, user and post-office functions can both be needed at the same location, but such dual-function requirements are not the general rule. In fact, it will become

evident later that geography, economic and population patterns, and system geometry usually conspire against such a joint function.

The residue of the function of trunk-line station-pair interconnection that remains essential for world-wide communication is reduced under the present concept to the relaying of traffic between central

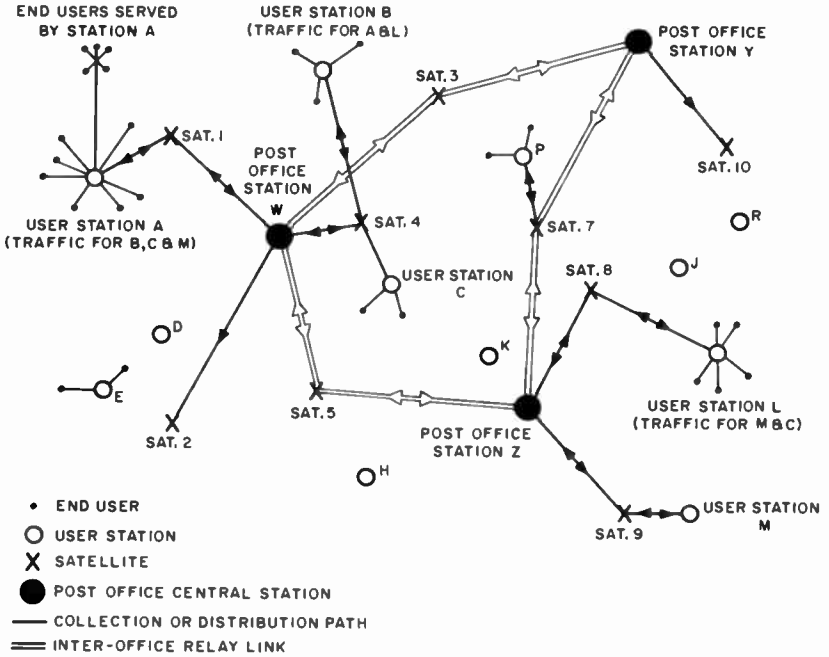


Fig. 1—Character of system organization and traffic flow.

post-office stations. Because this highly concentrated function is so important for smooth working of the total system, a third general requirement to be met in organizing the system emerges. This is that no post-office station shall ever be isolated entirely from communication with other post-office stations. If the net of post offices is rich in connectivity, alternate inter-office routes will exist. These will provide redundance to increase the level of assurance that the entire system will, indeed, function as a single unit at all times. The task of assuring continuous relay service between post offices, nevertheless, is usually the one that lays the severest requirement on the set of satellites in the sky. Should a normal, minimum-length inter-office link break down and longer alternate routes pick up its task, the only evident penalty

would be an increase in total transmission-time lag. In extreme emergencies, some user stations might even be pressed into service as auxiliary relay facilities, to preserve operating integrity of the system.

Each central post-office station must have a sufficient complement of trainable antennas, transmitters, and receivers to maintain continuous contact with all satellites visible from that station. This is in contrast to the requirement of even the heavy-traffic user stations for one antenna, one transmitter, and one receiver actually in use at each station. The post-office station, of course, must have some provision of antennas beyond those actually in use at any given instant, in order that interruption of service may be avoided during changeover between satellites and during routine maintenance of antennas. Extensive traffic-sorting and route-switching equipment is also needed at each post office. Overall system economy can be good, despite the heavy complement of equipment for each post-office station, because the number of such stations needed to serve the entire world can be kept very small.

Operation of a world-wide communication system organized along postal-system lines simplifies both the tasks of system coordination and the procedures and equipment at user stations, but there is a penalty. This is the double handling of all traffic, once into and once out of a central post-office station. There is no direct penalty on the user stations, only on the central system entity. Given adequate traffic capacity in the satellites, with efficient utilization of spectrum, it seems likely that the marked streamlining of operations for many user stations would more than justify the added burden on the central system. Inter-post-office relaying is a type of function that is not peculiar to the sort of system discussed here; it is needed to connect very widely separated users in any system (though some systems may one day substitute direct inter-satellite relays).

It has been customary in many preliminary system studies on communication via satellites to pick out one or two worst-case trunk-line paths, forcing the overall design to perform adequately under the difficult conditions of these cases. This tends to result in over-design of the system for its less difficult tasks in providing service over other, shorter trunk lines. The approach followed here is to exploit the symmetry of a rotating sphere in locating post-office stations, so as to avoid isolated most-difficult cases. If relay-path lengths between all pairs of nearest-neighbor post-office stations can be made the same, all relay tasks become equally demanding, and a balanced system design can result. Local overpopulation of post-office stations is avoided in this way, so that the total number of such stations necessary to serve the world can be kept low.

Satellite Pattern Considerations

System concepts based on full user-station utilization of any visible satellite, wherever it may be in the sky (above a specified minimum elevation angle), make the single-station satellite-visibility area an important working parameter. As indicated in Figure 2, this area is

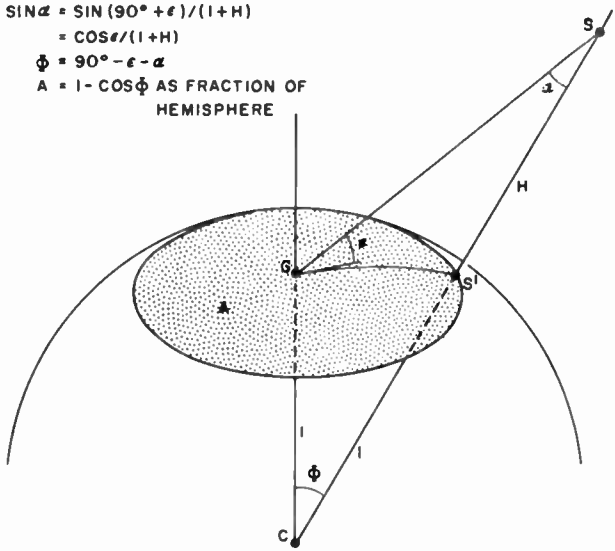


Fig. 2—Satellite viewing area of single ground station.

conveniently measured by the angle ϕ at the center of the earth C, between the radius through the viewing ground station G and the radius through the satellite S, when the latter is viewed at just the specified minimum elevation angle ϵ. All satellites with smaller central angles from G than ϕ are seen at elevations greater than ϵ, all those with larger central angles than ϕ appear below ϵ. ϕ evidently depends both on the threshold elevation angle ϵ and on the altitude H of the satellite, in accordance with the expressions shown in Figure 2. It is convenient to express H in units of one earth radius (3440 nautical or 3960 statute miles). Area A of a spherical polar cap of angular radius ϕ, the coverage area shown stippled in Figure 2, is conveniently expressible as a fraction of the area of one hemisphere, in the simple way shown on the figure. This is the area of useful satellite visibility from a single ground station or, conversely, the ground area usefully illuminated by a single satellite. α is the angular radius of the usefully illuminated area of the earth as seen from the satellite.

Areas of visibility turn out to be extremely large, even for rather low altitudes. The maximum possible visibility area for satellites at infinite altitude viewed down to the horizon is just one full hemisphere, with ϕ of 90°. An infinitely distant satellite can be seen above 5-degree elevation (ϕ of 85 degrees) from more than 91 per cent of one hemisphere. Figure 3 shows, as an example, A and ϕ versus H for 5-degree minimum elevation. At earth-synchronous altitude (5.62 earth radii),

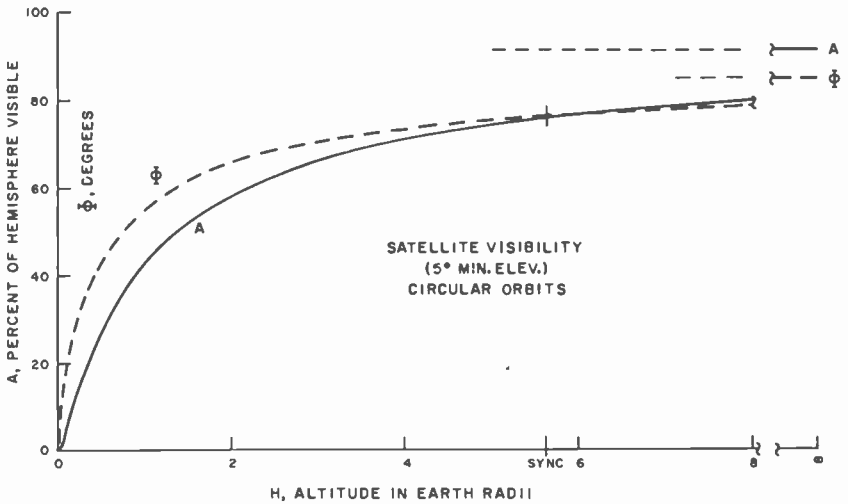


Fig. 3—Visibility-radius angle and viewing area versus satellite altitude.

with satellite period just 24 sidereal hours, ϕ is 76.4 degrees and A is 76 per cent of the hemisphere — ⅜ of the whole earth. Further reduction of H to 1.9 earth radii (6500 nautical miles) brings ϕ down to 65 degrees, and leaves viewing area at 57½ per cent of the hemisphere, still considerably more than ¼ of the entire earth. Even at only 0.36 earth radius (1240 nautical miles), ϕ is 38 degrees and A is still 21 per cent of the hemisphere, or more than 1/10 of the entire earth. Thus each satellite, as an outlying collection and distribution station of the post-office system, can always render very widespread service.

For the particularly demanding task of inter-post-office relaying, as for other pair-interconnection tasks, it is the mutual-visibility area over which a single satellite is simultaneously visible above threshold elevation from both of the stations to be connected that is important. This lens-shaped area (M in Figure 4) can readily be expressed in terms of the single-station visibility radius ϕ for a satellite and the

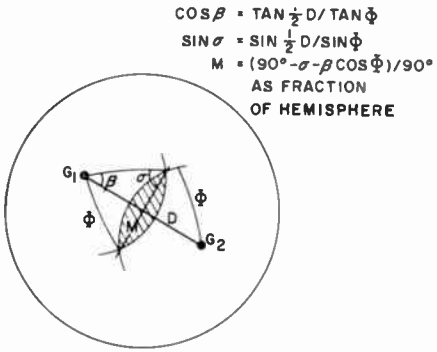


Fig. 4—Mutual satellite viewing area for pair of ground stations.

geocentric angular distance D between the ground stations G_1 and G_2 that require simultaneous visibility. The simple relations found, giving M as a fraction of one hemisphere, are shown in Figure 4. Dependence of mutual-visibility area M, in per cent of a hemisphere, on station distance D and visibility radius ϕ is shown graphically in Figure 5. Because of possible short dwell of moving satellites in the sharp cusps

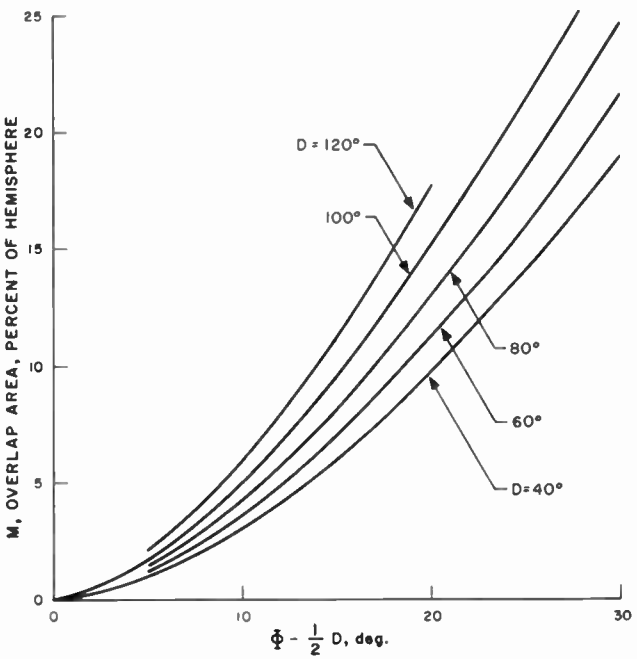


Fig. 5—Mutual-visibility area versus angular station separation and single-station visibility radius.

at the ends of the lens-shaped area, the practically useful area of mutual visibility tends to be somewhat less than the values given here.

Certain values of single-station viewing angle ϕ mark boundaries between somewhat different behavior regions of sets of satellites. Values of 30, 36, 45 or 60 degrees, for example, set the satellite-altitude boundaries above which a single station on the equator could be assured of always seeing at least one satellite in an equatorial orbit if there were, respectively, 6, 5, 4, or 3 satellites moving equidistantly along that orbit. Symmetry considerations in laying out sets of post-office central stations will be seen later to set up other sets of bounding values of viewing angle, lying between 20 and 71 degrees, with the values 38 and 55 degrees of particular significance. A single-station angle of view of 68 degrees, comfortably above both ground-determined 55-degree and sky-determined 60-degree bounding values, is used later to generate an example of a system configuration that is economical both on the ground and in the sky.

Patterns of satellite motion in the sky can be described in various ways that are physically equivalent but provide different points of view. It is usual to think of a satellite in a circular orbit about a spherical earth as moving at constant speed along a circular path in a plane that is fixed among the stars and passes through the center of the earth. This orbital plane is in general inclined at some angle i with respect to the equatorial plane of the earth and crosses the equator at two nodal points characteristic of the particular orbit. Rotation of the earth complicates the apparent path of this satellite in the sky, as seen by observers at points fixed on the surface of the earth. The satellites are seen from the rotating earth as weaving basket patterns in a succession of passes across the sky.

Alternatively, considering a nonrotating earth, attention can be fixed on the motion of the satellite relative to an earth-meridian line that is moving uniformly in longitude, with a period just equal to the sidereal period of the satellite. Figure 6 indicates the result; the satellite is seen to progress along a figure-eight locus centered on the moving meridian, completing one full circuit of the locus in each full rotation of the reference meridian. The extent in latitude of the figure eight, equal for its north and south loops, is just i, the angle of inclination of the orbit. The maximum width of the figure eight in longitude is $\sin^{-1} [\sin^2 i/(1 + \cos^2 i)]$, and this maximum width occurs at latitudes, north and south, of $0.71i$. Placing more than one satellite in the orbit depicted results simply in generating a separate figure-eight locus for each satellite, the loci following one another in their steady march around the earth. Placing one satellite each in

several orbits at the same inclination, which cross the equator at various longitudes, with all satellites phased to the same moving reference meridian, results in one figure eight, with all satellites following one another around it as it wends its uniform way around the earth. Moderate eccentricity of orbits merely distorts the figure-eight loci. Rotation of the earth merely makes the apparent period of travel of a

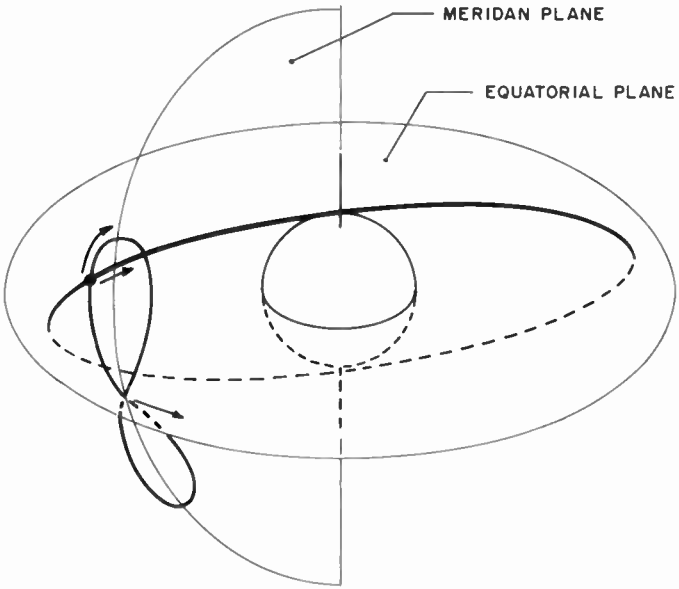


Fig. 6—Satellite motion relative to moving meridian.

figure eight around the earth differ from the period of a full satellite circuit of the figure eight. Oblateness of the earth modifies slightly the period of earth rotation relative to the satellite orbit, and adds some distortion to the shape of the figure eight.

Visualization of satellite-visibility coverage often can be facilitated greatly by the above alternative formulation of satellite paths in the sky. Take the case of 3 equally inclined orbits, with ascending nodes equally spaced around the equator, and with 3 satellites equally spaced in each orbit (which calls for relative-station-keeping capability). For equal satellite phasing in each orbit, the total satellite configuration may be visualized as in Figure 7. The satellites are seen to be grouped in 3 equally spaced figure-eight loci, with 3 satellites to each locus, and

with the loci progressing uniformly from west to east across the sky, as the individual satellites circulate around their respective loci.

By drawing a few spherical polar caps of 65-degree angular radius at various locations on a globe, it can be verified readily that no point on earth ever fails to have in view at least one of the 9 satellites of Figure 7, for the case of orbits inclined at 50 degrees with respect to

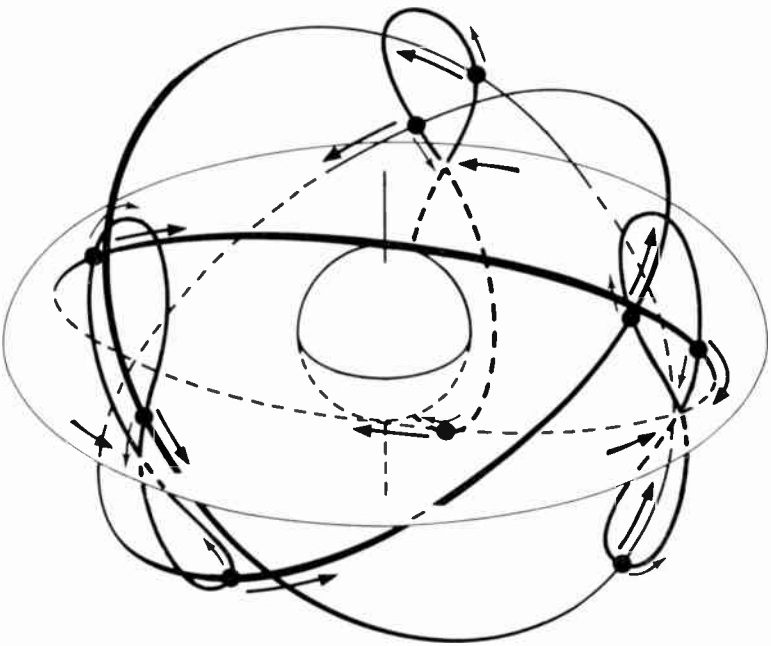


Fig. 7—Over-all motion of 9-satellite isochronous configuration.

the equator (with maximum figure-eight width 25 degrees in longitude, at 35-degree latitude). Thus, this simple 3-orbit, 9-satellite isochronous configuration, with all satellites keeping station relative to each other, fulfills with certainty the stated system requirement of making satellite service continuously available to every user station, wherever it may be. This is accomplished even at the modest altitude of 6500 nautical miles. It may be noted that the satellite pattern shown provides at all times 3 satellites, spaced 120 degrees in longitude, in the north-latitude range of 25 to 50 degrees, 3 in the corresponding range of south latitude, and 3 satellites 120 degrees apart in longitude in the equatorial belt between 25 degrees north and 25 degrees south latitude.

At the other extreme from the fully determinate procession of rising

and setting figure-eight loci just described, satellite-population problems for world-wide service can be approached on a purely statistical basis. For the one-dimensional case of a single belt of satellites in substantially circular orbits in the earth's equatorial plane, with orbital periods distributed randomly with constant density over a tolerance range, the problem of specifying the statistics of seeing or not seeing a satellite from a given ground station is fully soluble. As a matter of fact, the problem becomes determinate rather than statistical after the periods of the satellites have been determined to extreme accuracy, but the statistical description of the outages of satellite visibility may remain a useful one.

When one is concerned with outages of visibility from a given ground station for satellites all over the sky, the statistical solution becomes more involved, and is not known to have been worked out completely to date. Again, once all orbital parameters of all satellites have been determined accurately, the problem is really a fully determinate one, but one may hope that the properties of a statistical solution may again provide a simple but useful summary description, approximating the overall "pseudo-random" behavior of the real system. The situation is complicated by the fact that there probably exists no very simple recipe for arranging satellites and their orbits in such a way that the long-time average density of satellites per unit area of the sky is uniform over the whole sky.

For any given satellite, in an oblique orbit of specified inclination, the entire satellite motion is confined between limits of latitude, north and south, equal to the orbit-inclination angle. Within this latitude band, the probability of finding the satellite within a given small area of the sky depends strongly on, and increases with, the latitude of that area. It is not dependent, however, on the average, on the longitude of the area, so long as the period of the satellite and the length of the day are not simply commensurable. The concentration of satellites in high latitudes is particularly notorious in the case of polar orbits. Given a sufficiently large desired population of satellites, it should be possible to find a recipe for so distributing the inclinations and nodes of their orbits as to approximate well a uniform average density over the sky. For modest numbers of satellites, however, no assurance is yet at hand that such a recipe can exist to the degree of providing a practically valuable pseudo-random result.

What is often done in the face of the difficulties just described, and will be done here, is to assume that an effectively random long-term satellite distribution with uniform mean density over the entire sky can in fact be produced, even for a moderate number of satellites.

While far from rigorous, this assumption does permit some very simple results to be attained. If V is the satellite-visibility area as a fraction of a hemisphere, whether the single-station visibility area A as defined in Figure 2 or the two-station mutual-visibility area M as defined in Figure 4, then the number N_0 of such areas required just to cover the sphere is $2/V$. One satellite placed at random in the sky has a probability $V/2$ of being usefully visible (above the chosen elevation

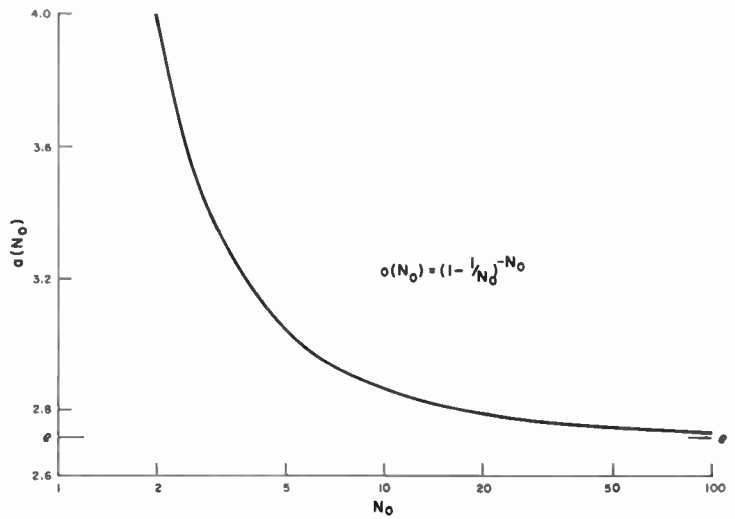


Fig. 8—Dependence of $a(N_0)$ on N_0.

threshold) from any particular ground point or pair of points and a probability $1 - V/2$, or $1 - 1/N_0$, of not being visible from there. If there are just N_0 satellites placed at random in the sky, the probability that no satellite is visible from the selected ground point or point pair is $(1 - 1/N_0)^{N_0}$. This quantity, which may be called $1/a(N_0)$, has the property that as N_0 becomes large, it approaches the limit $1/e$, where e is the base of natural logarithms. For a full hemisphere of visibility, N_0 has its smallest significant value of 2, and $a(N_0)$ is just 4; for a visibility area of 20 per cent of a hemisphere, with an N_0 of 10, $a(N_0)$ becomes 2.868, which is still 5.5 per cent above the asymptotic value e. Figure 8 shows the variation of $a(N_0)$ with N_0.

If N satellites are placed in the sky at random, the probability that a given ground station or station pair for which each satellite at the altitude used provides a hemisphere-fractional coverage area V will not see even one satellite is just $(1 - V/2)^N$, or $(1 - 1/N_0)^{N_0(N/N_0)}$,

where N/N_0 may be termed an "overfilling factor", F. It does not matter that N_0 and F may not be integers, so long as the actual number of satellites, N, is an integer. For large N_0, the probability of a service outage due to lack of any satellite in view is, thus, approximately e^{-F}. If 1 per cent of outage is tolerable, F must then be made at least 4.6, or 6.9 for 0.1 per cent tolerable outage. For more modest N_0, outage probability is $[a(N_0)]^{-F}$, a slightly more favorable condition, since $a(N_0)$ always exceeds e.

Consider the task of relaying between ground stations 64 degrees apart via satellites with a single-station visibility radius of 42 degrees. The reason for choosing the 64-degree transmission-path length will become evident later; the 42-degree visibility radius is close to an economic optimum for a passive-satellite system working over the stated distance. Figure 3 indicates that the satellite altitude needed for 42-degree visibility at 5-degree minimum elevation is 1570 nautical miles, while Figure 5 shows the mutual-visibility area to be 3.7 per cent of a hemisphere, so that N_0 becomes 54. $a(N_0)$ is then 2.74, very close to e, so that the overfilling factor F needed for 99-per cent continuity of relay-link service is 4.6, and the total random population N of satellites required is 250 (for 99.9-per cent continuity, 375 satellites would be needed). For a population this large, a considerable degree of uniform-density randomness may be achievable.

Consider, finally, the task of relaying between stations separated 90 degrees, with satellites which provide 65-degree single-station visibility. Figure 5 indicates a mutual-visibility area of 14 per cent of a hemisphere, or an N_0 of 14.4. Achieving 99-per cent service continuity with an N_0 of 14.4 for which $a(N_0)$ is 2.81, calls for an overfilling factor of 4.45, or a total of 64 random satellites scattered over the entire sky, a number that may still be large enough to lend some credibility to the assumption of uniform probability density. Comparing this random sky population with the 9-satellite isochronous configuration of Figure 7, the great advantage to be gained in satellite economy by using the technique of relative-station keeping is evident. Comparing further with the 250 satellites needed at much lower altitude, the incentive for longer-haul, higher altitude operation is also evident. These economies, most readily attainable with active satellites, which can individually be very light, tend to offset the high traffic capacity and complete availability of access so readily attainable with passive satellites, which tend also to be very heavy, especially when they must work over very long distances. Because two-station areas of common visibility are usually much smaller than single-station visibility areas, the task of maintaining intact relay paths between post offices is

usually the most demanding one. This more difficult task has been examined above rather than the easier task of keeping one satellite visible to each user.

Location of Post-Office Stations

Desirable location patterns for the few large post-office central stations emerge from interaction between the possible regular arrays of points on a sphere, the axial symmetry of the rotating earth, and the haphazard location (on the predominantly watery earth) of land on which to base such stations. Because of the very large region of the world served by each such station, local concentrations of population and commerce can have little effect on the proper choice of the station locations. Gross global population patterns can bias system configuration, however, particularly by indicating large regions of the sky in which rigorous adherence to the stated system goal of keeping every satellite continuously a live element of the central system may not be defensible economically.

Regular arrays of points possible on a sphere are limited to 7, of which 5 are the patterns of vertices of the only nondegenerate regular solid figures that are possible in Euclidean solid geometry: the tetrahedron (4 faces, 4 vertices, 6 edges), cube (or hexahedron, with 6 faces, 8 vertices, and 12 edges), octahedron (8 faces, 6 vertices, 12 edges), dodecahedron (12 faces, 20 vertices, 30 edges), and icosahedron (20 faces, 12 vertices, 30 edges). No more-complex or finer-meshed regular arrays of points are possible on a sphere. To these 5 must be added the vertex patterns of 2 degenerate regular solids, not usually mentioned as such because their volumes are zero: the equilateral triangle (2 faces, 3 vertices, 3 edges), and the finite line segment (0 faces, 2 vertices, 1 edge). The 5 regular solids are complementary in pairs, in that an icosahedron and a dodecahedron can be so intermeshed that all vertices of one lie at the centers of the faces of the other, and the same holds true for an octahedron paired with a cube, and for a tetrahedron paired with another tetrahedron. The 2 degenerate cases can also be paired off.

Characteristic angular dimensions of the 5 regular spherical figures are displayed in Figure 9. Each sketch in Figure 9 shows one face of one figure; arrows point toward adjacent vertices not shown. It is convenient here to remember that one minute of geocentric angle corresponds to one nautical mile along the surface of the earth. Edges of polyhedron faces represent lengths of inter-post-office relay paths, the D values needed to find mutual-visibility areas from Figure 5. Radii

of faces, from center to vertex, represent coverage radii necessary for gapless traffic collection and distribution service. Each station at a vertex of a complete icosahedron has 5 nearest neighbors, distant by relay paths of equal length, while a station at a corner of a dodecahedron relays to 3 equally distant nearest neighbors. Correspondingly, stations located at vertices of an octahedron each have 4 nearest neigh-

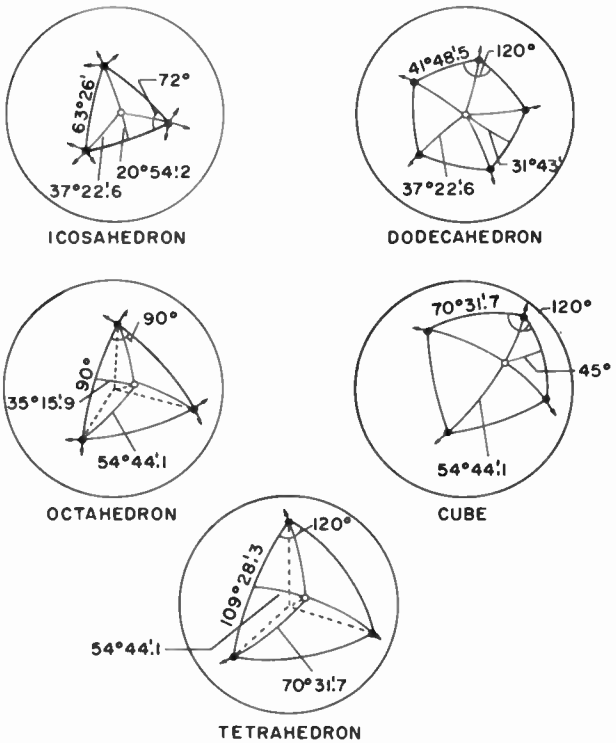


Fig. 9—Face dimensions of regular figures on sphere.

bors; stations at vertices of a cube have 3 nearest neighbors, as do those at vertices of a tetrahedron.

Angular dimensions given in Figure 9 set boundary values on regions of visibility radius at which conditions to be met by satellite configurations change somewhat, as was mentioned earlier. With a full octahedral net of post-office stations, for example, all satellites above the altitude for which Φ is 55 degrees can be continuously illuminated, wherever they may be; at lower altitudes, some satellites will lose post-office contact. Again, for a full octahedral ground net,

satellites with visibility radius over 45 degrees may provide relay links between post offices; satellites with less visibility cannot do so.

Regular point arrays can be made to interact with actual world distribution of islands and other land masses by constructing 5 symmetrical strip harnesses, representing the sides of the 5 regular polyhedra, to fit snugly but not tightly over a geographic globe. Availability of land bases for stations at the vertices of a chosen figure can then be investigated by trial and error, by sliding the harness for that figure around on the globe. This was the method followed in working up system examples for this paper. It should be noted, however, that station locations so discovered are subject to the errors in drafting and construction of the globe used. While a search for land bases is in progress, simultaneous attention to the graticule of latitude and longitude coordinate lines on the globe permits study also of the relation of the symmetrical arrays of points to the polar-axial symmetry of the rotating earth.

Configurations given particular study by the above method, for the communication-system problem, have been the octahedron, with vertices marking out 6 uniformly spaced stations that lie at the centers of the 6 regular quadrangular faces of a spherical cube, and the icosahedron, with vertices marking out 12 uniformly spaced stations at the centers of the 12 regular pentagonal faces of a spherical dodecahedron. The station configurations complementary to these, at the vertices respectively of a cube and a dodecahedron, require each station to serve a face of the same maximum radius as the one studied, but take more stations for total coverage. Relay paths are shorter in the less-studied cases, however. Tetrahedral station arrangements have faces too large and sides too long for effective use with medium-altitude satellites. The degenerate cases, with 3 or 2 station locations, tend to give incomplete coverage even for satellites at synchronous altitude.

Many seemingly useful ways to configure general-use communication systems have been found in these studies, and it is difficult to choose among them. Two particular system arrangements have been selected, however, for presentation as examples here. One of these is suited for altitudes that seem to require active-repeater satellites; the other is configured to accept the altitude limitation appropriate to use of passive-reflector satellites.

System Examples

One example of a system which fits the general operating concepts developed here has been well known for some years. It employs 3 satellites maintaining equal spacing in a circular equatorial orbit, at such

an altitude (19,300 nautical miles) that their motion in orbit is synchronous with the rotation of the earth. On the ground, 3 major central stations equally separated in longitude are provided for relaying. With the exception of triangular polar areas that receive no service at all, any user station anywhere on earth can see at least one satellite at all times. All 3 satellites are illuminated at all times by the 3 post-office stations, and the post-office stations are connected in pairs by relay links that are never interrupted. This system has the special virtue that ground antennas need not have tracking capability, and each post-office needs only 2 working antennas. If it is desired that the polar gaps be filled, at least 2 additional satellites are required, in inclined orbits, and at least some ground antennas must be made capable of tracking.

As a first example of a medium-altitude system that follows the concepts developed in previous sections of this paper, some lines of thought brought out earlier will now be examined further. It was pointed out that if the satellite altitude exceeds 6500 nautical miles (single-station visibility radius 65 degrees), a configuration of just 9 satellites, keeping relative stations in 3 orbits inclined 50 degrees, or equivalently circulating around 3 figure-eight loci equally spaced and uniformly progressing in longitude, as in Figure 7, will assure that no point on earth ever lacks line of sight to at least one satellite, at elevation always in excess of 5 degrees. It was also pointed out earlier that 6 ground stations located at the vertices of a regular octahedron will provide jointly a line of sight, elevated more than 5 degrees, to any satellite anywhere in the sky, for any satellite altitude above 3500 nautical miles (single-station visibility radius 55 degrees). Here, evidently, is a basis for a very economical system, though special care proves necessary to assure compatibility between maximum economy of satellites and good continuity of relay links.

Symmetrical placement of 6 ground stations on the rotating earth would require stations at both poles, but no land on which to base such a station exists at the North Pole (another symmetrical alternative, seemingly less attractive than the polar one, will not be pursued here). On the real earth, the land nearest the North Pole is Peary Land, at the northern tip of Greenland, with maximum latitude 83 degrees. This is a most unusual place, being essentially a polar desert. It is completely free of ice overlay over a considerable area, and has negligible snowfall because of its arid climate. The area seems remarkably well suited for a major ground installation. This location being 7 degrees away from the pole means that a visibility radius of 65 degrees in the direction of Australia would only extend down to latitude 32

degrees north. With 9 satellites grouped at 3 equally spaced longitudes, in 3 orbits inclined 50 degrees, even so small a polar offset prevents the 3 northernmost satellites from being always illuminated by this ground station. Merely increasing the orbit inclination, to insure illumination of one satellite in each locus at all times by the polar station, would spread the satellites out so that there would be times at which some points on earth could not see even one satellite.

In seeking a satisfactory system compromise, a number of parameters available for adjustment are: number of satellites, grouping of satellites, orbit inclination, orbit altitude, number of post-office stations, and location pattern of post-office stations. Since placing satellites in orbit and replacing them as they fail is very costly, the number of satellites used will be held to its minimum value of 9 for this example. They will still be grouped, for maximum serviceability, in 3 figure-eight loci, 120 degrees apart, with phase of circulation in the locus advanced uniformly from locus to locus. Orbit inclination will be made 58 degrees, and single-station visibility radius (for minimum elevation 5 degrees) will be made 68 degrees. This slight increase in visibility radius requires that satellite altitude be increased from 6500 to 8260 nautical miles. Every point on earth can again see at least one satellite at all times. The central line of a figure-eight locus will pass overhead every 4 hours and 40 minutes, at any particular ground location.

The stippled area in Figure 10 shows the latitude belt from 58 degrees north to 58 degrees south, within which satellites can pass directly overhead. Division of the satellite belt into 3 zones, in each of which there is always one satellite in each figure-eight locus, is indicated by the dotted parallels of latitude in Figures 10 and 11. These zones run from 58 to 29 degrees north latitude, 29 degrees north to 29 degrees south, and from 29 to 58 degrees south latitude. The 3 equally spaced figure-eight loci of satellite circulation, progressing uniformly across the sky from west to east, are also indicated; 3 satellites are shown circulating in each figure eight.

Figure 10 shows also, as a cross hatched area, the 68-degree-radius visibility pattern of a single north-polar post-office station at Cape Morris Jessup, Greenland. It is evident that the entire northernmost zone of the satellite belt is illuminated at all times. At least one satellite in this zone can be seen at all times from any user station located anywhere further north than 8 degrees north latitude, the parallel shown dashed in the figure. That is, unbroken service can be rendered by this one post-office station, through the 3 satellites which it is currently illuminating, to an overwhelming majority of the population, and even more of the industry, of the entire earth. The capability for

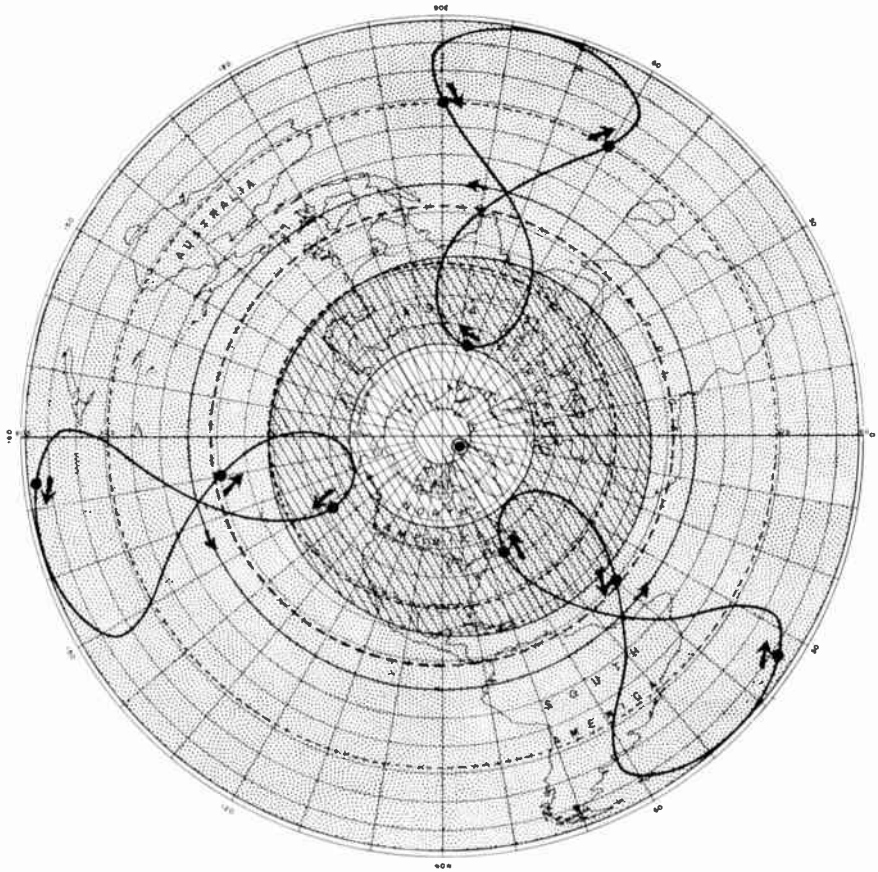


Fig. 10—Coverage of system using 9 station-keeping satellites at 8260-nautical-mile altitude and one post-office station.

service of one far-northern ground station working through satellites is something quite unique in communication.

Consider next the addition of a ring of 4 post-office stations along the equator, equally spaced at 90-degree intervals of longitude, to provide the full coverage shown in Figure 11. Land on which to place these stations is not plentiful over wide stretches of the watery equator, but geometrically suitable locations do exist at Christmas Island in the Pacific Ocean, at the center of Borneo, in the Congo near Stanleyville, and near the northwestern corner of Brazil. Together, these 4 additional stations suffice by themselves to illuminate continuously every one of the 9 satellites in the sky. Relay linkage over all paths among the 5 stations is not continuous, however. Any one of the equatorial stations can temporarily lose contact with the polar station,

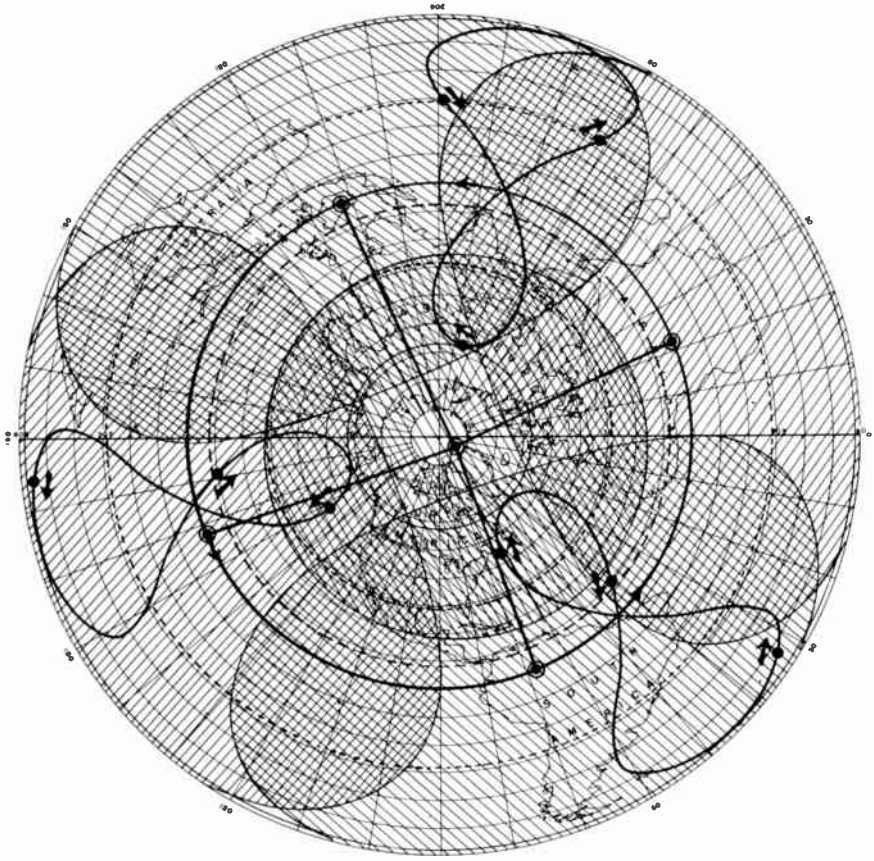


Fig. 11—Coverage of system shown in Figure 10 when four post-office stations equally spaced around the equator are added to the system.

or with its equatorial neighbors. Only one equatorial station can lose polar contact at a time, however, and the station that is out of contact with the pole is solidly connected to its equatorial neighbors for just the time that it lacks polar contact. All conditions for operation with the central system of satellites and post-offices appearing to all users as a single entity, always available, are met in this example.

