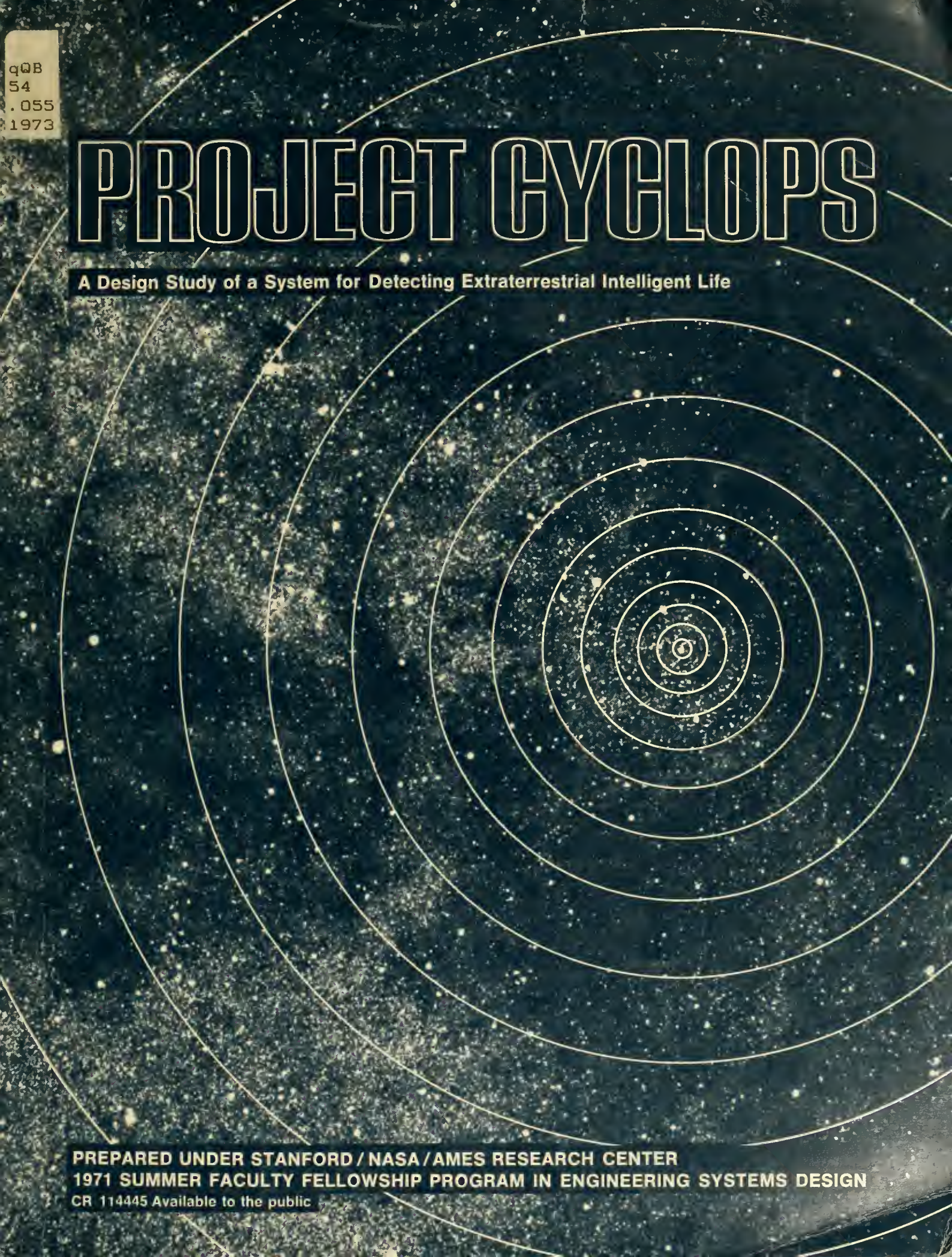
PROJECT CYCLOPS

A Design Study of a System for Detecting Extraterrestrial Intelligent Life



PREPARED UNDER STANFORD / NASA / AMES RESEARCH CENTER
1971 SUMMER FACULTY FELLOWSHIP PROGRAM IN ENGINEERING SYSTEMS DESIGN
CR 114445 Available to the public

CR 114445
(Revised Edition 7/73)
Available to the public

PROJECT CYCLOPS

A Design Study of a System for Detecting Extraterrestrial Intelligent Life

**PREPARED UNDER STANFORD / NASA / AMES RESEARCH CENTER
1971 SUMMER FACULTY FELLOWSHIP PROGRAM IN ENGINEERING SYSTEMS DESIGN**

Further copies of this report may be obtained by writing to
Dr. John Billingham
NASA/ Ames Research Center, Code LT
Moffett Field, California 94035

CYCLOPS PEOPLE

Co-Directors:

Bernard M. Oliver	Stanford University (Summer appointment)
John Billingham	Ames Research Center, NASA

System Design and Advisory Group:

James Adams	Stanford University
Edwin L. Duckworth	San Francisco City College
Charles L. Seeger	New Mexico State University
George Swenson	University of Illinois

Antenna Structures Group:

Lawrence S. Hill	California State College L.A.
John Minor	New Mexico State University
Ronald L. Sack	University of Idaho
Harvey B. Sharfstein	San Jose State College
Alan I. Soler	University of Pennsylvania

Receiver Group:

Ward J. Helms	University of Washington
William Hord	Southern Illinois University
Pierce Johnson	University of Missouri
C. Reed Predmore	Rice University

Transmission and Control Group:

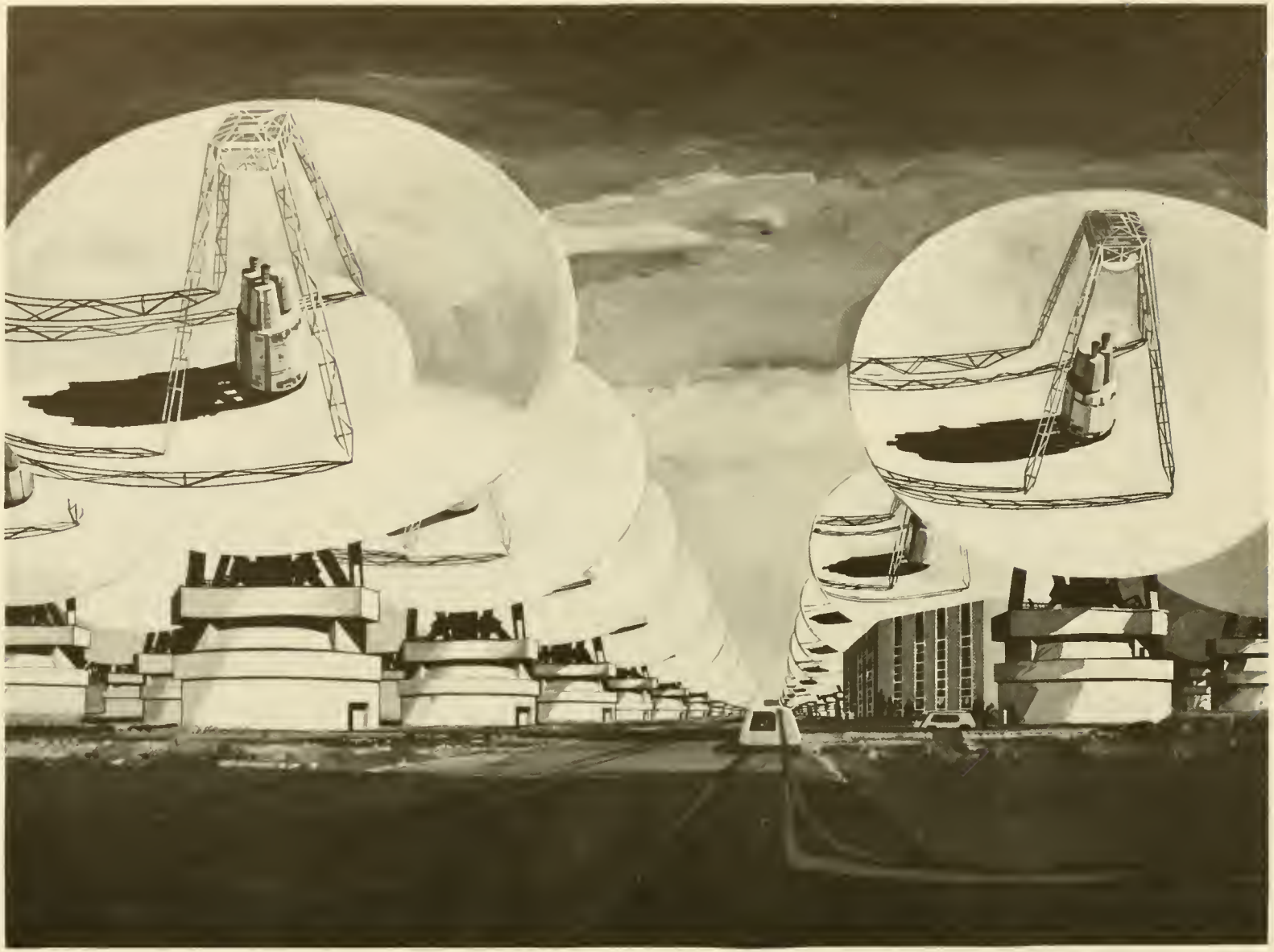
Marvin Siegel	Michigan State University
Jon A. Soper	Michigan Technological University
Henry J. Stalzer, Jr.	Cooper Union

Signal Processing Group:

Jonnie B. Bednar	University of Tulsa
Douglas B. Brumm	Michigan Technological University
James H. Cook	Cleveland State University
Robert S. Dixon	Ohio State University
Johnson Luh	Purdue University
Francis Yu	Wayne State University

At this very minute, with almost absolute certainty, radio waves sent forth by other intelligent civilizations are falling on the earth. A telescope can be built that, pointed in the right place, and tuned to the right frequency, could discover these waves. Someday, from somewhere out among the stars, will come the answers to many of the oldest, most important, and most exciting questions mankind has asked.

—Frank D. Drake,
(Intelligent Life in Space
The MacMillan Co.)



View 1. *Artist's concept of ground level view of Cyclops system antennas, showing the central control and processing building.*

FOREWORD

In addition to the findings of the Cyclops summer study, this report contains much tutorial material added for the benefit of those readers unacquainted with the subjects of exobiology and interstellar communication. This introductory material is found in Chapters 2, 3, and 4 and parts of Chapters 5 and 6. These chapters may be omitted or skimmed by readers already familiar with the subject. Parts of Chapters 5 and 6 together with Chapters 8 through 14 contain the main results of the summer study. Chapter 7 is a condensed description of the Cyclops system, while Chapter 15 presents some conclusions and recommendations that are the consensus of the System Design Group but do not necessarily reflect the *unanimous* opinion of all the fellows.

In my opinion, the major contributions of this study are:

1. The reaffirmation of the microwave window as the most suitable part of the spectrum for interstellar communication and search.
2. The identification of a particularly likely band in the microwave window.
3. The design of a data processing system capable of simultaneous search of at least half of this band and the detection of coherent signals 90 dB below the total noise power in the band.
4. The design of an improved local oscillator distribution system, an IF transmission and delay system, and low-noise remotely tunable receivers that appear to guarantee the feasibility and practicality of an array with several square miles of collecting area.
5. The identification and estimation of the major costs of a search system and program.

Taken together, these contributions indicate that we now have the technological capability of mounting an effective search for extraterrestrial intelligent life. Before any major commitment is made in this direction a great deal of additional study should be devoted to all aspects of the problem. It is hoped that this report can serve as a basis for that further study.

Although the editing of some sections and the writing of other sections of this report have consumed all my spare time for several months since the summer program formally ended last August, I have found the effort both fascinating and challenging. I only regret that I did not have the time nor skill to do a better job, and I must ask the reader to be forgiving of inconsistencies, poor organization, and errors that have eluded me.

Finally I want to express my gratitude to Hans Mark, Director of Ames Research Center and to John Billingham, Chief of the Biotechnology Division, for having conceived the idea of a summer engineering systems design study on this subject, and for having invited me to participate. I am also grateful to Hewlett-Packard for granting me a leave of absence and for excusing many hours devoted to the program both before and after. The summer experience was in itself exciting and rewarding, but if the Cyclops report stimulates further study that results in a full-scale search, I shall consider this to have been the most important year of my life.

Bernard M. Oliver

ACKNOWLEDGMENTS

The Cyclops study was greatly assisted and enriched by the informative and often inspirational contributions of a large number of people over and above those listed as participants. The opening lecture series, listed below, was particularly valuable in stimulating the group and acquainting the Fellows with the latest developments in the technologies involved in large array design.

PROJECT CYCLOPS SEMINAR SERIES:

What Cyclops Might See—*Philip Morrison, Massachusetts Institute of Technology*

The Cyclops Concept—*B.M. Oliver, Hewlett-Packard*

Design Constraints on Angle Pointing and Tracking Performance of Large Antennas—*R.J. Wallace, Jet Propulsion Laboratory*

Large Orbital Antennas with Filled Apertures—*H. Schuerch, Astro Research Corporation*

Electronics System for a Very Large Array—*Sander Weinreb, National Radio Astronomy Observatory*

Design Concepts of the Stanford Five Element Non-Redundant Array—*Ronald Bracewell, Radio Astronomy Institute, Stanford University*

Radio Astronomy Instrumentation Needs—*David S. Heeschen, National Radio Astronomy Observatory*

Some New Design Principles for Large Radio Telescopes—*S. von Hoerner, National Radio Astronomy Observatory*

Large Ground Antennas for Space Communication—*Earl Lewis, Philco-Ford Corporation*

Large Antenna Structures—*William D. Merrick, Jet Propulsion Laboratory*

Radio Astronomy in NASA—*William E. Brunk, National Aeronautics and Space Administration*

Low Noise Microwave Receivers—*Walter H. Higa, Jet Propulsion Laboratory*

Signal Processing in Radar Astronomy—*Richard Goldstein, Jet Propulsion Laboratory*

Advanced Radar Exploration of the Solar System—*Gordon H. Pettingill, Massachusetts Institute of Technology*

Microwave Ground Antennas for High Efficiency Low Noise Performance—*D.A. Bathker, Jet Propulsion Laboratory*

Array Configurations for Gathering and Processing Data—*Joseph Goodman, Stanford University*

Radio Astronomical and Cosmological Value of a Very Large Array—*Martin Rees, Cambridge University*

In addition to the scheduled series, the group heard talks by S.M. Katow of Jet Propulsion Laboratories and Herbert Weiss of Lincoln Laboratories. These contributed further valuable inputs on antenna design. Many private firms contributed consulting help. In particular Robert Markevitch of the Ampex Corporation and Robert Hall of the Rohr Corporation were of great assistance.

The Cyclops team wishes to acknowledge this support and to thank all those who participated for their professional help.

In addition we would be remiss if we did not acknowledge the hours of overtime work by Marilyn Chinn and Lorraine Welk who typed the final and earlier versions of this report. Any future job they may have will seem peaceful and simple by comparison. Finally we wish to thank the graphics department at Ames for the skillful rendering of this report and for the innumerable slides and drawings they prepared for earlier drafts and for our final briefing session.

TABLE OF CONTENTS

APPENDIX

CYCLOPS SYSTEM

1. INTRODUCTION

Each summer for the last five years the National Aeronautics and Space Administration (NASA) has funded eight faculty fellowship programs in cooperation with the American Society for Engineering Education (ASEE). The selection of four sites is governed by the proximity of a NASA facility to a university. Two programs are conducted at each site—one research oriented and the other engineering-design oriented. Project Cyclops is the 1971 NASA-ASEE Summer faculty Fellowship Program in Engineering Systems Design conducted jointly by Stanford University and Ames Research Center.

A major objective of these programs is to provide an educational experience for young faculty members of American universities. The Engineering Systems Design programs, for example, usually involve several disciplines, each represented by individuals involved in the program. The interaction of the disciplines, as well as the tradeoffs available for optimizing the overall system, becomes apparent. Each specialist inevitably learns something about the other fields involved in the total system design.

While the overall objectives of the summer programs are educational, it is important that any system studied have a clearly stated objective or set of objectives. The primary objective of Project Cyclops was:

To assess what would be required in hardware, manpower, time and funding to mount a realistic effort, using present (or near-term future) state-of-the-art techniques, aimed at detecting the existence of extraterrestrial (extrasolar system) intelligent life.

By a "realistic effort" is meant one having a high probability of success, based on our best estimates of all the factors involved. Because our present knowledge does not allow accurate estimates of the density of intelligent and especially of communicative life in the universe and therefore of the volume of space that must be searched, the Cyclops system design must take this uncertainty as to the magnitude of the problem into account at the outset. This suggests starting with the minimum system that would probably be needed and increasing its size till success is achieved or until a new and superior technique is discovered. Cyclops is therefore not a fixed static system but a flexible expandable one.

This report presents the conclusions reached in this summer study including a possible system design (with some alternatives) and some cost estimates. However, it should be emphasized that, because of time and manpower limitations, this study represents only about two to three man-years of effort. Before cost estimates can be made that are accurate enough to permit a go-no-go decision, at least ten times this effort will be needed. And before a final detailed design can be specified at least a thousand times the present Cyclops effort will have been expanded. Thus the present report must be taken as only a very preliminary and rough set of estimates.

Although the Cyclops system as here described would have certain important applications in both radio astronomy and communication with deep space probes both the extremely preliminary nature of this study and the great uncertainty as to whether a system of this type will be built in the reasonably near future force us to emphasize that presently contemplated expenditures in radio astronomy, radar mapping and deep space communication facilities should be evaluated without regard to the content of this report. It would be extremely unfortunate if funding for any worthwhile projects in these fields were to be deferred or reduced because of a mistaken impression regarding the imminence of a Cyclops-like system. Rather we would hope that the fact that many people are willing to consider the magnitude of expenditures involved in Cyclops might serve to emphasize how greatly under financed radio astronomy, for example, is at the present time.

2. LIFE IN THE UNIVERSE

"Absence of evidence is not evidence of absence."
Martin Rees

It is only recently that speculation about the existence of intelligent extraterrestrial life has turned into serious scientific study. As recently as 40 years ago, almost all scientists, speaking *ex cathedra,* would have argued that life, if not unique to earth, was at least exceedingly rare. The last 15 years have brought about an almost complete change of opinion. Today a large segment of the scientific community is willing to accept the beliefs that life is common in the universe, that many civilizations have existed and still exist in our Galaxy, and that many may be more advanced than our own. Indeed, the debate is now concerned primarily with how best to discover these civilizations rather than with the plausibility of their existence.

We still do not have "hard" scientific evidence for the existence of intelligent extraterrestrial life, evidence amounting to proof--evidence in the sense Rees uses the word in the quotation above. In fact, it is unlikely that we will ever have this kind of evidence unless and until we actually search for and succeed in contacting other life. To justify such an effort, which may require billions of dollars and decades of time, we must truly believe that other intelligent life exists and that contact with it would be enormously stimulating and beneficial to mankind. What is it that gives us such faith, and the audacity to suggest such an undertaking? What has caused such a reversal of conventional scientific thought over the last few years?

CONVERGING DISCIPLINES

Historically, science has been concerned with the gathering of data, with careful measurement and analysis, and with the construction of models in a number of seemingly unrelated disciplines. Little connection was seen, for example, between astrophysics and biology or between these and communication theory. The last two decades have witnessed a synthesis of ideas and discoveries from previously distinct disciplines. Out of this synthesis have grown new fields of research. Thus we now have exobiology, which represents a synthesis of discoveries in astronomy, atmospheric physics, geophysics, geochemistry, chemical evolution, and biochemistry.

This burst of interdisciplinary thought has produced an understanding of life in the universe based, for the first time, on scientific information. To be sure, the picture is incomplete; many details are missing but the broad picture is emerging. It is as if science, having spent three centuries doing its humble and arduous homework, had finally become competent to address itself to the most fundamental questions of all, the questions asked by children and philosophers: How did it all begin? What am I? *Quo vadimus*?

The picture we behold is one of *cosmic evolution:* a universe that was born, has evolved into galaxies of stars with planets, and into living things, and that will continue to evolve for aeons before vanishing into oblivion or collapsing toward rebirth. We owe our very existence to strange and wonderful processes that are an inherent part of this cosmic evolution. For example, we know that heavy elements were not present in the beginning but are formed in the fiery interiors of stars, many of which die in a titanic explosion. The dust of the heavy elements hurled into space becomes the raw material from which new stars with earthlike planets can form. The calcium in our very bones was manufactured in the cores of countless long dead stars not unlike those that sparkle in our night sky.

The ability of intelligent life to contemplate its own existence has always fascinated the philosopher. Now we see the even more fascinating picture of the entire universe contemplating itself through the minds and eyes

of the living beings evolved by the universe itself.

Out of the findings of the sciences that have converged to give us our present picture of cosmic evolution we draw certain conclusions that are the premises of a plausibility argument for the prevalence of technologically advanced extraterrestrial life.

1. Planetary systems are the rule rather than the exception. Our understanding of the process of star formation leads us to expect planetary systems around most stars, and to expect a favorably situated planet in most systems. There are therefore probably on the order of 10^{10} potential life sites in the Galaxy.
2. The origin and early evolution of life on Earth is apparently explicable in terms of the basic laws of physics and chemistry operating in the primitive terrestrial environment.
3. The laws of physics and chemistry apply throughout the universe. Moreover, the composition of the primordial material, from which life on Earth arose in accordance with these laws, is commonly repeated elsewhere.
4. The factors causing natural selection, which led to the evolution of more complex species and ultimately to intelligent life on Earth, may reasonably be expected to be present on any planet on which life originated.
5. Intelligence, if it develops, is generally believed to confer survival value and therefore to be favored in natural selection. On Earth, intelligence led to attempts to modify and utilize the environment, which were the beginnings of technology. This course of development seems a highly probable one on any earthlike planet.

It will be seen that in their totality these premises argue that the Galaxy contains a tremendous number of planets on which life could start, and evolve, and that there is nothing special about Earth to favor it over the others. Sagan refers to this as the "assumption of mediocrity." It is important to note that the assumption of mediocrity does not imply that living systems elsewhere will have compositions and forms identical to those of Earth. Indeed, such a coincidence would be too improbable to expect. Nor does it imply that life began on *every* suitable planet, or that it evolved *inevitably* to possess intelligence and technological prowess. It does imply that the basic processes of stellar, chemical, biological, and cultural evolution are universal and, when carried to fruition, lead to technologies that must have close similarities to ours today and in the future.

Regardless of the morphology of other intelligent beings, their microscopes, telescopes, communication systems, and power plants must have been at some time in their history, almost indistinguishable in working principles from ours. To be sure there will be differences in the order of invention and application of techniques and machines, but technological systems are shaped more by the physical laws of optics, thermodynamics, electromagnetics, or atomic reactions on which they are based, than by the nature of the beings that design them. A consequence of this is that we need not worry much over the problem of exchanging information between biological forms of separate origin. For we share with any intelligence we might contact a large base of technology, scientific knowledge, and mathematics that can be used to build up a language for conveying subtler concepts.

Finally it is important to note that the premises listed above, while they do not exclude totally different biological chemistries, such as ammonia-based life, do not require these for life to be common in the universe. If we can become convinced that carbon- and water-based life is to be expected, anything further is a bonus.

Cocconi and Morrison pioneered the new era of serious scientific study of the possibility of interstellar communication with their 1959 paper in *Nature* (ref. 1). In that paper we find many of the arguments still thought to be central in the quest for signals originated by other intelligent life. Within a few years several other papers appeared by Bracewell, Drake, Dyson, Handelsman, Huang, Oliver, Sagan, Shklovskii, and others. These were followed by five major works on interstellar communication: Cameron's 1963 collection of original papers (ref. 2), Sagan and Shklovskii's fundamental and lucidly written review of 1966 (ref. 3), Dole's 1964 (revised 1970) study of habitable planets (ref. 4), and two Russian contributions—a 1964 conference report (ref. 5) and Kaplan's 1969 compendium (ref. 6). Drake (ref. 7) and Sullivan (ref. 8) have written excellent popular books on the subject. The former is a compact summary for students, the latter a comprehensive review addressed to the educated layman. A very extensive bibliography is also available (ref. 9).

It is perhaps significant to note that while the Cyclops study examined many advances in technology that have occurred in the interim, the basic approach selected does not differ essentially from that proposed by Cocconi and Morrison twelve years ago, and espoused by Drake. The intervening years have greatly increased the search capability of feasible systems, but have not altered the part of the spectrum to be used. We do not feel this represents a lack of imagination in the present study, but rather a validation of the original thinking on the subject. There is also an implication that this

thinking has attained some status of stability, if not maturity. Some competing viewpoints are examined in Appendix B.

As a prelude to a detailed examination of the Cyclops system and search strategy we feel compelled to offer the reader a fuller picture of the story of cosmic evolution, and of the uncertainties that remain at each step. Since our case rests on a plausibility argument, the reader should weigh the evidence and reach his own conclusions as to the merits. In the following sections we have tried to summarize present thinking about the origin of the universe and matter, the evolution of galaxies and stellar populations, and the formation of planetary systems. From this point on our narrative necessarily becomes geocentric, for we have no evidence about the evolution of life elsewhere. However, by examining how life evolved on Earth we can decide whether the factors that caused it are in any way peculiar to this planet.

ORIGIN AND EVOLUTION OF MATTER

All stars are suns. Some are larger and some smaller than the star that warms this earth; some are younger, but most are older. Stars are not distributed uniformly or randomly throughout space, but instead occur in huge aggregations called galaxies; some of which show a spiral form, others an ellipsoidal shape. Our Galaxy, the Milky Way, is a spiral galaxy containing several hundred billion stars. The universe contains over a billion galaxies, or, in all, more stars than there are grains of sand on all the beaches of Earth.

Any theory of cosmology must account for the observed fact that the galaxies (and particularly clusters of galaxies) appear to be receding from one another with a velocity proportional to their separation. We are not at the center of this expansion any more than any other galaxy is; rather we consider that all observers anywhere in the universe would see the same recession of distant galaxies. This assumption, known as the "cosmological principle" implies space curvature. There is no center of the universe and therefore no outer limit or surface.

Twentieth century cosmology has been concerned primarily with two diametrically opposed views: instantaneous creation and continuous creation, both of which account for the expansion in different ways. According to the instantaneous creation view, now known as the "big bang" cosmology, the universe began as an awesome primordial fireball of pure radiation. As the fireball expanded and cooled, pair production yielded the fundamental nuclear particles of matter and antimatter and thermonuclear reactions produced helium nuclei. Still further expansion dropped the temperature to the point where hydrogen and helium atoms formed by combination of the electrons with the protons and helium nucleii, but elements heavier than helium were not produced in appreciable quantities. During the early phase of the expansion, when the ionized matter was strongly coupled to the radiation field, the matter distribution was controlled by the radiation field. Only later, when the radiation density dropped below the matter density (related by $E = mc^2$) and the matter deionized, could gravity forces act to enhance any nonuniformities of density that may have existed and thus begin the hierarchy of condensations that resulted in galaxies, stars, planets, and satellites.

The continuous creation or "steady state" theory proposed by Hoyle applies the cosmological principle to time as well as to space, thereby making it a "perfect" cosmological principle. It assumes that the universe not only appears the same to observers anywhere, but also at any time. As the galaxies recede, the steady state theory maintains the average density by having neutrons appear spontaneously in space. These then decay to protons and electrons and form new hydrogen out of which new galaxies condense. Thus, on the average, the density of galaxies in space remains constant in spite of their recession.

In addition to being consistent with the observed expansion of the universe, the big bang theory makes several predictions, such as (1) the primordial and most abundant elements should be hydrogen and helium in the (mass) ratio of about 3 to 1; and (2) there should be an isotropic background radiation corresponding to a black body at about 3° K. Findings in the last decade seem to have established the validity of both (1) and (2), which are hard to explain with steady state cosmology. A great deal of other data also conflict with the steady state hypothesis, and current opinion has swung heavily toward big bang cosmology.

While the big bang theory answers many questions, others still remain to be answered, among them:

1. What, if anything, preceded the initial fireball?
2. Where is the antimatter half of the universe?
3. How did galactic clusters and galaxies form out of the supposedly homogeneously dense fireball in which gravitational clumping was suppressed by radiation coupling?

A partial answer to (1) may found if future observations (or past observations handed down to us over billions of years by interstellar communication) show that the present expansion is slowing down rapidly enough to turn into a contraction aeons hence. We can then contemplate a cyclic universe, which defeats the second law of thermodynamics by being reborn. If the expan-

Figure 2-1. M*31, the great galaxy in Andromeda. (Photo courtesy Lick Observatory)*

sion is not going to stop, we must view the creation of the present universe as a unique event.

Questions (2) and (3) may be resolved by further study over the next few years. Although the coupling of radiation to the plasma of the fireball would prevent gravitational clumping, it might be that local regions dominated alternately by matter and antimatter could form. Pairs formed by radiation in the no man's land between these regions would be more likely to be accepted by the regions and increase their density, if of the dominant matter type, or to be annihilated and reconverted to radiation if of the minority type. Conceivably, this could provide an amplification process capable of developing millions of matter-antimatter regions out of the initially small statistical fluctuations in matter-antimatter preponderance. If so, the matter and antimatter in equal amounts could have emerged from the fireball state in separate blobs, already clumped to form galactic clusters and galaxies. The matter and antimatter between blobs would continue to annihilate into radiation, clearing out what was to become intergalactic space until the drop in temperature, the large distances, and the action of gravity prevented further collision.

We simply do not know the answer to these questions yet. We do know that galaxies formed, and that much of their matter condensed into stars. Any uncertainties we may have about prior epochs have little or no bearing on the prevalence of life in any of the billion or more galaxies that now exist.

GALACTIC EVOLUTION AND STELLAR POPULATIONS

Our Galaxy, the Milky Way, is essentially a flat disk with an oblate spheroidal nucleus. Ragged spiral arms twist around in the plane of the disk between the nucleus and the rim. The diameter of the disk is about 100,000 light-years and the thickness tapers smoothly from thick near the nucleus to thin at the rim, its thickness at the half radius point being about 1600 light-years. We believe that our Galaxy and the great nebula (or galaxy) in Andromeda, M31, shown in Figure 2-1 are very similar. The nucleus and disk are surrounded by a halo of stars and globular clusters of stars in highly elliptical orbits about the galactic center, but by far the largest part of the total mass is in the nucleus and the disk.

The Milky Way was originally spheroidal and many times its present size. Its evolution from spheroid to disk as it underwent gravitational contraction was probably due to conservation of angular momentum. We believe this pattern is typical of galaxies having a comparable amount of initial angular momentum. Most galaxies have less spin and have retained their spheroidal shape.

The history of our Galaxy spans at least 12 billion years and is traced through the study of the different chemical compositions, kinematics, and ages of the different stellar populations. Broadly speaking, there are two major populations of stars: old (population II) and young (population I). The oldest stars dominate in the halo and in the nucleus; the young stars dominate in the disk, and the very youngest stars are found in the spiral arms closely associated with the gas clouds out of which they recently formed. Elliptical galaxies seem primarily composed of old (population II) stars and do not contain the very bright young stars that mark the arms of the spiral galaxies.

The roughly spherical, slowly rotating, pregalactic clouds of hydrogen and helium collapsed under their own gravity until this centripetal force was balanced by gas pressure and the centrifugal force of the rotation. In such a gas cloud, we might expect a great deal of initial turbulence. But this turbulence would die out with time, while the systematic rotation would be maintained. The rate of star formation is expected to be proportional to the square of the gas density. Thus when the contraction had proceeded sufficiently, stars began to form and the rate of formation increased rapidly. Many of the larger stars that were formed initially went through their complete life cycle and vanished as supernovae while the Galaxy was very young and still contracting.

The observed distributions of stellar populations agree qualitatively with this general picture of contraction, decreasing turbulence, and increasing spin and flatness of the evolving galactic gas mass. The oldest stars (the smaller mass stars, which have survived) are found as expected in a spheroidal distribution, while progressively younger generations are found in increasingly flattened distributions. The youngest are confined to the present disk. The orbits of the older generations are highly eccentric and show greater velocity dispersion, reflecting the greater gas turbulence during the early phases. Young stars show much smaller velocity dispersion and quietly orbit the galactic center along with their neighbors in the disk.

Old (population II) stars have a much smaller heavy element content than young (population I) stars. By studying the elemental abundances in star clusters belonging to different age groups, we can plot the increasing abundance of heavy elements with time as shown in Figure 2-2. Although the ratio of hydrogen to helium has remained essentially constant, the heavy element abundance has risen steadily from essentially zero 12 billion years ago to about 2 percent by weight at the present time. The only known mechanism for heavy element production is thermonuclear fusion in the very hot cores of stars that have exhausted their central hydrogen. Most important as heavy element factories are the large stars that end their lives in violent supernova explosions. These explosions enrich the interstellar gas with the fusion products formed during the life of the star and with very heavy elements formed by neutron capture and β-decay during the supernova explosion it-

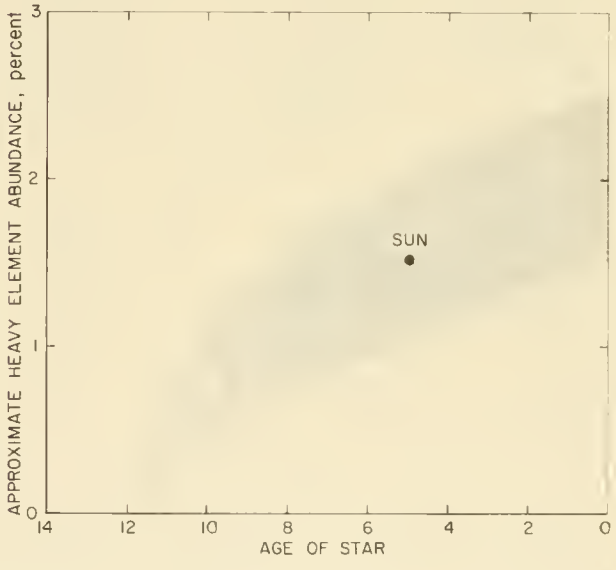


Figure 2-2. *Approximate heavy element abundance in stars of various ages.*

self. Thus, we can interpret Figure 2-2 as a curve proportional to the number of supernovae that have occurred in the Galaxy.

The curve of Figure 2-2 must be considered only approximate; many astronomers now feel the initial rise of heavy elements may have been more rapid. If so, the abundance could have reached 1 to 1-1/2 percent in the first 2 to 4 billion years. In any case, it is clear that the very oldest stars cannot have earthlike planets. Such planets could only have formed in large numbers after the first 2 to 4 billion years of galactic history, and owe their composition to earlier supernovae.

STELLAR CHARACTERISTICS AND EVOLUTION

In estimating the likelihood of life in the universe we need to know the range of stellar types that could serve as life-engendering and life-supporting suns to a favorably situated planet, and how numerous stars in this range are. The physical properties of stars vary over a wide range as shown in Table 2-1. The distribution within these limits is nonuniform.

TABLE 2-1

STELLAR CHARACTERISTICS

Parameter	Observed Limits	Ratio	Solar Value
Luminosity*	$10^{-6}L_\odot$ to $10^{5}L_\odot$	10^{11}	3.9×10^{26} watts
Surface temperature	2000°K to 60,000°+K	30+	5800°K
Mass	$0.05M_\odot$ to $100+M_\odot$	2000+	2×10^{30} kg
Age	$<10^{4}$ years to 10^{10} years	10^{6}	$\approx5\times10^{9}$ years
Heavy elements (mass)	0.01% to 5%	500	1.5%

*Luminosity is the total power radiated by a star. The subscript ⊙ refers to the Sun. See Appendix A for a glossary of astronomical terms.

Stars of small mass are much more common than stars of large mass; old stars are more common than young ones. Nor are the physical properties independent; mass and luminosity, for example, are highly correlated.

Classification of Stars

There are several ways of classifying stars. One method is by absolute visual magnitude (see Appendix A). Another is by size—that is, supergiant, giant, subgiant, or dwarf. A traditional classification that has proved very useful is by certain features of the stellar spectrum. The principal characteristics of each *spectral class* are listed in Table 2-2.

The relative intensities of the spectral lines are determined primarily by the surface temperature and only secondarily by such factors as true elemental abundance, luminosity, etc. Thus, the sequence from type *O* to type *M* is one of *decreasing surface temperature* and is accompanied by a color change from brilliant electric blue, through white, to yellow, orange, and finally dull red. Our Sun fits into the sequence about two-tenths of the way from type *G* to type *K* and is therefore known as a *G*2 (or "early" *G*) star.

Three graphical relations play an important role in understanding stellar types and stellar evolution: The *Hertzsprung-Russell (HR) diagram,* the *mass-luminosity relation,* and the *luminosity function.* Our estimates regarding the frequency of suitable planetary systems are primarily based on these three relations and the dynamics of the Galaxy.

TABLE 2-2

SPECTRUM CHARACTER BY CLASS

Spectral Class*	Spectral Characteristics
O	Very hot stars with He II absorption
B	He I Absorption; H developing later
A	Very strong H, decreasing later; Ca II increasing
F	Ca II stronger; H weaker; metals developing
G	Ca II strong; Fe and other metals shown; H weak
K	Strong metal lines; CH and CN bands developing
M	Very red, TiO bands developing

*Sequence of letters remembered by generations of beginning astronomy students by the mnemonic: "Oh, Be A Fine Girl, Kiss Me!"

HR Diagram. The HR diagram is a scattergram of stars. The ordinate is luminosity. Usually the abscissa is spectral class, but because spectral class and surface temperature are related, the abscissa can equally well be chosen as the latter. The effective temperature T_e can be defined by the relation

$$L_* = 4\pi R_*^2\, \sigma T_e^4 \qquad (1)$$

where L_* is the luminosity, R_* is the star's radius, and σ is the Stefan-Boltzmann constant. The abscissa in Figure 2-3 is chosen as T_e, and so lines of constant radius have a slope of 4 and are spaced as shown.

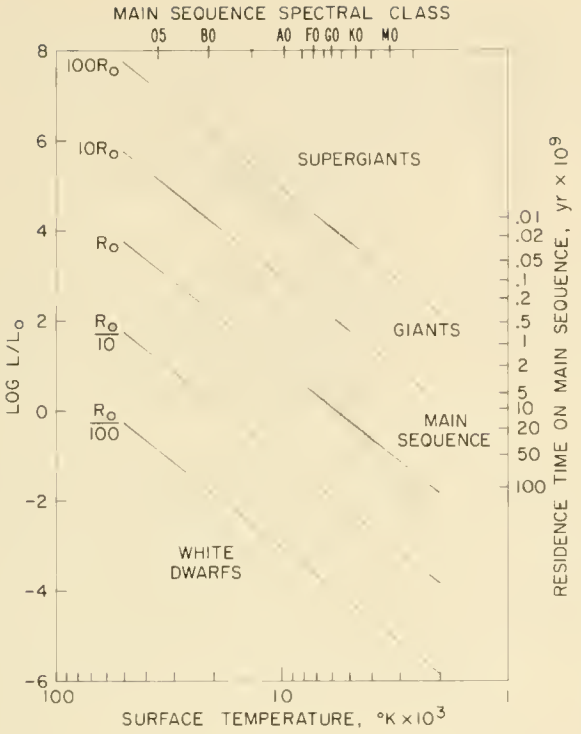


Figure 2-3. *The Hertzsprung-Russell (HR) diagram.*

The significant feature of the HR diagram is that almost all stars fall into the four shaded areas shown. This grouping indicates a restricted set of relationships between luminosity, temperature, and radius–relationships that are now understood to represent different epochs in the life of a star. The main sequence accounts for about 90 percent of all stars, giants and supergiants for about 1 percent, and white dwarfs for about 9 percent. (see Appendix A).

Mass-Luminosity Relation. The masses of binary stars can be determined by applying Newton's derivation of Kepler's Third Law relating mass, mean separation, and orbital period. If the luminosities of the stars are plotted against their masses, determined in this way, the distribution shown in Figure 2-4 is found for main sequence stars. The stars shown in this plot have the usual spread of chemical composition associated with different ages and populations. Nevertheless, the points do not depart very far from the dashed line. This indicates that the luminosity of a main sequence star is determined primarily by its mass. Although the true relation is more complicated, we can approximate it quite well over most of the range by the simple expression:

$$\frac{L_*}{L_\odot} \approx \left(\frac{M_*}{M_\odot}\right)^{3.5} . \qquad (2)$$

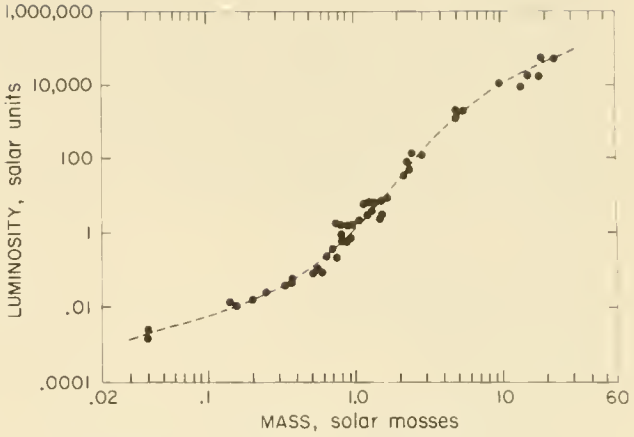


Figure 2-4. *The mass-luminosity relation.*

Luminosity Function. This is the generic term applied to the distribution of number of stars per unit volume of space versus their luminosity or absolute magnitude. The observed distribution is shown in Figure 2-5. At the high

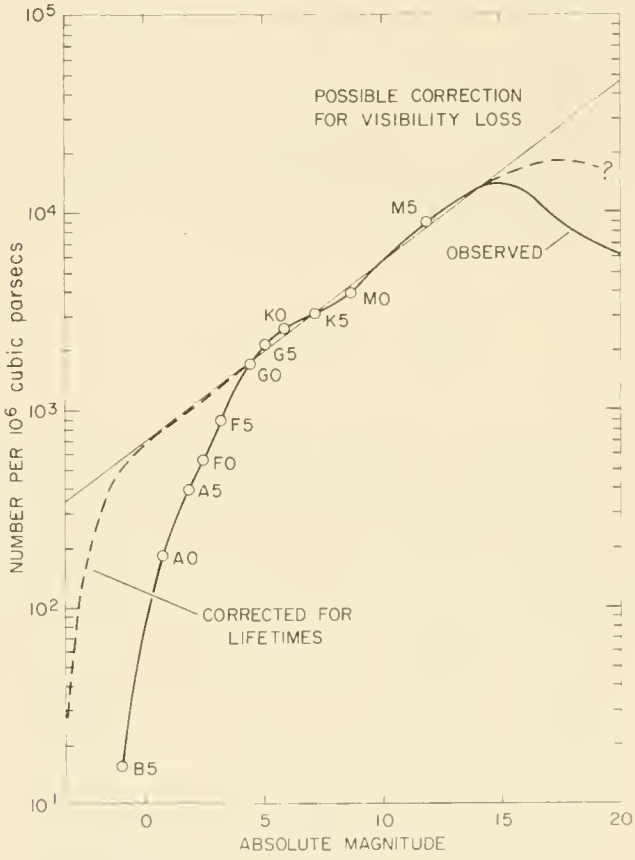


Figure 2-5. *The luminosity function.*

magnitude (low brightness) end, the observed curve falls off sooner than would be the case if the abscissa were log-luminosity. This is because, with decreasing surface temperature, the peak of the black-body radiation curve shifts below the visible part of the spectrum and brightness drops more rapidly than luminosity. A given luminosity range is thus spread over a greater magnitude range, thereby reducing the number of stars per magnitude. If the abscissa were bolometric magnitude (proportional to log-luminosity) the low brightness end falloff would be extended as indicated by the dashed curve.

If we wish to deduce the relative frequency with which stars of different masses are formed, we must also make a correction of the luminosity function at the high brightness end. *G* stars and smaller (magnitude greater than about 4.5) have lifetimes greater than the age of the Galaxy. All such stars ever formed are still on the main sequence. For larger stars, only those born not longer ago than their lifetimes will be found. Older stars will have left the main sequence thus depleting the present number. The correction factor becomes rapidly larger with increasing mass because of the rapid decrease in lifetime. When this correction is applied, the high brightness end is extended as indicated by the dashed line.

We now see that over a considerable range the corrected luminosity function can be approximated by a power law—that is, by a straight line on our logarithmic plot. Using the mass-luminosity relation (2), the corrected luminosity function over this range can be expressed in terms of mass as

$$\frac{dN}{dM} = \text{constant} \times M^{-7/3} \qquad (3)$$

This relation, which was first deduced by Salpeter in the early 1950s, is a description of how a typical gas cloud will fragment and condense into stars of various masses.

We see that larger stars are less likely to be born than smaller stars. For main sequence stars the number relative to solar *G*2 stars, as given by equation (3) is shown in Figure 2-6. Since the smallest mass that can start and sustain thermonuclear reactions is about $0.05M_\odot$ it is clear that, if equation (3) holds, the overwhelming majority of stars born onto the main sequence will be smaller than the Sun. In fact, the average stellar mass is about $0.2M_\odot$.

If we assume the frequency vs. mass relation of equation (3) to hold down to substellar masses ($<0.05M_\odot$) we might expect the galaxy to contain a very

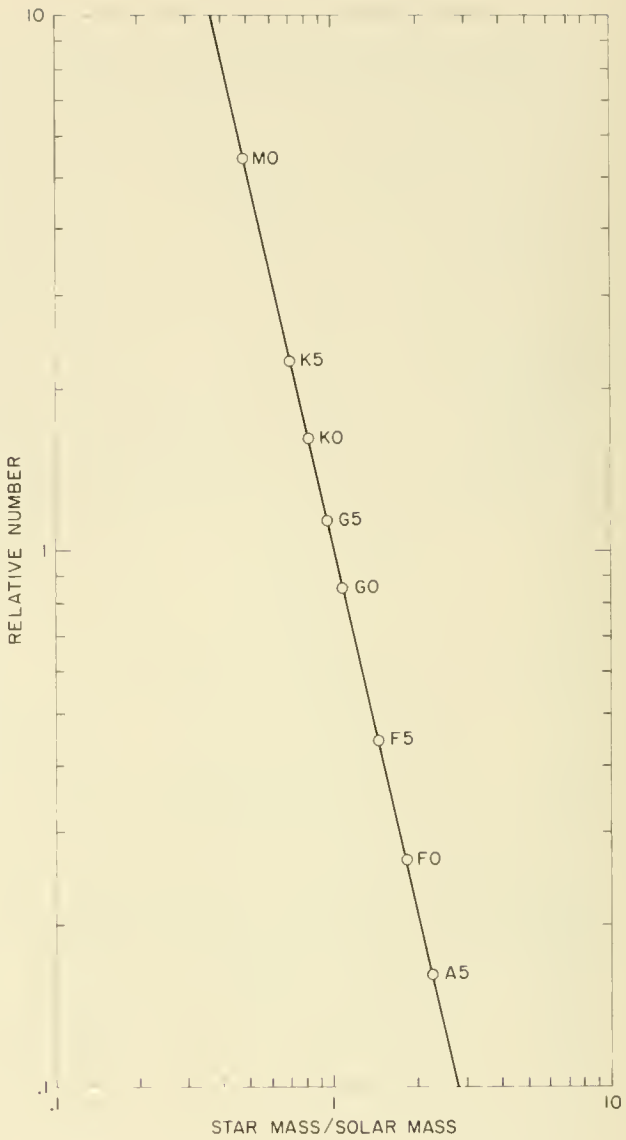


Figure 2-6. *Relative number of stars versus mass.*

large number of nonluminous bodies—stars too small to have ever ignited thermonuclear reactions. These may explain some of the "missing mass" of the Galaxy—mass not present as visible stars or interstellar dust and gas, but needed to account for the gravitational dynamics of the galaxy.

Equation (3) has also been used as an argument for expecting large numbers of planets. This represents an extrapolation of the observed luminosity function to masses less than $10^{-3}\ M_\odot$, and is questionable on these grounds alone. There well may be a lower mass cutoff below which gravitational forces cannot overcome the dispersive effects of turbulence in a gas mass. Further, if the present theories of planetary formation are correct,

planets are formed in a second fragmentation phase, after the globule that was to become the star and its planets had already fragmented from the main cloud and had contracted by several orders of magnitude. We see no reason for the statistics of the primary fragmentation to carry over into this later phase.

Stellar Evolution

Thanks to advances in nuclear physics we now have a detailed understanding of stellar evolution. Stars in the different regions of the HR diagram are stars in different phases of their life histories. We may divide stellar evolution into four principal phases:

Birth. The star begins as one of many globules about a light-year in diameter into which a larger gas cloud has fragmented. The globule contracts under its own gravity compressing the gas. The star is born when the gas becomes heated to incandescence. The luminosity steadily increases as gravitational potential energy is converted into heat. The contraction phase is very short—well under 1 percent of the main sequence lifetime.

Main Sequence. When the internal temperature has become high enough to initiate proton-proton fusion or a carbon-nitrogen-oxygen cycle (or both), the contraction stops. The temperature needed in the core to maintain hydrostatic equilibrium and to supply the radiation losses is now obtained from nuclear energy. The star has now taken its place on the main sequence at a point determined by its mass. The lifetime on the main sequence is proportional to the amount of nuclear fuel (i.e., to the mass) and inversely proportional to the power radiated (i.e., to the luminosity). The Sun's lifetime is about 12 billion years so the residence time of any star on the main sequence is approximately:

$$t_{ms} \approx (12\times10^9)\frac{M_*}{M_\odot}\frac{L_\odot}{L_*} \qquad (4)$$

In view of the mass-luminosity relation (2) this can be written

$$t_{ms} \approx (12\times10^9)\left(\frac{M_\odot}{M_*}\right)^{5/2} = (12\times10^9)\left(\frac{L_\odot}{L_*}\right)^{5/7} \qquad (5)$$

The larger the star, the shorter its life. Figure 2-7 is a plot of (5) in the spectral range of interest to us. We see from Figures 2-6 and 2-7 that the vast majority of stars live long enough for intelligent life to evolve if the evolution rates are comparable to that on Earth.

As the hydrogen fusion continues, a growing core of almost pure helium is produced, inside which all energy release has stopped. When this core reaches about one-tenth solar mass it collapses, releasing a large amount of gravitational energy in itself and in the surrounding shell where hydrogen burning is still occurring. The increased temperature increases the hydrogen fusion rate with the result that the outer layers of the star expand to absorb and eventually radiate the increased energy output. The expansion increases the surface area so much that the surface cools even though the total luminosity is greater than in the main sequence phase.

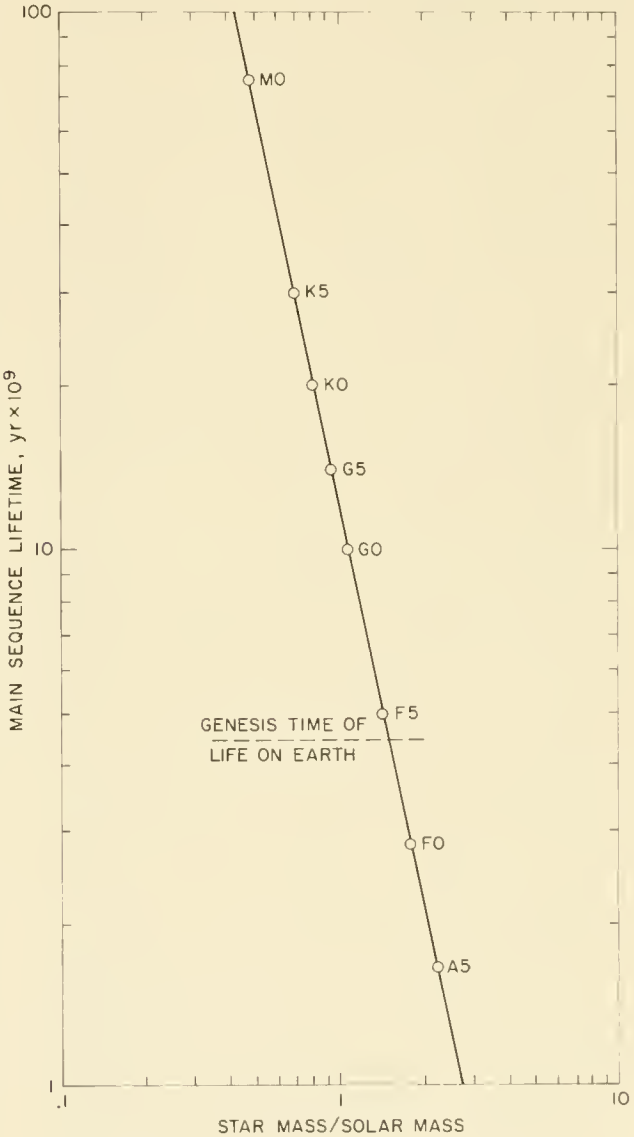


Figure 2-7. *Stellar lifetimes versus mass.*

Red Giant. The transition from main sequence star to red giant is very short, as evidenced by the scarcity of stars observed in this phase. Hydrogen burning in the shell stabilizes the star as a red giant for a time, but as the shell grows the rate of burning drops and the star again contracts. This time the contraction raises the temperature in the core to the point where helium fusion can commence and the star, now obtaining energy from both hydrogen and helium fusion, again expands. Eventually, a core of elements heavier than helium forms and the hydrogen and helium burning shells approach each other. Here the interaction becomes complicated and computer modeling has not as yet been successful beyond this stage.

Advanced Age. When gravitational contraction can no longer increase the core temperature and density sufficiently to initiate the next series of reactions, the star evolves away from the red giant region. Its path is to the left in the HR diagram of Figure 2-3 at roughly constant luminosity. During this time, the star may enter a pulsating phase, return briefly to the red giant regime, then conclude by slowly or violently ejecting its surface layers. Most frequently, the end result is a white dwarf. It is a configuration in which the material is almost completely degenerate from center to surface. The physical size of these objects is about the same as the Earth, but their mass is between 0.2 and 1.2 solar masses. The surface temperature is typically 6,000 to 12,000° K. Because of the small radius, the luminosity is very low compared with the Sun, even though the temperature is higher [see eq. (1)].

A more violent, explosive end for a star is called a supernova (L_{max} typically 10^8 $L_\odot$) and produces large quantities of heavy elements. These elements are ejected with velocities of roughly 10^4 km/sec and so soon become well mixed with the interstellar medium. The star may not always be entirely destroyed by the process: At the center of the expanding clouds, rotating neutron stars (pulsars) have been discovered. Such explosive fates await stars of 10 solar masses or larger. They are the primary source of elements heavier than neon and provide a large fraction of the elements between helium and neon.

Clearly when a star leaves the main sequence the effects on its planets are catastrophic. It is doubtful if even *very* intelligent life could remain in the planetary system and avert disaster. (Our Sun will not enter the giant phase before approximately seven billion more years.) So far as intelligent life is concerned we should therefore confine our attention to main sequence stars, and probably only to those cooler than *F*0. When we come to consider atmospheric evolution and tidal effects we will want to exclude *M* stars. Thus, the stars of greatest interest are *F*, *G*, and *K* main sequence stars, which amount to about 25 percent of all stars. If we exclude binary and multiple star systems, because of unstable planetary orbits, this still leaves over 20 billion stars in the galaxy as possible suns for intelligent life.

The Solar Neighborhood

Is ours in any way a special neighborhood? Photographs of spiral galaxies give the impression that most stars are concentrated in the nucleus and along the spiral arms. Our Sun is located midway between two spiral arms—actually, on the fringe of a spur extending off one arm. We are about 33,000 light-years from the galactic center or two-thirds of the way to the rim. The disk is about 1500 light-years thick at this radius, and the Sun is only a few tens of light-years off the midplane. At first glance, it might appear that we are in a somewhat isolated neighborhood but this is not the case.

It is true that the star density in the nucleus is extremely high—a hundred or more times that in the solar neighborhood. This means that the average separation between stars is only 20 to 25 percent as great as here. But only 10 percent of the total stellar mass is in the nucleus.

Contrary to common belief, the spiral arms are *not* regions of higher star density. They are regions of high gas and dust density, the maternity wards of the Galaxy, where new stars are being born at a higher rate. The gas clouds fragment to form clusters of hundreds of thousands of new stars including a large number of short-lived *O* and *B* stars. These stars die before their motions carry them very far from their birthplace. It is these very bright stars scattered along the spiral arms that give the arms their extra brightness.

The rotation about the galactic center of the stars in the disk is quite different from that of the spiral arm pattern. The Sun, along with most other stars in our neighborhood, circles the galactic center once each 240 million years and in one orbit will cross all components of the spiral arms. Stars born in the spiral arms drift out of the arms and, because of their velocity dispersion, become thoroughly mixed with older stars. While we can find the common age of stars in a cluster by noting the mass and hence the age of those leaving the main sequence, we can only estimate the general population class of isolated stars. Thus, if we should decide that 4 billion years were needed to evolve a technological civilization, we have no way of determining whether a particular population I star is too young or too old. The thorough mixing of stars with time means that there is

nothing special about the solar neighborhood. Ours is as good as any other in the disk, and conversely. Furthermore, the disk is the typical environment of the Galaxy.

FORMATION AND EVOLUTION OF PLANETARY SYSTEMS

One of the most striking features of the solar system is the extreme orderliness of the motions of the planets and satellites. The orbits of the planets are all nearly circular and, except for far-out Pluto and, to a lesser extent, near-in Mercury, lie very nearly in a common plane (see Appendix A). This plane makes only a small angle with the plane of the Sun's equator. All the planets revolve in the direction of the Sun's rotation. Within the solar system are two well-developed "subsolar systems": the satellite systems of Jupiter and Saturn. The major satellites of these two planets also exhibit the same regularity of motion with respect to their primaries as the planets have with respect to the Sun.

These facts strongly suggest that the collapsing gas cloud, which was to become the Sun, contracted to an initial disk shape with a concentrated center just as was the case for the Galaxy as a whole. In the galactic cloud about a trillion local condensations occurred to form stars. In the Sun's disk, because of the vastly smaller scale, only a few local condensations occurred to form planets, and the larger of these repeated the process to form satellites. Thus, although we have only one planetary system to study, we see certain aspects of its formation process repeated in miniature within it, and other aspects repeated on a much larger scale for the Galaxy as a whole.

Nebular vs. Catastrophic Theories of Planetary Formation

The earliest theories of planetary formation were the nebular hypotheses of Kant and LaPlace. These pictured a large cloud condensing into a disk much as we have indicated. These theories were brought into question in the mid-1800s by the observation that the Sun, which contains the overwhelming majority of the total solar system mass, has only a small fraction of the angular momentum (see Appendix A). Since no way was known for the Sun to get rid of angular momentum, this disparity constituted a real objection to any condensation, or *nebular* theory.

Thus the question was inverted: If the Sun, however it formed, had so little angular momentum, how did the planets acquire so much? To explain this excess, various *catastrophic* theories of planetary formation were devised. There were two major versions. In one version, an original binary companion of the Sun destroys itself (or is destroyed by collision with a third body) leaving debris, which condenses into the planets. In another version, a wandering star brushes past the Sun at such close range that tidal forces draw out filaments of matter, which separate into globules and cool to form the planets.

The catastrophic theories, popular in the early part of this century, failed for a variety of reasons. Self-destruction of a star produces a nova or supernova in which much of the matter is cast into space in excess of escape velocity and the core remains behind as a white dwarf, or a neutron star. The violence of a collision with a third body would scatter matter in all directions and would hardly produce coplanar and circular planetary orbits. The same objection applies to the globules condensed from tidal filaments. These could hardly all have had exactly the right initial velocities to produce the observed nearly circular orbits of the planets. The *coup de grace* was Spitzer's demonstration, in the late 1930s, that hot gases extracted from the Sun in a tidal filament would explode rather than condense into planets.

The implication of all catastrophic theories is that planetary systems must be extremely rare. Stars are so widely separated, and their orbital motions around the galactic center are, for the majority, so nearly concentric, that close encounters might be expected to occur only a few tens of times in the entire history of the Galaxy. Thus in the early twentieth century astronomers would have rejected the prevalence of extraterrestrial life solely on the basis of the scarcity of planetary systems.

The last 30 years have seen a return to the nebular hypothesis, with the newly discovered principles of magnetohydrodynamics playing a key role in solving the angular momentum problem. We should be clear that the problem is not, as the catastrophic theories supposed, to explain any angular momentum excess of the planets, but rather to explain the angular momentum *deficiency* of the Sun. The gas globule that fragmented out of the galactic gas cloud to become the Sun was on the order of one light-year across. If it merely shared the general rotation rate of the Galaxy its period of rotation would be about 240 million years. After contracting to a diameter of 10^{-3} light-years, which is roughly the diameter of the solar system, its rotation rate would be 10^6 times as fast, or one revolution in 240 years. This is about the orbital period of Pluto and so is an appropriate rotation rate for the planets.

At this stage the cloud would be greatly flattened by its rotation and would consist of a disk with a denser, rapidly rotating, hot central region. Irregular density distributions in the disk produced nucleation centers for the formation of planets. In the course of its contrac-

tion, as the central part of the cloud become ionized by heating, the galactic magnetic field was trapped by the conducting gas and compressed along with the matter. Thus, we visualize a disk with a slowly rotating cold rim about to condense into the gas-giant planets, and progressively more and more rapidly rotating inner parts until we come to the hot ionized central region. Our problem is to slow down the rapid central rotation so that this part may shrink to become a spherical star.

Several mechanisms have been proposed. Simple viscosity is one. Because of the velocity gradient with radius, collisions between particles will have the overall effect of slowing down the inner ones and speeding up the outer ones. This mechanism does not appear adequate. If one invokes large-scale turbulences, the effect can be increased but other difficulties arise. We now know that any ionized gas in the inner part of the disk would be trapped by, and forced to move along, rather than across, the flux lines of the central rotating magnetic field. This constitutes a powerful angular momentum transfer mechanism. Matter is spun out of the central region picking up angular momentum in the process and carrying it away from the Sun. It now appears that the Sun itself could have ejected much of this material.