Every satellite can provide traffic collection from and distribution to user stations; those over singly cross hatched areas of Figure 11 can do no more. Satellites over areas showing double cross hatching are in view of two post offices, so can serve also for relay links. Triply cross hatched areas indicate satellite serviceability in any of 3 relay

links. The sixth post-office station, near the South Pole, that would complete the octahedron, is not needed, unless the strong redundance of relay paths it could provide is considered worth the added cost. Unlike the north-polar station, which single-handedly can serve most of the world's communication needs without need for any relaying, a south-polar station would itself provide very little traffic-collection and distribution service.

With economy in number of satellites there goes economy in antennas at post-office stations. Each station can see at most 4 satellites at one time, and that number is reached only briefly; all needs, including satellite changeover, can be met fully by 4 antennas in working condition per station. Economy of satellites also means severe disruption of service when one satellite fails. Some safety factor in satellites aloft is a practical necessity. Geometrical requirements for economical system design can be seen to be strongly incompatible with the idea of locating post-office stations at major traffic-originating centers.

As a final example, a system constrained to work at limited altitude and with uncontrolled satellite motion will be examined. This is appropriate for use of passive-reflector satellites. The example is based on the symmetry of an icosahedron-vertex array of 12 ground stations, for which the inter-post-office relay-link length is 64 degrees. Altitude-dependent trade-offs affecting total cost of satellite launching indicate a single-station visibility radius of 42 degrees, with a satellite altitude of 1570 nautical miles, to be particularly economical in linking station pairs separated 64 degrees. The cross hatched areas in Figure 12 indicate coverage patterns for a mid-latitude belt of 5 stations configured on this basis. Suitably related land bases for these stations are found at Lisianski Island in the Hawaiian chain, on the northwest corner of the island of Mindoro in the Philippines, in the southeast corner of Arabia north of Murbat, at the island of Madeira off the west coast of Africa, and near Wichita in Kansas. Since a uniformly dense random distribution of satellites over the whole sky is assumed, no satellite-coverage pattern is shown. However, each 42-degree station-coverage circle shown is just the size of the area of the earth usefully seen by one satellite.

Satellites over the areas not cross hatched on Figure 12 cannot be illuminated by the 5 post-office stations shown. However, user stations within 22 degrees of the perimeter of illumination have a probability of seeing and using illuminated satellites that is equal to the probability chosen for existence of relay links between post-office stations. The north-polar illumination gap thus gets full pick-up and delivery service, equal in quality to the relay-link service. The same is true of the area

within the 5-sided perimeter shown in heavy lines to the south of the illuminated area. Going south from this perimeter, in the stippled area, continuity of service degrades rapidly. Substantially the whole world population can be seen to receive complete communication service, through local user stations placed as needed, from these 5 central

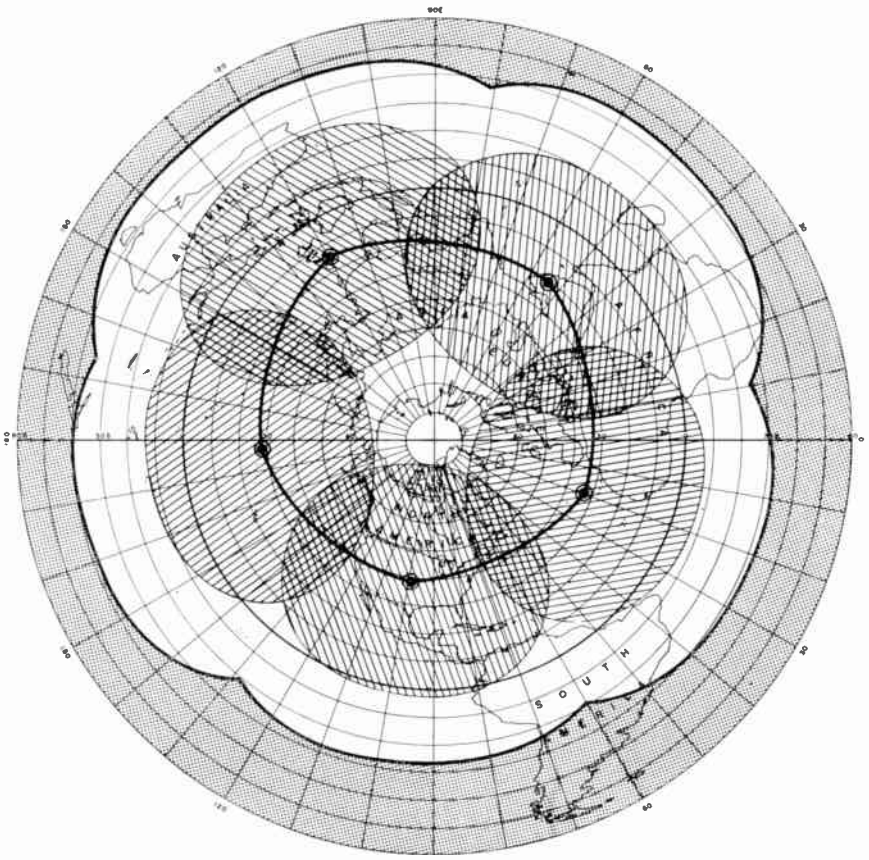


Fig. 12—Coverage of system using 250 random satellites at 1570-nautical-mile altitude and 5 post-office stations.

ground stations alone. This is accomplished in spite of the quite low altitude of the satellites, but does require the large number of satellites (250) needed to be consistent with random distribution over the whole sky.

While solid coverage of the entire northern hemisphere is provided by just 5 post-office stations, service to the more northerly regions

overland links; any post office has 0.01 per cent probability of isolation.

c. (Optional) Three additional post offices, at other icosahedron vertices. These will remedy all service lacks to the entire world, and reduce the probability of belt-post-office isolation to 10^{-6} per cent.

The well-known 3-satellite synchronous system with 3 ground relay stations also meets the stated goals, with a minimum of the double traffic handling necessitated by goal 11.

SUPERCONDUCTIVE ASSOCIATIVE MEMORIES*

By

Richard W. Ahrons

RCA Laboratories,
Princeton, N. J.

Summary—This paper covers the broad area of superconductive associative memories using cryotrons. A survey of existing read and write processes and circuit arrangements is presented, followed by specific circuit improvements and some logic suggestions. These include: (1) a cryotron implementation of a destructive-readout associative memory with parallel readout that contains a minimal number of cryotrons, (2) a cryotron implementation of a new read-out scheme called the Modified Lewin, (3) a comparison of speeds of operation inherent to the Modified Lewin and parallel-readout schemes, (4) circuit methods of increasing this speed, and (5) algorithms for use with a parallel-readout associative memory in greater-than, less-than, and between-limits comparisons.

IN THE INTERNAL MEMORY of most present-day computers, the random-access type memory employing an address to locate data (e.g., the magnetic-core memory) has replaced the serial-access memory employing a serial scan for the appropriate data (e.g., magnetic-drum or disc memories). Random-access memories have not as yet replaced the magnetic tape in semi-permanent external stores because of economics (cost per bit). The next step in basic memories may be the "more-intelligent" associative memory which does not require a fixed position, or address, for the stored information.

Two observations may be made in regard to random-access memories: (1) a large percentage of programming time and effort is usually spent in assigning and keeping track of addresses and (2) in so far as is known, the human memory allows information to be retrieved by an associative process without regard to physical location of information. Thus, the conventional technique of requiring an address for each physical word location may be neither the most natural nor the most convenient technique for storing information.[1] Associative memories should not be regarded as attempts to extend the random-access

* Taken from the dissertation submitted to the Faculty of the Polytechnic Institute of Brooklyn in partial fulfillment of the requirements for the degree of Doctor of Philosophy, 1963. Supported by the Office of Naval Research (NONR-3879-(00)).

[1] J. A. Weisbecker, personal communication.

memory process, since it eliminates the address and the bookkeeping of the addresses. Thus the associative-memory concept should lead to a change in the philosophy of memory programming and not merely an extension of present techniques. It is reasonable to believe that the associative form of memory will be applicable to future computers that are intended for the more-sophisticated learning and thinking processes.

The basic function of an associative memory (also called catalog, parallel-search, content-addressed, or data-addressed memory) could be defined as a memory with the ability to answer this question: Is this piece of information in the memory? However, since this interrogating information is already known, the information generally

Table I—Types of Associative Memories

	Name	Portion of Word Interrogated	Portion of Interrogated Segment Required for Interrogation
(1)	Fixed Tag	Fixed Segment	Total
(2)	Multiple Fixed Tags	Any Set of Fixed Segments	Total
(3)	Variable Tag	Fixed Segment	Any Part or Total
(4)	Fully Interrogable	Any Segment	Any Part

desired is that which is associated with the original interrogating information. Thus the memory that only answers the above question is the most elementary form of associative memory. McDermid and Petersen[2] have divided the more-advanced forms of associative memories into four categories. A modified list of these categories in order of their complexity is as follows (see Table I):

(1) *The fixed-tag associative memory* has one fixed segment of the word set aside for interrogation and all of that segment must be used in the interrogation. The remainder of the word is read as the desired information. Thus only a part of the memory is associative.

(2) *A multiple fixed tag* describes a memory whose words are divided into segments any of which can be used as the interrogated portion. All the bits in the interrogated segment must be used in the interrogation.

[2] W. L. McDermid and H. E. Petersen, "A Magnetic Associative Memory System," *IBM Jour. of Res. and Dev.* Vol. 5, p. 59, Jan. 1961.

within the 5-sided perimeter shown in heavy lines to the south of the illuminated area. Going south from this perimeter, in the stippled area, continuity of service degrades rapidly. Substantially the whole world population can be seen to receive complete communication service, through local user stations placed as needed, from these 5 central

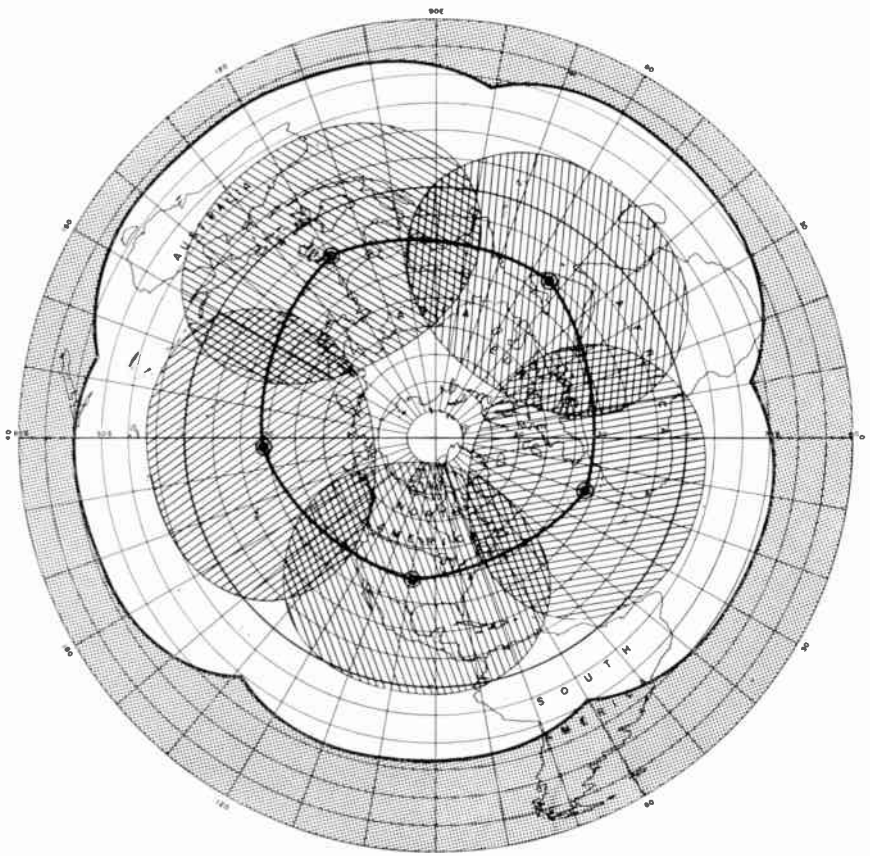


Fig. 12—Coverage of system using 250 random satellites at 1570-nautical-mile altitude and 5 post-office stations.

ground stations alone. This is accomplished in spite of the quite low altitude of the satellites, but does require the large number of satellites (250) needed to be consistent with random distribution over the whole sky.

While solid coverage of the entire northern hemisphere is provided by just 5 post-office stations, service to the more northerly regions

such as Siberia, Scandinavia, and northern Canada requires relaying around the mid-latitude ground-station belt. Also, simultaneous outage of only two relay links due to poor satellite positions can isolate a post office from the system, creating a serious hiatus in service. New Zealand, Argentina, and Chile are not well served, and must make

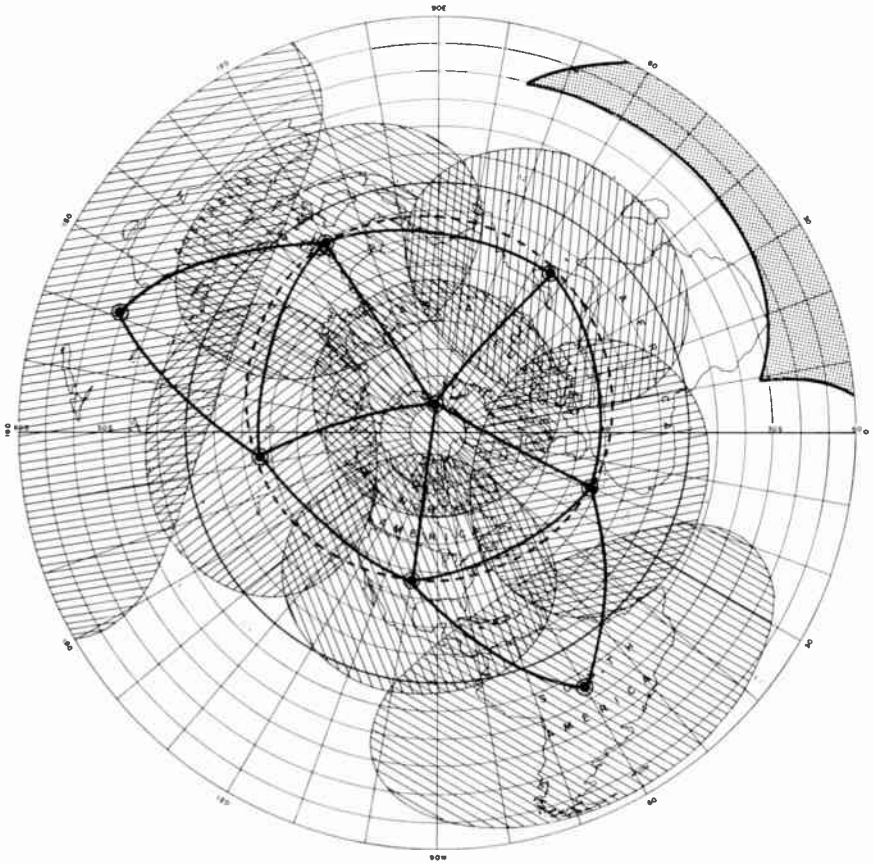


Fig. 13—World coverage with system using 250 random satellites at 1570-nautical-mile altitude and 8 post-office stations.

heavy use of land lines. Addition of 3 more post-office stations, at other vertices of the basic icosahedron pattern, can remedy all these shortcomings, if this is adjudged worth the added cost.

Figure 13 shows 8-station coverage, with the added stations located on the Taimyr Peninsula of northern Siberia, on Lord Howe Island off the east coast of Australia (which is a little out of symmetry, being

too close to Mindoro), and on the border of Brazil and Bolivia, between Trinidad and Cuiaba. The unilluminated sky area around the north pole has been eliminated, and the south-polar unilluminated area so reduced that the remaining region of less than relay-link-grade service has shrunk to a small antarctic triangle, which includes no land but a small part of Antarctica; only a portion of this triangle shows in the figure (stippled area). The entire area north of the dashed line in the figure can now be served directly by the Taimyr Peninsula station alone, with no relaying at all.

Perhaps even more important is the added assurance of system integrity that is provided by the additional stations. Taimyr is served by 5 relay links. Of the 5 belt stations, still the backbone of the system, only Arabia is served by less than 4 links (and that could be remedied, if desired, by filling in the missing station at the St. Paul Island vertex, in the southern Indian Ocean). If the probability of single-link outage is designed to be one per cent, the probability of isolation of a belt station by simultaneous random outage of 4 links becomes one in 10^8. Coupling such basic statistical reliability with the very great traffic capacity and complete availability of access provided by passive-reflector satellites would result in a system of truly remarkable capability.

When large numbers of randomly distributed satellites are used to provide statistical sky coverage, strict adherence to the concept of illuminating all satellites at all times leads to an excessive total complement of ground antennas for all post-office stations. What is really essential is assurance that no gaps in coverage occur because satellites are left unactivated. In those portions of the sky which happen to have their normal quota of random satellites or more, most of the satellites can be left inactive without degrading system continuity. Only in statistically sparsely populated sky areas is full illumination needed. Illumination of one satellite for each relay link is an essential minimum, with equivalent additional coverage required to serve user stations in sectors with no relay links.

The system of the last example, with land bases available for 10 out of a possible 12 stations, does permit practically total sky illumination if desired. The 42-degree-radius cap of sky covered by one post office, however, covers 25½ per cent of a sky hemisphere. With 250 satellites needed for 99-per-cent relay-link continuity, each post office can see on the average 32 satellites, and random density fluctuations can easily produce occasions with 45 satellites in view. To provide full illumination under such over-dense conditions would require at least 360 trainable antennas for 8 post offices, with little advantage gained over the total of about 80 antennas (plus spares) needed for the

minimum essential illumination of about 10 satellites per post office.

With only partial satellite illumination, user stations would have to select working satellites on about a one-out-of-three basis. Once an initial acquisition was accomplished, this could be made easy by broadcasting from the post offices their planned schedule of selective illumination, over the same channel used for supplying up-to-date satellite-ephemeris data to all user stations.

Many other ways can be thought of to configure systems in accordance with the operating and geometrical concepts used here. It is hoped, however, that the examples given may suffice to indicate the scope of applicability of these concepts. Another particular area of applicability lies in exploitation of the special geometrical characteristics of equatorial-orbit satellite belts and of the circumpolar concentrations of satellites achievable with polar-orbit satellite belts.[4]

Conclusions

A system to provide general-utility communication service via satellites directly to many major user entities, such as large cities or small countries, should be capable of attaining the following operational goals:

1. At least one satellite is visible to each user station at all times (at an angular elevation exceeding a chosen minimum value).
2. Every satellite is an integral, active part of the total communication system at all times.
3. Every satellite has communication channel capacity adequate in quantity and in flexibility to handle all proffered traffic, of whatever kind.
4. Access to the satellite capacity is fully available to any user. Any user station, anywhere, can obtain at any time, from any satellite visible to it, the full service of the total communication system.
5. Traffic dispatching and monitoring functions are concentrated at a few centers on the ground, to insure smooth and flexible operation of the whole system under heavy load.
6. These system centers need not originate traffic; they exist to provide the service functions that are characteristic of post offices, namely the collection, sorting, forwarding, and distribution of communication traffic.

[4] H. J. Weiss, "Communication-Satellite-System Handover Requirement and Associated Design Problems," *RCA Review*. Vol. XXIV, p. 421, Sept. 1963.

7. Jointly, the post offices keep all satellites in sight and in communication at all times.
8. Mutually visible satellites provide relay trunks between adjacent pairs of post offices.
9. No post office is ever isolated from the system by simultaneous opening of all relay trunks, due to temporarily unfavorable locations of all satellites.
10. Post offices are located symmetrically on the earth, equalizing all inter-office relay-span lengths, to avoid local overdesign of the system.
11. User stations communicate directly only with post-office stations, not with other user stations, to place the entire burden of route selection on the post offices.
12. Post-office stations provide up-to-date satellite-ephemeris data to all user stations, and provide system supervision as needed.
13. Orbit patterns and satellite population, whether pseudo-random or station-keeping, are as sparse as goals 1, 7, and especially 9 will permit. Implementation of goals such as 2 may be held down intentionally to the level justified by actual needs for service.

Examples of systems consistent with these goals are:

1. A system suitable for active-repeater satellites, using
   a. Nine satellites keeping relative stations in orbits inclined 58 degrees to the equator, at 8260 nautical miles altitude.
   b. One post-office station at the (snow-free) northern tip of Greenland. This provides uninterrupted service to all user stations located anywhere north of 8 degrees north latitude.
   c. (Optional) Four additional post-office stations equally spaced along the equator. These will extend service meeting fully all stated goals to users at any point on earth.
2. A system suitable for passive-reflector satellites, using
   a. 250 uncontrolled satellites at 1570 nautical miles altitude, in orbits placed to approach uniform probable satellite density over the whole sky; this provides 99 percent relay-link continuity.
   b. Five post-office stations in a northern-latitude belt, at 5 of the 12 vertices of a regular icosahedron on earth. This serves well the entire populated world except lower South America and Antarctica, though Africa and New Zealand must use

overland links; any post office has 0.01 per cent probability of isolation.

c. (Optional) Three additional post offices, at other icosahedron vertices. These will remedy all service lacks to the entire world, and reduce the probability of belt-post-office isolation to 10^{-6} per cent.

The well-known 3-satellite synchronous system with 3 ground relay stations also meets the stated goals, with a minimum of the double traffic handling necessitated by goal 11.

SUPERCONDUCTIVE ASSOCIATIVE MEMORIES*

By

Richard W. Ahrons

RCA Laboratories,
Princeton, N. J.

Summary—This paper covers the broad area of superconductive associative memories using cryotrons. A survey of existing read and write processes and circuit arrangements is presented, followed by specific circuit improvements and some logic suggestions. These include: (1) a cryotron implementation of a destructive-readout associative memory with parallel readout that contains a minimal number of cryotrons, (2) a cryotron implementation of a new read-out scheme called the Modified Lewin, (3) a comparison of speeds of operation inherent to the Modified Lewin and parallel-readout schemes, (4) circuit methods of increasing this speed, and (5) algorithms for use with a parallel-readout associative memory in greater-than, less-than, and between-limits comparisons.

IN THE INTERNAL MEMORY of most present-day computers, the random-access type memory employing an address to locate data (e.g., the magnetic-core memory) has replaced the serial-access memory employing a serial scan for the appropriate data (e.g., magnetic-drum or disc memories). Random-access memories have not as yet replaced the magnetic tape in semi-permanent external stores because of economics (cost per bit). The next step in basic memories may be the "more-intelligent" associative memory which does not require a fixed position, or address, for the stored information.

Two observations may be made in regard to random-access memories: (1) a large percentage of programming time and effort is usually spent in assigning and keeping track of addresses and (2) in so far as is known, the human memory allows information to be retrieved by an associative process without regard to physical location of information. Thus, the conventional technique of requiring an address for each physical word location may be neither the most natural nor the most convenient technique for storing information.[1] Associative memories should not be regarded as attempts to extend the random-access

* Taken from the dissertation submitted to the Faculty of the Polytechnic Institute of Brooklyn in partial fulfillment of the requirements for the degree of Doctor of Philosophy, 1963. Supported by the Office of Naval Research (NONR-3879-(00)).

[1] J. A. Weisbecker, personal communication.

memory process, since it eliminates the address and the bookkeeping of the addresses. Thus the associative-memory concept should lead to a change in the philosophy of memory programming and not merely an extension of present techniques. It is reasonable to believe that the associative form of memory will be applicable to future computers that are intended for the more-sophisticated learning and thinking processes.

The basic function of an associative memory (also called catalog, parallel-search, content-addressed, or data-addressed memory) could be defined as a memory with the ability to answer this question: Is this piece of information in the memory? However, since this interrogating information is already known, the information generally

Table I—Types of Associative Memories

	Name	Portion of Word Interrogated	Portion of Interrogated Segment Required for Interrogation
(1)	Fixed Tag	Fixed Segment	Total
(2)	Multiple Fixed Tags	Any Set of Fixed Segments	Total
(3)	Variable Tag	Fixed Segment	Any Part or Total
(4)	Fully Interrogable	Any Segment	Any Part

desired is that which is associated with the original interrogating information. Thus the memory that only answers the above question is the most elementary form of associative memory. McDermid and Petersen[2] have divided the more-advanced forms of associative memories into four categories. A modified list of these categories in order of their complexity is as follows (see Table I):

(1) *The fixed-tag associative memory* has one fixed segment of the word set aside for interrogation and all of that segment must be used in the interrogation. The remainder of the word is read as the desired information. Thus only a part of the memory is associative.

(2) *A multiple fixed tag* describes a memory whose words are divided into segments any of which can be used as the interrogated portion. All the bits in the interrogated segment must be used in the interrogation.

[2] W. L. McDermid and H. E. Petersen, "A Magnetic Associative Memory System," *IBM Jour. of Res. and Dev.* Vol. 5, p. 59, Jan. 1961.

(3) *A variable tag* describes a memory with only a fixed segment of the word used for the interrogation. Any part of this fixed segment can be used in the interrogation. Thus only a part of the memory is associative.

(4) *The fully interrogable memory* allows any choice of bits in the word to be used in the interrogation, and the remainder of the word is read. This is the most general form of the associative memory. The remainder of the paper is devoted to this last category.

Memory Readout

The desired information, i.e., that contained in other than the interrogated portion of the word, which is called the data portion, may be read from the memory in serial (bit-by-bit) or in parallel (all bits in a word) depending on the design. A superconductive associative memory using a serial read is described by Slade.[3,4] This memory is of the fully interrogatable type. One portion or segment of this memory is interrogated and a sensing signal (in this case a voltage) appears if there is a match present in the memory. After finding a match, the first digit of the data portion of the word is interrogated with a 1 along with the original interrogating segment. If a match is present, then the new bit is known to be a 1. If a match is not present, the bit is obviously a 0. With this known piece of information, the next unknown bit is interrogated with a 1. This procedure is repeated serially until all bits are known. It is assumed, in this procedure, that the interrogated word segment appears only with respect to one word in the memory.

Frei and Goldberg[5] have shown that the associative memory described by Slade[3,4] can be used when more than one word contains the interrogated word segment (henceforth termed "multiple responses"). All the data can be retrieved in serial fashion using the algorithms of Frei and Goldberg. This method permits retrieval of all data in ordered serial form. The words may be retrieved beginning with either the highest number or the lowest number. An example of this method

[3] A. E. Slade and H. O. McMahon, "A Cryotron Catalog Memory System," *Proc. Eastern Joint Computer Conf.*, New York, N. Y., Dec. 10-12, p. 115, 1956.

[4] A. E. Slade and C. R. Smallman, "Thin Film Cryotron Catalog Memory Symposium on Superconductive Techniques for Computing Systems," *ONR Symposium*, Washington, D. C., May 17-19, p. 213, 1960.

[5] E. H. Frei and J. Goldberg, "A Method for Resolving Multiple Responses in a Parallel Search File," *Trans. IRE on Electronic Computers*, Vol. EC-10, p. 718, Dec. 1961.

showing the retrieval of three words when the data portion of the word has 5 bits is given in Figure 1a. Each digit location can be interrogated by one of three signals, 1, 0 or ∅. There is an output signal when all interrogating digits match a stored word. The ∅ ("don't care") signal is used for digits which are not interrogated. It takes 23 cycles to obtain the three words of Figure 1a.

If, in addition to a yes or no answer, the memory can yield the information that more than one word exists in the memory for a set of interrogating bits, the number of retrieval steps can be reduced significantly from that of the method described by Frei and Goldberg. An example of this plurality sense method is shown in Figure 1b; it requires only 16 cycles compared with 23 for the method described by Frei and Goldberg. This plurality sense is similar to an ordered readout suggested by Seeber and Lindquist for parallel readout memory.[6]

A unique method of reading out of a memory in a series–parallel fashion was devised by Lewin.[7] One can read out in an ordered list. The requirement for this readout is that the memory provide two sense outputs for each digit of the word; one is used to detect any 1's present in any of the interrogated words and the other is used to detect any 0's present in the interrogated words. Thus each digit-sensing operation will reveal: (1) only 1's, (2) only 0's, (3) a combination of 1's and 0's (labeled X), or (4) neither 1's or 0's (labeled Y). A Y result indicates that the word does not exist in the memory. An example is shown in Figure 1c. Lewin's method requires only $2w$-1 cycles for complete readout of the w words with the same initial interrogation segment. Lewin's method may be modified to one sense output per digit if the sense can determine a match condition either between 1's or 0's. This Modified Lewin method, shown in Figure 1d, requires $2(2w\text{-}1)$ cycles.

The parallel readout scheme requires all data digits of a word to be read in parallel by appropriate word logic. In addition, when the memory is interrogated the words are read in sequence until all words with the same interrogating segment are read from the memory. This sequencing does not necessarily mean that the words are in an ordered list. They are usually read according to geometrical position in the memory.

---

[6] J. R. Kiseda, H. E. Petersen, W. C. Seelbach, and M. Teig, "A Magnetic Associative Memory," *IBM Jour. of Res. and Dev.*, Vol. 5, p. 106, April 1961.

[7] M. H. Lewin, "Retrieval of Ordered Lists from a Content-Addressed Memory," *RCA Review*, Vol. 23, p. 215, June 1962.

Example: 5 data bits 3 word retrieval

Data bits of word						
	(1)	0	0	0	0	1
	(2)	1	0	0	0	0
	(3)	1	1	1	1	0

a) Frei and Goldberg

Cycle	*Data Interrogation*	*Sense Output*	*Words Sensed*
1	0 0 0 0 0	1	1, 2, 3
2	0 0 0 0 0	1	1
3	0 0 0 0 0	1	1
4	0 0 0 0 0	1	1
5	0 0 0 0 0	1	1
6	0 0 0 0 0	0	
7	0 0 0 0 1	1	1*
8	0 0 0 1 0	0	
9	0 0 1 0 0	0	
10	0 1 0 0 0	0	
11	1 0 0 0 0	1	2, 3
12	1 0 0 0 0	1	2
13	1 0 0 0 0	1	2
14	1 0 0 0 0	1	2
15	1 0 0 0 0	1	2*
16	1 0 0 0 1	0	
17	1 0 0 1 0	0	
18	1 0 1 0 0	0	
19	1 1 0 0 0	1	3
20	1 1 0 0 0	0	
21	1 1 1 0 0	0	
22	1 1 1 1 0	1	3*
23	1 1 1 1 1	0	

b) Plurality Sense

Cycle	*Data Interrogation*	*Sense Output*	*Words Sensed*
1	0 0 0 0 0	P (plurality)	1, 2, 3
2	0 0 0 0 0	1	1
3	0 0 0 0 0	1	1
4	0 0 0 0 0	1	1
5	0 0 0 0 0	1	1
6	0 0 0 0 0	0	
7	0 0 0 0 1	1	1*
8	1 0 0 0 0	P	2, 3
9	1 0 0 0 0	1	2
10	1 0 0 0 0	1	2
11	1 0 0 0 0	1	2
12	1 0 0 0 0	1	2*
13	1 1 0 0 0	1	3
14	1 1 0 0 0	0	
15	1 1 1 0 0	0	
16	1 1 1 1 0	1	3*

Fig. 1(a,b)—Examples of Frei and Goldberg and plurality sense methods for ordered retrieval. (Asterisks indicate words retrieved.)

Comparison of Fully Interrogable Systems

A fully interrogable associative memory has the following characteristics:

(1) Any or all segments of the word can be interrogated and the remaining digit locations can be read.

(2) Any or all digit locations of the interrogated word can be written into. This includes the interrogated segment. Empty word

c) Lewin

Cycle	*Data Interrogation*	*Sense Output*	*Words Sensed*
1	Ø Ø Ø 0 0	X X X X X	1, 2, 3
2	0 Ø Ø 0 0	0 0 0 0 1	1*
3	1 0 0 Ø Ø	1 X X X 0	2, 3
4	1 0 Ø Ø Ø	1 0 0 0 0	2*
5	1 1 0 Ø Ø	1 1 1 1 0	3*

d) Modified Lewin

Cycle	*Data Interrogation*	*Data Interrogation*	*Sense Output*	*Words Sensed*
1	Ø Ø Ø Ø 0	0 0 0 0 0	0 0 0 0 0 0	1, 2, 3
2	0 0 0 0 0	1 1 1 1 1	1 1 1 1 1	
3	0 Ø Ø Ø Ø	0 0 0 0	0 0 0 N	1*
4	0 0 0 0 Ø	1 1 1 1	N N N 1	
5	1 0 0 Ø Ø	0 0 0 0	0 0 0 0	2, 3
6	1 Ø Ø 0 0	1 1 1 1	1 1 1 N	
7	1 0 Ø 0 0	0 0 0	0 0 0	2*
8	1 0 0 0 Ø	1 1 1	N N N	
9	1 1 Ø 0 0	0 0 0	N N 0	3*
10	1 1 0 0 Ø	1 1 1	1 1 N	

N = not 1 for a 1 interrogate or not 0 for a 0 interrogate

Fig. 1(c,d)—Examples of Lewin and Modified Lewin methods for ordered retrieval. (Asterisks indicate words retrieved.)

locations are designated by 00 . . . . 0 or a separate bit, and new information is read into a first empty word location.

(3) All word locations with a common interrogated segment can be read or written into sequentially.

Before discussing readout systems, consideration must be given to the write process. The associative memory may have a high-speed write, fixed store, or a slow-speed semipermanent write (such as mechanically fixing the word). This last appears to be basically impractical for superconductive memories, since the memory must be

sealed in a refrigeration unit and is not easily altered. It is possible to use a superconductive memory with a fixed store. However, the most versatile and attractive application for a superconductive associative memory is one with high-speed read and write with provision for multiple response.

An associative memory with a serial readout requires the least logic outside the cells. This "outside" logic is in two categories: the logic that must be reproduced for each word is called "word logic," and the logic that must be reproduced for each digit of all the words is called "digit logic". A serial readout requires very little word logic but does require some digit logic to operate the serial scan of the digits and record the words. If the serial readout uses plurality sense, one can save some of the serial scan cycles at the cost of additional word and digit logic. Lewin's method requires a more-complex cell with dual senses and more digit logic than either of the two preceding methods. The parallel readout requires a greater amount of word logic but less digit logic than the above three readout systems. Since the words in a practical memory vastly outnumber the digits, the parallel readout system usually requires more components. A high-speed parallel readout requires finding one of many interrogated words and reading it. The word logic used for writing is similar to that for reading, and thus the word logic can efficiently be used for both processes. Thus the over-all addition of complexity for parallel readout, assuming the necessity of write logic for the high-speed write, may be less than the other three schemes. A procedure for obtaining an ordered list when using parallel readout is given by Seeber and Lindquist.[6]

Another factor that must enter the discussion is the inherent speed of the various schemes. It is possible that a reading scheme such as Lewin's may be faster than parallel readout, although the Lewin method requires more cycles to retrieve a given number of words. Thus, a system based on Lewin's method may become very attractive.

In summary, the following areas must be considered in evaluation of the memories:

(1) Speed—write and read.

(2) Complexity—based on cost-per-bit and the ability to batch fabricate.

(3) Capacity—memory-capacity limitations.

The Associative Memory Cell

The associative concept is a matching and scanning process of the stored data not unlike the one employed by the human mind or, in the

simplest form, a catalog file. In general, the associative process is connected with large-capacity storage. It is estimated that high-speed associative memories commence being generally attractive at about 10^5 words of 10^2-10^3 digits, or 10^7-10^8 bits. Presently the most advanced random-access memories are in the 10^6-10^7 bit range. Because of the large number of bits involved, it is necessary that the cells be designed so as to permit production of large arrays at a low per-bit cost.

In a conventional random-access memory, the cell need have only the capability to store the binary bit. However, an associative memory cell must perform comparison logic in addition to storage at the cell. The logic performed at the cell between the storage state and the interrogate signal is of two categories: (1) binary, and (2) ternary or double binary. The first type of logic is used for interrogable memories of Table I and requires that the total portion of the interrogated segment be used. This logic produces a signal when there is not a match between storage and interrogation bit and does not produce a signal when there is a match; this is the "exclusive-or" function. The complement of the exclusive-or may be used as an alternative. If any portion of the interrogated segment can be used for interrogation, as in the fully interrogable case, the data bits not being interrogated must give a signal equivalent to a match, independent of the stored state. If this function is to be performed by one variable representing one interrogating line, this signal must have three states. The cell therefore must perform ternary logic. The interrogate signal has three states: 1, 0, and $\emptyset$ (don't care). Alternatively, two lines with binary information can form the interrogate, hence the name double binary. One of these lines performs the 1, 0 matching function; the other acts as a gating line, gating only the digits interrogated.

The requirements of the sensing arrangement are also an important consideration in the design of the cell. In the case of the serial readout, such as in the memory described by Slade,[4] the memory must have a sense arrangement in the direction of the word, which will be termed a "word-read" organization. This word-read organization must be such that it is possible to determine if there is a match between the interrogation segment and the interrogated segment in the word. The detection can represent a match of all the interrogated digits or, alternatively, it can represent the complement, which is one or more mismatches. For the series readout, all the word senses are connected to give a single sense readout for the memory.

In the case of parallel readout each word-read sense is separate and connected to word logic. In a three-dimensional array in which two directions form the nth digit plane and the third forms the word

direction, the parallel readout requires a plane (or a plane of threaded lines) as a readout plane. This is called a "digit-read" organization. Since only one word is read at each sequence in the parallel readout scheme, the digit-read may be destructive and the word rewritten into the memory. The conflict between word-read and digit-read organizations for the fully interrogable memory necessitates: (1) a simultaneous word and digit-read organization, (2) a switchable read organization, or (3) two memory arrays in parallel, one with digit-read organization and the other with word-read organization. If readout of the fixed segments in other than the fully interrogable memory

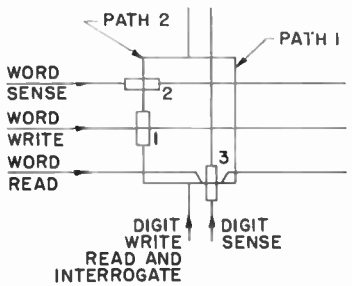


Fig. 2—Associative memory cell.

of Table I is not required, these memories could use word-read organization in the fixed segment portions and digit-read organization in the remaining or information portion.

When using Lewin's scheme, the cell requires a word-read organization and a two-digit-read organization. One read organization is sensitive to 1's and another to 0's. The modified Lewin scheme would require only one digit-read at the cost of twice the number of cycles. However, this digit-read organization must be capable of sensing 0's or 1's on command.

An extension of a simple cryotron cell has served for the majority of proposed cryotron associative memory systems described in the literature.[3,4,6,8–11] One form of this cell is shown in Figure 2. The cell can

[8] A. E. Slade, "A Cryotron Memory Cell," *Proc. IRE*, Vol. 50, p. 81, Jan. 1962.

[9] V. L. Newhouse and R. E. Fruin, "A Cryogenic Data Addressed Memory," *Proc. Western Joint Computer Conf.*, San Francisco, p. 89, May 1962.

[10] P. M. Davies, "A Superconductive Associative Memory," *Proc. Western Joint Computer Conf.*, San Francisco, p. 79, May 1962.

[11] R. F. Rosin, "An Organization of an Associative Cryogenic Computer," *Proc. Western Joint Computer Conf.*, San Francisco, p. 203, May 1962.

store (1) a clockwise current, (2) a counterclockwise current, or (3) no current. Two of these three states can be used for binary storage. The choice will depend largely on the type of word and digit logic. If a 1 is considered as a clockwise current and a 0 as a counterclockwise current, the 1 is written into the memory by the combination of a negative current in the digit drive line and a current of either polarity in the word write, the letter current being sufficient to keep cryotron 1 in the normal state. For a stored 0, a positive pulse is used as the digit drive. If a negative current with the same amplitude as that for write is also used as an interrogate 1 on the digit drive, the

Interrogate	*Store*	*Word Sense (Current in Path 2)*
0(—I)	0(—I_s)	0(0)
1(0)	0(—I_s)	1(—2I_s)
0(—I)	1(I_s)	1(2I_s)
1(I)	1(I_s)	0(0)
ϕ(I)	0(I_s)	0(—I_s)
ϕ(0)	1(I_s)	0(I_s)

Fig. 3—Truth table.

match condition would be all current I_0 in path 1 and no current in path 2. For a mismatch, the current in path 2 is equal to $2I_s$, where I_s is the stored current. For an interrogate 0 using a positive pulse, the same process occurs, and cryotron 2 is excited only by a mismatch. The don't care function (ϕ) is represented by no current in the digit interrogate line. The truth table for the interrogate–storage process is shown in Figure 3.