Certain young stars such as T-Tauri, which (from their positions on the HR diagram) appear to be entering their main sequence phase, are observed to be ejecting large amounts of matter. Violent flares occur that can be detected even on present radio telescopes. The Sun is now believed to have ejected a substantial fraction of its initial mass in this fashion. This emission would easily have been enough to slow down its rotation and to dissipate most of the gases in the disk out to the orbit of Mars or beyond. We note that the Sun *still* emits matter at a greatly reduced rate, in the form of the solar wind.

In the outer regions of the disk most of the primordial gases of the disk probably went into forming the dense atmospheres of the Jovian planets, with their high abundance of hydrogen and helium. In the inner part of the disk it is not clear whether the terrestrial planets formed early atmospheres that were then largely dissipated by the Sun in a "T-Tauri" stage, or whether most of the nebular gases had already been dissipated before these planets formed.

The formation of the disk and its evolution into planets is obviously a very complicated problem in statistical mechanics involving the interplay of just about every known property of matter and radiation. Very likely the exact sequence of events will not be known until the whole problem can be modeled on a large computer and the evolution followed step-by-step. When we consider that only one seven-hundredth of the mass of the solar system is outside the Sun the real mystery appears to be, not why a few planets were formed, but rather why a great deal more debris was not left behind. How, in fact, could the cleanup process have been so efficient?

The return to the nebular hypothesis is a key factor in the developing scientific interest in extraterrestrial life. For, with this mechanism of planetary system formation, single stars *without* planets should be the exception. (We cannot say this about binary stars, but we see no reason to rule out terrestrial planets even for them, if the separation of the stars is some 10 AU or more.) Thus in the last two decades many astronomers have become convinced that there are some billion times as many potential sites for life in the Galaxy as were thought to exist earlier in this century. This opinion is not unanimous. Kumar (ref. 10) argues that the condensation process may have several outcomes and that planetary systems, while far commoner than catastrophic theories would predict, may not be ubiquitous.

Some Evidence For Other Planetary Systems

Planets around other stars have never been observed directly with telescopes. The huge brightness difference between the planet and the star, together with the close separation of the images, make such observation all but impossible. Nevertheless we do have a little evidence for other planets or planetlike objects.

Figure 2-8 shows the observed positions of Barnard's star, the nearest star after the α-Centauri system. Van de Kamp reported in 1963 that the observed wobbling could be partly accounted for by a dark companion of roughly Jupiter's size in a highly elliptical orbit. He later reported an alternate solution based on two masses in circular orbits. In 1971, Graham Suffolk and David

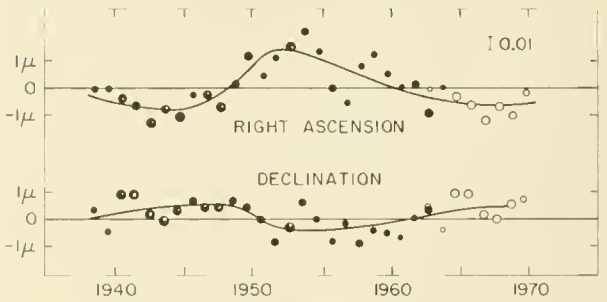


Figure 2-8. *Observed motion of Barnard's star. (Right ascension and declination are the coordinates used by astronomers to measure motion against the star background.) The error bar in the upper right hand corner indicates the average error associated with any point.*

Black at NASA's Ames Research Center reanalyzed Van de Kamp's data and found that an even better fit between the theoretical and observed wobbles could be obtained by assuming three bodies in circular orbits. Van de Kamp's two planet and Suffolk and Black's three planet solutions are given in Table 2-3.

Barnard's star is an *M*5 star with only 15 percent of the Sun's mass. It is tempting to regard Suffolk and Black's solution as evidence for a scaled-down solar system with three gas-giant planets at 1.8, 2.9, and 4.5 AU, rather than four at 5.2, 9.5, 19, and 30 AU, as in our Sun's family. However, the data do not appear to support this simple interpretation.

A single planet in orbit about a star causes the star to execute an orbit of the same shape (same eccentricity) about the common center of gravity. If the planetary orbit is circular, the observed stellar motion should be simple harmonic in both right ascension and declination, and the Fourier spectrum of both motions should therefore contain a single frequency. If the orbit is highly elliptical, the motion of the primary will occur rapidly when the two bodies are near periapsis, with relatively slow motion for the long intervening period. Thus, the spectrum of the motion will contain harmonics of the fundamental frequency of revolution.

TABLE 2-3

CHARACTERISTICS OF POSSIBLE PLANETS ORBITING BARNARD'S STAR

Planet	Distance (AU)	Mass (Jupiter = 1)	Orbit Period (yrs)	Source
*B*1	4.7	1.1	26	van de Kamp
	4.5	1.26	24.8	Suffolk & Black
*B*2	2.8	0.8	12	van de Kamp
	2.9	0.63	12.5	Suffolk & Black
*B*3	1.8	0.89	6.1	Suffolk & Black

Examination of the periods of the "planets" in Table 2-3 shows them to be very nearly in the ratio of 2:4 for Van de Kamp's solution and 1:2:4 for Suffolk and Black's solution. This raises the question as to whether the perturbations ascribed to planets *B*2 and *B*3 may not in reality be harmonics produced by a highly elliptical orbit for *B*1 as originally proposed by Van de Kamp.

It is significant that the "harmonic content" of the wobble of Barnard's star is different in right ascension and declination. This could not be the case for multiple planets in *coplanar circular* orbits. The reduction in residual errors in Suffolk and Black's solution as compared with Van de Kamp's solutions is obtained only if the orbit of *B*1 is steeply inclined ($\gtrsim 40°$) to the orbits of *B*2 and *B*3.

Thus, it appears likely that Barnard's star has more than one orbiting companion and that either the orbits are not coplanar or at least one is highly elliptical. The evidence for three bodies will not be conclusive until a longer observation time shows that the periods of *B*1, *B*2, and *B*3 are indeed incommensurate. A solution that allowed ellipticity but required coplanarity would be interesting to pursue.

Table 2-4 lists several other stars known to have dark companions. For the first six the companion has from 1/10 to 1/35 the mass of the visible star. These we might classify as binary stars in which the smaller companion has too little mass to initiate thermonuclear reactions in its core. (It is generally agreed that this requires about 0.05 solar masses.) The last examples, 61 Cyg A and Proxima Centauri, are borderline cases where the star is 55 or more times as massive as the companion. (The Sun is 1000 times as massive as Jupiter.) Like Barnard's star, 70 Ophiuchi appears to have more than one unseen companion. We infer from these examples that a more or less continuous spectrum of systems exists between symmetrical binaries at one extreme and single stars with a giant planet, or planets, at the other.

It will be noticed that with one exception all the stars listed in Table 2-4 are *K*5 or smaller. The size of the wobbling to be expected from planets orbiting stars of one solar mass or larger is too small to be detected with present instruments. That we find an unseen companion of giant planetary size about nearly every star for which we could hope to detect the perturbations argues that most single stars have their retinue of planets.

ATMOSPHERIC EVOLUTION

If all planets had simply condensed from the early disk one would expect them to have essentially the same composition as the Sun: less than 2 percent heavy elements and the remainder hydrogen and helium. The giant planets appear to have approximately this composition. The inner planets, on the other hand, are composed almost entirely of heavy elements. (It is perhaps worth noting that, without their hydrogen and helium atmospheres, the outer planets would be comparable in size to the inner planets. Jupiter would be about 4 times as massive as Earth, Saturn about 2 times, Uranus 0.3 times, and Neptune 0.35 times.) Earth has almost no helium and relatively little hydrogen, most of it in the form of water. It is believed that the inner planets either formed without hydrogen or helium atmospheres or else lost these gases soon after formation.

TABLE 2-4

SELECTED NEARBY STARS WITH UNSEEN COMPANIONS

Star	Spectral Class	Distance (LY)	Mass (Star) Sun = 1	Mass (Companion) Sun = 1	Mass (Companion) Jupiter = 1	Separation (AU)	Orbit Period (yrs)
Cin 2347	*M*1	27.0	0.33	0.02	20	5.9	24.0
Cin 1244	*M*4	15.0	0.3 (?)	0.02	20	?	9.0
Cin 2354	*M*3-*M*5 (?)	18.0	0.2 (?)	0.02	20	?	10.8
70 Ophiuchi	*K*0	17	0.89	0.01	10	6.4	17
				0.012	12	4.4	
Lal 21185	*M*2	8.2	0.35	0.01	10	2.8	8.0
Kruger 60A	*M*3	13	0.27	0.009 (?)	9	4.1	16
61 Cyg A	*K*5	11.0	0.58	0.008	8	2.4	4.8
Proxima Centauri	*M*5	4.28	0.1	0.0018	1.8	0.8	2.5

The ability of a planet to hold an atmosphere depends on the energy required for a molecule to escape the planet's gravity compared with the average thermal energy per molecule in the exosphere. The exosphere is the outer region of the atmosphere where the mean free path is greater than the depth, so that an upward traveling molecule has little chance of collision. The biggest uncertainty is the exosphere temperature, which depends largely on the ultraviolet and x-ray flux from the solar corona. Some computed values of mean lifetimes in years for various gases and planets are given in Table 2-5.

TABLE 2-5

MEAN LIFETIMES OF ATMOSPHERIC COMPONENTS

Gas	Moon	Mars	Earth	Venus	Jupiter
Hydrogen	10^{-2}	10^{3}	$10^{3\pm.5}$	$10^{3\pm}$	$10^{200\pm50}$
Helium	10^{-1}	10^{6}	$10^{8\pm1}$	$10^{8\pm3}$	$>10^{200}$
Oxygen	10^{2}	10^{20}	$10^{32\pm3}$	$10^{32\pm6}$	$>10^{200}$
Argon	10^{10}	10^{49}	$10^{70\pm7}$	$10^{70\pm14}$	$>10^{200}$

It is apparent that none of the inner planets with their present masses can hold a hydrogen or helium atmosphere. Earth can retain argon and neon, but these gases are far below their cosmic abundance relative to the other heavy elements on Earth; thus, we infer that any primitive atmosphere was either blasted away by early intense solar radiation or else that the solar nebula had been dissipated (out to perhaps 2 AU) by the time the inner planets formed.

If this is true, the present atmospheres are the result of outgassing of the crust by the heat of impacting material during planetary formation, or volcanic action, or both. Volcanoes discharge large quantities of water vapor (originally physically trapped in rocks or present as water of crystallization) and carbon dioxide, some nitrogen, sulphur dioxide, hydrogen sulphide, and traces of hydrochloric and hydrofluoric acids. It is believed that the present iron-nickel core of the Earth formed later after radioactivity had heated the interior to a molten phase, and that the early crust therefore contained large amounts of free iron. Carbon dioxide and water are reduced by iron to hydrogen and methane.

Some hydrogen and nitrogen combined to form ammonia. Thus, we visualize the early atmosphere to have consisted primarily of H_2, N_2, CH_4, NH_3, H_2O, CO_2 with small amounts of other gases. The total volcanic activity that has occurred over the history of the Earth can easily account for all the oceans and for many times the present mass of the atmosphere.

Presumably, a parallel outgassing process took place on Venus, which has a mass almost equal to that of Earth. The present atmosphere of Venus is about 100 times as massive as ours and consists primarily of CO_2 with very little water. The heavy infrared absorption of this atmosphere has produced a runaway greenhouse effect, resulting in a surface temperature of about 700° K.

The solar flux is about twice as great on Venus as on the Earth. Theoretical models, developed by Rasool (ref. 11) argue that, because of the higher temperature, the Cytherian water vapor was unable at any time to condense into oceans, but remained in the atmosphere. There the high ionizing flux of UV and X radiation from the Sun disassociated the molecules and the hydrogen component escaped. The oxygen then combined with other crustal elements as the temperature rose from the increasing blanket of CO_2.

On Earth, the story was very different. Rasool's evaluations show that with the water largely condensed

into oceans, and with the smaller UV and X-ray flux at 1 AU, the early tectonic atmosphere of the Earth would have been stable. However, even Earth could not have engendered life, given as much CO_2 buildup as occurred on Venus. Most of the carbon in the Earth's crust is locked up in limestone. Calcium and magnesium silicates react with CO_2 in water to form silica and calcium or magnesium carbonates. On Venus, the lack of oceans prevented this reaction.

A great deal of limestone as well as all coal and oil are the fossil remnants of early life. It may well be that life appeared quite early and, through photosynthesis, started removing CO_2 from the atmosphere before tectonic activity had released anywhere near the amount that has accumulated on Venus. We must remember that volcanic activity is due to heat produced by the decay of long-lived istopes and, while it may have been greater initially, it has continued for billions of years. Vegetation and plankton are responsible for the extremely small percentage of CO_2 in the present atmosphere despite all our burning of fossil fuels. Living systems have certainly assisted in the CO_2 removal and may have been the decisive factor. If so, then life may be what saved Earth from the heat death of Venus, and is as much responsible for our fleecy skies and blue seas as it is beholden to them. Let us hope this symbiosis will continue!

ECOSPHERES AND GOOD SUNS

The ecosphere is the region surrounding a star within which planetary conditions can be right to support life. Too close to the star, any planet will be too hot; too far away, any planet will be too cold. We do not know precisely the limits of insolation that define the ecosphere. Dole, in his study of the probability of planets being habitable by man (ref. 4), choses the limits at 65 percent and 135 percent of the solar flux on Earth on the basis that 10 percent of the surface would then be habitable. If we consider the adaptability of life and do not require man to survive the environment these limits may be pessimistic. On the other hand, the factors responsible for appropriate atmospheric evolution may be more limiting than the temperature that would later exist on an Earthlike planet at various distances.

We can say that for an Earth-size planet, the inner ecosphere limit in the solar system lies somewhere between Earth and Venus. Rasool feels that the limit is not closer to the Sun than 0.9 AU, corresponding to 123 percent of Earth's insolation. On the other hand, Mars, at 1.5 AU and with 43 percent as much solar flux, is probably barren, not because of too little sunlight, but because it is too small to have had enough tectonic activity to give it much of an atmosphere. With a heavier atmosphere than ours, the greenhouse effect could make conditions on Mars quite livable.

The greater the ratio of the outer to inner radius of ecosphere the greater the probability of finding one or more planets within it. Dole has computed these probabilities using the data on planetary spacings of the solar system. Taking $(r_{max}/r_{min}) = 1.5$, his curves show a 66 percent probability of one favorably situated planet and a 5 to 6 percent probability of two. Since we do not have statistics for other planetary systems, all we can do is fall back on the assumption of mediocrity and say that something over half of all planetary systems should have at least one favorably situated planet.

As we consider stars of increasing luminosity, the ecosphere moves out and widens. As Cameron observes, if Bode's law holds generally—if the planetary spacing is proportional to distance from the primary—this exactly compensates for the widening of the ecosphere and gives a *constant* probable number of planets per ecosphere, regardless of the size of the primary star. Assuming that the more luminous stars were able to dissipate their nebulae to proportionately greater distances so that terrestrial planets could evolve, the upper limit of luminosity for a good sun is reached when the stellar lifetime becomes too short. If intelligent life typically requires the 4-1/2 billion years it has taken to evolve on Earth we see from Figure 2-7 that we should exclude all stars hotter than *F*4 stars.

As we consider stars of decreasing luminosity, the ecosphere shrinks and a new difficulty arises. If the tidal braking becomes great enough to stop the planet's rotation, the entire atmosphere will probably freeze out on the dark side and all life will die. To estimate the limit imposed by this effect, let us divest the Earth of its moon and bring the Earth closer to the Sun, decreasing the Sun's luminosity (and mass) to keep the insolation constant, until the solar tide equals the present combined solar and lunar tide. Since the solar tide is 46 percent of the lunar tide we can afford to increase the solar tide by the factor:

$$k = \frac{\sqrt{1 + (0.46)^2}}{0.46} = 2.4 \qquad (6)$$

without changing the tidal drag Earth has experienced. The tide raising force is proportional to M/r^3, while the insolation is proportional to $M^{3.5}/r^2$. Thus if we keep the insolation constant, M will vary as $r^{2/3.5}$ and the

tide raising forces as $r^{-\alpha}$ where $\alpha = 3 - (2/3.5)$. We can therefore reduce r until $(r/r_0)^{-\alpha} = k$ or

$$\frac{r}{r_0} = k^{-1/\alpha} = k^{-7/17} \tag{7}$$

and

$$\frac{M}{M_\odot} = k^{-2/3.5\alpha} = k^{-4/17} \tag{8}$$

Taking $k = 2.4$ we find $r/r_0 = 0.7$ and $M/M_\odot = 0.81$. This corresponds to being at the orbit of Venus around a $K0$ star.

So long as we restrict ourselves to the same pattern of seas, continents, and basins, it is reasonable to suppose that tidal drag varies as the square of the tidal amplitude—that is, as k^2. However, differing continental configurations and coastal features can radically affect the total drag. There is evidence that the tidal friction on the Earth has varied widely during its history. If Gerstenkorn's theory (refs. 12, 13) of the evolution of the Moon is correct, the Earth underwent a period of enormous tides on the order of 10^9 years ago and was slowed greatly by the Moon at that time. Without the Moon, the Earth might well have been able to have endured solar tides two or four times as great as the present total tide and still have the present length of day. Also, a planet with less water and perhaps isolated seas would experience far less tidal drag.

If we let $k = 5$ we find $r/r_0 = 0.51$ and $M/M_\odot = 0.68$, which corresponds to a body 0.5 AU from a $K5$ star; while if we let $k = 10$ we find $r/r_0 = 0.387$ and $M/M_\odot = 0.58$, or a body in the orbit of Mercury around a $K7$ star. We concede that, as tidal forces increase, we are restricting more and more the planetary configuration that can survive. But we see no reasons to drop the probability of surviving tidal drag to zero at $K2$ as Dole has done.

Some current theories predict that the UV and X-ray flux of a star, which depend on coronal activity, may actually increase as we go from G to K to M stars. The effect of this on atmospheric evolution may well place a more stringent lower limit on likely main sequence stars than tidal braking. For the present, we feel that all F, G, and K main sequence stars should be included in a search, with G stars given priority.

THE ORIGIN OF LIFE

It is now believed that chemical evolution leading to life began on the Earth between four and four and a half billion years ago. At that time the atmosphere was probably composed primarily of a mixture of hydrogen, nitrogen, carbon dioxide, methane, ammonia, and water vapor. Ultraviolet radiation from the Sun was able to penetrate the Earth's atmosphere because ozone had not yet been formed in its upper layers. The Earth's crust was undergoing early differentiation. Earthquakes and volcanic activity were probably intense and frequent. The primitive oceans had formed, and contained, in contrast to the oceans of today, only small quantities of dissolved salts. Most scientists believe that life began in the sea.

A major requirement for the development of living systems is the presence of organic molecules of some complexity. Until recently it was thought that these molecules could be produced only by the activity of living systems, and there was no satisfactory explanation of where the living systems came from in the first place. In 1938, the Russian biochemist Oparin (ref. 14) put forward his theory of chemical evolution, which proposed that organic compounds could be produced from simple inorganic molecules and that life probably originated by this process. In 1953, Miller (ref. 15) working in Urey's laboratory, showed that organic molecules could indeed be produced by irradiating a mixture of hydrogen, ammonia, methane, and water vapor. Chemists began to experiment with all the different forms of energy thought to have been present early in the Earth's history. Ponnamperuma and Gabel (ref. 16) soon showed that ultraviolet radiation, heat, electric discharges, and bombardment by high energy particles also worked. Recently sonic cavitation produced in water by wave action has been shown by Anbar (ref. 17) to be effective in producing organic molecules.

A number of the organic compounds produced in this way are identical with those found in the complex biochemical structures of present-day organisms. The first compounds synthesized in the laboratory by Miller were amino acids, necessary for the formation of proteins.

Later experiments, including those of Ponnamperuma, have produced sugar molecules. Specifically, he isolated among others the two sugars with five carbon atoms: ribose and deoxyribose. These are essential components of the deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) molecules, which carry the genetic code of all Earth-based life. The purine and pyrimidine bases, whose locations in the DNA molecule represent the information carried by the gene, were also synthesized in the same simple apparatus. Finally, Hodgson has recently shown (ref. 18) that porphyrin compounds can be isolated. These substances are an

essential component of some of the molecules responsible for the transfer of energy in living systems.

It is clear that similar processes were taking place in the atmosphere of the Earth four billion years ago. Many of the organic molecules so produced subsequently dissolved in the primeval sea. Further organosynthesis took place in the upper few millimeters of the primitive ocean as a result of absorption of ultraviolet light. Beneath the surface, organic molecules were modified at a much slower rate since they were largely protected from the energy sources responsible for their origin.

So far the story of the origin of life is reasonable, easy to understand, and well supported by present-day experimental evidence. We visualize the primitive ocean containing in dilute solution a wide variety of organic compounds suitable as precursors for living systems. The environment is fairly stable over millions of years. As the compounds degrade they are replaced by more of their kind falling into the sea from the atmosphere, and they are modified at the surface by ultraviolet radiation. The process is called chemical evolution since living systems were not yet present.

In South Africa, Barghoorn found in a rock formation called Figtree Chert, microstructures that could be the fossils of single-celled organisms (ref. 19). The age of the formation, obtained from radioisotope dating, is 3.1 billion years. The sequence of events between the time when only the mixture of organic precursors existed in the early oceans and the time when the first living cell appeared, 3.1 billion or more years ago, is still unclear. It is the only portion of the entire chain of events constituting biological evolution that is not yet understood. It is a crucial step, for it marks the transition from the nonliving to the living system. Somehow the organic molecules of the primitive ocean were assembled into that complex unit of life, the cell.

The cell is a highly organized structure containing a large amount of information in a very small volume. It has all the characteristics we associate with living systems. It is self-reproducing. It contains within its nucleus a molecular code for the ultimate control of all cellular activities and structure. It degrades high energy food or light into lower forms of energy and extracts useful work to maintain its activities. It has sophisticated systems for maintenance and repair, and is able to remove waste products and to protect itself from a variety of environmental threats.

In recent years it has become more difficult to arrive at a rigorous definition of a living system. The chemical and physical basis of life, now generally accepted, makes it hard to find any characteristics peculiar to living systems as a whole. The most apposite definition, suggested by Schrödinger in 1943 (ref. 20), is that living systems maintain and propagate localized states of decreased entropy over very long periods of time. At first sight living systems appear to be contradictions of the second law of thermodynamics. The *total* system in which life occurs does not disobey the second law: decreased entropy represented by the high level of organization of the biological part of the system is achieved and maintained at the expense of a greater increase in entropy in the rest of the system.

The critical problem remains. What was the sequence of events in the evolution of the cell from the mixture of organic precursors in the ocean? One possibility is that autocatalysis played a significant role. In a sequence of reactions in which compound A → compound B → compound C, if compound C turns out to be a catalyst for the reaction A → B, it will accelerate the yield of compound B and of itself. Many examples exist. Particulate radiation can cause polymerization of certain classes of compounds. Today certain polymers, such as the DNA chain, can be made to produce replicas of themselves in solution given the raw materials, an appropriate enzyme, and an energy source. It is likely that the evolution of the cell required a long sequence of individually improbable events. Nevertheless, given the millions of years available and the vast supply of organic raw materials, these infrequent crucial events must have occurred many times.

The numerical values for the probabilities must await the construction of models of the chemical interactions, in a variety of environments, at a level of complexity unobtainable with present knowledge and computer capacity. As will be seen later, the numbers are very important because they will help to answer three fundamental questions, each crucial to the problem of interstellar communication:

1. On those planets of solar systems throughout the Galaxy that had evolutionary histories similar to that of Earth, what is the probability that life began? Clearly if the number is small, say 10^{-10}, then Earth has the only life forms in the Galaxy. However, it is generally believed that the number is high, perhaps high enough that life could have begun on many separate occasions on the Earth, and correspondingly on all planets similar to the Earth.
2. To what extent do the probabilities of the development of life change as the geophysical, atmospheric, and chemical characteristics of other planets deviate from those of Earth? If the numbers remain high, then life is the rule in the universe. There is less general agreement about the answer to this question.

3. Is it possible that living systems could be constructed from molecular subunits different from those making up our biosphere? Carbon, hydrogen, oxygen, and their numerous compounds are eminently suited to serve as the raw materials for living structures. It is not considered probable that living structures could be based on other chemistries or that they would survive the competition with carbon-, hydrogen-, and oxygen-based life if they were. It is noteworthy that all living forms on Earth not only have a common elemental composition, but are in fact composed of very similar molecular units—for example, amino acids, nucleotides, porphyrins and riboses—arranged in different ways. Living systems on other planets could well have similar fundamental components.

Although the steps leading from the organic precursors to DNA are presently unknown, biologists are confident that one or more likely routes will be found. There is thus a general consensus among exobiologists that life has originated and is now present on the planetary systems of a large proportion of the stars in the Galaxy.

BIOLOGICAL EVOLUTION

At the time of the emergence of the first living cells on Earth the atmosphere contained little or no oxygen, and the energy needs of the cell had to be derived from reactions other than direct oxidation. The energy was mostly derived from the pool of organic compounds built up by the processes already described, but may have included energy derived from a few inorganic compounds. Such anaerobic processes are still used today by a variety of primitive organisms. At some point in the early evolution of cellular systems, a mutation, or genetic change, gave certain organisms the ability to transform the energy of visible light into chemical energy stored in an oxidation-reduction system. The critical biochemical compound for this reaction is chlorophyll, in which the absorbed photons raise electrons to a higher energy level. An accompanying series of reactions enabled the plants to extract CO_2 molecules from the air, and reduce them with the new found energy to form the complex compounds making up the structure of the organism. The combined process is known as photosynthesis. The date at which photosynthesis began is still unclear, but Kvenvolden (ref. 21) has recently shown that pre-Cambrian rocks, in which microfossils are found, show an unusually high ratio of ^{12}C to ^{13}C. Photosynthesis is known to discriminate between these carbon isotopes in the same way. This evidence suggests that photosynthesis was present on Earth 3.4 billion years ago.

One of the major results of the evolution of the photosynthetic system was an increasing amount of molecular oxygen, formed by a by-product of the reaction, in the early atmosphere. At first the oxygen was probably used up fairly rapidly in oxidizing a wide variety of chemical substances. Later free oxygen became available for the oxidation of nutrients in living systems. Such oxidation possesses a clear advantage over the anaerobic system in that a great deal more energy per molecule of the organic nutrient is available to drive cellular reactions.

Over the next three and a half billion years the primitive organisms slowly evolved into the vast array of living systems we see today. The basic mechanism underlying this long epoch of biological evolution is the chance modification of the chemical structure of the DNA molecule known as a mutation. Most mutations are detrimental. They give rise to lethal chemical structures and reactions in the cell or prevent vital ones. Very occasionally the reverse is true and the mutation is said to be favorable. A favorable mutation confers a greater chance of survival, since the cell can now compete more successfully for the available energy sources and can more successfully withstand the rigors of the environment. Over many generations the organisms possessing a favorable mutation will gradually displace those without the new characteristic. This process, known as natural selection, was proposed by Darwin (ref. 22) and independently by Wallace in the middle of the nineteenth century as a rational explanation for the whole history of the evolution of the widely differing species that make up the plant and animal kingdoms, although the genetic mechanism by which it operates was then unknown.

The theory of natural selection at once explained the phenomenon, previously attributed to divine will, that the vast majority of living organisms seem so perfectly fitted for the particular environments they occupy. Experimental evidence supporting the Darwinian hypothesis is now so complete that few have any major reservations as to its correctness.

Evidence is available from the fossil record, and more recently from the studies of the comparative biochemistry of present-day species, to give us a reasonable picture of the sequence of events as new species emerged. Many died out en route either as a result of changes in the physical environment, or by competition with superior mutants better able to use the available resources. A number of species reached a certain stage of

development and remained little changed until the present day: these are the so-called "living fossils."

Gradually, as multicellular organisms evolved, different groups of cells became adapted to performing special functions that increased the survival capability of the organism. In the animal kingdom, certain of these cells formed into a nervous system. These communication links between different parts of the animal enabled a more coherent behavior in response to environmental threat. Skeletal structures developed for support, respiratory systems for increasing the efficiency of gas exchange, circulatory systems for carrying the gases and nutrients to the tissues, digestive systems for preprocessing organic foods, excretory organs for removing the end products of metabolism, and reproductive systems for increasing the odds of survival and for ensuring a constant mixing of genetic material.

In the animal kingdom, a crucial event was the gradual emergence of animals able to exist on the land. Successful adaptation to the land demanded not only alterations in the respiratory apparatus, but also the development of limbs to permit foraging for food and escaping from danger. It is generally thought that the great increase in complexity of the nervous system, which was necessary in the land animals, was a stimulus to the further evolution of the central nervous system essential for the later development of intelligence.

As animals and plants colonized the land, species diversified into a wide variety of organisms including birds, flowering plants, amphibians, and giant reptiles, the dinosaurs. The dinosaurs lacked two critical physiological systems that were combined in an evolving group of small animals called mammals. The mammals had developed a control system for the regulation of internal temperature and a mechanism for allowing the protection of the young inside the body of the female during early development. The conditions responsible for the extinction of the dinosaurs are not known, but the mammals survived them. At some point, perhaps a hundred million years ago, the mammals developed the capability of living in trees, as their descendants, the lemurs, marmosets, and monkeys, do today.

Further evolution of the central nervous system accompanied the arboreal mode of life. Control systems for vision, orientation, balance, and movement are necessarily complex, and demanded further sophistication of the central processing system in the brain. Twenty million years ago, the brain had enlarged to a volume of about a hundred cubic centimeters and contained millions of nerve cells functioning as a highly integrated control system. It is usually thought that environmental processes, including competition with other species, now favored the emergence of small groups of monkeylike species, increasingly able to survive in the open grasslands. Many of the adaptations needed for tree living, such as stereoscopic vision, precise orientation and balance, and great precision of movement, paved the way for the evolution of early man on the plains. (An interesting question about extraterrestrial biological evolution is whether trees are in fact necessary for this phase of development of the central nervous system.)

Existence on the plains favored further anatomical changes: adoption of the upright posture, and adaptation of the teeth and the jaws to the new and tough animal and plant foods. It is generally supposed that the hands, now freed, became adapted for the precise manipulation of tools. The thumb, for example, could be opposed to the fingers for grasping, and was capable of a much greater range of movement. Along with the changes in the body came further enlargement of the brain, principally in the cerebral hemispheres. Success at hunting in the plains was enhanced by the evolution of early speech patterns, and further by coordinated behavior between members of a group.

The fossil record shows that some two million years ago the brain size had increased, in a species called Australopithecus, to about 500 cubic centimeters. Australopithecus lived in the African plains, used tools, and was carnivorous. Early forms of group behavior and effective communication between individuals had probably emerged.

Two events of major significance were occurring as a result of the progressive enlargement of the brain. The first was the development in the cerebral hemispheres of an *internal model of the external world* and the second was the ability to pass on these models to the young and to other adults by communication and reinforcement.

CULTURAL EVOLUTION AND DEVELOPMENT OF INTELLIGENCE

In the lower animal species, the brain controls the behavior of an individual in a fairly stereotyped way. Reactions to the stresses of the environment are limited in terms of the options available. The brain acts in response to a genetically determined and comparatively simple model of the external world. In early man, the model was slowly developing into a more complex form. Association between events could be recognized and instated in the internal model so that more intelligent responses to the environment became feasible. An Australopithecine individual endowed with a cerebral hemisphere that allowed better association processes would be more efficient in hunting and, according to the

laws of natural selection, more likely to survive and pass on the accidental increase in computing capability to descendants. An example would be the observation that certain weather patterns, or seasons, coincided with the appearance of desirable game in different localities. Hunting becomes more efficient when this relationship is put to use. Such individuals were learning in a primitive way how to plan for the near future.

In earlier times, had such associations arisen in an individual, they would not have survived in descendants because they were not contained in the genetic code of that individual, and could not be passed on in any other way. With the ability to communicate, early man was not faced with this genetic limitation. By actions, by signs, or by speech, the new knowledge could be passed on to other members of the group, including the young. We see the first rudiments of training and learning, now possible because the young not only have inborn reflexes, but also a flexible central nervous system ready to be programmed by adults. The cost of this development was that the newborn were increasingly unable to fend for themselves, and required months, and later, years of protection by parents. The cost was outweighed by the ability to impart to the young the legacy of knowledge accumulated over many previous generations.

In lower forms of animal life, characteristics acquired by an individual cannot be passed on to its descendants. This is a central and largely unchallenged tenet of Darwinian evolution. The development of the cerebral hemispheres, and the sophisticated model of the external world, together with the ability to communicate, bypass this rigid restriction. Acquired characteristics can be transmitted in this special fashion.

It is clear that further development of associative areas in the brain, together with better ability to utilize the knowledge residing in the internal model, would confer selective advantages. Brain size began to increase rapidly. Between the time of Australopithecus, some two million years ago, and the present day, the brain has increased in volume from about 500 cubic centimeters to 1300 or 1400 cubic centimeters.

As man evolved he began to develop many of the characteristics we classify as "human." The ability to retain crucial components of the internal model led to the increasing complexity of short- and long-term memory. The model of the external world included a model of the individual himself. Consciousness and self-awareness are the words we use to describe this portion of the model. Individuals began to act in a concerted fashion, exhibiting group behavior. Codes of behavior were established, later to be refined into laws. The groups were better able to defend themselves, to counter environmental changes, and to obtain food. Home bases were established, and the females came to carry out specialized tasks associated with the maintenance of the base and rearing of the young. The males hunted in groups and were thus able to kill larger game animals. Hierarchies of command developed along with the specialization of function in individuals.

The events we are describing represent the beginnings of human society. The knowledge and behavior of the individual is a distillation of minute advances occurring occasionally over thousands of generations. The specialization of function among individuals was no longer a strictly genetic trait, but a consequence of an increasingly planned and deliberate training to better serve the group as a whole. The character, skills, and behavioral patterns of an individual were as much influenced by upbringing and directed education as by inborn capabilities. Occasionally today a child develops without any education, with the result that his behavior is very "animal-like," without speech, skills, apparent intelligence, or understanding.

The transmission of knowledge accumulated over many generations is known as cultural evolution. The now very sophisticated internal models of the external world allowed reasoning and the establishment of complex patterns of individual and group behavior based on that reasoning. The primitive instincts or "goals" of self-preservation, hunger, and reproductive drives still dominated, but the solutions to the daily problems of living were becoming more complex and successful. Man was becoming intelligent.

One of the characteristics of intelligence is that action can be taken to control the environment and hence to decrease its hazards. Early man developed a facility for the use of tools. At first, these were the simple instruments of stone and wood that litter dwelling sites. Gradual refinement brought the axe, the spear, and instruments to process game, to make clothes, and to manufacture further tools. Some half million years ago man learned to use fire. The materials of the environment were being manipulated to extend the physical capabilities of the body, and to supply energy for a variety of needs. Such activities could be described as the beginnings of technology.

Around fifty to one hundred thousand years ago, the evolution of the human race had reached the point where Homo sapiens was emerging in something like his present form. Another variety, Neanderthal man, with a somewhat different appearance, but with an equally large brain, existed until the Ice Ages, but then disappeared. The reasons are not clear. Two possibilities are an unfavorable geographic location at a time of

intense climatic change, or competition with Homo sapiens sapiens. Natural selection was still at work.

CIVILIZATION, SCIENCE, AND TECHNOLOGY

Some few thousand years ago, cultural evolution was accelerated by the development of new methods of controlling the environment. Today we might refer to these changes as "technological breakthroughs." The catalog reads: weaving and other improvements in clothing, house building, the development of agriculture, transportation by horse and boat, the invention of the wheel, and the discovery of metals (the Iron and Bronze ages). Along with these technical changes there developed an ultimately more powerful human capability, which was in turn to permit further technological sophistication.

The new achievement was the use of symbols, which can be viewed as an external manifestation of the internal model of the world now developing so rapidly. The symbols were those of writing and measurement, and they permitted more complete communication and planning. Early sign writing evolved into the alphabets. Numbers could be written down, communicated, and preserved, and the precision of the manufacture of tools, buildings, vehicles, and other devices were improved by measurement.

The rapid cultural evolution at the beginning of recorded history brought a much greater control of the environment and utilization of its resources. It became possible for larger populations to support themselves in relative stability. Instead of living in small groups, human societies began to congregate in large cities, and the early civilizations evolved in the Middle East, China, India, the Mediterranean, and in Central America. The process has continued to the present day and has been accelerated, particularly since the Renaissance, by the growth of modern science and technology.

In these few paragraphs it is not possible to review the development of science and technology in any detail. It is, however, important to single out some of the landmarks, and to describe their influence on the behavior of society.

Some gains in the understanding of natural processes were achieved by the ancient Egyptians, and during the classical age of Greece. For the most part these were limited to macroscopic events that could be seen and measured with simple instruments. Some predictions could be made of major astronomical events, and mathematical proofs were found for geometrical problems. The Greek philosophers struggled to find logical explanations for natural phenomena, including human behavior. While their conclusions were often wrong because of false premises, their methods of logical argument, together with the increasing respect for the power of analysis, foreshadowed the growth of the scientific method in Europe after the Dark Ages.

Post-renaissance inventions such as the telescope and microscope extended the range of man's senses and brought increased and more quantitative understanding of the world. The invention of printing made the communication of ideas comparatively simple. Improvements in agriculture, shipping, and manufacture, together with colonization and trade, led to an increased wealth so that it was possible for a few gifted individuals to devote time to the construction of theories and to the development of experimental techniques. Scientific societies were formed, and the scientific method was established. Biology, physics, chemistry, and mathematics progressed rapidly. Basic laws governing the behavior of matter and energy were discovered and refined. In their wake, and demonstrating their power, came the major technological advances that in turn, permitted a more precise control of the environment and an expanding population. Energy from steam, petroleum, and the atom became available, and made possible large-scale engineering projects.

This fragmentary chronicle describes the results of recent advances in the precision of the internal model of the external world. The modern computer represents an extension of such internals models, and is now capable of carrying out calculations necessary for a more complete description of the external world at speeds far in excess of those that could ever be achieved by the human brain in its present form. We begin to talk of artificial intelligence and of processes of machine computation that are coming to have some of the characteristics of human thinking. The boundary between the living and the nonliving (machines) is becoming less clear in any absolute terms, in the same way that the boundary between the first cell and its complex biochemical precursors is a little vague.

The capability to control and use resources has permitted the organization of modern society into larger and larger groups. The amount of localized decreased entropy in this corner of the solar system continues to increase. Life is flourishing as never before on the planet Earth.

We are concerned in this study particularly with the rapid growth in the recent understanding of the physical universe. The description of the evolution of galaxies, stars, and planetary systems found earlier in this chapter testifies to this knowledge. The description of the evolution of life on this planet, now fairly complete, forces us to ask whether living systems on the planets of

stars throughout the Galaxy have also evolved into advanced technological civilizations. If so, what are the civilizations like, and how did they evolve?

Many will presumably be far advanced in comparison with human society on Earth. Since no two planets are exactly alike it is improbable that intelligent beings on other planets would closely resemble man. On the other hand, physical and chemical laws suggest that the anatomy and physiology of the major human characteristics have self-evident evolutionary advantages. Perhaps other extraterrestrial intelligent species have some similarities to ourselves. What is their science and technology like? What new understanding would be revealed if we were able to communicate with them? These are among the most challenging scientific questions of our time.

CONCATENATED PROBABILITIES

We have outlined the development of technologically competent life on Earth as a succession of steps to each of the which we must assign an *a priori* probability less than unity. The probability of the entire sequence occurring is the product of the individual (conditional) probabilities. As we study the chain of events in greater detail we may become aware of more and more apparently independent or only slightly correlated steps. As this happens, the *a priori* probability of the entire sequence approaches zero, and we are apt to conclude that, although life indeed exists here, the probability of its occurrence elsewhere is vanishingly small.

The trouble with this reasoning is that it neglects alternate routes that converge to the same (or almost the same) end result. We are reminded of the old proof that everyone has only an infinitesmal chance of existing. One must assign a fairly small probability to one's parents and all one's grandparents and (great)^n-grandparents having met and mated. Also one must assign a probability on the order of 2^{-46} to the exact pairing of chromosomes arising from any particular mating. When the probabilities of all these independent events that led to a particular person are multiplied, the result quickly approaches zero. This is all true. Yet here we all are. The answer is that, if an entirely different set of matings and fertilizations had occurred, none of "us" would exist, but a *statistically undistinguishable* generation would have been born, and life would have gone on much the same.

It is not important that the particular sequence of events leading to intelligent life on Earth be repeated elsewhere, but only that some sequence occur that leads to a similar end result. Thus, the key question is not whether the precise conditions causing a particular sequence are replicated elsewhere, but *whether the forcing functions are present and whether enough alternate routes exist.* The origin and evolution of life would appear to be favored if a planet provides a wide variety of environments—that is, a *range* of values for every important parameter. Since all planets will have a range of climate from poles to equator, most will have tilted axes and therefore seasons, and many will have both seas and land and therefore solar tides, we can expect the variety of environments found on Earth to be common. For all we know, life on Earth may have had serious setbacks and been forced to re-evolve (refs. 12, 14). If this be true, the genesis time on many planets may have been much shorter than four billion years.

THE NUMBER OF COEXISTING ADVANCED CIVILIZATIONS

Our general picture of cosmic evolution, together with the assumption of mediocrity, suggest that life has evolved or will evolve on suitable planets around a large number of stars in the galaxy. In assessing the problems of interstellar contact between civilizations, we would like to know how many advanced civilizations are present *now* and how the number of coexisting cultures has varied over the past history of the galaxy.

Schmidt (refs. 23, 24) has studied the rate of star formation in the Galaxy as a function of time. His analysis, like that of other workers, begins with the assumption that the rate of star formation is proportional to the interstellar gas density to some positive power n. The gas density rose steadily during the collapse phase until hydrostatic and kinematic equilibrium were reached. During and subsequent to the collapse phase, numerous population II stars were born. The gas density was then gradually depleted as more and more stars with increasing heavy element content (population I) were formed. Thus, the rate of star formation should have reached a peak early in galactic history and have declined ever since.

In his 1959 paper, Schmidt examined three possible cases for which the exponent $n = 0,1,2$. Figure 2-9 shows the resulting buildup in total stellar population for these cases under the assumption that one-third of the interstellar material is recycled by supernova explosions. If we take the age of the population I disk to be ten billion years, then each subdivision of the abscissa is one billion years. For $n = 2$ we see that at half the present age of the disk, or 5 billion years ago, over 90 percent of all stars now in existence had already been born. In the 1959 paper, Schmidt found that the observational evidence favored $n = 2$ for all stellar types. In his 1963 paper, however, he argues that the past average birth rate of solar mass stars could not have exceeded three times

the present rate. This means that the true curve should approach the final point in Figure 2-9 with a slope not less than one-third as indicated by the dashed line.

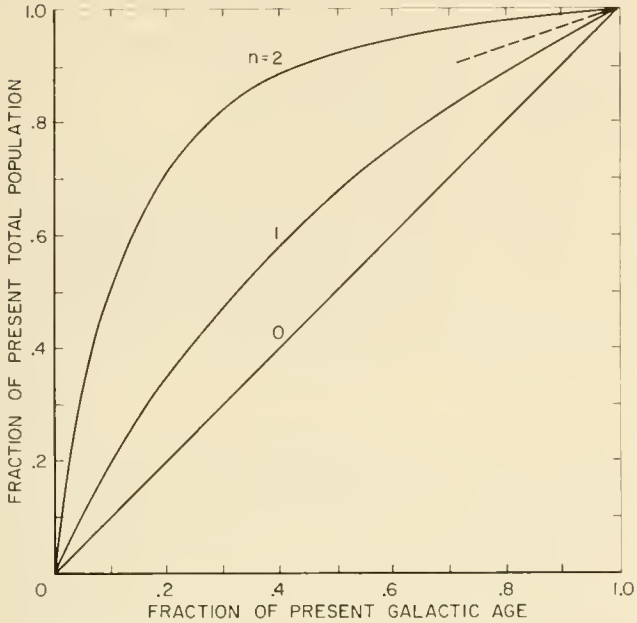


Figure 2-9. *Number of stars versus galactic age.*

Also, since the heavy element content has risen with time we might expect that a smaller fraction of the early population I stars had enough heavy elements to form terrestrial planets large enough for the evolution of life. Thus, the shape of the actual buildup curve for stars having suitable planets is somewhat uncertain at present, but is probably a curve intermediate to those for $n = 1$ and $n = 2$ in Figure 2-9.

Let us define a "life site" as a planet destined for the evolution of life regardless of whether that life has yet developed, exists at the time considered, or has already flourished and vanished. Let N_S be the present number of life sites and $n(t)$ be the number at galactic age t. Then the fraction of the present number $f(t) = n(t)/N_S$ will have built up with time in accordance with the population growth of F, G, and K main sequence stars as indicated qualitalively by curve (1) in Figure 2-10. Let us now assume that, after a genesis time G, life evolves to an advanced technological state on each of these sites and has a longevity L. Then the fraction of sites on which advanced life has evolved will be zero for $t \leqslant G$ and will be $f(t - G)$ for $t > G$. Similarly, the fraction of sites on which life has flourished and perished will be zero for $t \leqslant G + L$ and will be $f(t - G - L)$ for $t > G + L$. Thus, the fraction $g(t)$ of sites with life will be:

$$g(t) = \begin{cases} 0 & , \ t < G \\ f(t-G) & , \ G < t < G+L \\ f(t-G) - f(t-G-L), & G+L < t \end{cases} \qquad (9)$$

and is shown by the heavy line (4) labeled "sites with life" in Figure 2-9.

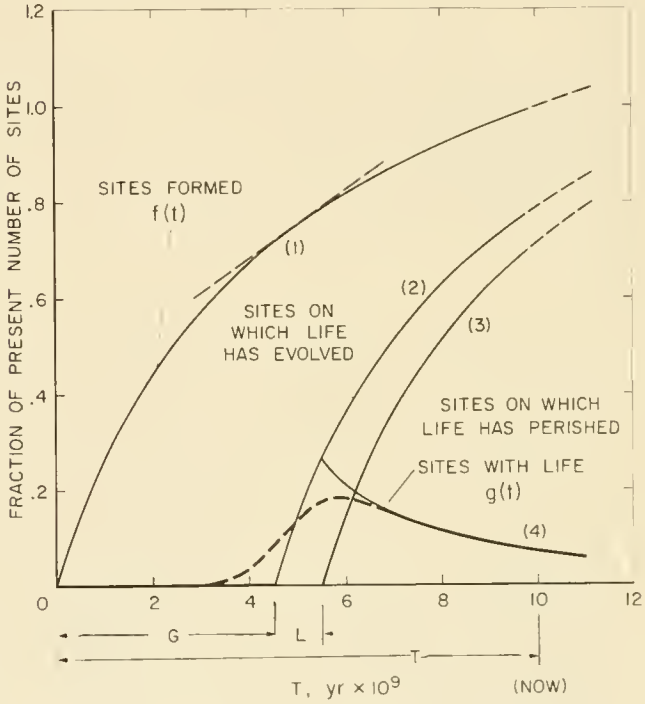


Figure 2-10. *Status of life sites in Galaxy.*

We do not expect the genesis time to have the same value for all sites but rather to be distributed about the mean value G. Similarly, we expect the longevity to be distributed about the mean value L. Thus curve (2) in Figure 2-10 should really be the convolution of curve (1) with the probability density function of G, while curve (3) should be the convolution of (2) with the probability density function of L, as Kreifeldt has pointed out (ref. 25). The effect of this convolution will be to produce "fillets" at the beginnings of these curves; that is, the slope will build up smoothly rather than having a discontinuity at the outset. Qualitatively, the effect on the difference curve (4) will be to smooth it as indicated by the dotted curve. So long as the dispersions are less than one or two billion years, the present value of $g(t)$ will be unaffected.

We see from Figure 2-10 that, if the genesis time of life on Earth is typical, the Galaxy has been populated with advanced life for five or six billion years, and that,

if the longevity of that life is typically less than one or two billion years, *advanced life was more common in the past than it is now.* Of course, if $G + L$ is greater than the present age of the Galaxy then the density of life in the Galaxy is still increasing.

If we assume that L is small, say on the order of one billion years, then $G + L$ is less than the present age of the Galaxy and the third form of (9) applies. In this case we have

$$g(t) \approx f'\,(t - G - \frac{L}{2})\,L \tag{10}$$

The number of sites in the Galaxy can be expressed as the product of the number of stars N_* and several "selectivity factors":

$$N_s = N_* F = N_* f_p n_e f_\ell f_i \tag{11}$$

where

$$F \equiv N_s/N_* = f_p n_e f_\ell f_i$$

and

f_p = fraction of stars having planets

n_e = number of suitable planets per ecosphere

f_ℓ = fraction of suitable planets on which life starts

f_i = fraction of life starts that evolve into intelligence

Multiplying (10) and (11) we get for the number N of intelligent civilizations in the Galaxy:

$$N = R_* FL \tag{12}$$

where

$$R_* = N_* f'\,(t - G - \frac{L}{2})$$

$$= \text{rate of star formation } G + \frac{L}{2} \text{ years ago}$$

If we wish N to represent the number of "communicative" civilizations we must interpret L as the longevity of the *communicative* phase and include an additional selectivity factor in F:

f_c = fraction of intelligent civilizations that attempt communication.

This yields Drake's expression

$$N = R_* f_p n_e f_\ell f_i f_c L \tag{13}$$

which has been described as a way of compressing a large amount of ignorance into small space. Nevertheless, this expression identifies the important factors and allows us to assess them independently. Let us enter some very approximate and probably optimistic values into (13).

R_*	= 20/year	(= N_*/T = average rate)
f_p	= 1/2	(excludes multiple star systems)
n_e	= 1	(correct for solar system)
f_ℓ	= 1/5	(only certain F, G, K main sequence stars suitable)
f_i	= 1	(Darwin-Wallace evolution inevitable)
f_c	= 1/2	(only land-based life develops technology)

With these values

$$N \approx L \tag{14}$$

which says that the number of communicative races in the Galaxy is roughly equal to the average number of years spent in the communicative phase. This turns out to be the most uncertain factor of all!

The substitution of N_*/T for R_* made above does *not* assume a constant rate of star formation, but only that the rate $G + (L/2)$ years ago was about the average rate. This may be slightly optimistic; in Figure 2-10 the slope of $f(t)$ at $t = 5$ billion years is about 3/4 rather than 1, but this is a small error compared with all our other uncertainties. If we make this same substitution in (12) we find

$$N = N_* F \frac{L}{T} \tag{15}$$

and if we now divide both sides by N_* we obtain the *a priori* probability $p = N/N_*$ that a given star selected at random is in the communicative phase

$$p = F \frac{L}{T} \tag{16}$$

If we confine our attention to what we consider particularly likely stars, say single main sequence F, G, and K stars, p is increased because the selectivity factors contained in F are presumably greater for this subset of stars. If our knowledge about planetary system statistics, atmospheric evolution, genesis times, etc., were complete, we might be able to be so selective in choosing target stars as to make $F \to 1$. At our present state of knowledge, all we can say is that probably $10^{-2} < F < 1$, but we cannot guarantee the lower limit.

The probability p is used in Chapter 6 to estimate the probability p_c of achieving contact as a function of the range to which we carry the search.

THE PROBABILITY OF INTERSTELLAR COMMUNICATION

To have a 63 percent chance of success we must search $1/p$ target stars. If the search takes τ years per star, the mean search time will be

$$L_S = \frac{\tau}{p} = \frac{\tau T}{F L_r} \tag{17}$$

where we have written L_r for L to indicate the length of the radiative phase.

If interstellar communication is not an already established reality in the galaxy and various races, like ourselves, make sporadic attempts at both searching and radiating, we might assume that $L_r = L_S$ in (17) and obtain

$$L_S \approx \frac{\sqrt{\tau T}}{\sqrt{F}} > \sqrt{\tau T} \tag{18}$$

Taking $\tau = 1000$ sec $= 3 \times 10^{-5}$ years, we find

$$L_S > 560 \text{ years}$$

This represents a truly formidable effort, but one we might be willing to make if we were sufficiently convinced of the existence of extraterrestrial intelligent life, of the potential value of contact, and of the validity of our search technique.

On the other hand, if interstellar communication already exists, the situation is likely to be very different as a result of what Von Hoerner (ref. 26) has called the "feedback effect." The radiative history of a typical civilization might then be as depicted in Figure 2-11. After an initial search phase of duration S, during which the civilization radiates a beacon signal, with probability p_r, contact is established for reasons that will soon become clear. The civilization then finds itself part of a galactic community that shares the obligation to facilitate acquisition by other civilizations. This might mean radiating a beacon for a fraction σ of the time for a very long period L. (Or such beacons might be established for reasons we are completely unable to foresee or understand.) The radiative phase would then have an effective duration:

$$L_r = p_r S + \sigma L \tag{19}$$

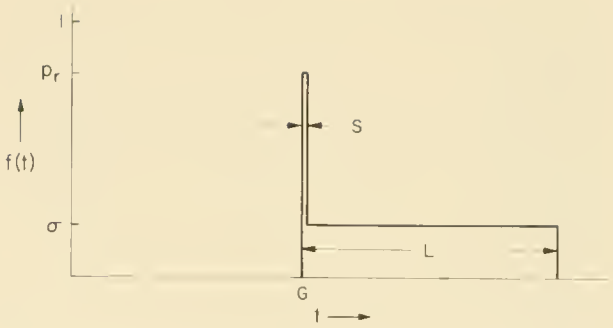


Figure 2-11. *Radiative history of a civilization assuming interstellar communication exists.*

We might, not unreasonably, expect σL to be as great as 10^8 years. Taking $\tau = 3 \times 10^{-5}$ years we then find from (17)

$$L_S = \frac{3 \times 10^{-3}}{F} \text{ years} \tag{20}$$

which means we might have to search 100 to 10,000 stars, requiring only from 1 day to 4 months, depending on F. If interstellar communication already exists and its participants have good reason to continue it, the acquisition of our first contact may be far easier than we would otherwise expect.

We thus come to the viewpoint that interstellar communication is a phenomenon with some of the attributes of life itself. It is difficult to explain how it got started, but once started it tends to perpetuate itself. Just as for life itself, interstellar communication is unlikely to have originated in any *single* trial (by some earlier race), but there have probably been millions of attempts over billions of years only one of which needs to have been successful to start the whole process. We can think of many situations that may have triggered interstellar communication initially:

1. A race realizes its primary star is nearing the end of its main sequence lifetime and simply transmits its history and knowledge with no hope of reply. Detection of these messages would provide other races with the existence-proof of other life needed to justify a prolonged search.
2. Planetary systems exist for which $n_e > 1$, and advanced life on one planet discovers reasonably evolved life on another planet, thus encouraging the effort at contact with other planetary systems.
3. The random distribution of stars places two or more advanced cultures fortuitously close to each other so that first efforts quickly bear fruit. Perhaps leakage signals are detected first. This situation is more likely in dense regions such as star clusters or the galactic nucleus.

The imaginative reader will have no difficulty adding to this list.

We postulate that interstellar communication, having spread rapidly throughout the Galaxy once it began, is now a reality for countless races. We postulate also that participation in this linked community of intelligent beings confers advantages that greatly outweigh the obligations. We view the qualifications for admission as a combination of technical prowess, which the Cyclops study shows we have, plus faith that the expectation of success justifies the effort, which we may or may not have—yet.

REFERENCES

1. Cocconi, G. and Morrison, P., *Nature,* 184, 844 (1959).
2. Cameron, A.G.W., (ed), Interstellar Communication, Benjamin Press (1963).
3. Sagan, C., and Shklovski, I.S., Intelligent Life in the Universe, Holden-Day Inc., 1966.
4. Dole, S.H., Habitable Planets for Man, Blaisdell (Ginn & Co.) 1964 (revised 1970).
5. Vnezemnye tsivilizatssi (Extraterrestrial Civilizations) Proceedings of a conference. Byurakan 20-23 May 1964—Izd. AN Arm SSR, 1965.
6. Kaplan, S.D., (ed), Extraterrestrial Civilizations (Problems of Interstellar Communication), NASA TT F-631 (1971).
7. Drake, F.D., Intelligent Life in Space, Macmillan, 1967.
8. Sullivan, W., We are not Alone, McGraw-Hill, 1966.
9. Mallove, E.F., and Forward, R.L., Bibliography of Interstellar Flight, Hughes Research Lab Report No. 439.
10. Saslaw, W., and Jacobs, K., (eds), The Emerging Universe, University Press of Virginia (1972).
11. Rassool, S.I., Physics of the Solar System, NASA SP300, (1972).
12. Gerstenkorn, H., Uber die Gezeitenreibung beim Zweikorperproblem, Z. Astrophys. *36,* 245 (1955).
13. Goldreich, P., History of the Lunar Orbit, Rev. Geophysic *4,* 411 (1966).
14. Oparin, A.I., Life: its nature, origin and development. Oliver and Boyd, London (1961).
15. Miller, S.L., Production of some organic compounds under possible primitive earth conditions., J. Amer. Chem. Soc., *77,* 2351-61 (1955).
16. Ponnamperuma, C., & Gabel, N.W., Current status of chemical studies on the origin of life. Space Life Sciences, *1,* 64-96 (1968).
17. Anbar, M., Cavitations during impact of liquid water on water: geochemical implications. Science, *161,* 1343-44 (1968).
18. Hodgson, G.W., and Ponnamperuma, C., Prebiotic porphyrin genesis—porphyrins from electric discharge in methane, ammonia and water vapor. Proc. Nat. Acad. Sci. U.S., *59,* 22-28 (1968).
19. Barghoorn, E.S., and Schopf, J.W., Microorganisms three billion years old from the early preCambrian of South Africa. Science, *156,* 508-512 (1967).
20. Schrödinger, E., What is Life? Macmillan, New York (1946).
21. Oehler, D.Z., Schopf, J.W., and Kvenvolden, K.A., Carbon isotopic studies of organic matter in preCambrian rocks. Science, *175,* 1246-1248 (1972).
22. Darwin, C., Origin of Species, J. Murray, London (1859).
23. Schmidt, M., The Rate of Star Formation, Astrophysical Journal, vol. 129, Number 2, page 243 (1959).
24. Schmidt, M., The Rate of Star Formation, II Astrophysical Journal, Vol. 137, Number 3, page 759 (1963).
25. Kreifeldt, J.G., A formulation for the number of communicative civilizations in the Galaxy, Icarus, Vol. 14, No. 3, p. 419-30 (June 1971).
26. Von Hoerner, S., The Search for Signals From Other Civilizations, Interstellar Communication, A.G.W. Cameron (ed), W.A. Benjamin, Inc., New York (1963) pages 272 et seq.

3. SOME REASONS FOR THE SEARCH

It is clear that establishing contact with intelligent life on another world is a major undertaking involving expenditures comparable to those of the Apollo program. There are many reasons for mounting an effort on this scale, but most if not all of them are based upon the assumption that the human race will survive as a socially and scientifically evolving species for a very long time. If this is not true, if our society is going to exhaust our mineral and fuel reserves, or overpopulate our planet, or extinguish itself through atomic war in the next few decades, or centuries, or even millennia, then there is very little reason other than curiosity to carry out the quest. Indeed, there is very little reason to carry out, on any large-scale, socially supported programs such as our present space program or research into biochemistry, medicine, cosmology, particle physics, or the mysteries of the brain, if today's world—the product of four and one half billion years of evolution on Earth—is the end of the road. Rather, like so many of our youth are doing to our distress, we should drink our nepenthe, turn inward to lives of sensuality and subjectivity, and only occasionally weep for glory unattained.

Underlying the quest for other intelligent life is the assumption that man is not at the peak of his evolutionary development, that in fact he may be very far from it, and that he can survive long enough to inherit a future as far beyond our comprehension as the present world would have been to Cro-Magnon man. To do this, man must, of course, solve the ecological problems that face him. These are of immediate and compelling importance and their solution must not be delayed, nor effort to solve them be diminished, by overexpenditures in other areas. But if the cost of the quest is, in one sense, competitive with the cost of solving ecological problems, we can nevertheless do both; and, in a larger sense, the quest gives more significance to survival and therefore places more, not less, emphasis on ecology. The two are intimately related, for if we can survive (and evolve) for another aeon the chances are that many other races have also, and this reduces the problem of making contact.

CONTINUING ADVENTURE

The sixteenth and seventeenth centuries brought an excitement and stimulation that we are prone to forget today. The discovery of the New World, the circumnavigation of the Earth, and the development of trade routes to the East brought cultures into contact that had long been isolated. There followed a period of trade and cultural enrichment in which change and growth were much more rapid than in the centuries before. With the explorations of the last three centuries, and particularly with the advent of modern technology, there are no longer any significant frontiers on earth for geographic or cultural discovery. But the tradition of cultural expansion is still strong, still exciting.

To a considerable extent the possibility of exploring other planets, of possibly finding them habitable, or of finding other life there led to the enthusiastic support of the space program. When the surface temperature of Venus was found to be too hot for life and when the Mariner photos showed the surface of Mars to be too similar to that of the airless Moon to cause one to expect advanced life there, the conviction grew that except for Earth itself the solar system is probably barren. Our planet seen from space was so beautiful to behold, so in contrast to the scarred faces of the Moon or Mars that the importance of saving our own world before exploring others grew larger in the popular mind. Support for the space program dwindled, and the clamor for ecology grew.

Looking far ahead, suppose we are successful in controlling our fecundity, in recycling our wastes, and in developing new energy sources. Will we then accept forever the ceiling on our growth imposed by the finite size of Earth? Will we be content to know forever only one advanced life form—ourselves? Or will such stasis sap us of our lust? In other words, after ecology, what?

BIO-COSMOLOGY

We can, of course, continue to probe the universe, the atom, and ourselves. But so long as our cosmology is limited to unraveling the evolution of the physical universe and our molecular biology to unraveling the complex chemistry of Earth-based life alone, so long as we are limited to physical cosmology and to geocentric biology, many enormously exciting and very fundamental questions will remain unanswered; these include:

1. Are we alone? Is Earth unique not only in the solar system but in the universe? Failure to discover other life could never answer this question in the affirmative, but the discovery of only *one* other race would provide the needed counterexample and supply at least partial answers to questions that follow.
2. How prevalent is life in the universe? Over what range of environmental conditions can it occur?
3. Is it the result of independent starts, or are planets "seeded"?
4. Is the biochemistry of life unique or are there alternatives to DNA?
5. What is the typical longevity of planetary cultures?
6. Are evolving life forms very sensitive chemically and morphologically to small differences in environmental factors, or does there tend to be an optimum design for a highly evolved species? In other words is evolution highly divergent or convergent?
7. Is interworld communication common or exceptional; does a galactic community of cultures exist?
8. Is there interstellar space travel, or merely intercommunication?
9. Does life itself serve a role in the evolution of the physical universe, perhaps modifying it in some way; or does it exist completely at the mercy of the latter?
10. What is our destiny? Do cultures survive the death of their primary stars? Of a collapsing universe?

These are only a few of the many questions that might be answered by, and perhaps *only* by, establishing contact with another race. The day this happens will be the birthdate for us of a new science, which we might call biocosmology.

OUR GALACTIC HERITAGE

Since we are only recently able to signal over interstellar distances, or to detect such signals, any race we contact will be at least as advanced as we in the technologies involved, and probably more so. There are thus some potentially valuable tutorial benefits to science and technology from contact with other cultures. However, the round trip delay times are apt to be on the order of a century or more, so any information exchange will not be in the form of a dialogue with questions asked and answers given, but rather in the nature of two semi-independent transmissions, each a documentary exposition of the salient facts about the society doing the sending—its planetary data, its life forms, its age, its history, its most important knowledge and beliefs, any other cultures it may have already contacted, etc. Thus, although over the course of a century or more we might receive a tremendous amount of information, the rate of reception, the gaps in the picture, and the effort needed to construct a model of the other race from the data received would prevent any violent cultural shock.

It seems virtually certain that if we are successful in establishing interstellar contact we will not be the first civilization to have done so. In fact it may well be that interstellar communication has been going on in our Galaxy ever since the first intelligent civilizations evolved in large numbers some four or five billion years ago. One of the consequences of such extensive heavenly discourse would be the accumulation by all participants of an enormous body of knowledge handed down from race to race from the beginning of the communicative phase. Included in this galactic heritage we might expect to find the totality of the natural and social histories of countless planets and the species that evolved: a sort of cosmic archeological record of our Galaxy. Also included would be astronomical data dating from several aeons ago, perhaps pictures of our own and neighboring galaxies taken by long dead races that would make plain the origin and fate of the universe.

If such a heritage exists it will not only illuminate our future, but the past as well. Access to such a treasury would certainly be worth the cost of Cyclops many times over.

Far more important in the long run than the "synchronization" of the scientific development of the

cultures in contact would be:

1. The discovery in one another of the social forms and structures most apt to lead to self-preservation and genetic evolution.
2. The discovery of new aesthetic forms and endeavors that lead to a richer life.
3. The development of branches of science not accessible to one race alone but amenable to joint efforts.
4. The end of the cultural isolation of the human race, its entry as a participant in the community of intelligent species everywhere, and the development of a spirit of adult pride in man, rather than childish rivalry among men.

Indeed the salvation of the human race may be to find itself cast in a larger role than it can at present visualize, one that offers a cosmic future but one that requires a reorientation of our philosophy and of our mores to fulfill.

POSSIBLE HAZARDS OF CONTACT

We have suggested several potential benefits of contact with extraterrestrial intelligence. We should also consider and attempt to evaluate any possible risks that might attend exposing our existence to an alien culture or cultures more advanced (and therefore more powerful) than ours. These risks range from annihilation to humiliation, but can be conveniently grouped into categories: (1) invasion, (2) exploitation, (3) subversion, and (4) cultural shock. Let us consider these in turn.

Invasion

By revealing our existence, we advertise Earth as a habitable planet. Shortly thereafter we are invaded by hordes of superior beings bent on colonizing the Galaxy. Mankind is annihilated or enslaved. Although this is a recurrent theme of science fiction, the facts do not appear to justify it as a real danger.

If, as we suspect, interstellar travel is enormously expensive even for an advanced culture (see Chap. 4), then only the most extreme crisis would justify mass interstellar travel. We feel we can dismiss the quest for additional living space as a motivation since any race capable of interstellar emigration would have already solved its population problems long ago by internal means. It is not inconceivable that a race might seek to avert extinction by mass exodus before its primary star leaves the main sequence. If so, we would conjecture that they would not wish to add the problems of combat to those of the journey itself and would seek habitable but uninhabited worlds. Such planets might have been located long in advance by the galactic community or by probes sent by the race in question. If so, affiliation with a galactic community might confer security rather than risk.

If, on the other hand, interstellar travel is much easier than we predict, we would argue that to maintain radio silence is no real protection, for in this case a galactic survey would not need to depend on beacons. The question to be answered in this case is Enrico Fermi's: *Where are they?*

Exploitation

The possibility has been voiced that to a very advanced race we might appear such a primitive life form as to represent delightful pets, interesting experimental animals, or a gourmet delicacy.[1] The arguments against invasion as a threat apply with even more force to these fears, for the motivations are less compelling. In addition, we might argue, albeit anthropocentrically, that compassion, empathy, and respect for life correlate positively with intelligence, though counterexamples are not hard to find.

Subversion

A more subtle and plausible risk is that an alien culture, under the guise of teaching or helping us might cause us to build devices that would enable the alien culture to gain control over us. A computer-controlled experiment in biochemistry, for example, might be used to create their life form here. There is no limit to the kinds of threats one can imagine given treachery on their part and gullibility on ours. Appropriate security measures and a healthy degree of suspicion are the only weapons.

Cultural Shock

Finally, there is the possibility that mere contact with an obviously superior race could be so damaging to our psyches as to produce retrogression rather than cultural advancement even with the best intentions on the part of the alien culture. Although many scientists might accept with equanimity positive proof of superior life on other worlds, is mankind as a whole prepared for this? The concept is certainly anathematic to most religions.

Sociologists point out that historically contact between two terrestrial cultures has usually, if not always, resulted in the domination of the weaker by the stronger. We would argue that there is no example where such domination has occurred *by radio* only. The domination has always involved physical contact and

[1] Differences in biochemistry might equally well make us deadly poisonous.

usually territorial expansion by the stronger culture. Where such aggression has been absent the lesser culture has often survived and prospered. The natives of certain South Sea islands have greatly improved their well-being as a result of improved skills and medical knowledge gained through contact.

In the interstellar communication case, the long delays and initially slow information rate should allow us to adapt to the new situation. After all, generations might be required for a round trip exchange.

We cannot assert that interstellar contact is totally devoid of risk. We can only offer the opinion that, in all probability, the benefits greatly outweigh the risks. We cannot see that our security is in any way jeopardized by the *detection* of signals radiated by other life. It is when we *respond* to such signals that we assume any risks that may exist. Before we make such a response or decide to radiate a long-range beacon,[2] we feel the question of the potential risks should be debated and resolved at a national or international level.

[2] It is worth noting that we are already detectable out to perhaps 50 light-years, or will be when our UHF TV signals have propagated that far.

4. POSSIBLE METHODS OF CONTACT

Several methods of achieving contact with intelligent life beyond our solar system have been proposed. These include actual interstellar space travel, the dispatching of interstellar space probes, and the sending and detection of signals of some form. Many other suggestions involving as yet unknown physical principles (or a disregard of known principles) have also been made but are not considered here.

INTERSTELLAR TRAVEL

The classical method of interstellar contact in the annals of science fiction is the spaceship. With our development of spacecraft capable of interplanetary missions it is perhaps not amiss to point out how far we still are with our present technology from achieving practical *interstellar* space flight, and indeed how costly such travel is in terms of energy expenditure even with a more advanced technology.

Chemically powered rockets fall orders of magnitude short of being able to provide practical interstellar space flight. A vehicle launched at midnight from a space station orbiting the Earth in an easterly direction and having enough impulse to add 10-1/2 miles/sec to its initial velocity would then escape earth with a total orbital speed around the Sun of 31-1/2 miles/sec. This would enable the vehicle to escape the solar system with a residual outward velocity of 18 miles/sec, or about 10^{-4} c. Since the nearest star, α-Centauri, is 4 light-years away, the rendezvous, if all went well, would take place in 40,000 years. Clearly we must have at least a thousandfold increase in speed to consider such a trip and this means some radically new form of propulsion.

Spencer and Jaffe (ref. 1) have analyzed the performance attainable from nuclear powered rockets using (a) uranium fission in which a fraction $\epsilon = 7\times10^{-4}$ of the mass is converted to energy and (b) deuterium fusion for which $\epsilon = 4\times10^{-3}$. The mass ratios required for a two-way trip with deceleration at the destination are given in table 1 for various ratios of the ship velocity v to the velocity of light c.

Table 1

$\frac{v}{c}$	Uranium fission	Deuterium fusion
0.1	3.8×10^{4}	8.1×10^{1}
0.2	2.3×10^{9}	6.2×10^{3}
0.3		1.1×10^{6}
0.4		1.5×10^{8}

From these figures we would conclude that with controlled fusion we might make the trip to α-Centauri and back in 80 years, but that significantly shorter times are out of the question with presently known nuclear power sources.

Let us ignore all limitations of present day technology and consider the performance of the best rocket that can be built according to known physical law. This is the photon rocket, which annihilates matter and antimatter converting the energy into pure retrodirected radiation. The mass ratio μ required in such a rocket is:

$$\mu = \sqrt{\frac{1 + v/c}{1 - v/c}} = \frac{v_{eff}}{c} + \sqrt{1 + \left(\frac{v_{eff}}{c}\right)^2} \qquad (1)$$

where

$$v_{eff} = \frac{v}{\sqrt{1 - v^2/c^2}} = \text{coordinate distance travelled per unit of ship's proper time.}$$

Figure 4-1 shows μ, μ^2, and μ^4 as a function of v_{eff}/c.

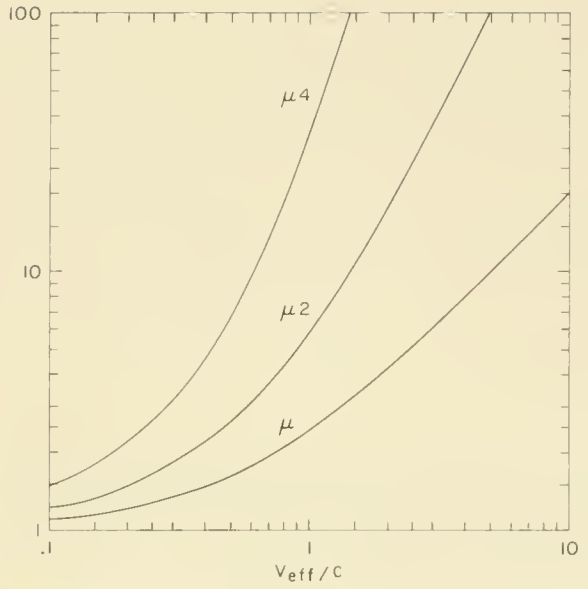


Figure 4-1. *Performance of the ideal photon rocket.*

If we choose $v_{eff}/c = 1$, then to reach the α-Centauri system, explore it, and return would take at least 10 years' ship time. It is hard to imagine a vehicle weighing much less than 1,000 tons that would provide the drive, control, power, communications and life support systems adequate for a crew of a dozen people for a decade. To accelerate at the start and decelerate at the destination requires a mass ratio of μ^2 and to repeat this process on the return trip (assuming no nuclear refueling) requires an initial mass ratio of μ^4. For $v_{eff}/c = 1$, $\mu = 1 + \sqrt{2}$ and $\mu^4 \approx 34$.

Thus the take-off weight would be 34,000 tons, and 33,000 tons would be annihilated enroute producing an energy release of 3×10^{24} J. At 0.1 cent per kWh this represents one million billion dollars worth of nuclear fuel. To discover life we might have to make well over 10^4 such sorties. The energy required for a single mission, used in another way, would keep a 1000-MW beacon operating for over 30 million years.

Even disregarding the cost of nuclear fuel, there are other formidable problems. If the energy were released uniformly throughout the trip, the power would be 10^{16}W. But the ship is μ^3 times as heavy for the first acceleration as for the last and the acceleration periods are only a fraction of the total time; hence, the initial power would be about two orders of magnitude greater, or 10^{18}W. If only one part in 10^6 of this power were absorbed by the ship the heat flux would be 10^{12}W. A million megawatts of cooling in space requires about 1000 square miles of surface, if the radiating surface is at room temperature. And, of course, there is the problem of interstellar dust, each grain of which becomes a miniature atomic bomb when intercepted at nearly optic velocity.

We might elect to drop v_{eff}/c to 0.1 and allow 82 years for the trip. But this would undoubtedly require a larger payload, perhaps a 10,000 ton ship, so our figures are not changed enough to make them attractive.

Ships propelled by reflecting powerful Earth-based laser beams have been proposed, but these decrease the energy required only by the mass ratio of the rocket they replace, and they require cooperation at the destination to slow them down. In addition, the laser powers and the mirror sizes required for efficient transmission are fantastically large.

Bussard (ref. 2) has proposed an ingenious spaceship that would use interstellar hydrogen both as fuel and propellant. After being accelerated by some other means to a high initial velocity, the ship scoops up the interstellar gas and ionizes and compresses it to the point where proton-proton fusion can be sustained, thereby providing energy for a higher velocity exhaust. Essentially the ship is an interstellar fusion powered ramjet. Bussard's calculations show that, for a 1 g acceleration, a 1000 ton ship would require about 10^4km^2 of frontal collecting area. No suggestions are given as to how a 60 mile diameter scoop might be constructed with a mass of less than 1000 tons. (10^4km^2 of 1 mil mylar weighs about 250,000 tons.) Solutions to this and other engineering details must await a technology far more advanced than ours.

A sober appraisal of all the methods so far proposed forces one to the conclusion that interstellar flight is out of the question not only for the present but for an indefinitely long time in the future. It is not a physical impossibility but it is an *economic* impossibility at the present time. Some unforeseeable breakthroughs must occur before man can physically travel to the stars.

INTERSTELLAR PROBES

Bracewell (ref. 3) has suggested that an advanced civilization might send probes to other stars. These would be designed to park in an appropriate orbit and scan the radio spectrum for radiation leaking from any planet circling the star. Assuming an increase of an order of magnitude or more in electronic reliability and sophistication over the present state of the art, and given sufficient energy from "solar" panels or a nuclear plant in the probe, such monitoring could continue for centuries. When radiation was detected, the probe would attract attention to its presence by, say, repeating the detected signal on the same frequency thus producing a long delayed echo. When it became clear that the probe

had been discovered, it could begin a series of transmissions conveying information about the sending civilization, and about how to contact it.

Bracewell also suggests we be alert for such probes in our own solar system. Villard (ref. 4) has suggested that long delayed echoes, which are in fact occasionally heard, conceivably could originate from such a probe. Until the source of such echoes can be definitely ascribed to some mechanism such as slow propagation in the ionosphere near the plasma cutoff frequency, this will continue to be an intriguing, albeit an unlikely, possibility. The phenomenon deserves further study.

An interstellar monitor probe could be much smaller than a spaceship and could take longer in flight. But although there would be no crew to face the psychological barriers or physiological problems of generations spent in space, there are still good reasons to require a short transit time. If the probe were to require a thousand years (or even only a century) to reach its destination, serious doubt would exist that it would not be obsolete before arrival. Thus even probes should be capable of velocities on the order of that of light. With this in mind probe weights in excess of a ton would almost certainly be needed.

To "bug" all the likely stars within 1000 light-years would require about 10^6 probes. If we launched one a day this would take about 3,000 years and an overall expenditure on the order of well over $10 trillion. Interstellar probes are appealing as long as *someone else sends them*, but not when we face the task ourselves.

The simple fact is that it will be enormously expensive, even with any technological advance we can realistically forecast, to send sizable masses of matter over interstellar distances in large numbers.

SERENDIPITOUS CONTACT

We cannot rule out the possibility that we might stumble onto some evidence of extraterrestrial intelligence while engaged in traditional archeological or astronomical research, but we feel that the probability of this happening is extremely small. Not everyone shares this view. Dyson, for example, has suggested (refs. 5, 6) that very advanced civilizations may have such powerful technologies that they can engage in engineering efforts on a planetary or stellar scale. He visualizes civilizations rebuilding their planetary systems to create additional habitable planets, planetoids, or huge orbiting space stations and thus provide more *lebensraum*. (See Appendix B.) Noting that this would increase the surface area at which radiation was occurring at temperatures on the order of 300° K, he suggests that such civilizations might be detectable as a result of the excess radiation in the 10μ range, and concludes that a rational approach to the search for extraterrestrial intelligence is indistinguishable from an expanded program in infrared astronomy. While we admire Dyson's imagination we cannot agree with this conclusion, much as we might like to see more infrared astronomy for its own sake.

In the first place we would argue that Dyson's premises are far more speculative than our conclusions. It is one thing to admit the *possibility* of Dyson's civilizations; it is quite another thing to consider them *probable* enough to base our entire search strategy on their existence. We submit that, to a civilization capable of the feats he describes, the construction of an extremely powerful beacon and interstellar search system would be child's play and that Dyson's civilizations would have pursued this course long before engaging in the remodeling of their planetary systems. (Further discussion of the motivations for, problems associated with, and detectability of Dyson's civilizations is given in Appendix B.)

However, let us assume that a super-civilization *does* rebuild its planetary system, not into the artificial structures Dyson describes, but into replicas of the home planet. These at least would offer living conditions to which the species was already adapted. Would such systems be detectable? The Sun radiates about 500,000 times as much energy in the 10μ region as the Earth and about 100,000 times as much as all the planets combined. If all the heavy elements in the Sun's planets were reassembled at about 1 AU from the Sun, about nine more Earthlike planets could be constructed and the present planetary 10μ radiation would be doubled to become only 50,000 times less than that of the Sun. This is still a long way from being a detectable increase.

Let us further assume that on each of these new earths the civilization releases one thousand times as much energy as we do. Earth traps about 10^{17} watts of sunlight, which it must then re-radiate. Geothermal and tidal heat is negligible by comparison. At present our world-wide rate of release of energy from coal, oil, natural gas and nuclear sources totals about 6×10^{12} watts, or about one sixteen thousandth part of the sunlight we receive. Even a thousandfold increase in this rate on each of the one old and nine new earths would raise their total infrared radiation only 6% leaving it still negligible compared with the Sun.

Finally, even if we were to detect a star with more than twice as much 10μ radiation as we would expect from the visible light output, would we not conclude merely that it was surrounded by a lot of dust, instead of a super civilization? The excess IR radiation lacks the hallmark of intelligence which combines a high degree of

order with what Morrison has described as an "extraordinary richness of combinatorial complexity." Even with the very sophisticated coding, an information-bearing microwave signal would stand out as an artifact. Excess incoherent IR radiation certainly would not.

Naturally we should keep our eyes open for artifacts—for unusual radiation of any kind. But while some may feel pessimistic about the chances of success of a concerted search for signals of intelligent origin, the very magnitude of the effort needed for the search to be effective convinces us that the chances of accidental contact are negligible.

INTERSTELLAR COMMUNICATION ALTERNATIVES

Although no one can deny the excitement that would accompany a physical visit to another inhabited world, most of the real benefit from such a visit would result from communication alone. Morrison has estimated that all we know about ancient Greece is less than 10^{10} bits of information; a quantity he suggests be named the *Hellas*. Our problem therefore is to send to, and to receive from, other cultures not tons of metal but something on the order of 100 Hellades of information. This is a vastly less expensive undertaking.

Fundamentally, to communicate we must transmit and receive energy or matter or both in a succession of amounts or types or combinations that represent symbols, which either individually or in combination with one another have *meaning*—that is, can be associated with concepts, objects, or events of the sender's world. In one of the simplest, most basic, types of communication the sender transmits a series of symbols, each selected from one of two types. One symbol can be the presence of a signal, the other can be the absence of a signal, for example. This type of communication is called an *asymmetric binary channel*. For the receiver to be able to receive the message, or indeed, to detect its existence, the amount of energy or number of particles received when the signal is present must exceed the natural background. Suppose we knew how to generate copious quantities of neutrinos and to beam them. And suppose we could capture them efficiently (sic !) in a receiver. Then, with the signal present, our receiver would have to show a statistically significant higher count than with no signal.

Even if the natural background count were zero, the probability of receiving no particles when the signal is in fact present should be small. Since the arrivals during a signal-on period are Poisson-distributed, the expectation must be several particles per on-symbol. Thus to conserve transmitter power we must seek particles having the least energy. The desirable properties of our signaling means are:

1. The energy per quantum should be minimized, other things being equal.
2. The velocity should be as high as possible.
3. The particles should be easy to generate, launch, and capture.
4. The particles should not be appreciably absorbed or deflected by the interstellar medium.

Charged particles are deflected by magnetic fields and absorbed by matter in space. Of all known particles, photons are the fastest, easiest to generate in large numbers and to focus and capture. Low-frequency photons are affected very little by the interstellar medium, and their energy is very small compared with all other bullets. The total energy of a photon at 1420 MHz is one ten billionth the kinetic energy of an electron travelling at half the speed. Almost certainly electromagnetic waves of some frequency are the best means of interstellar communication—and our *only* hope at the present time.

REFERENCES

1. Spencer, D. F., and Jaffe, L. D., Feasibility of Interstellar Travel, NASA TR32-233 (1962).
2. Bussard, R.W., Galactic Matter and Interstellar Flight, Astronautica Acta, Vol. VI, Fasc. 4 (1960).
3. Bracewell, R. N.: Communications from Superior Galactic Communities. *Nature, 186,* 4726, May 28, 1960 (670-671).
4. Villard, O. G., Jr.; Fraser-Smith, A. F.; and Cassan, R. T.: LDE's, Hoaxes, and the Cosmic Repeater Hypothesis QST, May 1971, LV, 5, (54-58).
5. Perspectives in Modern Physics, R. E. Marshak (ed.) John Wiley & Sons, 1966, p. 641.
6. Interstellar Communication, A. G. W. Cameron (ed.) W. A. Benjamin, Inc., New York (1963), pp. 111-114.

5. COMMUNICATION BY ELECTROMAGNETIC WAVES

Having decided on electromagnetic waves as the only likely interstellar communication means, we then must ask: "Is it possible with reasonable power and antenna sizes to signal over the enormous distances involved? and: Is there an optimum region in the spectrum?" The answers to these questions involve the interplay of a number of factors that determine the performance of a communication link. In this chapter we discuss these factors as they relate to systems from radio through optical frequencies.

ANTENNA GAIN AND DIRECTIVITY

We shall use the term *antenna* to mean any coherent collector or radiator of electromagnetic waves. At optical frequencies, antennas take the form of lenses or concave mirrors that focus the received energy from distant point sources into diffraction-limited spots in an image plane. The angular dimensions and intensity distribution of the optical image of a point source bear the same relation to the wavelength and the telescope objective diameter that the beam width and pattern of a large radio antenna bear to its diameter and operating wavelength. However, because there is no radio counterpart of photographic film, radio telescopes typically are built to examine only one resolvable direction at a time rather than to form an extended image.

A universal antenna theorem states that the effective area of an isotropic radiator is $\lambda^2/4\pi$, and that this is the effective area of any antenna averaged over all directions (ref. 1). An antenna whose area is greater than $\lambda^2/4\pi$ in some direction must have an effective area less than $\lambda^2/4\pi$ in other directions, and is therefore *directive.* For a uniformly illuminated aperture of an area A, the power gain, g_0, on axis is the ratio of A to $\lambda^2/4\pi$, that is,

$$g_0 = \frac{4\pi A}{\lambda^2} = \left(\frac{\pi d}{\lambda}\right)^2 \tag{1}$$

where the last equality applies if the antenna is circular and of diameter d.

For a uniformly illuminated circular aperture, the ratio of gain, g, or intensity, I, at an angle θ off axis to the gain, g_0, or intensity, I_0, on axis is

$$\frac{g}{g_0} = \frac{I}{I_0} = \left[\frac{2J_1\left(\frac{\pi d}{\lambda}\theta\right)}{\frac{\pi d}{\lambda}\theta}\right]^2 = \left(\frac{2J_1(\theta\sqrt{g_0})}{\theta\sqrt{g_0}}\right)^2 \tag{2}$$

where J_1 is the first order Bessel function. The gain and intensity fall to one half their on-axis values when $\theta = \theta_{1/2} = (0.5145\ldots)(\lambda/d)$ so the beamwidth between half power points is

$$2\theta_{1/2} = (1.029\ldots)\frac{\lambda}{d} \text{ radians} \tag{3}$$

THE FREE SPACE TRANSMISSION LAW

Assume a power P_t is radiated isotropically. At a distance R this power will be uniformly distributed over a sphere of area $4\pi R^2$. The amount received by an antenna of area A_r is therefore $P_r = P_t A_r/4\pi R^2$. If the transmitting antenna is now made directive, with a power gain g_t in the desired direction, we will have

$$\frac{P_r}{P_t} = \frac{g_t A_r}{4\pi R^2} \tag{4}$$

The quantity $g_t P_t$ is often called the *effective radiated power* P_{eff}. Thus an alternate form of equation (4) is

$$P_r = \frac{g_t P_t A_r}{4\pi R^2} = \frac{P_{eff} A_r}{4\pi R^2} \tag{5}$$

Making use of equation (1), we may write equation (4) either in the form given by Friis (ref. 1):

$$\frac{P_r}{P_t} = \frac{A_t A_r}{\lambda^2 R^2} \tag{6}$$

or in terms of the gains of both antennas

$$\frac{P_r}{P_t} = \left(\frac{\lambda}{4\pi R}\right)^2 g_t g_r \tag{7}$$

The appearance of λ^2 in the denominator of equation (6), shows that *for fixed antenna areas*, the transmission increases as the square of the operating frequency. However, larger antenna areas are more readily realized at lower frequencies. As one scales all *linear* dimensions of a given structure by a factor k, the deflections of the structure under its own weight vary as k^2. Since we can tolerate deflections proportional to λ, we can let k^2 vary with λ, which means that both A_t and A_r can be proportional to λ, and the λ^2 disappears. But this is only part of the story. Both in the optical region and by the use of phased arrays in the microwave region, we can build antennas having the same *gain*—that is, the same diameter measured in wavelengths. In both regions the maximum practical gain is limited by the directivity becoming so high that atmospheric turbulence or pointing errors prevent us from reliably keeping the beam on target. Typically this difficulty occurs in both spectral ranges when $2\theta_{1/2}$ is on the order of 1 arc sec and $g_0 \approx 10^{11}$. Thus equation (7) is a more appropriate form of the transmission equation to use when we are considering the ultimate realizable performance. The factor λ^2 now appears in the numerator and makes *low* frequencies appear more attractive. However, we are paying for the increased performance in the cost of the increased antenna areas needed to realize a given gain.

When the transmitting system is not under our control, as is the case when we are searching for radio leakage or beacons radiated by extraterrestrial life, equations (4) and (5) are most appropriate and we shall use this form of the transmission law in developing our range equations.

NOISE IN COHERENT RECEIVERS

The noise level of any ideal coherent amplifier of electromagnetic waves is:

$$\psi = h\nu \left(\frac{1}{e^{h\nu/kT} - 1} + 1 \right) \tag{8}$$

where

ψ = noise power per Hertz

h = Planck's constant

k = Boltzmann's constant

ν = frequency

T = temperature of source or field of view

The first term in the parentheses is black-body radiation in a single propagation mode while the second term arises from spontaneous emission in the amplifier. At high frequencies such that $h\nu/kT >> 1$ the black-body radiation or thermal noise term disappears and $\psi \approx h\nu$. At low frequencies where $h\nu/kT << 1$, we have $\psi \approx h\nu + kT \approx kT$ and since this is the usual situation in radio engineering, it has become customary to describe the system noise in terms of an equivalent noise temperature $T_n \equiv \psi/k$. Thus from equation (8)

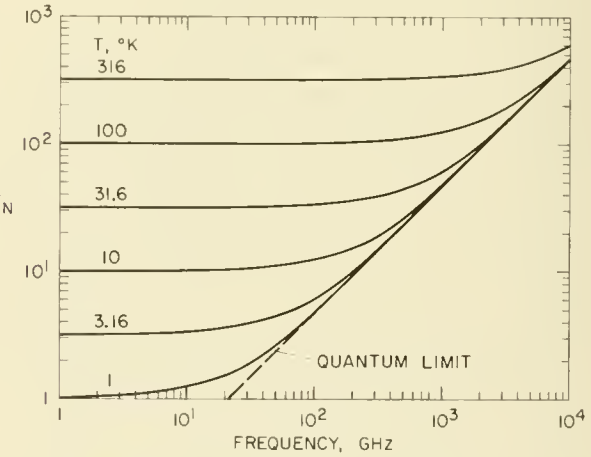


Figure 5-1. *Noise in ideal receivers.*

$$T_n = \frac{h\nu}{k} \left(\frac{1}{e^{h\nu/kT} - 1} + 1 \right) \tag{9}$$

Figure 5-1 shows the behavior of $T_n(\nu)$ for various source temperatures, T. We see that $T_n(\nu)$ is a monotonically increasing function of frequency.

In general, if the receiver is not ideal, but generates noise of its own, this noise may be included by increasing T in equations (8) and (9). Strictly speaking, this is not quite correct. Several noise sources at temperatures T_1, T_2 . . . T_m do not have the same total blackbody radiation as a single source at a temperature $T = T_1 + T_2 + \ldots T_m$; the latter has a higher frequency quantum cutoff. However, the total noise at low frequencies is proportional to the sum of the temperatures, while at high frequencies the spontaneous emission term dominates, so the overall result is not greatly in error.

COHERENT DETECTION AND MATCHED FILTERING

To realize the maximum signal to noise ratio in a coherent receiver a synchronous (or homodyne) detector must be used; in addition, the transmission band must be shaped to match the signal spectrum (Appendix C). When both these techniques are employed, the effective noise power spectral density is half that given by equation (8) or (9). Neither technique can be used until the signal has been discovered in the first place, so in the search process we must rely on square-law (energy) detectors. But once a coherent signal has been found, a local oscillator may be locked to it, and the nature of the signal spectrum can then be determined. Thus, coherent detection and matched filtering would be employed in any communication link once contact had been established.

In a synchronous detector, the coherent RF or IF signal is mixed with a local oscillator signal of the same frequency and phase and any modulation of the former is recovered by low pass filtering. Any noise wave present may be resolved into two statistically independent waves $a(t)\cos 2\pi\nu_0 t$ and $b(t)\sin 2\pi\nu_0 t$, where $\cos 2\pi\nu_0 t$ is the local oscillator wave. Both $a(t)$ and $b(t)$ will, in general, be gaussian variables whose mean square values each represent one-half the total noise power. Only $a(t)$ will produce an output after the low pass filter, so that only half the noise power present at the input appears in the output. But since (in double sideband reception) the postdetection bandwidth is half the predetection bandwidth, the noise power spectral density is unaffected.

If we know in advance the waveform or spectrum of the signal we are trying to receive (or if we can determine either of these), we can design the receiver to give the best estimate of some property of the signal (e.g., its amplitude or phase or time of occurrence) in spite of the added noise. In Appendix C we show that if the signal consists of a series of pulses, the best ratio of detected peak signal amplitude to the rms noise fluctuations in that amplitude will be obtained if the receiver band limiting filter has a complex amplitude transmission $K(\nu)$ given by

$$K(\nu) = m\frac{\overline{F(\nu)}}{\psi(\nu)} \tag{10}$$

where m is an arbitrary real constant, $\overline{F(\nu)}$ is the complex conjugate of the signal amplitude spectrum, and $\psi(\nu)$ is the noise power spectral density. This equation applies to filtering done either before or after synchronous detection, provided the appropriate spectra are used.

If $\psi(\nu)$ is a constant (white noise), the matched filter has a transmission everywhere proportional to the conjugate of the signal spectrum. The conjugacy aligns all Fourier components of the signal so that they peak simultaneously, while the proportionality weights each component in proportion to its own signal-to-noise ratio.

With a matched filter the output peak-signal to noise power ratio is

$$\frac{S}{N} = 2\int_{-\infty}^{\infty}\frac{|F(\omega)|^2}{\psi(\nu)}\,d\nu \tag{11}$$

Again, if the noise is white, so that $\psi(\nu)$ has the constant value ψ_0, equation (11) becomes simply

$$\frac{S}{N} = \frac{2E}{\psi_0} \tag{12}$$

where E is the energy per pulse. We see from this expression that the ultimate detectability of a signal depends on the ratio of the received energy to the spectral density of the noise background. The energy of the signal can be increased, of course, by increasing the radiated power, but once a practical limit of power has been reached, further increase is only possible by increasing the signal duration. This narrows the signal spectrum. In the limit, therefore, *we would expect interstellar contact signals to be highly monochromatic.*

As a simple example of a matched filter assume that the signal consists of a train of pulses of the form

$$f_i(t) = A_i\frac{\sin \pi B(t-t_i)}{\pi B(t-t_i)} \tag{13}$$

which might result from synchronous detection of pulses of the form

$$g_i(t) = A_i \frac{\sin \pi B(t-t_i)}{\pi B(t-t_i)} \cos 2\pi\nu_0 t \qquad (14)$$

The spectrum of a single detected pulse is constant from 0 to $B/2$ Hz while that of a single pulse prior to detection is flat from $\nu_0 - (B/2)$ to $\nu_0 + (B/2)$ Hz. Both spectra are zero outside these limits. The matched filter is either a predetection ideal bandpass filter of width B centered at ν_0 or a postdetection ideal low pass filter of cutoff frequency $B/2$. (In this case, because the two in tandem are idempotent, both may be used.) Since in this case the matched filter does not affect the signal spectrum the signal shape is unchanged.

If $t_i = i/B$ in equations (13) and (14) the peak pulse amplitudes will be independent. Thus independently detectable pulses may be sent at a separation of $\tau = 1/B$ sec, and we may consider τ to be the effective pulse duration. We thus have

$$B\tau = 1 \qquad (15)$$

a relation that is approximately true for any matched filter where B is the RF bandwidth and τ is the pulse duration, both appropriately defined.

With white noise, the signal to noise ratio from equation (11) or (12) is

$$\frac{S}{N} = \frac{2E_i}{\psi_0} = \frac{2A_i^2\tau}{\psi_0} = \frac{A_i^2}{\psi_0(B/2)} \qquad (16)$$

If a long train of such pulses is sent with a constant $A_i = A$, the transmitted signal amplitude will be constant at the value A representing a signal power A^2 and the received signal will show fluctuations about this amplitude of mean square value $\psi_0 (B/2)$ representing the noise power of spectral density ψ_0 in the band $B/2$.

We see that the combination of synchronous detection and matched filtering gives an output signal-to-noise ratio that is twice the received signal-to-noise ratio if the latter is defined as the ratio of the signal power to the total noise power in the RF band.

NOISE IN ENERGY DETECTION

The energy or power of a received signal may be determined by using a *square law detector* (a square law *device* followed by a low pass filter) to rectify the output of a coherent receiver. Alternatively, at infrared and optical frequencies the radiation may be allowed to fall directly onto a photon counter. A photocell, like a square-law detector, gives a response (current) proportional to the incident instantaneous power. In principle, the limiting noise performance of both types of detector is the same; however, at radio frequencies most of the output noise is produced by fluctuation in the random noise input and shot noise is usually negligible, whereas at optical frequencies the reverse may be true.

If a steady coherent signal of power P_r and an incoherent (black-body radiation or thermal noise) background of power P_0 are both applied to a square law detector or photocell the output signal-to-noise power ratio is shown in Appendix D to be

$$\frac{S}{N} = \frac{\tau P_r^2}{(h\nu/\eta)(P_r + P_0) + (P_0/B)[(2/m)P_r + P_0]} \qquad (17)$$

where

τ = integration time

B = predetection bandwidth

m = 1,2 = number of orthogonal polarizations reaching detector

η = quantum efficiency of detector

The ratio of P_0/B to $h\nu/\eta$ determines the relative importance of the ratio of fluctuation noise to shot in the detector output. If we let ϕ represent this ratio, then

$$\phi = \frac{P_0/B}{h\nu/\eta} = \frac{\text{spectral density of fluctuation noise}}{\text{spectral density of shot noise}}$$

$$= \frac{(\eta P_0)/B}{h\nu} \qquad (18)$$

In the last form we see that ϕ is the expected number of background photons counted in the time $1/B$. If $\phi >> 1$ fluctuation noise predominates; if $\phi << 1$ shot noise predominates.

At radio frequencies $m = 1$ and $P_0/B \approx kT >> h\nu$. Equation (17) then becomes

$$\frac{S}{N} = (B\tau)\frac{(P_r/P_0)^2}{1 + 2(P_r/P_0)} \qquad (19)$$

In this expression ($B\tau$) is the number n of independent samples that are averaged; P_r/P_0 is the *input* signal-to-noise ratio and S/N is the *output* signal-to-noise ratio. If $(P_r/P_0) >> 1$ we see that $(S/N) \approx (n/2)(P_r/P_0)$. At high signal levels the square-law detector gives one fourth the signal-to-noise ratio of the synchronous detector, for which $(S/N) = 2n(P_r/P_0)$. If $(P_r/P_0) << 1$ then $(S/N) \approx n(P_r/P_0)^2$ so that to hold a given output signal to noise ratio as the input signal level drops below the noise, n must vary inversely as the *square* of (P_r/P_0).

At optical frequencies ϕ may be much less than unity, and fluctuation noise may be unimportant. The only effect of background light is then to add to the dispersion in the photon count when a signal is present, and to produce a count in the absence of signal.

In the absence of any background ($P_0 = 0$) equation (17) becomes

$$\frac{S}{N} = \eta \frac{P_r \tau}{h\nu} \qquad (20)$$

which is simply the expected signal photon count $\bar{n}$ in the time τ. If P_r is constant, there is a constant probability per unit time of detecting a signal photon. The counts therefore have a Poisson distribution, for which the rms fluctuation in count is $\sqrt{\bar{n}}$. The output signal-to-noise ratio is therefore $(\bar{n}/\sqrt{\bar{n}})^2 = \bar{n}$ as given by equation (20).

In a signaling system τ must equal the symbol duration (Nyquist interval) which in a coherent detection system is $1/B$. In this sense the spontaneous emission noise of $h\nu B$ watts in a coherent receiver and the shot noise in a perfect photon detector ($\eta = 1$) are comparable. We cannot escape the quantum noise limit of a coherent receiver by going to direct photodetection.

THE MICROWAVE WINDOW

Any antenna pointed at the sky sees the 3° K isotropic background radiation that fills the universe. The background radiation begins to fall off at about 60 GHz, but by then quantum noise has taken over and the total noise level rises above 60 GHz, ultimately becoming proportional to frequency.

Galactic noise, mainly due to synchrotron radiation, varies with direction, being strongest in the direction of the galactic center and weakest near the galactic poles. Galactic noise falls rapidly with increasing frequency as shown in Figure 5-2. Note that at a galactic latitude of $|b| = 5°$, the galactic noise is about twice that near the poles ($|b| = 90°$). Thus for 90% of the sky the galactic noise lies between the limits shown.

Together these three noise sources (galactic, 3° K background, and quantum noise) define a broad quiet region of the spectrum extending roughly from 1 to 100 GHz that is nearly the same for observers anywhere in the solar neighborhood or similar regions of the galactic disk. This is the free-space microwave window. In addition, if the atmosphere of the receiving planet contains oxygen and water vapor (as is very likely if "they" are like us), there will be absorption lines at 22 and 60 GHz. These increase the noise temperature markedly above 10 GHz. Except at 60 GHz the absorption from these lines is not serious but the reradiation noise is. This noise varies with elevation angle so the curves on Figure 5-2 are only approximate typical values.

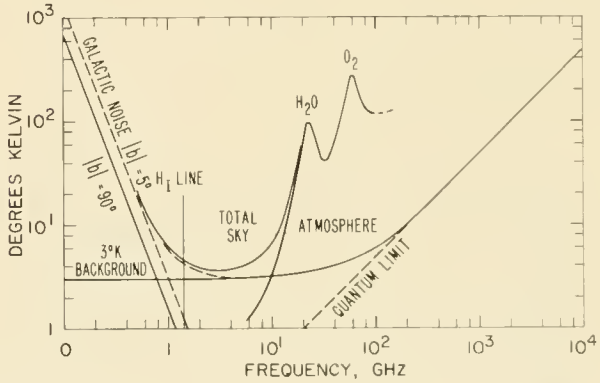


Figure 5-2. *Sky noise temperature for coherent receivers.*

The different noise contributions have been added in Figure 5-2 to give the curve marked "Total Sky." We note that the sky noise temperature is about 4° K in the vicinity of the hydrogen line and that the minimum value of about 3.6° K occurs at twice this frequency. On planets with heavier atmospheres the minimum temperature would be shifted to slightly lower frequencies. Since the power required for signaling is proportional to the total noise temperature, the microwave window from 1 to 10 GHz is clearly an attractive region for a ground-based receiver on an Earth-like plant.

STAR NOISE

In attempting interstellar communication, or in attempting to detect radiation from another technologically advanced civilization, we will in general be pointing our receiving system directly at the primary star of the planetary system because:

1. We will not have enough resolving power to separate the planet from the star at the distances involved. (One AU subtends 1 arc sec at a distance of 1 parsec or 3.26 light-years.)
2. Even if we had the resolution, we would not know the position of the planet relative to the star and would not like to add another dimension to the search.

As a result we need to determine how much noise the star itself adds.

One convenient measure of the star noise is the increase in system noise temperature it produces. This temperature increase is given by

$$\Delta T = \frac{1}{4\pi} \int_{\theta=0}^{\pi} \int_{\phi=0}^{2\pi} T(\theta,\phi) g(\theta,\phi) \sin\theta \, d\theta \, d\phi \qquad (21)$$

where

T = temperature of field of view

g = antenna gain

The quantity ΔT is simply the average over all sources in the field, with each direction weighted according to the antenna gain in that direction. For a star on axis, $g(0,\phi)$ will be constant at the value $4\pi A_r/\lambda^2$ from $\theta = 0$ out to $\theta = d_*/2R$, where d_* is the diameter of the star and R is the range. Over this same range of θ, $T(\theta,\phi)$ will have the value T_*, then drop to the normal background value (which we have already included) for $\theta > d_*/2R$. Thus for a star

$$\Delta T = \frac{2\pi A}{\lambda^2}\left(1 - \cos\frac{d_*}{2R}\right) T_* \approx \frac{\pi d_*^2 A_r}{4\lambda^2 R^2} T_* \qquad (22)$$

$$\frac{\Delta T}{T_*} = \frac{A_* A_r}{\lambda^2 R^2}$$

where $A_* = \pi d_*^2/4$ is the projected area of the star. This relation has a familiar appearance; it is the free space transmission law with A_* now playing the role of the transmitting antenna aperture area and T replacing power.

The nearest stars of interest are at about 10 light-years range. Let us therefore compute ΔT for the sun at a distance of 10 light-years. Because of its corona and sunspot activity, the Sun's effective temperature at decimeter wavelengths greatly exceeds its surface temperature. From curves published by Kraus (ref. 2) we take the following values to be typical:

TABLE 5-1

λ	$T_\odot$ (quiet)	$T_\odot$ (active)
1 cm	6000°K	7000°K
3.16	1.6×10^4	6×10^4
10	5×10^4	1.5×10^6
31.6	1.7×10^5	4×10^7
100	5×10^5	8×10^8

If we assume an antenna diameter of 10 km and take $d_* = 1.392\times10^9$ m and $R = 9.46\times10^{16}$ m = 10 light-years, we derive from equation (22) the values of ΔT given in Table 5-2.

TABLE 5-2

λ	ΔT (quiet sun)	ΔT (active sun)
1 cm	0.8° K	0.93° K
3.16	0.214° K	0.8° K
10	0.067° K	2° K
31.6	0.023° K	5.3° K
100	0.0067° K	10.7° K

We conclude that for antenna diameters up to 10 km and for stars beyond 10 light-years, star noise at microwave frequencies, though detectable, will not be a serious problem unless the stars examined show appreciably more activity than the Sun.

The ability of Cyclops to detect normal stars is studied in Appendix R. We estimate that the 3-km equivalent array should be able to see about 1000 stars at λ = 10 cm and about 4000 at λ = 3.16 with a 1-min integration time.

Another measure of star noise, and a particularly convenient one for determining the radiated intensity necessary to outshine the star, is the total radiated power per Hertz of bandwidth. Since this measure will be important in the optical and infrared part of the spectrum, we may assume that only blackbody radiation is involved. The total flux radiated by a star having a diameter d_* and surface temperature T_* is

$$\Phi(\nu) = \frac{2\pi^2 d_*^2}{c^2} \frac{h\nu^3}{e^{h\nu/kT_*} - 1} \text{ W/Hz} \qquad (23)$$

Figure 5-3 is a plot of this function for the sun. We see that to equal the brightness of the sun the total power of an omnidirectional beacon (or the effective radiated power of a beamed signal) at 10μ would have

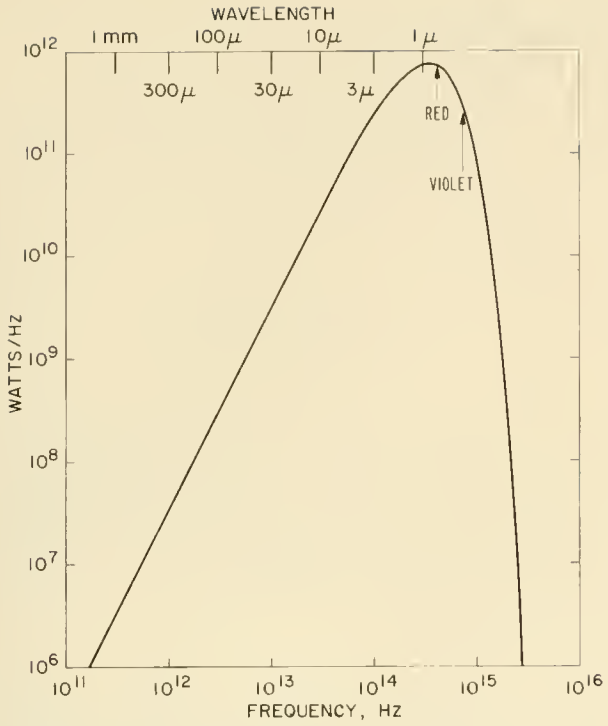


Figure 5-3. *Radiation of the Sun versus frequency.*

to be 2.5×10^{10} W/Hz of receiver bandwidth, while at 1μ the required power would be 7.4×10^{11} W/Hz.

RANGE LIMITS

The maximum range over which communication can take place with a given system depends on the coding used and on our definition of an acceptable error rate. Since our objective is to find expressions for the maximum range that will enable us to compare different systems fairly, the exact standards we choose are not too important so long as we are consistent. We shall assume that the message is transmitted in a binary code. For systems using incoherent detection, a "one" is sent as a pulse and a "zero" as a space. This is known as an asymmetric binary channel. With a coherent receiver and a synchronous detector we can detect pulses of either polarity and can therefore send a "one" as a positive pulse and a "zero" as a negative pulse. This type of communication is the symmetric binary channel. The *phase* of the transmitted signal is reversed to change from a one to a zero or vice versa.

We will *define* the range limit for a coherent receiver as the range at which the received signal-to-noise ratio is unity, and the range limits for all other systems as the range at which these systems have the same probability of error per symbol received.

Assuming the receiving antenna to be a clear circular aperture of diameter d_r, and allowing for losses, equation (4) may be rewritten as

$$P_r = \frac{d_r^2}{16R^2}\,\eta P_{eff} \tag{24}$$

where η is the overall quantum efficiency (i.e., the product of the transmission of both atmospheres, optical or radio receiving antenna efficiencies, photodetector quantum efficiency).

Sychronous Detection

Here, by our definition, we simply set P_r equal to the total system noise referred to the input. Since we may wish to synchronously detect signals heterodyned down from the optical region, we set

$$P_r = (\psi + \psi_*)\,B = \psi B + \frac{d_r^2}{16R^2}\frac{\eta}{\alpha_1}\frac{m}{2}\,\Phi(\nu)B \tag{25}$$

where ψ is given by equation (8) using the appropriate system noise temperature, $\Phi(\nu)$ is given by equation (23) and α_1 (which is also a factor included in η) is the transmission of the atmosphere of the sending planet. The coefficient m = 1,2 is the number of orthogonal polarizations reaching the detector. If we now define the star noise background ratio b_* as

$$b_* = \frac{\text{star noise power}}{\text{signal power}} = \frac{m\,\Phi(\nu)B}{2\,\alpha_1 P_{eff}} \tag{26}$$

then we find from equations (24) and (25)

$$R = \frac{d_r}{4}\left(\eta\frac{P_{eff}(1-b_*)}{\psi B}\right)^{1/2} \tag{27}$$

We note that if $b_* > 1$ the maximum range is imaginary; that is, there is no *real* range at which the signal-to-noise ratio is unity.

Bit Error Rate. We must now determine the bit error rate for the synchronous detector with *unity input signal-to-noise ratio.* Since the output signal to noise power ratio is 2, and the noise is gaussian, the probability density function when a positive output is present will be

normal with the mean at $\sigma\sqrt{2}$. When the signal is negative, the mean will be at $-\sigma\sqrt{2}$ as shown in Figure 5-4. If the *a priori* probability of either polarity is one

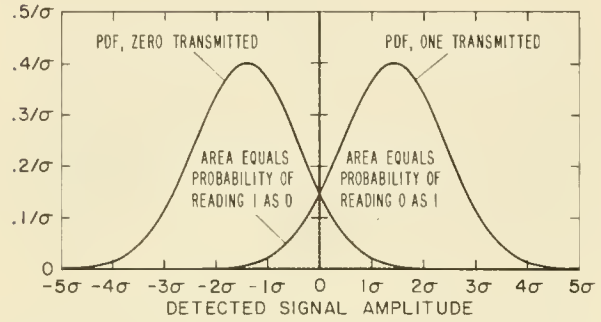


Figure 5-4. *Probability density functions–symmetric binary channel (S/N = 2).*

half, the decision threshold should be set at zero amplitude. The probability of a pulse being incorrectly read is the area of the tail of either distribution that extends across the threshold; that is,

$$
\begin{aligned}
p_e &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{(x-\sigma\sqrt{2})^2}{2\sigma^2}}\, dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{(x+\sigma\sqrt{2})^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{2}}^{\infty} e^{-z^2/2}\, dz \qquad (28) \\
&= 0.07865 \ldots
\end{aligned}
$$

This is the bit error rate we will require of all our detection methods in determining the range limits.

The n-Fold Average. When n signal samples are averaged, the signal amplitudes (being correlated) remain unchanged, while the rms noise fluctuation, if uncorrelated, is reduced by $\sqrt{n}$. Receiver noise is uncorrelated between different receivers, but star noise may be spatially correlated.

If we assume the star noise is spatially correlated but temporaly uncorrelated, then combining simultaneous signals from n antennas gives a slightly different result from averaging n successive samples from the same antenna or array. For n antennas, we find

$$R = \frac{d_r}{4}\left[\eta \frac{P_{eff}\,(1-b_*)}{\psi B}\, n\right]^{1/2} \qquad (29)$$

If successive time samples are averaged we find

$$R = \frac{d_r}{4}\left[\eta \frac{P_{eff}\,(n-b_*)}{\psi B}\right]^{1/2} \qquad (30)$$

so that time averaging (or averaging the signals from sufficiently widely separated antennas) does permit an adverse background ratio to be overcome. In the systems we shall consider, $b_* \ll 1$ and the distinction between equations (29) and (30) is negligible.

Square Law Detection

Short Averaging Time. If the averaging time, τ, at the output of a square law detector is comparable to or less than $1/B$, then as is shown in Appendix D, the detector output amplitude (which is proportional to the RF power, averaged over an RF cycle) when noise alone is present is Boltzmann distributed. If $x = P/P_0$ where P_0 is the noise power, the probability that P is exceeded is simply

$$p_0(x) = e^{-x} \qquad (31)$$

To get a given error rate (of detecting a pulse when none is present) we set $p_0(x) = p_e$ and solve for x. If $p_e = 0.07865$, then $x = 2.543 \ldots$, which means the threshold must be set 4.053 dB above the mean noise power level, P_0.

When a coherent signal is also present, the probability density function is given by

$$q_1(x) = c^{-(r+x)}\, I_0(2\sqrt{rx}) \qquad (32)$$

where $r = P_r/P_0$, P_r being the received coherent signal power and P_0 being the average noise power; I_0 is the zero-order modified Bessel function of the first kind. The probability that the output fails to exceed a threshold level x_T when a signal is present is therefore

$$p_1(x_T) = \int_0^{x_T} q_1(x)\, dx \qquad (33)$$

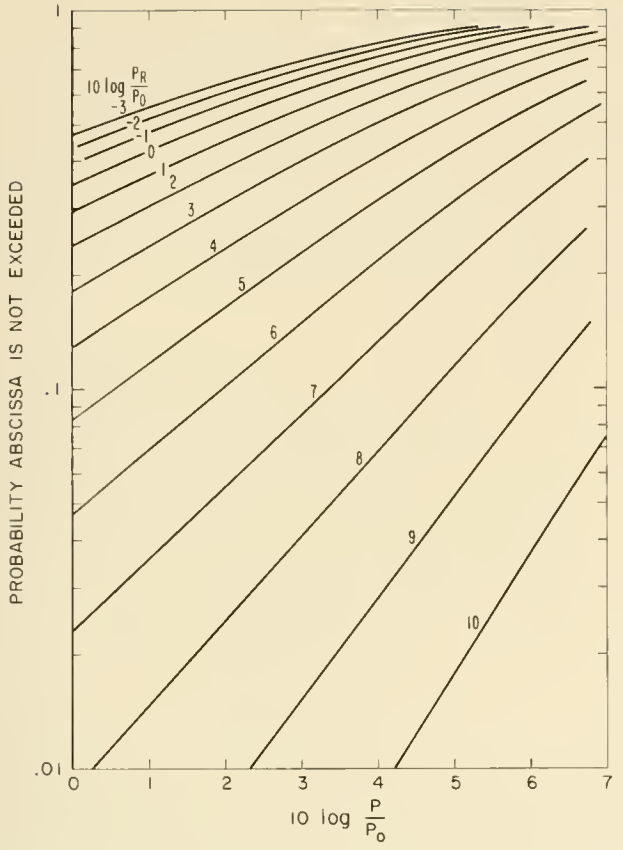


Figure 5-5. *Probability that signal plus noise fails to exceed power* P.

Figure 5-5 shows $p_1(x)$ plotted against 10 log x for various values of r expressed in decibels. By noting the value of $p_1(x)$ for 10 log x = 4.053 dB and plotting this versus the corresponding value of r, we obtain the curve shown in Figure 5-6. We see that if the probability of failing to exceed the threshold is to be 0.07865, then r, the *input* signal to noise ratio must be 6.06 (i.e., 7.83 dB).

Thus for a square law detector with short time averaging ($\tau \lesssim 1/B$), the limiting range is

$$R = \frac{d_r}{4}\left[\eta \frac{P_{eff}\,(0.165 - b_*)}{\psi B}\right]^{1/2} \tag{34}$$

where we have written 0.165 for $(6.06)^{-1}$.

The n-Fold Average. If $y = \frac{1}{n}\sum_{i=0}^{n} x_i$ is the average of n samples of the output of a square law detector, then, as shown in Appendix D, if the noise in successive samples is uncorrelated, the probability that this average exceeds a given threshold value of $y = y_T$ is

$$p_0(y_T) = e^{-y_T} \sum_{k=0}^{n-1} \frac{{y_T}^k}{k!} \tag{35}$$

If both signal and noise are present, the probability that the same threshold is *not* exceeded is:

$$p_1(y_T) = \int_0^{y_T} n\left(\frac{y}{r}\right)^{\frac{n-1}{2}} e^{-n(r+y)}\, I_{n-1}(2n\sqrt{ry})dy \tag{36}$$

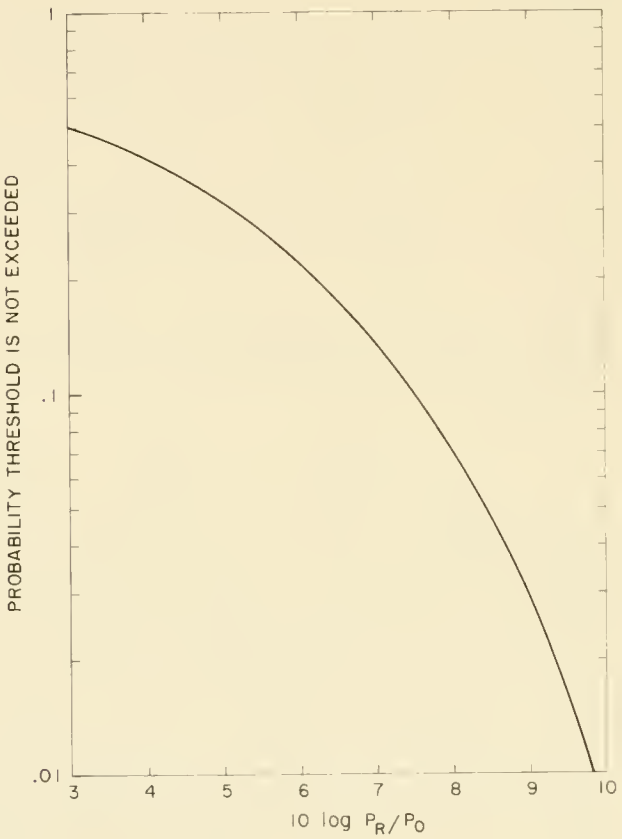


Figure 5-6. *Probability of missing signal.*

The same graphical method can be used with these expressions that was illustrated with equations (31) and (33) for the case n = 1. (This is the procedure used in

Chapter 11 in analyzing the data processing system proposed for Cyclops.)

When the number of samples averaged is large and the noise is uncorrelated, equations (35) and (36) both approach gaussian distributions. We then can find the separation between the means of the two distributions required for a given bit error rate. If the samples from n_r receivers are averaged, the star noise will be correlated. If then $n_s = B\tau$ independent samples of this composite signal are averaged, none of the noise will be correlated, and an overall improvement is produced.

If

$$\frac{2m^2}{n} b_* << 1 \tag{37}$$

where

$b_* \equiv P_*/P_r$ = ratio of star noise power to signal power

m = number of standard deviations threshold is separated from mean of either distribution

n_r = number of receivers averaged

n_s = number of time samples averaged

n = $n_s n_r$

then the required received signal-to-noise ratio is given by

$$\frac{P_r}{P_n} \approx 2\left[\frac{m^2}{n}(1 + 2b_*) + \frac{m}{\sqrt{n}}\right] \tag{38}$$

For the bit error rate allowed at our range limit, $m = \sqrt{2}$ and

$$\frac{P_r}{P_n} \approx \frac{4(1 + 2b_*) + \sqrt{8n}}{n} \tag{39}$$

If $b_* = 0$ and $n = 1$, expression (39) gives $(P_r/P_n) = 6.828$ rather than the value 6.060 we found from the actual statistics. Thus expression (39) should give reasonably accurate results for all $n > 1$.

If we substitute expression (39) into (24) we obtain

$$R = \frac{d_e}{4}\left[\eta\frac{P_{eff}}{\psi B}\,\frac{n}{4(1 + 2b^*) + \sqrt{8n}}\right]^{1/2} \tag{40}$$

In *this* expression d_e is the diameter of each array *element* since it is the signals from these elements that we are averaging. We see that with no time averaging ($n_s = 1$, $n = n_r$) the performance, provided $2(m^2/n_s)b_* << 1$, approaches that of an array having the diameter $(d_e/4)(n/8)^{1/4}$ rather than $(d_e/4)n^{1/2}$ as is true for a coherent array. Thus the advantage of going to a large array is much less if the signals, each containing uncorrelated receiver noise are added after square law detection.

Photon Detection

Short Averaging Time. In equation (17), if $P_0/B << h\nu/\eta$ most of the output noise will be shot noise and very little will be due to fluctuations in the incoherent background power. However, we may still include part of the latter by using the actual statistics for the photon count with incoherent illumination.

If the integration time $\tau \lesssim 1/B$, the background alone will give a count having a probability distribution

$$q_0(n) = \frac{1}{1 + \bar{n}_0}\left(\frac{\bar{n}_0}{1 + \bar{n}_0}\right)^n \tag{41}$$

where $\bar{n}_0$ is the expectation count due to the background light in the time τ; that is, $\bar{n}_0 = \eta(P_0\tau/h\nu)$, where η is the quantum efficiency, and P_0 is the received background power. The corresponding distribution with signal alone present would be

$$q_r(n) = \frac{\bar{n}^n e^{-\bar{n}}}{n!} \tag{42}$$

where $\bar{n}$ is the expectation count of signal photons. This is the Poisson distribution resulting from the constant probability per unit time of receiving a signal photon.

With both signal and noise we have approximately

$$q_1(n) = q_0(n) * q_r(n) \tag{43}$$

where the $*$ signifies the discrete convolution of the two distributions. This result ignores the cross-power term between the signal and the in-phase component of the noise amplitude, and will therefore yield a slightly optimistic range.