If I_{ws} is the maximum current in the word sense and I_s is the stored current, the combination of I_s (control) and I_{ws} (gate) must be insufficient to excite cryotron 2, but $2I_s$ must be sufficient to excite cryotron 2 with no current in the word sense. If cryotron 2 is represented by an ideal crossed-film cryotron with a characteristic parabolic phase curve, then the superconducting–normal boundary can be represented by

$$I_s^2 = I_{cc}^2 - \frac{I_{ws}^2}{G^2}, \tag{1}$$

where

$$G = \frac{I_{gc}}{I_{cc}}. \tag{2}$$

and I_{ws} is the gate current, I_s is the control current, I_{gc} is the critical value of gate current for 0 control current, and I_{cc} is the critical value of control current with 0 gate current. Thus from Equation (1), during the don't care operation,

$$I_s^2 < I_{cc}^2 - \frac{I_{ws}^2}{G^2}.$$

For mismatch operation

$$2I_s > I_{cc},$$

or

$$I_s^2 + \frac{I_{ws}^2}{G^2} < I_{cc}^2 < 4I_s^2. \tag{3}$$

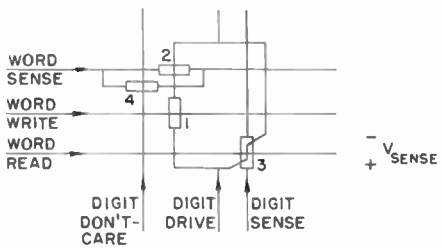


Fig. 4—Associative memory cell.

Thus

$$\frac{I_{ws}}{G} < \sqrt{3}\, I_s. \tag{4}$$

One also must be aware of the tolerances in each parameter, and the worst case must be considered.

The addition of a cryotron and another digit line can circumvent a possible tolerance problem indicated by Equation (4). Such a circuit[9] is shown in Figure 4. A current is fed to the digit don't care whenever the bit is to be interrogated by a 1 or a 0, thereby exciting cryotron 4. For digits with don't-care interrogate functions, cryotron 4 is superconducting and acts as a short bypassing cryotron 2.

The combination of the word-read current and the stored current can excite cryotron 3 which forms an AND gate for the above two

variables. The excitation depends on the relative directions of the stored current and the sense current. A word-read current and a stored 1 excites cryotron 3 when negative sense current is used. A stored 0 will not excite cryotron 3 whenever a positive current is used. A negative voltage across cryotron 3 signifies a 1 read from that digit of the memory.

If a destructive read were used as allowed for the parallel readout case, the cell complexity and size could be reduced as shown in Figure 5. The word drive is used for sensing the read in addition to write. A current in the word drive destroys the storage in the cell; a positive V_s is a 1 and a negative V_s is a 0.

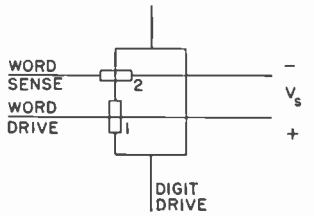


Fig. 5—Associative memory cell.

The number of digit lines determine the horizontal width of the cell in the evaporated circuit. The speed of some operations is an inverse function of this width. The degree to which the reading speed is related to the width is dependent upon the type of read system chosen and the word logic.

Word Logic for Parallel Read System

The word logic can be divided into three areas: (1) word sense, (2) word read and write, and (3) word sequencing. Word sequencing involves sequential readout of the multiple responses. Davies[10] and Newhouse and Fruin[9] have described associative memories using cells similar to that of Figure 4 with parallel readout. The following discussion in general applies to any of the three cells in Figures 2, 4, and 5, and specifically shows word logic design for the cell in Figure 5. The word logic dominates the speed of a parallel-readout associative memory. Logical designs therefore are discussed with reference to operating speeds.

Word Sense Logic

The basic problem is to logically distinguish the superconducting

lines, which represent a match condition, from the nonsuperconducting lines, which represent a mismatch. Since there may be a multiplicity of these superconducting lines, it is necessary to select one from all the others by geometrical preference. A series network first detects all superconducting lines simultaneously. This series network in turn controls a ladder network which chooses a word by geometrical position, usually the first from top or bottom of the memory.

The series circuit as shown in Figure 6 simultaneously detects all superconducting lines. W_s is the word sense line which is totally superconducting only for a match. $\overline{W}_s$ is a parallel line which represents the

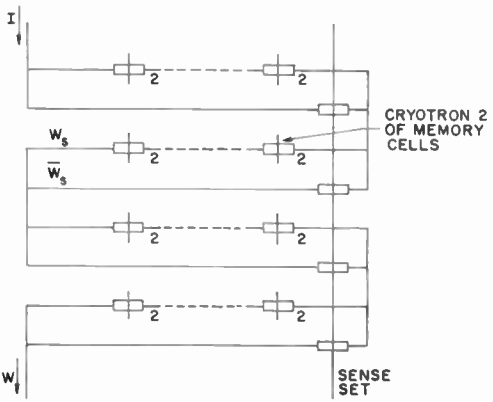


Fig. 6—Series sense.

complementary function of W_s. Each word sense set comprising W_s and $\overline{W}_s$ is in series with the corresponding sets of all other words. The sense-set line initially sets all the current into the W_s lines. If any mismatch occurs, the current is shifted from W_s to $\overline{W}_s$ in the mismatched words. This operation has a time constant for the worst case of $2L/R$, where L is the inductance of W_s or $\overline{W}_s$, and R is the resistance of the cryotron controlling the mismatch. The lines W_s and $\overline{W}_s$ control the logical operation of read, write, and sequencing.

Once the information is obtained for words whose integrated segments match the interrogating word, the read or write signal must decide which of all the matched words to excite first. This decision necessitates a geometrical choice, usually either the first match word from the top or bottom of the memory. A ladder network[10] controlled by the word sense as shown in Figure 7 can find the first match word from the top of the memory. A current, I, is fed to X. The first match

word (current in W_s) causes the current to be blocked by cryotron 1, which is excited by W_s. When cryotron 1 is excited, I switches to $\overline{X}$ at the corresponding word location. All preceding mismatch words had their respective cryotron 2 excited and thus I remains in X. The current, I, in branch B will control the read and write logic. In large-capacity memories, the ladder or logic of finding the first superconducting sense line dominates the overall speed. Two possible worst-case

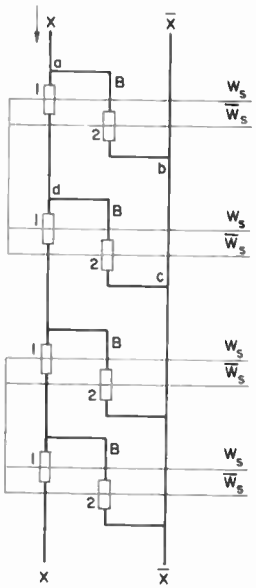


Fig. 7—Ladder network.

situations can be considered: all mismatches except the first word, and all mismatches except the last word. The latter case has a time constant approximated by nL_{ab}/R_1 and the former case $n(L_{bc} + L_{ad})/R_2$, where n is the number of word locations in the memory, L_{ab}, L_{bc} and L_{ad} are the inductances of line segments a-b, b-c, and a-d, respectively, and R_1 and R_2 are the resistances of cryotrons 1 and 2, respectively. Two points should be mentioned: (1) the X and $\overline{X}$ lines can be widened considerably at small expense in overall memory size; however, segment a-b is severely limited in width, and (2) it is less likely that the first (or near the first) word be the only match in the memory than that the last (or near the last) word be a match. Because of sequencing from the top, it is known that there are no other matches when the last word

is read. Thus nL_{ab}/R_1 can be considered the worst-case time constant. Thus this time constant is a function of the number of word locations, n. The current, I is always either in X or $\overline{X}$. If I has remained in X beyond the last word, then all (or none, if on the first-sequence step) of the matched words in the memory were read.

At the expense of increased complexity, the above time constant may be reduced in a ladder tree circuit. Such a circuit is shown in

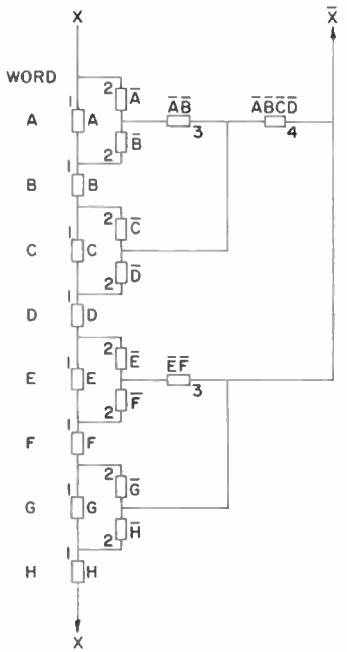


Fig. 8—Ladder-tree network. (Letters represent functions which cause resistance in cryotron. The letter such as A also represents a match; its complement a mismatch.)

Figure 8. Each letter such as A represents a word with current in W_s, and $\overline{A}$ represents a word with current in $\overline{W}_s$. Thus $\overline{A}\,\overline{B}$ forms an AND function when both words A and B are mismatched to the interrogating segment. The cryotron 3 excited by $\overline{A}\,\overline{B}$ becomes resistive when both A and B are mismatched. All the other cryotrons operate similarly. The circuit of Figure 9 shows the controls corresponding to the gates shown in Figure 8. Cryotrons 1 are in the same position as in the ladder of Figure 6 and perform the same function. Cryotrons 2 are supplemented by the addition of cryotrons 3 and 4 for an 8-word tree.

The number of levels, as defined by cryotrons 2, 3, 4 in the 8-word tree, is the smallest integer greater than $\log_2 n$. This ladder-tree changes the worst-case approximate time constant from nL_{ab}/R to $(\log_2 n)\ L_{ab}/R$. Cryotrons 3 and 4 (and higher numbers in larger memories) operate as AND gates.

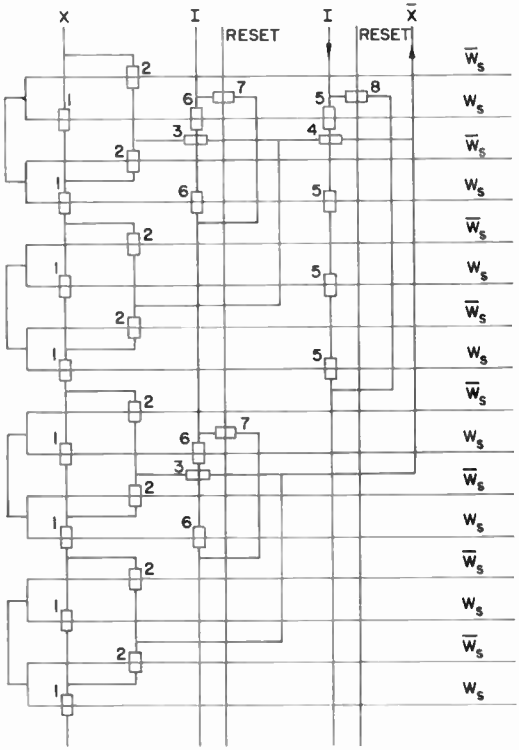


Fig. 9—Ladder-tree network with control network.

Since, for example,

$$\bar{A}\,\bar{B}\,\bar{C}\,\bar{D} = \overline{(A + B + C + D)}, \tag{4}$$

one can use a control circuit for the ladder tree as shown in Figure 10. This function corresponds to the control circuit controlling cryotron 4 in Figure 9. Cryotron 8 sets the current, I, initially into path 1. Cryotron 4 is excited only if A, B, C, D, or any combination is present. This type of control circuit is incorporated at each level of the tree with the

exception of the first, as shown in Figure 9. The time constant of the control used at the last level is the largest control time constant, and in large-capacity memories exceeds the time constant of the gates of the ladder tree. The worst-case time constant is for a match in the first word in the memory and for a mismatch in all other words. The time constant is $2L/R$. R is the resistance of a single cryotron gate and L is equal to the inductance of a line that extends about 1/2 the length of the memory. Thus this time constant is a function of the number of word locations. However, for multiple readout this ladder tree can be made faster than the ladder.

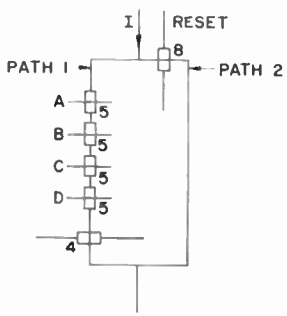


Fig. 10—Cryotron control circuit.

Word Read and Write Logic

The word read and write are controlled by a current present in the branch of cryotron 2 in either the ladder (Figure 7) or the ladder tree (Figure 9). A series network is used to maximize operating speed. The series network defines a network in which the word function line contains a dummy in parallel, and this set is in series with all other sets. A current is switched from the function line used for the function of read, write, or sense to a dummy path if the word is not to be read or written. The parallel network defines a network of function lines that are all connected in parallel, and the chosen line is the only superconducting line. The series network has a $2L/R$ time constant which is independent of the number of series sets. However, the parallel network has an nL/R time constant, where n is the number of word locations. In large-capacity memories, the latter time constant, which is dependent upon the number of words in memory, can become excessive. Newhouse and Fruin[9] use, for writing, a parallel network that will limit the writing speed considerably in large arrays.

Two series geometries are shown in Figures 11a and b. The geom-

etry of Figure 11b would not satisfy the series connection for word sense, but does satisfy the word read or write, since the controlling cryotrons are located at one end. Either geometry in Figure 11 can be used at the designer's discretion for writing or reading. D can be the word drive line as labeled in the cell of Figure 5, and can be either the word read or word write of Figures 2 and 4. The series network is

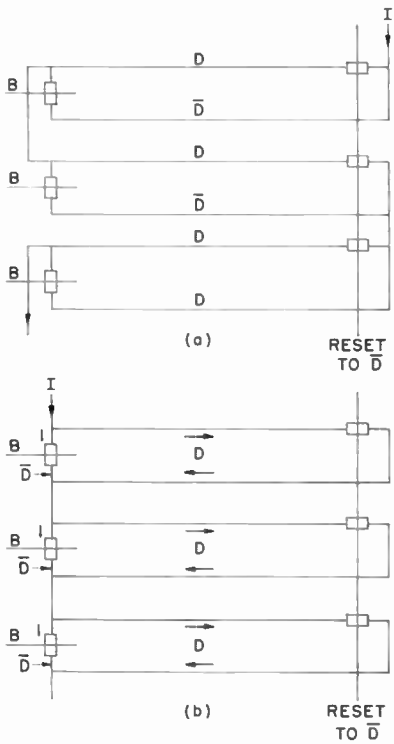


Fig. 11—Series read or write networks.

initially set with the current in the $\overline{D}$ line, which is the dummy line. B represents the branch which contains cryotron 2 of the ladder or ladder-tree network. Thus a current in B causes a current to switch from $\overline{D}$ to D line. Thus the word which has its current in D is activated for the write or read process. The time constant of either series geometry a or b is $2mL_c/R$, where L_c is the inductance of a line across each memory cell, R is the resistance of the cryotron, and m is the number of cells or binary digits in a word.

Word Sequencing

For multiple match responses, a sequential readout is necessary. As shown by Davies[10] the sequential operation can take advantage of the drive current in the cell of Figure 5 and the read current of the cell of Figures 2 and 4. The circuit is shown in Figure 12. A parallel set of cryotrons 1 and 2 (Davies[10] circuit) as shown in Figure 12a or a single cryotron acting as an AND gate as shown in Figure 12b, is

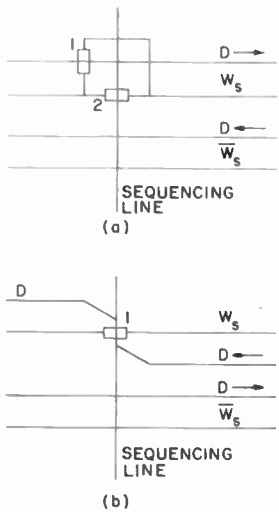


Fig. 12—Sequencing circuit.

inserted in series with the word sense, W_s. When the drive current is in line D, which can occur only in the one word being read or written, a current in the sequence line can cause the current in W_s to shift to $\overline{W}_s$. Only the word which is activated will have both cryotrons 1 and 2 of a or cryotron 1 of b excited. The word which was read now has current in its $\overline{W}_s$ and therefore acts as a mismatch word for the next read or write in the sequencing procedure. Thus one can sequentially read each word until the X line after the last word contains the current, which signifies there are no other words in the memory with the same interrogating segment.

The sequencing can also be accomplished by using a simple memory bit for the sequencing operation in each word. This added cell or digit can store the information that the word has been read or written and thus cause the word-sense line to assume a mismatch for the next step in the sequence. An example of this type of circuit is shown in Figure

13. The sequencing bit is interrogated with 0's along with the interrogation. A 1 is placed in this bit when a word is read from or written into the memory. Since there is no current in the D line for writing 0's into all sequencing bits, the 0 is written by a large negative current pulse that overdrives the cell, causing storage in the counterclockwise direction.

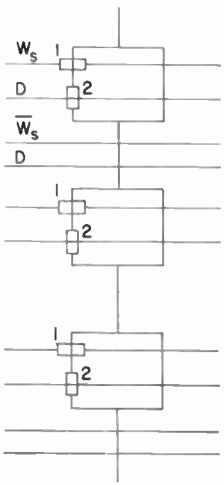


Fig. 13—Sequencing digit circuit.

The simple circuit of Figure 12 will operate with a nondestructive read, but will not operate with a destructive read. In the case of a destructive read, all sense information must be destroyed before or during the rewrite process, and the memory must be reinterrogated in the next sequence. The circuit of Figure 12 assumes that the sensing remains during the complete interrogation for multiple responses.

Parallel Readout Associative Memory

The circuit for one word of an associative memory with two of the many digits is shown in Figure 14. This word circuit incorporates the circuits of Figures 5, 6, 11b, and 13. Y is the drive line for the sequencing bit and X is the drive line for the ladder.

The operational steps are as follows: the currents I_{ws} and I_w are always in the memory as constant currents. I_w is always in $\overline{D}$ position at the beginning of the cycle.

Read and Retain

(1) Apply large negative current pulse I_y to store counterclockwise current in all Y cells.

(2) Apply current pulse to S_s to set I_{ws} to $\overline{W}_s$.

(3) Interrogate I_D's with 1, 0, or $\emptyset$ as represented by magnitudes $+I_0$, $-I_0$, and 0 current steps, respectively, and apply a current step I_y to Y. Current I_{ws} switches to $\overline{W}_s$ in all mismatched words.

(4) Apply current step I_x to X at top of memory. This current is routed through cryotron 2 of the first match word. I_w is switched

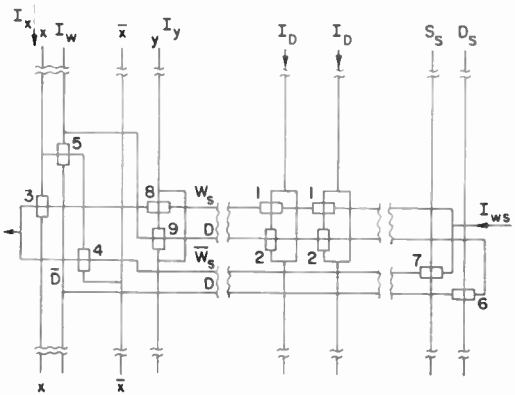


Fig. 14—Associative memory word with parallel readout.

from $\overline{D}$ to D in this first match word, and a voltage pulse appears across the I_D lines in the cells which are interrogated by $\emptyset$. A 1 is a negative voltage; a 0 is a positive voltage referred to the top.

(5) Remove current I_x; I_w remains in D.

(6) Replace $\emptyset$'s with the digits of the word just read. I_0 for a 1 and $-I_0$ for a 0.

(7) Apply positive current, I_y, to Y to write 1 into sequencing cells.

(8) Apply current to D_s to reset I_w to $\overline{D}$.

(9) Remove all I_D's. Word is now rewritten into memory. Remove I_y leaving a clockwise current stored in Y cell of word.

(10) Repeat steps 2 through 9 until all words are read from memory.

(11) Remove all I_D's from interrogation bits.

Read and Destroy

(1) Apply current pulse to S_s to set I_{ws} to $\overline{W}_s$.

(2) Interrogate I_D's with 1, 0, or ∅ as represented by $+I_0$, $-I_0$, and 0 current steps, respectively.

(3) Apply current I_x to X at top of memory. This current is routed through cryotron 2 of the first match word. I_w is switched from $\overline{D}$ to D in this first match word, and a voltage appears across the I_D lines that are interrogated by ∅. A 1 is a negative voltage; a 0 is a positive voltage.

(4) Remove current I_x; I_w remains in D.

(5) Replace all I_D's with 0's ($L_D = -I_0$).

(6) Apply current to D_s to switch I_w from D to $\overline{D}$.

(7) Repeat steps 1 through 6 until all words are read from memory.

(8) Remove all I_D's from all digits.

Write

(1) Apply current pulse to S_s to set I_{ws} to $\overline{W}_s$.

(2) Interrogate I_D's with 0's as represented by $-I_0$ current step.

(3) Apply current step I_x to X at top of memory.

(4) Remove current I_x; I_w remains in D.

(5) Replace I_D's with word.

(6) Apply current D_s to switch I_w from D to $\overline{D}$.

(7) Remove I_D's from all digits.

The timing for each step in the read process is approximated as follows:

Step No.	Process	Time Constant	
1	Switch Y bits to 0	$2\frac{L_c}{R}$	**
2	Switch I_{sw} from $\overline{W}_s$ to W_s	$2m\frac{L_c}{R}$	**
3	Switch I_{sw} from W_s to $\overline{W}_s$	$2m\frac{L_c}{R}$	
4a	Ladder routing time	$n\frac{L_c}{R}$	*
4b	Switch I_w from $\overline{D}$ to D	$2m\frac{L_c}{R}$	

Step No.	Process	Time Constant
4c	Stored current in cell	$2\frac{L_c}{R}$
5	Ladder decay time	$n\frac{R}{L_c}$ *, ***
6	Rewrite	$2\frac{L_c}{R}$ ***
7	Write 1 into sequence digit	$2\frac{L_c}{R}$ ***
8	Switch I_w to $\overline{D}$	$2m\frac{L_c}{R}$
9	Remove I_c's	Negligible
10	Repeat steps 2-9 until all words read	w (sum of 2-9)
11	Remove all I_b's	Negligible

*—worst case
, *—steps can be overlapped

In the above steps, L_c represents a magnitude of inductance equivalent to that possessed by a typical line crossing the width of a cell; m is the number of digits (bits per word); n is the number of word locations in the memory; and w is the number of matched words. The above values of the time constants give only order of magnitude values for comparison between the timing of various steps. As previously mentioned, in the case of practical memories the number of words, n, exceeds the number of digits, m, by two or more orders of magnitude. Thus it is easily seen that time constants in steps 4a and 5 dominate. However, it should be noted that this is the worst case, and the average time constant for 4a and 5 is much less. This time constant represents the most critical area of the associative memory. The addition of the ladder tree of Figure 9 would increase the speed somewhat at a cost in complexity.

The associative memory described above possesses all the characteristics necessary for a fully interrogable associative memory. However, the retrieval of multiple responses is in no particular order unless that order was originally stored in the memory by writing the words into the memory in a particular sequence.

The system described in this paper requires 2 cryotrons per cell and 7 cryotrons for the word logic. The system described by Davies[10] requires 5 cryotrons per cell and 9 for the word logic. The system described by Newhouse and Fruin[9] requires 5 cryotrons per cell and 12 for the word logic. The number of cryotrons per word is the product

of the number of cells per word and the number of cryotrons per cell; thus the latter number is of prime importance to memory size. An increase in the digit logic is necessary because of the destructive read. However, the savings in size as related to the number of cryotrons far outweighs the slight increase in digit logic. The digit logic is not described here since it can be designed to operate at room temperature outside the cryostat (refrigerator).

Associative Memory with Modified Lewin Readout

Although the Modified Lewin readout requires $2(2w-1)$ sequences to read out w words, whereas the parallel readout only requires w

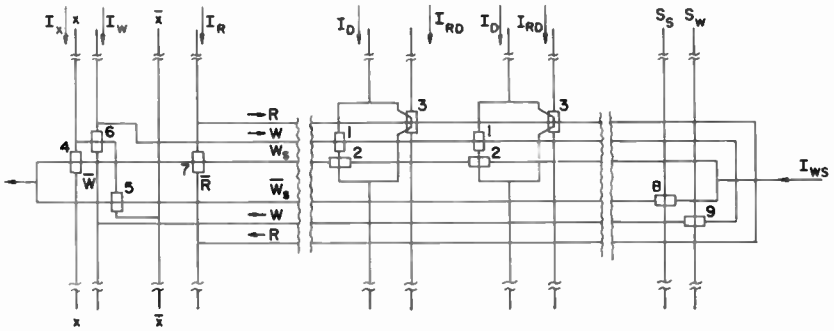


Fig. 15—Associative memory word with Modified Lewin readout.

sequences, each sequence of a Modified Lewin readout can require less time, and the system can basically read out faster than the parallel readout. A superconductive example of a circuit using the Modified Lewin readout is shown in Figure 15. This circuit incorporates a memory bit similar to that of Figure 4. A 1 is represented by a clockwise stored current and a 0 by a counterclockwise stored current. Cryotron 3 has in-line cryotron action with the control line from the cell and has crossed-film cryotron action with the word-write control line, W. The word-write current amplitude causes a readout which is a voltage across the gate of cryotron 3 when the stored current and the digit read current (I_{RD}) is in the same direction. Thus by using both polarities of digit read current, 1's can be read nondestructively without reading the 0's, and vice-versa, in sequence as required by a Modified Lewin scheme. The word-logic circuit represented by cryotrons 4, 5, and 6 is used for writing and is a combination of the circuits of Figures 7 and 11b. The R and $\bar{R}$ lines which contain cryotron 7 form a series read circuit (see Figure 11b) controlled by a series sense circuit (see Figure 6). The sense circuit also controls the ladder.

The steps for memory operation are as follows: I_w and I_{ws} are constantly in memory as constant supply currents. Current I_w is always in W at the onset of any cycle.

Read and Retain

(1) Apply current to S_s to set I_{ws} into W_s.

(2) Interrogate I_D's with 1, 0, or $\emptyset$ as represented by a I_0, $-I_0$, and 0, respectively. I_{ws} switches from W_s to $\overline{W}_s$ in all mismatched words.

(3) Apply current I_R. I_R will flow in R for all matched words and in $\overline{R}$ for all mismatched words.

(4) Where I_D's are $\emptyset$, apply negative I_{RD} to read 0's and remove.

(5) Apply negative I_{RD} to same digits as in step 4 to read 1's and remove.

(6) The first two cycles have been completed, and $4w$-4 cycles remain. Proceed according to the algorithm of the Modified Lewin process as indicated in Figure 1d until all accesses are complete. If a 0 must be changed to 1 in the interrogate, a reset of I_{ws} from $\overline{W}_s$ to W_s is necessary preceding the 0 to 1 change. If a $\emptyset$ must be changed to a 0, no reset is necessary.

(7) Remove I_R and all I_D's.

Write

(1) Apply current to line S_s to set I_{ws} into W_s.

(2) Interrogate I_D's with 0's. I_{ws} switches to $\overline{W}_s$ in all mismatched (non-empty) words.

(3) Apply current I_x to X. Current will be routed through cryotron 5 of first empty word location. I_w switches from $\overline{W}$ to W in this word.

(4) Change I_D's to represent word.

(5) Remove I_x.

(6) Reset I_w from W to $\overline{W}$ by a current in S_w.

(7) Remove I_D's.

Write and Destroy

(1) Read words as in read and retain, steps 1 to 7

(2) Apply current to S_s to switch $\overline{W}_s$ to W_s.

(3) Interrogate I_D's with word.

(4) Apply current I_x to X.

(5) Change I_D's to 0's.

(6) Remove I_x.

(7) Reset I_w from W to $\overline{W}$ by a current in S_w.

(8) Remove I_D's.

The timing for each step in the read cycle is as follows:

Step No.	Process	Time Constant
1	Switch I_{w_s} from $\overline{W}_s$ to W_s	$2m \frac{L_c}{R}$
2	Switch I_{w_s} from W_s to $\overline{W}$	$2m \frac{L_c}{R}$
3	Switch I_R from $\overline{R}$ to R	$2m \frac{L_c}{R}$
4	0's read	Negligible
5	1's read	Negligible
6	Sequencing	$7(w\text{-}1)\ 2m \frac{L_c}{R}$
7	Remove I_D's	Negligible

The time constants for the 6th step depend on the number of cycles which is $2(2w\text{-}1)$. However, only $2w$-1 represent changes in the interrogation as shown in the example in Figure 1d. After step 5, there are only $2w$-2 changes. At least one-half of those changes must be 0 to 0, which requires time for I_R to be reset, sense time, and time for I_R to be set again. At most, one-half the changes might be from 0 to 1, thus requiring in addition a reset of the sense. If, for order of magnitude results, the total time constant is considered to be the sum of the three or four time constants, the total time constant is $(7/2)\ (2w\text{-}2)\ 2m\ L_c/R$. The important point is that the timing is independent of n, the number of word locations in the memory. The parallel readout case appears as about $2wn\ L_c/R$. For large-capacity memories ($n \gg m$) there is no doubt that the Modified Lewin approach yields much faster reading time. However, the write in the Modified Lewin approach does utilize a ladder and must have a worst-case write time proportional to the number of word locations. Thus the associative memory with the Modified Lewin scheme can be considered as a memory with fast readout and moderately fast write.

The read and destroy command combines the read and the write command, and its speed is dependent upon the number of word locations in the memory. An addition of two cryotrons in the word logic can increase the speed for read and destroy. Figure 16 shows the added circuitry for rapid destroy. The circuit contains the addition of W_{s1}

line, cryotrons 10, 11, and 12, and the two control lines Y and $\overline{Y}$. Current I_y is always in Y or $\overline{Y}$. This rapid destroy circuit can also be added to the parallel readout of Figure 14. If a current is in Y, the memory is just that of Figure 15. However, for rapid destroy I_y is in $\overline{Y}$, allowing the word sense to control I_w and destroy the word. The read process is the same as read and retain. The procedure for destroy is as follows:

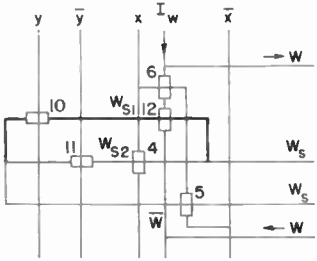


Fig. 16—Rapid destroy circuit.

Rapid Destroy

(Current I_y is in Y initially)

(1) Apply current to S_s to switch I_{ws} to W_s.

(2) Interrogate with word or word segment if all words with the same segment are to be destroyed.

(3) Switch I_y from Y to $\overline{Y}$. I_w switches from $\overline{W}$ to W in all matched words and destroys those words.

(4) Switch I_y from $\overline{Y}$ to Y.

(5) Change all I_D's to 0's.

(6) Apply current to S_w to reset I_w from W to $\overline{W}$. All 0's are now written into word location(s).

(7) Remove all I_D's.

The destroy process time does not use the ladder and thus is independent of the number of word locations in the memory. It requires the same order of magnitude of time as does the read process. Furthermore, all the words in the memory can be destroyed in one rapid cycle by using all Ø's as the interrogating bits. The memory can also alter any or all the digits of *one* word in one rapid cycle by writing the altered word in the memory step 4. A common segment of many words cannot be changed in one cycle without additional circuitry because all the unknown digits which are interrogated by Ø will be destroyed.

The use of a Lewin readout scheme instead of a Modified Lewin scheme requires two senses for each digit. This in turn requires three lines from each memory cell in the digit direction instead of two. The width of the cell must be widened by 3/2 and thus the inductance per cell, L_c, would be 3/2 greater, decreasing the speed of each switching process which is in the horizontal plane by 2/3. Comparing the two schemes in Figure 1, the process of changing the read from 0 to 1 requires negligible time as compared with the interrogate. Both schemes require the same interrogate changes. Thus the Modified Lewin scheme is faster, but requires more digit logic which may operate at room temperature.

An Algorithm for Comparison Responses for a Parallel-Readout System

Often the multiple responses of information from an associative memory are not particularly desired in an ordered retrieval form, but their responses are desired in the category of greater than, less than, or between two limits comparisons. If the multiple responses are in ordered form, this information can be automatically extracted by stopping and/or starting the readout with appropriate digit logic. In a parallel readout, multiple interrogations must be used to obtain all words dictated by greater than, less than, or between limits command. An algorithm governing the multiple interrogations for comparison responses is as follows:

Greater-than

(1) Choose the first 0 from the least significant digit (LSD) of the entry word and replace the 0 with a 1.

(2) Replace all lesser digits with don't cares, Ø, and all greater digits by the digit of the entry word.

(3) Repeat for all remaining 0's sequentially.

The number of interrogating words is equal to the number of 0's in the entry word.

Less-than

(1) Choose the first 1 from the LSD of the entry word and replace the 1 with a 0.

(2) Replace all lesser digits with don't cares, Ø's, and all greater digits by the digits of the entry.

(3) Repeat for all remaining 1's sequentially.

The number of interrogating words is equal to the number of 1's in the entry word.

Between-limits

(1) Perform a less-than comparison of the larger of the two entry words until all remaining non-Ø digits match the smaller of the two entry words.

(2) Perform a greater-than comparison of the smaller entry word until all remaining non-Ø digits match the greater entry word. All interrogating words which were formed preceding these matches are used as interrogating words. The two sets of interrogating words form the complete set necessary for the between-limits comparisons.

The rules of logic for these comparisons are demonstrated by the following examples:

Example 1: The word represents all U.S. planes with the following information: Serial No., Type No., armament, pilot, position latitude, and position longitude. Seven digits are used to describe the latitude (considering north latitude only). To obtain all information about planes greater than (north of) latitude 30° (binary 30 = 00011110), the interrogating words are:

Interrogating Words	*Decimal Equivalent*
00011111	(31)
001ØØØØØ	(32-63)
01ØØØØØØ	(64-127)
1ØØØØØØØ	(128-255)

Thus, it requires only four interrogating words to obtain all information about planes north of 30° north latitude. If there is one word in the memory for each interrogating word, the greater-than comparison is performed with speed equal to the case of having the greater than comparison logic built into the memory. If there will be many responses when a comparison is used, the costs of in-the-cell logic is unwarranted.

Example 2: To obtain all information, as in Example 1, related to all planes between latitude 30° (00011110) and latitude 90° (01011010).

Interrogating Words	*Decimal Equivalent*
01011ØØØ	(88-89)
01010ØØØ	(80-87)
01ØØØØØØ	(64-79)
00011111	(31)
001ØØØØØ	(32-63)

The first three interrogating words are part of the less-than comparisons of the upper limit, 90. If the less-than process is continued one more step, the non-0 digits of the interrogatory word would match the corresponding digits of the lower limit, 30. The next two interrogating words are part of the greater-than comparison of the lower limit, 30. If continued, the digits of the next interrogating word would match those of the upper limit, 90. Five interrogating words are required to address the memory between the limits of 30 to 90.

Comparison Among Systems

The comparison between parallel readout and Modified Lewin readout is:

(1) The Modified Lewin readout is a much faster readout. However, the write speeds are about the same.

(2) The Modified Lewin readout permits an ordered retrieval of information. The parallel readout does not unless the ordered retrieval coincides with the geometrical order.

(3) The Modified Lewin requires more digit logic to the interrogate, read, and logically interpret. This digit logic can be external to the cryostat.

The associative memory with Modified Lewin readout appears to have the edge in operation at the cost of some additional hardware.

Comparing the associative memory with Modified Lewin readout with a random-access memory, the random-access memory will have a faster write process. However, if one assumes a constant word length, there must be a crossover point as the memory size increases at which the read process of the associative memory is faster than that of the random access. The speed of the associative memory is independent of number of words, whereas the speed of the random-access memory is proportional to $n^{1/2} \log_2 n$. The associative memory has inherent redundancy in the system; the random-access memory does not. One can build an associative memory for example, 20% larger than specified for a given application. If 20% of the word locations do not operate, but these word locations are scattered geometrically throughout the memory, the memory operates within specification, since there is no addressing according to geometry. This latter point is a strong feature in batch fabrication. The associative memory is more complex and can be estimated as costing more per bit unless the inherent redundancy feature makes a sufficient difference in memory plane yield between the two systems.

HIGH-SPEED TRANSISTOR-TUNNEL-DIODE SEQUENTIAL CIRCUITS

By

J. J. Amodei and J. R. Burns

RCA Laboratories
Princeton, New Jersey

Summary—The high-speed current gain and the inherent memory of tunnel diodes, coupled with the unidirectionality and voltage gain of transistors makes their combination most attractive for use in sequential circuits of any type. This paper describes some register, counter, and shift-register configurations for operation at high repetition rates. The results show that operation of any of these circuits at clock rates between 100 and 200 mc can be achieved reliably under realistic conditions and using commercially available components. The circuit complexity is also minimized due to the added flexibility furnished by the combined properties of transistors and tunnel diodes. The tolerance requirements and fan-out capabilities of the circuits are adequate for most commercial applications.

Introduction

SEQUENTIAL CIRCUITS such as registers, counters, and shift registers stand out among the areas of computers in which the combination of transistors and tunnel diodes have proved most useful. This is due mainly to the fact that these circuits take full advantage of the memory property of the tunnel diode, whereas in other applications such as logic gates[1,2] this property proves to be a hindrance. Furthermore, the packaging and interconnection problems of iterative sequential circuits (shift registers and counters) are relatively simple, thus permitting efficient utilization of the high speed capabilities of hybrid circuits in an actual system. The purpose of this paper is to show some of the schemes with which these functions can be realized when utilizing transistor–tunnel-diode combinations and to discuss the performance features and relative merits of the different approaches.

---

[1] J. J. Amodei and W. F. Kosonocky, "High-Speed Logic Circuits Using Common-Base Transistors and Tunnel Diodes," *RCA Review*, Vol. 22, No. 4, p. 669, Dec. 1961.

[2] M. H. Lewin, "Negative Resistance Elements as Digital Computer Components," *Proc. Eastern Joint Computer Conference*, Dec. 1959.

Register

In order to achieve ring-counter or shift-register operation it is necessary to have a permanent storage module or register and means for transferring the information stored in such a module to a similar one on command of an external pulse.

When dealing with transistor–tunnel-diode circuits, the register part can be realized very readily with any of several configurations. Figure 1 shows a register with the transistor in the common-emitter configuration; this was chosen as the basic module for all the circuits which are described because it provides inversion as well as current and voltage gain. The performance features of this circuit are discussed in some detail because they play an important role in the choice of the configurations.

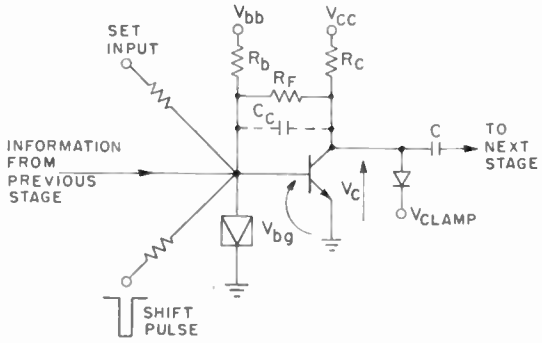


Fig. 1—Basic register module.

Figure 2 shows the static load line diagram for the circuit of Figure 1 with the two stable points for the cases with R_F out of the circuit and R_F in the circuit. The effect of R_F* is to decrease excess current flowing into the base when the transistor is in saturation, thus reducing storage time and minimizing reset power requirement.

The bias current, I_{bb}, is supplied by R_F and R_b and is designed to be as close to the nominal peak current of the tunnel diode as the tolerances permit. The tunnel diode in the high-voltage state corresponds to the register being "set" or loaded with a "one." The memory property is inherent in the arrangement in that once the register is "set," it is necessary to change the polarity of the input in order to "reset" it (i.e., in order to return the tunnel diode to the low voltage state).

* General feed back schemes for transistor-tunnel-diode circuits were originally suggested by Mr. W. F. Kosonocky of RCA Laboratories.

In the shift register and counter, this is done each cycle by the "shift pulse," whose function is to reset all the stages to the "zero" state so that they may be able to receive the information previously stored in the preceding stage. The behavior of the circuit during reset and the requirements for reliable information transfer between stages determines the type of coupling necessary between registers in order to achieve counter or shift register operation.

The use of an inverting register results in a low collector voltage in response to a high tunnel diode voltage and vice-versa, so that one additional inversion would appear to be needed in the coupling network for proper transfer of information. This problem is avoided when

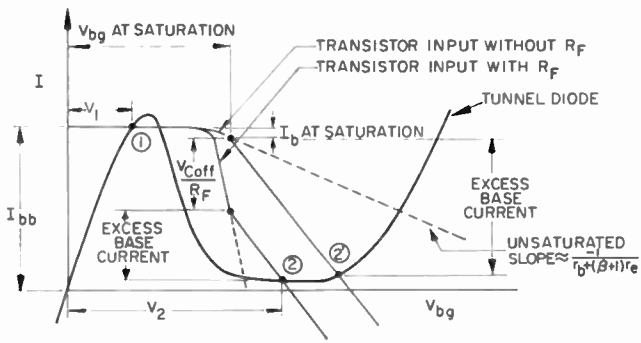


Fig. 2—Static *V–I* characteristics for register with and without R_F.

dynamic readout is used for information transfer, since the presence or the absence of a change of state is used to determine the prior information content of the stage. This method has the additional advantage of generating the signal that transfers the information only after the clear pulse has been applied, thus avoiding the burden of additional time gating to prevent information coupling during the storage portion of the cycle.

A differentiating capacitor in the interstage coupling link achieves this purpose by producing a current pulse of the proper polarity when a stage with a "one" is cleared. It is then only necessary to insure that the timing and amplitude of such a pulse be such that this information may be transferred reliably. In order to understand the parameters that influence the reliable transfer of information, it would be instructive to start with a brief discussion of the static and dynamic conditions encountered in this circuit during normal operation.