The probability that the background count equals or exceeds a certain threshold count in the time τ is

$$p_0(n) = \sum_{i=n}^{\infty} q_0(i) = \left(\frac{\bar{n}_0}{1 + \bar{n}_0}\right)^n \tag{44}$$

The probability that the threshold is *not* exceeded in the same observation time τ is simply $1 - p_0(n)$. The probability that the threshold is not exceeded at any time period τ in a much longer time T is then $[1 - p_0(n)]^{T/\tau} \to e^{-p_0(T/\tau)}$, and the probability p_e that the threshold *is* exceeded in this interval is

$$p_e = 1 - e^{-p_0(T/\tau)} \tag{45}$$

which if $p_0(T/\tau) << 1$ is approximately $p_0(T/\tau)$. Thus we must have

$$p_0(n) \leqslant \frac{\tau}{T} p_e \tag{46}$$

From equation (44) we see this requires that the threshold n_T be such that

$$n_T > \text{int}\left[\frac{\ell n(\tau/T)p_e}{\ell n(\bar{n}_0/1 + \bar{n}_0)}\right] > n_T - 1 \tag{47}$$

where "int" means the integer part of. We next require $\bar{n}$ to be such that

$$p_1(n) = \sum_{n=0}^{int\, n_T} q_1(n)dn \leqslant p_e \tag{48}$$

The required $\bar{n}$ can be determined by trial and error using numerical integration of expression (43). When the background is partly starlight and we change $\bar{n}$ by changing the range, we must preserve the constraint that

$$\bar{n}_0 = \bar{n}_b + b_* \bar{n} \tag{49}$$

where $\bar{n}_b$ is any background count that is independent of n. This means recomputing n_T each time using equation (47) but the process converges rapidly.

As an example, suppose $\bar{n}_b = 0$, $b_* = 1.2 \times 10^{-3}$, $(\tau/T) = 10^{-9}$ and $p_e = 0.07865$. Then if $\bar{n} = 10, \bar{n}_0 = 0.012$ and equation (47) shows that $5 < n_T < 6$. A plot of equation (48) with $\bar{n} = 9.7$ is given in Figure 5-7. We see that for $5 < n_T < 6$ the probability of missing a pulse is very nearly 0.07865 as required. The small change from $\bar{n} = 10$ to $\bar{n} = 9.7$ does not affect n_T.

When $b_* = \bar{n}_0/\bar{n} << 1$, changes in $\bar{n}_0$ do not affect the result very much. If in the previous example we let $\bar{n}_b = 0.008$ (which almost doubles $\bar{n}_0$), n_T is unaffected (because of quantization) and so is $\bar{n}$.

Having determined $\bar{n}$, the range is given by equation (51).

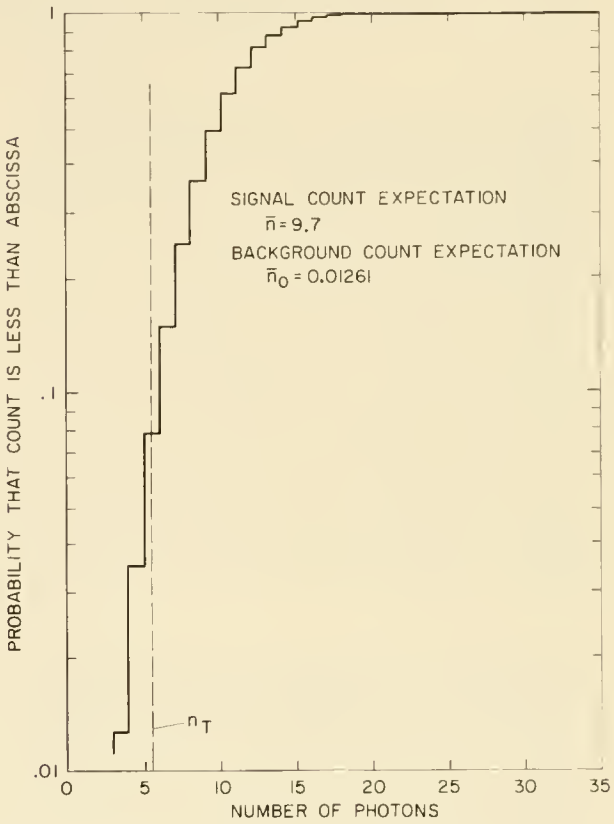


Figure 5-7. *Photon count distribution.*

Long Averaging Time. When the averaging time is long compared with the reciprocal optical bandwidth ($B\tau >> 1$) and the background count is large, the distribution approaches the gaussian form with its mean value at $\bar{n}_0$ and its standard deviation $\sigma_0 = \sqrt{\bar{n}_0}$. Simlarly, the distribution of the signal plus background is gaussian with its mean at $(\bar{n}_0 + \bar{n})$ and its standard deviation $\sigma = \sqrt{\bar{n}_0 + \bar{n}}$. To achieve the desired bit error rate the separation, $\bar{n}$, between the means must be some multiple m of the sum of the standard deviations; that is, $\bar{n} = m(\sigma_0 + \sigma)$. If we again let $\bar{n}_0 = \bar{n}_b + b_* \bar{n}$ we find

$$\bar{n} = m^2\left[1 + 2b_* + 2\sqrt{b_*(1 + b_*)} + (\bar{n}_b/m^2)\right] \tag{50}$$

For our assumed error rate ($p_e = 0.07865$), $m = \sqrt{2}$.

The range limit is then found by equating $\bar{n}h\nu$ to ηP_r and is given by

$$R = \frac{d_r}{4}\left[\eta \frac{E_{eff}}{\bar{n}h\nu}\right]^{1/2} \tag{51}$$

where $E_{eff} = P_{eff}\tau$ and d_r is the array diameter.

If $\bar{n}_b = 0$, then from (50) the required $\bar{n}$ is independent of the array size and the range increases in direct proportion to d_r, as is true for a coherent array. If $\bar{n}_b \neq 0$, and is proportional to area (as for sky light) or number of elements (as for detector dark current), then increasing d_r will increase $\bar{n}$, unless $\bar{n}_b$ is due to dark current and the array size is increased by increasing the *size of each element.* When we double d_r by using four times as many elements, $\bar{n}_b$ increases by a factor of four, so $\bar{n}_b = Kd_r^2$. When $\bar{n}_b$ becomes dominating, then $\bar{n}$ is proportional to d_r and the range increases only as $d_r^{1/2}$.

This is the problem with the array using heterodyne mixers (see eq. (40)) where the dominant noise is the spontaneous emission noise $h\nu B$ associated with each mixer.

COMPARISON OF SEVERAL INTERSTELLAR LINKS

We will now use the results of the last section to compare the ranges attainable with two near-optical (1.06μ), two infrared (10.6μ), and two microwave interstellar communication systems. *These are not search systems.* We assume that contact has already been established and that directive antennas can therefore be used at both ends of the link.

The wavelength of 1.06μ was chosen because the high power pulsed neodymium laser operates at this wavelength and there is a window in the atmosphere there. The high power CO_2 laser operates at 10.6μ and there is also a window in the atmosphere at this wavelength. The 3 cm microwave wavelength was chosen because it lies at the upper end of the microwave window, and when directive transmitter antennas are used, a sharper beam is produced for a given antenna size than at lower frequencies. From the data available, we estimate the transmission of the atmosphere at an elevation angle of 45° to be 70% at 1.06μ and 50% at 10.6μ. These figures represent good seeing conditions and drop rapidly with sky overcast. No appreciable attenuation exists in the microwave window under any but the most severe weather conditions.

Laser technology is about one third as old as microwave technology and may therefore be farther from its ultimate state of development. To avoid biasing the comparison because of this historical accident, we have included not only two near state-of-the-art laser systems but also two that represent several orders of magnitude improvement over present technology. These are the optical and infrared *B* systems.

To avoid a horsepower race we have assumed the same power of 100 kW for all systems capable of CW or long pulse operation. This power is at the limit of our present laser technology for the spectral purities we have assumed. Higher power lasers may be developed with higher spectral purity, but orders of magnitude greater power are already available at microwave frequencies. Ultimately the *energy cost,* rather than the technology of converting that energy to radiation at various frequencies, becomes the limiting factor. Hence we feel this equal power assumption is fair.

We have assumed a limiting beamwidth of 1 sec of arc at all wavelengths. This requires that the antennas be pointed to within ±0.3 sec of arc to avoid more than 1 dB signal loss. This may be too sharp a beam to use reliably in view of atmospheric turbulence. The use of a broader beamwidth would favor microwaves.

The same information rate of one bit per second and the same bit error rate has been assumed for all systems. All the incoherent detection systems use on-off pulse modulation; all the coherent detection systems use carrier phase reversal (known as phase shift keying, or PSK) to send the positive and negative pulses of a symmetrical binary channel.

Coherent (phased) arrays have not been included for the optical and infrared systems because the atmospheric resolution limit is easily achieved with single mirrors, and because even in space, pointing accuracy becomes limiting with single mirrors of practical size. In any case, optical phased arrays would pose extremely severe stability problems.

Out of ignorance, we have assumed an Earthlike atmosphere for the other planet. (Ground based laser systems run the risk that the atmospheric windows of another planet might be different.) Also, we have computed star noise on the basis of solar radiation, which is valid at least when we are transmitting. The assumption of polarizing filters in all cases reduces b_* by a factor of two.

Optical Systems

System A. Pulse energies of 600 J with durations of less than 100 picoseconds have been achieved with neodymium glass lasers using amplifier stages. We assume a pulse energy of 1000 J, a pulse duration of 10^{-9} sec, and a pulse repetition rate of 1 per sec, so this is a near state-of-the-art system.

Silicon photodetectors with about 80% quantum efficiency are available at 1μ and provide internal avalanche multiplication. Since they can be cooled to give very low dark currents, we have used photon detection to allow full advantage to be taken of an incoherent receiving antenna array. The array consists of 400 mirrors, each 5 m (200 in.) in diameter, giving an equivalent clear aperture of 100 m.

This system provides the example for required photon count given in the last section (p.47) where an $\bar{n}$ of 9.7 was found to be adequate. The product of the transmission of two atmospheres and the detector quantum efficiency gives an overall efficiency of $\eta = 0.7 \times 0.7 \times 0.8 = 0.4$ with no allowance for interstellar absorption or optical surface losses. The range limit was calculated using equation (51).

System B. Present neodymium lasers are capable of about 300 W CW output with a spectral linewidth of about 100 GHz. For optical system B we assume a 1μ laser with 3000 times the power output, or 100 kW, and a linewidth decrease of 3×10^4 to 3 MHz. Clearly, this system is far beyond the state of the art.

The assumed bandwidth of 3 MHz allows for some drift but nevertheless implies a frequency stability of a part in 10^8. This narrow receiver bandwidth may be achievable using parallel plate Fabry-Perot filters at the collimated outputs of each receiving telescope. With 5-m primary mirrors and a magnification of 50X, the exit pupil will be 10 cm in diameter. Pointing errors and image blur due to atmospheric turbulence will produce a range of input angles for the filter of about $\pm 0.6'' \times 50 = 30''$. At this angle the resonant frequency will shift about 1 part in 10^8 and the "walk-off" of the mode will amount to about 7 cm, so the operation is a little marginal. Nevertheless with a cavity 1-m long and mirrors having 99.4% reflectance, 0.1% absorption and 0.5% transmission, the required Q of 10^8 can be realized. Two cascaded filters and a mop-up interference filter would be needed to produce a single pass-band. The transmission of these would probably not exceed 25%, so η is reduced to 0.1.

Although the bandwidth is 0.003 as wide as in optical system A, the peak power is only 10^{-7} times as large. As a result, the solar background is increased by 3×10^4 to a value $b_* = 36$. Other characteristics are the same as optical A.

The necessary expected signal photon count (292) was calculated from equation (50) with $n_b = 0$, and the range was calculated from equation (51).

Infrared Systems

System A. Here we assume a CO_2 laser capable of CW operation at 100 kW with a line width of 3 kHz. Higher power CO_2 lasers have no doubt been built, but it is unlikely that the spectral power density we specify has been exceeded.

Self-oscillating lasers at this power level exhibit much broader spectral lines because of turbulent changes in index of refraction in the gas, mode jumping due to schlieren effects and cavity vibration. We assume a master oscillator power amplifier system so that these effects cause phase and amplitude, rather than frequency and mode, perturbations.

At 10.6μ no avalanche detectors are known. Quantum efficiencies of 80% appear feasible but, without avalanche multiplication, shot noise in the following amplifier dominates. The best known solution is to use photoelectric mixing to provide substantial power gain, and a noise level approaching $h\nu B$. Each antenna must deliver a spatially coherent signal, and the receiver antenna element size is thus limited to about 2-1/4-m diameter, so that 1975 units are needed for a 100-m array.

At 10.6μ we assume the atmosphere has 50% transmission at 45° elevation angle (on a clear night) so that $\eta = 0.5 \times 0.5 \times 0.8 = 0.2$. Again we have made no allowance for interstellar absorption or surface losses.

At 10.6μ thermal radiation from the (lossy) atmosphere produces only about 2° K increase in the system noise temperature. We find $(h\nu/k) + 2 = 1360°$. The range was calculated from equation (40) using $\eta = n_s n_r = 3000 \times 1975$ and ignoring b_*.

System B. All out for spectral purity! We assume the sky's *not* the limit and postulate a 100 kW CO_2 laser with a line width less than 1 Hz. This implies frequency stabilities of about a part in 10^{14} and is hardly practical in view of Doppler rates (see Chap. 6). But given this fantastic stability we could construct a complete coherent receiver with synchronous detection. However, we can now use only one 2-1/4-m antenna for our receiver. Equation (27) gives the range limit for this system and for the microwave systems to follow.

Microwave Systems

System A. This is a state-of-the-art system using two 100-m antennas, 100 kW of power, a 20° K system noise temperature, and a 1-Hz bandwidth obtained by synchronous detection. The beamwidth of this system is over one minute of arc, so no pointing difficulties should

TABLE 5-3

PARAMETER	OPTICAL A	OPTICAL B	INFRARED A	INFRARED B	MICROWAVE A	MICROWAVE B
Wavelength	1.06μ	1.06μ	10.6μ	10.6μ	3 cm	3 cm
Transmitter:						
Antenna diameter	22.5 cm	22.5 cm	2.25 m	2.25 m	100 m	3 km*
Number of elements	1	1	1	1	1	900
Element diameter	22.5 cm	22.5 cm	2.25 m	2.25 m	100 m	100 m
Antenna gain	4.4×10^{11}	4.4×10^{11}	4.4×10^{11}	4.4×10^{11}	1.1×10^{8}	9.8×10^{10}
Peak or CW power	10^{12} W	10^{5} W	10^{5} W	10^{5} W	10^{5} W	10^{5} W
Modulation	Pulse	Pulse	Pulse	PSK	PSK	PSK
Pulse duration	10^{-9} sec	1 sec	1 sec	1 sec	1 sec	1 sec
Energy per bit	10^{3} J	10^{5} J	10^{5} J	10^{5} J	10^{5} J	10^{5} J
Effective radiated power	4.4×10^{23}W	4.4×10^{16}W	4.4×10^{16}W	4.4×10^{16}W	1.1×10^{13}W	9.9×10^{15}W
Beamwidth (seconds of arc)	1"	1"	1"	1"	64"	1"
Receiver						
Antenna diameter	100 m	100 m	100 m	2.25 m	100 m	3 km*
Number of elements	400	400	1975	1	1	900
Element diameter	5 m	5 m	2.25 m	2.25 m	100 m	100 m
Atmosphere transmission	.7	.7	.5	.5	1	1
Overall Quantum efficiency	.4	.1	.2	.2	.9	.9
Solar background ratio	1.2×10^{-3}	36	1.7×10^{-3}	6×10^{-7}	—	—
Noise temperature	13,600° K	13,600° K	1360° K	1360° K	20° K	20° K
Effective RF bandwidth	1 GHz	3 MHz	3 kHz	1 Hz	1 Hz	1 Hz
Detection method	Photon	Photon	Sq. Law	Synch.	Synch.	Synch.
System:						
Range limit (l.y.)	26	24	22	41	500	450,000
State of the Art?	?	No	?	No	Yes	Yes
All weather?	No	No	No	No	Yes	Yes

*Array spread out to 6.4 km diameter to avoid vignetting.

arise. The required frequency stability is one part in 10^{10}, which is easily realized. Doppler rates would require correction but are only on the order of 1 Hz/sec before correction. This system has the same *collecting area* as the first three laser systems.

System B. This system is the same as microwave A, but the antennas have been enlarged to phased arrays 6.4 km in diameter to provide 1 sec of arc resolution. The array elements are assumed to be spaced a little more than twice their diameter to avoid shadowing at low elevation angles, so the equivalent clear aperture diameter is 3 km. Arrays of these general dimensions are probably needed for the search phase, so their communication capabilities are of more than academic interest. This system has the same *beamwidth* as the laser systems.

Table 5-3 summarizes the characteristics of the six systems studied. We see that all the systems, aside from practical considerations discussed below, can achieve communication over interstellar distances, but that the laser systems have more than one order of magnitude less range. The laser systems can span the distance to some 20 to 140 stars of interest, while microwave system A can reach about a quarter of a million stars of interest. Microwave system B can easily reach every star in the galaxy. If each transmitting antenna element had a 200 kW phased transmitter, the total power of microwave B would be 180 MW and the range would be 20 million light-years. With microwaves we could communicate not merely (sic!) over interstellar distances but over *intergalactic* distances.

Why do the laser systems show up so poorly by comparison? Basically, of course, all laser systems suffer the disadvantage of a higher energy per photon than microwave systems: their effective noise temperature is high. This disadvantage is partly compensated by the ease of obtaining narrow beams, but once we approach

practical limits of narrow beamwidths in the microwave region as well, the photon energy disadvantage shows through. But there are additional specific disadvantages that are revealed by each of the laser systems studied. Optical system A suffers from too little energy per bit. If its pulse energy were equal to that of the other systems, its range would increase over 10 times. Optical system B does not adequately override star noise, so about 30 times as many signal photons must be received per pulse as for optical system A and, because of the filter loss, the received energy per pulse is only 25 times as great at the same range. Infrared system A suffers from the inefficient array utilization because of incoherent addition and the spontaneous emission noise associated with each receiver. Finally, infrared system B suffers from the small size of receiving antenna possible in a coherent system at 10.6μ.

Except possibly for infrared system B, all the laser systems would appear to be far more expensive than microwave A. Certainly to provide two acres worth of optical collecting area in the form of precisely steerable 2- to 5-m telescopes would cost much more than a single 100-m dish. All the systems require cooled detectors. Optical B requires costly precision narrow band filters at each antenna. Infrared A requires precision local lasers and mixers at each of 1975 antennas, while the precision needed for the local oscillator of infrared B may be beyond the capabilities of even the most advanced technology.

The mechanical stability and precision problems of the laser systems have their counterpart in the precision of electrical control needed to compensate for Doppler drifts. The drift due to Earth diurnal Doppler could be as great as 3 kHz/sec at 10.6μ for stars near the meridian and a station near the equator. Locking infrared system B onto such a rapidly drifting signal would be a nightmare. The total interstellar Doppler shift is about $\pm 10^{-3}$ so the filters in the optical systems would require tunability over a ±300 GHz range while the infrared systems would need a ±30 GHz tuning range.

Finally, laser systems suffer from atmospheric absorption even in clear weather and are unusable in cloudy weather. Sky light does not materially affect the performance of any of the systems, except for optical system B. In the daytime, the background count from the sky would reduce the range of this system from 24 to about 20 light-years. Thus all systems would be usable in clear daytime hours as well as at night, but the microwave systems are the only all-weather systems.

In summary

1. The *sole* advantage of lasers is that they permit narrow beamwidths to be realized with small transmitting mirrors, but this advantage turns into a disadvantage at the receiver where we need a large collecting area.
2. The collecting surface is much cheaper and more durable in the microwave region, because the tolerances are much greater and polished surfaces are not needed.
3. Microwave systems offer substantially more range for the same power, even with wider beamwidths.
4. Because of the wider beams, automatic positioning and tracking are feasible at microwaves on a dead reckoning basis.
5. Doppler shifts and drift rates are orders of magnitude less at microwave frequencies and can easily be compensated.
6. Microwave systems are all-weather systems.

We have gone to some length to compare microwaves and lasers because the ease of obtaining sharp beams has led many people to propose laser systems (ref. 3). We have made the comparison for point-to-point communication links where the transmitting directivity of the laser is used. When we come to consider the search problem in the next chapter we will see that transmitter directivity is not an asset, and lasers lose out completely. In fact, we believe that if lasers had been known for the last hundred years and microwaves had only recently been discovered, microwaves would be hailed as the long sought for answer to interstellar communication.

COMMUNICATION RATE OF A MICROWAVE LINK

At the range limits given, the links compared in the last section are all 1 bit/sec systems with a bit error rate of 0.07865. As long as the received energy per bit is held constant, the error rate will not change. Thus the communication rate varies inversely as the square of the

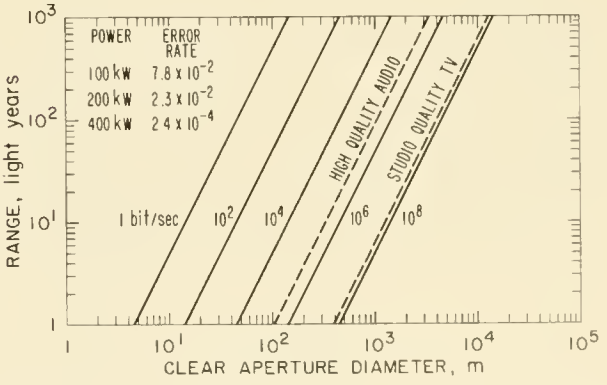


Figure 5-8. *Information rates of microwave links; λ = 3 cm.*

range and directly as the square of antenna diameter (if both antennas are scaled). Figure 5-8 shows, as a function of antenna size and bit rate, the ranges achievable with 100 kW of 3-cm power. A bit error rate of 0.07865 is assumed.

Error-correcting codes permit a dramatic decrease in bit error rate as the signal-to-noise ratio increases above unity (0 dB). As an example, we compare in Table 5-4 the uncoded performance with the coded performance using a 7 bit convolutional code and eight-level maximum likelihood Viterbi decoder. The code *rate* is 1/2 (i.e., the information bit rate is 1/2 the signal bit rate), so the duration and energy per transmitted bit are half as great with coding as without.

TABLE 5-4

S/N, dB	Uncoded		Coded	
	W/kT	Error rate	W/kT	Error rate
3	2.	2.3×10^{-2}	1	1×10^{-3}
4	2.5	1.25×10^{-2}	1.25	3.2×10^{-5}
5	3.16	6×10^{-3}	1.58	7×10^{-7}
6	4.	2.4×10^{-3}	2.	1×10^{-8}

Thus with 200 kW we can achieve the bit rates shown in Figure 5-8 but with an error rate of 10^{-3}.

The bit rate of a good audio channel with 15 kHz bandwidth and 60 dB S/N ratio is 3×10^{5}, while PCM TV with 7 bits per picture element (128 levels) is of good quality and requires about 6×10^{7} bits/sec. These rates have been indicated by the dashed lines. Clearly, if we and the other race were willing to construct microwave arrays representing from one to several kilometers of clear aperture, normal communication rates could be achieved over distances on the order of 100 to 1000 light-years. (With round-trip delays of several centuries, at least we would not be cursed with "talk-back" shows!) The rates are of more than academic interest because arrays of this size will probably be needed to establish contact in the first place and so may already exist in both worlds when that day arrives.

REFERENCES

1. Schelkunoff, Sergei A.; and Friis, Harald T.: Antennas: theory and practice, New York, Wiley, 1952.
2. Krauss, John D.: Radio Astronomy, McGraw Hill, 1966.
3. Schwartz, R.N. and Townes, C.H., Interstellar and Interplanetary Communication by Optical Masers, *Nature*, vol. 190, No. 4772, p. 205 (Apr. 15, 1961).

6. ACQUISITION: THE CENTRAL PROBLEM

The performance figures for microwave links given in the last chapter show that communication over interstellar distances is technically feasible and in fact that rates of 1 bit/sec over distances up to 1000 light-years or 100 bits/sec over distances up to 100 light-years are possible with antennas that are already in existence. But these performance figures assume that two highly directive antennas are pointed at each other (or, more precisely, are pointed in the proper directions at the proper time) and that a receiver is tuned and phase locked to precisely the proper frequency. The reason interstellar communication is not a *fait accompli* is that we do not know where to point our antennas nor on what frequency to listen or send. To determine these "rendezvous coordinates" we must engage in an extensive search program, scanning the sky and the spectrum for signals of a clearly artificial character and perhaps radiating such a signal for other races. In this chapter we examine some important aspects of this search phase.

PROBABILITY OF CONTACT VERSUS RANGE

The first question that naturally arises is: How far out into space must we carry the search to have a high probability of success? If we were able to give a definite answer to this question we would already know the answers to many of the very questions that goad us into making contact in the first place. We do not *know* how prevalent life is. We would like to know. We do not know how long technological civilizations survive, nor how long they attempt communication. We would like to know this too. Nevertheless, we need some idea, if only a rough idea, of how far we must be prepared to search.

Let us assume we have some method of interrogating every likely star to see if one of its planets is radiating artifact signals. Let p be the *a priori* probability that a planet around the star is radiating a signal. Then the probability of *not* finding a signal is $(1 - p)$. After interrogating n stars, the probability of still not finding a signal is $(1-p)^n$. If we assume $p \ll 1$ then

$$\log(1-p)^n = n\left[-p + \frac{p^2}{2} - \frac{p^3}{3} + \ldots\right] \approx -np$$

$$(1-p)^n \approx e^{-np}$$

and the probability of having made at least one contact is

$$p_c = 1 - (1-p)^n = 1 - e^{-np} \qquad (1)$$

To have a 63% chance of success we would have to interrogate $1/p$ stars; for a 95% chance of success, $3/p$ stars.

Allen (ref. 1) gives the total density of main sequence stars of spectral class range $F0$ through $K7$ as $0.088/\text{pc}^3$ or 5.4×10^{-4} / (light-year)^3. The Sun lies very nearly in the galactic plane, and the density of stars averaged over all galactic longitudes is very nearly constant out to a few thousand light-years. Toward the galactic poles, however, the density drops as we leave the galactic plane. From Elvius (ref. 2) we have plotted points showing the falloff in density with height h above or below the midplane of the galactic disk for main sequence stars of types $F8$ through $G8$, representing our prime range of interest (Fig. 6-1). We have approximated the distribution by the purely empirical relation

$$\frac{\rho}{\rho_0} = \left[1 + \left(\frac{h}{h_0}\right)^2\right]^{-3/2} \qquad (2)$$

with $h_0 = 850$ light-years. Figure 6-1 shows the fit of

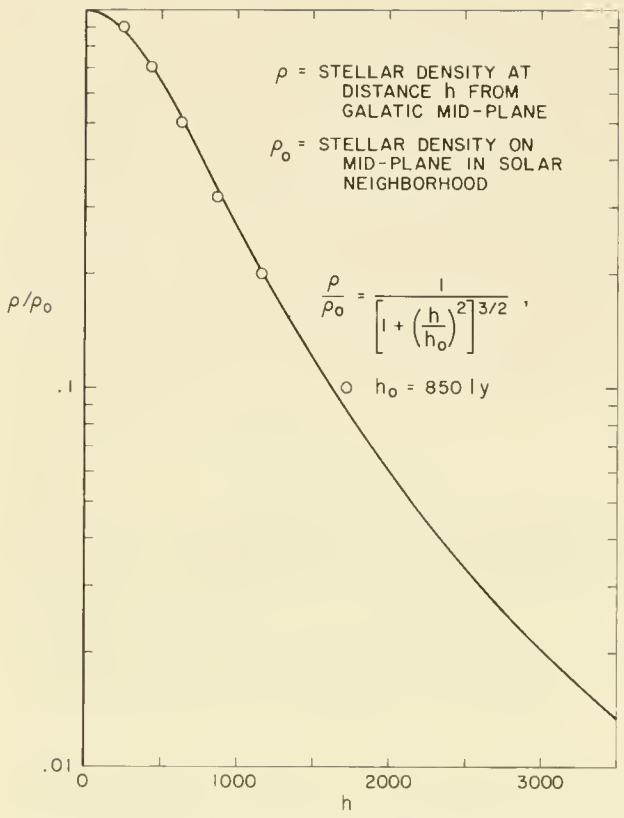


Figure 6-1. *Empirical approximation for number of stars per unit volume.*

this relation to points taken from the reference. Thus we find for the number n of stars within R light-years of the Sun

$$n = \rho_0 \int_{-\infty}^{\infty} \pi(R^2 - h^2)\,\frac{\rho}{\rho_0}\,dh$$

$$= 2\pi\rho_0 {h_0}^3 \left[\frac{R}{h_0}\sqrt{1 - \frac{R^2}{{h_0}^2}} - \log\left(\frac{R}{h_0} + \sqrt{1 + \frac{R^2}{{h_0}^2}}\right)\right] \tag{3}$$

Taking h_0 = 850 light-years and $\rho_0 = 5.4 \times 10^{-4}$ /(light-year)^3 (thereby stretching the approximation to include *F*0 through *K*7 stars) we obtain the curve shown in Figure 6-2. Using these values of $n(R)$ in equation (1) we plotted the curves of Figure 6-3. It is evident from Figure 6-2 that the simple cube law

$$n = \rho_0 \frac{4\pi}{3} R^3 \tag{4}$$

is a good approximation out to $R \approx 500$ light-years. So long as the cube law holds the range to which we must search varies only as the cube root of p. A thousand-to-one uncertainty in p produces as ten-to-one uncertainty in the maximum range.

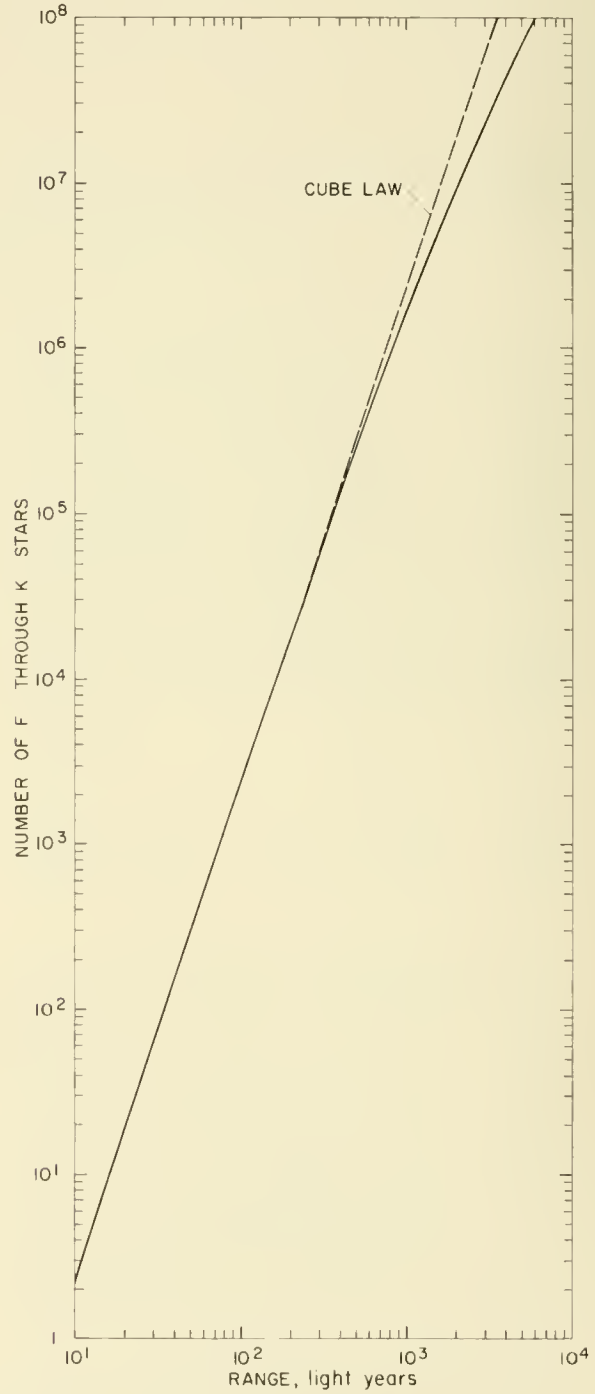


Figure 6-2. *Number of F0 through K7 stars versus range.*

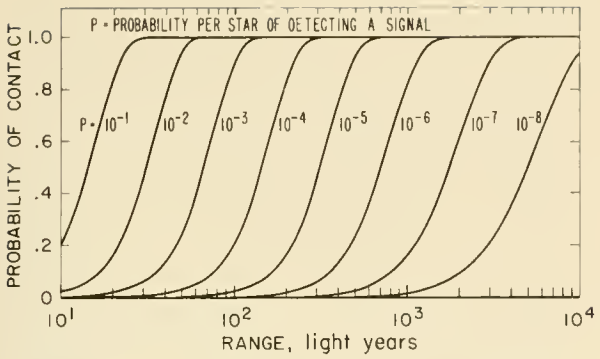


Figure 6-3. *Probability of contact versus range.*

However, we must concede at least this much uncertainty in p. If civilizations typically radiate megawatts of power for their own purposes for 10^7 years we might assign p the value 10^{-3} and be able to eavesdrop on their signals out to some 60 to 100 light-years. If, on the other hand, they typically radiate beacons for 10^4 years, then p might be as low as 10^{-6} and the beacons would have to be detectable at 600 to 1000 light-years.

Beyond 1000 light-years the situation becomes rather bleak. Not only does the cube law fail, but also the radiative epoch becomes shorter than the round-trip delay making two-way exchange unlikely. We cannot, however, exclude the possibility that very advanced races exist beyond this range, and have constructed very powerful beacons and have used them for unknown purposes for very long times, perhaps for aeons.

We must also point out that the curves of Figure 6-3 do not give a true picture of what happens as the sensitivity of a receiver is increased, for as we increase our receiver sensitivity, we not only extend the range for a given radiated power, we also permit the detection of weaker radiation from sources already within range, so to speak. To the extent that beacons are less likely than radio leakage, this capability increases the value we should assign to p for the nearer stars.

Since the range we must cover is so uncertain, only some general conclusions emerge:

1. We should start the search with a modest system capable of detecting beacons out to perhaps 100 light-years.
2. We should expand the system as the search proceeds (and is repeated) and continue the expansion until success is achieved or until we are able to eavesdrop on unintended radiation from 100 light-years range. The system should then be able to detect beacons of reasonable power at 1000 light-years range.
3. If technologically feasible at any time we may want to search for more distant powerful sources, perhaps even to scan other galaxies.

THE NUMBER OF RESOLVABLE DIRECTIONS

The number of distinct directions in which an antenna must be pointed to cover the sky is proportional to its gain. An isotropic antenna has a gain of 1, and would need to be "pointed" in only one direction; an antenna radiating uniformly into a hemisphere would have a gain of two and would need to be pointed in two directions. Similarly, any antenna radiating uniformly into a solid angle Ω would have a gain $4\pi/\Omega$ and would need to be pointed in this many directions. Actual antennas do not have a uniform gain inside a given solid angle and zero gain outside, and the number of directions in which we must point them depends on the loss we are willing to tolerate at the edge of each area covered by the beam.

If the aperture is circular the beam intensity is given by equation (2), Chapter 5, and if $(\pi d/\lambda)\,\theta = 1$ the loss at the edges is about 1.1 dB. If we accept this as tolerable then $\theta_{max} = \lambda/\pi d$ and the solid angle covered per beam is

$$\Omega = \pi\theta^2 = \pi\left(\frac{\lambda}{\pi d}\right)^2 = \frac{\pi}{g} \qquad (5)$$

The number of resolvable directions is therefore

$$N = \frac{4\pi}{\Omega} = 4g \qquad (6)$$

Our requirement of 1.1 dB maximum loss results in four times as many pointing directions as would be needed with a uniform conical beam.

Figure 6-4 shows N as a function of operating wavelength for circular antennas of different diameters. We see that N is a large number even for antennas of the size we might use for array *elements*. For example, a 100-m dish operating at 20 cm would have 10^7 fields of view in the sky. Figure 6-2 indicates that out to 1000 light-years there would be about 1.7×10^6 stars of interest in the sky. Even if we image the entire field of view of the array element, we will have only 0.17 stars of interest per field, on the average. With 10-m dishes we could average 17 stars per field, but we would need 100 times as many dishes to realize a given total array diameter. The simultaneous searching of several stars does not appear too feasible.

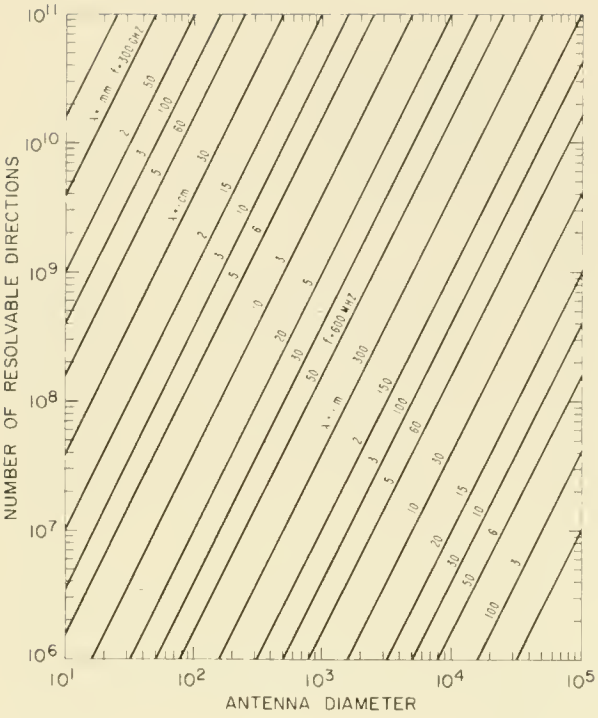


Figure 6-4. *Number of resolvable directions versus diameter and wavelength.*

SEARCH RANGE LIMIT

In radio astronomy and in most fields of measurement the minimum detectable signal amplitude is defined as being equal to the rms fluctuations due to noise. The signal-to-noise power ratio out of a square-law detector is given by equation (19) in Chapter 5 as

$$\frac{S}{N} = n\frac{(P_r/P_0)^2}{1 + 2(P_r/P_0)} \tag{7}$$

where

P_r = received signal power

P_0 = noise power

n = $(B\tau)$ = number of independent samples that are averaged

Setting S/N equal to unity we find for the *received* signal-to-noise ratio

$$\frac{P_r}{P_0} = \frac{1 + \sqrt{1 + n}}{n} \tag{8}$$

Substituting this relation into equation 5, Chapter 5, with $P_0 = \psi B$ we find the conventional range limit

$$R = \frac{d_r}{4}\left(\frac{P_{eff}}{\psi B}\,\frac{n}{1 + \sqrt{1 + n}}\right)^{1/2} \tag{9}$$

$$\approx \frac{d_r}{4}\left(\frac{P_{eff}}{\psi B}\right)^{1/2} n^{1/4},\ n >> 1 \tag{10}$$

where

d_r = receiving antenna diameter

P_{eff} = $P_t g_t$ = effective radiated power

ψ = noise spectral density = kT at radio frequencies

B = receiver bandwidth

τ = integration time

n = $(B\tau)$

We use equation (10) in estimating range limits when the data processing system is unspecified. However, in determining the performance of the Cyclops system we use the actual range limit at which the probability of missing the signal p_{ms} and the probability of a false alarm p_{fa} have specified values. That is, we take

$$R = \frac{d_r}{4}\left[\frac{P_{eff}}{kTB} f(n, p_{ms}, p_{fa})\right]^{1/2} \tag{11}$$

where $f(n,p_{ms},p_{fa})$ is evaluated for the proposed data processing method using the actual statistics, and will be a number on the order of unity. Anticipating this, we will use as a *reference* range limit for some of our curves the range R_0 given by

$$R_0 = \frac{d_r}{4}\left(\frac{P_{eff}}{kTB}\right)^{1/2} \tag{12}$$

which is simply the range at which the *received* signal-to-noise ratio is unity and is less than R, as given by equation (9), for $n > 3$.

DOPPLER SHIFTS AND RATES

Relative motion between transmitter and receiver can occur because of (1) radial velocity of the star with respect to the Sun, (2) orbital velocity of the Earth and

the other planet, and (3) the rotation of the Earth and the other planet.

Radial velocities of stars relative to the Sun are essentially constant over long periods of time and therefore produce fixed frequency offsets. The principal result is to broaden the frequency range over which we must search for a signal known (or assumed) to have been radiated at some particular frequency such as the hydrogen line.

A few stars have velocities in the range of 65 to 365 km/sec relative to the Sun. These stars are virtually all Population II stars, old stars that are deficient in heavy elements. Almost all the Population I stars that might have inhabited planets have velocities relative to the Sun that are less than 60 km/sec. Thus, the range of Doppler offsets we may expect is given by

$$\left.\frac{\Delta\nu}{\nu}\right|_{\max} \approx \pm 2\times 10^{-4} \tag{13}$$

Planetary orbital motion and rotation produce Doppler shifts that vary nearly sinusoidally with time. Assume the source is moving in a circle of radius a with an angular velocity Ω, and that the line of sight makes an angle θ with respect to the plane in which the motion occurs. Then the radial velocity will be $v = a\Omega \cos\theta \sin\Omega t$ and

$$\Delta\nu = \nu\frac{a\Omega^2}{c}\cos\theta \sin\Omega t \tag{14}$$

$$\dot{\nu} = \nu\frac{a\Omega}{c}\cos\theta\cos\Omega t \tag{15}$$

The frequency drift rate is greatest when the shift is zero and vice versa. Taking $\theta = 0$ and the appropriate times we find

$$\left.\frac{\Delta\nu}{\nu}\right|_{\max} = \frac{a\Omega}{c} \tag{16}$$

$$\left.\frac{\dot{\nu}}{\nu}\right|_{\max} = \frac{a\Omega^2}{c} \tag{17}$$

For the Earth, the values arising from orbital motion and diurnal rotation are:

	$\lvert\Delta\nu/\nu\rvert_{\max}$	$\lvert\dot{\nu}/\nu\rvert_{\max}$
Orbital motion	10^{-4}	2×10^{-11}/sec
Diurnal rotation	1.5×10^{-6}	1.1×10^{-10}/sec

The orbital velocity is higher and produces a larger total shift, but the acceleration along the line of sight due to daily rotation is greater and produces a greater drift rate.

Planetary rotation rates in the solar system show a large spread. There is considerable evidence that the rotation rate of the Earth has been slowed by the Moon. Without this slowing we might have an 8-hour day, like Jupiter. We might very well find an inhabited planet in another system with this short a day and thus a value of $|\dot{\nu}/\nu| \approx 10^{-9}$. Using this value and adding twice the Earth's orbital shift to the stellar radial velocity, we find for the total frequency ranges and rates the values given in Table 6-1.

TABLE 6-1

Wavelength	Frequency	$2\lvert\Delta\nu\rvert_{\max}$	$\dot{\nu}_{\max}$
21 cm	1420 MHz	1.14 MHz	1.4 Hz/sec
10 cm	3 GHz	2.4 MHz	3 Hz/sec
3 cm	10 GHz	8 MHz	10 Hz/sec
10.6μ	2.8×10^{13} Hz	23 GHz	28 kHz/sec
1.06μ	2.8×10^{14} Hz	230 GHz	280 kHz/sec

The lower the frequency the narrower the band that must be searched to find a signal of known original frequency, and the narrower can be the receiver bandwidth. In effect, Doppler rate degrades the spectral purity of the source and forces us to use larger receiver bandwidths. These are additional reasons for preferring the microwave over the optical region and for preferring the low end of the microwave window over the high end.

THE EFFECT OF FREQUENCY DRIFT ON RANGE

If the signal were truly monochromatic and if we momentarily ignore our own inability to generate a completely drift-free signal then, in principle, we could make the receiver bandwidth equal to the reciprocal of our observing time and thus extend the range limit indefinitely. But because of Doppler rates, source instability, and our own local oscillator instability, we can expect the signal to have a drift rate $\dot{\nu}$ Hz/sec. This means that it will remain in a channel B Hertz wide for a time $\tau = B/\dot{\nu}$ sec. The response time of the channel is on the order of $1/B$ sec. Thus if we want the narrowest

bandwidth that will still give near maximum response we must set $\tau = 1/B$ and find that $B_{min} = \dot{\nu}^{1/2}$. A more precise analysis would show that

$$B_{min} = \beta \dot{\nu}^{1/2} \tag{18}$$

where β is a constant near unity that depends on the shape of the RF filter. For a gaussian filter we show in Appendix E that $\beta = 1.24 \ldots$.

Substituting equation (18) into (12), we obtain for the reference range limit

$$R_0 = \frac{d_r}{4}\left(\frac{P_{eff}}{\psi \beta \dot{\nu}^{1/2}}\right)^{1/2} \tag{19}$$

Figure 6-5 shows R_0 as a function of d_r and $\dot{\nu}$ with $P_{eff} = 10^9$ W, $\beta = 1$, $\psi = kT$ and $T = 20°$ K.

As can be seen from the curves or from equation (16), a hundredfold reduction in $\dot{\nu}$ permits the same range with one tenth the antenna area. This is why monochromatic signals with high-frequency stability are desirable for beacons.

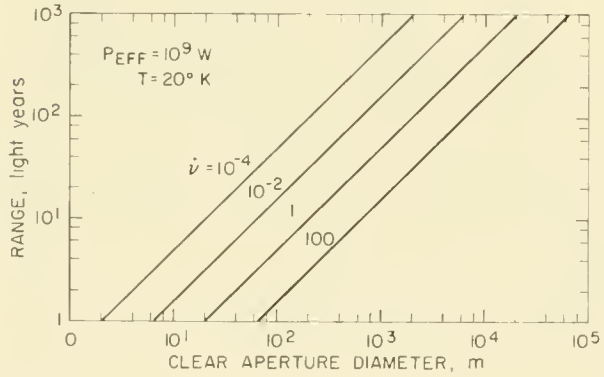


Figure 6-5. *Doppler limited ranges.*

The hydrogen maser has a stability of about 10^{-13}, so the curve marked $\dot{\nu} = 10^{-4}$ represents about the limit of our present technology for a 1 GHz signal. But to reap this benefit, Doppler rates would have to be compensated to an accuracy as high as a part in 10^4, depending on the planet's rotation rate. On Earth, the compensation would require an accuracy of a part in 10^3, which is not unreasonable. Thus it appears that the detection of a 1000-MW omnidirectional beacon at a range of 1000 light-years would require an antenna aperture of 2 to 6 km.

THE MAGNITUDE OF THE SEARCH

We have seen that we must have a receiving system with large antenna gain and, as a result, high directivity—so high that it will have on the order of 10^{10} resolvable directions. We have seen that the microwave window covers about 10^{10} Hz. We have seen that more nearly monochromatic signals are easier to detect and that we probably need to "comb" the spectrum into channels 1 Hz wide or less. If the signal is beamed at us for only a small fraction of the time, say 1 sec/day, we must search each direction *and* each 1 Hz channel for 10^5 sec or more. If we were to search the entire sky, blindly, with a receiver having a 1 Hz bandwidth the search would take us a time

$$\begin{aligned} T &= 10^{10}\text{ (directions)} \times 10^{10}\text{(channels/direction)} \\ &\quad \times 10^5\text{(sec/channel)} \\ &= 10^{25}\text{ sec} \\ &= 3 \times 10^{17}\text{ years} \end{aligned}$$

or 30 million times the age of the galaxy! What can we do to reduce this appalling number?

First, *we must not search blindly in all directions.* We must identify by optical or other means the million or so stars of the right spectral range that lie within 1000 light-years of the Sun. We may then search these in order of increasing range as we build our array, and not waste time on the voids between.

Second, *we ought to search all likely channels in the spectrum simultaneously.* At present we cannot do this for the entire microwave window, so we are forced to decide that a much narrower portion of the window is the place to look, and to design a system that analyzes *this* portion into some 10^8 or 10^9 simultaneous outputs.

Third, *we must decide that the signal we will detect is either on all the time or else comes from a nearby star.* There are only a thousand or so stars near enough to permit eavesdropping on leakage radiation from their planets. We can afford to spend days on each of these, and we should. But for stars beyond this range we must depend on beacons that are always shining.

Under these assumptions the search time is greatly reduced. If it takes us 1000 sec per star to search for spectral anomalies in the likely band or bands, then to search all the 1.7×10^6 *F* through *K* stars out to 1000 light-years would require

$$\begin{aligned} T &= 1.7 \times 10^9 \text{ sec} \\ &\approx 50 \text{ years} \end{aligned}$$

If we could eliminate the late *K* type stars this figure would drop by another factor of two.

But does by our technology and our assessment of the *joint* search strategy permit us to make these assumptions? We think so. We believe that we can pinpoint the likely stars both as to spectral type and range with a reasonable optical search program concurrent with the radio search. We believe we can construct receivers that will display power spectra over 100 MHz of bandwidth with a resolution better than 1 Hz. Finally, we believe that leakage radiation is most likely to be detected at the low end of the microwave window, and that beacon signals are most likely to be found in a relatively narrow range at the low end of the window, for reasons that are given below.

LEAKAGE SIGNALS

Electromagnetically, Earth is at present a noisy planet. We radiate hundreds of megawatts and much of this power is at frequencies for which the ionosphere is quite transparent. This, of course, raises the question of whether or not we might eavesdrop on the signals another race transmits for its own purposes. There is then no need to assume the existence of intentional beacons and at first glance the probability of detection appears to be greatly increased.

Our interstellar radiation became significant about 20 years ago with the advent of VHF TV broadcasting and more recently has increased with the expansion of TV allocations into the UHF band. Today we are surrounded in space with a sphere of radiation some 20 light-years in radius, and the energy density in this sphere is growing annually. How long this buildup will continue is anyone's guess. Cable TV is replacing direct reception in metropolitan areas where bad reflections and shadowing exist, and in many rural areas shadowed by mountains. However, the economics do not favor cable TV in the normal service areas of broadcast stations, where good reception exists. On this basis, one might conclude that powerful TV radiation would be an enduring phenomenon.

Satellite broadcasting appears to be a greater long-term threat to our TV leakage than cable TV. A UHF transmitter in synchronous orbit using present transmission standards need only radiate a few kilowatts to cover the entire United States. A fair fraction of this power would be reflected back into space by the Earth but, because far fewer stations of lower power would be needed, the resulting leakage would be negligible compared with our present leakage.

Nevertheless, it is of interest to calculate how far into space the present radiation level of Earth might be detectable. Our TV stations radiate about 50 kW. Assuming a grey field (signal amplitude halfway between black level and white level) the effective carrier power is about 20 kW. The antennas typically have 13 dB of gain as a result of vertical directivity, so the radiation is confined between a plane tangent to the Earth and a cone whose elements have an elevation angle of about 6°. As the earth rotates at 15°/hour this sheet of radiation sweeps across the celestial sphere. For a station at latitude θ the beam would take a time $(6/15)\sec\theta$ hours to scan a given star on the celestial equator. Thus 20 min is a reasonable average time. Using a receiver with a 20° K total noise temperature, an antenna 5 km in diameter, a bandwidth of 0.1 Hz[1] and integrating for 1200 sec, we find from equation (11) that the range limit is 50 light-years. Actually, this is a somewhat pessimistic figure since many stations, and often several on each channel, could be received at the same time and proper data processing could make use of the *total* power. We conclude that if we keep on broadcasting TV for another century, Earth will be visible out to something on the order of 100 light-years, which could announce our existence to beings on any of the 1000 or more likely stellar systems within that range.

To beings that detected us, there would not be doubt for very long that the signal was the work of man, not nature. They would observe that the signals (1) had highly monochromatic components, (2) were distributed systematically in slots across the spectrum, (3) appeared and disappeared with great regularity (in particular, a 24-hour cycle would stand out) and (4) exhibited a sinusoidal frequency modulation whose period was proper for the annual motion of a minor planet, and whose fractional frequency variation $\Delta f/f$ was the same for all signals. The annual Doppler and daily periodicity would identify the signals as being of planetary origin while the monochromaticity and regularity of spacing would identify them as artificial in origin.

We conclude that leakage signals are a possible means for the detection of other life. However, the longevity of their emission is very uncertain, and their low power compared with an intentional beacon restricts their detection range significantly.

BEACONS

One can imagine several reasons why an intelligent race might construct a beacon (or even many beacons) but perhaps the strongest reason is to facilitate the

[1] The Doppler shift would be less than 0.01 Hz during the observing period, but the frequency instability of the source might cause larger drifts.

acquisition phase for other races. Suppose that Earth does indeed become electromagnetically quiet as our technology becomes more advanced. Might we not then, realizing how undetectable we had become, construct a beacon to help other races find us? We would conclude that our own radiative history might be typical and that it would be pretty silly for everyone to listen with nobody transmitting.

There is a somewhat more speculative reason to expect beacons. Population I stars, which condensed from gas clouds enriched in heavy elements from earlier supernovae explosions, began to form in large numbers early in the history of the galaxy and were very numerous 9 billion years ago. If we take the 4 billion year gestation time for advanced life on Earth as typical, then as long ago as 5 billion years advanced cultures appeared in the galaxy in large numbers. If we assume that most of these attempted interstellar communication during their advanced lifetimes, then many of them may have been successful. One of the consequences of this success would be the accumulation of a large common body of knowledge that would include, in addition to all the biological knowledge of all races in contact, a rather complete picture of our galaxy, our neighboring galaxies, and of the universe as these appeared 5 billion years ago.

As new races came of age and made contact with the galactic community they would inherit this body of knowledge, add to it, and in turn pass it on to still younger races when they made contact. In fact, the transmission down through the aeons of this accumulated heritage of galactic knowledge could become one of the principal *raisons d'être* for interstellar communication. In this event beacons would very likely be used to ensure the survival of the "galactic heritage" by attracting the attention of young races.

Directivity of Beacons

In the search phase we very quickly conclude that we need a large receiving antenna with its high gain and directivity to (1) collect enough signal energy, (2) exclude local interference, and (3) tell us where the signal came from. In fact, the collecting area we will need is probably so large that, even at the lowest usable frequencies in the microwave window, we will be only able to search one star at a time, as we have seen.

Suppose now that a similar system were used as a beacon; that is, a single transmitter with a highly directive antenna is used to irradiate sequentially m stars over and over again. This beacon would be detectable by beings around any one star only $1/m$th of the time. In effect this replaces p in equation (1) by p/m to give

$$p_c = 1 - e^{-np/m} \tag{20}$$

Assuming an equal search effort $m \approx n$ and, since $p \ll 1$, we have

$$p_c \approx p \tag{21}$$

no matter how large m and n become (i.e., no matter how hard both races try). Clearly, a single highly directive beacon is only effective if it obviates the search at the receiving end; that is, if the range is small enough and the effective radiated power is great enough to allow the signal to be detected with an omnidirectional, or only slightly directive, receiver. We have not yet detected any such signals.

We, or the other races, might construct m beacons pointed at each of the m stars within range. This is, in fact, a practical solution for values of m less than 1000 or perhaps 2500, which is to say for ranges up to 70 to 100 light-years. In fact, the receiving array could double as a beacon, say on alternate years, as discussed in Chapter 12. However, for several reasons, this approach rapidly becomes impractical for ranges beyond 100 light-years.

The only reason for using directive transmitting antennas is to avoid wasting energy in the voids between target stars. When m is small we can save a great deal of energy this way. As m increases and we try to save energy in the same ratio, we must increase the area of each antenna in direct proportion to the number of beams, m. This means the *total antenna area will increase as* m^2, or *as the sixth power of range.* Even if we were to accept the expense of this rapid buildup we would ultimately run into another problem. Our beams would become so narrow that the proper motions of the target stars would carry them out of the beam in one round trip light time. Since we should direct the beam not where the star appears to be now, but where it will appear to be this far in the future, we would then have to know the ranges and proper motions of all the target stars in order to compute the lead angles.

If, to avoid this problem and the expense of the vast antenna area, we keep the area of each element constant as m increases, then very soon we will be flooding the sky with beams. We will, in fact, be radiating omnidirectionally, which we could do without any large antennas at all!

This situation is shown graphically in Figure 6-6 where the total transmitted power required (at the reference range limit for the assumed receiving system) is plotted as a function of range limit for an omnidirectional beacon, and for directive radiators of several sizes. For any size of directive radiator the total power, which

is less for short ranges, increases as the fifth power of range (the power per antenna increases as R^2 and m increases as R^3), and so intersects the power required for an omnidirectional beacon. The line labeled $\Sigma A_t = A_r$ is the locus for which the total transmitting antenna area equals the receiving antenna area of the assumed receiving system. The curves assume that m is equal to the number of F, G, and K stars within the indicated ranges; $m = R^3/440$. The scale of m is shown along the top.

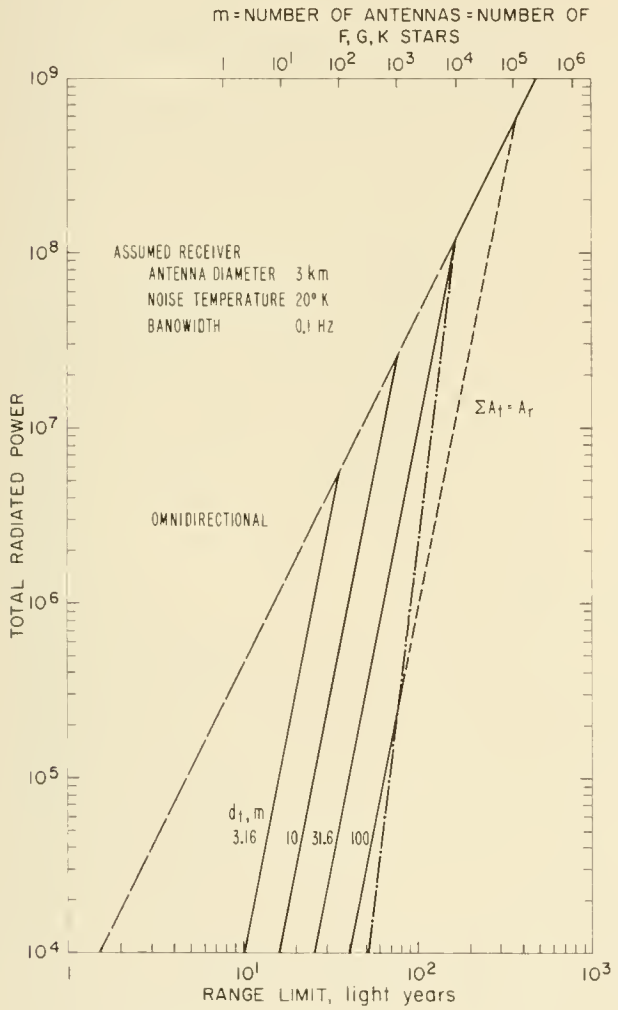


Figure 6-6. *Beacon power versus range.*

We see that if we had a receiving array of a thousand 100-m dishes, we might use it as a beacon to illuminate the stars out to about 76 light-years (or perhaps a narrower spectral class, say G stars, to a greater range) but beyond 100 to 200 light-years we would almost certainly construct a separate omnidirectional beacon. We conclude that long-range beacons will be omnidirectional and high powered.

Power Level of Beacons

How much power might we expect in an omnidirectional beacon? If we make the assumption that all races end up transmitting omnidirectional beacons as well as searching with directive receivers, then the proper beacon power is that which minimizes the total system cost for all races involved. The total system cost will increase as the search is carried to greater ranges, but at any assumed range limit, it can be minimized.

The efficacy of a beacon-search-receiver combination is proportional to $P_t A_r$ where P_t is the beacon power and A_r is the receiving antenna area. At both ends these factors represent the most costly items. Any race deciding to transmit as well as receive will decide that the other races will also be doing both for similar reasons, and that all would have the goal of minimizing the *total* system cost. The total cost includes many terms but a dominating one will be

$$C = K_P P_t + K_A A_r \tag{22}$$

where

K_P = cost per unit power capacity of transmitter

K_A = cost per unit area of antenna.

If the product $P_t A_r$ is held constant, C will be a minimum when

$$K_P P_t = K_A A_r \tag{23}$$

Nuclear power costs about \$300/kW of generating capacity. The conversion of this power to monochromatic microwave power costs about \$1200/kW. Let us therefore take K_P = \$2000/kW. If we are willing to spend $K_A A_r = \$4 \times 10^9$ for antennas we should also be willing to spend this same amount for beacons, which makes $P_t = 2 \times 10^6$ kW or 2000 MW. These figures are technology dependent, but it appears not unreasonable to expect beacon powers in excess of 1000 MW.

Other Properties of Beacons

In addition to having high powers, beacons would have other properties that make them easy to detect:

1. *They would transmit continuously.* Knowing the short search time available per star, no one constructing a beacon would build a low duty cycle one.

2. *They would be extremely monochromatic.* The detectability of a beacon is proportional to its spectral power density, not just to power alone. The inherent source instability would probably be better than a few parts in 10^{13}, which is our current state of the art and, in addition, steps would be taken to greatly reduce apparent Doppler shifts. Some possibilities are: (a) locating beacons at the poles of the planet, (b) radiating in a series of fan beams tangent to the planet near the equator, and (c) placing the entire beacon in stellar orbit at a large distance from the star.

   Alternative (a) leaves the signal with orbital Doppler. Alternative (b) can correct for orbital Doppler by appropriate frequency control on each transmitter but cannot completely correct for diurnal Doppler as each beam swings past the receiver. Alternative (c) can be essentially Doppler free but to be economically attractive it requires more advanced technology than we now have. With appropriate compensation, Doppler rates as low as 0.01 Hz/sec are not unreasonable. Since both source instability and Doppler rates produce drift rates that are a fixed fraction of the carrier frequency, reducing the carrier frequency to the low end of the microwave window is beneficial. In fact, this is a very strong reason for favoring the low end of the window in the search.

3. *They would probably be circularly polarized.* The polarization of a circularly polarized wave is unaffected by the interstellar medium. Such a wave arrives with only two equally likely alternatives: right-hand or left-hand polarization. Both of these need to be tested, and this doubles the data handling equipment or time. Linearly polarized waves, on the other hand, may arrive with a continuous range of polarization angle. To avoid having more than a 1-dB loss we would need to examine not only vertical polarization (V) and horizontal polarization (H) but also both 45° polarizations $V + H$ and $V - H$. Since this would require four additional tests, the admission of linear polarization into consideration increases the data processing equipment or time by another factor of three. To save the searching race time or expense, the transmitting race would choose circular polarization for a beacon. (In addition, such a choice facilitates later communication of what is meant by "left" and "right.")

4. *They would be information bearing.* If we received merely a monochromatic signal we could respond by beaming a return signal at or near the received frequency. (If our reply was *at* the received beacon frequency, it would be received at twice the one-way Doppler offset caused by interstellar motion.) We would then have to wait the round-trip light time before any information exchange took place. It would be much more satisfactory to have information transfer during this time. Even at a relatively slow rate such information transfer could be large, and could include among other things how best to reply to the beacon signal itself (e.g., what frequency and modulation method to use), or where and how to receive a signal having a faster information rate.

   The modulation of the beacon should not and need not jeopardize its detectability. Some possibilities include frequency shift or phase shift keying, or the transmission of side frequencies that would beat with the CW beacon to produce spectral lines in the error signal output of an oscillator phase-locked to the beacon. These recovered tones could appear and disappear and constitute a succession of code groups. Many other possibilities exist.

Lasers as Beacons

Omnidirectional beacons at optical frequencies must be very powerful to compete with star noise. At 10.6μ the Sun radiates 1.25×10^{10} W/Hz in each polarization. Even assuming no Doppler rate and a frequency stability of a part in 10^{13} we would need a receiver bandwidth of 3 Hz giving a beacon power of 4×10^{10} W to equal the solar background ($b_* = 1$). At 1.06μ the situation is even worse: 3.6×10^{11} W/Hz and a 30 Hz bandwidth giving a power of 10^{13} W. If these powers were possible and the bandwidth could be achieved without added noise, the star's brightness in the selected frequency band would appear to change by a factor of two as the beacon turned off and on. However, we know of no way of achieving such narrow bandwidths other than by optical heterodyning which introduces spontaneous emission noise and limits the receiver antenna size as well.

If we include the system quantum efficiency and the star noise, range equation (11) becomes:

$$R = \frac{d_r}{4}\left[\eta\frac{P_{eff}}{h\nu B}\left(\frac{n}{1+\sqrt{1+n}} - b_*\right)\right]^{1/2} \qquad (24)$$

where b_* = (1/2 star power in bandwidth B)/beacon power.

We see that it is not necessary, with long integration times, for b_* to be less than unity, but merely less than $\sqrt{n}$. However, we have computed the range limits for the high powers given above. These data are given in Table 6-2, which includes a microwave system for comparison. Even with the much higher powers used, the laser systems are entirely inadequate, primarily because of the severe limitation on antenna sizes imposed by the requirement for coherent first detection (in order to achieve the narrow bandwidths). We can add the outputs of separate receivers after detection, but, as we saw in Chapter 5, the range will then improve only as the square root of the array diameter even assuming the star noise were spatially uncorrelated. The use of a 100-m array would increase the range of system I to about 1.4 light-years (and would require 197,500 antennas), while the range of system II would increase to about 1 light-year. If we attempt direct photon detection at 1.06μ the receiver bandwidth becomes about 30 MHz at least, $b*$ becomes 10^5 while $\sqrt{n}$ becomes 5.5×10^4 and the range limit is imaginary.

TABLE 6-2

System	I	II	III
Wavelength	1.06μ	10.6μ	20 cm
Beacon power	10^{13} W	4×10^{10} W	10^9 W
Quantum efficiency	0.4	0.2	0.9
Receiving antenna diameter	0.225 m	2.25 m	3000 m
Receiver bandwidth	30 Hz	3 Hz	0.1 Hz
Integration time	1000 sec	1000 sec	1000 sec
Background ratio (b_*)	1	1	---
Noise temperature			20° K
Range (light-years)	0.07	0.16	1360

In going omnidirectional we have given up the major virtue of lasers—the ability to produce tight beams with small antennas—but still suffer the penalty of having to use small antennas at the receiver. One possible way to use lasers as beacons is to make a huge number—say half a million—of them and point one at each target star (or, as already noted at the places where these stars will appear to be one round trip light time from now). With 10^5 W per unit for a total power of 5×10^{10} W and with receivers like those of optical system A and Infrared System B (see Chap. V) we could achieve (because of the longer integration time of 1000 sec.) ranges on the order of 200 light-years. With greater antenna expense and 50 times as much power we are finally able to contact one three hundredth as many stars as with the microwave beacon.

Likely Beacon Frequencies

There are several reasons to prefer the low end of the microwave window for the acquisition phase. These include:

1. Smaller Doppler shifts as noted above
2. Less stringent frequency stability requirements
3. Greater collecting area for the narrowest usable beam
4. Reduced cost per unit area of collecting surfaces
5. Smaller power densities in transmitter tubes, waveguides, feeds and/or radiators, thus allowing higher powers per unit
6. Greater freedom from O_2 and H_2O absorption, which may well be more on some planets having our life forms

Reasons for preferring the high end of the window include:

1. Reduced spectrum clutter
2. Smaller transmitter antennas to get a given directivity (for Doppler corrected beams)

Neither of the last two reasons seems as compelling as the first six. The transmitting antennas are in any case much smaller than the receiving array, and spectrum utilization can be programmed to clear certain bands at a time as the search involves them. Thus it appears likely that the search should be concentrated in the region from perhaps 1000 MHz to about 3 GHz. This is still an enormously wide bandwidth to be combed into channels 1 Hz wide or less. So the question arises as to whether or not there is a more sharply defined region of the spectrum where there is a common reason to expect the search to be located.

Several years ago Cocconi and Morrison (ref. 3) suggested that the hydrogen line (1420 MHz) was a cardinal frequency on which to listen. Their argument was that this is a natural frequency known to all communicative races, and one on which a great many radio astronomers are busily listening. Interstellar Doppler shifts might amount to $\pm10^{-3}$ so we would still have to search on the order of a 3 MHz bandwidth, but this is still one thousandth the task of searching a 3 GHz bandwidth.

This suggestion provided the initial impetus for Project Ozma in which Frank Drake and his associates at the National Radio Astronomy Observatory at Green Bank, West Virginia, listened for about 400 hours in April and June of 1960 for evidence of artifact signals from two stars in the 10 to 11 light-year range: ϵ-Eridani

and τ-Ceti. No signals were detected except for a couple of very exciting false alarms.

The arguments for using the hydrogen line no longer seem quite so compelling. Many other spectral lines have been discovered. Nevertheless, the concept of a naturally identified frequency is a good one and should not be discarded at this time. Of all the spectral lines, the hydrogen line (1) is the strongest yet found, except for the hydroxyl line in certain regions; (2) is radiated by the most abundant element, which also happens to be element number one; and (3) lies at the low end of the microwave window where we have gravitated for other reasons.

There is a real problem, however, with choosing a *single* natural frequency for transmission and reception. If we do both simultaneously, we jam ourselves. We would have to send half the time, which halves the probability of our being detected, and listen the other half of the time, which doubles the search time. Obviously, we cannot all choose one natural frequency for transmission and another for reception. What then is the best strategy?

As we have seen, because of Doppler effects, the choice of an exact transmitted frequency does not eliminate the frequency dimension of the search; it merely reduces it greatly. To eliminate self-jamming what we are seeking is a natural interstellar contact *band* rather than a single frequency. This band should be many times broader than the Doppler broadening but small enough to permit simultaneous search over at least half the band. The band should contain no strong spectral lines where beacons would interfere with radio astronomy work and where reception would be interfered with by the noise of the spectral lines themselves. (These are good reasons not to choose the hydrogen line itself.) Nor should the band lie below the hydrogen line, since beacons could then interfere strongly with observations of red-shifted hydrogen lines from other galaxies. Furthermore, the unavoidable background noise rises sharply below 1400 MHz. There is one band that seems to fulfill all the requirements. It is the band from the hydrogen line (1420 MHz) to the lowest of the hydroxyl lines (1662 MHz). If we leave 10- or 20-MHz guard bands at either end, we find about 200 MHz of clear band, which is just about what we need for transmission and reception to go on simultaneously. This band is very nearly at the quietest part of our microwave window and would be at the quietest part if the H_2O and O_2 absorption bands were many times stronger, as they might be on a planet with an atmosphere several times as dense as that of Earth. There may be other such bands that deserve study, but surely the band lying between the resonances of the disassociation products of water is ideally situated and an uncannily poetic place for water-based life to seek its kind. Where shall we meet? At the water hole, of course!

SEMANTICS AND ANTICRYPTOGRAPHY

Once a signal (just one!) that is clearly of intelligent origin is detected the acquisition phase is over, at least temporarily. We would immediately suspend the search, lock on to the signal, and apply every means at our disposal to detect and record any modulation that might be present. What form this might take we cannot predict, but only a few likely alternatives exist and it would probably not take long to discover the modulation method.

We would then be faced with the problem of determining the *meaning* or significance of the modulation. Here we can be fairly sure of one thing: the sending race will attempt to make the job of deciphering and understanding the messages as simple and foolproof as possible. The coding will be simple, redundant, and full of clues as to how to reduce the message to clear form. Because this objective is exactly the reverse of what is wanted in cryptography, it has often been called the *principle of anticryptography*.

No optimum method of forcing the proper initial decoding and ensuring the rapid subsequent decipherment of the message content into understandable concepts has been developed. Given the enormous disparity among races as to what might be considered *obvious*, no single optimum may exist. Sukhotin (ref. 4) has presented an extensive analysis of the problem. Drake[2] and Oliver (ref. 5) have suggested that the initial messages of a sequence might be pictorial, since any intelligent race would very likely have vision. They have suggested sending binary information in repeated sequences having the same length in bits. If this number of bits is the square of a prime number p, the rearrangement of the message into a raster of p lines of length p is strongly suggested, whereupon the received pulses form a picture. If the picture has a "frame," that is, a complete border of pulses, there will be two long rows of pulses representing two opposite sides (such as the top and bottom) and repeated pairs of pulses at the same separation indicating the other two sides. This structure alone would suggest a raster without resorting to prime numbers.

Once the receiving race has cracked the code into pictures, the way is cleared for a series of "primer" messages that can convey a great deal of information in

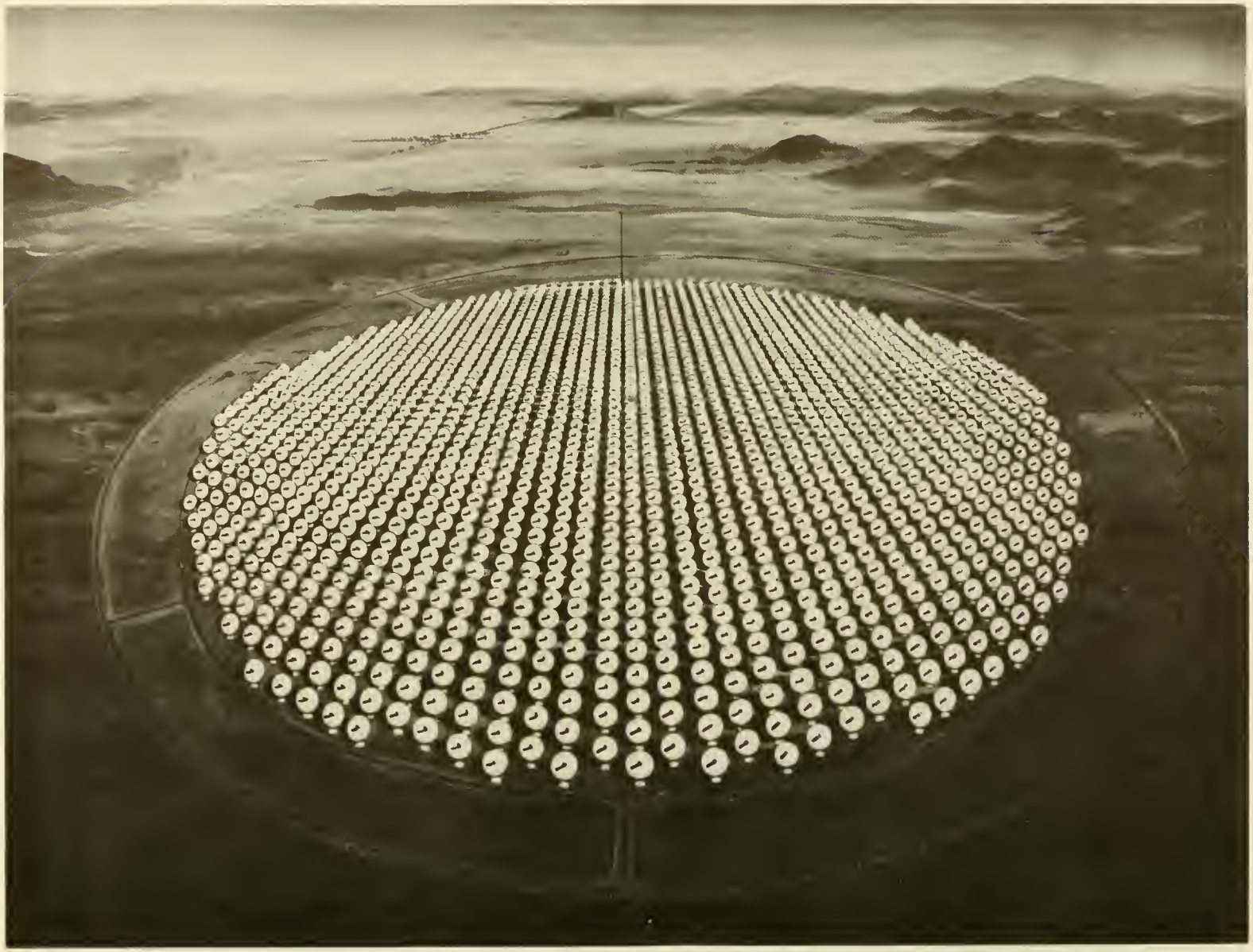
[2] Private communication to the Order of the Dolphin, Dec. 12, 1961.

themselves as well as define symbols to be used later. Instructions as to how to respond to the beacon messages would almost certainly be included. A year's worth of such messages would probably consist of interspersed primer, intermediate and advanced messages much like a year spent in a one room elementary school with all grades in one room.

The possibilities are endless and difficult to predict. The tendency to be anthropocentric in thought is very strong in this area. However, there does not seem to be any great difficulty associated with the semantics problem. Compared with the acquisition problem all else is easy. Acquisition of the first signal is the big hurdle and the central problem of interstellar communication.

REFERENCES

1. Allen, C.W.: Astrophysical Quantities, Oxford Univ. Press, N.Y., 1963.
2. Elvius, T.: Distribution of Common Stars in High Latitudes, in *Galactic Structure,* S. Blaauw and M. Schmidt, eds., University of Chicago Press, 1965, p. 484.
3. Cocconi, G.; and Morrison, P.: *Nature*, Sept. 19, 1959, *184*, 4690, (844-846).
4. Sukhotin, B.V.: Chapter IV, Extraterrestrial Civilizations, S.A. Kaplan ed., NASA TT F-631, 1971.
5. Oliver, B.M.: Radio Search for Distant Races. *Int. Science and Technology,* Oct. 1962 (p. 59).



View 2. *Artist's concept of high aerial view of the entire Cyclops system. Diameter of the antenna array is about 16 kilometers.*

7. THE CYCLOPS SYSTEM

Solutions to the major engineering problems posed by the Cyclops system are described at length in the next four chapters. In this chapter, we present some of the highlights in a condensed form so that the reader can obtain an overview of the system before studying it in detail.

THE ANTENNA ARRAY AND SYSTEM FACILITIES

From the calculations and discussions presented in Chapter 6 it appears likely that the detection of coherent signals of intelligent extraterrestrial origin will require a microwave antenna system with a total collecting area on the order of 7 to 20 km^2, which corresponds to a clear circular aperture diameter of 3 to 5 km. A single unit, Earth-based, steerable antenna of this size is out of the question. Even in space, where gravity forces could not crush it, nor winds overturn it, the cost of orbiting the thousands of tons of material needed, the problems of assembly and erection, and the logistics of maintenance, appear too formidable. The only practical method of realizing square miles of collecting area at microwave frequencies seems to be with phased arrays of smaller antennas.

Our estimates of the density of intelligent life in the universe and of the power levels this life might radiate as leakage, or use in beacons, are so uncertain that the required antenna area might be as little as one-tenth, or as much as ten times the figures given above. This uncertainty is in itself a strong argument in favor of phased arrays, which can be expanded if necessary as the search proceeds. A single-unit antenna involves a size decision at the outset and runs the risk of serious over- or underdesign.

A phased array is steered by turning the individual elements of the array so that they point in the desired direction, and by electrically shifting the phase and adjusting the delay of the signals received by each element, so that these signals add in amplitude at the combined output. While the beamwidth of the array may be 1 sec of arc or less, the beamwidth of the individual elements may be much wider, perhaps 1 min of arc. The mechanical pointing precision is then only on the order of 10 sec. The final steering, done electrically, provides the pointing precision required for the array beam.

The optimum size of the individual elements in an array depends on the rate of growth of element cost with diameter. If the cost increases as the square of diameter, or more slowly, the elements should be as large as possible. If the cost increases more rapidly than the square of the diameter, say as the xth power of diameter, then the optimum size is that which makes the structure cost $2/(x-2)$ times the fixed costs per element The Cyclops study indicated a value of x very near 2, so we have assumed the largest size of fully steerable element yet constructed (100-m diameter dishes) for the Cyclops antenna element. Further study might reveal that with mass production techniques, material costs would be a greater fraction of the total cost. This would raise x and might result in a smaller optimum size.

Several novel antenna mounts were studied but none of these designs was carried far enough to justify recommending it over a conventional mount with azimuth and elevation axes. Thus, at present, we visualize the antenna element as a steerable az-el mounted paraboloid of about 100-m diameter

The array elements must be separated by about three times their diameter to avoid shadowing each other at low elevation angles. In the final full sized array, the elements should be packed as closely as allowed by this minimum separation to (1) keep the array beam as broad as possible to avoid pointing error problems, and (2) reduce the cost and technical difficulties of phasing

and of transmitting signals to and from the element. Thus the final array would be roughly circular in outline with the elements disposed in a regular hexagonal lattice.

The initial array, which might consist of only 100 elements or less ($\leqslant$ 1 km clear aperture), could also be close spaced. Additional elements, as they were built, would then be added at the periphery. However, the initial array might be a more valuable tool for the radio astronomer if it had the full resolution of the final array. This could be achieved by spreading the initial elements over the full area, and then adding new elements, as they were built, to the interior of the array. At the intermediate sizes the element density could be tapered from the center to the rim so as to provide very low near in side lobes.

The control and data processing center is most economically located at the center of the array. Omitting the central element would allow room for a building 200 to 300 m in both directions, which should be adequate. From the air, the final Cyclops system would be seen as a large central headquarters building surrounded by an "orchard" of antennas 10 km to 10 miles in diameter and containing 1000 to perhaps 2500 antennas.

Below ground, these antennas would be connected by a system of service tunnels radiating from the central building. Through these tunnels would flow the power and the control signals to drive and position the antennas, and the standard frequencies from which the local oscillator signals are synthesized. Back through the same tunnels coaxial cables would carry the precious IF signals to be combined and processed in the central building. The tunnels would also carry telephone cables for communication with crews at the antenna sites. Conditioned air from the central building might be exhausted through these tunnels to the antennas and used there to stabilize the temperatures of the receiver house and structural members.

Cyclopolis, the community where the system staff and their families live, might be located several miles from the array, perhaps behind hills that provide shielding from radiated interference. Transportation to and fro could be by bus at appropriate hours. Alternatively, the central headquarters might be made large enough to provide the necessary housing, stores, schools, and so on. There would be ample room between the antenna elements for playgrounds and recreation facilities. The major problem would be interference from appliances, and adequate shielding might be too expensive. This problem has not been studied.

SKY COVERAGE

Ideally we would like Cyclops to be able to search the entire sky. This is not possible even if the array is on the equator since we are limited by noise and shadowing to a minimum elevation angle of about 20°. Figure 7-1 shows the percentage of the sky covered as a function of latitude for various values of the minimum elevation angle. The advantage of low latitudes is apparent, but we see that with a minimum elevation of 20° and a latitude of 33° the sky coverage is still 80% as against 94% at the equator.

Figure 7-1. *Sky coverage*

If only one Cyclops system is ever built, a nearly equatorial or slightly southern latitude is technically (though perhaps not politically) preferable. If two or more are built, complete sky coverage is possible with one in a northern and the other in a southern temperate latitude. The existence of two or more Cyclops systems widely separated in longitude opens the possibility of using very long base-line interferometry on normal stars to detect periodicities in their proper motions, from which the architecture of their planetary systems might be deduced, as has been done optically for Barnard's star.

The price of incomplete sky coverage is hard to assess; the one intelligent life form accessible to us might lie in the regions of the sky not covered. About all we can say, *a priori*, is that the probability of contact is roughly proportional to the fraction of the sky covered, with a slight premium attached to the portions of the sky toward the galactic center, because of the increasing stellar density in that direction. Under this premise, an element mount that reduces the sky coverage by a factor of two should result in a cost saving of at least a factor of two to be worth considering.

SITE SELECTION

If Cyclops were to be located in the territorial United States, sites along the southern border deserve primary consideration. In addition to having a low latitude, the site should have the following characteristics:

1. *Geologic stability.* The area should not be subject to earthquakes nor lie across faults. The land should not be subsiding nor warping. Otherwise, the geometry of the array will be disturbed and require resurvey and reprogramming of coordinate data for the phasing and delay.
2. *Low relative humidity.* Atmospheric turbulence in the microwave region is determined primarily by inhomogeneities in the water vapor content of the air. For good "seeing" the air masses should exist in horizontal layers each of uniform composition. High, dry plateaus are preferable to moist low lying regions prone to cumulous clouds and thunderstorms.
3. *Mild, calm climate.* High winds deform antenna surfaces and cause loss of gain. Heavy snow loads require stronger more expensive structures to avoid permanent deformation.
4. *A large plane area.* The array need not be level, in fact, in northern latitudes a slight tilt toward the south is an asset. However, to avoid distortions in the imaging process, and complications in the delay and phase shift computing programs, the array should be plane or nearly so.
5. *Remoteness from habitation and air routes, preferably ringed by mountains.* The Cyclops system is of necessity an extremely sensitive detector of weak radiation. Although the antenna gain would, in general, be small for radiation generated within several thousand kilometers of the array, the low noise receivers can pick up weak interference even without the antenna gain. The problem is worse for Cyclops than for the usual radio telescope because of the high spectral resolution and because interfering signals may have strong coherent components of just the type we are searching for.

All in all, a remote site in the southwestern United States seems indicated. The Very Large Array (VLA) siting study is pertinent for Cyclops, and any of the sites recommended in this study should be considered.

RECEIVER SYSTEM

Since we concluded that the low end of the microwave window is best suited for interstellar acquisition and communication, the Cyclops receiver system was designed to cover the frequency range from 0.5 to 3 GHz. The antenna elements and system design would very likely permit higher frequency receivers to be used for radio astronomy, deep space probe communication, and radar astronomy, but these higher bands are not believed essential to the primary mission.

To avoid having to transmit the entire RF spectrum back to the central station, heterodyne receivers are used at each antenna. These convert the received RF band to a fixed (IF) band for transmission. The use of heterodyne receivers requires that local oscillator signals of precisely known phase be available at each antenna. In the proposed Cyclops system these are synthesized, at each antenna, from two standard frequencies, which, in turn, are distributed from the central station. The technique used for distribution is an extension and refinement of the two-way transmission technique described in the VLA report. The Cyclops technique virtually eliminates the residual errors present in the VLA system. We believe that local oscillator signals with phase errors of only a few degrees at 10 GHz can be generated throughout the Cyclops array using the proposed method.

Two primary design requirements of the Cyclops receivers are low noise and remote tunability. Halving the receiver noise is equivalent to doubling the antenna area. The remote tunability is a practical requirement; with a thousand or more antennas, local retuning of each receiver would be virtually impossible. Band changing must occur at all antennas in response to a single command over the control system. The proposed design achieves both these requirements by the use of cooled up-converters followed by a fixed frequency maser. It is believed that with the appropriate cryogenic cooling this combination can yield noise temperatures as low as 20° K. To change receiver bands the local synthesizers at the antennas are simply commanded to synthesize a new pump frequency, and, if necessary, the up-converter and feed horn are switched. This approach to radio telescope front end design is believed to be novel and should find application in other telescopes and arrays where flexible

operation is important.

The proposed design also allows simultaneous reception of two orthogonal polarizations, and the antenna electronics includes automatic monitoring of several parameters indicative of the state of health of the antenna and receiver, and provides automatic alarms in the event of malfunctioning.

The proposed instantaneous bandwidth of the receivers is 100 mHz for each polarization. Further study might permit this specification to be broadened to 150 or even 200 mHz. This would allow searching the entire spectrum of the "water hole" (Chap. 6) at one time.

IF TRANSMISSION

The two 100-MHz IF bands representing the two polarizations are multiplexed onto a single coaxial cable per antenna for transmission to the central station. Because of the wide IF bandwidth and the long distances involved, delay changes with temperature cannot be ignored but must be compensated. The Cyclops design incorporates two novel features that permit this delay compensation, as well as the equalization of cable loss with frequency and temperature, to be achieved almost free of charge.

The two IF bands are sent, one of them inverted in frequency, symmetrically disposed about a pilot frequency. This pilot tone is generated with a precise phase by each local antenna synthesizer. One-quarter of the way along each cable and again at the three-quarter point the entire spectrum is inverted. High frequencies, which have been suffering the greatest attenuation, thus become low frequencies and are attenuated least. At the midway point along the cable the IF bands are transposed about the pilot without inversion. In this way each IF band edge occupies for an equal length of cable the same four frequencies as every other band edge. The result is very nearly flat transmission versus frequency for both IF bands. Slope equalization of the cable loss is unnecessary, and changes in total cable loss are corrected merely by holding the received pilot level constant.

At the central station the total IF spectrum of each cable is passed through a small variable delay unit after which the phase of the pilot signal is compared with that of a locally generated signal. Any phase error is taken to signify a cable delay error, and the delay unit is actuated to remove the error. In this way, the standard frequency distribution system is used not only to generate the local oscillator signals in their proper phases but also to compensate for delay variations in the IF distribution system.

These techniques should also find application in other arrays besides Cyclops. The trick is simply to make full use of the fact that in a properly phased array a time reference is available everywhere.

THE IF DELAY SYSTEM

To provide the IF delay needed to steer the array beam to any azimuth and elevation angle, the Cyclops design proposes the use of digitally controlled IF delay units. The fractional nanosecond delays are achieved with microwave stripline, intermediate delays with short coaxial cables, and the longer delays with acoustic surface wave delay lines. Acoustic surface waves are non-dispersive and provide the tens of microseconds of delay needed in only a few centimeters of length.

It is believed that by heterodyning the total IF signal received over each cable up to a frequency range centered at 1 GHz we can delay the IF signals for both polarizations in the same units.

The delay precision required is very high: about ±1/8 nsec in as much as 50 μsec of total delay. Since acoustic delay lines have a temperature coefficient of delay, this might be thought to be an impossibly tight tolerance. But in the Cyclops system, the phase of the pilot signal before and after passage through each delay unit is compared and any error is used to heat the delay line or allow it to cool. In this way the temperature coefficient of the line is actually *used* to obtain the desired stability.

CONTROL AND MONITORING

In its primary mission all antennas of the array are pointed in the same direction, so common azimuth and elevation control data are sent to all elements of the array. All receivers are tuned to the same frequency. However, individual local oscillator phase shift and rate information must be supplied to each antenna and each IF line.

For multiple uses at one time the array must be broken into several subarrays that can be independently steered and tuned. The proposed Cyclops control system provides for either the primary or multiple use mode of operation through the use of time division multiplexed control signals distributed over a common cable system.

In addition, the local monitoring systems at each antenna cause the central computer to report the trouble, if minor, or to drop the element out of service, if the trouble is serious. Periodically, the computer also checks the gain, phasing, and positioning of each element by cross correlating the signals obtained against those from a reference element when both are pointed at a known radio source. Noise temperature and receiver sensitivity tests are also included.

While the analysis has not been carried to great depth, it appears that the central computer need be only of

modest size. A large data base on tape or disk giving the coordinates of hundreds of thousands of target stars is needed to automate the search, but the data rates are slow enough so that only a relatively small central processor is sufficient. A great deal more fast storage capacity is needed for data processing than for controlling and monitoring the entire array.

The system power and communication (telephone) needs have also been assessed. The power requirements are substantial and, in view of the remote location, may require a dedicated power-generating plant. The costs of this have been included. The telephone costs have also been included but are a negligible part of the total system cost.

COMBING THE SPECTRUM FOR SIGNS OF LIFE

As described so far the Cyclops system is simply a huge phased microwave antenna array that could be used for many purposes. And indeed, if built, it would be! It is the signal processing system of Cyclops that both distinguishes it and qualifies it for its unique primary mission. The signals we are searching for are earmarked by their coherence. They will cause very narrow, needlelike peaks in the power spectrum that may drift slowly with time but can be seen if we can resolve the power spectrum sufficiently and follow the drift. As delivered, in the 100 MHz band, the needles are literally buried in the haystack of receiver and sky noise. Yet the proposed signal processing system will find a signal even if its coherent power is 90 dB below the total noise power in the IF band.

The first step in the signal processing is to transform the received signal (amplitude versus time) so as to obtain successive samples of its power spectrum (energy versus frequency). This converts the nearly sinusoidal waveform of any coherent signal to a "needle" in the frequency domain. The second step is to add the power spectra under a variety of offsets between adjacent samples to allow for any reasonable drift rate the signal may have had during the observation time. In one of these additions, the one that matches the drift rate, the signal "needles" (which in any given sample may *still* be inconspicuous compared with the noise peaks) will all add to form a spike that is clearly above the noise level. Thus, the final step is to determine whether any of the added spectra contain spikes above a certain threshold.

In the proposed system the IF signals are first subdivided by filters and heterodyne mixers into bands from 1 to 10 MHz wide, depending on the bandwidth capabilities of the subsequent equipment. The time signal in each of these subbands is then recorded as a continuous raster on photographic film. A constant bias is added to prevent negative values of amplitude. After processing, the film passes through the gate of an optical Fourier transformer where it is illuminated by coherent light. The amplitude distribution in the aperture plane (film gate) is transformed by a lens into the signal spectrum in the image plane. The intensity of the light in this image plane is the power spectrum of the signal sample in the gate at any instant of time. The power spectrum is automatically displayed in raster form also. Thus, the full two-dimensional Fourier transforming power of the lens is used for a one-dimensional signal.

Each line of the power spectrum represents a frequency band equal to the scanning frequency used in the recording process. The frequency resolution is the reciprocal of the time represented by the total segment of the signal in the gate. Thus, if a 1 MHz band has been recorded, using a 1-KHz sweep frequency, and if 1000 lines of the raster are in the gate at any time, this represents one second's worth of signal. The power spectrum will also consist of 1000 raster lines, each of which represents 1 kHz of the spectrum, displayed with a resolution of 1 Hz.

Optical spectrum analyzers with a time-bandwidth product of 10^6 are currently available. Two hundred such units would be needed to comb the 200-MHz total IF band (both polarizations) into 1-Hz channels. Analyzers with a time-bandwidth product of 10^7 are believed to be within the state of the art. No other known method of spectrum analysis even approaches the capability of the optical analyzer. (It is interesting to note that, in principle, the Fourier transformation can take place in about 10^{-11} sec. The time need only be long enough to allow about 5000 cycles of the coherent light used to pass through the film. Shorter times would broaden the spectrum of the coherent light too much. Thus, in principle, the optical spectrum analyzer can handle about 10^{18} data samples per second. No practical way of utilizing this speed is known.)

In the proposed system the power spectrum is imaged on a high resolution vidicon tube, where it is scanned and converted into a video signal, which is then recorded on magnetic disks. As many as a hundred or more complete power spectra, representing successive frames of film in the gate, are recorded for each observation. The power spectra are then played back simultaneously and added with various amounts of relative delay between successive spectra. This is accomplished by sending all the signals down video delay lines and adding the signals from taps on these lines that are disposed in slanting rows across the array of lines.

In each observation of a star we will thus record 200 MHz of IF signal for something on the order of 1000

sec, thereby obtaining 100 samples of the complete power spectrum with 0.1-Hz resolution. These are then added in perhaps 400 different ways to synchronize with the assumed drift rate, but since additions with almost the same slope are correlated, there are only about 100 independent signals obtained from the adder, each having 2×10^{9} independent Nyquist intervals per polarization. Thus, a total of 4×10^{11} tests is made per observation, and to keep the probability of a false alarm to 20% for the entire observation, the probability of a false alarm *per test* must be about 5×10^{-13}

With the threshold set so that the probability of the noise alone exceeding it in any one Nyquist interval is 5×10^{-13}, the probability of missing the signal is 50% if the received signal-to-noise ratio is unity (0 dB) in the 0.1 Hz band or −93 dB for the 200-MHz total band. For only a 1% chance of missing the signal an additional 2 dB of signal power is needed. We know of no other method of processing the signal that can even approach this performance. The technique is believed to be an optimum one, though of course other and better means of implementing the operations may be found.

IMAGING THE RADIO SKY

The proposed Cyclops array is orders of magnitude more powerful than any existing fully steerable radio telescope. As such it would be a magnificent tool for finding and studying distant weak radio sources and for mapping the structure of known sources. Because all the IF signals are returned independently to the central headquarters, they can be combined to form more than one simultaneous beam. In fact, with n elements we can form n *independent* beams, or more than n correlated beams. By arranging the beams in a closely packed array and portraying the sky brightness measured by each beam as the brightness of a corresponding point on a screen, we can form a real-time high-resolution map of a portion of sky within the beamwidth of the antenna elements. In this way the time required to map the sky or to search for new sources is only one nth as great as with a single beam.

Several ways of doing this imaging were studied, none of which is ideal in all respects. The proposed method involves reradiating the signals from the antenna elements as electromagnetic waves from antennas in a scaled down model of the receiving array. If the scale factor is σ, the angular magnification of the telescope is $1/\sigma$. The imaging, if broadband, must be done at the original received frequency; otherwise, the angular magnification is frequency dependent and changes appreciably across the received band.

The model, or signal, array focuses the radiation on a second array whose receiving antennas pick up samples of the signal in the image plane. The energy collected by each receiver is then detected and converted to a proportional brightness of an element of a display screen.

The useful angular field of the image is inversely proportional to frequency. Thus, to keep the number of points in the image constant, the separation between the two arrays must be proportional to the RF center frequency. In the proposed system, the signal and image arrays are 20-m in diameter and their separation varies from 40 to 120 m over the tuning range from 1 to 3 GHz. Imaging is not attempted from 0.5 to 1 GHz.

The size of the proposed imager is awkwardly large. To prevent coupling to the receiving array it must be housed in a completely shielded space, which, in turn, must be lined with microwave-absorbing material to prevent undesired reflections. The dimensions of the anechoic chamber needed are about 75 ft wide, 75 ft high, and 550 ft long, a building roughly the size of the turbine house of a large power plant.

The building volume needed is directly proportional to the number of antennas in the array, to the ratio of the high to the low frequency limit of operation, and to the cube of the wavelength at the low frequency limit. If we were content to do the imaging over the band from 1400 to 1700 MHz, which includes the *H* and *OH* lines, a chamber 60 by 60 by 125 ft would suffice. However, even the cost of the proposed imager is small compared the total system cost, so we have proposed the large version, leaving possible compromises to the judgment of future study teams.

The shielding requirements, though severe, involve only a few mils of copper. Most of the problems are associated with bringing leads into and out of the chamber. All shielding problems are greatly reduced if the imaging frequency is offset from the receiver band by the IF bandwidth. This introduces an appreciable, but tolerable, amount of lateral chromatic aberration.

The proposed imager can image both polarizations simultaneously. By combining the intensities, we can cut the integration times in half. If the selected polarizations can be varied, then *subtraction* of the intensities would reveal polarized sources. This might be a powerful search tool for unusual types of emissions in the universe. In any event, the imaging capability would reduce by about three orders of magnitude the time needed to make certain kinds of radio astronomy studies, and, for the first time, would place radio astronomy on a more nearly equal footing with optical astronomy so far as data collection rates are concerned. Who can say what major discoveries might ensue?

THE AUXILIARY OPTICAL SYSTEM

Although not studied in any detail by the Cyclops team, it is obvious that the Cyclops system should include several optical telescopes. These could be used independently or be slaved to look at the same area of sky as the antenna array or subarrays.

One of these telescopes, probably a 1-m diameter aperture Schmidt, equipped with the proper instrumentation, would be used to survey the sky for likely target stars. Target star coordinates would be automatically recorded in the master computer data file for subsequent use. This same instrument, or another, used at high magnification could provide visual confirmation of the tracking accuracy of the computer program. Indeed, tracking corrections might be automatically introduced into the system.

The advantages of being able to obtain simultaneous optical and radio observations of source should appeal to the astronomer. Pulsar radio and optical emissions are known to correlate. Do the pulsar "starquakes" cause optical phenomena? Do the optical and radio emissions of flare stars exhibit correlation? Improved instrumentation always facilitates research and sometimes opens up whole new and unsuspected areas of research.

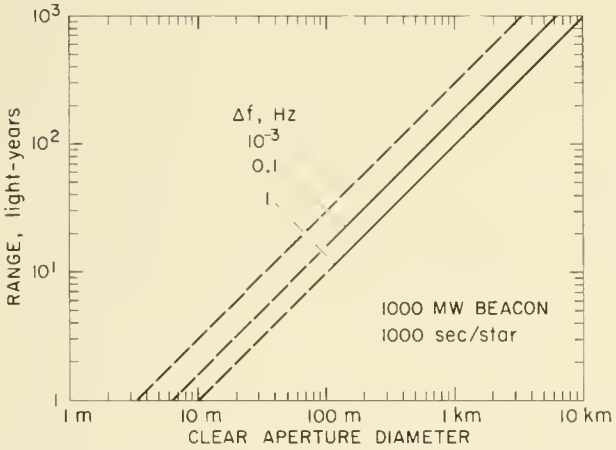


Figure 7-2. *Cyclops range capability.*

RANGE CAPABILITY

Figure 7-2 shows the range at which the Cyclops system could detect a 1000-MW beacon, assuming an observing time of 1000 sec per star. With a 1-Hz resolution in the optical analyzers the range is simply 100 light-years per kilometer of antenna diameter. Going to 0.1-Hz resolution increases this range by a factor of 1.6. The dashed curved marked $\Delta f = 10^{-3}$ is the performance of a system in which the receiver is matched to the observing time of 1000 sec. Its range is a factor of three greater than the 1-Hz system, but to achieve this performance the Doppler drift rate would have to be less than 10^{-6} Hz/sec, which is completely unrealistic.

THE COST OF CYCLOPS

The next four chapters contain cost estimates of the major subsystems of Cyclops. At this early stage, many of these estimates are quite crude. Accurate cost estimates cannot be made until detailed designs have been evolved, and even then are difficult unless production experience with similar systems is available. Nevertheless these estimates give a rough idea of the cost of Cyclops and how the cost divides among the various systems.

In Figure 7-3 the various costs have been plotted as a function of the effective clear aperture diameter of the array to permit easy comparison against the range performance curves of Figure 7-2. The clear aperture diameter is about one-third the physical diameter of the array, since the antennas are spaced by about three times their diameter.

Many of the hardware costs are roughly proportional to the number of antenna elements and therefore to the square of the clear aperture diameter. This is true for the tunneling, the receivers, the IF delay, and the imager. Aside from a fixed added amount it is true for the power system (where a half million dollars was added to account for other power uses) and for the antenna structures (where $200 million in initial plant and tooling was added). The cost of the IF transmission system varies as the 3/2 power of the number of antennas and therefore as the cube of the clear aperture diameter. This is because the array size increases as the square root of the number of antennas, and the average IF cable length increases accordingly.

The cost of the coherent signal detector depends principally on the IF bandwidth and the search time per star and is independent of the array size. We show a fixed figure of $160 million, which assumes optical analyzers with a 10^7 time-bandwidth product and 1000 sec observation time per star. If we were to use analyzers with a 10^6 time-bandwidth product and 100 sec observation time per star, this figure would drop to $45 million.

A rough guess of $100 million is shown for engineering costs. Building costs are shown at $10 million and the companion optical system at $2 million. Road and utility costs are highly site dependent and have not been included.

For apertures larger than a few hundred meters the cost of Cyclops is dominated by the antenna structural

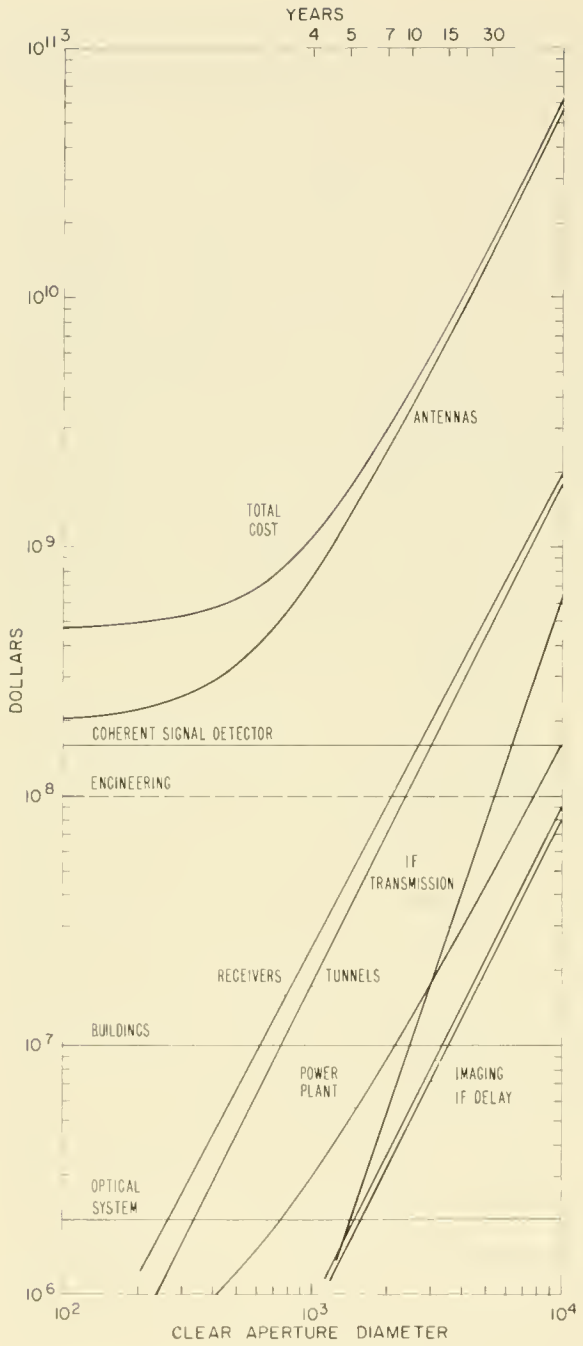


Figure 7-3. *Cyclops costs.*

cost. Here is where creative innovation would really pay. Unless the antenna costs can be greatly reduced, the other costs of the system, while large in their own right, are insignificant by comparison. Simply reducing the maximum operating frequency of the antenna elements from 10 GHz to 3 GHz might effect a factor of two saving in cost.

At the top of Figure 7-3 is a time scale that assumes 3 years of preliminary planning and site development followed by construction of antennas at the rate of 100 per year. At this rate of buildup, we see the cost of the Cyclops project is on the order of $600 million per year, on the average.

A COMPARISON OF THE CYCLOPS AND OZMA SYSTEMS

The only significant attempt ever made in this country to detect interstellar signals of intelligent origin was Project Ozma, mentioned in the last chapter. Table 7-1 compares the significant parameters of the Ozma system with those of the proposed Cyclops bogey system of 3.16 km clear aperture.

TABLE 7-1

Parameter	Symbol	Ozma	Cyclops
Antenna diameter	d	26m	3160 m
Antenna efficiency	η	0.5	0.8
System noise temperature	T	350°K	20°K
Resolved bandwidth	Δf	100 Hz	0.1 Hz
Integrating time	τ	100 sec	10 sec
Instantaneous bandwidth	B	100 Hz	200 MHz

Taking the sensitivity to be proportional to $(\eta d^2/T)(\tau/\Delta f)^{1/2}$ we see that the sensitivity *ratio* of the Cyclops to the Ozma System is

$$S_{c/o} = \left(\frac{0.8}{0.5}\right)\left(\frac{3160}{26}\right)^2\left(\frac{350}{20}\right)\left(\frac{10}{100}\,\frac{100}{0.1}\right)^{1/2}$$

$$= 4\times 10^{6}$$

Since the range limit for a given radiated signal varies as the square root of the sensitivity and since the volume that can be searched is proportional to the cube of range, we see that the range and volume ratios of Cyclops over Ozma are

$$R_{c/o} \approx 2000$$

$$V_{c/o} \approx 8\times 10^{9}$$

The target stars of Ozma, namely τ-Ceti and ϵ-Eridani, are about 10 to 11 light-years from us. Any signal that the Ozma system could have detected at this range could

be detected at 20,000 light-years by Cyclops, or could be 1/400 as strong and be detected at 1000 light-years.

In addition, the proposed Cyclops system searches two million times as broad a band in ten times the observation time. Its spectrum search rate is thus 200,000 times faster.

These comparisons are made not to disparage Ozma, but to build faith in Cyclops. Ozma cost very little and was a laudable undertaking, but the power of the Cyclops search system is so enormously greater that we should completely discount the negative results of Ozma. The τ-Cetacians or ϵ-Eridanians would have to have been irradiating us with an effective power of about 2×10^{12} W to have caused a noticeable wiggle of the pens of Ozma's recorders; 500 kw would be detected by Cyclops.

CONFUSION LIMITATION

If the number of detectable radio sources is so large that more than one is within the beam at all times, the system is said to be "confusion limited" (rather than sensitivity limited). Some concern has been expressed that Cyclops, with its enormous proposed area, might be hopelessly confusion limited. We do not believe this to be the case. In fact, we believe that Cyclops will be less prone to confusion limitation than smaller telescopes.

The crucial point in this question is the shape of the so-called log N versus log S relationship. This is the plot of the logarithm of the number N of detectable radio sources versus the logarithm of the limiting sensitivity S (measured in flux units) of the radio telescope used. The lower S is, the higher the sensitivity. This curve is shown in Figure 7-4. For relatively low sensitivities (S large) the number of detectable sources rises more rapidly with decreasing S than S^{-1}. For low values of S the magnitude of the slope decreases.

Obviously we do not know the shape of the curve for smaller values of S than our present instruments provide. This is one thing of great interest that Cyclops would tell us. However, we do know that the slope cannot continue indefinitely to have a magnitude greater than 1. As Martin Rees points out, if the slope were -1, then each new decade of log S to the left would contribute the same total radio flux from the sources picked up in that decade, and the total radio flux from the sky would be infinite.

For a telescope of a given aperture size, a decrease in system noise temperature may, and a sufficient increase in integration time always will, cause a changeover from sensitivity limitation to confusion limitation, unless the aperture is already large enough to resolve, at the operating wavelength, all the sources that in fact exist at

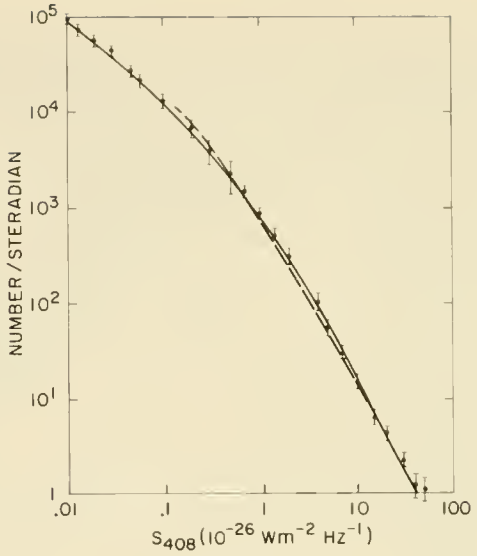


Figure 7-4. *Log* N-*log* S *relation at 408 MHz. Units of* S *are* 10^{-26} *W/m*^{2} *Hz. (Pooley and Ryle, 1968)*

that wavelength. In other words, so long as the log N-log S curve continues to rise as log S decreases, we can produce confusion limitation in a given system by extending the integration time sufficiently.

However, if we increase the system sensitivity by increasing the aperture, we simultaneously increase the resolution and, for confusion limitation to occur, the magnitude of the slope of the log N-log S curve must be unity or greater. For example, consider a telescope of filled aperture A that, for a given integration time, τ, is not confusion limited and can resolve n sources in the solid angle Ω. If the area is increased to kA the instrument can now resolve kn sources in the same solid angle, and this is exactly the number it could detect in the same integration time if the slope of the log N-log S were -1 between the two values of S.

In going from a single 100-m dish to an array of 1000 such dishes covering an area 10 km in diameter, we have reduced the filling factor by 10 to 1. Although the sensitivity has increased by 10^3, the resolving power has increased by 10^4. If, for a given integration time, the elements of the array are not confusion limited, the array will not be either unless the magnitude of the slope of the log N-log S curve over the range is greater than 4/3, which seems unlikely.

We predict that longer integration times will be needed to produce confusion limitation for the Cyclops array than for its elements.

View 3. *Artist's concept of low aerial view of a portion of the Cyclops system antenna array, showing the central control and processing building.*

8. ANTENNA ELEMENTS

A tremendous amount of effort has already been devoted to the design of large antennas for radio astronomy, radar, satellite tracking and other applications. The designs that have been developed over the past two decades reflect an increasing sophistication in the efficient structural use of materials, more recently as a result of the use of computer aided design. Each of these designs has had its own set of specifications as to surface tolerance, sky coverage, environmental conditions for operation and survival, and other factors. Thus intercomparison of designs is extraordinarily difficult.

The limited time and manpower available for the Cyclops study precluded the possibility of designing in detail an antenna element for mass production and the associated fabrication, assembly, and erection tooling required to substantially reduce the high labor content of one-of-a-kind designs. Instead, we have been forced to draw some rough estimates of savings that might result from the application of these techniques to state-of-the-art designs. The cost estimates arrived at may be pessimistic, but can only be improved with confidence by a much larger scale funded study.

CYCLOPS REQUIREMENTS

At the outset of the Cyclops study some tentative specifications were set down as guidelines. These were:

Total equivalent antenna diameter $\approx$ 10 km (max)
Frequency range $\approx$ 500 MHz to 10 GHz
Minimum elevation angle = 20° max.
Wind: 20 mph maximum for 10 GHz operation
100 mph minimum for survival

The total effective diameter of 10 km was based on an early estimate of what might be required to achieve a detection range of 1000 light-years. It now appears that this figure may be high and that a diameter as low as 3 km might suffice (Chap. 6).

The frequency range was chosen to cover the microwave window. If the arguments presented earlier for favoring the low end of the microwave window are accepted and if alternative uses of the array do not require operation at 10 GHz with full efficiency, then the surface tolerances required will not be excessive.

The minimum elevation angle was chosen on the following grounds:

1. Operation below 20° elevation increases the system noise temperature appreciably.
2. Operation below 20° elevation rapidly increases the element separation required to prevent self-shadowing.
3. Operation down to about 20° is required to permit continuous reception with three arrays spaced at 120° around the world or to permit very long base-line interferometry using pairs of such arrays.

The wind specifications ultimately would have to take into account the weather characteristics at the chosen site. The values used in this study would allow operation at 10 GHz for over 75% of the time with survival expectations of over a century at many suitable locations in the southwest United States.

TYPES OF ELEMENTS

The usual radio astronomy antenna designed for use in the microwave region consists of a paraboloidal dish so mounted that it can be directed at most of the sky that is visible at any one time. Many other types of elements have been proposed and some have been built that sacrifice full steerability or operating frequency range or beam symmetry to obtain a lower cost per unit of collecting area. Some examples are:

1. A fixed spherical dish pointed toward the zenith with tiltable feeds that illuminate a portion of the dish and allow limited steering of the beam. The

feeds must correct for spherical aberration and as a result tend to be narrow band. The Arecibo telescope is an example of this arrangement. Bandwidths can be increased by the use of triple mirror configurations.

2. Fixed parabolic troughs with north-south axes and line feeds along the line focus. Again the feeds are difficult to broadband, and without east-west tilting of the trough the instrument can only aim at the meridian.
3. Fixed paraboloidal sectors pointed at tiltable flat mirrors. Again, unless the flat mirror can rotate in azimuth, only near meridian observation is possible, and the sky coverage in declination is limited.

In addition to Earth-based elements, large single antennas have been proposed for space use (ref. 1). None of these has had the accuracy required for operation in the microwave region. Although not subject to gravity and wind stresses, space antennas are subject to thermal gradients from the sunlight. In addition, a large single dish 3 km or more in diameter would have to be rather rigid to avoid large amplitude, very low frequency modes of vibration, which would be excited under repointing maneuvers. Even the flimsiest of structures—a balloon of 1 mil mylar 10 km in diameter (which would permit a 3-km diameter spherical reflector as part of the surface)—weighs about 8000 metric tons. At a cost of $100 per pound to put it in *synchronous* orbit, the cost would be $1.8 billion. This does not include cost of assembly in space, nor the weight of receivers, transmitters, pointing rockets, servos, and all the other needed equipment. We do not mean to exclude space antennas from consideration, but we cannot consider a microwave antenna of 3-km diameter or more in space to be within the present or near-future state-of-the-art.

For these reasons we largely confined our thinking to more or less conventional steerable dishes to be used as elements of an array. If further study should reveal a less costly approach, the effect will be to *reduce* the cost of the Cyclops system.

TYPES OF MOUNTS

Steerable dishes, particularly in the smaller sizes, are often mounted equatorially. The equatorial mount has the advantages that (1) tracking of sidereal motion requires a constant rate of rotation of the polar axis only, and (2) there is no singularity in the sky through which tracking must be interrupted. For reasons of economy, large fully steerable dishes are almost always mounted in alt-azimuth mounts. In this form of mounting, rotation about both the azimuth (vertical) axis and the elevation (horizontal) axis occurs at nonuniform rates while a star is being tracked. In addition, a star passing directly overhead requires an abrupt rotation of 180° in azimuth to maintain tracking. These, however, are minor disadvantages. The axis motion is readily programmed into a computer and the zenith singularity can almost always be avoided. Because the az-el mount requires less counterweighting and permits a lighter, less complicated base structure and because more experience has been gained with large az-el mounts than with large equatorial mounts, the az-el mount has been selected for Cyclops.

Az-el mounts fall into two rough categories: the *king-post* design and the *wheel and track* design. In the former, a very massive rigid column, or king post, carries the azimuth bearings and supports a yoke or crosspiece, which serves as the elevation axis bearing support. In the latter, the elevation bearings are widely separated and carried by a truss structure, which in its entirety rotates on trucks carried on a circular track. Lateral constraint can be supplied by a simple central bearing at ground level.

The king-post design is suitable for small dishes or for dishes enclosed in radomes. For large dishes exposed to the wind, the wheel and track design appears to offer the required stiffness and strength against overturning wind moments with much lighter members and smaller bearings. In addition, the large-radius track permits simple angle encoders in azimuth while the open structure allows a large radius elevation drive and simple encoders on this axis as well. Sandstorms and ice pose problems for the wheel and track design but these appear controllable with appropriate seals or pressurization or both. Thus, if large dishes are used for Cyclops, wheel and track alt-azimuth mounts are probably indicated.

The element design group gave some thought to some novel base structure (see Appendix F). Because of the time limitations of the study, careful costing and, in certain cases, careful evaluation of the structural stability problems involved could not be carried out. Hence, the economies of these designs could not be ascertained, particularly if sky coverage is sacrificed.

SURFACE TOLERANCE

The function of the backup structure of the reflecting surface is to hold the shape of the surface within prescribed tolerances under all operating conditions of wind, gravitational, and thermal stress. In addition, buckling or misalignment of the surface panels can take their toll from the error budget. The allowable surface errors are proportional to the minimum wavelength and hence inversely proportional to the highest operating frequency at which a given efficiency is desired.

If the surface irregularities are large in lateral extent compared with a wavelength, but small compared with the diameter of the dish and are normally distributed, their effect is readily computable. If the surface of the reflector near the axis is displaced toward the focus by a distance δ, the phase of the received wave from that element of area will be advanced by an amount $\phi = 4\pi\delta/\lambda$, where λ is the wavelength. If δ is normally distributed, ϕ will be also and we have

$$p(\phi) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\phi^2}{2\sigma^2}} \tag{1}$$

where σ is the rms deviation in ϕ. The in-phase contribution of each element is proportional to $\cos\phi$, so

$$\frac{a}{a_0} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\phi^2}{2\sigma^2}} \cos\phi \, d\phi = e^{-\frac{\sigma^2}{2}} \tag{2}$$

where a_0 is the amplitude that would be received at the focus from a perfect surface, and a is the actual amplitude. The efficiency η, which is the ratio of the actual power gain g to the gain g_0 of a perfect surface, is thus

$$\eta = \frac{g}{g_0} = \left(\frac{a}{a_0}\right)^2 = e^{-\sigma^2} = e^{-\left(\frac{4\pi\delta_{rms}}{\lambda}\right)^2} \tag{3}$$

For $\eta = 1/2$,

$$\frac{4\pi\delta_{rms}}{\lambda} = \sqrt{\ell n 2}$$

$$\delta_{rms} \approx \frac{\lambda}{15} \tag{4}$$

Figure 8-1 shows a plot of equation (3) over the frequency range of interest. We see that if we allow $\eta = 0.5$ at 10 GHz, $\delta_{rms} = 2$mm (0.079 in.) and that the efficiency over the primary range of interest below 3 GHz is 0.95 or higher. Actually because the illumination at the rim is less and *normal* departures of the surface introduce less phase shift there, the errors at the rim of the dish can be somewhat greater.

SIZE LIMITS FOR SELF-SUPPORTING STRUCTURES

As the size of a given structure is scaled in all its linear dimensions by a factor k the linear deflections under its own weight vary as k^2. Since we can stand deflections proportional to wavelength we can scale a

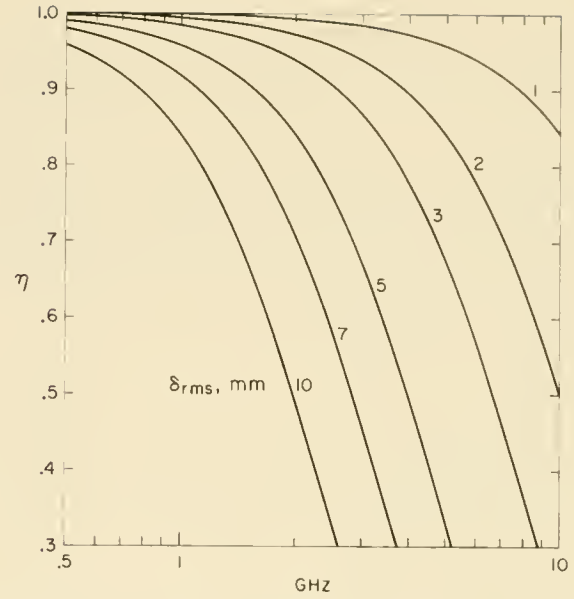
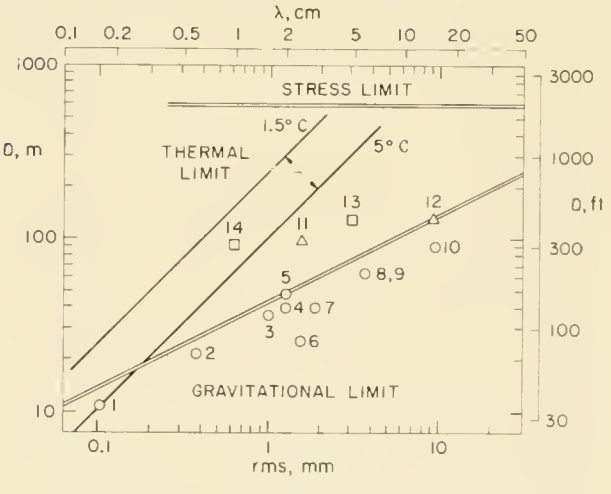


Figure 8-1. *Antenna efficiency versus surface tolerance.*

given structure in proportion to $\lambda^{1/2}$, where λ is the minimum operating wavelength. This accounts for the line with slope 1/2 marked "Gravitational Limit" in Figure 8-2. Thermal strains, produced by sunlight on one



○ EXISTING
1. 36 ft, NRAO KITT PEAK
2. 22 m, LEBEDEV, SERPUKHOF
3. 120 ft, MIT, HAYSTACK
4. 140 ft, NRAO, GREEN BANK
5. 150 ft, ARO, CANADA
6. VARIOUS 85 ft TELESCOPES
7. 130 ft, OWENS VALLEY
8. 210 ft, PARKES, AUSTRALIA
9. 210 ft, JPL, GOLDSTONE
10. 300 ft, NRAO, GREEN BANK

△ WITHIN 1-2 YEARS
11. 100 m, BONN, GERMANY
12. 450 ft, JODRELL BANK

□ IN PREPARATION
13. 440 ft, CAMROC
14. 300 ft, HOMOLOGOUS DESIGN

Figure 8-2. *Diameter D and shortest wavelength, λ. Three natural limits for tiltable, conventional telescopes.*

side of the dish and shade on the other, for example, cause deformations proportional to dish size and these account for the line marked "Thermal Limit." Finally the strength of materials sets a size above which the dish will collapse under its own weight. This is marked "Stress Limit" in the figure, and represents a dish 600 m in diameter. Also shown on the figure are points corresponding to a number of existing and proposed telescopes. Notice that three of these points exceed the gravitational limit. In the case of the proposed 440-ft NEROC (or CAMROC) telescope (point 13), this is achieved by compensation, that is, by mechanical change in the length of certain structural members as a function of elevation angle. In the case of the Bonn 100-m telescope (point 11) and to a much greater degree in the proposed 300-ft NRAO design (point 14), the gravitational limit is exceeded by making use of the principle of homologous design.