Since the tunnel diode is fully characterized for our purpose by the static *V-I* plot and its shunt capacitance, the only complexity is intro-

duced by the transistor, whose behavior is dependent on a multitude of variables. The concept of a switching load line will be used as a catchall to denote the very-high-frequency *V-I* characteristics that do not follow the static plot. This switching characteristic is a function of the instantaneous values of input current and voltage and also of the previous history of the device (i.e., the stored charge in the base of the transistor). Figure 3 illustrates the variety of input characteristics that one must consider depending on the state of the transistor

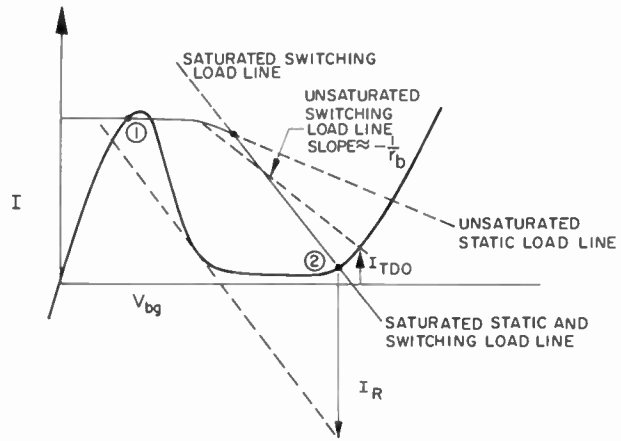


Fig. 3—Static and switching load line diagram neglecting R_F.

and the time constants involved. For simplicity, the effect of R_F is neglected. The unsaturated switching load line is essentially the one that governs the switching from the low- to the high-voltage state, provided that the transition occurs in a time so short that the stored charge in the transistor does not change appreciably (i.e., the tunnel diode switches before the collector current has a chance to build up significantly). The unsaturated static load line would be the one to consider if the switching occurred very slowly and the transistor was not permitted to saturate. When the collector voltage reaches saturation level, the incremental resistance looking into the base of the transistor will be determined no longer by the normal transistor parameters, but rather by the characteristics of the emitter diode and base region at the current levels in question.* It is because of this that an additional load line (the saturated switching load line) is shown in Figure 3; this is the load line that determines the switching behavior

* This assumes that the impedance of the load is large compared with this incremental resistance.

of the circuit while the transistor is in saturation. For our purposes, the slope of this load line was found to be approximately that of the emitter diode with the collector open, and is shown as a straight line to facilitate this purely qualitative description. The slope of this line is steeper than that of r_b, and this characteristic must be taken into account during resetting. It should be noted that until the stored charge is removed from the base the switching behavior during resetting will be governed by the saturated load line which, unlike the static load line, maintains a low resistance even when the flow of current in the base lead has been reversed. This implies that the reset current requirements determined from the static characteristics are not in general sufficient to achieve reliable turn off, particularly if the clear pulse duration is shorter than the storage time. In fact, with some transistors, it is possible to force the tunnel diode to the low-voltage state during the presence of the pulse and have the tunnel diode return to the high-voltage state after the pulse is removed.

Therefore it is necessary to consider the switching and saturated behavior of the transistor in determining reset pulse amplitude and duration. To achieve fast turn off, the reset current should be sufficient to force the tunnel diode to the low-voltage state immediately. An estimate of the minimum reset current required to achieve this can be obtained by drawing a line parallel to the saturated switching load line and tangent to the negative-resistance portion of the tunnel diode *V-I* characteristic. The current required is I_R, as shown in Figure 3.

If the above conditions are satisfied, the sequence of events following the application of a clear pulse to a circuit with the tunnel diode in the high-voltage state will be as follows: initially, the tunnel diode will be forced to switch to the low-voltage state; this will cause the transistor to begin to come out of saturation and finally turn off with a resulting sharp change in the collector voltage. The relative timing of these waveforms is shown in Figure 4 together with the current waveform through the coupling capacitor, C, of Figure 1. The significant times are the combined delay time, T_D, the storage time, T_S, and the fall time, T_F, which determines the width of the current pulse flowing through the coupling capacitance.

If we can assume that the intrinsic transistor switches in a time significantly shorter than the collector RC time constant, we can approximate the fall of the collector voltage with the following expression:

$$V_c = \frac{V_{cc}R_T}{R_C}\left(1 - \exp\left\{-\frac{t}{R_T C_T}\right\}\right), \tag{1}$$

where

$$R_T = \frac{R_F R_C}{R_F + R_C},$$

$$C_T = C + C_{cg},$$

and C_{cg} is the capacitance from collector to ground and C_c is the

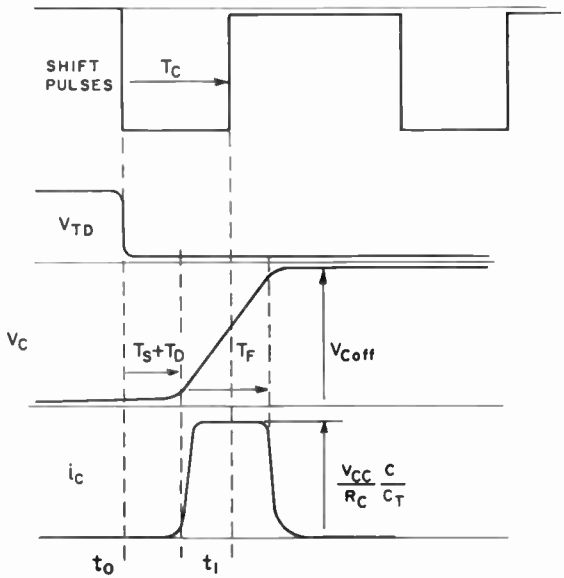


Fig. 4—Timing sequence at reset.

collector-to-base capacitance.

The fall time then will be given by

$$T_F = -R_T C_T \ln\left(1 - \frac{V_{c(off)}}{V_{cc}} \frac{R_C}{R_T}\right), \tag{2}$$

where $V_{c(off)}$ is the collector voltage when the transistor is off; $V_{c(off)}$ is slightly higher than V_{clamp}.

For $V_{c(off)} \ll V_{cc}R_T/R_C$, the fall time is

$$T_F \approx \frac{V_{c(off)}}{V_{cc}} R_C C_T. \tag{3}$$

The current pulse through the coupling capacitance is the signal which is available to transfer the information previously stored in a given stage to the succeeding stage. If this pulse is the current that actually loads the succeeding stage, its timing must be such that the pulse is still present at the input of this stage when the shift pulse has terminated; otherwise, the inhibiting action of the shift pulse would prevent the information from being written into this stage.

The magnitude of the current flowing through the total capacitance, C_T, during turn-off is given as a function of instantaneous collector voltage, V_c, by

$$i_{CT} = \frac{V_{cc}}{R_C} - \frac{V_c}{R_T}. \tag{4}$$

The current that flows into the coupling capacitance is

$$i_C = \frac{i_{CT}C}{C_T} = \left(\frac{V_{cc}}{R_C} - \frac{V_c}{R_T} \right) \frac{C}{C_T}. \tag{5}$$

For $V_{c(off)} \ll V_{cc}R_T/R_C$ the amplitude of this current pulse during turn-off may be approximated by

$$i_C \approx \frac{V_{cc}}{R_C} \frac{C}{C_T}. \tag{6}$$

During turn-off there is also an undesirable feedback current that will flow through the collector-to-base capacitance, C_c, the magnitude of which is given by

$$i_F = \left(\frac{V_{cc}}{R_C} - \frac{V_c}{R_T} \right) \frac{C_c}{C_T}. \tag{7}$$

It may also be approximated under the same assumption as in Equation (6) by

$$i_F \approx \frac{V_{cc}}{R_C} \frac{C_c}{C_T}. \tag{8}$$

This feedback current tends to rewrite the information in the same stage, which is particularly detrimental to the operation of the counter described. A means to remedy the situation is discussed.

The bias current of the tunnel diode is provided partly through R_b

and partly through R_F; this is done in order to minimize both storage time and reset current requirements. Thus, when the tunnel diode is in the low-voltage state, the total bias current, I_{bb}, is given by

$$I_{bb} = \frac{V_{bb}}{R_b} + \frac{V_{c(off)}}{R_F} - \frac{V_1}{R_B}, \tag{9}$$

where V_1 is the voltage from base to ground obtained from the static plot of Figure 2 and R_B is the total resistance from base to ground. The current flowing into the base of the transistor immediately following turn-on will be, from Figure 3,

$$I_{b0} \approx I_{bb} + I_{in} - I_{TD0}, \tag{10}$$

where I_{in} is the input pulse from the previous stage or from external sources and I_{TD0} is the current taken by the tunnel diode at the intersection of the "unsaturated switching load line" and the tunnel-diode characteristic. If this intersection is in the valley region, then

$$I_{TD0} \cong I_v,$$

where I_v is the valley current. After the transistor is fully turned on and the input pulse has terminated, the total current into the base of the transistor can be approximated as follows (neglecting saturated collector-to-ground voltage):

$$I_{bs} \approx \frac{V_{bb}}{R_b} - \left(I_v + \frac{V_2}{R_B} \right). \tag{11}$$

A very important condition on this current is that

$$I_{bs} \geq \frac{V_{cc}}{R_C} \frac{1}{\beta_{min}}. \tag{12}$$

This will insure that the circuit will retain its memory and that the transistor will be in saturation when the tunnel diode is in the high-voltage state.

COUNTER

As mentioned earlier, counting can be achieved by simply transferring a "one" along a chain of normally unloaded modules on command

of each clock pulse. This implies that the counter stage receiving the information will not have contained a "one" in the previous cycle and that the stage that transfers the information will not receive an input from the previous stage while this transfer is in progress. It is these properties of the counter that permit reliable coupling between stages without the additional interstage storage elements and other complications that are needed in a shift register.

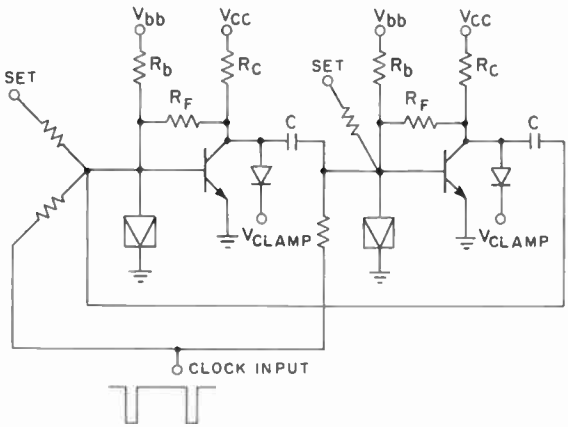


Fig. 5—Scale-of-two counter.

In order to describe the operation in detail it will be convenient to refer to Figure 5, in which two stages are shown connected in a ring to form a scale-of-two counter. A set pulse of short duration designed to occur between two successive clock pulses may be applied as a method of starting the counter. This sets one of the tunnel diodes in the high-voltage state and the corresponding transistor is turned on. When the clock pulse occurs, it resets the tunnel diode and the transistor is turned off, thus producing the signal which is used to set the tunnel diode of the succeeding stage. Since the counter has no intermediate storage, the information must be loaded in the succeeding stage within the duration of the current pulse i_C, Figure 4, which is also the duration of the collector turn-off voltage ramp (T_F). This condition, which is avoided when intermediate storage is available, also causes the tunnel diode in the stage that is being reset (and which should not receive any information) to be uninhibited while the feedback current, i_F, Equation (7), is still flowing in response to the changing collector voltage. It is necessary, therefore, to take this current into account and design I_{bb} to be lower than required by d-c considerations to avoid erroneous

operation. Since this feedback current is proportional to the collector-to-base capacitance, C_c, it reaches a peak at very low values of collector voltage and decreases as this voltage increases. The tunnel diode bias, on the other hand, is partially provided by R_F and thus is at a minimum during the time that the feedback current is at a maximum, which considerably improves the tolerance situation. Because of this, it is necessary to take into account only the values of collector capacitance at voltages near the clamp voltage rather than the larger capacitance encountered near zero volts. Thus, for a practical case,

$V_{oo} = 15.5$ volts,

$R_O = 1000$ ohms,

$R_F = 750$ ohms,

$R_T = 430$ ohms,

$C = 6.8 \times 10^{-12}$ farad,

$C_{co} = 10^{-12}$ farad,

$C_c = 2 \times 10^{-12}$ farad at $V_c = 2.5$ volts (or $C_T = 10 \times 10^{-12}$ farad).

From Equations (5), (7), and (3) we find

$i_O \approx 6.6$ milliamperes,

$i_F \approx 1.9$ milliamperes,

$T_F \approx 3 \times 10^{-9}$ second.

This indicates that the tunnel-diode bias must be designed so that under the above conditions it will be more than 1.9 ma from the peak current and, at the same time, be less than 6.6 ma from the peak if it is to switch in response to the signal from the previous stage. This condition is not difficult to meet with conventional circuit components.

As was mentioned in the section on information transfer, the timing requirements are as follows (see Figure 4):

$$T_S + T_D + T_F > T_C. \tag{13}$$

The fall time, T_F, is dependent on the circuit design parameters. The storage time, T_S, for the circuit used was found to vary only slightly from unit to unit and, in fact, because of the dynamic nature of the operation which does not permit the charge to build up, it is practically negligible under actual operating conditions. The total

Fig. 6—Coupling-capacitor current waveform.

storage time plus the combined delay through the transistor and the tunnel diode were of the order of 1 nanosecond. Figure 6 shows the phase relationship between the clock and the current through the coupling capacitor in a circuit similar to the counter. Although the circuit configuration and the parameters involved are somewhat different from those of the counter (the width and amplitude of the current pulse are larger for the counter), the delays are approximately the same.

Performance of the Counter

An eight-stage counter was built and tested as a ring counter and then was incorporated as part of a 100-mc, 8-bit shift-register system to control the number of shifts. The counter-circuit diagram is shown in Figure 7.

As a ring counter, the circuit is capable of circulating any pattern of pulses indefinitely provided that the pattern does not contain "ones" in consecutive stages. The system was tested as a ring counter at clock rates of 100 and 180 mc, and pictures of representative patterns are

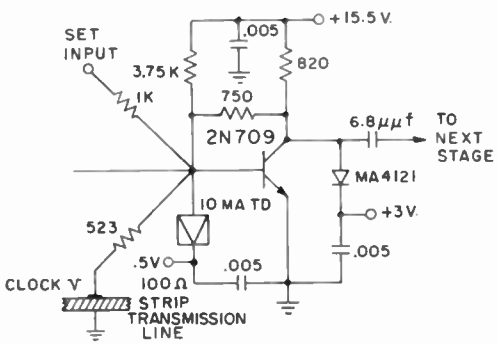


Fig. 7—Counter circuit.

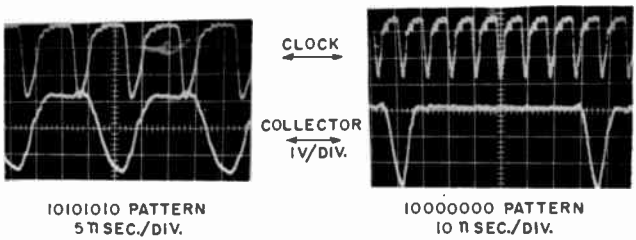


Fig. 8—100-mc ring-counter test.

shown in Figures 8 and 9. The same circuit and voltage values were utilized at both clock rates.

Shift Register

The shift register, whose function is to shift the location of information on command one or more spaces at a time, is one of the most versatile digital systems in existence and is, therefore, necessarily more complex than the counter previously described. Since its implementation requires the combination of several functions, each of which may be realized in many ways, the number of possible shift-register configurations is very large, each with different design and performance features.

The first requirement of a shift register, that of memory or permanent storage, is realized by the basic register of Figure 1. This register is used in all the shift-register systems described, so that the only differences in the various systems are the coupling networks. In general, information is shifted in three steps:

(1) The permanent storage registers are cleared, i.e., a pulse is applied to the registers which resets all stages to binary "zero."

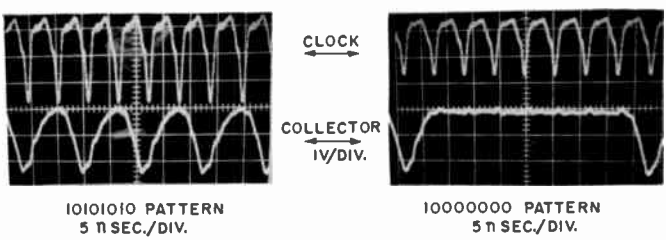


Fig. 9—180-mc ring-counter test.

(2) The change in state of the memory register is used to load an associated intermediate storage stage.

(3) The intermediate stage releases the information and stores it in the succeeding register.

The most straightforward way in which this sequense can be realized is by making the intermediate stage perform as a memory register, also. The "clear" pulse will then load the intermediate register, which retains the information until another pulse, applied to the intermediate registers only, transfers the information to the succeeding stage. The operation of such a system is extremely reliable, although this reliability is generally achieved at the expense of a reduced shifting rate, more hardware, and a more complex two-phase-shift pulse source.

In high-speed applications, intermediate storage is usually "dynamic"; i.e., information is retained for a predetermined time interval by a suitable reactive element, such as a charged capacitor or an energized inductor. At some specified time during this interval, the dynamic element will transfer its energy to the next memory register, completing the shifting cycle. The counter circuit is an example of a system utilizing a charged capacitor to store the information temporarily. However, as was previously mentioned, the utilization of the same simple counter arrangement to shift patterns which contain "ones" in adjacent stages is unreliable. If a stage containing a "one" is cleared and then loaded with another "one" before the succeeding stage has been loaded (which could happen due to slight differences in circuit elements) the transfer of information would not occur. It is seen that the operation of such a dynamic system is more unreliable than a two-phase system, due to the necessity for precise control of the timing elements and the shift pulse width.

The first two systems discussed use dynamic coupling stages, one being an ordinary coaxial transmission line, the other a tunnel-diode monostable circuit. The third system, a "feedback reset" system, uses a combination of dynamic and permanent storage in the coupling stage, and is designed to incorporate the desirable features of both modes of operation. All systems require only a single-phase pulse source, and have been operated at shift rates in excess of 100 mc.

Transmission-Line Shift Register

The most straightforward method of delaying the signal information is by a transmission line of adequate impedance connected as shown in Figure 10. To avoid multiple reflections, it is necessary that at least one end of the line be terminated with its approximate charac-

teristic impedance. Since the tunnel diode, upon switching, shows a highly nonlinear relationship between voltage and current, the only possible way to approximate a termination would be to use an impedance level such that, for the desired signal current, the voltage across the line is significantly larger than the tunnel-diode voltage swing. When 10-ma peak current tunnel diodes are used, the optimum impedance will lie somewhere between 200 and 400 ohms. Because of commercial availability, 185-ohm cable (RG 114/u) was utilized, although a higher impedance and smaller physical dimension would have been desirable.

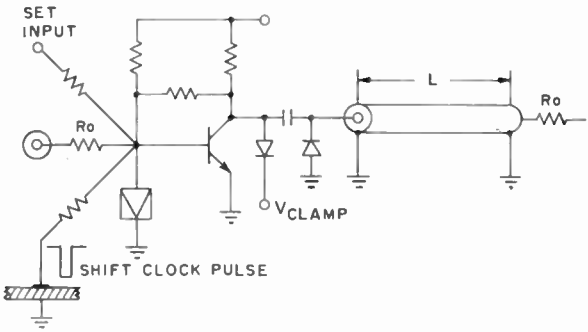


Fig. 10—Delay-line shift register.

The operation of the shift register is very similar to that of the counter except for the use of the transmission line as an interstage delay. This considerably simplifies the timing requirements, since it is no longer necessary to ensure that the pulse which is to transfer the information (namely, the current through the coupling capacitance) be wide enough to be present when the shift pulse is terminated. In fact, it is advantageous to change the timing requirements which were derived for the counter in reference to Figure 4 to the following:

$$T_S + T_D + T_F < T_C. \tag{14}$$

This ensures that there is no further collector-voltage change after the termination of the shift pulse, and therefore eliminates the problem of the capacitive feedback current (i_F) discussed in connection with the counter.

The pulse that is generated when clearing a stage which previously contained a "one" is delayed by the transmission line so that its leading

edge will arrive at the input of the succeeding stage immediately following the termination of the shift pulse.

The optimum delay time is then given by

$$T = T_O - (T_S + T_D). \tag{15}$$

Figure 6 shows the pulses that are generated across the coupling line each time that a stage containing a "one" is reset by a clock pulse in a circuit such as that shown in Figure 11.

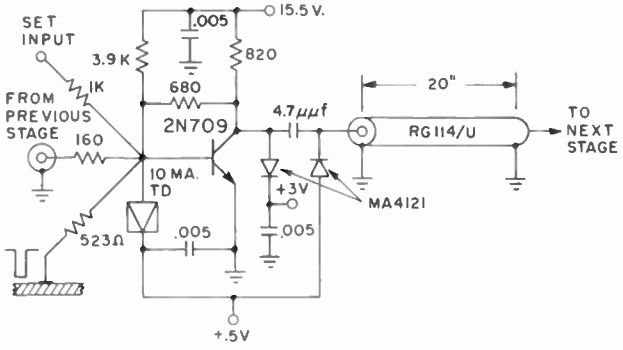


Fig. 11—Shift-register circuit diagram.

Performance

A seven-stage shift register was constructed using the circuit shown in Figure 11. The complete system was tested at 100 mc; some representative patterns are shown in Figure 12. In the photographs, it may be noticed that the collector swing of the stage under observation is slightly greater when it receives a "one" preceding a "zero" in the pattern. This is due to insufficient clipping of the negative spike generated when the preceding stage turns on. This spike is delayed by the transmission line and arrives in time to slightly inhibit the turn-on drive of the transistor, thus resulting in a small decrease of the collector swing. When the stage under observation receives a "one" that is to be followed by a "zero," the preceding stage is maintained in the "zero" state, so that the inhibiting spike is not generated and the collector achieves its full swing. This effect had very little influence on the performance; it can be completely eliminated by using a series diode or a clipping diode with higher forward conductance.

Tests of the completed 7-bit system were not made at frequencies above 100 mc. Results of tests on the individual circuits show that

operation at frequencies between 150 and 200 mc is possible with transistors presently available. An increase of the speed of the transistors (namely higher f_T and lower collector capacitance) would result in an approximately proportional increase in the repetition rates possible with either the counter or the shift register. In oven tests, the shift register operated between room temperatures and 120° C. Tests were not made at higher temperatures because the dielectric in the input and output cables would soften and short circuit the lines.

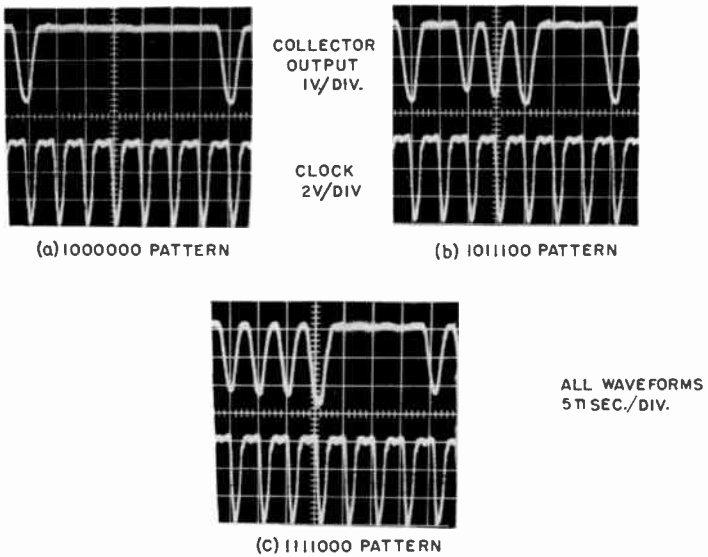


Fig. 12—7-bit shift-register patterns.

Monostable Shift Register

The shift-register circuit of Figure 13 again uses a common-emitter transistor–tunnel-diode gate for permanent storage with a tunnel-diode monostable circuit for dynamic storage and transfer of information. The monostable circuit[3] in the coupling network can provide some gain as well as temporary storage, so that slightly shorter turn-on times may be obtained. The operation of the circuit may be described by referring to Figures 14(a) and 14(b). The quiescent point "a" shown

[3] R. H. Bergman, "Tunnel Diode Logic Circuits," *Trans. I.R.E. PGEC*, Vol. EC-9, No. 4, Dec. 1960.

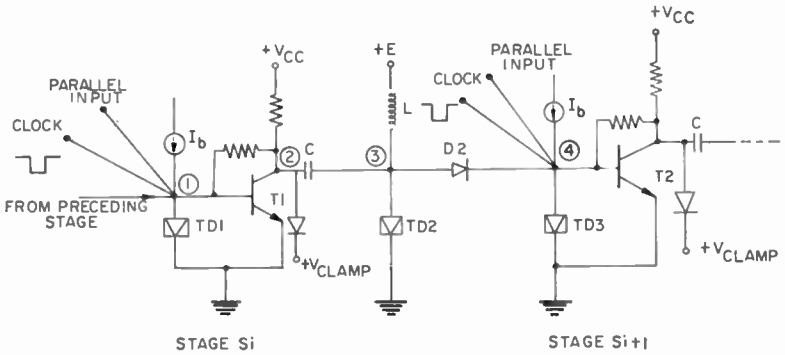
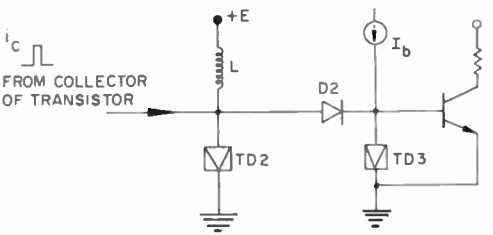
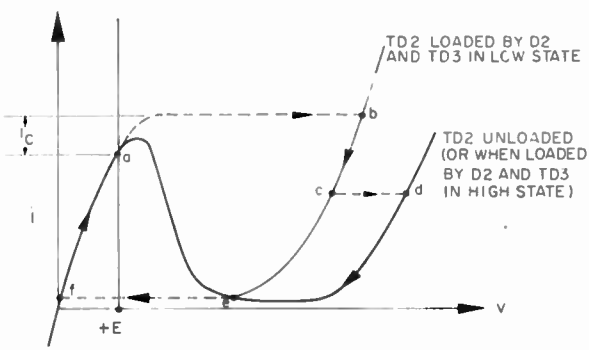


Fig. 13—Monostable shift-register circuit.



(a)



(b)

Fig. 14—Monostable circuit: (a) circuit diagram and (b) graphical operation.

in Figure 14(b) is the only stable point for the circuit under any condition. If the previous stage contained a "one," a positive current spike is applied to the monostable input when the register stage is cleared; this will switch the tunnel diode to its high-voltage state at point "b." Note that the current remains constant during switching due to the action of the inductance in preventing instantaneous changes of current. Since "a" is the only stable operating point, the current through the tunnel diode will decay with an inductive time constant, maintaining its voltage approximately constant in the region "b"-"c." The next stage will not be switched immediately, however, since the shift pulse, I_S, is still present at the input, thereby tending to keep all the stages in the reset, or "zero" state. The inductance, L, is chosen to make the pulse width greater than the shift pulse width, so that when I_S is terminated, the monostable will switch the next tunnel diode to the high-voltage state. At some point "c," the shift pulse terminates and TD3 will switch to its high-voltage state, thereby back-biasing D2 and causing the monostable to jump abruptly to point "d." The monostable tunnel diode continues the relax to point "e," at which time the circuit will switch to the low-voltage state at "f," and then to its equilibrium point "a." The cycle is now completed as each stage has been loaded with the information stored in the preceding stage prior to the application of the shift pulse, and the monostable circuit has returned to its quiescent point, "a."

Thus, from Figures 4 and 14, the maximum repetition rate is given by

$$f_{max} = \frac{1}{(T_S + T_D + T_{ON} + T_{OFF})_{max}} \tag{16}$$

where T_D = total delay of transistor and tunnel diode,

T_S = transistor storage time,

T_{ON} = "on" time of monostable circuit (i.e., time from "b" to "e" in Figure 14(b),

T_{OFF} = time for relaxation of monostable circuit to equilibrium (from "f" to "a" in Figure 14(b)).

The "on" time must be greater than T_C for proper operation, but T_C must be wide enough to bring the transistor out of saturation. Furthermore, the "off" time is fixed once the "on" time has been chosen, as it is a function of the inductance and tunnel-diode characteristic. The "off" time is usually much greater than T_{ON} because of the low

impedance of the tunnel diode in region "f"-"a" (Figure 14(b)) which reduces f_{max} considerably from that required for information transfer only.

Feedback Reset Shift Register

The feedback reset shift register is designed to eliminate the undesirable timing problems associated with the two systems previously discussed without any sacrifice in the speed of operation. In looking at the operation of the transmission line and monostable shift registers,

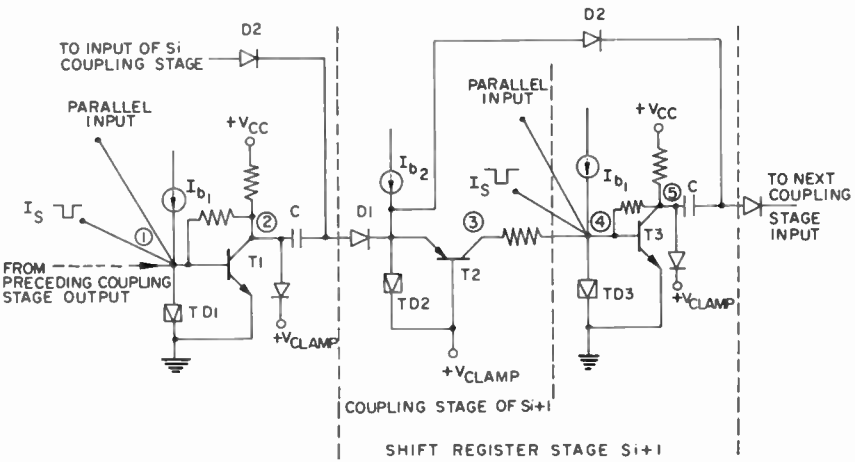


Fig. 15—Feedback reset shift register.

it is apparent that the major loss in reliability is due to the fixed timing intervals of the coupling networks. The main function of the intermediate stage is to maintain the information long enough so that all the register stages will receive the information. However, component tolerances make the timing condition slightly different for each stage, making optimum operation with a fixed timing interval virtually impossible. The optimum interstage network should, therefore, maintain the information only until the stage it is loading has switched to the required state, at which time it should return to its off condition.

The feedback reset shift register shown in Figure 15 exemplifies one way in which this type of coupling network may be realized. The memory is provided by the basic register as described previously, the high and low states of the tunnel diodes TD1 and TD3 representing binary "one" and "zero," respectively. The main feature of this

approach is the fact that it makes use of a bistable common-base hybrid register instead of a monostable circuit for temporary storage. This method of intermediate storage would noramlly require an additional pulse to reset the coupling stage after the proper information has been gated into the next register. However, an external gating

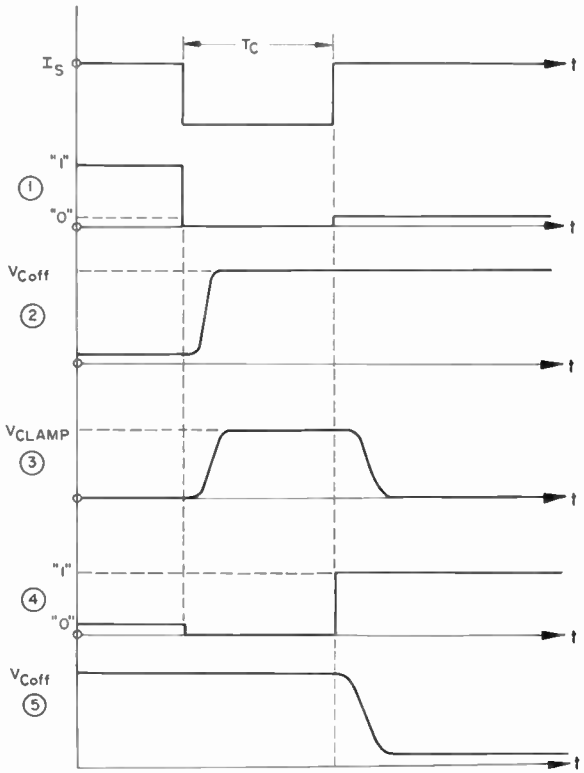


Fig. 16—Timing diagram of feedback reset shift register.

pulse is unnecessary in this system since the resetting pulse is generated internally when the succeeding register transistor turns on.

The operation of the system can be described by assuming that register S_i contains a "one" while S_{i+1} contains a "zero." A negative shift pulse, I_s, is applied simultaneously to all register inputs, resetting all stages to state "zero" as shown in the timing diagram of Figure 16. As the transistor turns off, the resulting positive step produces a current pulse through C in Figure 15 which is routed by D1 into the succeeding coupling stage, switching TD2 to the high-voltage state.

This will cause the bias current, I_{b_2}, to flow through T2 and into the next register tunnel diode, TD3, tending to switch it to the high-voltage state ("one"). As the coupling gate is bistable, this current will continue to flow irrespective of any imposed timing interval; such was not the case in the monostable and transmission-line shift registers. Upon termination of the shift pulse, stage S_{i+1} will switch to state "one," the previous state of stage S_i. The collector voltage ("5" in Figure 16) will then fall toward saturation, producing a negative pulse through diode D2 of sufficient amplitude to reset the preceding coupling stage. Proper information has been stored in all stages, and the coupling networks have been reset so that another shift cycle may now be initiated. The magnitudes of the capacitive set and reset currents to the coupling stage are given by the relationships derived in the section on the register.

The feedback-reset property of the circuit takes full advantage of transistor speed, since as soon as the register is turned off and information is loaded from the coupling stage, this stage is automatically reset, thereby immediately readying the circuit for another shift cycle. This is very desirable in high-speed work in that the maximum shift rate is determined solely by the turn-on and turn-off times of the register transistors which, in the above circuits, is of the order of 3 to 4 nanoseconds. Furthermore, the tunnel diode in the coupling stage provides current gain at very high speed, thereby enabling faster turn-on of the register and improving the circuit tolerances. Thus, the system combines "asynchronous" and "dynamic" operation, achieving high reliability without sacrificing repetition rate.

Performance

An eight-stage shift register using feedback reset was constructed with 10-ma germanium tunnel diodes, 2N709 n-p-n silicon transistors in the register, and 2N769 p-n-p germanium transistors in the coupling stage. The system was built in a ring, i.e., the last stage connected to the first so that the binary number would circulate indefinitely at the input clock rate. Figure 17 shows the output waveform of one register stage for two representative binary numbers circulating in the register at a 100-mc clock rate. The negative pulses used for shifting are shown in the same figure.

The entire eight-stage shift register was placed in an oven to check operation at elevated temperatures. The system operated correctly at 100 mc for temperatures as high as 90° C. The output waveform at this temperature is shown in Figure 18. Note that the voltage swing is reduced under these conditions. However, for a shift register this

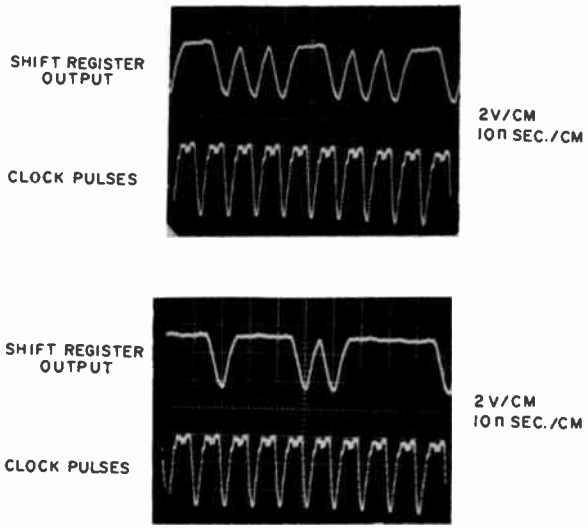


Fig. 17—Output waveforms of feedback reset shift register.

is not important as long as the information can be transferred at this speed. It should be noted that, although operation was possible at 90° C, it is recommended that the maximum temperature should not exceed 60° C, as the input characteristics of the germanium transistor in the coupling stage would vary enough to impair the over-all reliability.

Each register may drive two equivalent register stages (4 ma for each register) as well as the succeeding stage of the shift register. Furthermore, since information transfer depends on a-c coupling, it is possible to obtain larger fan-out by decoupling the loads during switching. In practice, this can be done by placing an inductance in series with the load as shown in Figure 19. The resultant inductive time constant should be large compared to the transistor fall time.

The complete circuit diagram of the feedback reset system is shown in Figure 20. The 0.5-volt supply at the cathode of the register tunnel diode is necessary to turn on the register transistor when the tunnel

Fig. 18—Output waveform at 90°C.

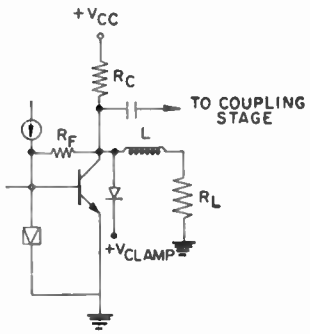


Fig. 19—Register load inductive decoupling.

diode is in its high-voltage state, since the silicon transistor has a conduction voltage of 0.8 volt as compared to 0.5 volt for the germanium tunnel diode. To check the circuit tolerance, binary numbers were loaded in at 100 mc and the three supply voltages were varied individually until the number was destroyed. The variations tolerable on the supplies under these conditions were at worst ±15%. These variations seem adequate for most applications, and they may be improved upon by using precision 1% resistors rather than the 5% resistors used throughout this system. A worst-case tolerance analysis of the circuit, assuming ±5% peak current tunnel diodes and 5% overdrive current for switching, has shown that the sum of the tolerances on voltages and resistors can be ±15%.

Programmed Shift-Register System

To test the performance of the counter and feedback reset circuits under actual conditions, an eight-stage, 100-mc, programmed shift-

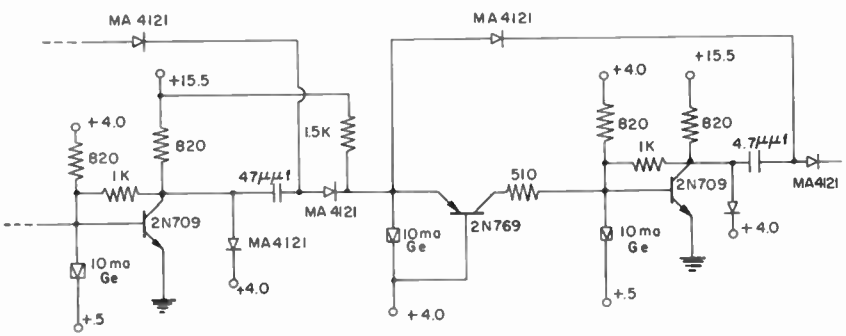


Fig. 20—Schematic diagram of feedback reset shift register.

register system was constructed. The system was designed to load any selected eight-bit binary number, shift the number around a given number of positions, and retain the shifted number in the register. The block diagram of the system is shown in Figure 21 and is seen to consist of an eight-stage shift register, an eight-stage counter, two 100-mc clock sources that are synchronized and drive the shift register and counter, input control circuitry for selecting and loading the proper information in the shift register, and a power gate which shuts off the

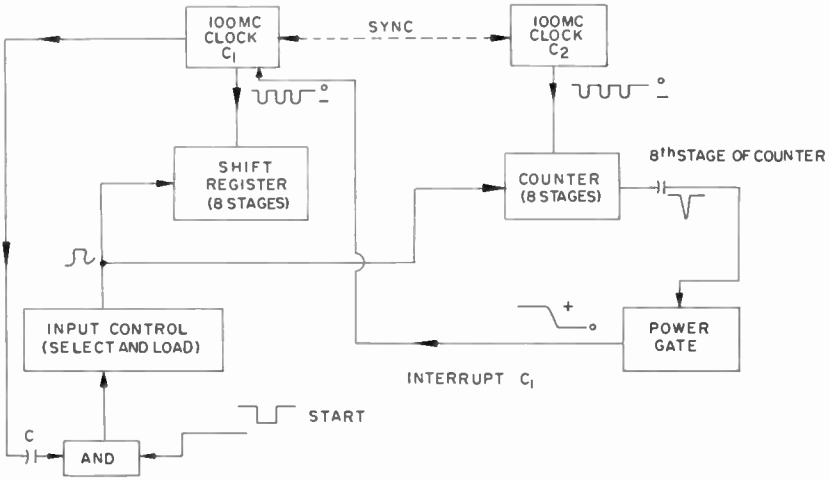


Fig. 21—Block diagram of shift-register system.

shift register clock when the required number of shifts have been executed.

A complete cycle of operation is as follows:

(1) The input control gates are set up according to the eight-bit binary number to be loaded into the shift register and the number of shifts to be performed.

(2) The clock pulse C_1 is differentiated giving an indication of when C_1 is initiated by the resultant negative spike of current. This is AND gated with a negative START pulse which is normally generated by the operator and the output is used to trigger the selected input control gates. The control gates load the shift register in the time interval between clock pulses, and also load a binary "one" into the $(8\text{-}n)^{th}$ stage of the counter, where n is the number of shifts required.

(3) The clock pulses, C_1 and C_2, then shift the binary information in the shift register and the binary "one" in the counter, respectively, once for each applied clock pulse. This process continues until, after n clock pulses, the 8^{th} counter stage receives the binary "one." The output (differentiated) of this stage then triggers a power gate which interrupts the shift-register clock, C_1.

Each phase of the system operation was found reliable, thereby demonstrating the feasibility of control for the counter and shift register at these high repetition rates.

CONCLUSIONS

The circuits discussed in this paper represent some of the simplest and best performing of a large number of possible choices available using the transistor–tunnel-diode combination. Although speed was a very important consideration in the design, reliability and compatibility with all-transistor circuitry played an important role in the choice of configurations and components. Emphasis on different objectives or specifications may well show that alternative designs using the combination of the two devices are more advantageous than those shown. The results that were obtained with these circuits, as well as the tests made in many other approaches, have shown that the combination of transistors and tunnel diodes is ideally suited for performing sequential functions. The very fact that so many choices are available to the designer in realizing each function is an indication of the flexibility of the approach.