In homologous design, the object is not to eliminate deflections by making the structure as stiff as possible everywhere, but to allow greater deflections in certain regions than would normally be present in conventional design. Adding this compliance makes it possible to control the deflections so that, under changing direction of the gravity vector, the paraboloidal surface deforms into a new paraboloid. The new surface may be simply the old surface shifted and tilted slightly, and the feed is repositioned accordingly. In this case, the deflections merely produce an elevation angle error as a function of elevation angle, which may be removed by calibration.

Thus the gravitational limit shown in Figure 8-2 applies only to "conventional" designs and may be exceeded by active structures or structures designed by modern computer techniques. We conclude that dishes 100 m or more in diameter operable down to 3-cm wavelength are well within the present state of the art. However, it should be noted that the application of the homologous design principle to date does not seem to yield a configuration well suited for quantity production. This is not a criticism of the procedure, for the homologous design approach has been used to date only for "one of a kind" antennas. If this design technique were to be useful for Cyclops, further study is needed to assure that the design evolved using homology exhibits the required features for mass production efficiencies.

Some of the techniques for the design of the back up structure are listed in Appendix G. The backup structure cannot be designed independently of the base, since the positions of load support points, and the force vectors and torques introduced at these points, must be known.

OPTIMUM SIZING

In an array, the total area can be obtained with a certain number of large dishes or a larger number of smaller ones. We would like to choose the dish size so that the total cost is minimized. Following an analysis by Drake we let the cost per element be

$$C = ad^x + b \tag{5}$$

where

d = dish diameter

a = a constant

b = fixed cost per element, that is, cost of receivers, control equipment, IF transmission and delay circuit, etc.

To realize a total equivalent antenna area, A, we need a number of dishes

$$n = \frac{4}{\eta}\frac{A}{\pi d^2} \tag{6}$$

where η is the efficiency of utilization of the dish surface. Thus, the total cost is

$$nC = \frac{4A}{\pi\eta}(ad^{x-2} + bd^{-2}) \tag{7}$$

Differentiating equation (7) with respect to d, we find that nC is a minimum when

$$ad^x = \frac{2}{x-2}b \tag{8}$$

that is, when

$$\text{Structural cost} = \frac{2}{x-2}\text{ fixed channel cost} \tag{8a}$$

The optimum size thus depends heavily on x, that is, upon the exponent that relates the structural cost[1] to the diameter. If $x > 2$ there will be an optimum size given by equation (8). If $x \leqslant 2$ there is no optimum and one uses the largest dishes that can be built. It is therefore of great interest to determine x, and a serious effort was made to do this by correlating the costs of available antenna designs that were believed to represent the full usage of modern design technique.

[1] We include in the "structural cost" the cost of the dish surface the servodrive system, and the foundation, as well as the structural support itself.

ELEMENT COST VERSUS SIZE

Here we present the results of earlier studies and a description of the study made by the Cyclops group to determine the relation between cost and antenna size.

Previous Studies

Reference 2 presents cost data on existing and some proposed designs, with the costs updated to reflect 1966 prices. The information pertinent to our study is shown in Figures 8-3 and 8-4, which are graphs taken directly from reference 2. Figure 8-3 indicates that the weight of

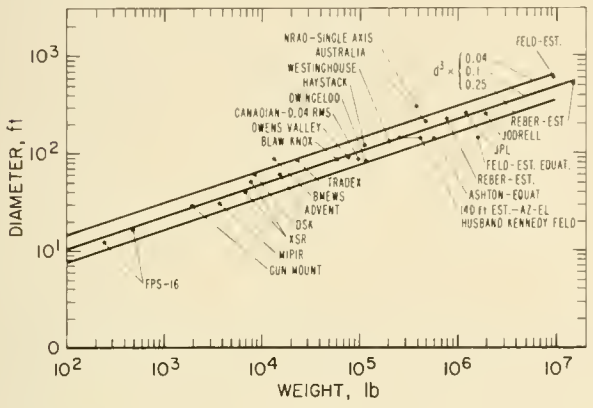


Figure 8-3. *Antenna weight versus diameter.*

the structure increases roughly as the cube of the diameter. Figure 8-4 shows that, on the basis of the data available at that time, the antenna *cost* for az-el mounts also increased roughly as the cube of diameter. There is, however, a considerable dispersion, which may reflect the different accuracy specifications of the various designs as well as the type of mount used. Reference 2 concluded that at that time (1966) published data were insufficient to establish any exact formula relating cost

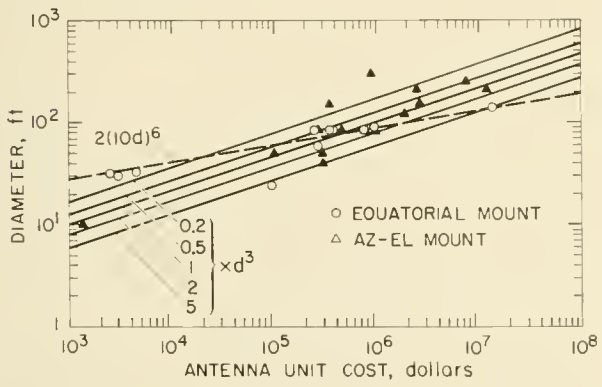


Figure 8-4. *Antenna cost versus diameter.*

to surface accuracy. It did note that all the equatorial mounted dishes lay on a line (the dashed line in Fig. 8-4 indicating cost proportional to the *sixth* power of diameter.

Figure 8-5 reproduces cost data from Potter et al. (ref. 3). This curve, for az-el mounts only, indicates that cost varies as diameter to the 2.78 power.

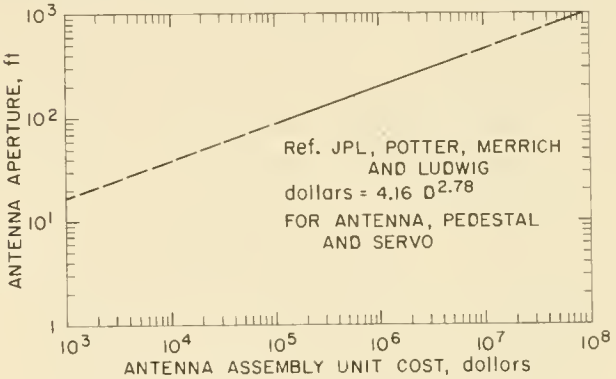


Figure 8-5. *Diameter versus antenna unit cost.*

On the basis of the above results it would seem that $2.78 < x < 3$ in equation (8) and that an optimum diameter (smaller than the largest practical diameter) might be found for the Cyclops array elements.

Updated Study

Since the publication of reference 2 a number of antenna element design studies have been made, and reasonably accurate cost estimates have been obtained (refs. 4-6). In addition the Bonn 100-m antenna has been built (ref. 7), and actual cost figures for that structure are available. Some of these newer systems have design requirements comparable to those for Cyclops antenna element, while others have more stringent accuracy (NRAO 300 ft dish) or environmental (Bonn 100 m, NEROC 440 ft dishes) specifications. The latter were nevertheless included as examples of modern computer aided design. The following Table 8-1 lists the units included in the study and the raw cost data.

The "Basic Structural Cost" column includes the costs of the antenna structure and surface, the drives and servosystems, the secondary reflector and support, the position encoders, foundations, etc. The actual basic structural costs for the NRC and Bonn dishes were not available. The figures shown were obtained by multiplying the total cost figures by 0.62. This factor is the average ratio of basic structural cost to total cost for the

TABLE 8-1

UNADJUSTED UNIT COST DATA

Unit	Total Cost $	Basic Structural Cost $	Reference	Dollar Year
NRC 150 ft	3.5×10^6		6–vol. 2, p. 131	1966
NEROC* 440 ft	39.4×10^6	24.68×10^6	6–vol. 2, p. 132	1970
NRAO 300 ft		7.29×10^6	5–chap. 6 p. 2	1969
VLA I** (Av. Unit Cost) 32-82′	$.808\times10^6$	$.468\times10^6$	4–vol. 1 chap. 10	1966
VLA II** (Av. Unit Cost) 27-82′	$.892\times10^6$	$.442\times10^6$	4–vol. 3, chap. 9	1968
PARKES***	2.5×10^6	1.705×10^6	2–p. 4-58 (graph)	1966
210 ft	2.2×10^6	1.5×10^6	8 83	1961
BONN 328 ft	9.8×10^6		7	1969

*Includes 15% contingency
**Less transports and trackage
***Adjusted from 1961 costs at 2.5% per year

other systems shown with the two VLA designs each given half weight because VLA figures were considered less firm and the two designs are not independent.

The basic structural cost figures were then corrected to 1970 dollars by applying an escalation factor of 9% per year. Further, the foreign-built units were corrected to U.S. prices by multiplying by 1.5 except for the Canadian (NRC) unit where the factor 1.25 was used. Finally a 15% contingency was added to all units in the "proposed" status, except for the NEROC design, which already included this factor. The results of all these operations are shown in Table 8-2.

TABLE 8-2

BASIC STRUCTURAL COST COMPARISON
1970 U.S. COSTS

Unit	Diam., ft	Basic Structural Cost ($ million)
NRC	150	3.84
NEROC	440	24.68
NRAO	330	9.15
VLA I	82	.763
VLA II	82	.605
PARKES	210	3.62
BONN	328	9.94

Figure 8-6 indicates an excellent fit by a line showing cost proportional to diameter *squared.* A regression analysis showed that the data of Table 8-2 are best fitted by the cost-diameter relation $C = 107.4\, d^{2.001}$ where C is the cost in dollars and d is the diameter in feet. Rounding off the exponent to 2, and *converting to meters* we get

$$C = \$1156\, d^2 \qquad (9)$$

The exponent value, $x = 2$, in this relation implies that the total structural cost of an array is independent of the element size, and that the system cost will therefore be minimized by using the largest possible elements.

Equation (9) is surprising in view of the larger exponents found by earlier studies and in view of the approximate cube law for the weights of structures (cf. fig. 8-3). If one accepts equation (9) as correct, at least for the size range of 25- to 150-m antennas, the explanation must be that the material costs, which vary more like d^3, are diluted by labor costs that increase with size *less rapidly* than d^2. The material costs are a small fraction of the total cost, so this is easily possible. For example, the total moving weight of the Bonn 100-m dish is 3200 tons. At 10 cents per pound this represents a raw material cost of only $640,000 out of

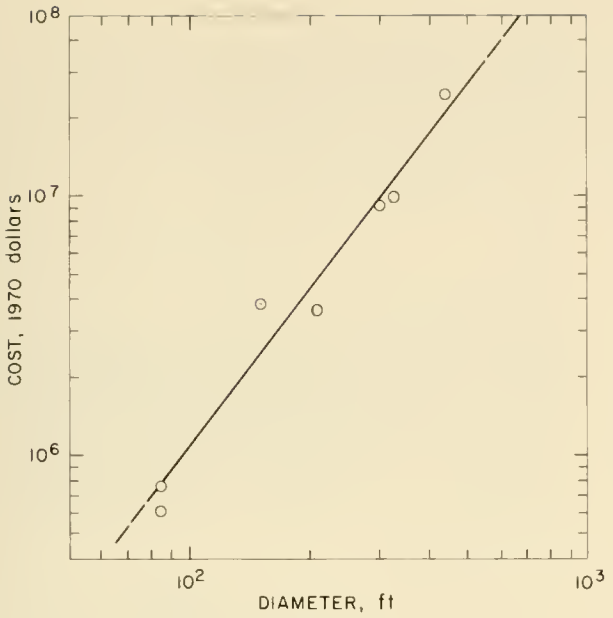


Figure 8-6. *Antenna cost versus diameter.*

$10 million. The foundation and the mechanical drive system (i.e., servos, gears, bearings) might add another $900,000 to bring the total material and purchased parts to around $1.5 million. This still leaves roughly $8.5 million for the labor of fabrication, assembly, erection and testing.

Another reason for the low value of the exponent is the increased sophistication in design capability from the use of computers. It should be realized that the constancy of the exponent breaks down at some value of diameter. It is obviously not as cheap on an area basis to make a 10 km diameter steerable dish as a 100-m diameter dish. However, we believe that the exponent remains constant through the sizes considered in this study.

Based on equation (9) the total structural cost for an array will be simply

$$C = \$1156\, d_a^{\,2}\, \frac{R}{\eta} \qquad (10)$$

where

d_a = equivalent clear aperture diameter of the array

R = cost reduction factor from mass production

η = aperture efficiency of the elements.

The efficiency η is the product of the efficiency due to surface tolerances and the illumination efficiency. With careful feed horn design and for frequencies up to about one-third the 3 dB cutoff frequency due to surface tolerances, we can expect $\eta \geqslant 0.8$.

MASS PRODUCTION SAVINGS

Except for the VLA design (which involved small quantity production) the costs that were used in establishing the cost-versus-size relationship (9) were the material and labor costs for producing a single unit. With large volume production substantial reductions in the unit cost are to be expected. Two cases need to be considered: volume production of an existing design, and semiautomated production of a design adapted to mass production methods.

When an existing design is produced in quantity at a constant rate, cost reductions can occur through

1. Contract purchase of large volumes of materials
2. Direct factory purchase with scheduled delivery of prefabricated purchased parts
3. Efficient layout of fabrication and assembly lines
4. Efficient work scheduling and labor deployment
5. Reduction of fabrication and assembly labor through tooling, jigs, and fixtures
6. Reduction of fabrication and assembly time from accrued experience in doing each operation
7. On-site production

Contract purchases of materials could easily reduce the material costs by 10% while factory purchases of prefabricated purchased parts (bearings, gears, servomotors, etc.) can save up to 40% for these items. Accurate estimates of labor cost reductions cannot be made without an exhaustive detailed study, or past experience. However, it is typical in a wide variety of products for start-up costs (which represent initial production of one to several units) to exceed final costs (which represent steady-state experienced production) by 150% to 200% or more.

When mass production is anticipated, additional savings are possible by:

1. Designing the structure to take advantage of well known low cost processes
2. Integrating the design of the product and the factory to produce it
3. Eliminating all selective assembly and hand adjustment through extensive tooling
4. Replacing hand fabrication by stamping, die forming, die casting and other suitable processes
5. Making widespread use of automated numerically controlled machines both for piece-part production and assembly
6. Using automatically fabricated material-saving tapered structural sections

7. Using huge specially designed jigs for final assembly and erection

An antenna designed for mass production could use cigar shaped major structural members rolled and welded from plate or sheet stock. All structural members could be cut to finished length and drilled on numerically controlled machines. Radial trusses could be assembled in large jigs that eliminate the need for any alignment or measurement. Partially automated welding or even complete one stop brazing of the entire truss might be possible. Surface panels could be stamped into double curved surfaces and formed with stiffening edge lips in large single-shot presses. Stamped channels with preformed profiles could be affixed to the rear of the panels by multiple-head, automatically sequenced spot welders. Completed trusses and panels could be assembled at the antenna site on a large lazy susan jig which could also serve to raise a completed dish into position on the mount. Special vehicles would convey the finished parts from the on-site factory to the antenna location.

The structures group discussed the possible savings due to quantity production of the Cyclops antenna element (assumed to be a 100-m az-el mounted dish) with a large shipbuilding firm, two leading engineering firms, a prominent "think-tank," a large aerospace corporation and a number of antenna manufacturers. As a result of these discussions and application of the sophistries of learning theory a total cost *reduction* from mass production of 20 to 40% was estimated.

Others feel that this estimate is too conservative and that full-scale inventive application of the arsenal of mass production techniques could result in a cost reduction of 60 to 70%. This question can be resolved only by a full scale design study, which of course was impossible to accomplish in the summer study period.

The following table gives estimated total structural costs for the Cyclops array as computed from equation (10) with $\eta = 0.8$ and for mass production cost reduction factors $R = 0.7$ and $R = 0.4$. For the latter case a tooling cost of $200 million has been added.

TABLE 8-3

ESTIMATED STRUCTURAL COSTS FOR CYCLOPS ARRAYS

Equivalent diameter, km	Cost in $ Billions $R = 0.7$	$R = 0.4$*
1	1.	0.78
2	4.	2.5
3	9.	5.2
5	25.	14.

*Includes $200 million tooling costs

Assuming an ultimate size of 5 km for Cyclops we see that the structures cost is in the $10 to $25 billion range. Since this is the dominating cost of the entire Cyclops system, a large study aimed at reducing this figure and refining the accuracy of the estimate would appear to be the first order of business.

ACKNOWLEDGMENTS

During the course of this study, the element design group had the benefit of many helpful discussions from members of the industrial community. We sincerely appreciate and acknowledge the interest, information, and suggestions received from our meetings and/or written and phone conversations with the following firms and individuals:

Philco-Ford Corporation (Palo Alto, Calif.)
I.E. Lewis, R. Melosh

Rand Corporation (Los Angeles, Calif.)
Milton Kamins, Sue Haggart

Lockeed Corporation (Sunnyvale, Calif.)
R.M. Rutledge, V. Wise

Bechtel Corporation (San Francisco, Calif.)
David J. Goerz, Jr.

Bethlehem Steel Company (San Francisco, Calif.)
E.J. Stuber, L.A. Napper

Nippon Electric Company America, Inc. (N.Y.C., N.Y.)
Robert Alarie

Tymeshare Corporation (Mt. View, Calif.)
C. Love

Stanford University (Palo Alto, Calif.)
Ronald Bracewell

Synergetics Corporation (Raleigh, N.C.)
T.C. Howard

Rohr Corporation (Chula Vista, Calif.)
Robert Hall

REFERENCES

1. Schuerch, Hans U.; and Hedgepeth, John M.: *Large Low Frequency Orbiting Telescope.* NASA CR-1201, 1968.
2. *VLA Antenna Construction and Emplacement Study.* Final Report, prepared by R.C.A. Defense Electronics Products Missile and Surface Radar Division, Moorestown, N.J., Nov. 1966.
3. Potter, P.D.; Merrick, W.D.; and Ludwig, A.C.: Big Antenna Systems for Deep Space Communication. *Astronautics and Aeronautics,* vol. 4, no. 10, Oct. 1966, pp. 85-95.

4. *The VLA: A Proposal for a Very Large Array Radio Telescope.* (3 vol) National Radio Astronomy Observatory, Green Bank, West Virginia, Jan. 1967.
5. *A 300 Foot High High-Precision Radio Telescope.* National Radio Astronomy Observatory Corp., June 1970.
6. *A Large Radio–Radar Telescope Proposal for a Research Facility.* (4 vol) Northeast Radio Observatory Corp., June 1970.
7. Wielebinski, R.: 100-m Radio Telescope in Germany. *Nature,* vol. 228, Nov. 7, 1970, pp. 507-508.
8. *Large Steerable Radio Antennas–Climatological and Aerodynamic Considerations.* Vol. 116 of Annals of the New York Academy of Sciences, Art. 1, Proceedings of a 1963 Conference, E. Cohen, ed.

REFERENCES USED BUT NOT CITED

Cost Savings Through Realistic Tolerance. Space Systems Division Manufacturing, Lockheed Aircraft Corp., Sunnyvale, California.

Bauer, C.J.: Analysis of Manufacturing Cost Relative to Product Design. ASME Paper 56–5A-9.

Johnson, H.A.: From One Tenth to Nothing–The Gaging Tightrope. ASTME 403.

Structures Technology for Large Radio and Radar Telescope Systems. J.W. Mar and H. Liebowitz, eds., MIT Press, 1969.

Hooghoudt, B.G.: *Cost Consideration–Synthesis Radiotelescope.* Westerbrook, Holland Design Report, Aug., 1966.

Asher, H.: *Cost-Quantity Relationships in the Airframe Industry.* The RAND Corporation, R-291, 1956.

Concepts and Procedures of Cost Analysis. The RAND Corporation, RM-3859, 1969.

Russell, J.A.: Progress Function Models and Their Deviations. J. Industrial Engineering, vol. XIX, no. 1, Jan. 1968.

Engineering Study and Analysis of a Large Diameter Antenna System. Philco Corp., Contract No. NASA-10-452, May 22, 1963.

Weiss, H.G.: Design Studies for a 440-Ft-Diameter Radio and Radar Telescope. *Structures Technology for Large Radio and Radar Telescope Systems,* J.W. Mar and H. Liebowitz, eds., MIT Press, 1969, pp. 29-54.

Weidlinger, P.: Control of RMS Surface Error in Large Antenna Structures. pp. 287-310.

Processing of Vectorvane Data. Meteorology Research, Inc., Contract No. 951472 JPL, Dec. 15, 1967.

Special Issue in Satellite Power Station and Microwave Teams Mission to Earth. J. Microwave Power, vol. 5, no. 4, 1970.

9. THE RECEIVER SYSTEM

The receiver system consists of three subsystems: (1) the antenna feed system, (2) the signal conversion system, and (3) the local oscillator system, as shown in Figure 9-1. The following were considered to be the most important requirements for the Cyclops receiver design:

1. *High aperture efficiency.* Because the major expense of the Cyclops system is the cost of the collecting area, it is imperative that the area be used efficiently. A 10% increase in aperture efficiency could represent a saving of a billion dollars in the full-scale system.
2. *Low noise temperature.* A two to one reduction in system noise temperature is equivalent to a two-fold increase in antenna area. The cost of liquid helium cryogenics is saved many times over in antenna cost, and only the lowest noise front ends were considered. For the same reason only very low ground spillover feeds can be used.
3. *Wide instantaneous bandwidth.* The wider the bandwidth that can be searched simultaneously, the less will be the search time per star. Ideally, we would like to search the entire microwave window at one pass but present technology and interfering signals from Earth do not allow this.
4. *Rapid remote band switching.* Because no low noise receiver can cover the entire band of interest, and because on-site receiver tuning is out of the question in an array of a thousand or more elements, the receivers must be rapidly and remotely tunable and capable of being band-switched from the control center.
5. *Universal polarization capability.* Cyclops must be able to receive with minimum loss a signal having an arbitrary polarization. Beacons may be expected to be circularly polarized (Chap. 6) but may be either left or right, and leakage signals could have any polarization.
6. *Automatic phasing, calibration, and fault detection.* The Cyclops array must remain properly phased at all times, with all elements delivering the same signal amplitude without operator attention. This means the system must incorporate computer-controlled calibration and fault-location systems.

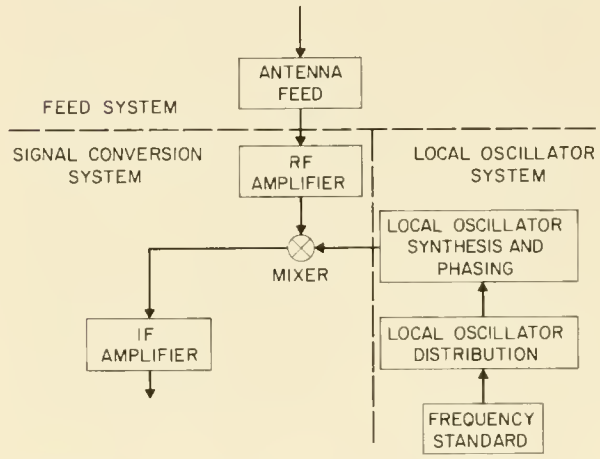


Figure 9-1. *Major elements of the Cyclops receiver.*

The receiver design outlined here is believed capable of meeting the above criteria as well as the present state of the art allows. The design is not complete in detail and, of course, in some areas, such as optimization of feeds, a great deal of experimental work would be needed to arrive at the best final design. Nevertheless, the design permits some rough estimates to be made of performance and cost.

The proposed design covers the low end of the microwave window from 500 MHz to 3 GHz and is capable of being extended to 10 GHz if desired. Six

bands, each covering a 1.35 to 1 frequency range are used, giving the following ranges:

Band	Range (GHz)	Coverage (MHz)
1	0.5 -0.675	175
2	0.675-0.91	235
3	0.91 -1.23	320
4	1.23 -1.66	430
5	1.66 -2.24	580
6	2.24 -3.02	780

The portion of the receiver associated with each band comprises a circularly symmetric feedhorn coupled to two up-converters, one for each orthogonal polarization mode. The up-converters in use are cooled to 20° K and their outputs are fed to two masers operating at 10 GHz (or higher) and cooled to 4° K. The maser outputs are further amplified and then down converted to the IF frequency. Precision frequencies derived from a hydrogen maser are used for the up- and down-conversions, and the array phasing is done in the final down conversion.

Two IF channels per receiver are proposed, one for each polarization. As shown in Appendix H, the choice of orthogonal polarizations is arbitrary. Thus vertical and horizontal linear, or right and left circular polarization may be used. From the two polarizations selected four additional polarizations may be resolved at the central processing station to give a total of six: V, H, $V + H$ (or 45°), $V - H$ (or 135°), $V + jH$ (left circular) and $V - jH$ (right circular). If all six are processed the average loss, for a signal of unknown polarization, is 0.4 dB and the maximum loss is 1 dB. If only four polarizations are processed the average loss is 0.7 dB and the maximum loss is 3 dB. Since 3 dB represents a loss of half the antenna surface and since processing costs are much less than antenna costs, it is recommended that all six polarizations be processed.

A detailed block diagram of the dual polarization Cyclops receiver is shown in Figure 9-2. Below the horizontal dashed line in the figure the blocks shown are used for all bands. Above the dashed line a complete set of the blocks shown is needed for each of the six proposed bands. The various blocks and their functions are described in greater detail in the sections to follow.

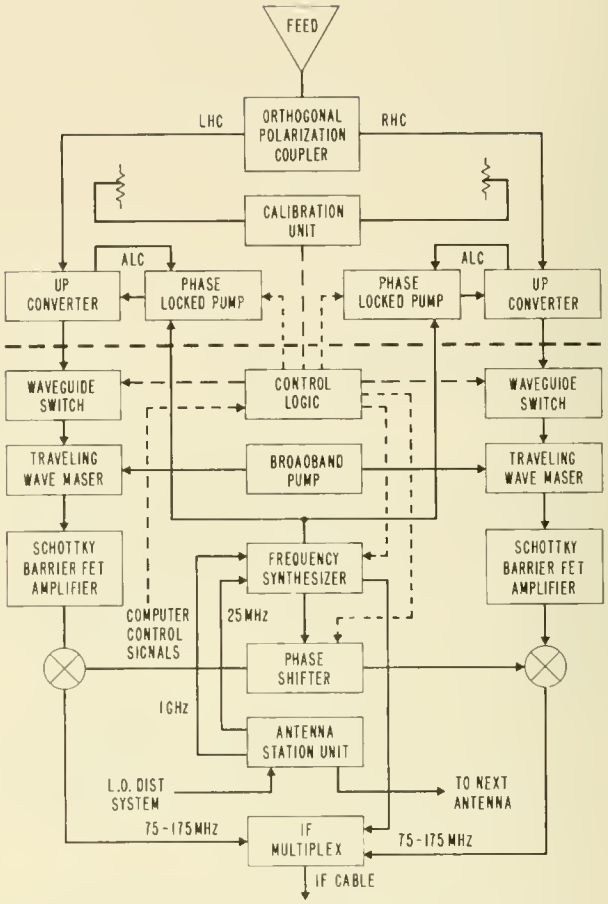


Figure 9-2. *Dual polarization receiver.*

ANTENNA OPTICS AND FEED SYSTEM

Because antennas and the associated feed systems generally obey reciprocity, we may analyze the operation in either the transmission or the reception mode as best suits our convenience. Although we are primarily concerned with reception, we may, for example, choose to speak of the uniformity with which the feed horn "illuminates" the antenna surface realizing that this also describes the uniformity of receiver sensitivity to signals received from various portions of the antenna surface.

If $U(\rho,\phi)$ is the amplitude of the field over a plane normal to the axis of the paraboloid and bounded by its rim, then the far field on-axis radiation intensity at the distance R is

$$|u|^2 = \frac{1}{\lambda^2 R^2} |\int U\, dA|^2 \qquad (1)$$

where dA is an element of the surface and the integral is over the whole surface. The amplitude U is a complex vector quantity and $|u|^2$ will be reduced if the phase and polarization are not constant over the surface. If the antenna were isotropic the intensity at the same far field point would be

$$|u|^2 = \frac{P}{4\pi R^2} \qquad (2)$$

where P is the radiated power. Dividing equation (1) by (2) we find the antenna gain to be

$$g = \frac{4\pi}{\lambda^2} \frac{|\int U\, dA|^2}{P} \tag{3}$$

The effective area of any antenna is $\lambda^2 g/4\pi$, so from equation (3) we have

$$A_{eff} = \frac{|\int U\, dA|^2}{P} \tag{4}$$

Obviously, the gain and effective area are reduced if any of the radiated power P spills past the dish, for then P is greater than necessary to produce a given $U(\rho,\phi)$. If there is no spillover, then all the power is reflected by the dish so

$$P = \int |U|^2\, dA \tag{5}$$

and

$$A_{eff} = \frac{|\int U\, dA|^2}{\int |U|^2\, dA} \tag{6}$$

From equation (6) we see that A_{eff} is greatest and equal to the physical projected area A if U is constant, since of all functions a constant has the smallest mean square value for a given mean value. For the greatest effective area the feed should illuminate the dish uniformly. For any illumination pattern having no spillover the aperture efficiency is

$$\eta \equiv \frac{A_{eff}}{A} = \frac{|\int U dA|^2}{[\int dA]\ [\int |U|^2\, dA]} \tag{7}$$

A convenient figure of merit for a feed is the ratio of the aperture efficiency to the total system noise temperature, T, of which the antenna noise temperature, T_a, is a significant part. The antenna noise temperature T_a, given by equation (21) of Chap. 5, will be greater than the sky temperature to the extent that the antenna can receive radiation from the (hot) ground. The amount of the radiation received from the ground when the antenna is pointed at the zenith is proportional to the amount of radiation from the feed horn that spills past the dish in the transmission mode. Thus,

$$T_a = (1 - \alpha_1)\, T_S + \alpha_1 T_0 \tag{8}$$

where T_S is the sky temperature, T_0 is the ground temperature, and α_1 is the fraction of the total power represented by spillover. Solving for α_1 we find

$$\alpha_1 = \frac{T_a - T_S}{T_0 - T_S} \tag{9}$$

If $T_S = 4^\circ$ K and $T_0 = 300^\circ$ K and we wish $T_a \leqslant 7^\circ$ then

$$\alpha_1 \leqslant \frac{1}{102} \approx 1\%$$

Obviously a very small amount of spillover can raise the noise temperature significantly.

The goal of the feed designer therefore is to design a wave-guiding structure that will accept energy in a single propagation mode (in either a coaxial or waveguide) and distribute this power so as to produce a plane reflected wave of constant polarization having as uniform an intensity as possible over the surface without spilling an appreciable amount of the radiation past the reflector. Since radiation patterns are analytic functions, they have a finite number of zeros or nulls and it is impossible to have the illumination constant over the dish and vanish identically everywhere beyond the rim.[1] Moreover, the higher the illumination near the rim the greater will be the spillover. It is therefore of interest to evaluate the effect of nonuniform illumination on η.

A convenient family of illumination patterns are the so-called Sonine distributions

$$U(r) = \frac{\nu}{\pi a^2} \left(1 - \frac{r^2}{a^2}\right)^{\nu - 1} \tag{10}$$

which have the radiation patterns

$$f(\theta) = \nu! \left(\frac{\pi a}{\lambda}\theta\right)^{-\nu} J_\nu \left(\frac{2\pi a}{\lambda}\theta\right) \tag{11}$$

[1] The closest approach is to carry the feed horn clear to the dish surface as in the horn-reflector antenna, but even here there is some spill around the necessary aperture.

where θ is the angle off axis, and J_ν is the νth order Bessel function. Now $f(0) = \int U\,dA$ and is unity for all ν. If $\nu = 1$, we have a constant illumination over the circular aperture. If $\nu = 3/2$ the amplitude has a hemispherical distribution, while the intensity $|U|^2$ is a paraboloid. If $\nu = 1$ the amplitude has a paraboloidal distribution. Several distributions of this family are shown in Figure 9-3 and the corresponding radiation patterns are shown in Figure 9-4. For all $\nu > 1$ the amplitude and intensity fall to zero at the rim $r = a$. As $\nu \to \infty$, the distributions approach a gaussian form.

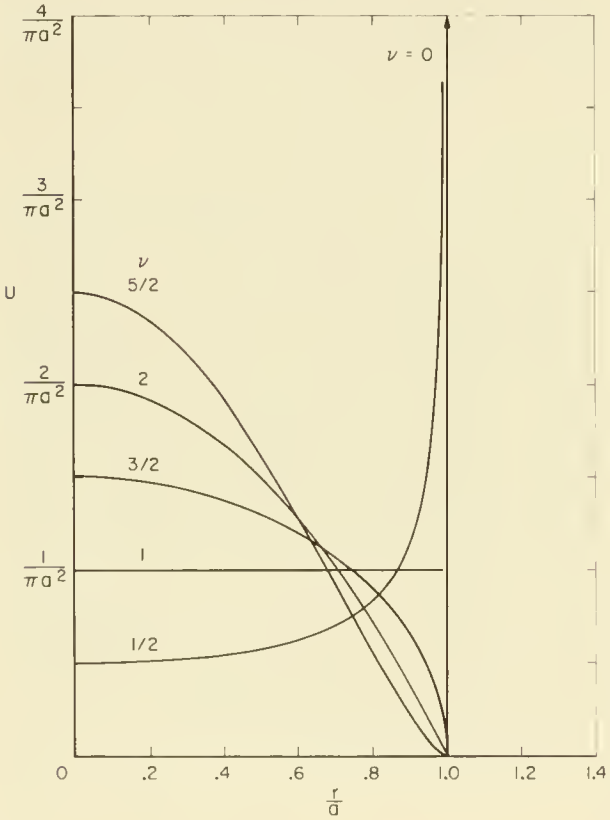


Figure 9-3. *The Sonine distributions.*

Applying equation (7) to these distributions we find that

$$\eta = \frac{2\nu - 1}{\nu^2} \qquad (12)$$

which gives the following values:

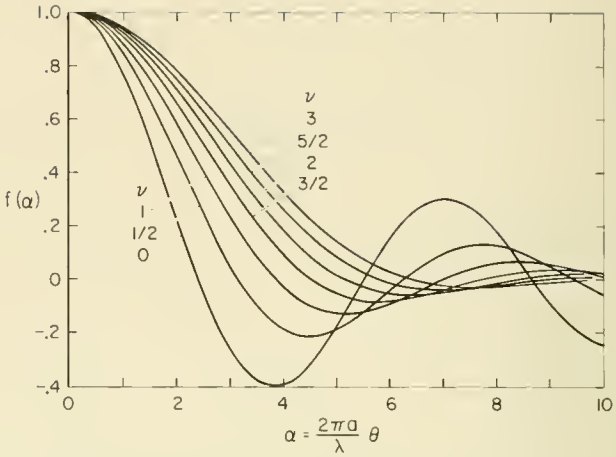


Figure 9-4. *Sonine distribution radiation patterns.*

Amplitude distribution	Intensity distribution	ν	η
Uniform	Uniform	1	1
Spherical	Parabolic	3/2	8/9
Parabolic	Bell-shaped	2	3/4

We see no fundamental reason why, in large dishes (50 to 100 m), distributions at least as uniform as the case $\nu = 3/2$ cannot be obtained, and with enough effort we feel that illumination efficiencies on the order 90% are realizable.

The simplest way to illuminate a large paraboloid is with a feed horn at the prime focus, and many radio astronomy antennas have prime focus feeds. More recently, Cassegrainian and Gregorian systems involving a secondary reflector have come into use. The secondary reflector is typically some 10% to 20% of the diameter of the primary mirror and thus shadows some 1% to 4% of the collecting area. Also, just as in optical telescopes, the hollow pupil produced by the central stop has somewhat greater near-in side lobes. Finally, the secondary mirror represents some additional expense. For Cyclops these disadvantages are more than offset by several important advantages.

First, the cluster of feed horns and receivers can be located in the shadow of the secondary mirror near the vertex of the main dish. This eliminates the shadowing and scattering that these horns and a prime focus receiver house would produce, allows convenient access to the receiver house, and greatly reduces the cabling and piping costs. Figure 9-5 shows various possible arrangements for clusters of feed horns disposed in the shadow of the secondary mirror. Off-axis feeds are obtained by tilting the secondary mirror.

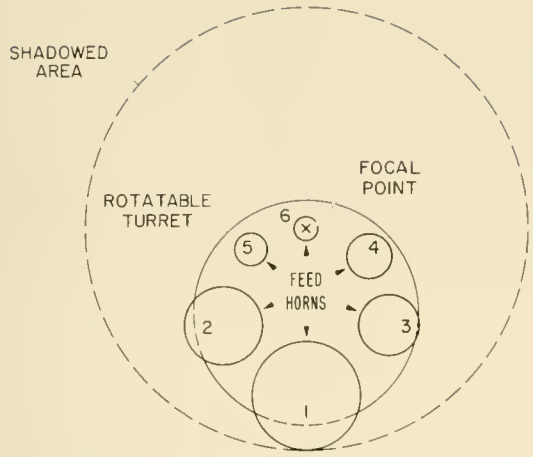


Figure 9-5a. *Rotating turret for axial feeds.*

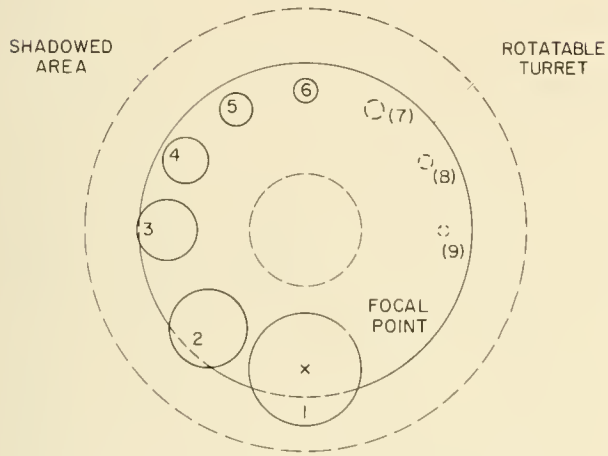


Figure 9-5b. *Rotating turret for off-axis feeds.*

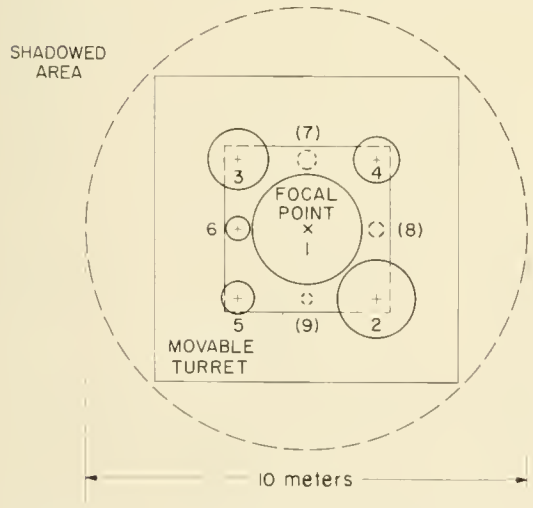


Figure 9-5c. *XY displaceable receiver turret for on-axis feeds.*

Second, large feed horns that illuminate the secondary rather uniformly with little spill can be used without increasing the total shadowing. These large feed horns have lower response at large off-axis angles than the smaller horns that would be needed at the prime focus to cover the wide subtended angles of the primary mirror. This reduces the sensitivity of the system to local interference picked up directly by the feed.

Third, the feed-horn spillover at the secondary is directed at the sky so radiation received past the rim of the secondary comes from the sky rather than from the (hot) ground. Feed-horn spillover thus causes far less elevation of the antenna temperature than in a prime focus feed.

Fourth, in a large antenna, the secondary can be many many wavelengths in diameter and can thus produce a sharper edged illumination pattern on the primary mirror than a small horn. This allows the ground spill to be reduced or a more uniform primary illumination to be realized, or both.

Finally, a Cassegrainian system having a magnification m allows an f/d ratio $1/m$ times as large to be used for the main reflector for a given feed-horn pattern. (See Appendix I.) This larger ratio greatly reduces the length of the legs of the supporting tripod or tetrapod, which in turn permits a smaller cross section with reduced shadowing. In fact, with a primary f/d ratio of 1/4 or less, one might consider supporting the secondary mirror with tension members only.

If an isotropic radiator is placed at the prime focus of a paraboloid or at the secondary focus in a Cassegrainian system, the intensity of illumination of the primary reflector falls off as

$$\frac{|U|^2}{|U_0|^2} = \frac{I}{I_0} = \cos^4 \frac{\theta}{2} \tag{13}$$

where θ is the angle subtended at the feed horn by the ray considered. This ray strikes the primary reflector at a radius

$$\rho = 2mf \tan \frac{\theta}{2} \tag{14}$$

where f is the focal length of the main reflector and m is the magnification of the secondary mirror (Appendix I). If θ_0 is the angle that the marginal rays subtend at the feed horn, we see from equation (14) that

$$\tan \frac{\theta_0}{2} = \frac{d}{4mf} \tag{15}$$

We also find from equation (7) that

$$\eta = 2\left[\frac{\log\cos(\theta_0/2)}{\sin(\theta_0/2)\tan(\theta_0/2)}\right]^2 \qquad (16)$$

Figure 9-6 shows a plot of I/I_0 vs. θ and of mf/d and η vs. θ_0. We see that even for values of θ_0 as high as 70°, where the illumination intensity at the rim has fallen to 45% of its central value, the efficiency is still almost 99%. If the f/d ratio of the primary mirror is 0.25 (which places the focus in the plane of the rim), and the magnification $m = 3$, then $mf/d = 0.75$ and at the Cassegrain focus $\theta_0 = 37°$. At this angle $I/I_0 = 0.81$ and $\eta \approx 1$.

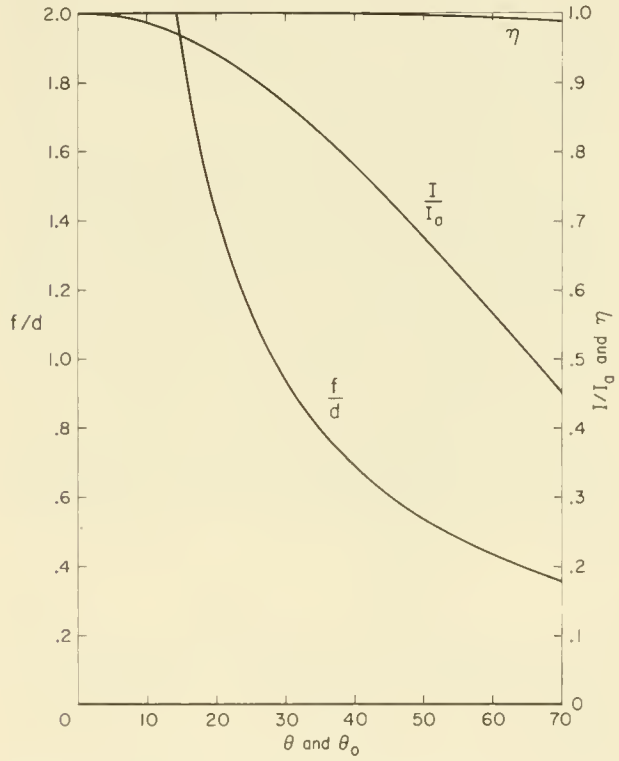


Figure 9-6. *Illumination of a paraboloid by an isotropic feed.*

We conclude that with a properly designed Cassegrainian system there is very little reason to attempt to increase the off-axis radiation of the feed to compensate for the illumination falloff given by equation (13). It is far more important to achieve as abrupt a decrease in illumination at the rim and as little spillover as possible.

There appear to be two distinct approaches that can be used. The first is to make a feed that, in the transmission mode, generates (with reversed E × H) the diffraction pattern produced in the image plane by an infinitely distant point source. This is illustrated schematically by feed horn *A* in Figure 9-7. The second is to make a large feed horn that, in the transmission mode, generates (with reversed E × H) the spherical wavefront produced by the same distant source many wavelengths "upstream" from the focal plane. This is illustrated by feed horn *B* in Figure 9-7.

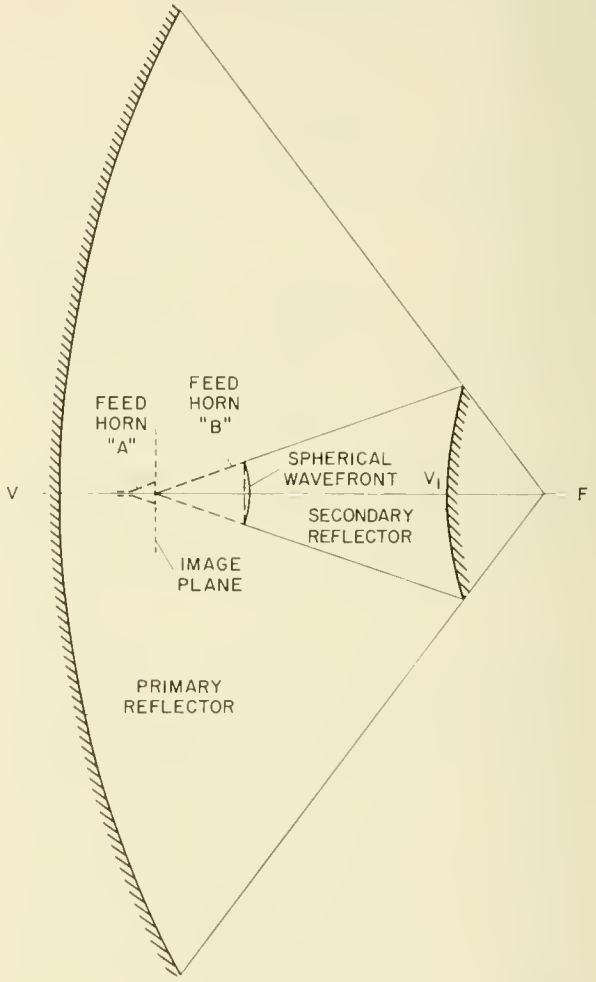


Figure 9-7. *Two methods of feeding a Cassegrainian antenna.*

Minnett and Thomas (ref. 1) have investigated the fields in the vicinity of the focal plane, and have shown that they are composed of HE_{1n} balanced hybrid modes. Such modes can be propagated in waveguides and horns whose inner surfaces are corrugated—that is, carry circumferential grooves having a spacing less than about $\lambda/10$ and a depth of $\lambda/4$ at the center of the band.

Corrugated horns supporting balanced hybrid modes have equal E-plane and H-plane patterns with zero cross polarization component and are therefore well suited for illuminating reflectors having axial symmetry (refs. 2-4). In the focal plane the field amplitude as a function of radius is very nearly given up by the classical Airy distribution

$$u(\rho) = u_0 \frac{2J_1(2\pi\theta_0\rho/\lambda)}{2\pi\theta_0\rho/\lambda} \qquad (17)$$

for values of the (maximum) convergence angle θ_0 from zero to 30°. For $\theta_0 > 30°$ the energy in the rings increases at the expense of energy in the central spot. In addition, for $\theta_0 \gtrsim 60°$ regions of reversed energy flow appear in what, for lower values of θ_0, were the dark rings. Thus for wide angle feeds the focal plane matching must include several rings to obtain high efficiencies.

Thomas (refs. 5, 6) has shown that it is possible to match the focal plane fields of a reflector having $\theta_0 = 63°$, and to obtain efficiencies within 0.1% of the theoretical values of 72.4%, 82.8%, 87.5%, 90.1% and 91.9% for feed diameters capable of supporting one to five hybrid modes, respectively. The match requires that the relative amplitudes of the modes of different orders be correct. Mode conversion can be accomplished in various ways such as by irises or steps in the waveguide or horn. Whether the ratios between the mode amplitudes can be made to vary in the appropriate fashion to preserve the distribution of equation (16) as λ varies is another question. The efficiencies cited above assume that the guide radius at the mouth equals the radius of the first, second, third, fourth or fifth null of equation (16), a condition that is only possible at discrete frequencies. Thus, although bandwidth ratios of 1.5 to 1 have been reported (ref. 7) for single mode horns, it is not clear that multimode horns designed to match focal plane fields over several rings of the diffraction pattern can achieve high performance over such bandwidths.

For broadband operation the second approach of generating a spherical cap of radiation several wavelengths in diameter to match the field some distance in front of the focal plane appears more promising. Higher order modes are involved here also, but since their role now is simply to maintain the field at a nearly constant value over the wavefront at the mouth of the horn, the higher order mode amplitudes are less than in the focal plane horn where field *reversals* are required. Thus the mode conversion process is less critical and might take the form of a dielectric cone lining the horn as shown purely schematically in Figure 9-8. The dielectric slows the wave off axis, increases the convexity of the wavefront at the axis and thus serves to guide the energy flow away from the axis. Since all transitions can be many wavelengths long, very low standing wave ratios should be expected. The major problem with such a horn might be the loss of the dielectric and the consequent elevation of the noise temperature. If this proves to be the case, artifical dielectrics might be preferable.

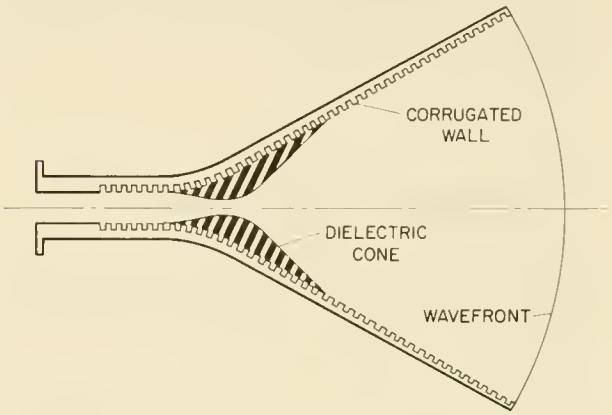


Figure 9-8. *Dielectric loaded corrugated feed horn.*

With this type of feed horn, the spherical cap of radiation illuminates the secondary mirror in the same way that the spherical cap of radiation generated by the secondary illuminates the primary mirror, except that for the feed horn the cap dimensions are less and the secondary spillover is therefore greater for a given illumination at the rim. However, the secondary spillover represents side lobe response aimed at the sky, so the effect on noise temperature is negligible.

The reradiation reflected in the transmission mode by the secondary onto the primary in the shadow region of the secondary is reflected off the primary as a parallel wavefront that is intercepted by and re-reflected by the secondary. After this second reflection off the secondary, most of this radiation is reflected by the primary to a distant focus on the beam axis after which the radiation diverges. The net result is a general rise in the side lobe level at modest off axis angles. By applying spherical wave theory to the design of Cassegrainian systems, Potter (ref. 8) has shown that it is possible to shape the secondary mirror near the vertex so that the radiation that would normally be reflected into the shadow region is redirected into the unshadowed region where it combines constructively with the rest of the radiation. This reduces the shadowing loss and improves the side lobe pattern.

Through the use of the techniques described above, we are confident that illumination efficiencies on the order of 90% and overall aperture efficiencies (including shadowing by the secondary and its supports) in excess of 80% can be realized over a series of 1.35 to 1 (or wider) bands. Because an improvement of only 1% in feed efficiency is equivalent to many millions of dollars of antenna area a major engineering design effort aimed at improving feed efficiency is justified for the Cyclops system, and is strongly recommended. Improved performance in radar and satellite tracking systems would be an important byproduct.

THE SIGNAL CONVERSION SYSTEM

The ultimate sensitivity of the Cyclops array is directly proportional to antenna area and inversely proportional to the system noise temperature. Halving the noise temperature is equivalent to doubling the collecting area. Within rather definite limits, set by our technology, it is far more economical to increase the system sensitivity by reducing the noise temperature than by increasing the area. The lowest noise receivers require cyrogenic cooling and yield total system noise temperatures that are about four to five times the sky temperature in the microwave window. Future advances in technology could thus conceivably double the effective diameter of the Cyclops array.

The major sources of noise are the sky itself, ground spillover, waveguide loss and RF amplifier noise. If T_a is the noise temperature measured at the antenna feed, α_2 is the fraction of the power lost to the waveguide walls (at a physical temperature T_0) and T_r is the noise temperature of the RF amplifier, the system noise temperature T' is given by

$$T' = T_a + \frac{\alpha_2}{1-\alpha_2} T_0 + \frac{T_r}{1-\alpha_2}$$

$$= (1-\alpha_1) T_S + \alpha_1 T_0 + \frac{\alpha_2}{1-\alpha_2} T_0 + \frac{T_r}{1-\alpha_2} \qquad (18)$$

provided we include the power lost through spillover as part of the antenna aperture efficiency η. If we do not include this loss in the aperture efficiency we can use a corrected system temperature

$$T = T_S + \frac{\alpha_1}{1-\alpha_1} T_0 + \frac{\alpha_2 T_0}{(1-\alpha_1)(1-\alpha_2)} + \frac{T_r}{(1-\alpha_1)(1-\alpha_2)}$$

$$= T_S + \frac{\alpha}{1-\alpha} T_0 + \frac{T_r}{1-\alpha} \qquad (19)$$

where

$$\alpha = \alpha_1 + \alpha_2 - \alpha_1 \alpha_2 \qquad (20)$$

If the antenna ground spillover is 1% then $\alpha_1 = 0.01$. If the waveguide loss is 0.1 dB (corresponding to about 5m of guide at 3 GHz) $\alpha_2 = 0.023$ and $\alpha = 0.033$. Taking $T_0 = 300^\circ$ K we find from equation (19) that the noise temperature contributed by the losses is

$$T_\ell = (\alpha/1-\alpha)\, T_0 \approx 10^\circ \text{ K} \qquad (21)$$

Since this is 2-1/2 times the sky temperature, T_S, the importance of minimizing spillover and waveguide losses is clearly evident.

The lowest noise receiver yet developed is the helium-cooled maser (ref. 9). One solution for the Cyclops receiver would be to associate two masers (one for each polarization) with each feed horn. However, the bandwidth achievable with masers is only about 150 MHz between 1 dB points, so to cover the band from 500 MHz to 3 GHz we would need 16 feed horns and 32 masers. The masers operate at liquid helium temperatures (~4° K) and if rapid band switching is required the refrigerator power requirements become exhorbitant.

Another low noise amplifier is the parametric up-converter. Theoretically, the up-converter can be noiseless; it has no spontaneous emission noise and achieves its power gain by converting the input photons to higher frequency photons on a one-for-one basis (ref. 10). However, to get the signal back down in frequency without adding excessive noise, the up-converter must be followed by a low noise amplifier operating at the high frequency. Figure 9-9 compares the noise performance of masers and of up-converters followed by a maser or parametric amplifier (refs. 11, 12). We see that over the band of interest the up-converter followed by a maser is only slightly noisier than the maser alone. To achieve this performance the up-converter must be cooled, but only to 20° K rather than 4° K.

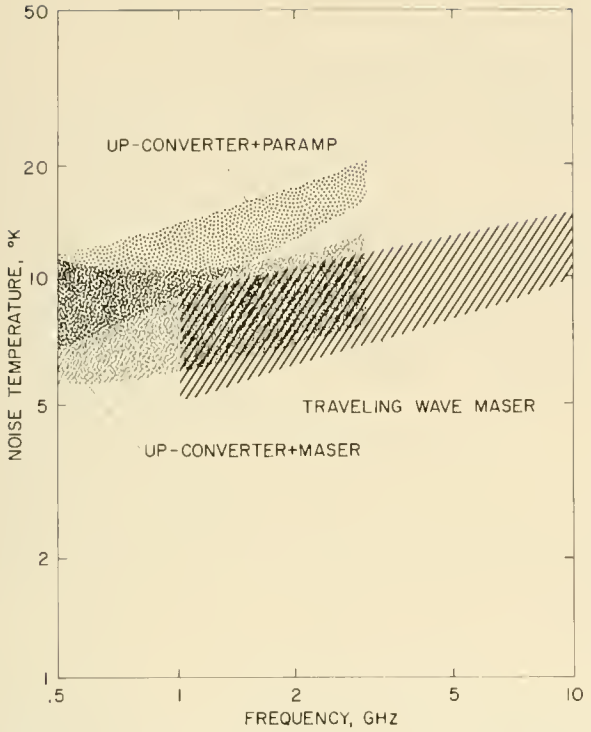


Figure 9-9. *Up-converters and maser noise temperature.*

The up-converter can have an instantaneous bandwidth of 10% and in addition is tunable over a 30% band. Thus only six feed horns and 12 up-converters are needed to cover the 0.5 to 3 GHz range. When the up-converters are driven with the appropriate pump frequencies, the outputs of all bands are in the same high frequency band and can be selected by waveguide switches as the input to a single high frequency maser (for each polarization) as shown in Figure 9-10.

As shown, the six up-converters (three for each polarization) associated with three bands are clustered in a 10 W capacity, 20° K closed cycle cryostat, while the other six up-converters are in a second cryostat. If a larger unit is used and if the waveguide runs, particularly for the higher bands, can be kept short enough, then all 12 up-converters and the 10-GHz masers can be enclosed in this single unit. Because the 10-GHz masers are relatively small in size, and can be thermally guarded in a 20° K ambient, the 4° K cooling load is much less than for an all maser system.

In addition to having low RF loss, it is imperative that the waveguides place as small a thermal load as possible (< 0.2 W per input line) on the 20° K cooling system. Refrigerators with simultaneous cooling capacity at 20° K and 4° K are not presently available. Thus the masers may require a separate cryostat with about 1-W capacity at 4° K.

A water-cooled helium compressor with output at 500 psi will be needed to run the refrigerators. A small additional compressor may be needed to operate the Joule-Thompson circuit in the 4° K units. These compressors may be located at the base of each antenna. The supply and return lines carry helium at near-ambient temperature. These stainless steel lines would require helical spring sections to negotiate the telescope axes.

Rather than having compressors at each antenna, a single large unit might serve a group of four to six antennas. The economics of this have not been studied. A single central unit is probably undesirable because of the large piping cost and because a failure of this central unit would shut down the entire array. Even with individual units, high reliability and freedom from maintenance is imperative.

Following the masers the signals are further amplified before final down conversion to the intermediate frequency. In Figure 9-10 this further amplification is

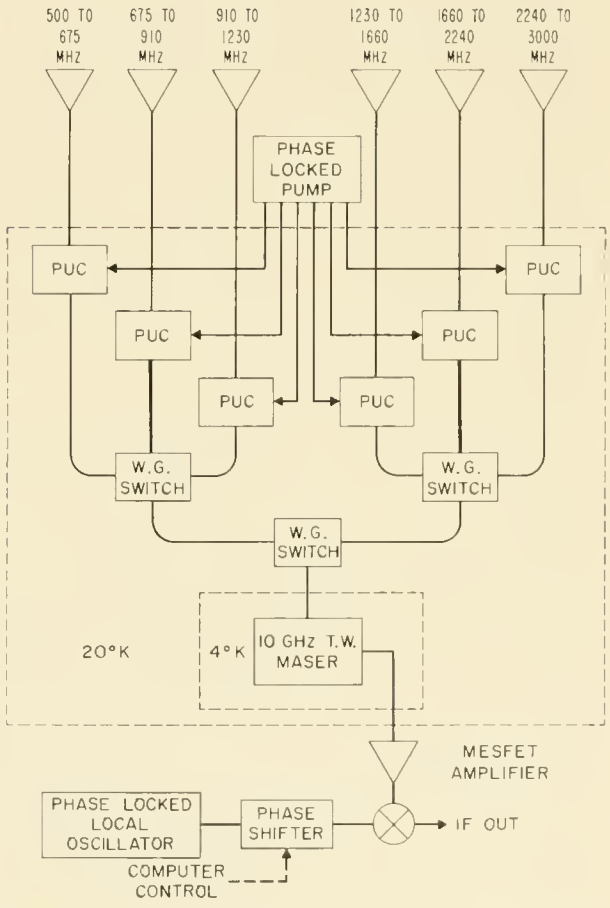


Figure 9-10. *Receiver block diagram (one polarization).*

shown as a room temperature MESFET (Metal gate Schottky barrier field-effect transistor) amplifier. Alternatively, parametric down-converters followed by parametric amplifiers all operating in the 20° K environment might offer better gain stability and good noise performance.

A fixed frequency local oscillator signal synthesized from the standard reference frequencies is needed for the down conversion. Ferrite phase shifters programmed by the central computer are used to shift the phase of this local oscillator at each antenna to provide beam steering and tracking capability. Further discussion of the phase shifter is given in Appendix J.

Using the receiver system described above, the estimated system noise temperature would vary from about 40° K at 500 MHz to about 24° K at 3 GHz. The receiver noise and waveguide loss decrease with frequency but the spillover loss and sky temperature increase at lower frequencies. The minimum noise temperature is expected to occur around 1.5 GHz and to be about 20° K.

THE LOCAL OSCILLATOR SYSTEM

The three major functional components of the Cyclops local oscillator system shown in Figure 9-1 are the primary frequency standard, the distribution system, and the frequency synthesis and phasing systems located at each antenna. In principle, one could generate centrally the local oscillator frequencies needed for reception of a particular band rather than synthesizing them at each antenna. However, it does not appear practical to transmit frequencies over a large frequency range to the individual antennas and maintain the phase tolerances required. For this reason only two precise frequencies are generated centrally and distributed, and the synthesis is done locally at each antenna.

In a system the size of Cyclops with its thousand or more antenna elements, the cost of any item associated with each antenna must be kept low, and the cost of transmission circuits is quite appreciable. However, the cost of a single item such as the central frequency standard is an insignificant part of the total system cost, and to the extent that increased expense in this item will improve system performance it is easily justifiable. The spectral purity of the central frequency standard limits the minimum bandwidth that can be used in the signal processing, which in turn is directly related to the minimum detectable signal strength. Since the detection of very weak signals is the whole object of building so large an array, only frequency standards of the highest possible spectral purity and stability should be considered. Also, if very long base-line interferometry is to be undertaken using the Cyclops array, good frequency stability is essential.

The best frequency standard available today is the hydrogen maser phase locked to a high quality crystal oscillator (refs. 13-15). Table 9-1 is a comparison of several state-of-the-art frequency standards. Figure 9-11 shows the spectral purity obtainable from the major kinds of frequency standards. The ordinate is the rms frequency deviation divided by the operating frequency, while the abscissa is the averaging time over which the measurement of Δf_{rms} is made. The hydrogen maser gives the best performance for all averaging times and, for times on the order of 100 seconds, which are involved in the Cyclops signal processing, the hydrogen maser is some 30 times better than a rubidium maser and some 500 times better than cesium. Hydrogen masers are not commercially available at present, but several have been built by Varian Associates, by the Smithsonian Astrophysical Observatory, and by others.