The circuits used in the systems described were designed for operation at 100 mc, although tests were made at higher clock rates where possible. On the basis of tests on the systems and on individual circuits, it seems safe to say that reliable operation of any of these circuits in a system can be obtained at 200 mc with only minor design modifications, but considerable improvements in the packaging. Operation at much higher rates could be achieved with transistor–tunnel-diode combinations provided that tolerances are made tighter and that logical fan-out and other requirements are decreased.

Some other important features of the hybrid circuits described in this paper, from the standpoint of their potential use in digital computing systems, are summarized below:

(1) Single-phase clocking is all that is required to shift information.

(2) The registers may drive a resistively coupled load equivalent

to a logical fan-out of 2 with no effect on the shifting operation. When decoupled or gated, the loads may be increased considerably since loading does not affect the circuit tolerances if one is willing to wait 3-5 nanoseconds for power transfer to the load.

(3) Tolerance requirements on circuit components are very lenient, as exemplified by the tolerances required on the feedback reset shift register.

(4) The tunnel diodes, transistors, and diodes used are readily available commercially, making the use of the systems feasible in large-scale applications.

PARAMETER OPTIMIZATION OF AN FM/FM MULTICHANNEL TELEMETRY SYSTEM

By

David H. Sapp

RCA Communications Systems Division
Camden, N. J.

Summary—This paper discusses the r-f equipment parameters and parameter optimization of the r-f portion of a wideband telemetry system containing a combination PCM and FM/FM multiplex. The various sources of distortion in FM systems are discussed and graphs are given to show typical saddle points by means of which parameters can be selected to maximize the channel signal-to-noise ratios. Empirical methods are presented that give a systematic approach to the selection of the parameters that maximize the amount of data transmitted over a telemetry link of a given bandwidth or minimize the bandwidth for a given amount of transmitted data.

Description of the Telemetry System

THIS PAPER DESCRIBES the development of the airborne telemetry equipment, ground station receiving equipment, and data recovery equipment portions of a wideband telemetry subsystem containing a combination PCM and FM/FM multiplex.† The emphasis of this paper is on the r-f equipment parameters and parameter optimization of the r-f portion of the telemetry system. The various sources of distortion in FM systems are discussed and graphs are given to show typical saddle points by which the selection of parameters can be obtained to maximize the channel signal-to-noise ratios.

Figure 1 is a functional block diagram of the air-to-surface telemetry link in its simplest form. The transducers, signal conditioners, and multiplexing equipment sense and prepare the data for transmission. The composite baseband frequency modulates the carrier in the SHF transmitter. On the surface, the high-gain receiving antenna directs its main beam at the airborne vehicle by virtue of its self-acquisition and tracking capabilities. The amplifier–frequency converter consists of a low-noise parametric amplifier, down converter,

† Electro-Mechanical Research, Inc., of Sarasota, Florida, was responsible for the airborne multiplex and ground multiplex equipments while RCA developed the airborne transmitter and ground r-f receiving equipment.

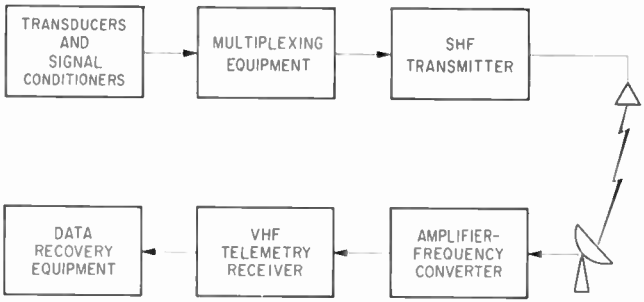


Fig. 1—Block diagram showing telemetry-system equipment.

and preamplifier, where the signal is translated to the VHF band and is fed to the VHF telemetry receiver. The output of this receiver is essentially the same as the input to the airborne SHF transmitter. The data recovery equipment contains PCM decoders and subcarrier discriminators which demodulate each of the FM subcarriers to obtain the data which is then processed for recording and display.

Figure 2 shows the 400-kc baseband configuration which is transmitted. It consists of PCM data at 144,000 bits per second, which is filtered at 72 kc or one half the bit rate, a 3-kc FM voice channel, and 35 FM analog-data channels, whose modulation varies from 100 cps to 2 kc. The height of each subcarrier relates to the deviation allotted to each subcarrier. To transmit the data using a single transmission carrier, all the data channels must form a single composite signal. A frequency translation method is used to combine the FM data, and the filtered PCM and translated FM signals are then mixed to form the composite video signal.

The primary objective in the development of the system was to transmit a large amount of data over a maximum range. Considerable design effort was spent maximizing the transmitter power and antenna gains and minimizing the equipment losses and the receiving system

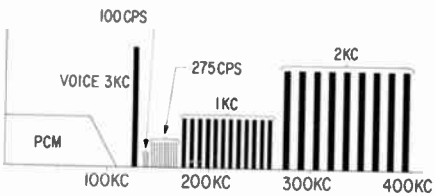


Fig. 2—Baseband configuration.

noise figure. A parametric amplifier was developed at the SHF frequency in order to obtain a low noise figure. The effect of receiver bandwidth on telemetry data is quite complex. For a bandwidth which is wider than the data spectrum, the telemetry link is thermal noise limited. As the receiving bandwidth is decreased, the thermal noise decreases and the nonlinear noise due to i-f distortion becomes more and more important until, for small bandwidths, it is the limiting factor and the total noise increases for further decreases in bandwidth. A saddle point is reached in varying the bandwidth where the signal-to-noise ratio in the data channels is maximized. If a receiving bandwidth is then selected which is near this saddle point, the system may be further optimized by varying the SHF carrier deviation with proportional changes in the individual subcarrier deviations.

Analysis of Telemetry System Parameters

One of the main technical requirements in the design of a telemetry system is the establishment of a satisfactorily low level of noise in the output of the various channels. This noise may be separated into thermal and nonlinear noise components. Thermal noise is related to those factors associated with transmission loss, while nonlinear noise, which arises from distortion of the multichannel signal, is related to the capabilities of the system. The over-all performance can, therefore, be expressed as a function of system parameters that utilizes these relations. Thus, it is possible to determine those conditions for which the total noise is a minimum.

The analysis of this telemetry system is based on the number of subcarriers and their bandwidths comprising a baseband signal that is used to modulate the carrier. The decision of FM modulation was based mainly on compatibility with present techniques and system equipments. More-sophisticated modulation techniques that reduce the baseband requirements and provide somewhat better signal-to-noise ratios do not favor simplicity as much as FM.

In telemetry systems, the thermal noise limits the performance under weak received signal conditions. When the signal is strong, the nonlinear noise, sometimes called the intermodulation noise, becomes the limiting factor. It is therefore necessary to control the intermodulation distortion due to the system components by careful equipment design and that due to the medium by appropriate choice of system parameters.

The distribution of the nonlinear noise that results from the transmission of a complex wave through a nonlinear device has been

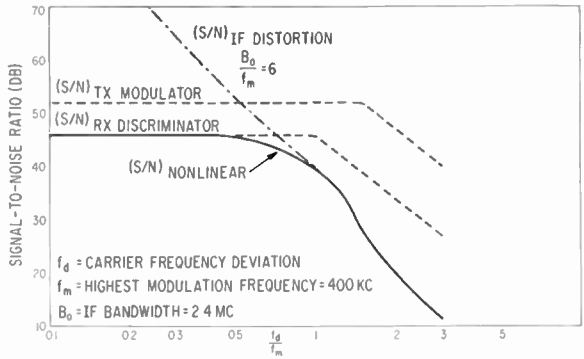


Fig. 5—Signal-to-noise ratio in top FM channel due to nonlinear noise.

tortion characteristics; a series of one-tone, two-tone, and noise-loading tests were conducted on the receiver and other r-f portions of the system. Theoretically, any of the three tests could be used for a complete system as well as its components, but from a practical point of view, the noise loading test is more appropriate for system evaluation as it simulates approximately the actual operating condition, while the two-tone test is more suitable for subsystem tests, as it is more meaningful to the equipment designer than the noise-loading test, and requires less elaborate testing equipment.

Figure 5 shows the signal-to-nonlinear-noise ratio in the top channel due to all nonlinear noise for a selected bandwidth. For low deviation, the discriminator is the limiting factor. Above $f_d/f_m = 1$, the prevalent noise is the phase distortion in the final i-f.

Figure 6 shows the total effect of all noise both at threshold and

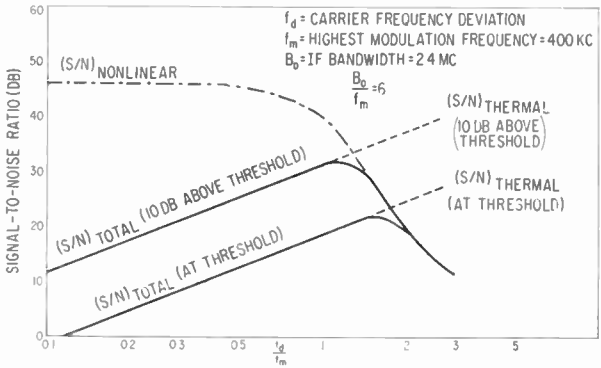


Fig. 6—Signal-to-noise ratio in top channel due to thermal and nonlinear noise.

at 20 db above threshold on the top-channel signal-to-noise ratio. For small deviations, the output is thermal-noise limited, and the S/N will increase linearly with channel deviation. Beyond a certain point, the nonlinear noise becomes prevalent and it becomes worse as the deviation increases. Saddle-points are reached where the channel performance is maximized with regard to signal-to-noise. The carrier deviation is usually adjusted below the saddle-point for the threshold curve in order to obtain adequate channel improvement for signal strengths above threshold. For this telemetry system, an f_d/f_m of about 1.5 was selected.

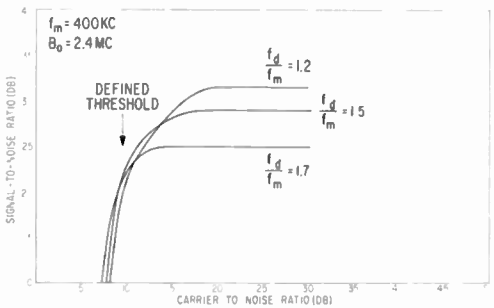


Fig. 7—Signal-to-noise ratio in top FM channel due to thermal and nonlinear noise.

Figure 7 is a plot of signal-to-noise in top channel versus carrier-to-noise input to the receiving system for f_d/f_m deviations of 1.2, 1.5, and 1.7. This graph shows that an f_d/f_m of 1.5 is nearly optimum for the selected i-f bandwidth. Similar curves could be drawn for other i-f bandwidths to obtain optimum deviations. However, for this telemetry system, the optimum parameters calculated for the 400-kc baseband were a 2.4-mc i-f bandwidth and a carrier deviation of ± 600 kc.

System Tests

Due to the complexity of the telemetry equipment design, several development tests were performed at intervals during the progress of the design and construction of the equipment.

An over-all system development test of the RCA and EMR equipment was performed as soon as engineering models were available to determine the system equipment compatibility, to isolate design limitations, to insure interface control, and to verify the accuracy requirements.

Several data-bandwidth configurations were employed during the system test, but the nominal 400-kc baseband, composed of the PCM signal and 36 FM subcarriers, was considered to be of primary importance.

The initial system tests consisted of transmitter-deviation sensitivity measurements and receiver-system noise figure. These tests were made to insure satisfactory operation of the equipment and to provide data to be used as a baseline for evaluation of the complete system.

To optimize the individual subcarrier deviation schedule and the total carrier deviation and to determine the system threshold, it was necessary to conduct a series of tests in which the optimum deviation was obtained by an iteration type method.

The purpose of these tests was to determine the unique subcarrier deviation schedule which allowed all information channels to "threshold" at the same receiver-system input signal level. Threshold was determined for the PCM channel as a 10^{-3} bit error rate. The method of setting the deviation schedule was to set the SHF signal level at the predicted threshold level for the particular bandwidth and to adjust the various baseband signal levels until the PCM indicated an error rate of 10^{-3} and all the FM subcarriers were at threshold. Threshold for the FM subcarriers was determined as the condition where occasional full-bandwidth spikes appeared in the demodulated signal when a triangular wave was passed through the channel. A random word train was employed for the PCM simulation, and all FM channels other than the one being modulated by the triangular waveform were modulated from the random noise generators. Visual observation was a sufficiently accurate means because of the very sharp threshold characteristic of a double FM system.

A series of such tests was conducted for various SHF carrier levels and carrier deviations to establish the proper deviation schedule. The deviation which gave adequate data at the lowest SHF carrier level and also gave good improvement above this level was considered optimum. A probability analyzer was used to measure the level of the carrier deviation. The carrier deviation was defined as the positive and negative peak levels which are exceeded only 0.2% of the time.

Figure 8 shows the output signal-to-noise ratio versus SHF carrier level for two of the channels. The top curve is for a channel which was located near the middle of the modulation spectrum and had a subcarrier modulation index of five. The signal-to-noise requirement on this channel was 30 db. It shows little improvement for carrier levels above threshold due to the presence of nonlinear noise. The bottom curve is

for a channel which was located near the top of the modulation spectrum and had a subcarrier modulator index of two. The signal-to-noise requirement for this channel was 15 db. The effect of variation of carrier deviation can be seen for this channel. Similar curves were obtained for many other channels in the optimization process, but this channel illustrates the effect. The results agree very closely with the analytical calculations on the system. When the carrier deviation ratio was decreased from 1.5 to 1.37, the output signal-to-noise ratio decreased 1 db at threshold. When the deviation increased to 650 kc or f_d/f_m to 1.62, the signal-to-noise ratio again decreased at threshold due

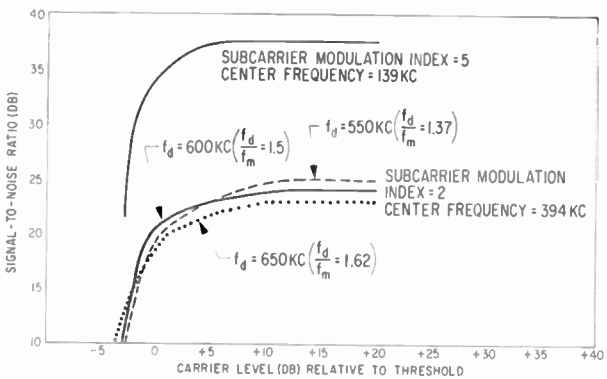


Fig. 8—Experimental signal-to-noise ratio of FM channels versus carrier level.

to the added nonlinear noise. Thus, the choice of carrier deviation of 600 kc appeared to be optimum.

The effect of the nonlinear noise on the PCM channel was quite evident. Near threshold the error remained relatively constant as the carrier deviation was varied, but above threshold the error rate increased as the deviation increased. For the final configuration of the deviation schedule, the error rate was 3×10^{-4} at threshold and decreased to 5×10^{-6} at 4 db above threshold, but no further improvement occurred as the carrier level was increased further.

The PCM channel required a very high deviation compared to the other channels. Measurements of the output baseband made with a frequency selective voltmeter indicated that intermodulation products of the FM subcarriers fell in the PCM passband. When only a few subcarriers were present the required deviation was close to that which would be calculated by standard FM/FM equations to give the measured signal-to-noise ratios. Then, as more and more subcarriers were

added, additional deviation was required for the PCM channel. Analysis of the data showed that the intermodulation present could be explained by the nonlinear noise analysis which was made previous to the system test.

Conclusions

It may be concluded that the design of the telemetry system from the "paper stage" through the experimental system tests that were conducted met the required objective, which was to transmit a large amount of data for a maximum range. Most telemetry systems are designed with bandwidths that are wider than necessary in order to obtain very low nonlinear noise. However, this approach is wasteful of bandwidth and does not utilize the full capacity of the system. The empirical methods presented in this paper give a systematic approach to the selection of parameters that maximize the amount of data transmitted over a telemetry link of a given bandwidth or minimize the bandwidth for a given amount of data.

ON A PROBLEM IN SINGLE-SIDEBAND COMMUNICATIONS

By

Jacques Dutka†

Summary—*A problem in the design of a single-sideband modulation system is considered which reduces to the determination of the distribution function of the sum of an infinite series of independent random variables. An approximation is obtained in the form of a distribution of a finite sum of independent uniform (but not necessarily equidistributed) random variables which is useful here and for broader classes of problems. Five-decimal-place tables of the original distribution function are computed, and some analytic properties of the distribution function are obtained.*

Physical Problem Description

BECAUSE of the constantly increasing demands for the limited available radio spectrum, it has been suggested that pulse code modulation (PCM) signals be transmitted using single-sideband (SSB) instead of double-sideband (DSB) FM to save bandwidth. In general, an SSB signal is of the form

$$s(t) = a(t) \cos \omega_0 t - b(t) \sin \omega_0 t$$

where $s(t)$ and $a(t)$ denote the transmitted and modulating signals, respectively, $b(t)$ denotes the Hilbert transform of $a(t)$ and ω_0 is the carrier frequency. Now $a(t)$ and $b(t)$ are uncorrelated, and it may sometimes happen that the amplitude of $b(t)$ is large relative to that of $a(t)$. The SSB equipment must be designed to accommodate such cases. Thus a question with important consequences for design is "What is the probability distribution for the amplitude of $b(t)$?"

The question can be put more precisely as follows: The modulating signal $a(t)$ is assumed to be constructed by choosing a convenient fundamental pulse form $f(t)$, which is transmitted at regular intervals, T, with either a positive or negative polarity. For each pulse, the polarity is chosen at random, independently of the polarities of other pulses. Thus the basic modulating signal has the form

† RCA Communications Systems Division, New York, N. Y., and Columbia University.

$$a(t) = \sum_{n=-\infty}^{+\infty} c_n f(t - nT),$$

where $\{c_n\}$ is a sequence of independent identically distributed random variables and c_n assumes the values ± 1, each with probability 1/2. The corresponding Hilbert transform is (formally)

$$b(t) = \sum_{n=-\infty}^{+\infty} c_n g(t - nT),$$

where $g(t)$ is the transform of $f(t)$. Now for accurate data transmission $f(t)$ is selected such that it has a negligible amplitude outside the interval $|t| < T/2$. But the transform pulse $g(t)$ has a considerable amplitude outside this interval, and it follows that the amplitude of $b(t)$ is affected by many pulses. For a number of convenient choices of $f(t)$ such as the rectangular pulse, the gaussian pulse, and other symmetric forms, it can be shown, that for t large, $g(t)$, to a first approximation, has the form K/t where K is a constant. (When $t = 0$, $g(t) = 0$.)

For instance, if $f(t)$ is a unit rectangular pulse centered at the origin, its Hilbert transform is

$$g(t) = \frac{1}{\pi} \ln \left| \frac{t + \frac{T}{2}}{t - \frac{T}{2}} \right|,$$

which for large t is asymptotic to $T/(\pi t)$. At a time corresponding to $t = T$, $g(T) = \pi^{-1} \ln |3| = 1.099\pi^{-1}$ is approximated by π^{-1} with a relative error of about 10%. Similarly, $g(2T) = \pi^{-1} \ln |5/3| = 0.511\pi^{-1}$, $g(3T) = \pi^{-1} \ln |7/5| = 0.336\pi^{-1}$ are approximated by $0.5\pi^{-1}$ and $0.333\pi^{-1}$, respectively, with relative errors of about 2% and 1%. The relative errors of later pulse magnitudes are negligible. For many engineering purposes this degree of approximation is sufficiently accurate.

Thus at time $t = 0$, corresponding to a pulse center, the function $b(t)$ may be approximated by the random variable

$$b(0) = \sum_{n=-\infty}^{+\infty'} c_n \frac{K}{nT},$$

where the prime denotes that the term corresponding to $n=0$ in the summation is omitted. The constant K is chosen so that the r-m-s values of $b(0)$ and $b(t)$ are equal. Thus the r-m-s value of $b(0)$ is

$$b(0)_{\text{r-m-s}} = \frac{K}{T}\sqrt{2\sum_{n=1}^{\infty}\frac{1}{n^2}} = \frac{\pi K}{T\sqrt{3}},$$

since, as is well known,

$$\sum_{n=1}^{\infty} n^{-2} = \frac{\pi^2}{6}.$$

For the engineering design problem which has been described, the problem reduces to finding the distribution function of

$$Y = \sum_{n=1}^{\infty}\frac{c_n}{n}, \tag{1}$$

where the sequence of random variables $\{c_n\}$ is defined above. More precisely, the design problem actually requires the determination of the distribution of the sum of two independent random variables each of which is distributed as Y in Equation (1). This is essentially the approximate distribution of the SSB quadrature voltage which occurs at the center of pulse intervals in a pulse train.

Of greater practical interest is the distribution of the SSB envelope. This distribution is readily obtained, in principle, by a simple transformation, using the variable

$$e(0) = \sqrt{a^2(0) + b^2(0)},$$

where $e(0)$ is the magnitude of the envelope of the SSB signal at pulse centers. Because the quadrature voltage can have large amplitudes (although rarely), the SSB power amplifier must occasionally clip the peaks. This clipping will inevitably cause some "sideband splatter." Using the distribution function for the envelope, a designer can estimate the amount of clipping that will be tolerable, and thereby determine the voltage range over which the SSB power amplifier must maintain linear amplification.

If the basic pulse shape $a(t)$ has a negligible amplitude outside the interval $-T/2 < t < T/2$, then at pulse centers $a^2(0)$ is a constant. The probability that the normalized envelope $e(0)/a(0)$ will

exceed any given value, y, is

$$P\left[\frac{e(0)}{a(0)} > y\right] = P\left[\frac{b(0)}{a(0)} > \sqrt{y^2 - 1}\right].$$

From this relation and the distribution of $b(0)$ which is derived below, the following numerical results can be obtained:

y	$P\left[\frac{e(0)}{a(0)} > y\right]$
1.00	1.0
1.65	0.1
2.40	0.01
2.88	0.001
3.20	0.0001
3.47	0.00001

From this it follows, for example, that if the SSB equipment is designed for linear amplification up to 2.4 times the peak amplitude of the inphase component, then clipping will occur on only one in 100 pulses. (The problem of deciding on the tolerable clipping level is a special study in itself and will not be discussed here. However, in any such study, the statistical distribution of the signal is an essential starting point.)

Another application for the quadrature distribution derived below is in the study of bit error probability. To achieve the theoretical advantage of SSB pulse transmission, it is necessary to use coherent detection at the receiver. That is, the receiver must obtain a phase-locked replica of the transmitted carrier. When the demodulating carrier has a small phase error, α, the detected signal voltage in the absence of noise is

$$r(t) = a(t) \cos \alpha + b(t) \sin \alpha.$$

The term $b(t) \sin \alpha$ is a disturbing voltage which interferes with bit detection. Evidently, the probability distribution of both $b(t)$ and α must be known in order to calculate the bit error rate.

A discussion of the error rate in such systems is given by Gardner and Rabinowitz (see Ref. (5) in the bibliography).

Mathematical Formulation

On examination of the series of Equation (1), it is found that for some distributions of values of the sequence $\{c_n\}$ the series converges, but for other distributions of values, the series diverges. Now series of this form have been investigated in the mathematical literature, and it may be shown that if $\sum_{n=1}^{\infty} (c_n/n)^2$ converges then $\sum_{n=1}^{\infty} c_n/n$ converges with probability one. (See Ref. (2), Alexits, p. 53.)

An expression for the distribution function of Y will be obtained by application of the method of characteristic functions. (See bibliography, Ref. (3), Cramér, Ch. 10.) Let $X_n = c_n/n$. Then the characteristic function of X_n is $\mathcal{E}\,[\exp\{itX_n\}] = \cos(t/n)$. The characteristic function of $Y_N = \sum_{n=1}^{N} X_n$ is $\mu(t) = \prod_{m=1}^{N} \cos(t/m)$. The characteristic function of $Y = \lim_{N\to\infty} Y_N$ is therefore given (formally) by

$$\mathcal{E}[\exp\{itY\}] = \mu(t) = \prod_{m=1}^{\infty} \cos\frac{t}{m}. \tag{2}$$

(The existence of the distribution function (and the characteristic function) of Y is a consequence of the convergence of $\sum_{n=1}^{\infty} (c_n/n)^2$. (See Ref. (6), Wintner, Ch. 12.)

The distribution function of Y can, in principle, now be obtained by calculation of the Fourier transform of the characteristic function, but it is convenient to digress briefly at this point.

The general characteristics of the distribution function of Y have been determined by A. Wintner and his collaborators by an analysis of the properties of the characteristic function $\mu(t)$ above. (See Ref. (6), Wintner, p. 38 and references to a series of papers there.) They proved that the distribution function of Y is analytic on the entire real axis, $-\infty < y < \infty$, and is convex. That is, the symmetric density function of Y is a nonincreasing function of $|x|$.

The distribution function of Y represents an infinite convolution of discrete independent random variables $\{X_n\}$. It is convenient, for purposes of calculation below, to show that this distribution function can also be regarded as the infinite convolution of continuous independent random variables. Now from the infinite product formula for $\cos(t/m)$, it is seen that the characteristic function of Y is

$$\mu(t) = \prod_{m=1}^{\infty} \cos\left(\frac{t}{m}\right) = \prod_{m=1}^{\infty} \prod_{n=1}^{\infty} \left[1 - \frac{4t^2}{(2n-1)^2\pi^2 m^2}\right]$$
$$= \prod_{n=1}^{\infty} \prod_{m=1}^{\infty} \left[1 - \frac{4t^2}{(2n-1)^2\pi^2 m^2}\right],$$

where the inversion of the order of the double product is justified because of the absolute convergence of the double series

$$\sum_{m,\,n=1}^{\infty} \frac{4t^2}{(2n-1)^2\,\pi^2\,m^2}.$$

Also, since

$$\sin u = u \prod_{n=1}^{\infty} \left[1 - \left(\frac{u}{n\pi}\right)^2\right],$$

it follows on substitution above that

$$\mu(t) = \prod_{j=1}^{\infty} \frac{\sin a_j t}{a_j t}, \tag{3}$$

where $a_j = 2/(2j-1)$. Now $(\sin a_j)/(a_j t)$ is the characteristic function of a random variable W_j which is uniformly distributed on the interval $(-a_j, a_j)$. Thus $\mu(t)$ is the characteristic function corresponding to an infinite convolution of independent uniformly distributed random variables $\{W_j\}$ and $Y = \sum_{j=1}^{\infty} W_j$. It is this result which enables a convenient formula for the distribution function of Y to be obtained.

On applying the inversion formula for characteristic functions, it is readily seen that $H(y)$, the distribution function of Y, is given by

$$H(y) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin ty}{t}\, \mu(t)\, dt$$
$$= \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin ty}{t} \left(\prod_{j=1}^{\infty} \frac{\sin a_j t}{a_j t}\right) dt, \tag{4}$$

where $a_j = 2/(2j - 1)$, $j = 1, 2, 3 \ldots$. In principle, therefore, the original problem stated in the first section can be solved if this integral can be evaluated.

Numerical Solution

It does not appear possible to obtain a closed-form evaluation of the integral above; therefore recourse must be had to approximate methods. Now it can be shown that as $n \to \infty$, $H_n(y)$, the distribution function of $\sum_{j=1}^{n} W_j$, converges uniformly for all y to $H(y)$. Moreover, as is proved in the Appendix, an explicit expression for $H_n(y)$ can be obtained in the form

$$H(y) = \frac{1}{(2^n a_1 a_2 \ldots a_n)\, n\,!} \left[(z)_+^n - \sum_{j}^{n} (z - 2a_j)_+^n + \sum_{j>k}^{n} (z - 2a_j - 2a_k)_+^n - \sum_{j>k>1}^{n} (z - 2a_j - 2a_k - 2a_1)_+^n + \ldots + (-1)^n (z - 2a_1 - 2a_2 - \ldots - 2a_n)_+^n \right] dt \tag{5}$$

where $z = a_1 + a_2 + \ldots + a_n + y$, and x_+ denotes the ramp function. That is $x_+ = x$ for $x \geq 0$ and $x_+ = 0$ for $x < 0$. Thus $H(y)$ can, in theory, be found as accurately as desired by choosing n sufficiently large and evaluating Equation (5). Unfortunately, however, this involves the calculation of 2^n terms on the right in the equation above, so that the procedure is not practical if n is even tolerably large. (It was found that by choosing $n = 8$, an approximation to $H(y)$ was obtained which was valid to about three decimal places.) Another approach is to evaluate the integral formula

$$H_n(y) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin ty}{t} \left(\prod_{j=1}^{n} \frac{\sin a_j t}{a_j t} \right) dt \tag{6}$$

numerically for a suitable value of n and to use this as an approximation to $H(y)$. Under the direction of M. S. Corrington, $H_n(y)$ was calculated on a large scale computer for $n = 1000$ and is summarized in the table given below. A check for particular values of y was made by M. Landis, who employed the approximate formula

$$H(y) \doteq \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} H_8(x) \exp\left\{-\frac{(y-x)^2}{2\sigma^2}\right\} dx, \tag{7}$$

where $\sigma^2 = \sum_{j=9}^{\infty} a_j^2/3$. It is believed that the results given here are the correct values of $H(y)$ rounded off to five decimal places. In addition, the convolution of two Y variables, (which is necessary for the computation of the distribution of $b(0)$ in the first section), which is given by

$$I(y) = \int_{-\infty}^{+\infty} H(y-x)\, dH(x), \tag{8}$$

was calculated and checked and is believed to be the correct values of $I(y)$ rounded off to five decimal places.

Table I—Distribution Functions $H(y)$ and $I(y)$

y	$H(y)$	$I(y)$
0	.50000	.50000
0.1	.52500	.52102
0.2	.55000	.54199
0.3	.57499	.56283
0.4	.59998	.58349
0.5	.62495	.60391
0.6	.64988	.62404
0.7	.67475	.64383
0.8	.69951	.66324
0.9	.72408	.68222
1.0	.74838	.70074
1.1	.77228	.71876
1.2	.79564	.73627
1.3	.81830	.75323
1.4	.84009	.76963
1.5	.86082	.78545
1.6	.88032	.80068
1.7	.89844	.81531
1.8	.91505	.82934
1.9	.93006	.84276
2.0	.94340	.85556
2.1	.95506	.86775
2.2	.96505	.87931
2.3	.97344	.89027
2.4	.98032	.90061
2.5	.98582	.91033
2.6	.99009	.91945
2.7	.99331	.92797

Table I—Distribution Functions $H(y)$ and $I(y)$—*(continued)*

y	$H(y)$	$I(y)$
2.8	.99564	.93590
2.9	.99728	.94324
3.0	.99838	.95001
3.1	.99908	.95622
3.2	.99951	.96188
3.3	.99975	.96702
3.4	.99988	.97166
3.5	.99995	.97581
3.6	.99998	.97951
3.7	.99999	.98277
3.8	1.00000	.98563
3.9	1.00000	.98812
4.0	1.00000	.99026
4.1		.99210
4.2		.99365
4.3		.99494
4.4		.99602
4.5		.99690
4.6		.99762
4.7		.99819
4.8		.99864
4.9		.99900
5.0		.99927
5.1		.99948
5.2		.99963
5.3		.99974
5.4		.99983
5.5		.99988
5.6		.99992
5.7		.99995
5.8		.99997
5.9		.99998
6.0		.99999
6.1		.99999
6.2		1.00000
6.3		1.00000
6.4		1.00000
6.5		1.00000

This table was computed under Signal Corps contract DA-36-039-SC-87240.

Further Properties of the Distribution Function

From Equation (3), it follows that

$$\log \mu(t) = \sum_{j=1}^{\infty} \log \frac{\sin a_j t}{a_j t},$$

and, on expanding the right-hand side by means of the standard ex-

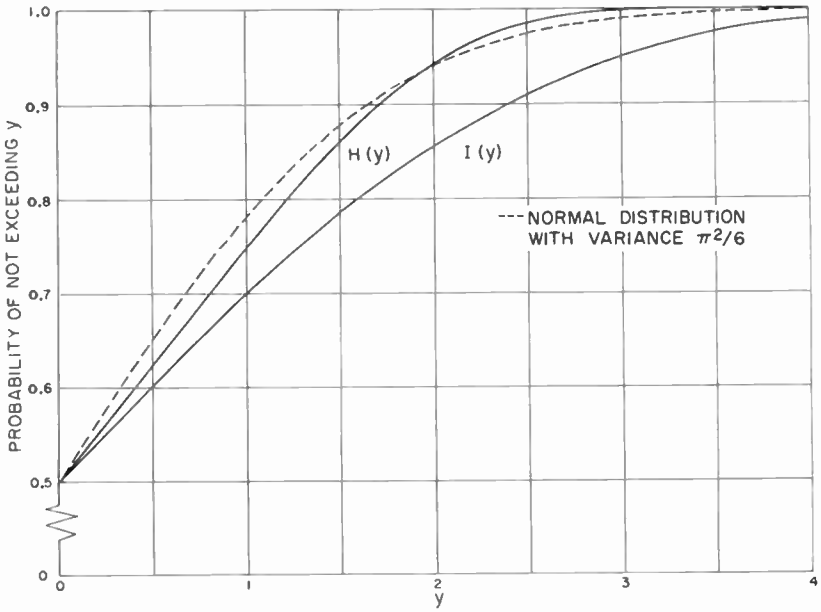


Fig. 1—Distribution functions.

pansion*

$$\log \frac{\sin u}{u} = -\sum_{n=1}^{\infty} \frac{2^{2n-1}}{n(2n)!} B_n u^{2n}, \qquad -\pi < u < \pi,$$

it is seen that

$$-\log \mu(t) = \sum_{j=1}^{\infty} \sum_{n=1}^{\infty} \frac{S_{2n}}{n} \left[\frac{2t}{\pi(2j-1)} \right]^{2n},$$

where $2^{2n-1}B_n/(2n)! = \pi^{-2n} \sum_{k=1}^{\infty} k^{-2n} = \pi^{-2n} S_{2n}$. Thus, since the terms in the double sum on the right are all positive, the order of summation may be interchanged, and it follows that

$$\log \mu(t) = -\sum_{n=1}^{\infty} \frac{S_{2n}\, \sigma_{2n}}{n} \left(\frac{2}{\pi} \right)^{2n} t^{2n}, \qquad |t| < \frac{\pi}{2}, \tag{9}$$

* See, e.g., Ref. (1), Smithsonian Tables, 6.43.1.

where $\sigma_{2n} = \sum_{k=1}^{\infty} (2k-1)^{-2n}$. The cumulants $\{\kappa_n\}$ of Y are defined as the coefficients of $(it)^n/n!$ in this expression. Thus we see that all odd-order cumulants of Y vanish, and the even-order cumulants are

$$\kappa_{2n} = (-1)^{n+1} \frac{S_{2n}\,\sigma_{2n}}{n} \left(\frac{2}{\pi}\right)^2 (2n)! \tag{10}$$

The moments of Y (which are the coefficients of $(it)^n/n!$ in the Maclaurin expansion of Equation (3) can be expressed in terms of the cumulants, and this is a particularly convenient method of calculation here. (See Ref. (3), Cramér, p. 186.) Many qualitative characteristics of the distribution function of Y can be obtained from a knowledge of the moments (or cumulants) as explained, for instance, in Cramér, Ch. 15, but this will not be considered further here.

The characteristic function associated with the right-hand side of Equation (7) is $\exp(-\sigma^2 t^2/2) \prod_{j=1}^{8} (\sin a_j t)/(a_j t)$. If we take the logarithm of this expression, we see that all odd-order cumulants vanish, the first three cumulants are identical with those of Y, and the higher-order even cumulants differ little from the corresponding cumulants of Y. It may therefore be expected that the right-hand side of Equation (7) will furnish a good approximation to $H(y)$.

It was shown in the second section that Y can be regarded as the sum of an infinite number of independent random variables, and it might be expected that Y is normally distributed. That this is actually not the case follows from the fact that the cumulants of order three and above for a normally distributed random variable all vanish, whereas from Equation (10) it follows that the even cumulants of Y, for all orders, do not vanish. Thus Y is not normally distributed.

What can be said regarding the behavior of $H(y)$ for large values of $|y|$? The best result which has been found is that of Wintner who showed that $H(y)$ approaches its limits as $|y| \to \infty$ at least as rapidly as any normal distribution function. More precisely, Wintner showed that

$$\int_{-\infty}^{+\infty} \exp\{\lambda y^2\}\, dH(y) < \infty$$

for $\lambda > 0$ sufficiently small and that for every $c > 0$

$$1 - H(y) = O(\exp\{-cy^2\}) \text{ for } y \to \infty,$$

$$H(y) = O(\exp\{-cy^2\}) \text{ for } y \to -\infty. \tag{11}$$

For further details, see Ref. (6), Wintner, p. 35.

ACKNOWLEDGMENT

The author wishes to express his appreciation to M. S. Corrington, R. Gardner and H. Staras for helpful discussion regarding various points in this paper. Mr. Gardner, in particular, is responsible for the statement of the communications problem and for various editorial contributions.

APPENDIX

Let $W_1, W_2, \ldots, W_n$ be a set of independent random variables where the probability density of W_j is

$$f_j(w) = \frac{1}{2a_j}, \quad |w_j| \leq a_j, \quad a_j > 0. \tag{12}$$

What is the distribution function $H_n(y)$ of the sum $\sum_{j=1}^{n} W_j$? The procedure given here follows one developed by the author in Ref. (4).

Let $h_n(y)$ denote the probability density corresponding to $H_n(y)$. Then evidently the probability density $h_{k+1}(y)$ is the convolution of $h_k(w)$, the probability density of $\sum_{j=1}^{k} W_j$, with $f_{k+1}(w)$, the probability density of W_{k+1}. Thus the recursive convolution formulas

$$h_{k+1}(y) = \int_{-\infty}^{+\infty} h_k(y-w) f_{k+1}(w)\, dw, \quad k = 1, 2, \ldots, n-1, \tag{13}$$

where $h_1(w) = f_1(w)$, are obtained. The required distribution function $H_n(y)$ then follows on integrating $h_n(y)$.

Before proceeding with the evaluation of $H_n(y)$, it is convenient to introduce the ramp function x_+. That is $x_+ = x$ for $x \geq 0$ and $x_+ = 0$ for $x < 0$. Thus, on integrating Equation (12) it is found

that $F_j(w)$, the distribution function of W_j, may be written as

$$F_j(w) = \frac{1}{2a_j}\left[(a_j + w)_+ - (a_j + w - 2a_j)_+\right]. \tag{14}$$

From Equation (13), it follows that

$$h_2(y) = \int_{-\infty}^{+\infty} f_1(y-w) f_2(w)\,dw = \frac{1}{2a_2}\int_{-a_2}^{a_2} f_1(y-w)\,dw$$

$$= \frac{1}{2a_2}\int_{y-a_2}^{y+a_2} f_1(u)\,du = \frac{1}{2a_2}\left[F_1(a_2+y) - F_1(a_2+y-2a_2)\right],$$

and, on substituting Equation (14),

$$h_2(y) = \frac{1}{2^2 a_1 a_2}\left[(a_1+a_2+y)_+ - (a_1+a_2+y-2a_1)_+ \right.$$
$$\left. - (a_1+a_2+y-2a_2)_+ + (a_1+a_2+y-2a_1-2a_2)_+\right].$$

To obtain $h_3(y)$, it follows from Equation (13) that

$$h_3(y) = \int_{-\infty}^{+\infty} h_2(y-w) f_3(w)\,dw = \frac{1}{2a_3}\int_{y-a_3}^{y+a_3} h_2(u)\,du,$$

and, on integrating,

$$h_3(y) = \frac{1}{2^3 a_1 a_2 a_3 . 2!}\left[(a_1+a_2+a_3+y)_+^2 - \sum_{j=1}^{3}(a_1+a_2+a_3+y-2a_j)_+^2 + \right.$$
$$\left. \sum_{j>k}^{3}(a_1+a_2+a_3+y-2a_j-2a_k)_+^2 - (a_1+a_2+a_3+y-2a_1-2a_2-2a_3)_+^2\right].$$

By mathematical induction it follows that if

$$z = a_1 + a_2 + \ldots a_n + y,$$

then

$$h_n(y) = \frac{1}{2^n a_1 a_2 \ldots a_n (n-1)!} \left[(z)_+^{n-1} - \sum_j^n (z - 2a_j)_+^{n-1} + \sum_{j>k}^n (z - 2a_j - 2a_k)_+^{n-1} - \sum_{j>k>l}^n (z - 2a_j - 2a_k - 2a_l)_+^{n-1} + \ldots + (-1)^n (z - 2a_1 - 2a_2 - \ldots - 2a_n)_+^{n-1} \right], \quad (15)$$

where the variables j, k, l, . . . in the summations are taken over all possible selections of integers from 1 to n which are consistent with $j > k > l > \ldots$. If $h_n(y)$ is integrated, it is seen that the distribution function of $\sum_{j=1}^n W_j$ is

$$H_n(y) = \frac{1}{2^n a_1 a_2 \ldots a_n n!} \left[(z)_+^n - \sum_j^n (z - 2a_j)_+^n + \sum_{j>1}^n (z - 2a_j - 2a_k)_+^n - \sum_{j>k>l}^n (z - 2a_j - 2a_k - 2a_l)_+^n + \ldots + (-1)^n (z - 2a_1 - 2a_2 - \ldots - 2a_n)_+^n \right]. \quad (16)$$

The terms on the right in Equations (15) and (16) are either polynomials in y of the $(n-1)$th and nth degrees, respectively, or they are equal to zero. The graphs of $h_n(y)$ and $H_n(y)$ are thus composed of a series of osculating polynomial arcs.

$H_n(y)$ has an interesting geometrical interpretation; it represents a ratio whose denominator is the content of the n-dimensional rectangular parallelopiped R: $|w_1| \leq a_1$, $|w_2| \leq a_2$, . . . , $|w_n| \leq a_n$, and whose numerator is the content of the region R_y bounded by the half-space $x_1 + x_2 + \ldots + x_n \leq y$ and the parallelopiped R.

If n is large, it is not possible as a practical matter to evaluate Equation (16) exactly, and recourse must be had to approximations. For example, the following result is a consequence of a form of the central limit theorem due to J. L. Lindeberg:

> Let $\sigma = \max(\sigma_1, \sigma_2, \ldots \sigma_n)$, where $\sigma_j^2 = a_j^2/3$, the variance of W_j, and let $s_n^2 = \sum_{j=1}^n a_j^2/3$, the variance of $\sum_{j=1}^n W_j$. If as $n \to \infty$ the ratio $\sigma/s_n \to 0$, then

$$H_n(s_n y) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left\{-\frac{w^2}{2}\right\} du$$

uniformly for all y.

That is, if $\sigma/s_n \to 0$ as $n \to \infty$, then $\sum_{j=1}^{n} W_j$ is asymptotically normally distributed.

Bibliography

1. E. P. Adams, *Smithsonian Mathematical Formulae and Tables of Elliptic Functions*, Washington, 1947.

2. G. Alexits, *Konvergenz probleme der Orthogonalreihen*, Deutsche Verlag der Wissenschaften, Berlin, 1960.

3. H. Cramér, *Mathematical Methods of Statistics*, Princeton, 1946.

4. J. Dutka, "The Distribution of a Sum of Independent Uniformly Distributed Errors," Norden Laboratories Corp., Report 077 D0036, July 15, 1954.

5. R. Gardner and J. Rabinowitz, "An Analysis of PCM Transmission via Single Sideband," 1 May 1963, Signal Corps Contract No. DA-36-039-SC-87240.

6. A. Wintner, *Asymptotic Distributions and Infinite Convolutions*, The Institute for Advanced Study, Princeton, 1938.

DISCUSSION AND APPLICATIONS OF ELECTROSTATIC SIGNAL RECORDING

By

Irwin M. Krittman and John A. Inslee

RCA Astro-Electronics Division
Princeton, N. J.

Summary—A proposed electrostatic signal recorder incorporating principles and techniques underlying the photodielectric tape camera is described. General expressions for the recorder signal-to-noise ratio and packing density are derived. Experimental results of a study of high-resolution camera tubes are used to predict typical recorder characteristics. Possible applications—namely a tape loop, wide-band recording, and unequal recording and playback rates—are also discussed.

Introduction

THE LIMITATIONS of magnetic signal recording have prompted investigations of other recording techniques. One of these, electrostatic signal recording, entails the deposition (recording) and detection (playback) of fine charge patterns by electron beams. Recent research and development efforts have produced advances in electron-beam technology.[1,2] The possible application of improved electron guns to electrostatic recording has led to new studies of the technique.

General Description

In electrostatic signal recording, recording may be accomplished by depositing charges on the surface of an insulating target from an electron beam modulated by the input signal. Playback can be accomplished by retracing the recorded area with an unmodulated electron beam.

A schematic representation of a proposed electrostatic (tape) signal recorder is shown in Figure 1. Because it would employ electron guns, the device is shown housed in a vacuum enclosure. The recording medium consists of a flexible base on which insulating and conducting films are deposited. At present, a 70-mm tape can be produced with polystyrene insulator, copper–gold conducting backplate, and Cronar (duPont Mylar) base.

---

[1] E. C. Hutter, J. A. Inslee and T. H. Moore, "Electrostatic Imaging and Recording," *Jour. S.M.P.T.E.*, Vol. 69, No. 1, Jan. 1960.

[2] Contract AF33(657)-7939: Applied Research on High Resolution Camera Tubes.

Recording Process

Prior to recording, the tape passes before an electron flood beam; old information is "erased" and the insulating surface is brought to equipotential. During the recording process, a fine, modulated electron beam deposits charges along parallel, transverse tracks on the tape in accordance with the input data. High-velocity-beam operation (i.e., operation resulting in a greater-than-unity secondary-emission ratio of the target surface) produces more-positive voltages on those areas of the tape struck by more primary electrons.

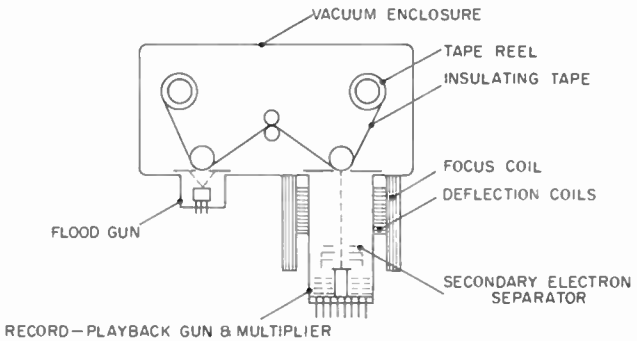


Fig. 1—Schematic representation of proposed electrostatic signal recorder.

Playback Process

During playback, impinging high-velocity electrons from a fine, unmodulated primary beam liberate secondary electrons from the tape. The voltage modulation of the tape surface produces a corresponding modulation of the originating potentials of the secondary electrons. The secondaries are removed from the target, brought back to a separating aperture near the record–playback gun, and separated according to their originating potentials. Higher-energy electrons are collected by an electron multiplier to form the output signal.

The playback process partially destroys the stored information, but several acceptable readouts may be achieved. Recorded information can be stored for many months* and can be easily erased. The storage tape is reusable through hundreds of record–playback cycles.

---

* The available storage time is determined by the resistivity of the tape insulator. For polystyrene targets, this time may be as great as one year.

Beam Tracking

The playback process employs mechanical and electrical servo systems to insure tape-speed control and beam-tracking accuracy, respectively. The mechanical servo, for example, may incorporate a magnetic strip on the reverse side and along one edge of the recording tape. A magnetic head would sense a pulse train recorded on the magnetic strip during the recording process. A comparison between this and a standard pulse train would generate an error signal for the servo. A servo with frequency response below 100 cps would maintain the tape speed constant to within less than 1 per cent r-m-s—sufficient for coarse correction.

Fine correction may be achieved with an electronic servo employing a return-beam technique. A high-frequency wobble perpendicular to the recorded tracks on the tape would be impressed upon the playback electron beam. This would produce amplitude modulation of the secondary electron beam. After amplification and separation, the resulting amplitude-modulated signals would be compared with the output from the wobble-frequency generator in a standard phase-comparison circuit. The magnitude and direction of the phase errors would indicate the magnitude and direction of the beam-positioning errors with respect to the centers of the recorded tracks. The amplitude of the impressed wobble need not be large to provide sufficient modulation. The wobble frequency would be slightly greater than the highest recorded signal frequency.

Signal-to-Noise Ratio

Expressions for the theoretical signal-to-noise ratio of the proposed electrostatic signal recorder can be derived with the aid of the current-flow diagrams and approximate secondary-electron separator characteristic, shown in Figures 2 and 3, respectively.[3]

Recording Process

The recording process peak signal to r-m-s noise ratio is

$$\left(\frac{S}{N}\right)_r = \frac{m r_r r_t (\delta - 1)}{\sqrt{\delta^2 - \delta + 1}} \sqrt{\frac{i_r}{2ef_r}}, \tag{1}$$

where e is electronic charge (1.6×10^{-19} coulomb),

[3] Figure 2 is patterned after similar diagrams and associated analyses employed by A. D. Cope and H. Borkan. See, for example, "Isocon Scan—A Low-Noise, Wide-Dynamic-Range Camera Tube Scanning Technique," *Applied Optics*, Vol. 2, No. 3, March 1963.

f_r is the recording bandwidth, in cycles/second,

i_{r0} is the minimum recording beam current, in amperes,

i_r is the maximum recording beam current, in amperes,

m is the ratio of $i_r - i_{r0}$ to i_r,†

r_r is the recording beam sine-wave (aperture) response at the recorded signal frequency,

r_t is the target sine-wave (aperture) response at the recorded signal frequency,

δ is the target surface secondary-emission ratio.

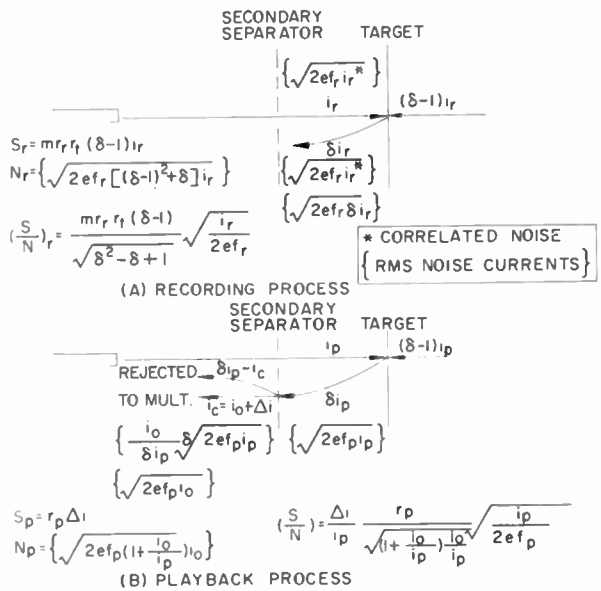


Fig. 2—Electrostatic signal recorder current-flow diagrams.

Playback Process

The playback process peak signal to r-m-s noise ratio is

$$\left(\frac{S}{N}\right)_p = \frac{r_p \Delta i/i_p}{\sqrt{\left(1+\dfrac{i_0}{i_p}\right)\dfrac{i_0}{i_p}}}\sqrt{\frac{i_p}{2ef_p}}, \qquad (2)$$

† If M is the recording beam modulation, then

$$M = \frac{i_r - i_{r0}}{i_r + i_{r0}} = \frac{m}{2-m}.$$

where f_p is the playback bandwidth, in cycles/second,

i_p is the playback beam current, in amperes,

i_0 is the separated return-beam current incident to the multiplier, originating from a background-level area on the target, in amperes,

Δi is the peak swing of the separated return-beam current above its background level, in amperes,

r_p is the playback beam sine-wave (aperture) response at the recorded signal frequency.

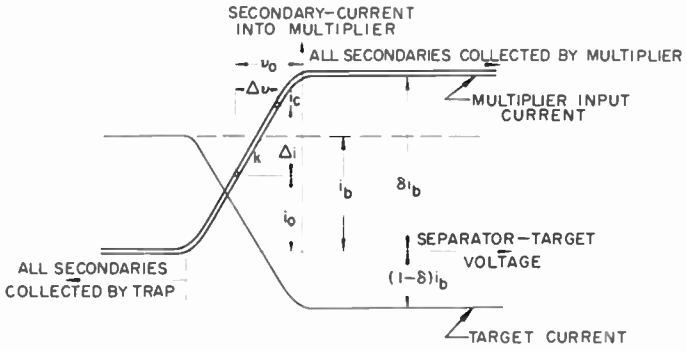


Fig. 3—Approximate secondary-electron separator characteristic (i_b = primary current and δ = target surface average secondary-emission ratio).

From the separator characteristic,

$$\Delta i = k\Delta v, \tag{3}$$

where k is the slope of the characteristic, in mhos, and

Δv is the peak swing of the target voltage above its background level, in volts.

But

$$\Delta v = \frac{mnr_r r_t (\delta - 1) i_r L^2}{2 f_r c}, \tag{4}$$

where c is the target capacitance, in farads/cm^2,

n is the fraction of the recording beam current landing on the bit area,

L is the horizontal density, in half-cycles/cm; thus L^{-2} is the area of a target bit, in cm^2.

Therefore

$$\left(\frac{S}{N}\right)_p = \frac{mnr_p r_r r_t(\delta-1)i_r L^2 k/i_p}{2f_r c\sqrt{\left(1+\dfrac{i_0}{i_p}\right)\dfrac{i_0}{i_p}}}\sqrt{\frac{i_p}{2ef_p}}. \tag{5}$$

From experimentally determined separator characteristics,*

$$i_0 \approx \frac{\delta i_p}{3}, \tag{6}$$

$$k \approx \frac{\delta i_p}{15}, \tag{7}$$

so that

$$\left(\frac{S}{N}\right)_p \approx \frac{0.1\, mnr_p r_r r_t \delta(\delta-1)i_r L^2}{f_r c\sqrt{\delta(\delta+3)}}\sqrt{\frac{i_p}{2ef_p}}. \tag{8}$$

General Output Expression

The peak signal to r-m-s noise ratio of the cascaded record–playback process can be expressed in terms of the individual process ratios;

$$\left(\frac{S}{N}\right)_c^{-2} = \left(\frac{S}{N}\right)_r^{-2} + \left(\frac{S}{N}\right)_p^{-2}. \tag{9}$$

The ratio $(S/N)_c$ is also the signal-to-noise ratio associated with the separated return-beam current entering the electron multiplier. If the effect of the multiplier on the recorder output signal-to-noise ratio is neglected, then $(S/N)_c$ is the output signal-to-noise ratio. The error introduced by this simplification is usually less than 10 per cent.

From Equations (1), (8), and (9),

$$\left(\frac{S}{N}\right)_c \approx \frac{0.1\, mnr_p r_r r_t(\delta-1)\sqrt{\delta/(\delta+3)}\,L^2/c}{\sqrt{1+\dfrac{n^2\delta(\delta^2-\delta+1)}{\delta+3}\left(\dfrac{r_p L^2}{10c}\right)^2\dfrac{i_r}{f_r}\dfrac{i_p}{f_p}}}\frac{i_r}{f_r}\sqrt{\frac{i_p}{2ef_p}}. \tag{10}$$

* The experimental characteristics were obtained for polystyrene surfaces under 300-volt primary-electron energy bombardment. It is assumed that all suitable insulators will display similar characteristics.

Output Expression for Polystyrene

Results of the electrostatic image and signal recording research efforts to date suggest the use of polystyrene as the recorder target insulator material. Storage tapes can be readily fabricated by depositing polystyrene on a suitable backplate from a glow discharge in styrene vapor.

For polystyrene,

$$\delta \approx 1.5 \text{ (at 300 volts)},$$

$$c \approx 2 \times 10^{-9}/t \qquad \text{farads/cm}^2,$$

where t is the storage-target insulator thickness, in microns.

Letting

$$F_p = f_p \times 10^{-6} \text{ megacycles/sec},$$

$$F_r = f_r \times 10^{-6} \text{ megacycles/sec},$$

$$I_p = i_p \times 10^{6} \text{ microamperes},$$

$$I_r = I_r \times 10^{6} \text{ microamperes},$$

$$m = 0.75,$$

$$n = 0.80,$$

the theoretical output signal-to-noise ratio of the proposed electrostatic signal recorder with polystyrene insulator is

$$\left(\frac{S}{N}\right)_c \approx \frac{0.153 r_p r_r r_t t L^2}{\sqrt{1 + \dfrac{(r_p t L^2)^2}{1.07 \times 10^9} \dfrac{I_r}{F_r} \dfrac{I_p}{F_p}}} \frac{I_r}{F_r} \sqrt{\frac{I_p}{F_p}}. \tag{11}$$

The values of m and n chosen are based on estimates of recorder performance.

Packing Density

The bit packing density of the proposed electrostatic signal recorder would be equal to the product of the horizontal and vertical (linear) densities. The ratio of the vertical density to the horizontal density, the packing factor, would be limited by crosstalk considerations.

If L is the horizontal density, in half-cycles/cm,

M is the vertical density, in half-cycles/cm,

P^2 is the packing density, in bits/cm^2,

ρ is the packing factor,

then
$$P^2 = LM = \rho L^2, \tag{12}$$

where
$$0 < \rho = M/L \leq 1. \tag{13}$$

The horizontal density is defined as that spatial frequency at which the overall or cascaded recorder sine-wave response, $r_p r_r r_t$, is 50 per cent. The packing density, therefore, would be determined by the target and beam responses.

Target Response

Studies of the detailed operation of electrostatic image and signal recording devices have revealed an effect which can limit their performance. The effect applies to image orthicons and vidicons (under transient conditions), photodielectric tape cameras, and electrostatic signal recorders.

In each of these storage devices, a playback electron beam senses the potential distribution rather than the recorded charge distribution on a target. In each case, the potential distribution—determined by the thickness of the storage target—is a degraded image of the charge distribution.

The transformation of a charge pattern into a potential pattern on the surface of a dielectric can be considered an imaging process. The theoretical sine-wave (Fourier) response of an aperture defined by this process may be computed from its line transmittance.[4]

For the electrostatic signal recorder, the line transmittance of the storage target may be determined from the voltage distribution at its surface due to a point charge on the surface.

The voltage at the point x,y due to q, in volts, is

$$V(x,y) \approx \frac{q}{4\pi\epsilon\sqrt{x^2 + y^2}} + \frac{-q}{4\pi\epsilon\sqrt{(2t)^2 + x^2 + y^2}}, \tag{14}$$

where q is the point charge at the center of the surface of an electrostatic signal recorder target, in microcoulombs,

t is the storage target insulator thickness, in microns,

[4] O. H. Schade, "Image Gradation, Graininess and Sharpness in Television and Motion-Picture Systems, Part IV, A & B: Image Analysis in Photographic and Television Systems (Definition and Sharpness)," *Jour. S.M.P.T.E.*, Vol. 64, Nov. 1955.

x,y are the coordinates of a point on the surface of the target as measured from the center of the surface, in microns,

ϵ is an equivalent permittivity, in farads/meter.

The second term in the expression for $V(x,y)$ is the voltage at the surface of the target due to the image charge, $-q$, simulating the insulator–conductor equipotential interface.

The line transmittance of the storage target, in volt-microns, is

$$f(x) = \lim_{\alpha\to\infty} \int_{-\alpha/2}^{\alpha/2} V(x,y)\, dy = \frac{q}{4\pi\epsilon} \ln\left[\frac{x^2 + (2t)^2}{x^2}\right], \qquad (15)$$

where α is the width of the storage target, in microns.

The sine-wave response of an aperture is a normalized Fourier transform of its line transmittance. The Fourier transform of the line transmittance, in volt-micron^2, is

$$F(\omega) = \int_{-\infty}^{\infty} f(x)\, \exp\{-j\omega x\} dx = \frac{q}{\epsilon} \frac{1 - \exp\{-2\omega t\}}{2\omega}, \qquad (16)$$

where ω is the spatial frequency corresponding to the variable x, in radians/micron.

The sine-wave amplitude, in volt-microns^2, is

$$\psi(N) = F(\pi N \times 10^{-4}) = \frac{q}{\epsilon} \frac{1 - \exp\{-2\pi N \times 10^{-4} t\}}{2\pi N \times 10^{-4}}, \qquad (17)$$

where N is the number of half-cycles (or television lines) per centimeter, given by

$$N = \frac{\omega}{\pi} \times 10^4. \qquad (18)$$

The sine-wave response factor, or relative sine-wave amplitude normalized to unity at $N = 0$, is

$$r\psi(N) = r_t = \frac{\psi(N)}{\psi(0)} = \frac{1 - \exp\{-2\pi N \times 10^{-4} t\}}{2\pi N \times 10^{-4} t}. \qquad (19)$$

The theoretical sine-wave response of an electrostatic signal recorder target is plotted in Figure 4.

Beam Response

The target response, which specifies the theoretical limitation on recorder packing density, is a function of tape geometry only. Beam response, on the other hand, is dependent upon the state-of-the-art of electron optics and component technology. The resolution properties of electron beams do not readily lend themselves to analytical description.

Substantial improvements in high-velocity electron-beam resolution have been demonstrated in the laboratory. Some results of experimental efforts performed under contract have already been published.[5] To date, the most impressive electron guns have utilized surface-smoothed oxide cathodes on passive nickel cups.

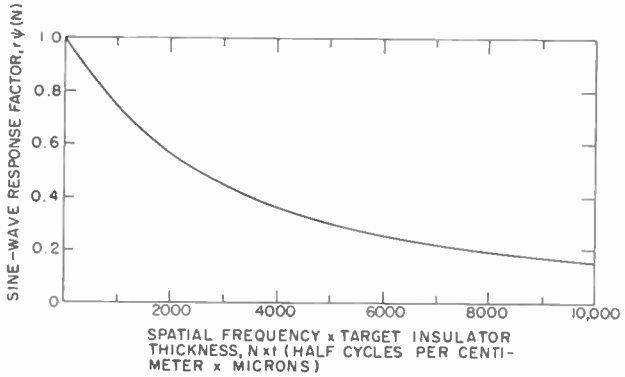


Fig. 4—Theoretical sine-wave response of electrostatic signal recorder target.

Sample Calculations

The experimentally determined sine-wave response of a 0.1-microampere, high-velocity electron beam (from a surface-smoothed oxide cathode on a passive nickel cup) is given by curve a in Figure 5. The cascaded beam response for an electrostatic signal recorder employing this same beam for recording and playback is given by curve b in Figure 5. The corresponding target responses required to produce an overall, cascaded recorder response (at each spatial frequency) of 50 per cent are plotted in curve c, Figure 5.

[5] S. Gray, P. C. Murray and O. J. Ziemelis, "Improved High Resolution Electron Guns for Television Camera Tubes," presented at 93rd S.M.P.T.E. Convention, Atlantic City, N. J., April 21-26, 1963.

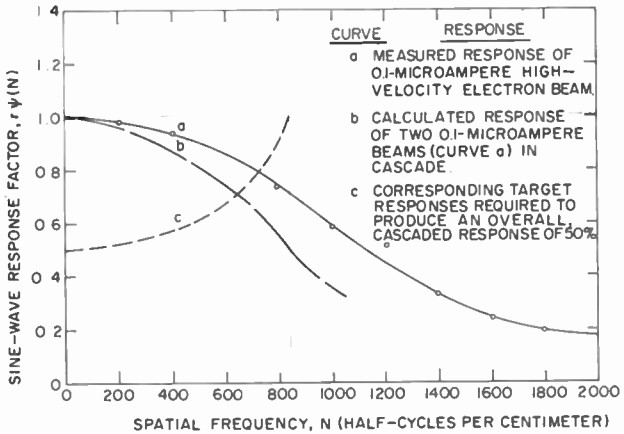


Fig. 5—Sample sine-wave responses for electrostatic signal recorder.

The insulator thicknesses yielding these target responses at the respective spatial frequencies can be obtained from Figure 4. Inspection of Equation (11) shows that a maximum output signal-to-noise ratio will occur for a maximum product tL^2. The product of insulator thickness and the square of horizontal density is plotted in Figure 6.

For the electrostatic signal recorder described above (and in Figures 5 and 6), maximum signal-to-noise ratio can be achieved with

$$t = 3.4 \text{ microns},$$

$$L = 500 \text{ half-cycles/cm}.$$

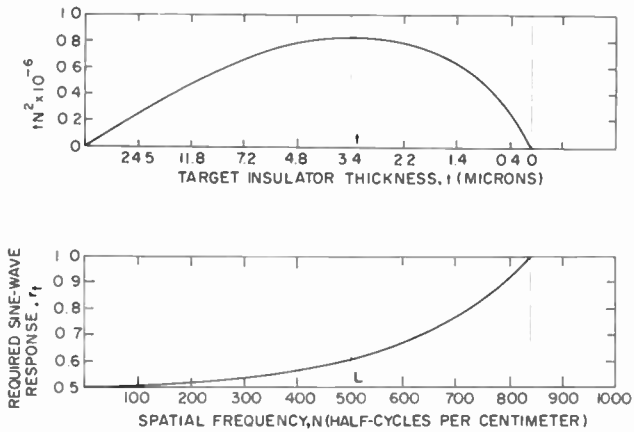


Fig. 6—Sample calculations for electrostatic signal recorder.

From Equation (12),

$$P^2 = 250{,}000\rho \text{ bits/cm}^2.$$

For a 50 per cent packing factor, the packing density would be 125,000 bits/cm^2. The required tape speed, in inches/sec, is then

$$S = \frac{2F \times 10^6}{2.54\, WP^2}, \tag{20}$$

where F is the operating bandwidth of the recorder, in mc,

P^2 is the recorder packing density, in bits/cm^2,

W is the active width of the insulating tape, in cm.

For 70-mm (nominal width) tape and a packing density of 125,000 bits/cm^2, $S \approx F$.

Applications

The effects of beam current density and operating bandwidth on the performance of the proposed electrostatic signal recorder can be illustrated graphically. If it is assumed that the recording and playback beams to be utilized by such a recorder would exhibit sine-wave responses as shown by curve a in Figure 5, then, from above,

$$t = 3.4 \text{ microns},$$

$$L = 500 \text{ half-cycles/cm}.$$

At the signal frequency F_r, the theoretical output signal-to-noise ratio of the recorder, as given by Equation (11), would be

$$\left(\frac{S}{N}\right)_{F_r} \approx \frac{6500/\sqrt{F_p/F_r}}{\sqrt{1 + \dfrac{550}{F_p/F_r}\left(\dfrac{F_r}{I}\right)^{-2}}} \left(\frac{F_r}{I}\right)^{-3/2}, \tag{21}$$

where I is the primary (recording and playback) beam current, in microamperes.

The "average" signal-to-noise ratio of the recorder would be

$$\left(\frac{S}{N}\right) \approx \left(\frac{S}{N}\right)_{F_r} + 3 \text{ db}. \tag{22}$$

Equation (22), including Equation (21), is plotted in Figure 7. When multiplied by the beam current, the variable F_r/I becomes the recording bandwidth. Thus Figure 7 contains a set of plots of theoretical signal-to-noise ratio versus bandwidth, with the playback-to-recording bandwidth ratio a parameter. The three abscissas at the top of the figure correspond to three values of beam current density; the references $I = 0.1$, 0.2 and 0.5 microampere suggest factors of 1, 2 and 5, respectively, times the current density achieved with an experimental image orthicon gun.

For each of the applications of electrostatic signal recording discussed below, the recording medium is assumed to be 70-mm, polystyrene insulating tape and the packing density is taken as 125,000 bits/cm^2 (500 half-cycles/cm $\times$ 250 tracks/cm). The sets of specifications are derived from Figure 7.

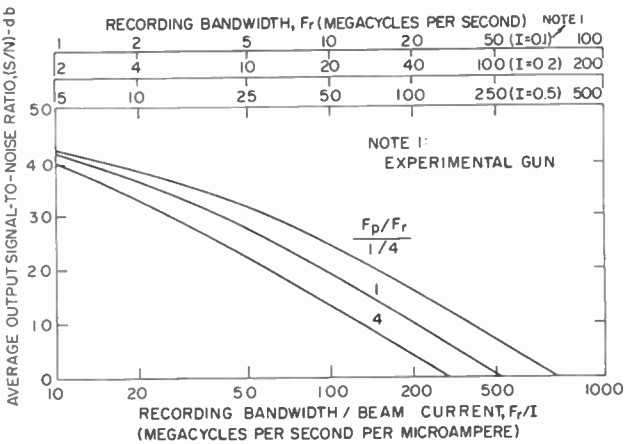


Fig. 7—Typical characteristics of proposed electrostatic signal recorder.

Tape Loop

An electrostatic tape loop could incorporate one flood electron gun and two high-resolution electron guns in a device similar to that pictured in Figure 1. Simultaneous recording and playback (the latter of a previously recorded signal) would be performed by the two high-resolution guns. Delay times as great as one year might be achieved. The loop would also permit multiple readout of recorded data.

Wide-Band Recording

A schematic representation of a possible wide-band electrostatic signal recorder is shown in Figure 1. The record–playback gun would be a high-resolution electron gun. If it were an airborne recorder, the

device would have integral means for providing an electrical signal for vehicle-to-ground transmission.

Typical Specifications—Tape Loop

Beam Current (I)	0.1 μa*	0.2 μa	0.5 μa
Bandwidth, Recording (F_r)	8 mc	8 mc	8 mc
Bandwidth, Playback (F_p)	8 mc	8 mc	8 mc
Tape Speed, Recording (S_r)	8 ips	8 ips	8 ips
Tape Speed, Playback (S_p)	8 ips	8 ips	8 ips
Signal-to-Noise Ratio (S/N)	22 $\pm$ 3 db	30 $\pm$ 3 db	38 $\pm$ 3 db

Typical Specifications—Wide-Band Recording

Beam Current (I)	0.1 μa*	0.2 μa	0.5 μa
Bandwidth, Recording (F_r)	32 mc	32 mc	32 mc
Bandwidth, Playback (F_p)	32 mc	32 mc	32 mc
Tape Speed, Recording (S_r)	32 ips	32 ips	32 ips
Tape Speed, Playback (S_p)	32 ips	32 ips	32 ips
Signal-to-Noise Ratio (S/N)	4 $\pm$ 3 db	13 $\pm$ 3 db	24 $\pm$ 3 db

Unequal Recording and Playback Rates

As a first-order effect, the theoretical output signal-to-noise ratio of an electrostatic signal recorder varies inversely with the square-root of the playback-to-recording bandwidth ratio. Thus, playback at four times the recording speed would result in only (approximately) a 6-db loss in signal-to-noise ratio. An electrostatic signal recorder employing unequal recording and playback rates would resemble the wide-band recorder.

Typical Specifications—Slow Record, Fast Playback

Beam Current (I)	0.1 μa*	0.2 μa	0.5 μa
Bandwidth, Recording (F_r)	8 mc	8 mc	8 mc
Bandwidth, Playback (F_p)	32 mc	32 mc	32 mc
Tape Speed, Recording (S_r)	8 ips	8 ips	8 ips
Tape Speed, Playback (S_p)	32 ips	32 ips	32 ips
Signal-to-Noise Ratio (S/N)	16 $\pm$ 3 db	25 $\pm$ 3 db	35 $\pm$ 3 db

* Experimental gun.

Typical Specifications—Fast Record, Slow Playback

Beam Current (I)	0.1 μa*	0.2 μa	0.5 μa
Bandwidth, Recording (F_r)	32 mc	32 mc	32 mc
Bandwidth, Playback (F_p)	8 mc	8 mc	8 mc
Tape Speed, Recording (S_r)	32 ips	32 ips	32 ips
Tape Speed, Playback (S_p)	8 ips	8 ips	8 ips
Signal-to-Noise Ratio (S/N)	11 ± 3 db	19 ± 3 db	29 ± 3 db

Conclusions

The principles and techniques underlying the photodielectric tape camera have been shown to apply to electrostatic signal recording as well. The performance of a theoretical recorder has been analyzed and shown to improve with advances in the state-of-the-art of electron optics and component technology. Characteristics of three possible recording systems have been presented.

The electrostatic recording technique is more versatile and has inherent capability for much-wider-bandwidth recording and playback than magnetic recording techniques. Also, the mechanical (tape) speeds predicted for electrostatic recorders are far slower than those required by magnetic recorders. However, it is unlikely that electrostatic signal recorders could exhibit greater than 40-db signal-to-noise ratios.

Simple electrostatic recording of analog signals has been experimentally demonstrated with single-target vacuum tubes. Demonstrations incorporating refined operating techniques, high-resolution components, long-length insulating tapes and beam tracking servos, however, have not been attempted.

Acknowledgment

The authors would like to thank T. H. Moore for his valuable assistance throughout the investigation. They are also grateful to E. C. Hutter, S. Gray and C. D. Deyerle for their continued interest and encouragement.

* Experimental gun.

COMMUNICATION-SATELLITE-SYSTEM HANDOVER REQUIREMENT AND ASSOCIATED DESIGN PROBLEMS

By

H. J. Weiss

RCA Communications Systems Division
Camden, N. J.

Summary—A concept for the possible implementation and operation of a satellite communication system is developed in various steps. Design "ground rules" to define limitations and capabilities of the concept are derived and applied to a real, typically limited example—a system comprised of station-keeping satellites in equatorial orbits. The concept is then extended to nonequatorial orbits which retain a particular characteristic of the equatorial orbit—the invariance of the satellite subtrack. In a qualitative comparison with recently proposed systems, the advantages of the concept are pointed out; in particular it is noted that a system designed under observation of the derived rules may be a serious competitor for the synchronous satellite.

Introduction

With the imminent advent of operational satellite communications, the search for the "best" system to expand the dwindling channel space in existing communications media poses immediate and quite real problems. There are a multitude of criteria that must be considered in order to select, from the many conceivable satellite–ground-station configurations, that which satisfies the largest number of requirements.

The number of analyses and recommendations already published is staggering. However, many of the conclusions arrived at by the various authors are based on a priori selection of specific systems, and therefore lead to optimizations only within the systems framework defined by the specified characteristics.

In the following paragraphs an attempt is made to outline a satellite-system concept which is highly adaptable to communication requirements of any kind and admits of optimization by many conceivable criteria. The "specific" features required for the treatment of this concept do not go beyond those tacitly stipulated for many other analyses of this kind; the satellites involved have station-keeping capabilities over the projected service period and they permit loading with a limited number of independent carriers (independent access).

Derivation of Design Rules

Design considerations are developed in this and the next three sections for general-use satellite-communication systems based on orbital subsystems of the following special character:

Orbit: Circular, equatorial; single orbital path common to all satellites.

Orbital Period: Identical for all satellites.

Satellite Spacing: Maintained uniform along the common (equatorial) orbit.

In this section, some rules are worked out that govern the utilization of ground-station antennas for such a satellite array. In the final section, consideration is extended to systems using more-general orbital subsystems.

Given a set of n ground stations S_i with the coordinates λ_i (latitude) and μ_i (longitude), each station will see a satellite in an equatorial circular orbit (satellite $\lambda = 0$, always) over an angular portion of this orbit centered on μ_i and of length θ_i given by

$$\theta_i = 2\cos^{-1}\left[\frac{r_0\cos^2\psi + (r_0^2\cos^2\psi + 2r_0h + h^2)^{1/2}\sin\psi}{(r_0 + h)\cos\lambda_i}\right] \quad (1)$$

where $r_0 =$ radius of (spherical) earth,

$h =$ satellite altitude above mean sea level,

$\psi =$ minimum useful elevation angle of ground-station antennas.

The n stations S_i permit a two-station connectivity C_{ij} ($i = 1 \cdots n$, $j = 1 \cdots n$, $i \neq j$) whenever a pair θ_i/θ_j overlaps. The width of the overlap is given by

$$\Delta\theta_{ij} = \min\left[\frac{\theta_i + \theta_j}{2} - \Delta\mu, \theta_i, \theta_j\right] \quad (2)$$

Let K be an orbital angle interval, entirely arbitrary within the limits $0 < K < \Delta\theta_{max}$, chosen to yield a reasonable number of satellites and reasonable contact times in a practical system. To provide completely interconnective communication, a C_{ij} will be said to be accepted, if $\Delta\theta_{ij} \geq K$. Any group of stations out of the set S_i which yields only accepted C_{ij} (in other words, is entirely interconnective) will be called a subset $S_{(i)}$.

Given a "reasonable" value of K centered on μ_K, a totally connective subset $S_{(i)}$ out of any set S_i will consist of all stations for which

$$|\lambda_i| \leq \cos^{-1}\left[\frac{\cos\left(\frac{\theta_0}{2}\right)}{\cos\left(\frac{K}{2} + |\Delta\mu_{Ki}|\right)}\right], \tag{3}$$

where $$0 \leq |\mu_i - \mu_K| = |\Delta\mu_{Ki}| \leq \frac{\theta_0}{2} - \frac{K}{2},$$

and $$\frac{\theta_0}{2} = \cos^{-1}\left[\frac{r_0\cos^2\psi + (r_0^2\cos^2\psi + 2r_0h + h^2)^{1/2}\sin\psi}{r_0 + h}\right]$$

Equation (3) describes the well-known lenticular shape of the area common to two circles of equal radius $\theta_0/2$ on the earth's surface, their centers located on the equator and separated by an arc K.

Among all the $\Delta\theta_{ij}$ of the accepted C_{ij} there is a smallest one, $\Delta\theta_{\min}$, defined by

$$K \leq \Delta\theta_{\min} \leq \Delta\theta_{ij}. \tag{4}$$

If uninterrupted contact is now stipulated, this requires that two consecutive satellites be within $\Delta\theta_{\min}$ at least long enough to perform handover, which effectively reduces $\Delta\theta_{\min}$ by an angle θ_h. The number of satellites required to serve the accepted C_{ij} thus becomes:

$$N = \frac{2\pi}{K - \theta_h} \geqq \frac{2\pi}{\Delta\theta_{\min} - \theta_h} > N - 1, \tag{5}$$

N being an integer.

Since the choice of K determines the minimum number of satellites required in the system, the same value of K must be valid for any totally connective subset making use of the same set of satellites. Hence, all such subsets $S_{(i)1}$, $S_{(i)2} \cdots S_{(i)l}$ permit an arc of width K to be placed inside the corresponding $\Delta\theta_{\min 1}$, $\Delta\theta_{\min 2} \cdots \Delta\theta_{\min l}$. Once placed in such a manner their midpoints will lie at μ_{K1}, $\mu_{K2} \cdots \mu_{Kl}$.

Visibility arcs, visibility overlaps, K arcs, accepted connectivities, connectivity boundaries, handover-time losses, and required satellite population have now been laid out as working tools. These tools can be used to develop some rules for the design of ground systems to meet the needs of various situations.

Cost considerations suggest that ground stations be planned to operate with the lowest possible number of antennas per station. Due to the handover requirement, this number cannot be less than two. Obviously, the two-station case will never require more than two antennas at each ground terminal; the three-station case requires more-searching analysis. Attention will first be concentrated on the case of two-antenna stations only, to define conditions under which they can provide adequate service.

Let S_1, S_2, and S_3 be three two-antenna ground stations, and let θ_1, θ_2 and θ_3 be their respective visibility zones on the equator. Two configurations are possible: (1) the accepted connectivity is C_{12}, C_{13}, and C_{23}, or (2) the accepted connectivity is C_{12}, and C_{13} (the sub-indexing is arbitrary and, therefore, fully descriptive of the configuration).

Configuration 1 stipulates that each station have full-time contact with both of the other stations. In order to achieve this, the minimum number of satellites required is determined by the common overlap of all three stations which turns out to be just the smallest two-station overlap $\Delta\theta_{\min}$. Hence, the number of satellites required is that obtained from Equation (5).

Handover for any one of the three links may be performed independently if one satellite participating in the handover is within the common overlap zone of all three stations, $\Delta\theta_{\min}$, and the other satellite is a neighbor and is anywhere in the overlap zone of the two stations involved. Furthermore. handover in any two links must not be performed during the same time (time separation of consecutive handovers in two different links must at least equal the duration of a satellite's passage through an arc θ_h). If any two links are required to perform handover during the same period, it follows that the third link must do so as well, and the condition for this is that the two satellites involved be within the common overlap angle $\Delta\theta_{\min}$.

If handover in all three links is performed during the same period, one antenna at each station will be idle during a time equal to a satellite's passage through an arc $\Delta\theta_{\min} - \theta_h$, so that it is available for contact with any station outside the three-station complex during a time equivalent to a satellite's passage through an arc $\Delta\theta_{\min} - \beta$, defined by

$$\Delta\theta_{\min} - \theta_h \geq \Delta\theta_{\min} - \beta \geq \Delta\theta_{\min} - \theta_h - 2\theta_s, \tag{6}$$

where θ_s is the arc through which a satellite will move in its orbit while the antenna is slewing. In any link, both antennas at each station are busy with handover and therefore not available in another

link during a period equivalent to a satellite's passage through an angle β, defined by

$$\theta_h \leq \beta \leq \theta_h + 2\theta_s, \tag{7}$$

depending on how much slewing may be required.

The above reasoning can be applied to any n-station complex (two antennas per station) with the accepted connectivity C_{ij}—again the minimum number of satellites required is found from Equation (5); $\Delta\theta_{\min}$ is defined as in Equation (4) and is the common overlap zone of all stations which constitute the accepted connectivity. Again, handover in any one link may be performed independently, observing the same rules as with the three-station case. If two links are required to perform handover during the same period, all links again must do so, and with two satellites present in $\Delta\theta_{\min}$.

In general, a two-antenna station common to two or more totally connective subsets can be considered an element of one of these subsets only. In other words, with the limitation of two antennas per ground station no two subsets can communicate with one another on a full-time basis unless they are identical (trivial case), are contained within each other or within a totally connective subset of higher order, or meet a specific, further-to-be-discussed requirement. This requirement is particularly significant for configuration (2) for which the accepted connectivity of the three-station complex is C_{12} and C_{13}.

The special requirement is brought about by the notable exception which occurs when, during the time equivalent to β when both antennas at either station of a link are tied up in handover, one of these antennas is simultaneously providing contact over the other link. Obviously, for this to be possible, an overlap of $\Delta\theta_{12}$ and $\Delta\theta_{13}$ is essential. The question is, how much overlap is required, and how can the two antennas at S_1 be utilized to permit uninterrupted service over both C_{12} and C_{13}.

Each of the two antennas of S_1 must track one of two successive satellites during an angle $\Delta\theta_{\min} - \theta_h$. One of the two antennas of S_1 will track a satellite through the coverage zones of both stations, or an angle $2\Delta\theta_{\min} - \theta_x$, where θ_x is the unknown minimum overlap angle. If we assume that any antenna at S_1 tracks first over C_{12} and then over C_{13}, then, during a satellite's passage through θ_x, the following actions must take place:

1. Antenna 1 of S_1 must turn over the C_{13} traffic to antenna 2 (handover; angle required θ_h).

2. After handover, antenna 1 slews back in preparation to take

over the C_{12} traffic from antenna 2 (slewing time sees the satellites through an angle θ_s).

3. After slewing, antenna 1 takes over the C_{12} traffic from antenna 2 (handover; angle required θ_h).

The same sequence is repeated with antennas 1 and 2 exchanging roles, after the satellites have moved by an angle $\Delta\theta_{\min} - 2\theta_h - \theta_s$, or the passage minimum angle $\Delta\theta_{\min}$ diminished by two handover angles θ_h and one slewing angle θ_s. (The various angles θ may be equated to corresponding time increments, since the relative angular velocity of all satellites is assumed to be constant and identical.)

From this it follows that each antenna in turn must hold both the C_{12} and C_{13} traffic over an angle

$$\theta_x = 2\theta_h + \theta_s. \tag{8}$$

The minimum number of satellites required to handle the three-station dual-connectivity traffic is, therefore,

$$N = \frac{2\pi}{K - \theta_x} \geqq \frac{2\pi}{\Delta\theta_{\min} - 2\theta_h - \theta_s} > N - 1. \tag{9}$$

It must be remembered that $\Delta\theta_{\min}$, in this case, is defined as $\min(\Delta\theta_{12}, \Delta\theta_{13})$, since C_{23} was excluded from the accepted connectivities.