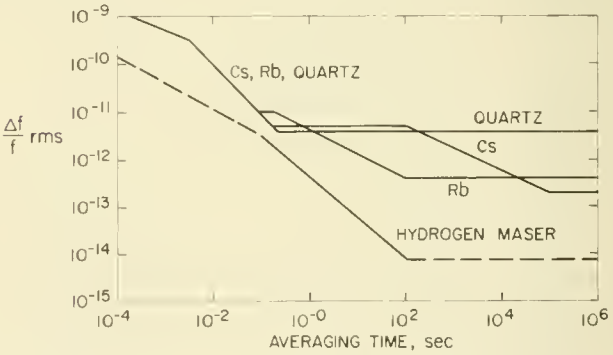


Figure 9-11. *Fractional frequency fluctuation versus averaging time.*

Since the reliability of the hydrogen maser, as with any other device, is not perfect, and since the entire Cyclops array would be inoperative if the frequency standard were to fail, at least two hydrogen masers should be provided. In fact, if two standards disagree, a third is needed to determine which of the two is in trouble, so it is recommended that the control center include three hydrogen masers.

The standard frequency distribution system must distribute the reference frequencies required to synthesize all local oscillator and pilot frequencies at each antenna site with an unambiguous phase relationship. This requires that *one of the standard frequencies be equal to the smallest desired tuning increment*, for the following reason. The usual frequency synthesizer contains frequency dividers as well as multipliers and mixers. Any frequency divider that divides a reference

TABLE 9-1

A COMPARISON OF VARIOUS PRECISION FREQUENCY SOURCES

Type of signal source	Hydrogen maser plus phase-locked oscillator	Passive Cs Atomic Beam	Passive Rb vapor cell	5-MHz crystal oscillator
Long-term stability (magnitude of drift)	Not detectable (less than two parts in 10^{12} per year)	Not detectable (less than two parts in 10^{12} per year)	Less than two parts in 10^{11} per month	Less than five parts in 10^{10} per day
Flicker level of fractional frequency fluctuation	Probably less than 1×10^{-14}	Probably less than 1×10^{-13}	3.6×10^{-13}	1×10^{-12}
Weight (lb)	600	60	37	1.1
Volume (cu ft)	16.4	1.38	.93	.029
Power requirement (W)	200	43, ac or 27, dc	48, ac or 35, dc	4
Cost (dollars)	Not available commercially at present time	14,800	7,500	950
Unusual features	Best phase spectral density	High intrinsic accuracy and reproducibility; second is defined by the cesium 0.0 transition	Least expensive atomic signal source	Low cost, high reliability
Magnitude of ambient temperature coefficient	Less than one part in 10^{13} °C	Less than five parts in 10^{12} change from 25° C value for ambient temp. between 0° and 50° C	Less than five parts in 10^{11} change for ambient between 0° & 50° C	Less than 2.5 parts in 10^{9} change for ambient between 0° and 50° C
Magnitude of sensitivity to magnetic fields	Less than three parts in 10^{13} under normal laboratory conditions (±20 milligauss changes)	Less than two parts in 10^{12} for any orientation in a 2 gauss field	Less than one part in 10^{11} for a one gauss change	Not applicable
Advantages	Very good low Fourier frequency phase spectral density	High intrinsic accuracy and reproducibility, portable, relatively low power, low sensitivity to magnetic field	Low cost, small size & weight, portable, moderately good phase spectral density	Low cost, small size and weight
Disadvantages	Size, weight, magnetic field sensitivity. Not fully developed at the present time	Low Fourier frequency phase spectral density is poor	Frequency drift requires calibration	Frequency drift, requires calibration, acceleration sensitivity

frequency f by n may lock in n possible states in which the phase of the output frequency f/n differs by multiples of $2\pi/n$. This is inherent in the division process since for every cycle of the output frequency there are n *indistinguishable* cycles of the input frequency. If one is to use the synthesized frequency at each antenna as the local oscillator signal for a phased array this ambiguity cannot be tolerated. If all the synthesizers at each antenna were in phase at some output frequency f_i, they must remain in phase when switched to some new frequency f_j. If there were phase differences to begin with, these differences must scale in proportion to f_j/f_i. This means that the synthesizers must contain no dividers. Thus only multiplication and mixing processes can be used and this implies that one of the reference frequencies must be the smallest frequency increment obtainable from the synthesizer.

For the Cyclops system with its IF bandwidth of 100 MHz or more a minimum tuning increment of 50 MHz seems reasonable. This means that no phenomenon of interest need be more than 25 MHz from the center of the band. One possible design for the Cyclops synthesizer is shown in Figure 9-12. Because of the final ×2

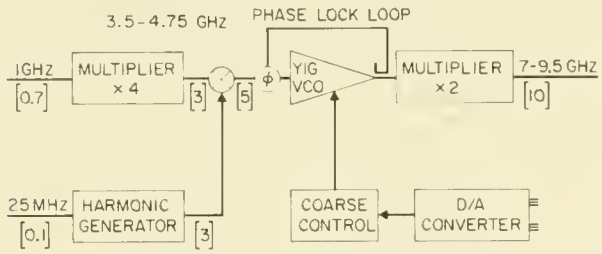


Figure 9-12. *Cyclops frequency synthesizer.*

multiplier, the lowest standard frequency is 25 MHz rather than 50 MHz. The other standard is 1 GHz. As will be described later, both these standard frequencies are obtained by transmitting signals at one fourth the final frequency over a coaxial distribution system.

The 25-MHz frequency is applied to a harmonic generator, such as a step recovery diode, to generate sharp spikes at the 25 MHz rate. The spectrum of the pulse train contains components every 25 MHz extending out to at least 750 MHz (30th harmonic). This signal is mixed with a 4-GHz carrier obtained by multiplying the 1-GHz reference signal by 4. The result is a comb of lines with 25-MHz spacing centered about 4 GHz. The dc component of the pulse train (or imperfect mixer rejection) contributes the line at exactly 4 GHz. A YIG tuned voltage-controlled oscillator is then phase locked to the desired line using a technique described by Barnum (ref. 16). The output of this oscillator is then doubled to provide the local oscillator signal for the up-converters.

A second similar fixed frequency synthesizer is used to generate the local oscillator signal to heterodyne the amplified 10-GHz signal down to the IF frequency. The two synthesizers may share some components (such as the harmonic generator) in common.

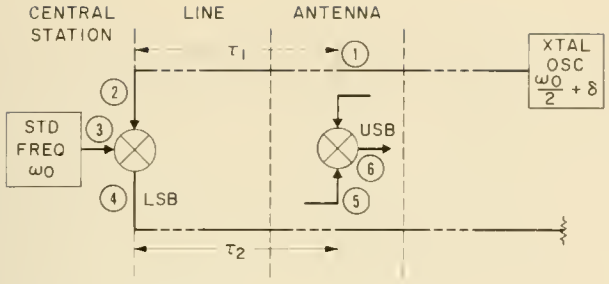
The numbers in the square brackets in Figure 9-12 are the phase stabilities required at each point to ensure no more than 10° shift at the 7- to 9.5-GHz synthesized output. Converting the permissible reference input phase shift to path length changes and assuming a 5 km maximum distance from the central distribution point to any antenna reveals that the 1-GHz path must be held to about 0.5 mm or one part in 10^7 and the 25 MHz path must be held constant to about 3 mm or six parts in 10^7. Since the line length variation over an assumed 15°C temperature range is about three parts in 10^4 both frequencies will require compensation for delay variation.

A very ingenious method for compensating for propagation delay (and therefore for delay variations) is presented in the National Radio Astronomy Observatory proposal for the Very Large Array (VLA) (ref. 17). The basic principle of this method is shown in Figure 9-13. Instead of simply transmitting the standard frequency, ω_0, from the central station to the antenna along the cable, a frequency $(\omega_0/2) + \delta$ is generated by a crystal oscillator at the far end of the line and is sent *toward* the central station, where it is mixed with the standard frequency, ω_0. The lower sideband $(\omega_0/2) - \delta$ is then returned over this same cable, to be absorbed in the far end termination. (In Fig. 9-13, the incoming and outgoing signals are shown as separate lines, but in reality they are the same line and the termination is the match provided by the output impedance of the crystal oscillator.)

At each antenna station, the two oppositely traveling waves are picked up by directional couplers and, after appropriate amplification, are fed into a mixer. The standard frequency ω_0 is recovered as the upper sideband (sum product). Taking the phase of the standard frequency at the central station to be zero, the phase of the recovered signal at the antenna is

$$\theta = \frac{\omega_0}{2}(\tau_1 - \tau_2) + \delta(\tau_1 + \tau_2) \qquad (22)$$

If the cable is dispersionless then $\tau_1 = \tau_2 = \tau$ and



POINT	FREQUENCY	PHASE
①	$\frac{\omega_0}{2} + \delta$	θ_0
②	$\frac{\omega_0}{2} + \delta$	$\theta_0 - (\frac{\omega_0}{2} + \delta)\tau_1$
③	ω_0	0
④	$\frac{\omega_0}{2} - \delta$	$-\theta_0 + (\frac{\omega_0}{2} + \delta)\tau_1$
⑤	$\frac{\omega_0}{2} - \delta$	$-\theta_0 + (\frac{\omega_0}{2} + \delta)\tau_1 - (\frac{\omega_0}{2} - \delta)\tau_2$
⑥	ω_0	$\frac{\omega_0}{2}(\tau_1 - \tau_2) + \delta(\tau_1 + \tau_2)$

IF $\tau_1 = \tau_2 = \tau$, PHASE AT ANTENNA IS $2\delta\tau$

Figure 9-13. *Standard frequency delay compensation (VLA proposal).*

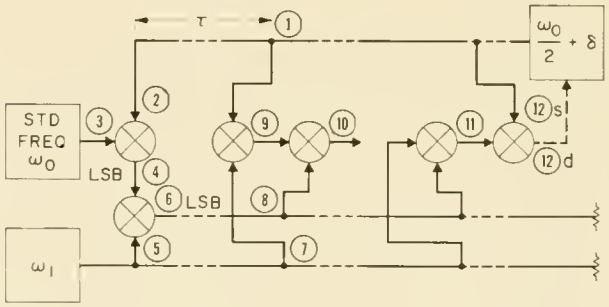
$$\theta = 2\delta\tau \qquad (23)$$

In other words the phase error is the frequency offset δ times the round trip delay time and, in principle, would be zero if δ were zero. However, if $\delta = 0$, the only means of separating the incoming and outgoing frequencies is by the directivity of the directional couplers. Since practical coupler directivities are limited to about 40 dB, a 30-dB level difference in the two signals on the line (as a result of different distances from the generators) would result in one mixer input containing the other unwanted signal only 10 dB down. Since level differences greater than 30 dB may occur, the imperfect coupler directivity seriously degrades the mixer balance.

If both signals appear at one input to the mixer, or if the mixer is not balanced, or both, then in addition to ω_0 the output will contain the frequencies $\omega_0 \pm 2\delta$. These will, in general, produce amplitude and phase modulation of the recovered output. Even if $\delta = 0$ we will simply have arrested this phase modulation at some point in the cycle and thus have a phase error.

One possible solution is to phase lock an oscillator to the recovered output and make δ greater than the bandwidth of the phase lock loop. If we choose $\delta = 2\pi \times 10^4$ (10-kHz offset), then over a 5 km path, $2\delta\tau \approx 2$ radians. The *variation* in θ will be about $3 \times 10^{-4}\ \theta$ or about 6×10^{-4} radians = 0.034°, which is quite satisfactory However, unless we use crystal filters in addition to the directional couplers to discriminate between the two mixer inputs, one of the spurious output frequencies might be stronger than the desired output. The oscillator might then lock on this frequency with disastrous results.

By any of several possible modifications of the VLA phase cancellation technique we can avoid having to have an offset δ and at the same time greatly ease the frequency selectivity needed to separate the signals along the cable. Figure 9-14 shows one alternative. As before, the incoming signal is subtracted from the standard frequency at the central station, to obtain the frequency $(\omega_0/2) - \delta$. Then, in a second mixer a new frequency ω_1 is subtracted. Both the difference frequency $(\omega_0/2) - \delta - \omega_1$ and the frequency, ω_1 itself



POINT	FREQUENCY	PHASE
①	$\frac{\omega_0}{2} + \delta$	θ_0
②	$\frac{\omega_0}{2} + \delta$	$\theta_0 - (\frac{\omega_0}{2} + \delta)\tau$
③	ω_0	θ
④	$\frac{\omega_0}{2} - \delta$	$-\theta_0 + (\frac{\omega_0}{2} + \delta)\tau$
⑤	ω_1	θ_1
⑥	$\frac{\omega_0}{2} - \delta - \omega_1$	$-\theta_0 - \theta_1 + (\frac{\omega_0}{2} + \delta)\tau$
⑦	ω_1	$\theta_1 - \omega_1\tau$
⑧	$\frac{\omega_0}{2} - \delta - \omega_1$	$-\theta_0 - \theta_1 + (2\delta + \omega_1)\tau$
⑨	$\frac{\omega_0}{2} + \delta + \omega_1$	$\theta_0 + \theta_1 - \omega_1\tau$
⑩	ω_0	$2\delta\tau$
⑪	$\frac{\omega_0}{2} - \delta$	$-\theta_0 + 2\delta\tau$
⑫ s	ω_0	$2\delta\tau$
⑫ d	2δ	$2(\theta_0 + \delta\tau)$

Figure 9-14. *Standard frequency delay compensation (Cyclops proposal I).*

are now sent out over the cable. At the antenna station the incoming frequency and the two outgoing frequencies are added to obtain ω_0. In principle, the order of addition is immaterial, but greater freedom from spurious products will obtain if the two inputs to any mixer differ considerably in frequency. Thus we can add ω_1 to $(\omega_0/2) + \delta$ and then add the sum to $(\omega_0/2) - \delta - \omega_1$ as shown, but we should not add ω_1 to $(\omega_0/2) - \delta - \omega_1$ and then add the sum, $(\omega_0/2) - \delta$ to $(\omega_0/2) + \delta$, since δ will be small, or zero.

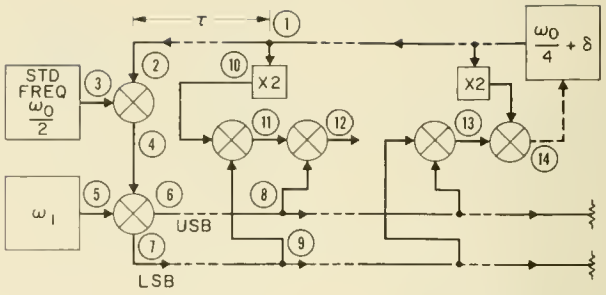
However, we can *subtract* these two inputs in a mixer as shown at the right of Figure 9-14 to obtain the frequency 2δ. If the remote oscillator is made voltage controllable, we can now apply the output of the rightmost mixer, after appropriate low-pass filtering, as the control signal to this oscillator and thus phase lock the entire system, which sets $\delta = 0$. In the absence of cable dispersion and reflections, there will be zero phase error.

If, for example, $\omega_0/2 = 500$ GHz and $\omega_1 = 200$ GHz, then $(\omega_0/2) - \omega_1 = 300$ GHz and the three frequencies on the cable are widely separated; no narrow filters are needed to isolate them. The problem is that they are *too* widely separated. The directional couplers must cover a wide band and dispersion in the cable produces a significant error.

A second alternative, which avoids these difficulties, is shown in Figure 9-15. The remote oscillator now has the frequency $(\omega_0/4) + \delta$, while the standard frequency is $(\omega_0/2)$. The difference product, $\omega_0/4 - \delta$, is then modulated with ω_1 in a balanced mixer, and both the upper and lower sidebands are returned over the cable. At the antenna station units these frequencies are added to *twice* the incoming frequency, either by using a doubler as shown, or a third mixer so that $(\omega_0/2) + \delta$ is added twice. The fundamental phase error in this system is $4\delta\tau$, or twice as great as before, but since we are setting $\delta = 0$ by phase locking the remote oscillator, this difference is immaterial.

With the phase lock in operation the incoming frequency is $\omega_0/4$ and the outgoing frequencies are $(\omega_0/4) \pm \omega_1$. We can now choose ω_1 large enough to avoid highly selective filters, yet small enough to avoid cable dispersion problems. Since the two outgoing frequencies are symmetrically disposed about the incoming frequency, cable dispersion causes only a second order effect.

In coaxials (and open wire lines) the conductor losses cause the attenuation of the line to increase as the square root of frequency. This loss causes an added phase, over and above that associated with the propagation time of a TEM wave, which amounts to one radian



POINT	FREQUENCY	PHASE
①	$\frac{\omega_0}{4} + \delta$	θ_0
②	$\frac{\omega_0}{4} + \delta$	$\theta_0 - \left(\frac{\omega_0}{4} + \delta\right)\tau$
③	$\frac{\omega_0}{2}$	0
④	$\frac{\omega_0}{4} - \delta$	$-\theta_0 + \left(\frac{\omega_0}{4} + \delta\right)\tau$
⑤	ω_1	θ_1
⑥	$\frac{\omega_0}{4} - \delta + \omega_1$	$-\theta_0 + \theta_1 + \left(\frac{\omega_0}{4} + \delta\right)\tau$
⑦	$\frac{\omega_0}{4} - \delta - \omega_1$	$-\theta_0 - \theta_1 + \left(\frac{\omega_0}{4} + \delta\right)\tau$
⑧	$\frac{\omega_0}{4} - \delta + \omega_1$	$-\theta_0 + \theta_1 + (2\delta - \omega_1)\tau$
⑨	$\frac{\omega_0}{4} - \delta - \omega_1$	$-\theta_0 - \theta_1 + (2\delta + \omega_1)\tau$
⑩	$\frac{\omega_0}{2} + 2\delta$	$2\theta_0$
⑪	$\frac{3\omega_0}{4} + \delta - \omega_1$	$\theta_0 - \theta_1 + (2\delta + \omega_1)\tau$
⑫	ω_0	$4\delta\tau$
⑬	$\frac{\omega_0}{2} - 2\delta$	$-2\theta_0 + 4\delta\tau$
⑭	4δ	$4(\theta_0 + \delta\tau)$

Figure 9-15. *Standard frequency delay compensation (Cyclops proposal II).*

per neper of attentuation. Since this phase term varies as $\omega^{1/2}$ rather than directly as ω, it does not represent a constant delay. The effect of this cable dispersion is analyzed in Appendix K, where it is shown that with $\delta = 0$, the system of Figure 9-14 will have a phase error

$$\Delta\theta_1 = \left[\frac{1}{\sqrt{2}} - \left(\frac{1}{2} - \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0 \qquad (24)$$

while that of Figure 9-15 will have an error

$$\Delta\theta_2 = \left[1 - \left(\frac{1}{4} - \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{1}{4} + \frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0 \qquad (25)$$

where N_0 is the cable loss at ω_0 *in nepers*. Figure 9-16 is a plot of $\Delta\theta_1$ and $\Delta\theta_2$ vs. ω_1/ω_0. The second system ($\Delta\theta_2$) is clearly superior, since the first requires ω_1 to be extremely small (or very near $\omega_0/2$) to avoid large phase errors from dispersion. If we choose the second system with ω_0 = 1 GHz and ω_1 = 25 MHz, the frequencies on the line will be 225 MHz, 250 MHz, and 275 MHz. These are easily picked up by a single coupler and easily separated by filters. The dispersion phase error is then 0.00125 radians/neper or about 0.72° for the 14 nepers of a 5 km run of 1-5/8 in. diameter cable. This represents 7.2° error at 10 GHz, but since the error will not vary by more than ±5%, and since fixed errors are calibrated out of the system, the performance should be excellent.

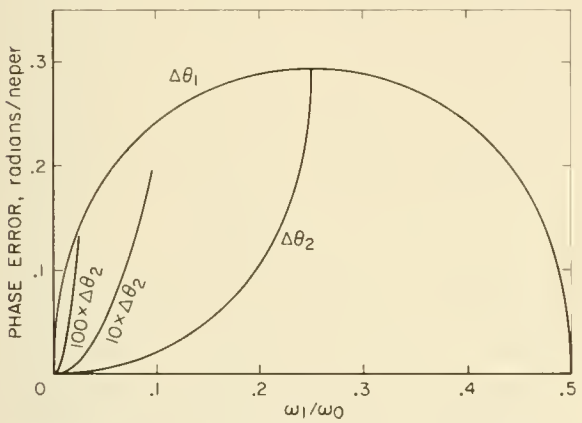


Figure 9-16. *Phase error due to cable dispersion.*

The system of Figure 9-15 has the further advantage that the signals on the line are at one quarter rather than one half the standard frequency output ω_0 They are well removed from the RF bands of the receivers. The line losses are less, so fewer repeaters are needed. Furthermore line reflections are easier to control at the lower frequency.

Although not shown in Figure 9-15, directional couplers are needed to pick the signals off the line, not merely to help in separating the signals, but to reduce reflection errors. Directional couplers not only introduce less reflection than bridging amplifiers, they discriminate against single reflections from downstream discontinuities. With perfect directivity, only double reflections can cause trouble. Single reflections must override the front to back ratio of the coupler. With coupler directivities of 40 dB and (single) reflections on the order of 3% (30 dB down, VSWR = 1.06) we can expect rms phase errors of about 4×10^{-2} radian—that is, about 0.2° at 1 GHz or 2° at 10 GHz. These phase errors may be subject to considerable variation but are tolerable in absolute magnitude.

For the reasons given above, the phase compensation method shown in Figure 9-15 is recommended for the 1-GHz standard frequency distribution system. There appears to be no reason not to use the same system for distributing the 25-MHz standard frequency. If all frequencies are scaled down by 40 to 1, the phase errors from cable dispersion are reduced by 6.3 to 1. Reflection errors should also reduce by at least the same amount. Although simple directional couplers designed for minimum forward loss at 250 MHz will have about 28 dB more forward loss at 6.25 MHz, the line loss (assuming 30 to 35 dB loss between repeaters) is also reduced by about the same amount, so the minimum signal levels obtained from the coupler should be comparable in both bands. The coupler directivity should also be comparable. Thus, it appears that we can use the same cable and couplers for both standard frequencies without resorting to a carrier system for the lower frequency.

Figure 9-17 is a block diagram of the central standard frequency unit. A majority vote of the three hydrogen masers is taken as the reference frequency of 1 GHz. This frequency is successively divided to obtain the various frequencies needed at the central station as well as the lower of the two standard frequencies to be distributed. The 1 GHz and 25 MHz standard frequencies are used to drive several modulator units each of which feeds a main trunk cable in a main tunnel.

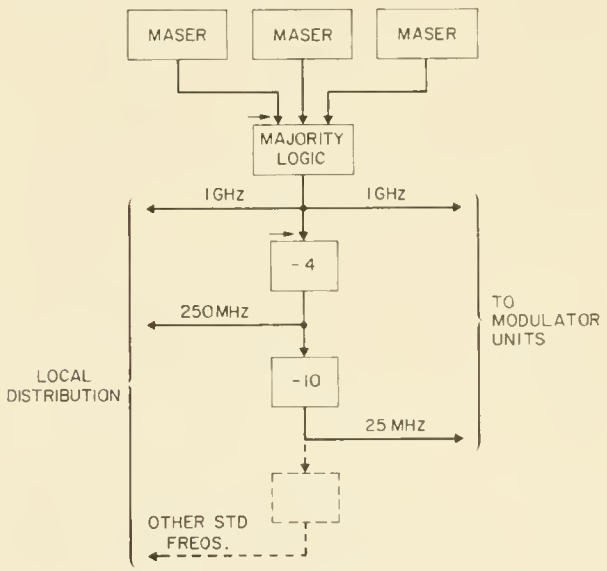


Figure 9-17. *Frequency standard.*

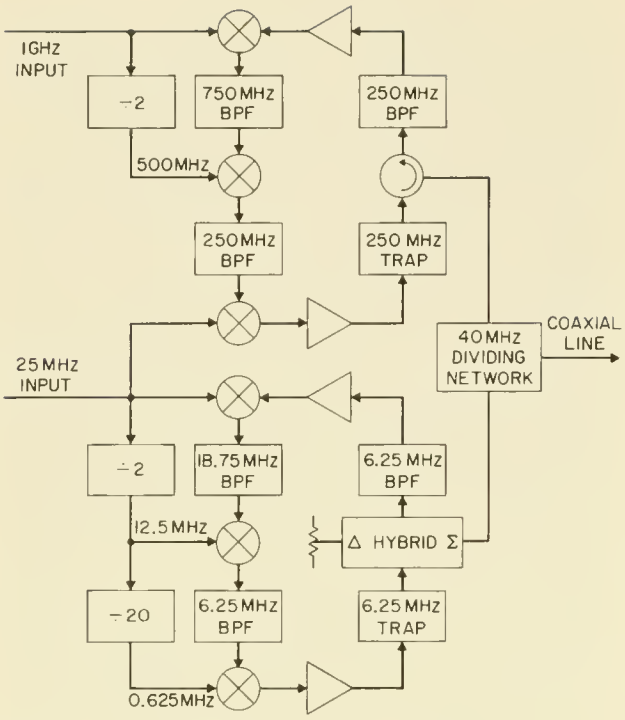


Figure 9-18. *Modulator unit.*

A modulator unit is shown in Figure 9-18. The signals received over the coaxial line are separated by a dividing network with a crossover at about 40 MHz, the 250-MHz signal going to the upper modulator and the 6.25-MHz signal going to the lower modulator. For the upper band a circulator is used to improve input-output isolation and power transfer. For the lower band a hybrid transformer is used. The received signals are selected by band pass filters and then mixed with the standard frequencies. The lower sideband is selected and mixed with one half the standard frequency. Again the lower sideband is taken. These first two mixers replace the single first mixer in Figure 9-14 and 9-15. A double mixing process is used to keep all input and output frequency bands separated, thus avoiding balance problems. Finally, the lower sideband is mixed with $\omega_1 = \omega_0/40$ and, after rejecting the carrier, the two sidebands are sent back over the cable. At the central station the standard frequencies are supplied by the central frequency standard; at branch points along the cable they are supplied by an antenna station unit, and at repeaters by a terminator unit.

Figure 9-19 shows an antenna station unit. The signals picked off the line by the directional coupler are adjusted in level and separated by filters. The incoming frequencies are doubled and added to the outgoing frequencies to recover the two standard frequencies.

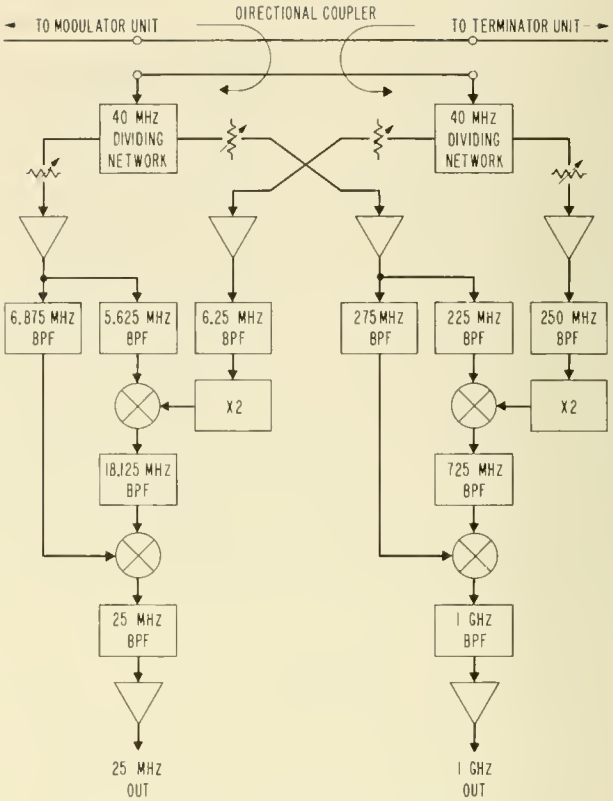


Figure 9-19. *Antenna station unit.*

A terminator unit is shown in Figure 9-20. It is similar to the antenna station unit except that, in addition to recovery of the standard frequencies, the two outgoing frequencies in each band are added and applied, along with double the incoming frequency, to a four-element mixer, which serves as a phase detector. The output of each phase detector, after appropriate low pass filtering and amplification, is applied as the control signal to a voltage controlled oscillator. This phase locks the entire system to the central standard so that the offset frequency δ is zero at all times. The signals from the two voltage controlled oscillators are combined in another 40-MHz crossover network and returned over the cable.

Figure 9-21 shows how the various units just described are used in the distribution system. The central frequency standard and its dividers drive as many modulator units as there are main or branch tunnels radiating directly from the central station. At each point along a main tunnel where two side tunnels branch, an "antenna" station unit drives two modulator units. At each repeater point in a main tunnel a terminator unit drives three modulator units; or one, if in a side tunnel. Finally at the end of all tunnels there is a terminator

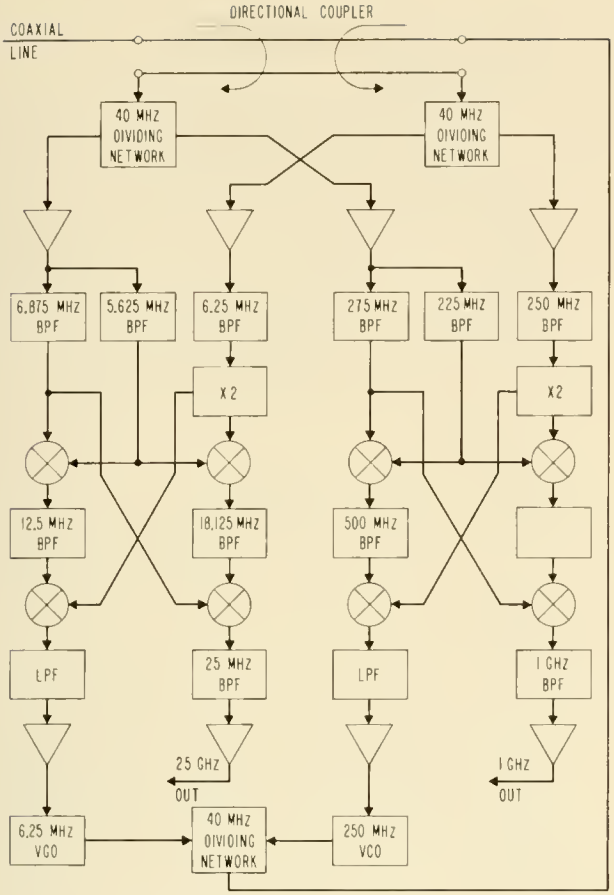


Figure 9-20. *Terminator unit.*

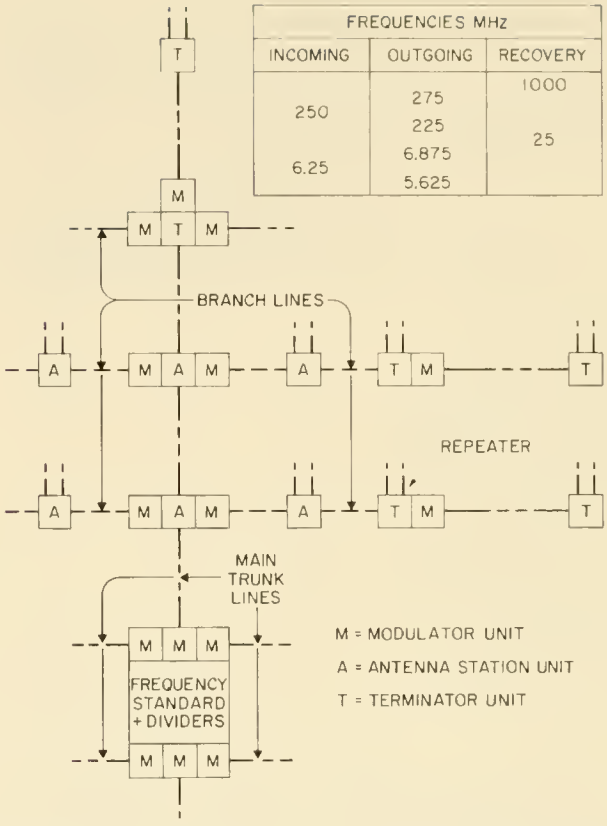


Figure 9-21. *Standard frequency distribution system.*

unit. The repeaters thus subdivide the system into a series of individually phase compensated links.

It is proposed that 1-5/8 in. diameter cable be used to distribute the standard frequencies. At 250 MHz the loss in this cable is about 12 dB/km. In a 10-km diameter array, one repeater will be needed per main tunnel, and one for each of the longer side tunnels that branch prior to the main tunnel repeater. Thus there will be no more than two repeaters between any antenna and the central station. This does *not* mean that the dispersion errors, calculated earlier, must be tripled, since these were computed for the total cable length. Reflection errors, on the other hand, may add.

It is not necessary for the modulator, antenna station and terminator units to have zero phase shift, though this could be achieved. All we ask of the distribution system is that it have *phase stability*. Fixed phase errors can and will be calibrated out by phasing the elements on a radio source. Since the same standard frequencies are used to synthesize the pump frequencies for up-conversion and the later down conversion to IF, any phase errors in the standard frequencies tend to cancel out in these two processes, leaving only the phase error *at the received frequency*. Thus a phase stability of 10° at 10 GHz implies a stability of 1.5° at 1.5 GHz. The tolerances we have assumed are probably tighter than necessary for reception up to 3 GHz.

While the Cyclops array contains many more antennas than the proposed VLA system, the overall dimensions of the array are smaller. The VLA report concluded that local oscillator phasing was feasible for the VLA, and we conclude that the same is true for Cyclops.

COST ESTIMATES

Almost all of the receiver system can be considered in terms of a cost per antenna element. All the receiver electronics and cryogenics represent a cost per antenna as do the distribution systems. The cost of the IF distribution system is included in the next chapter. The only part of the receiver system that represents a fixed cost per array is the central frequency standard and the associated electronics. The cost of this is estimated at $200,000. For a thousand-element array this is equiv-

alent to $200 per antenna and is so much less than the total uncertainty in the per site costs that we will neglect it. Table 9-2 shows the estimated costs in small quantities of the receiver components:

TABLE 9-2

Item	Number/ Site	Unit Cost (thousands of dollars)	Total Cost (thousands of dollars)
Antenna Feeds	6	10	60
Up-Converters	12	5	60
Maser Amplifier	2	50	100
Mixers, IF Amplifier	2	5	10
Phase Shifter, Driver	1	1	1
Cryogenic 20° K	2	12.5	25
Cryogenics 4° K	1	20	20
1-5/8 in. coax cable	300 m	7/m	2
LO distribution units	≈ 1	2	2
Synthesizer	1	5	5
System Integration	1	15	15
Total			300

With quantity production it is almost certain that the total receiver system cost could be reduced from $300,000 to $200,000 or less, but in any event it is obvious that the receiver cost is a small fraction of the antenna structural cost.

REFERENCES

1. Minnett, H.C.; and Thomas, B. Mac A.: Fields in the Image Space of Symmetrical Focusing Reflectors. Proc. IEE, vol. 115, 1968, pp. 1419-1430.
2. Jeuken, M.E.J.: Experimental Radiation Pattern of the Corrugated Conical-Horn Antenna With Small Flare Angle. *Electronics Letters,* vol. 5, Oct. 1969.
3. Brunstein, S.A.: A New Wideband Feed Horn With Equal E- and H-Plane Beamwidths and Suppressed Sidelobes. JPL Space Programs Summary 37-58, vol. II.
4. Clairicoats, P.J.B.: Analysis of Spherical Hybrid Modes in a Corrugated Conical Horn. *Electronics Letters,* vol. 5, no. 9, May 1969, pp. 189-190.
5. Thomas, B. Mac A.: Matching Focal-Region Fields With Hybrid Modes. IEEE Trans., AP, May 1970.
6. Thomas, B. Mac A.: Prime-Focus One- and Two-Hybrid-Mode Feeds. *Electronics Letters,* vol. 6, July 1970.
7. Thomas, B. Mac A.: Bandwidth Properties of Corrugated Conical Horns. *Electronics Letters,* vol. 5, Oct. 1969.
8. Potter, P.D.: Application of Spherical Wave Theory to Cassegrainian-Fed Paraboloids. IEE Trans., AP, Nov. 1967.
9. Siegman, A.E.: Microwave Solid-State Masers. McGraw-Hill, N.Y., 1964.
10. Manley, J.M.: and Rowe, H.E.: Some General Properties of Nonlinear Elements—Part I. General Energy Relations. Proc. IRE, vol. 44, 1956.
11. Sard, E.; Peyton, P.; and Okwitt, S.: IEE Trans. on Microwave Theory and Techniques, vol. MTT-14, Dec. 1966.
12. Smith, G.: Parametric Sun-Frequency Upconverter. *IEEE Spectrum,* Dec. 1969, p. 5.
13. Cutler, L.S.; and Vessot, R.F.C.: Present Status of Clocks and Frequency Standards. NEREM Record, 1968.
14. Baugh, R.A.; and Cutler, L.S.: Precision Frequency Sources. *Microwave J.,* June 1970, pp. 43-55.
15. Barnes, J.A.; Chi, A.R.; Cutler, L.S.; Healey, D.J.; Leeson, D.B.; McGunigal, T.E.; Muller, J.A., Jr.; Smith, W.J.; Sydnor, R.L.; Vessot, R.F.C.; and Winkler, G.M.R.: Characterization of Frequency Stability. IEEE Trans. on Instrumentation and Measurement, vol. IM-20, no. 2, May 1971, pp. 105-120.
16. Barnum, L.J.: A Multioctave Microwave Synthesizer, *Microwave J.,* Oct. 1970.
17. *A Proposal for a Very Large Array Radio Telescope.* Vol. 2, "Systems Design," National Radio Astronomy Observatory, Green Bank, West Virginia, Jan. 1967.

REFERENCES NOT CITED

A Proposal for a Very Large Array Radio Telescope. Vol. 1, "The VLA Concept," National Radio Astronomy Observatory, Green Bank, West Virginia, Jan. 1967.

A Proposal for a Very Large Array Radio Telescope. Vol. 3, National Radio Astronomy Observatory, Green Bank, West Virginia, Jan. 1969.

Beazell, H.L.: VLA Local Oscillator Distribution Systems—Summary Report. Report IRDC-4787-101-68U, Univ. of Virginia, Charlottesville, Aug. 1968.

Cuccia, C.L.; Williams, T.G.: Cobb, P.R.; Small, A.E.; and Rahilly, J.P.: *Microwaves,* vol. 6, no. 6, June 1967, p. 27.

Davis, R.T.: *Microwaves,* vol. 10, Apr. 1971, p. 32.

DeJager, J.T.: IEEE Trans. on Microwave Theory and Techniques, vol. MTT-12, Jul. 1964, p. 459.

Henock, B.T.: IRE Trans. on Microwave Theory and Techniques, vol. MTT-11, Jan. 1963, p. 62.

Kliphuis, J.; and Greene, J.C.: AIAA 3rd Communications Satellite Systems Conference, April 6-8, 1970.

Kraus, J.D.: Radio Astronomy. McGraw-Hill, N.Y., 1966, p281.

Smith, G.; and DeBruvl, J.: *Microwave J.*, Oct. 1966.

Rumsey, V.H.: Horn Patterns With Uniform Power Patterns Around Their Axes. IEEE Trans., AP, Sept. 1966.

Vu, T.B.; Vu, Q.H.: Optimum Feed for Large Radio Telescopes: Experimental Results. *Electronics Letters*, vol. 6, Mar. 19, 1970.

Zucker, H.; and Ierley, W.H.: Computer-Aided Analysis of Cassegrain Antennas. *Bell System Technical J.*, July-August 1968.

10. TRANSMISSION AND CONTROL

Several major systems are needed to tie the array and data processing equipment together into a functioning unit. These include:

1. The standard frequency distribution system
2. The power generation and distribution system
3. The IF transmission system
4. The IF delay system
5. The control, monitoring and calibration system
6. The communication system
7. The master computer control

The standard frequency distribution system is an integral part of the receiver system and is discussed in Chapter 9. The power distribution system appears to present no serious problem, and only some rough cost estimates have been made. These are included at the end of this chapter. The same is true for the master computer control. Thus the primary concern of this section is with items 3 through 6 above.

In a fixed array of definite size the distribution systems could be buried in inaccessible conduits. However, in the Cyclops system, which is visualized as expanding with time over a span of 10 to 20 years, additional feeders and transmission lines will be continually added to supply new antenna elements as they are brought on stream. Although cities endure the continual excavation and re-excavation of streets as services are expanded, a more intelligent approach, cheaper in the long run, would be to anticipate this problem by providing a set of service tunnels, which is extended as the array grows. Since all the distribution systems would use these tunnels, let us first consider the best design for them.

PROPOSED TUNNEL PATTERNS

As discussed in Chapter 7, the Cyclops array will be roughly circular in outline with antennas disposed in a square or hexagonal lattice, each antenna separated from its neighbors by a distance

$$s \geqslant \frac{d}{\sin \epsilon} \tag{1}$$

where d is the antenna diameter and ϵ is the minimum elevation angle.

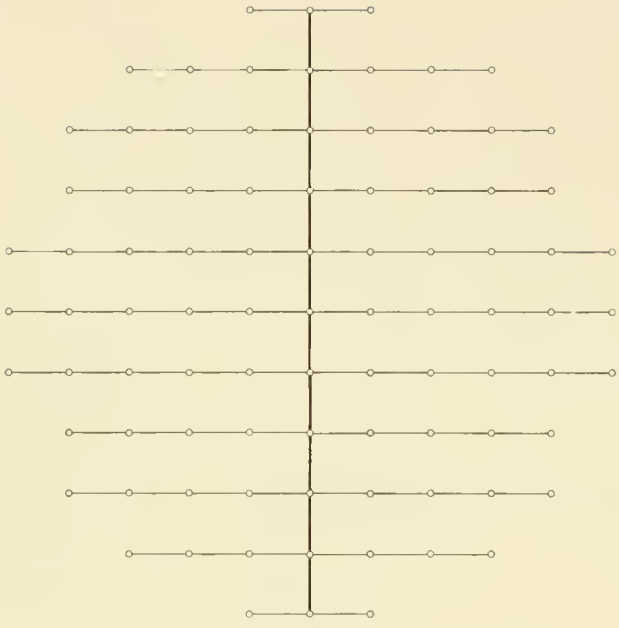
Regardless of the array or tunnel pattern, $(n - 1)$ unit tunnels between pairs of antennas are needed to connect n antennas. If one antenna in the array is replaced by the control center, then the number of unit tunnels is n. The control center can then be placed anywhere within the array without materially affecting the total tunnel length.

However, the average length of an IF cable is increased as the control is displaced from the physical center of the array. Since this factor increases the cost of the IF system, and since there appears to be no reason *not* to locate the headquarters at the array center, we will consider only symmetrical tunnel patterns.

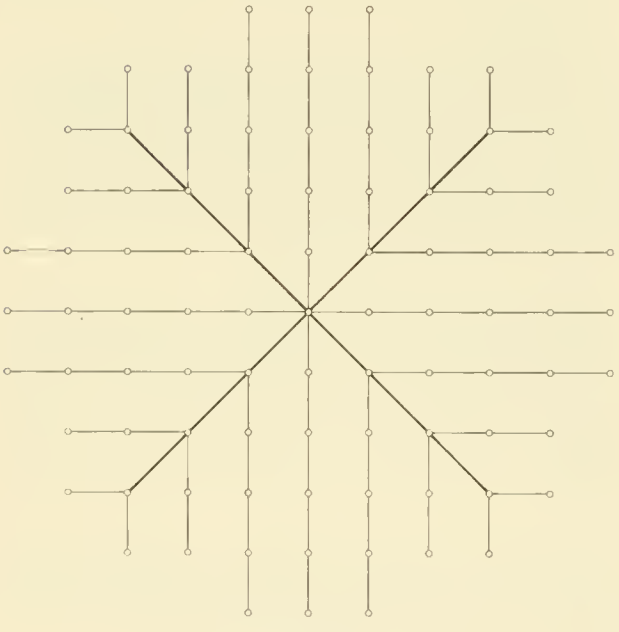
Two obvious tunnel patterns for a square lattice array and two patterns for a hexagonal lattice array are shown in Figures 10-1 and 10-2, respectively. In Appendix L the cost of tunnelling is given by

$$C = ns\gamma f$$

where n is the number of antennas, s is their spacing, γ is the cost per unit length of tunnels and f is a factor that depends on the patterns in the manner shown in Table 10-1.



Pattern (1a)



Pattern (1b)

Figure 10-1. *Tunnel patterns for square lattice arrays.*

Pattern 2*a* is the most economical and would be preferred if tunnel cost were the only cost and the only consideration. However, the differences are not great and decrease as the array size increases.

In Appendix L it is shown that the maximum distance through any of the tunnel patterns from the control center to an antenna element is

$$\ell_{max} = \frac{a}{\cos(\theta/2)} \tag{2}$$

and that the average distance is:

$$\bar{\ell} = \frac{2a}{3} \frac{\tan(\theta/2)}{(\theta/2)} \tag{3}$$

where a is the array radius and θ is half the central angle of the sector served by any main feeder tunnel; that is, $\theta = \pi/m$ where m is the number of main feeder tunnels (2,4,6) radiating from the control center. Data for the patterns shown in the figures are given in Table 10-2:

TABLE 10-1

Relative tunnel costs for Patterns 1*a*, 1*b*, 2*a*, 2*b*

n	f_{1a}	f_{1b}	f_{2a}	f_{2b}
100	1	1.050	0.987	1.120
200	1	1.041	0.989	1.090
500	1	1.026	0.993	1.060
1000	1	1.020	0.995	1.048
2000	1	1.014	0.997	1.033
∞	1	1.	1.	1.

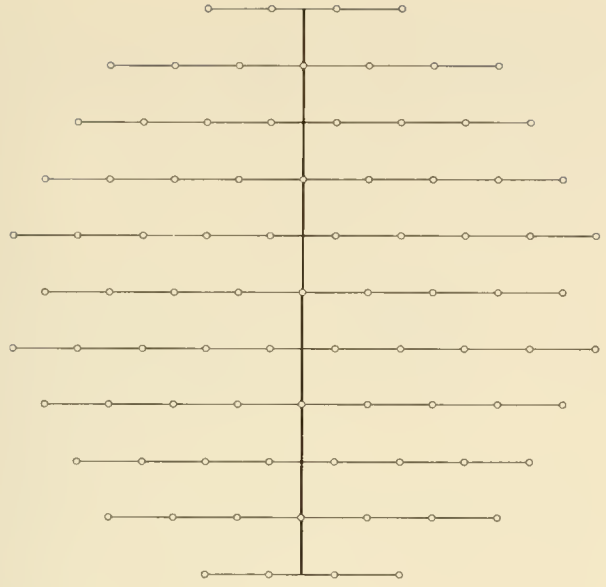
TABLE 10-2

CABLE LENGTHS FOR DIFFERENT TUNNEL PATTERNS

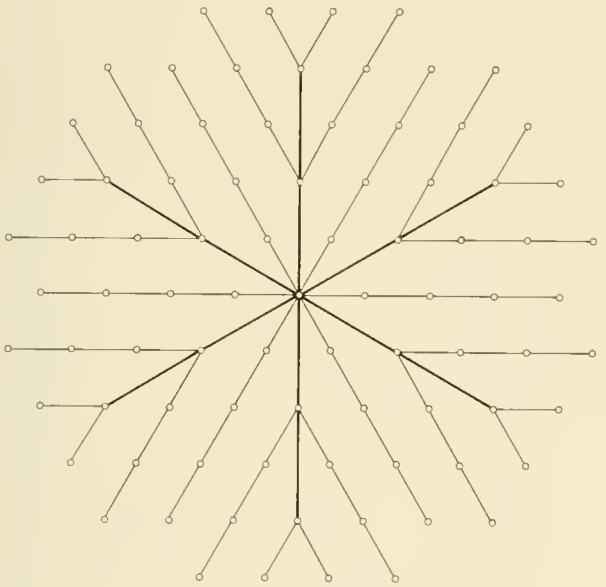
Pattern	θ	m	ℓ_{max}/a	$\bar{\ell}/a$	$3\bar{\ell}/2a$
1a,2a	$\pi/2$	2	1.4142	0.8488	1.2732
1b	$\pi/4$	4	1.0824	0.7032	1.0548
2b	$\pi/6$	6	1.0353	0.6823	1.0235
	0	∞	1.	0.6666	1.

In comparing the figures in Table 10-2 one must also remember that the array radius a is $\sqrt{\sqrt{3}/2}$ or 0.93 times as great for the hexagonal patterns as for the square patterns.

The *b* patterns also have the advantage that, since there are more main feeder tunnels, the array size can be enlarged more without increasing the main tunnel cross section over that required for the branch tunnels. Nevertheless, since most of the cables picked up by the main tunnels in the *b* patterns are collected in the first few sections of main tunnel, it might be prudent to



Pattern (2a)



Pattern (2b)

Figure 10-2. *Tunnel patterns for hexagonal lattice arrays.*

make these sections somewhat larger. In our cost estimates we assume 8-ft diameter tunnels, which would allow ample room for a man to stand, for carts to run on rails in the aisle, and for the services to be shelved at either side. However, 7-ft diameter tunnels would suffice at the periphery, where the cables are few. The saving made in this way would more than compensate for the difference among the different patterns and for the tapering of the diameter of the main tunnels from perhaps 10-ft in diameter near the control center to 8-ft farther out. We have not studied such refinements. For the present we estimate the cost of 8-ft tunnels at $570/m and, allowing 5% for the *b* patterns, we take the tunnel cost to be

$$C_T = \$600 \times n \times s \qquad (4)$$

IF TRANSMISSION SYSTEM

Several media were considered for the IF transmission system:

1. Coaxial cables
2. H_{01} mode low-loss circular waveguide
3. Free-space laser links
4. Guided laser links
5. Optical fibers
6. Microwave links

Of all these, the coaxial cable is recommended as the best alternative *at the present time.* Some of the other media hold great promise, but are not well enough developed to commit the Cyclops system design to them. Below are some of the reasons they were rejected for Cyclops.

H_{01} mode circular waveguide is very expensive per foot and depends on a high degree of multiplexing to make it cost competitive. Since this multiplexing must take place at millimeter waves it is not at all apparent that the hardware required would have the necessary gain and phase stability. In addition, only very long radius bends can be tolerated; bends such as those required at tunnel junctions would probably require demodulation and remodulation of the signal.

Free-space laser links (atmospheric) appear very questionable for paths several kilometers long near the ground. Atmospheric turbulence and thermal stratification on hot days could cause loss of signal and severe phase fluctuations. So could dust and fog. Fundamentally, Cyclops should be an all weather system in which an automated search goes on without regard for the elements.

Guided laser links—that is, laser beams confined to controlled atmosphere tubes—avoid the atmospheric problems referred to above. However, the beam cross sections needed to prevent serious diffraction losses are several centimeters in diameter and would require larger tunnels. Severe alignment problems exist for lenses and for prisms or mirrors at bends in the path.

Optical fibers are a rapidly improving means of guiding light. Although remarkably low losses (16-20

dB/km) have been reported, a great deal more work is needed before the transmission properties of optical fiber links can be specified with confidence. All optical systems share the problems of unproven hardware.

Microwave links (above ground) do not have enough directivity to avoid multipath problems, fading, and crosstalk from reflections off the antenna elements, etc. Unless operated outside the microwave window, they would constitute an intolerable source of interference.

While the above problems seemed to limit the choice of medium to coaxial lines, further developments could change the picture drastically and these (or new) media deserve review, especially after a few years.

The principal problems associated with coaxial lines are: (1) loss, (2) variation of loss with frequency, (3) variation of loss and delay with temperature, and (4) periodic variation of line constants with distance.

In telephony, coaxial lines may be a few thousand miles long and have losses of thousands of dB. The usual practice is to make up for this loss by amplifiers (repeaters) at appropriate intervals. Because of skin effect the loss in coaxial lines, expressed in dB per unit length, increases with the square root of frequency; and the usual repeater design provides a gain that varies in approximately this way also. The gain of the repeaters must vary with temperature to compensate for the loss variation with temperature of the line. These compensations are not perfect, so pilot frequencies of known amplitude are introduced at various points in the band of interest and mop-up equalizers are inserted at intervals along the line. These are servoed to restore the pilot tone amplitudes (and hence the signal) to the proper value. Delay variation with temperature is ordinarily of little concern in telephony, but is very important in Cyclops.

Periodic variations of line impedance causes stop bands at frequencies for which the period is an integral number of half wavelengths. These can be avoided by careful design of the fabrication machinery and by careful handling of the cable. For example, the cable must not be wound on a drum so as to cause a kink every turn. Filler sheets between layers should be inserted to prevent this.

Coaxial lines were studied as an IF medium for the VLA and it was concluded that 7/8-in. diameter line and 1-5/8-in. diameter line (because of its greater multiplexing capability) were about equally attractive. Experiments made at NRAO and elsewhere have demonstrated that coaxial lines buried deeply enough (below the frost level) show negligible diurnal delay variation. However, the tunnel system proposed for Cyclops would not necessarily have the temperature stability of the earth at the same depth. Variations in power dissipation in the tunnels and variations in the temperature of the ventilating air (even though conditioned) could cause cable temperature variation of a few degrees.

Assuming a 10-km diameter array (3-km clear aperture) the longest IF cables in Cyclops would be on the order of 6 km. With a 10° C variation in ambient temperature their length would vary about 1 m, or more than 1/2 wavelength at 175 MHz. Thus some delay compensation is needed. The proposed system includes such compensation as well as automatic equalization of the gain versus frequency characteristic. Except for the inverting repeaters, which can be well shielded, all frequencies lie below the RF band of 0.5 to 3 GHz and should cause no interference problems.

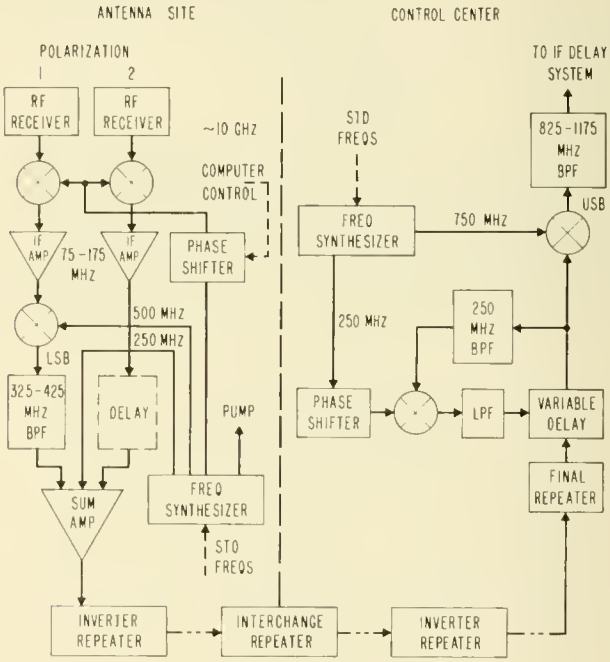


Figure 10-3. *IF transmission system.*

The proposed IF system is shown in block schematic form in Figure 10-3. At the antenna site the amplified RF signals for each polarization are mixed with a signal of approximately 10 GHz obtained from the local synthesizer to obtain two IF signals both extending from 75 to 175 MHz. A phase shifter in the 10-GHz local oscillator path accomplishes the necessary array phasing. One of the two IF channels is then mixed with a 500 MHz signal, also obtained from the synthesizer, and the lower sideband from 325 to 425 MHz is selected. To compensate for the delay of the bandpass filter, delay is

added to the other IF channel. The two IF bands, one extending from 75 to 175 MHz and the other from 325 to 425 MHz, are then combined with the 250-MHz pilot frequency for transmission over the coaxial cable.

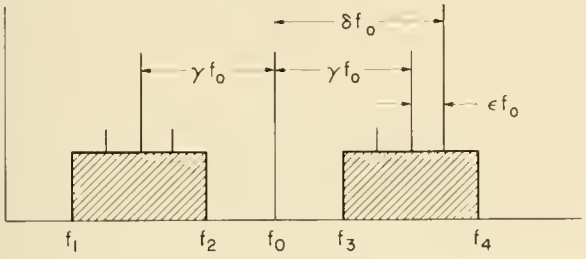


Figure 10-4. *IF transmission spectrum.*

The spectrum occupancy for the IF cable is indicated schematically in Figure 10-4. The shaded blocks extending from f_1 to f_2 and from f_3 to f_4 represent the two IF bands. Centered between these at f_0 is the 250-MHz pilot. It would certainly be possible to send these signals in the relationship shown over the entire distance; this was one of the alternatives seriously investigated. This requires the design of equalizers to compensate for the cable loss versus frequency characteristic and probably the provision of other pilot frequencies (at the outer edges of the band) to servo those equalizers as the cable temperature changes. The attenuation of 7/8-in. coaxial cable is about 24 dB/km at the 250 MHz pilot frequency, or 120 dB for a 5-km length. Since the loss varies as $f^{1/2}$ the total loss in 5 km would be 66 dB at f_1 (75 MHz), 100 dB at f_2 (175 MHz), 137 dB at f_3 (325 MHz) and 156 dB at f_4 (425 MHz). Thus about 90 dB of equalization is needed from f_1 to f_4 with about 34 dB of this occurring between f_1 and f_2 and 19 dB between f_3 and f_4. These loss figures are at room temperature and are roughly proportional to the square root of absolute temperature.

While the equalization needed is well within the state of the art, the symmetry of the Cyclops transmission problem causes us to favor another approach. The two IF channels representing orthogonal polarizations are to be combined in various ways to obtain other polarizations (Chap. 9). If a signal is received in a mixed polarization it might be transmitted as a frequency near f_1 in one IF channel and near f_4 in the other. We would like these signals to have suffered identical loss, phase shift, and delay in each channel, so that when they are demodulated to the same frequency at the central station, the polarization represented by their relative amplitudes and phases is the same as would have been determined at the antenna. This suggests that we should arrange to have each IF band experience the same cable loss and dispersion. In fact, we would like this to be true for each frequency within each IF band.

With straight through transmission the loss experienced by any frequency is

$$\alpha = \alpha_0 \left(\frac{f}{f_0}\right)^{1/2} = \alpha_0 \sqrt{1+\delta} \tag{5}$$

where $\delta = (f/f_0) - 1$, and α_0 is the cable loss at f_0. Rather than using the total loss α, we define a normalized relative loss $\rho = (\alpha/\alpha_0) - 1$. Then in terms of ρ, equation (5) becomes

$$\rho = \sqrt{1+\delta} - 1 \tag{6}$$

$$\approx \frac{\delta}{2} - \frac{\delta^2}{8} + \frac{\delta^3}{16} - \frac{5\delta^4}{128} + \cdots \tag{6a}$$

If now we arrange to invert the entire spectrum at the midpoint of the line, as indicated in Figure 10-5, so that a frequency transmitted as f for the first half is sent as $2f_0 - f$ for the second half, we will have handled both IF channels in a similar fashion. The relative normalized loss is then

$$\rho = \frac{1}{2}\left(\sqrt{1+\delta} + \sqrt{1-\delta}\right) - 1 \tag{7}$$

$$\approx -\frac{\delta^2}{8} - \frac{5\delta^4}{128} + \cdots \tag{7a}$$

which shows far less variation, and is symmetrical about f_0.

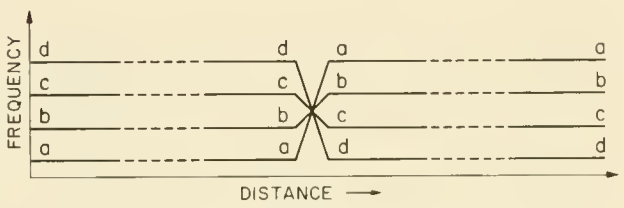


Figure 10-5. *Midpoint frequency inversion.*

Finally, a further improvement is realized if we perform this inversion twice at the one-quarter and three quarter points along the line and, at the midpoint, swap

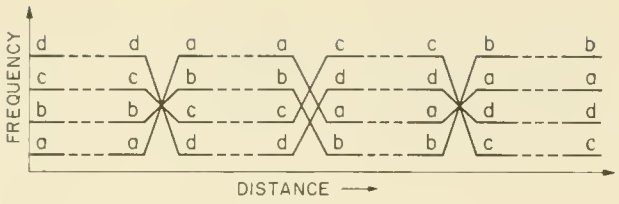


Figure 10-6. *Two inversions and one interchange.*

the two IF bands without inversion as shown in Figure 10.6. Let

$$\gamma = 1 - \frac{f_1 + f_2}{2f_0} = \frac{f_3 + f_4}{2f_0} - 1$$

be the fractional frequency interval from the pilot to the center of each IF band, and let $\epsilon = |\delta| - |\gamma|$ be the fractional departure of any IF frequency from its band center. Then, for the transmission system depicted in Figure 10-6, a frequency introduced as $f = f_0 (1 + \gamma + \epsilon)$ will be sent one quarter of the distance, converted to $f_0 (1 - \gamma - \epsilon)$, sent one quarter of the distance, converted to $f_0 (1 + \gamma - \epsilon)$, sent one quarter of the distance, converted to $f_0 (1 - \gamma + \epsilon)$, and sent the remaining quarter of the distance. The total relative normalized loss is therefore

$$\rho = \frac{1}{4}\left[\sqrt{1+(\gamma+\epsilon)} + \sqrt{1-(\gamma+\epsilon)} + \sqrt{1+(\gamma-\epsilon)} + \sqrt{1-(\gamma-\epsilon)}\right] - 1 \qquad (8)$$

$$\approx -\frac{\gamma^2}{8} - \frac{5\gamma^4}{128} + \ldots - \frac{\epsilon^2}{8}\left(1 + \frac{15}{8}\delta^2\right) - \frac{5\epsilon^4}{128} + \cdots \qquad (8a)$$

Because γ is constant, only the latter terms in ϵ^2 and ϵ^4 involve variation in transmission across the band. Since only even powers occur, each band is symmetrical about the center, This is to be expected since in Figure 10-6, for example, one observes that each band edge successively occupies all four frequencies f_1, f_2, f_3, f_4.