It can now be seen that two subsets $S_{(i)1}$ and $S_{(i)2}$ do permit interset traffic via any one two-antenna station common to both subsets, *provided* the two K-arcs overlap by at least the angle θ_x as defined in Equation (8). Any two such stations that are common to both subsets can in that case communicate with each other continuously over at least one of two satellites; in other words, they might actually skip every other satellite for purposes of mutual communication. However, the common stations are required to track all the satellites consecutively if they are also to maintain their respective intra- and inter-subset connectivity to all other stations.

Attention may now be turned from the rules governing capabilities achievable by two-antenna stations to the characteristics of more-complex situations. Logical processes similar to those already pursued show some of the rules applicable to more-elaborate systems to be as follows:

Stations common to two subsets with non-overlapping K-zones generally require three antennas; the center separation of the two K-zones must meet the condition:

$$nK \leq |\mu_{K1} - \mu_{K2}| \leq (n+1)\,K - 2\theta_x. \tag{10}$$

The factor n is a small positive integer ≥ 1. If the condition cannot be met, four antennas at such stations are required.

If two such stations alternately provide inter-subset relay functions but otherwise are not connective within either subset, each station will require only two antennas.

Stations common to s subsets with consecutively overlapping K-zones require s antennas to provide both total connectivity within each subset and total inter-subset connectivity.

Stations common to t subsets with non-overlapping K-zones require in general $t + 1$ antennas to provide both total intra- and inter-subset connectivity. The condition for this is

$$nK \leq |\mu_{Ki} - \mu_{Kj}| \leq (n+1)\,K - 2\theta_x. \tag{11}$$

Representative System-Design Example

The design rules previously arrived at were established with the objective of achieving maximal network connectivity on a global basis with the smallest number of antennas at each ground station. Access to the global connectivity is guaranteed to any country located within, or extending to within, a geographical belt of width $2\lambda_B$ (from λ_B south latitude to λ_B north latitude), with λ_B defined as

$$\lambda_B = \cos^{-1}\left[\frac{\cos\,(\theta_0/2)}{\cos\,(K/2)}\right]. \tag{12}$$

The global system might well be implemented with stations having only two antennas, but the inter-subset traffic would have to rely on overlapping K-arcs and might involve too many satellite links for longer routes. It is more reasonable to establish subset connectivities in high-density local areas and connect the subsets by relay stations with a larger number of antennas. Then, it will not be necessary to establish overlapping K-arcs, and savings can be made on a station basis rather than on an antenna basis.

A representative system design will be carried out assuming the following parameters:

- Orbit altitude — $h = 6400$ statute miles
- Relative angular satellite velocity — $\omega_r = 0.75°/\text{min.}$
- Number of satellites — $N = 12$

- Handover period $t_h = 2$ min. (equivalent to $\theta_h = 1.5°$ relative satellite movement)
- Slewing period $t_s = 2$ min. ($\theta_s = 1.5°$)
- Width of subset overlap $K = 34.5°$
- Satellite separation $2\pi/N = 30°$
- Minimum antenna "look" angle $\psi = 5°$
- Single-station coverage-zone radius $\theta_0/2 = 62.2°$
- Maximum station latitude $\lambda_B = 60.7°$

The λ_B value of 60.7° given above immediately excludes Iceland from the global system. Since no other sovereign nation is excluded, a real system should find a means of including Iceland also, even if no initial interest in a participation exists. The inclusion can be achieved by using satellites in higher orbits or removing the stipulation of equatorial orbits. In the former case, the systems design will proceed in the manner outlined below; in the latter case, modifications of the rules are necessary.

With the basic system design parameters determined, the major totally connective subsets should be identified first. The choice of their location along the equator is essentially arbitrary, but an intimate knowledge of communication traffic densities and patterns as they exist now aids considerably in the logical selection.

Both the European/African and the North/South American complex will require a large amount of short range (one-hop) traffic. In order to reduce single-satellite loading in these areas, local traffic should not be concentrated in the assigned K-zone but should reach as far West and East of it as a given station connectivity permits. This leads in effect to a number of overlapping K-zones in the subset region. As has been shown, overlapping K-zones permit stations common to both to act as relays without requiring more than two antennas and without loss of connectivity into either subset.

Figure 1 shows a global subdivision into major subsets, identified by the areas S-1 to S-5 with the corresponding K-zones K-1 to K-5.

The major subsets are interconnected by 5 relay stations, viz. (tentatively):

R-1. Recife (Brazil)
R-2. Aden (Aden Protectorate, Br.)
R-3. Manila
R-4. Honolulu
R-5. La Paz (Mexico)

The relay stations provide the communication potential for two-hop

inter-subset traffic, using satellites during their passage through the major K-zones only. According to the ground rules they do not require more than three antennas to do so, provided the condition of Relation (10) is met.

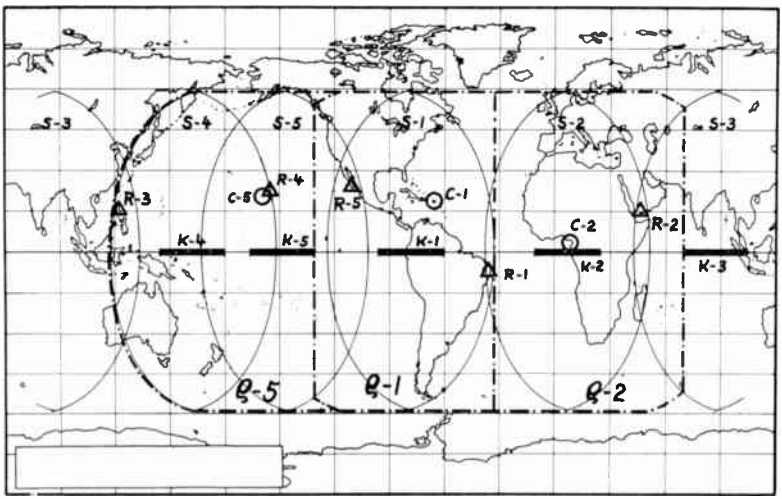


Fig. 1—Major subsets (S) and their K-arcs (K), relay stations (R), and control areas (ρ) with their control centers (C).

With the K-arcs K-1 to K-5 centered at the longitudes:

1: 78° W
2: 7° E
3: 88° E
4: 162° E
5: 148° W

all K-separations comply with Relation (10):

K-1:K-2	$69° < \lvert\mu_1 - \mu_2\rvert = 85° < 94.5°$
K-2:K-3	$69° < \lvert\mu_2 - \mu_3\rvert = 81° < 94.5°$
K-3:K-4	$69° < \lvert\mu_3 - \mu_4\rvert = 74° < 94.5°$
K-4:K-5	$34.5° < \lvert\mu_4 - \mu_5\rvert = 50° < 60°$
K-5:K-1	$69° < \lvert\mu_5 - \mu_1\rvert = 70° < 94.5°$

A fourth antenna will be required on a standby basis in order to meet emergencies. A fifth antenna permits a service rotation for pur-

poses of repairs and overhauls. Five antennas, incidentally, could handle all the satellites visible to one ground station, if this were required.

While the five relay stations previously mentioned will have to handle the bulk of inter-subset communication, which initially is seen to be heavy only between the subset pairs S-1/S-2 (U.S.-Europe) and S-5/S-1 (Hawaii/West Coast-East Coast), the manner in which satellite channels are used will determine whether these relay stations can also handle whatever intra-subset traffic control is required.

Two operational philosophies offer themselves as applicable to the global system.

(1) The system connectivity is fixed, and permanent channels are assigned each ground station on a rigid full-time or time-sharing basis. Reassignments would be relatively rare and established well in advance of their actual consummation. For this type of operation the control function may well be given to the relays, and all other stations could operate with two antennas each, unless certain individual stations require a better-than-single-subset connectivity. This procedure appears to offer the smoothest operational characteristics, but it will not provide for maximal use of the available satellite channels.

(2) The system connectivity is fixed on the inter-subset level, but random on the intra-subset level. The problem of channel assignment in such a configuration is quite complex, but not much more so than in modern switching facilities. However, coordination must be precise and reliable, and it is necessary to assign each subset a control center whose sphere of influence should extend over the entire equatorial area contained in the subset area *S*.

In a real system some compromise may be visualized which combines the simplicity of scheme (1) in low-traffic density-areas with the channel-saving characteristics of scheme (2) in the high-density-traffic zones.

For the system shown in Figure 1, subset control centers could be set up at

C-5	Honolulu
C-1	San Juan (Puerto Rico)
C-2	Fernando Poo (Spanish Guinea)

It may not be desirable to handle both inter- and intra-subset traffic at the same site. One array of antennas could perform this task easily; the complication would lie in the actual traffic supervision. If Hawaii is chosen as an intra-subset control station, the inter-subset relay may operate at Midway or Johnston Island.

The basic difference in the coverage characteristics between inter-

subset relays and intra-subset control stations is that the former serve to link subsets by using satellites in the *K*-zones of the two neighbor subsets only, thus requiring not more than three antennas at any time, while the intra-subset control stations are to handle, supervise and assign traffic over their entire coverage angle ρ_i, which, of course, includes the respective K-*i*. As has been pointed out, this would require five antennas at each intra-subset control station C-*i*. However, the ρ_i are likely to overlap somewhat, and a judicious assignment of the 360 equatorial degrees among the intra-subset control stations should define inflexibly the operational zone $\rho_{i\text{-}\mathrm{eff}}$. Subsets with high-density traffic, hence, should be assigned larger $\rho_{i\text{-}\mathrm{eff}}$ than others.

In the model system such intra-subset control stations are C-1, C-2, and C-5. C-1 and C-2 could handle all traffic originating and terminating in stations contained in the subsets S-1 and S-2, respectively. C-5 could handle all traffic for S-4 and S-5 simultaneously in a similar manner.

If C-1, C-2, and C-5 have maximal implementation (five active antennas at all times) their range of influence may be extended beyond the arbitrarily defined main subset limit to handle all traffic over the control zones defined by ρ-1, ρ-2, and ρ-5, defined in Figure 1 by heavy broken lines.

The manner in which a *C*-station handles traffic within a zone invites some discussion. Figure 2 shows the ρ-1 zone of Figure 1 isolated from the rest of the system. Both the main subzone S-1 and the corresponding *K*-zone, K-1, are shown. The entire zone ρ-1, however, is further subdivided into several displaced zones equal in shape and size to S-1. This subdivision leads to a network in which each zonal unit may be identified by a coordinate letter, a coordinate number, and the affix "North" or "South," as shown in Figure 2. Any two stations contained in an area of the size of S-1 can communicate with each other using satellites over an arc equal in length to, but displaced with respect to *K*, toward the east or west. The purpose of such a zoning of the ρ-1 area is to assign to each communicating station pair an arc K_i within ρ-1 to which it must confine its use of satellites. In most cases a station pair has quite a bit of freedom concerning the location of its assigned arc K_i.

Selecting, for example, a station in F8-North and one in H5-South, the entire arc within which the *K*-arc of this station pair may lie is considerably greater than *K* (as shown by the dark-pointed equatorial double arrow of width *E*).

The control station C-1 will assign the station pair a *K*-arc which lies entirely within the arc of freedom *E* of that pair, its chosen loca-

tion conforming to the prevailing channel and satellite occupancy.

If the channel capacity of the individual satellite is n, the mean duration of a call is t_0 (including request, addressing, assignment and switching), the total number of satellites in the system is N, and the entire control arc of a given C-station ρ, then, assuming uniform loading of all satellites contained in the arc ρ, the mean number of simultaneously existing single-channel conversations is

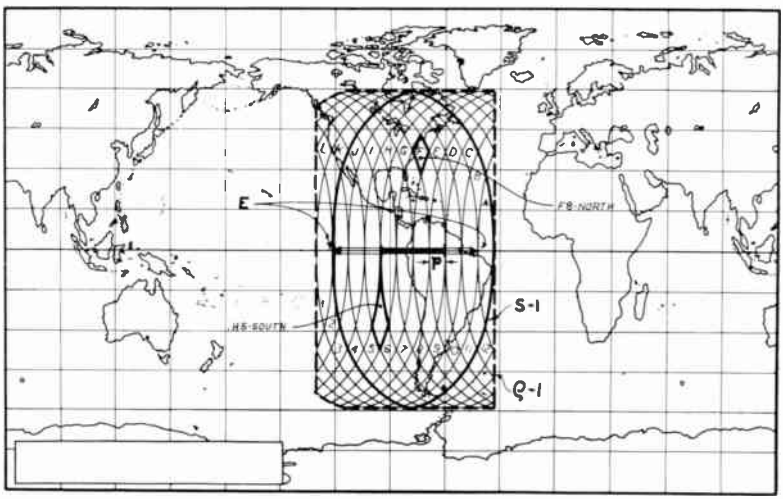


Fig. 2—Zoning of a control area (ρ-1).

$$\bar{c} = \frac{\rho N}{2\pi} n, \tag{13}$$

and the maximum possible number of calls in time t:

$$\bar{c}_t = \frac{\rho N t n}{2\pi t_0}. \tag{14}$$

Since during handover a given conversation must make use of one channel in each of two different satellites, the entire capacity cannot be realized. Assuming not only uniform channel loading on all satellites in ρ but also uniform handover density in that interval, then zonal staggering may be introduced in angular steps of width p along the equator (see Figure 2). To avoid triple loading of satellites (corresponding to the simultaneous use of a satellite for handover in

station pairs assigned to three different staggered K-zones) p must meet the condition

$$p \geq \theta_x. \tag{15}$$

In compliance with Equation (15), only double occupancy will occur and the total capacity is limited to an effective value

$$\bar{c}_{\text{eff}} = \frac{\rho N n}{2\pi}\left(1 - \frac{p}{K - \theta_x}\right). \tag{16}$$

Observing Equation (15), the maximum effective capacity of a ρ-zone becomes

$$\bar{c}_{\text{eff max}} = \frac{\rho N n}{2\pi}\left(1 - \frac{\theta_x}{K - \theta_x}\right), \tag{17}$$

where $K - \theta_x = 2\pi/N$, according to Equation (9), or

$$\bar{c}_{\text{eff max}} = \frac{\rho N n}{2\pi}\left(1 - \frac{N\theta_x}{2\pi}\right). \tag{17a}$$

The assumption of uniform satellite loading and uniform handover density within ρ tends to neglect "boundary traffic" during pickup and dropout after a satellite has entered ρ and before it leaves ρ. This can be circumvented by assuming that pickup starts when a satellite is still an angle $K/2$ outside ρ and increases linearly to a maximum when it reaches a point of $K/2$ inside ρ, and that a similar tapering off would occur when a satellite leaves ρ. This assumption places no restriction on the validity of the above equations and does not seem to increase operational complexity, since the traffic pickup and dropout can be done on a channel-block basis in such a way that on an incoming satellite the "low" blocks are occupied first, followed by successively "higher" blocks, with the same succession for the dropout. "Low" and "higher" may be agreed-upon numbered blocks or even single-channel combinations or codes, depending upon the modulation and loading characteristics of the satellite system.

Since in a given ρ-zone an arc of width K is reserved for intersubzone traffic, the local or ρ-zone traffic should preferably use arc portions that do not coincide with the K-arc traffic.

Figure 3 shows which traffic should use the extra K-arc satellites. The same ρ-zone (North- and South-America) was chosen as an exam-

ple. Remaining entirely inside ρ, the zones α and β may use the arcs K_α and K_β, respectively, to handle their internal traffic. Most of this traffic is apt to be North-South traffic, but zone α contains quite an East-West potential (for example Washington-San Francisco).

Two stations belonging to different zones α, β, and even γ, must make use of the inter-subzone K-arc K-1, either completely or partially; an exception is the case in which one or both stations are situated within the overlap region of α and β, and none anywhere in γ.

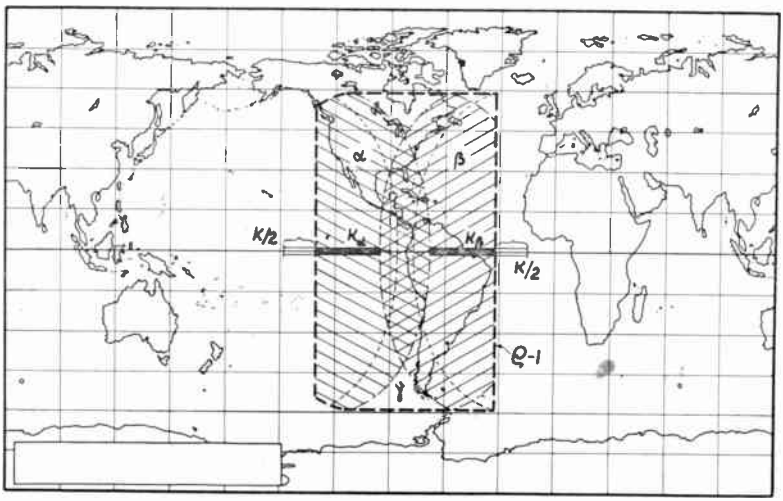


Fig. 3—Exploitation of the equatorial arc assigned to a control area (ρ-1).

On the other hand, if a station pair lies well within α or β, say within one of the zones outlined by the dotted arcs in Figure 3, then it may make use of the pickup or dropout capacity provided beyond the nominal limits of the ρ-arc mentioned above.

In the example of Figure 3, most U.S. traffic west of the Mississippi might use the pickup arc which extends with slowly increasing capacity from an arc K/2 west of the beginning of ρ-1. Similarly, most of the internal South American traffic may make use of the dropout capacity provided to K/2 east beyond ρ-1. Incidentally, South America has a considerable potential for tie-in with both Europe and Africa, and the arc K_β with its attached dropout arc of width K/2 may well be used to provide South America–Europe or South America–Africa traffic, by-passing the nominal K-1 traffic originally destined to provide all inter-subset traffic.

In general, a station pair should communicate with each other over an arc of width K which is located in ρ (in order of preference):

(1) in the pickup or dropout region,

(2) in the eastern-most or western-most K-arc entirely contained within ρ,

(3) in an arc of width K coincident with one-half or less of the inter-subzone arc K-i, and located entirely east or west of μ_{ki} (the center longitude of K-i),

(4) in an arc of width K coincident partially or completely with K-i and not subject to the restrictions of (2) or (3).

Any desired connection will be established using the lowest-numbered of the above four situations that is applicable.

Operational Aspects

The system developed in the previous section emphasizes *local* coherence in that it establishes with the S-zones and the corresponding K-arcs a fully interconnective communication capability within the subset.

The following rule, derived earlier, is necessary for the operation of such a system:

- Stations common to s subsets with consecutively overlapping K-arcs require s antennas to provide both total connectivity within each subset and also total inter-subset connectivity.

Application of this law to the typical ground station shows that when the ground station is to have an inter-subset potential, any intra-subset connectivity must be established over a K-arc that overlaps the nominal zonal K-arc, K-i, by at least θ_x; this guarantees the station's access to the system on both the inter- and intra-subset level with two antennas.

A station located within a nominal subzone S-i near the maximum latitude λ_B will essentially have *all* its traffic handled over the corresponding nominal K-arc, K-i, and will thus automatically have both inter- and intra-subset connectivity.

A station located outside any nominal subzone S-i but within a ρ-zone (e.g., Finland in Figure 1) has access to the system on an intra-subset level over one or more K-arcs which do not coincide with the nominal subset K-arc, and for inter-subset connectivity must make use of the assigned ρ-zone traffic central, C-i, which has access to *any* K-arc within the ρ-zone, thus requiring an additional hop.

This system is such that the nominal subzone K-arcs will bear full load during most of the local traffic hours, and it would seem desirable to shift the inter-subzone traffic to K-arcs with less intra-subzone traffic expectance. Thus, a K-arc for the U.S.–Europe traffic could be centered on 30°W. However, both in the U.S. and in Europe, there are only a few locations that could communicate with each other over this K-arc. If two of these are chosen, one in the U.S. and one in Europe, they could serve as relay stations for inter-subset traffic; but it is readily seen that such traffic would in almost all cases be three-hop traffic, whereas the prevailing scheme provides essentially the same capability over two hop links.

While it is desirable in the light of even-load distribution on all satellites to make use of inter-subset contact capabilities that avoid the use of a nominal K-arc, such as would be the case for stations in South America–Europe and North America–Hawaii links, a dilemma would ensue from such a mode of operation; a K-arc assigned to such a station pair could overlap only one of the nominal subset K-arcs, and might overlap none, so that one or both of the involved stations would lose the intra-subset connectivity (always assuming two antennas at the "typical" station). A station that "loses" its intra-subset connectivity regains it, of course, immediately via the ρ-zone control station, but only at the expense of an added hop.

The problem may be partially relieved if the relay stations R-i are implemented with a sufficient number of antennas to handle traffic outside the nominal subset K-arcs. In fact, the traffic coordination then could be placed with the relay stations, and the subset control stations could be eliminated. A maxim for this kind of system should be the avoidance of concentrating intra- and inter-subset traffic in the same K-arcs.

Such a design and a further system alternative are discussed in the next section.

Alternative System Designs

If the assumption can be made that a marked need for local (subset) traffic exists, for example, in parts of the world where geographical or political considerations or both preclude or hamper the installation of central stations that have access to the satellite system by means of other media, it appears reasonable to reserve the equatorial zones, which will have to bear the densest "local" traffic, for exactly that purpose, and place the "long distance" (inter-subset) traffic in equatorial portions that are not significantly occupied by intra-subset traffic.

It has been shown that the smallest equatorial arc necessary to permit full-time access of a station to the system is the K-arc. It has also been shown that optimal coherence of a system is obtained by placing successive K-arcs in such a manner with respect to each other that they overlap by at least an arc θ_r.

If one attempts to populate the entire equator with a minimum number of overlapping K-arcs, one finds that, to do so, just N K-arcs are required.

Hence, to obtain system coherence all around the equator, one needs N stations, each of which has access to two consecutive K-arcs with two antennas only, or $N/2$ stations with access to three consecutive K-arcs and three antennas each, or, generally, N/m stations with access to $m + 1$ consecutive K-arcs and $m + 1$ antennas each (provided, of course, that N/m is an integer).

The K-arcs thus subdividing the equator will be called "nominal." Similarly, the stations defined above that provide the global coherence will be called "nominal," or relay, stations.

Besides the nominal stations there will be a number of other stations. Many of these will have access to one or more of the nominal K-arcs, but some of them will have access only to K-arcs that are not nominal, i.e., that extend from a point in one K-arc to a point in one of the neighboring K-arcs. In accordance with the system design, each of the stations with non-nominal K-arcs can route its traffic at all times through one of the nominal stations, since two or more consecutive K-arcs are served by each nominal station. During part of the time a non-nominal station may have an alternative of routing its traffic over one of two nominal stations. This usage requires switching only at the nominal stations and may in some cases be sufficient to handle one or more calls over the not permanently available nominal station, at a potential saving of satellite channel space through reduction of the number of hops over which the calls in question may have to be routed.

Figure 4 shows schematically how this is done. A satellite traveling from left to right will enter the range of R-1 (the first nominal or relay station) at 0 and will be firmly acquired at 0′. Identifying this satellite as S_4, there will be satellites S_3, S_2, and S_1 going through the acquisition (and handover) phases at 2-2′, 4-4′ and 5-5′, respectively. S_3 will at this time bear all traffic from G-1, the non-nominal station, which had acquired that traffic while it was passing through 1-1′. Since R-1 tracks a satellite from 0-0′ through to 4-4′, it will remain in contact with G-1 during all of a satellite's passage from 1-1′ to 3-3′. However, a satellite passing through the acquisition interval 2-2′ will

at this time be able to service R-2 until it is dropped by R-2 at 5-5′, so that during a satellite's passage between 2-2′ and 3-3′, G-1 will be able to contact not only R-1 but also R-2. Thus, if G-1 is called upon to contact another non-nominal station G-2, with a K-arc extending from 4 to 5′ and thus in contact with R-2, then it will have to do so via both R-1 and R-2 (which are, of course, in constant contact with each other) whenever a satellite travels between 1-1′ and 2-2′. While

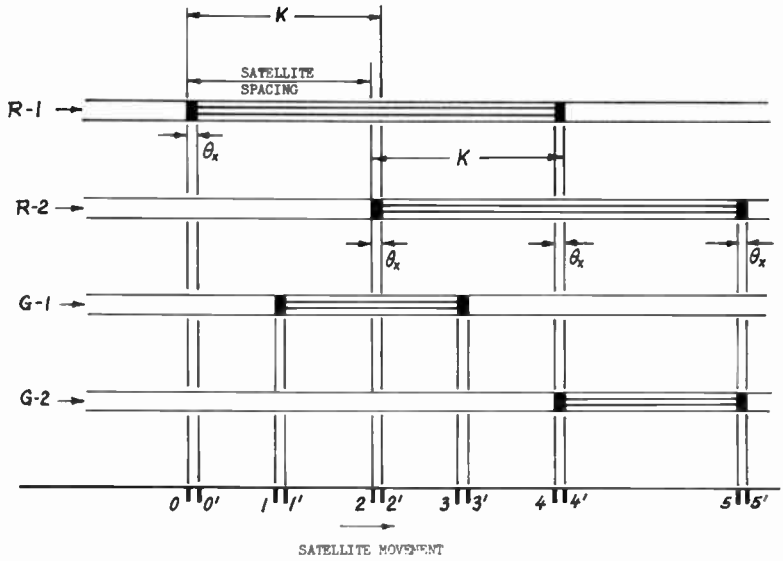


Fig. 4—Exploitation of the connectivity potential.

the satellite passes the handover interval 2-2′, the traffic between G-1 and R-1 may be switched directly to R-2, and during the satellite's passage from 2-2′ to 3-3′, the traffic between stations G-1 and G-2 may use only the nominal station R-2 as a relay, thus permitting part-time contact between G-1 and G-2 in a two-hop mode. During that time the channel occupancy of the satellite is relieved by just the amount of traffic exchanged between G-1 and G-2. If the non-nominal station G-1 covers an arc which exceeds the value K, but is contained entirely within the arc 0 to 4′, some choice exists as to where best to place its operational K-arc (1 through 3′) in order to conform with existing channel loading of the satellite and with possible direct (one-hop) contact requirements to one or more non-nominal stations besides G-2.

In Figure 5 an attempt is made to devise a configuration which at

the cost of intra-subset coherence permits a more-reasonable loading of the satellites and provides, specifically, two-hop traffic between adjacent major communication areas such as the U.S.–South America complex and the Europe–Africa complex.

The same system parameters were retained so as to permit comparison with the system presented earlier—12 satellites uniformly spaced in circular equitorial orbits at an altitude of 6400 miles.

Since the traffic within the U.S., South America, Europe, and Africa, and among these areas is particularly dense, the satellite utilization in this part of the system will be discussed as representative.

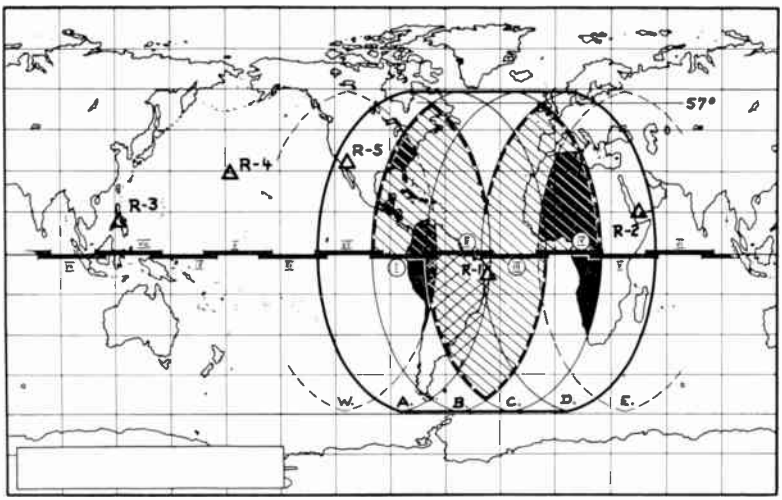


Fig. 5—Representative population of the equator with staggered nominal *K*-arcs.

The "nominal" station providing the traffic coherence in that region will best be located at Natal, Brazil; it will command coverage over four consecutive *K*-zones, designated I through IV in Figure 5. This permits access of any station within the heavily outlined circumference, and thus provides a zone within which any two stations may communicate with each other by no more than two hops.

In particular, any traffic between the U.S. and Europe will be two-hop traffic, since in this link only the "nominal" relay station R-1 will be assigned international relay status. A station which might have a one-hop contact potential with the U.S., say Madrid, will not handle any traffic other than that which originates in, or is directed to, Spain or its "reach of service." The reach of service may contain any country or countries not owning a satellite station but having access to

the Madrid station via non-satellite media. It is, therefore, implied that "many" ground stations will use the satellite system, and that the system essentially will *not* make use of traffic centers which concentrate international traffic intracontinentally by nonsatellite media. Thus, the problem of local international idiosyncrasies will be reduced to a minimum, and the discussion of the system will still be based on "area access," with emphasis on the inter-area connectivity.

Since Figure 5 is somewhat involved, it will be discussed step by step, and the results of the discussion will be summarized briefly.

The twelve K-arcs, identified on the equator by Roman numerals, are shown with the corresponding overlap arcs (width θ_x). Arcs I through IV must bear most of the traffic originating in, and leading to the U.S., South America, Europe, and Africa. As with the originally developed design each K-arc is assigned its lune, formerly called subset; the lunes are designated A through D. Any two overlapping lunes generate sublunes, as for instances the sublune AB which contains all of South America, the Caribbean, and the eastern U.S., or the sublune AC, which contains the eastern part of South America only.

R-1 is the nominal relay station for the four K-arcs I through IV, and it always has access to any of these. The entire area which can be serviced from satellites moving within the arc composed of the K-arcs I through IV is drawn out in heavy lines, and contains entirely the lunes A to D. Two main lunes, one west of lune A, designated W, and one east of lune D, designated E, are drawn in with broken lines; they serve to define certain areas to be discussed later.

Stations contained in sublunes generated by two adjacent lunes (e.g., BC) have access to at least 2 K-arcs (II and III for BC). Stations contained in sublunes generated by non-adjacent lunes, e.g., AC or AD, have access to at least 3 or 4 K-arcs, respectively, e.g., I, II, III for AC, or I, II, III, IV for AD. Obviously, the relay station must lie in sublune AD, with access to 4 K-arcs. The system parameters admit no more than four successive overlapping K-arcs to be visible to any one station.

The extreme latitude contained within a sublune of the type AB is 57°, within one of type AC is 47°; hence, any station with latitude less than 57° has access to the satellite chain over at least an angle of width $2K - \theta_x$, and a station below latitude 47° can see a satellite over an orbital arc of at least width $3K - \theta_x$.

All of these properties of the system are important for smooth traffic handling. Applied to a first-order analysis they permit the following statements:

- Lune A, containing most of the continental U.S., will require high-density local traffic concentrated over the *K*-arc I; *one-hop traffic only.*
- Similarly, lune D will concentrate intra-European traffic over *K*-arc IV; *one-hop.*
- Intercontinental traffic between the U.S. and Europe is handled in *K*-arcs II and III via R-1, connecting U.S. and European stations lying within B and C, respectively; *two-hop traffic.*
- Intercontinental traffic for stations in the U.S. and Europe, not lying within B and/or C, but within A and/or D, is handled essentially over *K*-arcs I and IV via R-1. *All Europe–U.S. traffic is two-hop.*
- Traffic within South America may be handled in either *K*-arc I or II, since all of South America is contained in the sublune AB. *One-hop traffic.*
- Traffic between South America and the U.S. may be handled over either *K*-arcs, I or, I and II, depending on whether the U.S. station lies in A only, or in AB. *All traffic is one-hop.*
- Traffic between eastern South America and western Europe or West Africa may be handled over *K*-arc III. *This is one-hop traffic.*
- Traffic within Africa is essentially handled over *K*-arcs IV and V, *in one-hop links.*
- Traffic between Africa and Europe will use IV. A certain amount of this traffic may be routed through V. *All traffic is one-hop.*
- Traffic originating in AB and leading to a station in CD, may be handled by any station in BC (which would replace the relay R-1) with only two antennas required at each of the three stations. This kind of traffic would use *K*-arcs II and III. The main areas which might participate in such a connectivity are outlined in heavy broken lines and marked by slanted straight lines.
- Stations in either the shaded areas or the crosshatched area have access to three *K*-arcs and therefore permit greatest flexibility in traffic routing.

The fact that most stations have access to an arc wider than *K* makes it possible to distribute handover intervals in a judicious manner. All stations at a latitude below 47° may enhance their connectivity potential by the installation of a third operational antenna; initially, however, this may not be required.

Since any station that is called upon to communicate with two or more other stations at the same time cannot do so over more than two

overlapping K-arcs, it should be assigned an arc of $2K - \theta_x$ on the equator within which it may normally communicate. Exceptions to this are stations above a latitude of 57° which are access-limited to less than two overlapping K-arcs anyway. Conversely, important stations located at latitudes of less than 47° may be implemented with an additional antenna and may be assigned a communications arc of $3K - \theta_x$. If this is done for each station, the one-hop connectivity will become quite rigid and some potential will be lost. In order to overcome this restriction, the relay station may temporarily dislodge the rigid pattern, but only when and where necessary. The nominal configuration should be resumed after completion of the temporary extranormal connection.

The rigid pattern may be changed with the region's rotation through the day's time zones to meet the changing traffic needs. Such a change may be preprogrammed and could be followed by each station without command from the central station.

Considerations similar to those for the U.S.–South America and Europe–Africa traffic apply to the rest of world-traffic. Conditions will initially be more relaxed due to the expected lower traffic density; however, the zonal-access principle is likely to prove valuable.

Figure 5 shows the rest of the nominal K-arcs, and also recommended locations of further relay stations:

R-5	San Diego	for XI, XII, and I,
R-4	Midway Island	for IX, X, and XI
R-3	Manila	for VII, VIII, and IX, and
R-2	Aden	for IV, V, VI, and VII

These K-assignments are only nominal; actually VII will not be entirely covered by Aden, but Manila overlaps the arc accessible to Aden by a sufficient amount to provide an overlap of K.

The number of operational antennas required at the relay-station is:

R-1	Natal	4
R-2	Aden	4
R-3	Manila	4
R-4	Midway Island	3
R-5	San Diego	3

In this system the nominal K-arcs lose the inflexibility they had in the originally designed system in which inter-subset traffic was handled rigidly over certain K-arcs.

Now the nominal K-arcs serve rather to define zones of different

types of traffic. In practice their limits will virtually disappear, and actual communication will take place over staggered arcs, in the same manner as explained in the earlier system design. The average station (below 57° latitude) will have access to an arc equal to or larger than $2K - \theta_x$ at the equator. Where a choice exists, the location assignment of these "arcs of operation" should be made on the basis of optimal connectivity into the system. As has been mentioned, such an assignment will essentially be fixed. *A station cannot use an arc wider than this "arc of operation" for continuous service with two antennas only, nor can this arc of operation be split into two arcs of width K.* Hence, once each station has been assigned its arc of operation, the only stations that may directly communicate with each other are those whose arcs of operation overlap by at least an arc of width K. So long as only a few stations take part in the system this rigidity will be tolerable; with growing system complexity some incongruities will occur, and some of the stations will be required to add one antenna to their existing two-antenna complement in order to save satellite capacity which will tend to become inadequate.

When a satellite dies, the ensuing gap in the satellite chain will have the following impact on the system:

- Stations above latitude 57° will lose contact during passage of the gap. The average loss will not exceed 100/N%, or, in this case, 8.3% of the time.
- Stations below latitude 57° will not lose access to the system although their one-hop connectivity will be drastically curtailed. Specifically, they will retain a one-hop contact potential with only those stations with which they have arc-of-operation overlaps of $2K - \theta_x$ or better. To those stations, of course, belong the nominal or relay stations for the area in question.

Thus the degradation of the system caused by the loss of a satellite presents mainly operational and loading problems, and those only while the gap is in the area of densest traffic. The loss of some communication links during the critical time is then more of a secondary effect (with the exception of links involving stations at latitudes above 57°) and a matter of choice. *Hence, no important link need suffer more than temporarily from the death of a satellite.* The possible outage time can never exceed the time it takes to recognize the outage and to switch to another satellite, possibly involving pre-emption of existing connections. Links involving stations at latitudes above 57° must not be given high priority status since they are sensitive to satellite outage.

The other system alternative using the same ground rules may be visualized as a system of ground stations, each of which has a complement of two antennas and commands an arc of width $2K - \theta_x$ on the equator. The individual arcs will be staggered in a suitable manner to permit uniform distribution of double loading during individual handover periods.

Successive stations have an arc overlap of at least width K to insure traffic coherence. Closer spacing of stations may result in an overlap by more than K; in fact, three or more stations may have a common overlap of K. This leads to a continuous one-hop connectivity of such stations around the equator.

The operational philosophy of such a configuration has been treated by others.[1] Satellite capacity is made use of in a rigid pattern, each station tracking over an arc $2K - \theta_x$, loading and unloading its traffic at certain pre-established points on the equator.

Any station operating in such a system will at all times track two satellites, one handling the station's traffic to the east (forward traffic), one handling the west-bound (back) traffic. During a satellite's passage through the west-traffic arc it will step by step enter west-traffic arcs of stations located east of the original station. Upon entering these arcs the satellite will acquire each of these stations' traffic. Simultaneously, the satellite bearing the east traffic must give up all the traffic which is being taken over by the next satellite. Traffic pickup and dropping must occur in this order to guarantee uninterrupted service.

If in this system one satellite dies, only station pairs whose access arcs overlap by $2K - \theta_x$ can continue communicating with each other on an uninterrupted basis. Station pairs that do not meet this requirement will lose contact with each other periodically.

Satellite Systems with Inclined Orbits

Coverage analyses for satellite systems to serve ground-station networks have been made in many papers; they indicate that optimal satellite orbits should be inclined rather than equatorial.

However, most of these analyses started out with a network of "worst" hops; i.e., length, location on the globe, and direction were worse than similar or neighbor links. The analyses led to optimizations in terms of one or more of such criteria as implementation cost of satellite system, number of satellites, mean passage duration of

[1] Dr. R. M. Wilmotte and H. H. Edwards, private communication.

satellites, and degradation vulnerability of systems. In most of these analyses no precautions were taken to insure cheapest ground-station implementation with maximal connectivity potential.

While the preceding sections of the present paper illustrate the implementation of a world-wide satellite communications system in just these terms, some thought should be given to the possibility of using inclined orbits and transferring some of the philosophy of area- or system-access with minimum ground-station implementation to a system with inclined orbits.

It may be said at the outset of this discussion that nothing in the *equatorial* system as presented above indicates that either 12 satellites or an orbit altitude of 6400 miles are required. The ground rules derived in the first part of the paper are valid for all circular orbits at any orbit altitude and with any commensurate number of satellites. With some modifications they may also be applied to elliptical orbits.

Subtrack Analysis

Freedom in the choice of certain system parameters is essentially lost when one wants to apply the area-access principle to an inclined satellite system.

Obviously, one of the most important conditions in the system developed previously is the *invariance of the satellite subtrack* with respect to the land masses. While this is provided automatically in an equatorial system it is, nevertheless, a basic requirement.

It is evident that any satellite whose period is commensurate with the earth's sidereal period will have nodes which repeat themselves identically every so often. The "every so often" is the crux of the matter; the ideal would be once a day, and this is just the synchronous satellite which may as well be excluded from these considerations, since its coverage characteristics are quite unique.

The only subsynchronous inclined orbit that offers a satellite subtrack that never crosses itself is the half-day orbit (day is here the sidereal period of the earth's rotation). The fact that the subtrack is a single one permits the same rules and design methods which were devised for the treatment of the equatorial orbit to be applied to this one.

Assuming a half-day *equatorial* orbit, five relay stations at Natal, Aden, Manila, Midway, and El Paso would require six satellites, uniformly spaced, yielding a K-arc of slightly more than 60°. A K-arc of 60° will permit access of stations with latitudes up to 63°. Stations below 48° latitude would have access to $2K - \theta_s$ (with θ_s assumed to be small), and no station could command $3K - \theta_s$. Loss of a satellite would seriously affect traffic continuity.

In the case of an inclined half-day orbit it becomes necessary to determine the optimal location of the subtrack to insure best world coverage. Figure 6 shows a half-day orbit subtrack of a 45° inclination system using satellites, suggesting a requirement for six consecutive K-arcs, as with the equatorial system. In Figure 7 the satellite subtrack is subdivided into 6 such K-arcs, K_A to K_F. Each K-arc has the same length as far as a satellite's travel time through it is concerned.

It is of interest to note that to populate the subtrack with 6 satellites, each one must occupy its own orbit plane. There are three times two satellites whose ascending nodes will occur simultaneously, but separated by 180° (Figure 6). Satellites travelling on the same subtrack generally are "alone" in their respective orbit planes with the exception of certain cases where two or more satellites of one orbit plane may travel along different branches of a multibranch subtrack (see Figures 9 and 10).

Each K-arc permits mapping of an area such that any station within this area can see a satellite while it is within the corresponding K-arc. Since the satellites are spaced by an angle equal to the geocentric equivalent plane width of a K-arc, there will always be a satellite in each K-arc, and hence each of the stations within one of the assigned coverage areas will always see a satellite. The areas thus assigned are designated A through F in Figure 7.

To permit global coherence of the system, each two neighbor areas must be connected by a relay station which is common to both areas and thus capable of scanning two consecutive K-arcs. Figure 7 shows land areas which are potential locations of such relay stations as darkened areas within overlap portions of the K-assigned visibility zones. Thus, relays might be located at

Recife (Brazil)	Overlap FA
Fernando Poo (Spanish Guinea)	Overlap AB
Colombo (Ceylon)	Overlap BC
Cooktown (Austr.)	Overlap CD
Loyalty Isl. (France)	Overlap DE
Mexico City	Overlap EF

In Figure 7 the coverage seems somewhat insufficient since the U.S. and Canada are excluded. However, this is due to the choice of the K-arc locations. Since a station has access to the system when it can command any arc on the subtrack of width K, the coverage is quite a bit better than is shown in Figure 7. To demonstrate this, Figure

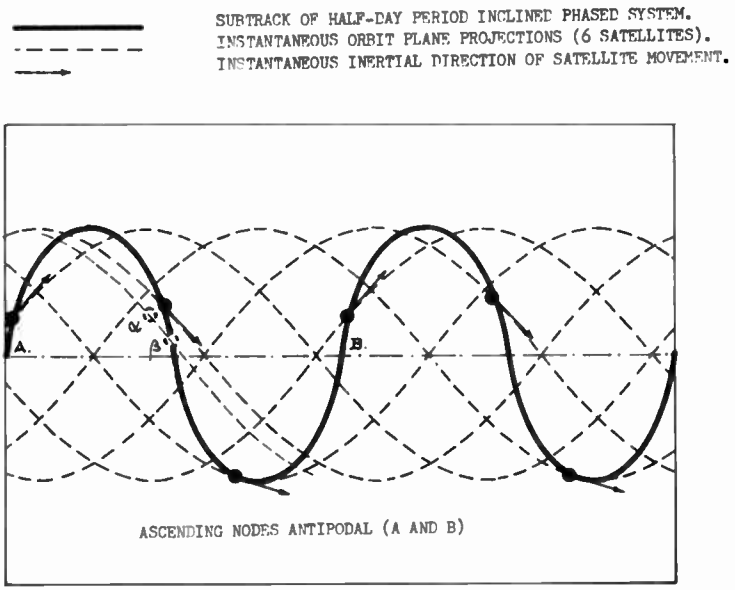


Fig. 6—The relationship between the orbital satellite movement and its stationary subtrack. Satellite in arbitrary orbit plane is either off subtrack (α) or in "wrong" position on subtrack (β). If its period is the same as that of the system, it will generate a parallel subtrack; if not, generally no closed subtrack is generated.

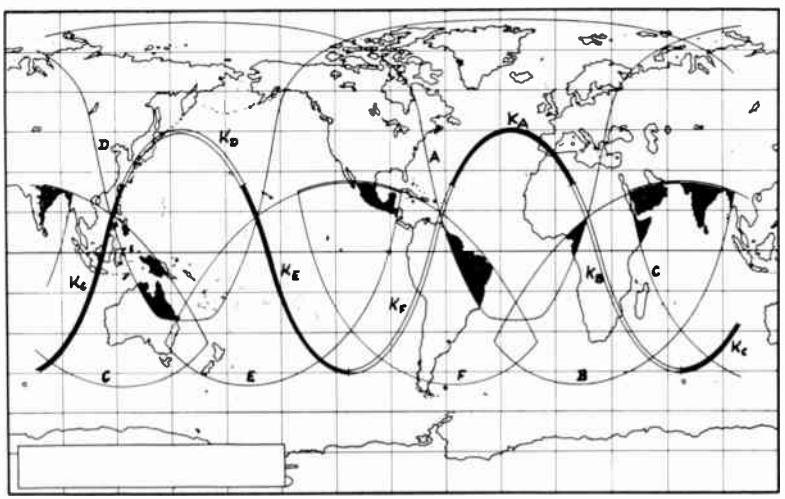


Fig. 7—Major subset areas (A . . . F) and corresponding *K*-arc equivalents for inclined orbits with a stationary subtrack.

8 shows the original zones of coverage plus such zones as are generated by shifting the *K*-arcs by just half their width, K/2. Thus a succession of areas is created whose *K*-arcs overlap by just 1/2 K. In this succession of zones, relay stations must be able to command two entire successive *K*-arcs and, hence, must lie in an area common to three successive coverage zones. (These coverage zones are designated I through XII in Figure 8; half of these are, of course, identical to the zones A through F in Figure 7). Immediately, two possibilities exist to place relay stations in locations where they command not 3, but 5 successive coverage zones or 3 successive *K*-arcs. Thus, complete global coherence is possible with the following relay stations:

La Paz (Mexico)	for IX, X, XI, XII, I
Natal (Brazil)	for I, II, III
Karachi (Pakistan)	for III, IV, V, VI, VII
New Caledonia (Loyalty Isl. Fr.)	for VIII, VII, IX

Four relay stations handle the entire system traffic.

In order to permit the traffic distribution to be shown, the orbit subtrack was subdivided into arcs of width K/2, their separation points numbered 1 through 12.

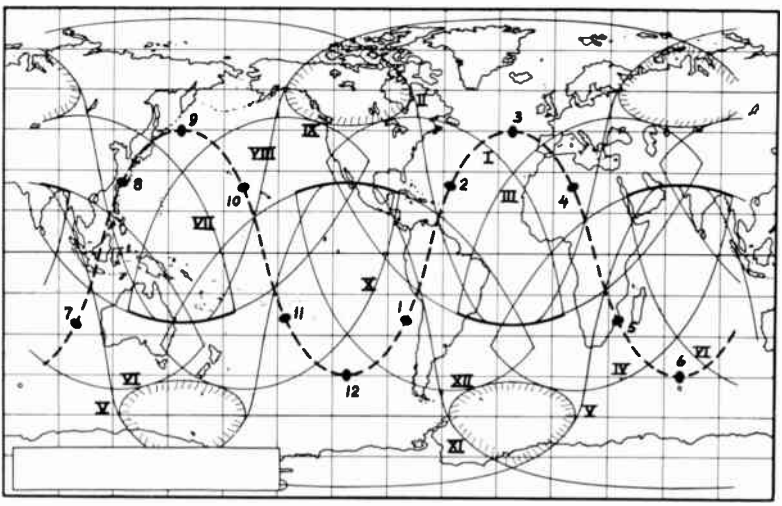
Thus, operation may be handled in the following manner (nominally):

- Internal U.S. traffic (Zone I) uses *K*-arc 1-3
- Internal South American traffic (Zone XII) uses *K*-arc 12-2
- Internal European traffic (Zones II and III) uses *K*-arcs 2-4 and 3-5
- U.S. to Europe traffic (Zones I and II) uses *K*-arcs 1-3 and 2-4, relaying through Natal
- South America to Europe (Zones XII and II) uses *K*-arcs 12-2 and 2-4, relaying through Natal
- U.S. to Far East traffic (Zones IX and VIII) uses *K*-arcs 9-11 and 8-10, relaying through New Caledonia

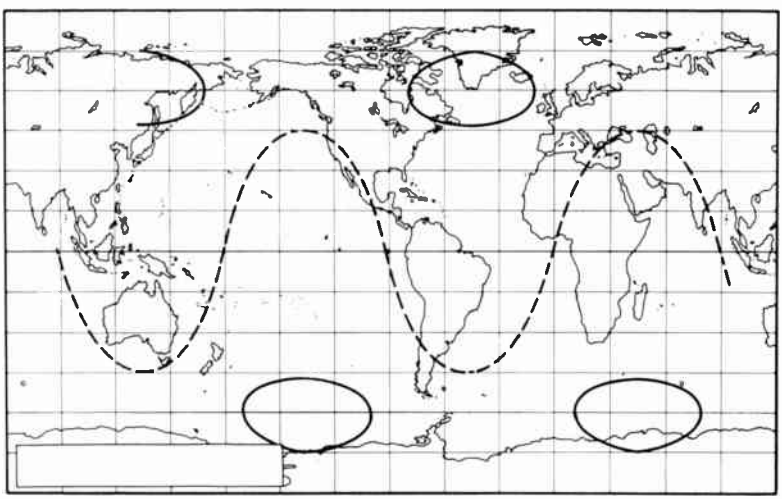
In Figure 8b the same 45° subtrack is shown displaced by 67.5°. World coverage is even better (in particular, Canadian coverage is improved), but the U.S.-to-Europe traffic would have to be handled over three hops via Port-of-Spain and Madagascar.

Variations of inclination and location permit the generation of any number of alternative schemes, but none seems really superior to the equatorial system.

There is some potential advantage to the inclined system—only



(a)



(b)

Fig. 8—System coverage in the inclined-orbit case. Alternative system coverage is shown at bottom.

four relay stations are needed, two of them requiring three active antennas, and two requiring only a pair. Since the passage of a 12-hour (sidereal) satellite through a 60-degree K-arc requires about 4 hours, the loss of communication of a few minutes at the end of this time may well be tolerable; this opens up the possibility of implementing many stations with only *one* antenna (in the equatorial 6-hour system the passage of a satellite through a 60-degree K-arc lasts for 80 minutes only). If a station has an access potential of 2K, it should be implemented with two antennas; one antenna could track through 8 hours without slewing.

Of particular interest may be the 30° orbit which, with a given booster, permits at this time the largest payload to be placed into a half-day orbit.

When all is said and done, the half-day orbit is very nearly as practical as the synchronous orbit, with the exception of the tracking requirement. Station keeping is needed in either case, and the first synchronous satellites will have a limited capacity that may well dictate the use of twice as many satellites as would be required from coverage minimum criteria alone.

Some advantage may be gained from the fact that the half-day orbit admits of some increase in system gain over the synchronous orbit, resulting in either a savings of power, antenna size, or an increase in capacity.

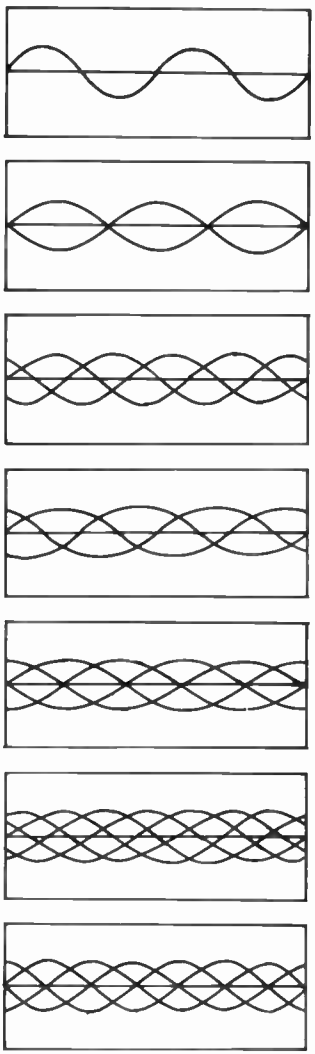
In discussing orbits with periods shorter than one half day, the simple subtrack no longer holds. Figure 9 shows the subtracks for several different orbit periods together with the corresponding orbit altitudes. Conforming to the stipulation that satellite orbit period and sidereal day yield a ratio equal to small integers, the sidereal day exceeding twice the satellite orbit period, i.e.:

$$T_{\mathrm{SAT}}/\mathrm{DAY} = n/(n+m), \quad m > n \tag{18}$$

then m is just the multiplicity of the subtrack, i.e., the number of subtrack branches cut by each meridian.

The advantage of handing over satellites which follow each other along the same subtrack will in general be lost with the multibranch patterns.

In order to regain this advantage the satellite population would have to be very dense and would result in a certain amount of coverage redundancy. If the single satellites have only limited capacity, so that two over each region at all times would be required to yield sufficient communication potential, then the multibranch patterns might come



$\frac{T_{SAT}}{T_{DAY}} = \frac{1}{2}$ T_{SAT} = 720 min, h ≅ 12,600 mi

m = 1

$\frac{T_{SAT}}{T_{DAY}} = \frac{1}{3}$ T_{SAT} = 480 min, h ≅ 9,850 mi

m = 2

$\frac{T_{SAT}}{T_{DAY}} = \frac{2}{5}$ T_{SAT} = 576 min, h ≅ 10,330 mi

m = 3

$\frac{T_{SAT}}{T_{DAY}} = \frac{1}{4}$ T_{SAT} = 360 min, h ≅ 6,400 mi

m = 3

$\frac{T_{SAT}}{T_{DAY}} = \frac{1}{5}$ T_{SAT} = 288 min, h ≅ 5,000 mi

m = 4

$\frac{T_{SAT}}{T_{DAY}} = \frac{2}{7}$ T_{SAT} = 411 min, h ≅ 7,550 mi

m = 5

$\frac{T_{SAT}}{T_{DAY}} = \frac{3}{7}$ T_{SAT} = 618 min, h ≅ 11,000 mi

m = 4

Fig. 9—Multibranch subtracks and characteristics of the associated orbits.

into their own, as they are particularly invulnerable to the loss of individual satellites due to their redundancy. Plainly, a system might then be based upon the same rules and modes as used and applied to the previous systems.

An evaluation of the coverage of multibranch subtracks or even free equi-inclined, equiperiodic orbits in the light of area access must

use a different yardstick than the K-arc, since no two consecutively visible satellites will move along the same subtrack branch unless the system contains a redundant number of satellites, a contingency which we do not have to be concerned with.

The quest now is for the best pattern, optimal inclination, and minimum number of satellites that permits global coherence with a minimum of relay stations, minimal implementation at the average station, and a given degree of area coherence.

In a rigid (or multibranch) pattern the satellite movements occur periodically over the same subtracks. This yields an advantage over

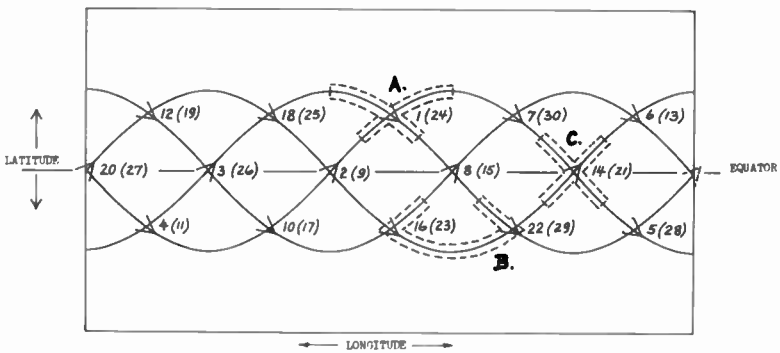


Fig. 10—Zones of permanent occupancy on a multibranch subtrack.

a system in which the subtrack pattern slowly wanders around the polar axis, as is the case when satellite period and sidereal day are incommensurate, or when its multiplicity is so high that the periodicity in satellite appearances over the same subtrack cannot be fruitfully utilized.

An example of the use of the coverage characteristics in the fixed pattern case is as follows: Populating a 4-branch subtrack corresponding to certain 5000-mile, 1/5-day orbits, with 15 satellites, one obtains the satellite distribution of Figure 10. It can be shown that the two crossing arcs, marked A, cut out of two of the subtrack branches and emphasized by a broken outline, will at all times contain a satellite. The individual satellites will traverse through the zone in the order of their numeration in Figure 10.

Similarly, the outlined arc configurations, B and C, will also at all times contain one satellite, and other such arc combinations can be found.

Since the system is symmetrical, each of the configurations, A or B, occurs 10 times in the entire pattern; C, five times.

A and C use satellites from two branches of the subtrack only; passages over each branch alternate. B contains one satellite out of three branches, and the passages occur in a cyclical manner over the three arcs.

Suitable superposition of the subtrack pattern upon the earth's land masses will yield areas which simultaneously have access to certain configurations (such as A, B or C from Figure 10) on the satellite orbit sphere. Such areas would be completely intraconnective.

Selecting relay stations which simultaneously have access to two or more of the "configurations" and implementing them with sufficient antennas to serve all visible satellites will provide the global coherence.

The fact that the subtrack branches are stationary permits the use of a minimal number of relay stations. On the other hand, the need for tracking over alternating subtrack branches does not permit the "average" station to track "through" the equivalent of overlapping K-arcs unless its coverage characteristics are such that it has access to a sufficiently large portion of one subtrack branch to permit handover of two consecutive satellites on that branch.

Satellite orbits with stationary subtrack patterns favor the area-access approach only when they have few subtrack branches, such as the half-day or one third-day pattern shown in Figure 9, with $m = 1$ and $m = 2$, respectively.

Satellite-Sphere Analysis

The intentional downgrading of single-hop regional service to ground stations paired at random leads to the approach of Reference (2), based upon the erection of a global-system skeleton consisting of balanced numbers of satellites and ground relays or "post offices." The approach deliberately disregards possible system limitation by limited traffic capacity of individual satellites.

In designing this skeletal system, the following conditions must be observed:

1. The system must be closed, i.e., each post office must at all times be able to contact each other post office, although not necessarily via the same route.

2. Each average or "consumer" station to be served by the system must have access to at least one satellite at all times.

---

[2] D. G. C. Luck, "System Organization for General Communication via Medium-Altitude Satellites," 8th National Communications Symposium (Sponsored by the I.R.E.), Utica, N. Y., Oct. 1-3, 1962.

3. Each satellite in use by a customer station must at the same time be in contact with at least one post office.

Implementation requirements call for two antennas at the customer station and a sufficient number of antennas at the post offices to comply with condition 3.

The major disadvantage of such a system is the neglect of single-hop connectivity potential and the ensuing multiple loading of satellites. The advantage, on the other hand, lies in the efficient central control of channel utilization, calling processes and traffic routing possible in such a system, as well as in the economy of needing only two antennas at any consumer station. Economy in numbers of post offices and satellites results from exploitation of net symmetry.

Obviously, the three conditions can be met by populating one of the stationary subtrack patterns of Figure 9 with a sufficient number of satellites. It may, however, be advantageous to avoid the requirement of a stationary subtrack which freezes a significant degree of freedom from the choice of orbits by limiting orbit altitude, and at the same time imposes an additional condition upon the station-keeping characteristics of the satellites.

In order to analyze the coverage characteristics of a station-keeping system that uses satellites in inclined orbits, it is useful to define areas in the orbital sphere which will always contain a satellite. To simplify this task, we shall assume that the satellites are distributed over k equally inclined orbit planes with n satellites in each; the ascending nodes of the orbit planes are distributed uniformly in the equatorial plane.

Following the example set by Reference 2, we choose as a reference system one which rotates around the earth axis at the angular velocity of the satellites, its rotation vector including an angle of less than 90° with that of the satellites. The satellites' movements in the rotating reference system occur along analemma-type subtracks. The number of satellites in one orbit plane (n) determines the number of subtracks. Satellites in different orbit planes can be made to occupy identical subtracks, and the number of orbit planes (k) determines the number of satellites following along the same subtrack; to achieve this, it is necessary that isochronous satellites in all orbits be properly phased with respect to each other.

Figure 11 illustrates the geometry involved. On each of $n = 3$ subtracks $k = 3$ satellites are shown. The angle η denotes the separation $2\pi/n$ of the uniformly spaced satellites in a single orbit, i is the inclination of the orbits, and δ denotes the meridional (*not* geocentric) width

of the retarded and the advanced limbs of the subtrack relative to the rotating reference system. This meridional width is given for circular orbits by

$$\delta = 2 \left| \max \left[\frac{2\pi t}{T_s} - \tan^{-1} \left(\cos i \tan \frac{2\pi t}{T_s} \right) \right] \right| . \qquad (19)$$

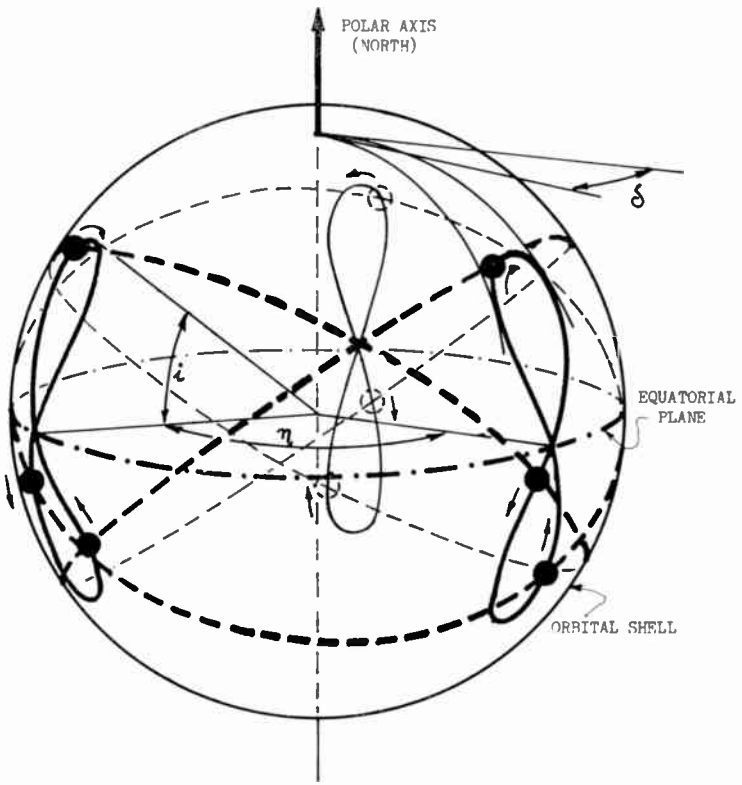


Fig. 11—The rotating figure-eight concept (by permission of Dr. D. G. C. Luck).

This angle δ becomes rather large for values of i approaching 90° (t is time in the same units as T_s, the satellite's period).

With n equispaced satellites at the same altitude in circular orbit in a common orbit plane, it is always possible to divide the orbit into n fixed sections or section pairs each of which contains at all times exactly one satellite which generates one figure-eight subtrack. Figure 12 illustrates this principle for n from 2 to 5. Each of the areas desig-

nated A, B, . . . E, fixed with respect to the line identified by EQ, isolates an arc or a pair of arcs, and in each of these arcs A, . . . E there will be just one satellite at all times. The intervals A . . . E are considered closed, i.e., a satellite on the dividing line between any two such intervals is considered to be *in* one of the two intervals.

Since the subtracks as shown in Figure 11 are projections of n equidistant satellites, the intervals A . . . E may likewise be projected

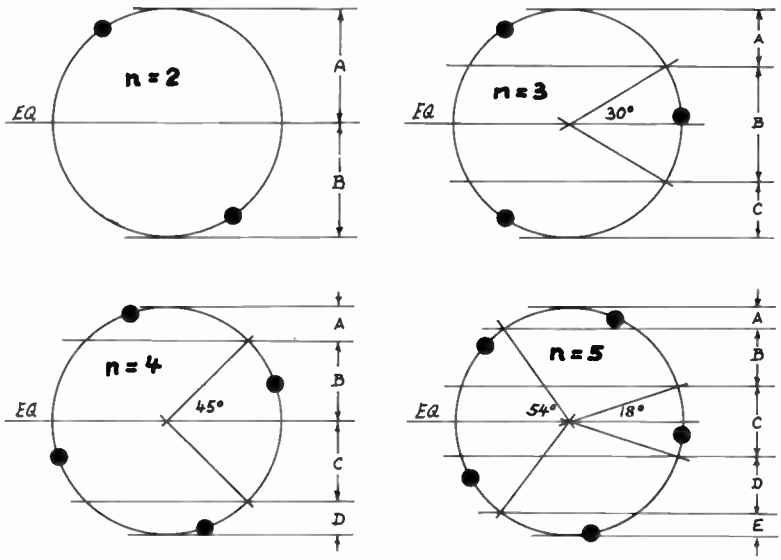


Fig. 12—Zones of permanent occupancy in orbit planes containing n equispaced satellites.

on the orbit shell of Figure 12, aligning the line EQ with the equatorial plane. The intervals A . . . E then each identify a pair of latitudes between which one satellite will be, on either the left or the right branch of the subtrack, for just $1/n$ of the time. The longitudinal separation $2\pi/n$ of two consecutive subtracks, together with the latitude intervals for the chosen value of k, define an area on the orbit shell in which at all times there is at least one satellite. Figure 13 shows the areas thus defined for a symmetrical 9-satellite system containing three satellites properly phased in each of 3 orbits. Coverage areas will adjoin closely as shown.

The size of the synchronized coverage areas are invariant towards their common rotation around the polar axis, and a station that can

see all of coverage area A-2, for example, will *always* see a satellite.

Relay stations placed in such a manner that any two adjacent ones can at all times see a common area of the size of one of the coverage areas are assured continuous contact.

The consumer station able to see all of one coverage area will also always see a satellite, but that satellite will in general pass through the visibility range of three relay stations. Hence, the customer's traffic will enter the system through at least two different post offices, though the customer need give this no attention.

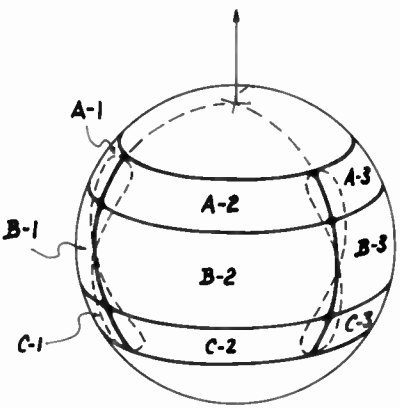


Fig. 13—Zones of permanent occupancy in a well synchronized nine-satellite inclined system.

A specific case has been analyzed in which one of the relay stations, with very large visibility area, is located at or near the North Pole and another four stations are uniformly spaced on or near the equator, so that complete global coverage is guaranteed. The satellite system consists of nine satellites in three orbits, totally synchronized at an orbit altitude of somewhat under 10,000 miles, to assure the relay stations of sufficient visibility area.

A similar system may be designed with 16 satellites in four orbits at lower altitude and a total of five or six relay stations. The orbit inclinations necessary to fulfill the coverage requirements of these systems are about 60°. If full global coverage is not required, the two systems comprising 12 satellites in four orbit planes, and 15 satellites in five orbit planes at correspondingly lower orbit altitudes have reasonably good characteristics. To determine which of the various

systems is the best, one would require a detailed comparative analysis of each one.

Surprisingly, there does not seem to be any significant advantage in coverage characteristics of these later systems (since high-latitude regions of the earth are sparsely inhabited) over the earlier developed equatorial systems—either in the number of relay stations or in the number of satellites. On the contrary, the use made at the "average" ground station of a two-antenna complement appears to be superior in the equatorial system. Effective use of single-hop modes in the equatorial systems suggested must be weighed against central control of highly miscellaneous traffic in the inclined-orbit post-office systems. The fact that some range advantage results from "low" orbits with reasonably long passage durations makes the isochronous but not earth-synchronous equatorial system a close contender to the earth-synchronous equatorial system; in either case a single antenna at the average ground station appears to be sufficient.

Acknowledgment

I am indebted to Dr. D. G. C. Luck, who went through the entire material contained in the original version of this paper and came up with some errors, now corrected, and numerous suggestions for the improvement of form and intelligibility of the concepts presented.

RCA TECHNICAL PAPERS†

Second Quarter, 1963

Any request for copies of papers listed herein should be addressed to the publication to which credited.

"Absorption Spectrum of Germanium and Zinc-Blende-Type Materials at Energies Higher than the Fundamental Absorption Edge," M. Cardona and G. Harbeke, *Jour. Appl. Phys.*, Part 1 (April) 1963

"Cerenkov Radiation and Leaky Waves," L. W. Zelby (Author's Comments), *Proc. IEEE* (Correspondence) (April) 1963

"Conditions Existing at the Onset of Oscillator Action," R. D. Larrabee, *Jour. Appl. Phys.*, Part 1 (April) 1963

"Design and Operating Characteristics of a High-Bit-Density Permalloy Sheet Transfluxor Memory Stack," G. R. Briggs and J. V. Tuska, *Intermag Conf. Proc.*, p. 1 (April) 1963

"Duo-Emitter Diode," K. G. Hernqvist and J. R. Fendley, Jr., *Jour. Appl. Phys.*, Part 1 (April) 1963

"Experimental Tunnel-Diode Electromagnetic Delay Line Storage Registers," C. M. Wine and L. S. Cosentino, *Proc. IEEE* (Correspondence) (April) 1963

"High-Field Study of a Hall-Effect Microwave Converter," K. K. N. Chang and R. D. Hughes, *Jour. Appl. Phys.*, Part 1 (April) 1963

"Magnetic Memories—Capabilities and Limitations," J. A. Rajchman, *Jour. Appl. Phys.*, Part 2 (April) 1963

"Measurements of the Density of GaAs," L. R. Weisberg and J. Blanc, *Jour. Appl. Phys.*, Part 1 (Communications) (April) 1963

"Negative Resistance in Constricted Semiconductors," K. Ando, M. C. Steele, and M. A. Lampert, *Jour. Phys. Soc. of Japan* (April) 1963

"Reasons for the Failure of Radio Interferometers to Achieve Their Expected Accuracy," D. K. Barton, *Proc. IEEE* (Correspondence) (April) 1963

"Switching Properties of a Single-Crystal Specimen of Nickel Ferrite," J. C. Miller and Coauthor, *Jour. Appl. Phys.*, Part 2 (April) 1963

"Theory of Double Injection into a Semiconductor of Finite Cross Section," R. Hirota, S. Tosima, and M. A. Lampert, *Jour. Phys. Soc. of Japan* (April) 1963

"Ultimate-Speed Adders," J. Sklansky, *Trans. IEEE PGEC* (Correspondence) (April) 1963

"Laser-Induced Emission of Electrons, Ions, and Neutral Atoms from Solid Surfaces," R. E. Honig and J. R. Woolston, *Appl. Phys. Letters* (April 1) 1963

"Observation of Paramagnetic Resonance Centers in GaAs in Unusually High Concentrations," N. Almelek and B. Goldstein, *Appl. Phys. Letters* (April 1) 1963

"Snap-Off Diodes," K. C. Hu, *Electronics* (Comment) (April 5) ... 1963

"Optical Maser Action in a Eu^{+3}-Containing Organic Matrix," N. E. Wolff and R. J. Pressley, *Appl. Phys. Letters* (April 15) 1963

"Plasma Quenching by Electro-Negative Gas Seeding," G. G. Cloutier and A. I. Carswell, *Phys. Rev. Letters* (April 15) 1963

"An Accurate Technique for Measuring Weakly Coupled Slots in Rectangular Waveguide," D. Miller and Coauthor, *Microwave Jour.* (May) 1963

† Report all corrections to *RCA Review*, RCA Laboratories, Princeton, N. J.

"An All-Electronic Method for Tuning Organs and Pianos," Part II, A. M. Seybold, *Audio* (May) 1963
"Analysis of the Arc Mode Operation of the Cesium Vapor Thermionic Energy Converter," K. G. Hernqvist, *Proc. IEEE* (May) 1963
"Cadmium Selenide Thin-Film Transistors," F. V. Shallcross, *Proc. IEEE* (Correspondence) (May) 1963
"Error Coefficients Ease Servo Response Analysis," S. Shucker, *Control Eng.* (May) 1963
"Investigation of the Electrochemical Characteristics of Organic Components—X. Sulfur Compounds," R. Glicksman, *Jour. Electrochem. Soc.* (May) 1963
"Measurement of Tunnel-Diode Parameters," F. M. Carlson, *Electronic Equipment Eng.* (May) 1963
"Photoconductivity Performance in Large Single Crystals of Cadmium Sulfide," R. H. Bube and A. B. Dreeben, *Jour. Electrochem. Soc.* (May) 1963
"A Study of Tracking-Angle Errors in Stereodisk Recording," J. G. Woodward and E. C. Fox, *Trans. IEEE PGBTR* (May) 1963
"Comparison of the Crystal Fields and Optical Spectra of Cr_2O_3 and Ruby," D. S. McClure, *Jour. Chem. Phys.* (May 1) 1963
"Lattice Vibration Spectra of Germanium-Silicon Alloys," R. Braunstein, *Phys. Rev.* (May 1) 1963
"The Range of Hot Electrons and Holes in Metals," J. J. Quinn, *Appl. Phys. Letters* (May 1) 1963
"Valence Band Structure of Germanium-Silicon Alloys," R. Braunstein, *Phys. Rev.* (May 1) 1963
"ESP," H. Sinofsky, *Electronics* (Comment) (May 10) 1963
"The Crystal Structure of Tetramethylammonium Mercury Tribromide, $N(CH_3)HgBr_3$," J. G. White, *Acta Crystallographica*, (May 10) 1963
"Paramagnetic Resonance of Divalent Holmium in Calcium Fluoride," H. R. Lewis and E. S. Sabisky, *Phys. Rev.* (May 15) 1963
"On the Paramegnetic Resonance Spectrum of the $Cr(CN)_5NO^{3-}$ Ion," I. Bernal and S. E. Harrison, *Jour. Chem. Phys.* (Letters to the Editor) (May 15) 1963
"Determination of Epitaxial-Layer Impurity Profiles Using Microwave Diode Measurements," H. Kressel and M. A. Klein, *Solid-State Electronics* (May-June) 1963
"AGREE-able Experience," P. J. Goldin, *Environmental Quarterly* (June) 1963
"An Analysis of the Characteristics of Insulated-Gate Thin-Film Transistors," H. Borkan and P. K. Weimer, *RCA Review* (June) 1963
"Determination of Oxygen in Gallium Arsenide by Neutron Activation Analysis," R. F. Bailey and D. A. Ross, *Analytical Chemistry* (June) 1963
"Efficiency Calculations of Thermoelectric Generators with Temperature Varying Parameters," R. W. Cohen and B. Abeles, *Jour. Appl. Phys.* (June) 1963
"An Experimental Parametric Tuner for UHF Television Receivers," L. A. Harwood and T. Murakami, *RCA Review* (June) 1963
"The Future of Semiconductor Devices," A. M. Glover, *Electronic Industries* (June) 1963
"Growth Rates of Epitaxial Gallium Arsenide," N. Goldsmith, *Jour. Electrochem. Soc.* (June) 1963
"High-Cutoff-Frequency GaAs Diffused-Junction Varactor Diodes," L. H. Gibbons, Jr., M. F. Lamorte, and A. E. Widmer, *RCA Review* (June) 1963
"High Voltage Epitaxial Gallium Arsenide Microwave Diodes," H. Kressel and N. Goldsmith, *RCA Review* (June) 1963
"Magnetic Load-Sharing Switches for High-Speed Applications," R. B. Lochinger, *RCA Review* (June) 1963

"A Nondestructive Measurement of Carrier Concentration in Heavily Doped Semiconducting Materials and Its Application to Thin Surface Layers," I. Kudman, *Jour. Appl. Phys.* (Communications) (June) 1963

"Operation of a Memory Element Based on the Maser Principle," H. J. Gerritsen, *Proc. IEEE* (Correspondence) (June) 1963

"The Performance of Sum and Difference Mode Parametric Amplifiers in Television Receivers," D. D'Agostini, *RCA Review* (June) 1963

"A Rigorous Analysis of Harmonic Generation Using Parametric Diodes," K. K. N. Chang and P. E. Chase, *RCA Review* (June) 1963

"Five-Layer Diode Guards Cables," H. R. Montague, *Electronics* (Components and Materials) (June 21) 1963

"S-Band Paramp Approaches Noise-Figure Minimum," P. Koskos, D. Mamayek, W. Rumsey, and C. L. Cuccia, *Electronics* (June 28) 1963

"Future Needs in Research and Training," M. M. Tall, *9th Nat. Symposium on Reliability and Quality Control* 1963

"A Procedure for System Maintainability Testing," B. L. Retterer and R. A. Miles, *9th Nat. Symposium on Reliability and Quality Control* 1963

"RCA's Experience with AGREE," P. J. Goldin, *9th Nat. Symposium on Reliability and Quality Control* 1963

"Reliability Evaluations by Computer Simulation," E. W. Veitch and G. Ashendorf, *9th Nat. Symposium on Reliability and Quality Control* 1963

"Reliability Models in Space Systems Planning and Decision Making," G. H. Sandler, *9th Nat. Symposium on Reliability and Quality Control* 1963

AUTHORS

RICHARD W. AHRONS received the B.S. and M.S. degrees from the Massachusetts Institute of Technology in 1954 in the field of Electrical Engineering. In 1963 he received the Ph.D. degree from the Polytechnic Institute of Brooklyn. In 1954, he joined the Research Staff of RCA Laboratories where he worked six years in the field of solid state circuits and television. In 1960, he received a Graduate Study Award from RCA Laboratories enabling him to pursue full time doctoral studies. He returned to RCA Laboratories in 1961 and began his research in the field of computer applications of superconductivity which led to his doctoral dissertation. Dr. Ahrons is a member of Sigma Xi, Tau Beta Pi, and Eta Kappa Nu, and the Institute of Electrical and Electronics Engineers.

JUAN J. AMODEI received the B.S. in Electrical Engineering from Case Institute of Technology in 1956 and the M.S. degree from the University of Pennsylvania in 1961. In June 1956 he joined Philco Corp., where he worked on transistor radio development. He joined RCA in March 1957 and spent the following three years in the Advanced Development Department of the Electronic Data Processing Division. His assignments included the development and design of several special purpose digital and non-digital systems. He transferred to RCA Laboratories at Princeton in December 1959, where he has been engaged in research in the field of solid-state digital circuits. He is a member of the faculty of the Evening Division of LaSalle College, Philadelphia, Pennsylvania. Mr. Amodei is a member of the Institute of Electrical and Electronics Engineers and Eta Kappa Nu.

JOSEPH R. BURNS received the B.S. and M.S. degrees in Electrical Engineering from Princeton University in 1959 and 1962, respectively. Since 1959, he has been employed at the RCA Laboratories in Princeton, New Jersey where he has engaged in research on magnetic switching and high-speed digital logic circuits. Mr. Burns is a member of the Institute of Electrical and Electronics Engineers.

JACQUES DUTKA received the B.S. degree from the City College of New York in 1939, the A.M. degree in Mathematics from Columbia University in 1940, and the Ph.D. degree in Mathematics from Columbia in 1943. From 1944 to 1946, he was a member of the Operations Research Group of the U.S. Navy, where he made studies on anti-submarine warfare as well as accuracy studies of aerial attacks on shipping. In 1946, he served as consultant to the office of Naval Research. From 1947 to 1953, Dr. Dutka was an Assistant Professor of Mathematics at Rutgers University, and from 1953 to 1956, he was employed at the Norton-Ketay Corp., where he specialized in military research and development projects. In 1956, he joined the RCA Surfcom Systems Laboratories as a senior engineer, and is currently directing the activities of a group engaged in operations and systems analysis of command communications problems. Since 1954, he has been teaching at Columbia University and is presently an Adjunct Professor of Electrical Engineering. Dr. Dutka is a member of the American Mathematical Society, the Institute of Mathematical Statistics, and Sigma Xi.

JOHN ADAMS INSLEE received the B.E. degree in Electrical Engineering from Yale University in 1955, and the M.S. in Electrical Engineering from Rutgers University in 1958. From 1951 to 1953 he was in the U.S. Army Signal Corps where he was an instructor of radar and guided missile electronics at Fort Monmouth, N.J. and Redstone Arsenal, Huntsville, Ala. In the summers of 1953 and 1954 he was employed as a Junior engineer by the Allen D. Cardwell Corporation of Plainville, Conn. Mr. Inslee joined RCA Laboratories in 1955 and since that time has been primarily concerned with television pick-up and image storage for satellite systems. He is now in the Physical Research Group of the Astro-Electronics Division.

IRWIN M. KRITTMAN received the B.E.E. degree, cum laude, from the City College of New York, N.Y., in 1957. He completed his course requirements for the M.S. degree in electrical engineering at the University of Pennsylvania, Philadelphia, Pa., during 1957-1959. During the summers of 1955 and 1956, he was an engineering trainee at the Naval Research Laboratory, Washington, D.C., and the Bell Aircraft Corporation, Buffalo, N.Y., respectively. He joined RCA Laboratories, Princeton, N.J., in 1957, and transferred to the Astro Electronics Division upon its formation in 1958. Since 1958, he has been a member of the Physical Research Group and associated with various electrostatic image and signal recording projects. Mr. Krittman is a member of Tau Beta Pi, Eta Kappa Nu, and the Institute of Electrical and Electronics Engineers.

DAVID G. C. LUCK received his B.S. degree in physics from Massachusetts Institute of Technology in 1927 and his Ph.D. in 1932. In 1927-28 he was a Swope Fellow in Physics, and a Malcolm Cotton Brown Fellow in 1928-29. He was an Assistant in the Department of Physics at Massachusetts Institute of Technology in 1929-32. Dr. Luck joined the Research Division of RCA Victor Company in 1932 and remained with the Victor Division of RCA Manufacturing Company until 1941, working on pulse communication, direction finding, and radio guidance of aircraft. He was transferred to RCA Laboratories Division upon its formation in 1941 and remained a Member of the Technical Staff of the Princeton Laboratories through 1953, working on special electronic systems for military and other aircraft and on color television systems. From 1954 to 1958, he served as staff engineer of the Airborne Systems Department of RCA Defense Electronic Products, acting as a technical adviser on a number of military-system proposals and projects. During 1958 and '59, on leave from RCA, Dr. Luck participated in the work of the Advanced Research Projects Agency of the Department of Defense as a member of their technical advisory staff. He is now a member of the Technical Staff of the Advanced Military Systems operation of RCA Defense Electronic Products.

Dr. Luck is a member of the Institute of Navigation and Sigma Xi, and a Fellow of the Institute of Radio Engineers. He is a recipient of the Ballantine Medal of the Franklin Institute and of the Pioneer Award of the IRE Professional Group on Aeronautical and Navigational Electronics.

DAVID H. SAPP received the B.S. degree in Physics from Ohio State University in 1955 and the M.S. degree in Systems Engineering from the University of Pennsylvania in 1960. In 1955, he joined RCA Airborne Fire Control Engineering in Camden, N. J. where he was engaged in the development of several airborne fire-control systems and airborne missile range and angle trackers. Since 1959, when he joined Airborne Systems Engineering, he has been engaged in the analysis and design of communication and telemetry systems for airborne and space vehicles. Mr. Sapp is a member of the Institute of Electrical and Electronics Engineers, Sigma Pi Sigma, and Tau Beta Pi.

HANS J. WEISS received the degree of Dipl. Phys. (Diploma of Physics) from the Technical University of Karlsruhe, Germany, in 1954. After several years' work as a systems and design engineer with RCA Victor S/A, Sao Paulo, Brazil, he joined the RCA Surface Communications Division, Camden, in 1959. Mr. Weiss has done extensive work in the design of tropospheric scatter communication systems. His more recent efforts have been almost exclusively concerned with systems aspects of military and commercial satellite communications.