By analogy to the transposition of open-wire lines to avoid crosstalk, we shall refer to this method of obtaining uniform response as *transposition equalization.* Figure 10-7 shows the relative normalized loss with no transposition, with one midpoint inversion and with two inversions and a band interchange. If $\alpha_0 = 120$ dB,

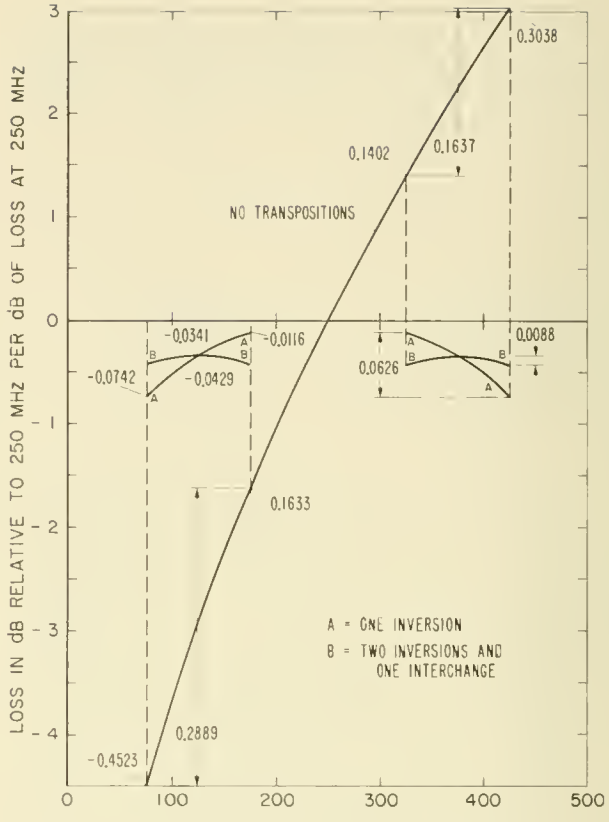


Figure 10-7. *Effect of transposition equalization.*

we see that with one inversion (*A* curves) both IF bands have a slope of 120 (−0.0116 + 0.0742) = 7.5 dB from the inner to the outer edges. Over a 20° C range, this slope would vary by (10/293) 7.5 ≈ 0.34 dB, which is probably small enough not to require compensation. Also, the equalizers to correct the slope across the band present no serious problem. However, a large level difference between IF bands will accumulate (on the longer lines) before the single inversion at the midpoint. On the longer lines, three (or five) inversions might be necessary.

If a midpoint band interchange is used, together with an inversion at the one-quarter and three-quarter points, we see from the *B* curves that the transmission in each IF band is symmetrical about the band center and is 120(−0.0341 + 0.0429) ≈ 1 dB greater at the band edges than at the centers. No temperature compensation is needed and the normal rolloffs of the IF amplifiers and filters can be used to flatten the passband to within a tenth of a decibel if desired.

If the pilot frequency, f_0, lies midway between the two bands—that is, if $f_0 = (f_2 + f_3)/2 = (f_1 + f_4)/2$, and if $f_3 - f_2 = 2f_1$—the inversions and interchanges are easily accomplished. (We are assuming that $(f_2 - f_1) = (f_4 - f_3)$,

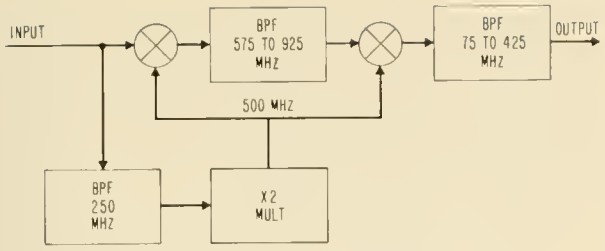


Figure 10-8. *Channel inverter.*

i.e., that the two bands are alike.) Figure 10-8 shows one way of accomplishing the inversion. The 250-MHz pilot is extracted by a narrow band pass filter and multiplied by 2 to obtain a 500-MHz carrier. Using this as the local oscillator (LO) in a first mixer and taking the upper sideband, we translate the entire IF spectrum upward by 500 MHz. Then the same LO signal is applied to a second mixer and the lower sideband is taken as the inverted output. Since the same frequency is used in the up and down conversion, the exact frequency is of no consequence and it may not be necessary to derive the 500-MHz signal from the 250-MHz pilot. However, there may be some advantage in suppressing higher order modulation products to have a particular phase relation between the LO signal and the 250-MHz pilot.

Figure 10-9 shows one method of accomplishing the swapping of the IF channels. Here the 250-MHz pilot is extracted and used as the LO frequency for two mixers and is reintroduced into the line through an isolation amplifier. One mixer is supplied with the upper IF band and the lower sideband is taken as the output. The other mixer is supplied with the lower IF band and the upper sideband is taken as the output. The two new IF bands and the pilot are added at the output.

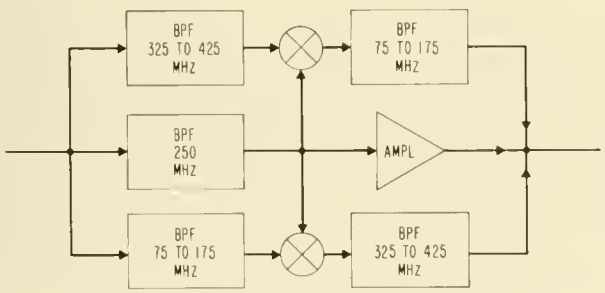


Figure 10-9. *Channel interchanger.*

In the interchange process, each IF band passes through the same two band pass filters, but in reverse order. Thus, the same delay is added to each band. If desired, the 250-MHz pilot can be shifted in phase by an amount corresponding to this same delay. Then, as in the inverter, the whole IF signal can be considered to have undergone an additional flat delay.

For array diameters up to 5 km a satisfactory solution to the IF transmission problem would be to use 7/8-in. diameter coaxial lines and three repeaters per line for all lines. (There are only a few lines so short that this combination would represent overengineering.) Two of the repeaters would be inverting and the middle repeater would interchange bands. Also, repeaters would have a maximum gain of about 30 dB. Automatic gain control would be used to set the level of the 250-MHz pilot at the output of each repeater to a standard level. If the AGC system cannot operate satisfactorily over a 30-dB range, input pads could be switched in on the shorter runs to center the AGC in its proper operating range. In this way all signals arriving at the control center will have undergone (except for total delay) substantially the identical transmission history.

For larger arrays, additional pairs of inverting repeaters may be added symmetrically about the midpoint of the long lines, which are now divided into sixths or eighths, etc., instead of quarters.

With the continued development of microelectronic hybrid integrated circuits the repeaters may become smaller and cheaper and have tighter tolerances. Thus, it may well be that, by the time the Cyclops system is built, a closer repeater spacing and smaller cable would be a more economical solution.

Referring once more to Figure 10-3 we see that at the control center each IF line ends in a final repeater having up to 30 dB gain. The entire IF signal then passes through a variable delay unit having a delay range of a few nanoseconds. This can be done with digitally switched lengths of strip line. Following the delay unit, the 250-MHz pilot is extracted and its phase is compared with a locally synthesized 250-MHz signal. Changes in phase of the pilot are taken to signify change in the IF delay and are servoed out with the variable delay unit. In this way, we make double use of the standard frequency distribution system; not only does it supply the reference phase for the local oscillators at each antenna, but it also stabilizes the delays of the entire IF transmission system.

For this compensation method to be operable, all the 250-MHz pilots and the 250-MHz references at the control center must be generated without phase ambiguity. A precise method of doing this is to divide the 1-GHz standard reference frequency by 4, and to permit the divider to lock only if the 250 MHz so obtained agrees in phase (to within say $\pm 5^{\circ}$) with the tenth harmonic of the 25-MHz standard reference frequency. In this way the

phase stability resulting from *dividing* the higher frequency is obtained and the lower frequency serves only to remove ambiguity. The 500- and 750-MHz signals used for mixing at the antenna and central station are conveniently obtained by doubling and tripling the 250 MHz signal.

At the central station, the 750-MHz signal is mixed with the IF signal and the upper sideband is taken. This translates the entire IF signal up in frequency to a range suitable for the IF delay system.

Since the array radius, and hence the average IF transmission line length increases as the square root of the number n of antennas, the total IF cable required varies as $n^{3/2}$. The total cable cost may therefore be expressed as

$$C = \gamma k s n^{3/2} \tag{9}$$

where γ is the cable cost per foot, s is the antenna spacing and k is a factor that depends on the tunnel pattern. In Appendix L, k is shown to have the following values:

Tunnel Pattern	k
1*a*	0.4789
1*b*	0.3967
2*a*	0.4457
2*b*	0.3582

If we assume 7/8-in. coaxial cable at \$3/m, a spacing s of 300 m, and configuration 2*b*, the average cable cost *per antenna* is C/n = \$322 $n^{1/2}$. For a 1000 element array this is \$10,200.

The repeaters are estimated at \$1800 each and the terminal electronics per line at \$5000 to give a total electronics cost of \$10,400 per IF line.

THE IF DELAY SYSTEM

Perfect phasing of the array requires that the local oscillators at each antenna have the same relative phases that would exist if the local oscillator frequency had been received from space along with the signal itself. The IF outputs of the mixers then have the relative phases and delays that would exist if the signal had been received at the IF frequency (without heterodyning). To phase the array, additional delay must now be added in each IF line to bring the total delay from the source to the summing point to the same value.

If the fractional IF bandwidth is small, the LO phasing may be omitted and the array may be phased by choosing a value of IF delay in each line that is nearest the proper delay and gives the right IF phase. This approach is discussed in Appendix M. However, in the Cyclops array with its relatively wide IF bandwidth it seems best to provide for proper LO phasing at each antenna as proposed in Chapter 9.

The task of providing the proper IF delay for each channel is a formidable one. If each IF band is finally demodulated against the (delayed) pilot signal, the resulting "baseband" will extend from 75 to 175 MHz. If the signals are to add in phase, we would like the phase difference due to delay error not to exceed some 8° at 175 MHz. This corresponds to a delay error of ±1/8 nanosecond.

If the processing and control headquarters is located at the center of the array, the total IF delay required is minimized and is fairly easy to compute to within a few percent. We can compute the needed delay exactly if we assume the IF transmission lines are radial rather than confined to the actual tunnel pattern. The IF transmission delay associated with an antenna at radius r is then

$$\frac{r}{v} \leqslant \tau_0 \leqslant \frac{r}{a}\tau_\ell \tag{10}$$

where v is the velocity of propagation, a is the array radius and $\tau_\ell = \ell_{max}/v$ is the delay of the longest transmission lines. With the recommended tunnel patterns, the two limits in equation (10) differ by only a few percent. We will take

$$\tau_0 = \frac{r}{a}\tau_\ell \tag{11}$$

and thus obtain a slightly pessimistic result. If we plot the delay associated with any array element as distance above the plane of the array, then equation (11) defines a "cone" of delay with its apex at the array center and having a height τ_ℓ at the rim of the array as shown in Figure 10-10.

The minimum delay needed to make the array operable at the zenith is that needed to fill this cone up to a level surface; that is,

$$\tau_1 = (1 - \frac{r}{a})\,\tau_\ell \tag{12}$$

To receive radiation arriving at an angle θ from the zenith and at an azimuth ϕ, this level delay plane must now be tilted so that the near point on the rim (at

azimuth ϕ) has a delay $(2a/c)\sin\theta$ greater than the far point (at azimuth $\phi + \pi$). For radiation arriving at zero elevation ($\theta = \pi/2$). The delay difference is $2a/c$, the propagation time across the array (Fig. 10-10). The delay plane is the tilted ellipse, and the delay for an element at A is $\tau_\alpha = \tau_\varrho + 2a/c$ while that for an element at B is $\tau_b = \tau_\varrho$.

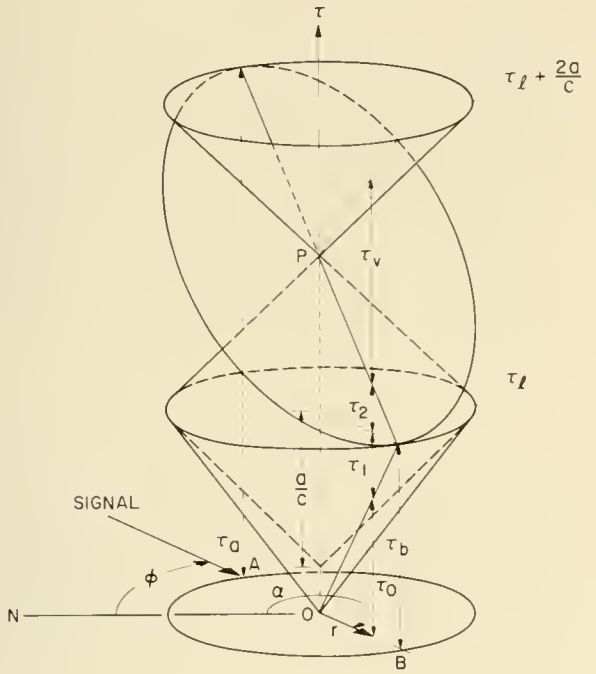


Figure 10-10. *IF delay requirements.*

As ϕ is varied from 0 to 2π, the tilted delay plane rotates around the central axis and is always tangent to two cones having their common vertex at P. If we arrange to pivot the delay plane at P, then for greater elevation angles it will always lie between these two cones. The required delay for any element is never greater than that defined by the upper cone and never less than that defined by the lower cone of the pair. For an antenna element at azimuth α and radius r from the center of the array

$$\tau = \tau_\varrho + \frac{a}{c} + \frac{r}{c} \sin\theta \cos(\phi - \alpha) \tag{13}$$

$$\tau_\varrho + \frac{a-r}{c} \leqslant \tau \leqslant \tau_\varrho + \frac{a+r}{c} \tag{14}$$

Thus the delay compensation can conveniently consist of two parts—a fixed part

$$\tau_f = \tau_{\min} = \tau_0 + \tau_1 + \tau_2$$

$$= (\tau_\varrho + \frac{a}{c})(1 - \frac{r}{a}) \tag{15}$$

and a variable part having a delay range

$$\tau_v = \tau_{\max} - \tau_{\min} = \frac{2r}{c} \tag{16}$$

For an array 10 km in diameter, $a/c = 16.67$ μsec and $\tau_\varrho \approx 20$ μsec. Thus, the fixed delay needed for the center element is 36.67 μsec and the variable delay needed for an element at the rim is 33.3 μsec. The maximum total delay $(\tau_1 + \tau_2 + \tau_v)$ needed for any IF line is thus about 37 μsec and must be accurate to ±1/8 nsec, an accuracy of about 3.4 parts per million. Clearly the delay system is going to require stabilization of some sort to achieve this accuracy.

The average value of τ_f is one third the sum of the altitudes of the lower two cones:

$$\tau_f = \frac{1}{3}(\tau_\varrho + \frac{a}{c}) \tag{17}$$

while the average value of τ_v is

$$\bar{\tau}_v = \frac{2}{3} \cdot \frac{2a}{c} = \frac{4a}{3c} \tag{18}$$

The total amount of delay required for n antennas therefore is

$$T = \frac{n}{3}(\tau_\varrho + \frac{5a}{c}) \tag{19}$$

Taking $\tau_\varrho = a/c$ gives the minimum possible total delay for a fully steerable array

$$T = n\,\frac{2a}{c} \tag{20}$$

If the delay is obtained with additional transmission line the total length needed is vT, where v is the velocity of propagation. Thus $L_{\min} = 2na(v/c)$. Since the total length of line used in the IF transmission system is about $(2/3)\,na$ (and $2/3 < v/c < 1$), we see that about two to three times as much line would be needed for the delay

system as for the IF system itself. For a thousand element 10-km-diameter array this would be about 5000 to 7000 miles of cable. Clearly, the use of IF cable for all the delay would be cumbersome and costly.

Nevertheless, it might be worth considering making all the IF cables in the IF distribution system the same length as the longest cable. This would add about 50% more IF cable and, using the equalization scheme proposed, would require no more repeaters. The additional cost would average about $5000 per line for the 10-km array, which may not be much greater than the cost of an electronic delay unit. The extra IF cable could be accommodated by running it out a side tunnel as far as necessary and then back to the antenna involved. We would then have an array already phased and stabilized for looking at the zenith and would need only to add the variable delay to tilt the delay plane.

Acoustic surface wave delay lines appear to be the best broadband delay devices available at this time. In these devices, a thin film interdigital electrode structure is used to launch a Rayliegh or Love type surface wave on a piezoelectric crystal substrate such as quartz or lithium niobate. A similar electrode structure detects the passage of the wave at the other end or at an intermediate point. The surface waves propagate with essentially no dispersion and with very little loss ($\approx$ 1 dB/μsec). The loss, bandwidth, and dispersion problems occur in the transducers (electrode structures) used to couple the electrical signals into and out of the crystal. A good survey of acoustic surface wave techniques is given by R.N. White in the August 1970 Proceedings of the IEEE.

Lithium niobate lines appear most attractive at this time. Some of the properties obtainable with this material are listed in Table 10-3.

TABLE 10-3

Velocity	3.45×10^5 cm/sec
Temperature coefficient	85 ppm/°C
Insertion loss	10 – 35 dB
Fractional bandwidth	75%
Center frequency	0.1 to 1 GHz
Delay precision	$<10^{-9}$ sec
Phase dispersion	$<5°$

To get 16 μsec of delay requires a lithium niobate crystal only 5.5 cm long. Thus, for delays of the magnitude required for Cyclops, the cost of a delay unit will be almost independent of the delay, and there is no great advantage in minimizing the amount of variable delay. It is proposed that the shorter IF lines in Cyclops be lengthened by an amount

$$\Delta\ell = v(\tau_\ell - \frac{a}{c})(1 - \frac{r}{a}) \qquad (21)$$

thereby filling in the basic delay cone in Figure 10-10 to a new cone (shown dotted) having its apex a/c below the delay τ_ℓ. The total delay that now must be added by the delay system to achieve full steerability is $2a/c$ *and is the same for each IF line.* Although full use will never be made of the variability of central delay units, several advantages including interchangeability accrue from having all units alike.

It is proposed that the delay system associated with each line consist of individual delay units connected in tandem, each unit having a delay

$$\tau_k = \tau_0 + 2^k \text{ nsec} \qquad (22)$$

when switched in, and a delay τ_0 when switched out. For a 10-km array, k ranges from -2 to $+14$. The delay required in any line (in units of 250 picoseconds) is then expressed as a 17-digit binary number and the delay units are switched in or out if the digit they represent is a 0 or 1 respectively.

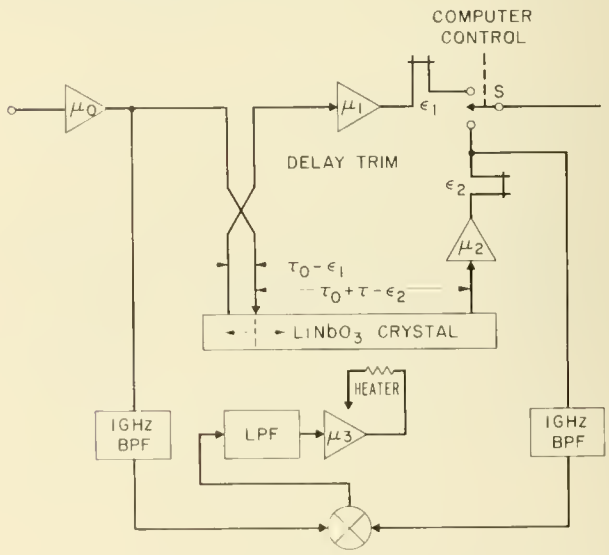


Figure 10-11. *Acoustic delay unit.*

For $-2 < k \leq 1$ – for the four shortest delay steps having delays of 0.25, 0.5, 1, and 2 nsec – the delay can easily be achieved by switching in appropriate lengths of

stripline. For $2 \leqslant k \leqslant 6$, appropriate lengths of coaxial line may be used. For $7 \leqslant k \leqslant 14$, it is probably more economical to use the acoustic surface wave delay units.

Figure 10-11 is a block diagram of an acoustic delay unit. The entire IF signal (825-1175 MHz) is amplified by the input amplifier μ_0 and drives the lithium niobate delay line. The wave launched propagates in both directions and is absorbed at both ends of the crystal. The wave propagating to the left is picked up after some convenient minimum delay $\tau_0 - \epsilon_1$. This signal is amplified by the amplifier μ_1 and is given an additional small trimming delay ϵ_1 in a loop of stripline. By adjusting ϵ_1 we can set the delay via this path at τ_0. The wave propagating to the right is picked up after a delay $\tau_0 + \tau - \epsilon_2$, amplified in amplifier μ_2, and trimmed by the additional delay ϵ_2, and arrives at the switch with the precise delay $\tau_0 + \tau$.

When the switch S is thrown by the computer, a delay $\tau = 2^k$ nanoseconds is inserted or removed. In either position of the switch, the signal is introduced and picked up from the acoustic line. Thus, switching the delay in or out does not affect any phase shift (or dispersion) caused by the transducer electrodes. Regardless of the delay, the total number of transducers through which every IF signal passes always remains the same. Hence any dispersion caused by these elements is common to all lines and does not interfere with the constructive addition of the signals.

To hold the switched delay of each unit at precisely an integer number (2^k) of nanoseconds, τ_0 is adjusted to be an odd number of quarter nanoseconds. Then the phase of the 1 GHz pilot should be 90° different at all times at the input and output. To assure this, the pilot is stripped at the input and output, and the two signals are applied to a phase detector whose output drives a heating element attached to the acoustic line. *We thus make use of the temperature coefficient of the line* to hold the delay precisely constant. This compensation may not be needed on the units having short delays; it is not expensive and may be included on all units to standardize them and to obviate the need for precise control of ambient temperature.

If further study reveals that indeed there is some dispersion in the acoustic wave propagation, frequency-dependent delay compensation may be added in the shorter units. The longer delay units may be divided into two equal parts (two of the next shorter delay units) and a spectrum inverting repeater introduced between the two halves. By combining these techniques we feel confident that the required delay can be obtained to the required high accuracy and stability.

As the Earth turns, the fine delay units associated with the antennas at the rim are switched rather rapidly. The maximum switching rate is

$$f = \frac{\Omega a}{c\tau_{\min}} \tag{23}$$

where $\Omega = 2\pi/86164$ is the angular rotation rate of the Earth, a is the array radius, c is the velocity of light and $\tau_{\min}$ is the minimum delay step. For a 10-km diameter array and $\tau = 1/4$ nanosecond $f = 4.86$ switch operations per second. For reliability and to minimize switching times it is recommended that all switching be done with diode gates.

With the digital delay switching described above the required delay slope across the array is approximated at all times by a series of steps. The *delay error* is therefore a sawtooth function. As a result the probability density function for the phase error, $p(\theta)$, is a constant $1/2\theta_0$ between the limits $-\theta_0$ and θ_0 where $\theta_0 = \pi f_{\max}\tau_{\min}$. Here $f_{\max}$ is the maximum demodualted IF frequency and $\tau_{\min}$ is the minimum delay step. As a result, the array efficiency is reduced by the factor

$$\eta = \left[\int_{-\theta_0}^{\theta_0} p(\theta)\cos\theta\, d\theta\right]^2 = \left[\frac{\sin\theta_0}{\theta_0}\right]^2 \tag{24}$$

Taking $f_{\max} = 175$ MHz and $\tau_{\min} = 1/4$ nsec we find $\theta_0 = 0.1375$ radian and $\eta = 0.994$. Clearly, the discrete delay steps do not degrade the performance appreciably.

To save central computer storage and time, it is suggested that the delay system for each IF line have associated with it a delay register, an increment register, and an adder. At the outset, the proper delay and increment (positive or negative) are computed for each line. (Both are stored to several more digits of accuracy than used by the delay units.) Periodically—say every 1/10 sec—the increments are added, thus updating the delays. Occasionally, perhaps once a minute, the computer makes its rounds, and updates each delay to the right absolute value and corrects the increment. (The same technique, incidentally, can be used to drive the antennas in azimuth and elevation.)

It is estimated that the cost of the IF delay system should not exceed $8000 per line. Since the number of delay bits increases only logarithmically with array radius we may, to the present accuracy of estimation, assume the cost per line to be independent of array size.

EFFECT OF GAIN AND PHASE DISPERSION

Equation (24) gives the reduction in array efficiency

as a result of a uniformly distributed phase error among the individual IF channels. There will, however, be many small independent sources of phase error besides the quantizing error of the IF delay system. These include

1. Position (not pointing) errors of the elements
2. Surface errors of the antenna elements
3. Feed position errors
4. Errors in the standard frequency phases
5. Phase shifter errors
6. Phase instability of all selective filters
7. Phase shifter errors
8. IF cable delay errors

and others. Many of these errors are greatly reduced by the techniques that have been described; others can be eliminated by overall calibration. Nevertheless, there will be residual errors which, since they arise from a large number of small causes, will tend to be normally distributed. In this case the system efficiency and signal to noise power ratio reduction is given by equation (3) Chapter 8, with $4\pi\delta_{rms}/\lambda$ replaced by θ_{rms}:

$$\eta = e^{-\theta^2_{rms}} \tag{25}$$

where θ_{rms} is rms phase error in radians.

If:

$\theta_{rms} = 0.1$ $\eta = 0.99$

$\theta_{rms} = 0.2$ $\eta = 0.94$

A similar situation exists for gain dispersion. If the gains of the various channels are not identical but are distributed about a mean value μ_0 with a variance $\delta \equiv \sqrt{\overline{(\mu - \mu_0)^2}}$, the noise will be greater than for the ideal case of all gains equal. The total power will be

$$N = \sum \mu^2 = \sum [\mu_0 + (\mu - \mu_0)]^2$$

$$= n[\mu_0^2 + 2\mu_0 \overline{(\mu - \mu_0)} + \overline{(\mu - \mu_0)^2}]$$

$$= n[\mu_0^2 + \delta^2] \tag{26}$$

where n is the number of channels added. For $\delta = 0$ we have $N_0 = n\mu_0^2$, so the reduction in efficiency or signal to noise ratio is

$$\eta = \frac{1}{1 + (\delta/\mu_0)^2} = \frac{1}{1 + \sigma^2} \tag{27}$$

where $\sigma = \delta/\mu_0$ is the normalized variance. If $\sigma = 0.1$, $\eta = 0.99$. Thus we see that overall dispersions of 0.1 radian and 0.1 neper or less are desirable.

ARRAY CONTROL AND MONITORING SYSTEMS

To steer, tune, and phase each antenna element, several control signals must be sent from the control center. Further, to ascertain that each element is functioning properly requires several monitoring, calibration, and alarm signals. In the normal search mode many of the control signals are the same for all antennas. However, Cyclops will be a more valuable tool if the array can be divided into several subarrays for simultaneous use on independent radio-astronomy projects and for the reception of signals from space probes. Designing this flexibility in at the outset adds a negligible amount to the total cost.

Because of the large number of alternative ways of handling the control and monitoring signals (e.g., coaxial lines, cable pairs, time or frequency division multiplex), and because the control and monitoring systems represent only a small part of the total system cost, an exhaustive analysis of this problem was not attempted. Rather than propose an optimum solution, we merely describe a possible solution from which rough cost estimates can be made.

Control Systems

The major control signals needed and the information content associated with each are listed in Table 10-4.

TABLE 10-4

CONTROL FUNCTIONS AND BIT REQUIREMENTS

Signal	Range	Accuracy	Ratio	Bits
Azimuth position	360°	2″	648,000	20
Azimuth rate	±5°/sec	2″	±18,000/sec	15/sec
Elevation position	90°	2″	162,000	18
Elevation rate	±5°/sec	2″	±18,000/sec	15/sec
LO phase	360°	1°	360	8
LO rate	±18,000°/sec	1°	±18,000/sec	15/sec
Assignment code	16	1	16	4
Receiver select	16	1	16	4
Pump frequency	10 GHz	25 MHz	400	9
RF test	10 GHz	5 MHz	2000	11
Operational command	256	1	256	8

The azimuth and elevation position and rate infor-

mation are common to all antennas in a particular subarray. The rest of the information is specific to a particular antenna. In principle, it is not necessary to send both position and rate information. However, unless a closed-loop data system is provided for continuous monitoring of all position information, it is impossible to correct the positions by rate signals alone. If only position information is sent, then in one instance at least (the local oscillator phase shift), the refresh rate would have to be several times per second. By sending both, and by incorporating a modest amount of logic and arithmetic capability at each antenna, we can greatly reduce the required refresh rate. In fact, the refresh rate will probably be set by the delay one can tolerate in having the array respond to new instructions. For the discussion to follow, we assume a refresh rate of once a second.

To address one of a thousand or more antennas we will need an address code of 10 or more bits. Let us assume a 12-bit code. We can then divide the array into natural sectors determined by the tunnel pattern and serve all the antennas in each sector from a common coaxial (or shielded pair) cable system. The first three bits of the antenna address would then determine over which of up to eight cable systems the remainder of the addresses and the data for the antennas in a given sector were sent. Thus, we can address any of 4000 antennas with only a nine bit *transmitted* address.

During each refresh cycle we would then send the address plus the LO phase and phase rate—a total of 32 bits of information—to each of as many as 512 antennas in each sector for a total of 16,384 bits.

In addition we propose to send over all cables all the position information needed for all possible subarrays. An antenna assignment code, previously transmitted to each antenna would determine which set of position data is read by each antenna. Each array requires 68 bits of information. If we assume that 16 subarrays are possible at any one time, an additional 1088 bits are required for a total of 17,472 bits of information per refresh cycle.

Synchronizing and punctuation information must also be transmitted. This can be done in several ways, such as by pulses of a different amplitude or polarity. One rather attractive method[1] uses a set of unique code symbols, no one of which can occur as a combination of the others under any time displacement. In addition:

1. The code is self-clocking; a transition occurs each Nyquist interval.
2. There is zero dc component in each symbol.
3. The code is *immediate;* each symbol is uniquely dicipherable in and of itself.

The method requires a little more than three Nyquist intervals per symbol on the average, but in our case this theoretically would mean only 26 kHz of bandwidth. In practice, 50 kHz of bandwidth would be ample.

The information in the last five items listed in Table 10-4 need only be sent occasionally—when an antenna is being calibrated, tested, or transferred from one subarray to another. It is proposed that this information, when needed, be sent *in lieu* of the phase shift information to the antenna or antennas involved, the latter information being unnecessary at such times. A special code group after the antenna address would signal the processer that the data was command, not tracking, information.

Thus in setting up a subarray, the operator would list the antennas wanted in the array, the operating frequency, and the right ascension and declination of the source. These data would be fed into the central computer, which would search for the first vacant array codes. It would then assign the array code to all the antennas involved and specify the proper receiver and pump frequency. It would then compute the azimuth, elevation and phase data, and start sending these in the proper slots. An operational command code would then be sent to slew to the new position. While this is happening position error and alarm signals are disabled, and no IF signal is transmitted. After all antennas have been interrogated (see the next section) to assure that they are in proper position and operating correctly, a start transmission command would be sent.

The logic and memory required at each antenna to make the above control system operable is very inexpensive. It could very likely all be put on one or two LSI chips, and its cost should not exceed $100 per unit in quantities of 1000 or more. However, the output circuitry and switching system required to interface the data with the units involved would raise the cost considerably. A total cost of $5000 per antenna control unit seems reasonable.

The cabling involved costs about 25 cents/m and its length is about equal to the total tunnel length. Probably no repeaters would be needed; only transformers at the branch points and bridging amplifiers at the antenna sites. For a 1000-element array with 300-m spacing the total cost should not exceed $125,000 or another $125 per antenna.

It thus appears that the cost of the control system is so small as to be lost in the cost uncertainty of the expensive parts of the system.

Monitoring and Calibration System

It is not very important in the search mode, if, say,

[1] S. Walther, private communication.

1% of the antennas in the Cyclops array are inoperative. It is very important, however, that an antenna in trouble not contaminate the signal with a large added noise, or with spurious signals. Further, for maintenance purposes, any difficulty should be reported to the control center as soon as it can be detected. Thus, a rather elaborate self-diagnostic routine should be an integral part of the design.

Some of the quantities that should be monitored are:

1. Position drive servo operation
   a. Position error
   b. Motor temperature
   c. Drive circuitry voltages and currents
2. Feed position confirmation
3. Cryogenics operation
   a. Compressor and motor temperatures
   b. Helium pressures
   c. Receiver and maser temperatures
4. Receiver voltages and currents
5. Pump and LO frequencies and levels
6. IF level
7. Ambient temperatures (fire alarm)
8. Receiver noise temperature
9. Receiver sensitivity
10. Antenna phasing

Most of the monitoring can be done with the antenna element in operation. It is proposed that all tests in this category be done automatically. All that is required is a number of transducers and a scanner that looks at the outputs of these (or the electrical signals from various test points) and routes the signals to a programmable digital voltmeter or frequency counter. In addition, a small amount of logic would be needed to program the test instruments to their proper settings, to sequence the tests, to compare the measurements with high-low limits, and to initiate the proper alarm sequences when a malfunction is detected.

As in the case of the control system antenna logic unit, the cost of the logic circuits would be small; most of the cost would be in the instrumentation. Both counters and digital voltmeters have dropped substantially in price with the advent of integrated circuits. We believe the local monitoring function could be performed with an equipment cost not exceeding $8000 to $10,000 per antenna.

When a malfunction or abnormal reading was detected, the local monitoring system could communicate this to the control center in one of two ways, neither of which requires an additional transmission system.

1. The detection of a fault could interrupt the 250-MHz IF pilot frequency, killing the IF transmission. A pilot sensor on the IF line could then sound an alarm and identify the antenna in trouble.
2. The monitoring unit could dial the computer over the communications (telephone) system.

The relative merits of these two methods have not been fully explored. Method 1 has the possible advantages of instant protection against disturbing signals from the faulty antenna and of providing a failure check on the IF transmission system itself. Method 2 requires little if any added equipment. In either case, the central computer could interrogate the antenna monitoring system and find out and print the nature of the trouble.

Monitoring functions 8, 9 and 10 cannot readily be performed with the antenna in service. Here it is proposed that the central computer make the rounds of all antennas, sequentially drop them from their immediate tasks, and make these measurements. The measurements of noise temperature and sensitivity could be made (with the antenna pointed at a known quiet part of the sky) by means of a noise diode and the test RF signal from the antenna synthesizer. The overall phasing and gain adjustment of the array and the calibration of its sensitivity could be made by correlating the output of the antenna under test with that of a reference antenna with both trained on a standard known radio source. This procedure is discussed in Appendix N. If a typical calibration took, say, 15 to 20 min including slewing times, all antennas would then receive an overall monthly checkup. Together with the local monitoring and closed-loop phase and delay regulating systems this operation should be adequate to ensure the health of the entire array at all times.

Naturally, all this automatic monitoring would have to be supplemented with a regular program of protective maintenance plus an emergency repair operation. If this routine maintenance took one day per antenna, a crew of 6 to 12 men could provide a 1000-element array with regular service at 6 month intervals.

The Intercom System

It is imperative that maintenance and repair crews be in direct contact with the control center when they are at any antenna site. For this function a complete telephone system with a central automatic exchange is proposed.

The cost of the central exchange is estimated at $200 per line or roughly $250,000 for a 1000-element array and for the office and other phones in the control center. The cost per station is estimated at $50 or $62,500 for the estimated 1250 telephones required.

A cable pair costs about 20 cents/m installed. If an individual pair were run from the control center to each antenna, the cabling cost would be about

$$C = \$0.2 \times 0.36\, n^{3/2}\, s$$

For a 1000-element array with 300-m spacing this is $680,000. On this basis we estimate the cost of the telephone system at about $1 million. Trunking techniques with remote substations could reduce the cable cost substantially and should be considered. The main consideration is that they not interfere with the monitoring function.

The Master Computer

In addition to computing all the position and phase shift data, controlling and monitoring the array, and reporting faults as described above, the central computer is expected to automate the search process. It must access its (tape) file of the stars to be searched and direct the array at these when they are in the sky. It must repeat searches when spectral anomalies are found, acting in response to the data processing equipment.

In spite of the large number of functions and their detailed nature, only a modest size computer is required. The data rates are slow and a medium size computer with microprogrammed trigonometric algorithms should be able to handle the job easily. Only a few trigonometric functions need be computed for each subarray each second. The rest is simple arithmetic (multiplication by stored coordinates, etc.). Only 64 bits of memory are needed to store the coordinates of the most distant antenna in a 10 km array with an accuracy of 1 mm. Thus for 1000 antennas, 4K of core memory (16 bits/word) is needed for this information to be entirely resident in core. Another 5K of memory should suffice to store all the computed control data, 1K should handle the monitoring function, and 6K should be adequate for interpretive interaction with terminals.

Doubling these values for a two to one margin results in only a 32K machine. With a small disk, tape deck, and other peripherals the system cost should not exceed $120,000. At this price we can afford 100% redundancy to minimize the risk of shutting down the entire system. Conclusion: $250,000 at most is needed for the computer control.

Power System

To avoid radio interference, the Cyclops array should be built in a remote location. As a result, a power plant will be needed to supply the electrical power for the system and for the associated community. A buried, shielded distribution system is recommended to avoid local radiation from arcing insulators and the like. The generating capacity needed may be estimated from the demands listed in Table 10-5 for the array system.

TABLE 10-5

	Peak (kW)	Average (kW)
Power requirements per antenna		
Antenna drives		
Azimuth (5 hp) (η = 75%)	5.	1.
Elevation (2 hp) (η = 75%)	2.	.25
Feed change (1 hp)	.75	—
Frame temperature control (5 hp)	3.5	1.5
	11.25	2.75
Cryogenics		
1 Unit @ 2 kW	3.	3.
1 Unit @ 5 kW	6.	5.
	9.	8.
Lighting	1.	.5
Local electronics		
Receiver front end	.1	.1
Mixers and LO synthesizers	.2	.2
Monitoring and control	.2	.2
	.5	.5
IF System		
Repeaters (3)	.012	.012
Delay system	.008	.008
	.020	.020
Addition of radar mode (1 kW per antenna)	3.	0.
Total power per antenna	25.	11.
Control and processing center power requirements		
Air conditioning	500	
Electronics	200	
Shops	200	
Lighting	100	
	1000 kW	

In addition, power needs of the community were estimated on the basis of an average consumption per person of about 1 kW, with a peak of perhaps 5 kW when averaged over many people. If we assume a population of 1 person per antenna, we may add these figures to the antenna demands.

If we assume a nuclear power plant (Cyclops will outlast our fossil fuels), the cost will be about $300 per kW of generating capacity. Distribution costs normally

run about $225/kW. We will assume $300 to allow for buried services. Roughly, then, the total power system capacity and cost will be as given in Table 10-6. The power plant should probably be designed at the outset to handle the ultimate needs, since on-site construction would raise the power consumption during this phase.

TABLE 10-6

Number of 100-m antennas	Equivalent aperture, km	Power, MW Peak	Power, MW Average	Cost, $ Million
100	1	4	2.2	2.4
200	1.4	7	3.4	4.2
500	2.2	16	7	10
1000	3.16	31	12	20
2000	4.4	61	25	36

11. SIGNAL PROCESSING

The Cyclops system as it has been specified in the last three chapters amounts to a very large radio telescope with an effective clear aperture of a few kilometers, capable of simultaneous reception of both orthogonal polarizations of a received signal over a 100-MHz band. The received band can be quickly tuned anywhere in the low end of the microwave window and, if desired, could be extended to higher frequencies in that window. The system up to this point is a very high resolution, high sensitivity instrument that would find many applications in radio astronomy, radar astronomy, and space probe communications. Each of these applications would require further processing of the signals delivered by the phased outputs of the array. Much of this processing would use standard equipment and techniques peculiar to the application involved. These will not be discussed here.

This chapter is primarily concerned with the signal processing techniques and equipment needed to allow the Cyclops system to carry out efficiently its primary mission of detecting signals originated by other intelligent life—spectrally narrow band information-bearing signals. A second concern of this chapter is the techniques and equipment needed to form wide band images of the radio sky, and thereby greatly speed the construction of detailed maps of natural sources. The chapter concludes with a discussion of interfering signals the Cyclops system must contend with and what might be done about these.

THE DETECTION OF NARROW BAND SIGNALS

We concluded in Chapter 6 that signals of intelligent origin, and particularly signals from intentional beacons, were most likely to contain strong, highly monochromatic components. We saw that these coherent components would probably be best detected using a receiver having a predetection bandwidth on the order of 0.1 to 1 Hz, but that to search sequentially across the spectrum with such a narrow band receiver would result in prohibitively long search times per star. What we are seeking, therefore, is a receiver with some 10^8 to 10^9 channels each 1 or 0.1 Hz wide so that we can monitor the entire 100-MHz IF band simultaneously. In this section, we describe such a receiver, but before doing so we will dispose of some alternative methods that have been proposed.

Total Power Detection

In principle we do not need to divide the spectrum into narrow channels to detect the increase in total power in the IF band produced by a coherent signal in one (or more) channels. All we need to do is to integrate for a long enough time to be able to measure the increase over the noise power alone that is produced by the signal. A little reflection shows that, while possible in principle, such an approach is totally impractical. Suppose there is a coherent signal that doubles the noise power in a 1 Hz band. This produces an increase of 1 part in 10^8 in the total noise in a 100 MHz band. To detect such a tiny increase would require integrating 10^{16} samples of the wide band noise. (See e.g., equations (8) (9) and (10) Chap. 6). This is an integration time of 10^8 seconds or roughly 3 years. Even accepting this time, the method fails, for we cannot assume we know the noise bandwidth to this accuracy nor that it is this constant, nor that the system noise temperature is this constant, nor even that the radio sky is this constant.

Cross Power Detection

Simpson and Omura[1] have proposed that, instead of a simple noise power measurement, a measurement of

[1] NASA unpublished report, 1970.

cross power be made using two signals each from half the array. Their hypothesis seems to be that only the received signal will be correlated in the two signals and therefore produce a cross power. Unfortunately, this is not the case. *Any signal* received by the two subarrays that lies in the beam of the combined array will also produce a cross-power term. Such signals include the 3°K background radiation and noise from the star whose planetary system is being searched. In addition, this approach halves the effective receiving area. At most a reduction of the effective system noise temperature (by a factor of 5 or 6) to that of the sky alone is achieved and the reduction in effective area reduces this improvement to a factor of 2-1/2 to 3. This is not enough to make cross power detection attractive.

Golay Detection

A method of distinguishing between coherent radiation (representing a signal of intelligent origin) and incoherent radiation (representing system and sky noise) has been proposed by Golay (refs. 1,2) and discussed by Bracewell (refs. 3,4) and Hodara (ref. 5). In the Golay detector an IF band is demodulated to baseband in two mixers supplied with local oscillator signals that differ in phase by $\pi/2$. The two mixer outputs are low pass filtered and applied to the X and Y axes of a cathode ray oscilloscope as shown in Figure 10-1. If a sinusoidal signal alone is present anywhere in the IF band, the CRT will display a circular trace. If gaussian white noise alone is present as the IF signal, each mixer will produce a gaussian noise output and the two outputs will be statistically independent. In this case the CRT display will be, circularly symmetrical two dimensional gaussian distribution. With both a sine wave and noise present,

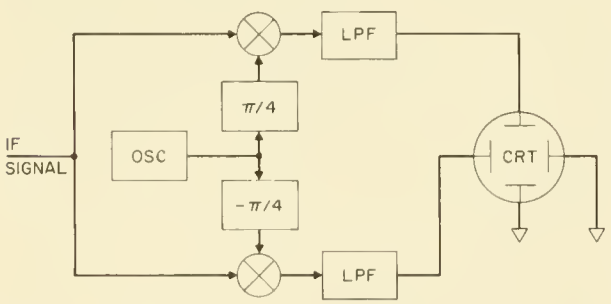


Figure 11-1. *A Golay detector.*

the distribution, if the sinusoidal signal is strong enough, is an annulus—a fuzzy ring whose mean radius is the amplitude of the sinusoid and whose radial spread is proportional to the rms noise amplitude. The distinctly different appearance of this ring or "mole run" and the gaussian central mound produced by noise alone has led many people to suppose that the Golay detector is an especially sensitive detector of coherent radiation. Not so.

First we note that since the distribution always has circular symmetry with or without the sinusoid present, all information about the presence of the coherent signal must be contained in the probability density versus radius. It is easy to show (see Appendix D) that this distribution is given by

$$p(r,\theta) = \frac{1}{2\pi\sigma^2} e^{-\frac{a^2+r^2}{2\sigma^2}} I_0\left(\frac{ar}{\sigma^2}\right) \qquad (1)$$

where

a = amplitude of sinusoidal component

r = radius

θ = angle

σ = variance of gaussian distribution in the absence of a signal ($a = 0$)

I_0 = zero-order modified Bessel function of the first kind

One might measure this distribution by measuring the total light from the CRT produced in a ring of radius r and width dr, and dividing by $2\pi r\, dr$.

If instead one divided merely by dr the result would be

$$q(r) = \frac{r}{\sigma^2} e^{-\frac{a^2+r^2}{2\sigma^2}} I_0\left(\frac{ar}{\sigma^2}\right) \qquad (2)$$

which is the probability density function for a simple linear envelope detector (or for a square-law detector followed by a square root law device).

Since $p(r,\theta)$ can be obtained from $q(r)$ and vice versa, the output of a Golay detector contains no more (and no less) information about the presence or absence of a coherent signal than the output of a simple linear or square law detector. Figures 11-2 and 11-3 show the probability density functions for the Golay detector (equation (1)) and for the linear detector (equation (2)). Both show a significant change between the cases $a^2/2\sigma^2 = 0$ and $a^2/2\sigma^2 = 1$ (unity received signal to noise ratio). Neither would show a perceptible difference between $a^2/2\sigma^2 = 0$ and $a^2/2\sigma^2 = 10^{-8}$ which represents a unity signal-to-noise ratio in one of the 10^8 1 Hz wide channels in a 100-MHz IF band.

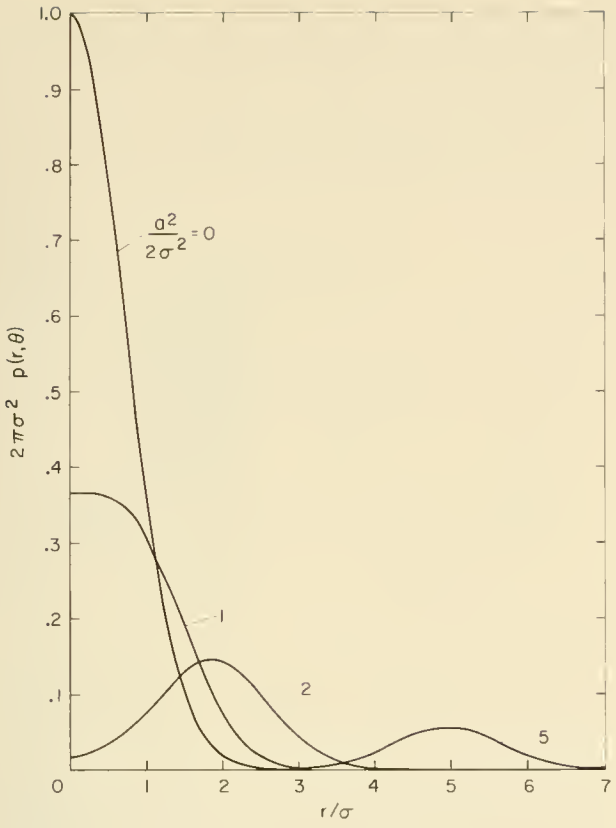


Figure 11-2. *Probability density versus radius (Golay detector).*

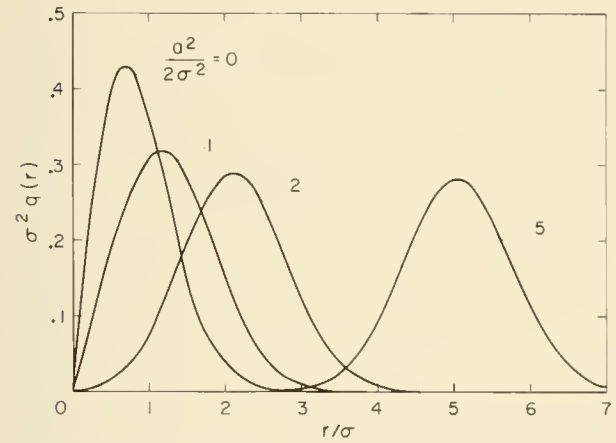


Figure 11-3. *Probability density versus amplitude (r) (linear detector).*

We conclude that the Golay detector offers no sensitivity advantages for its added complexity. Furthermore, even the "mole run" distribution is lost if more than one coherent signal of comparable amplitude exists in the received band. With several carriers present their asynchronously rotating vectors tend to produce a gaussian amplitude distribution just like the noise itself.

There is no way to cheat nature. If the hallmark of signals of intelligent origin is their spectral purity, we must perform a spectrum analysis of the received signal to the resolution necessary.

Scanning Spectrum Analyzers

Many spectrum analyzers sweep a narrow band receiver across the band of interest. If this narrow band receiver has a bandwidth Δf and if the band over which it is swept is B, then it receives the desired signal for a fraction of the time $\delta = \Delta f/B$. Thus, the integration time per scan is reduced in this ratio and must be compensated for by making the total observation time $1/\delta$ times as great. Scanning spectrum analyzers offer no advantage over simply tuning a single receiver of the same resolution slowly across the band. Whatever means of spectrum analysis is used—if it is to fully reduce the observing time—must use all the data in the received signal all of the time.

The Fast Fourier Transform

A method of spectrum analysis that makes full use of the input data, and one that is finding widespread application today is the fast Fourier transform. The received signal (or a selected portion of the band) is converted to digital form by an analog-to-digital converter. If a single unit cannot handle the entire band, parallel operation of several units is employed. The spectrum of the digitized data is then found using the Cooley-Tukey algorithm (ref.6), in which the number of complex arithmetrical operations required to analyze N data points is

$$P = 2N \log_2 N \tag{3}$$

Only hard-wired Fourier transformers are suitable for the data rates posed by Cyclops, and even with these many parallel units are needed. Units that can handle $8192 = 2^{13}$ data points with a speed of 1 μsec per operation, or about 0.2 sec for the $2 \times 8192 \times 13 = 213{,}000$ operations needed, are available commercially and cost about \$50,000 or about \$0.25 per operation performed. From this we infer that the cost of providing the capability of transforming N data points is

$$C \approx \$\ \frac{N \log_2 N}{2} \tag{4}$$

and the required calculation time is

$$T_0 \approx 2(N \log_2 N)\, 10^{-6} \text{ sec} \qquad (5)$$

Now let us determine what sort of bandwidths may be efficiently analyzed on a real time basis with a hard-wired fast Fourier transformer. For this let

T = recording time necessary to achieve the desired frequency resolution ($1/T$) in the spectrum
B = bandwidth handled per analyzer
N = number of data points per recording of signal $\approx 3BT$ (to prevent aliasing)

We assume that, for a given channel, two FFTs are used alternately to achieve real-time operation. One analyzes while the other is recording. It is clear that if $T_0 > T$ the analyzing unit will fall behind, while if $T_0 < T$ it will be idle some of the time. Thus, for most efficient operation $T = T_0$ and we have

$$T = 2(N \log_2 N)\, 10^{-6}$$

$$T = 2(3BT \log_2 3BT)\, 10^{-6}$$

or

$$5 \times 10^5 = 3B \log_2 3BT \qquad (6)$$

If $T = 1$ sec (to achieve 1 Hz resolution) then $B = 11{,}094$ Hz. Since $2N \log_2 N = 10^6$ we find from equation (4) an estimated equipment cost of $250,000 to analyze an 11,000 Hz band. The total system cost to analyze two 100 MHz bands would therefore be on the order of $4.5 *billion*.

Fast as the Cooley-Tukey algorithm may be, it is no match for the data rate of the Cyclops system. The above cost, comparable to the cost of the array itself, plus the expected downtime and maintenance cost of the 18,000 FFT computers required, cause us to reject this approach in favor of the inherently more powerful and less expensive method of optical spectrum analysis described in the next section.

THE OPTICAL SPECTRUM ANALYZER

It is commonly known that the complex amplitude distribution over the back focal plane of a lens is the two dimensional Fourier transform of the complex amplitude distribution over the front focal plane (ref.7). (Note: the front and back focal planes are each one focal length from the corresponding principal plane and are *not* the object and image planes.) The *intensity* distribution over the back focal plane is thus the two dimensional *power* spectrum of the front focal plane distribution. Because of these relationships and, because lenses have enormous information transmission capacity, coherent optical systems are widely used to obtain two dimensional power spectra. Not so well known is the fact that this large information rate of lenses can be used efficiently to obtain the power spectra of one dimensional signals (refs. 8,9).

The signal to be analyzed is recorded in a raster scan on a strip of photographic film or other suitable medium so that after development the transmittance of the film is directly proportional to the signal amplitude. A dc bias or offset is added to the signal (or in the light modulator) to avoid negative amplitudes. Figure 11-4 shows a laser beam recorder in which the laser beam is first modulated by the biased signal and then deflected horizontally by a sawtooth waveform. Blanking signals are applied to the modulator during flyback. A lens brings the modulated and deflected beam to a sharp focus on a strip of film moving downward at a constant rate. Each line of the raster thus represents a time segment of the input signal.

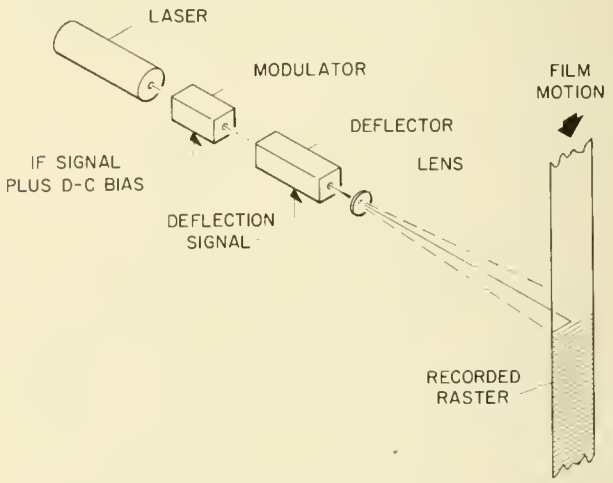


Figure 11-4. *Raster scan recorder.*

After exposure, the film is run through a rapid processor and into the optical spectrum analyzer. The processor introduces a fixed delay into the system but otherwise the analysis is done in real time. Figure 11-5 shows the developed film being pulled through the gate of the spectrum analyzer where it is illuminated with

collimated coherent light. A simple lens then takes the two dimensional Fourier transform of the amplitude distribution in the gate and produces an intensity distribution over the back focal plane which is the desired power spectrum *also mapped in raster form,* as shown in Appendix O.

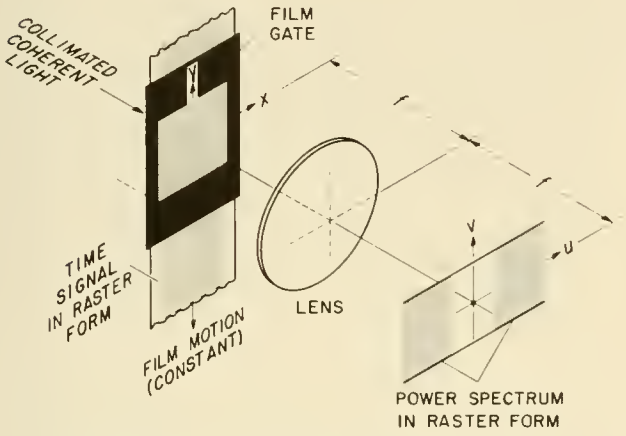


Figure 11-5. *Optical spectrum analyzer.*

To understand how this occurs, consider the case of a sinusoidal imput signal. If the frequency of this sinusoid is exactly n times the scanning frequency, then (neglecting the flyback time) n cycles of the input signal will be recorded per line and the cycles on any line will lie directly under those of the previous line. The result is that the developed film will contain (almost) n vertical clear and opaque pairs of strips as shown in Figure 11-6a. When placed in the gate of the analyzer this pattern acts like a diffraction grating. At the back focal plane three spots are produced: a central spot, O, produced by the dc component of the signal, and two side spots, A_1 and A_2 produced by the two exponential components of the sinusoid $\cos \omega t = (e^{i\omega t} + e^{-i\omega t})/2$. Since the grating axis is a–a' is vertical, the spots lie in a horizontal line with a separation proportional to ω.

If ω is now increased so that there is more than an integer number of cycles per line, the transmission peaks in each line will be shifted to the left of those in the line above. There are still as many cycles per line horizontally (in fact, slightly more) but now there are a number of cycles vertically in the frame. The grating axis is rotated clockwise through an angle θ as shown in Figure 11-6b. The left spot, A_1, shifts upward and the right spot, A_2, shifts downward so that the line on which the spots lie also rotates clockwise through an angle θ. Both spots shift slightly outward because the horizontal pitch of the grating has decreased. Note also that a new grating axis b–b' is beginning to appear.

If ω is further increased so that there are exactly $n+(1/2)$ cycles per scan, the transmission peaks of successive lines are staggered like successive courses of bricks, as shown in Figure 11-6c. Both grating axes a–a' and b–b' are now equally prominent. As the spots A_1 and A_2 (corresponding to axis a–a') leave the frame two new spots B_1 and B_2 (corresponding to axis b–b') enter the frame.

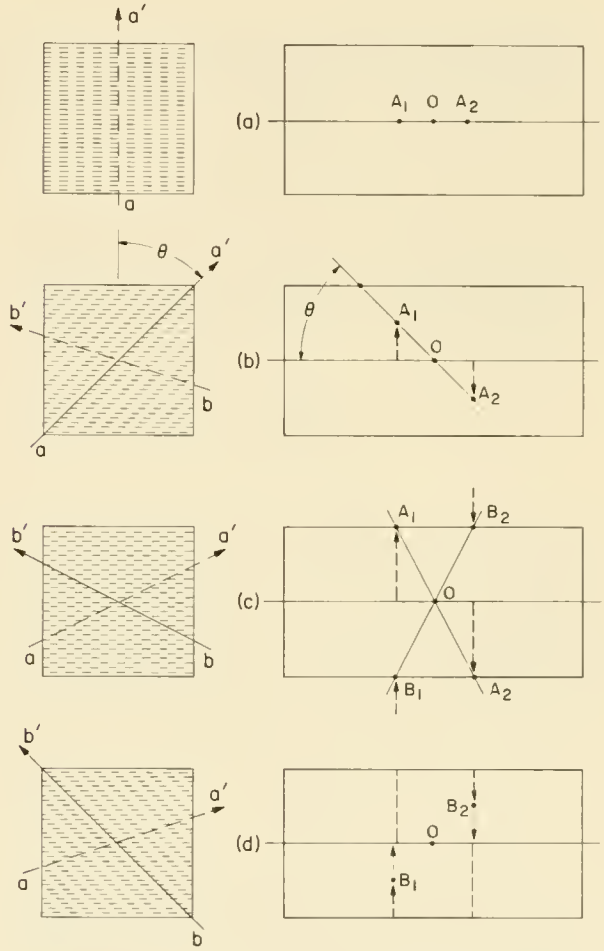


Figure 11-6. *How the power spectrum is generated in raster form.*

With further increase in ω these new spots approach the horizontal axis as indicated in Figure 11-6d. Finally, when ω is $n+1$ times the scan frequency, the grating axis b–b' becomes vertical as in Figure 11-6a, but now there is one more grating cycle per line so the spots B_1 and B_2

arrive on the horizontal axis a little farther from the central spot than the original spots A_1 and A_2.

As a single sinusoid is swept upward in frequency a succession of pairs of diffraction spots traces out a series of almost vertical lines to form two rasters, one to the left and one to the right of the central spot. Thus, frequency is mapped into two rasters in the transform plane. Each line of each raster maps a frequency interval of f_S Hz where f_S is the scanning frequency used in the recording process.

Time-Bandwidth Product

The resolution of the spots along the raster lines of the power spectrum is determined by the height of the gate, which in turn is proportional to the time duration of the recorded signal contained in the gate. The spot intensity versus distance along the raster lines is the square of the transform of the gate transmission versus height. Expressing the distance along the raster line in terms of frequency f and the gate height in terms of duration T of the signal sample contained in the gate, we find (see Appendix O) for a rectangular gate

$$\frac{I}{I_0} = \left(\frac{\sin \pi(f - f_0)T}{\pi(f - f_0)T}\right)^2 \tag{7}$$

where I is the intensity at a point on the raster line corresponding to frequency f, and I_0 is the intensity at the center of the spot, corresponding to the frequency of the sinusoid, f_0. We shall take the resolving power to be the frequency interval, Δf, from the center of the spot to the first null. Thus, from equation (7):

$$\Delta f = \frac{1}{T} \tag{8}$$

The gate transmission versus height can be tapered to produce other kinds of selective filters, such as gaussian, or ideal bandpass. The latter (and any filter having an oscillatory impulse response) would require half wave plates to produce phase reversals with height in the gate. We have not studied the merits, if any, of different equivalent filter shapes.

If B is the highest recorded frequency, then with a rectangular gate, there will be a maximum of $N = BT$ cycles in the gate. We shall call N the *time-bandwidth product*. From equation (8) we see that the total number of resolved frequency intervals is

$$\frac{B}{\Delta f} = BT = N \tag{9}$$

At present, time-bandwidth products of 10^6 have been achieved and 10^7 is believed to be possible.[2]

A single spectrum analyzer having $N = 10^7$ could resolve a 100 MHz band into 10^7 channels each 10 Hz wide (provided the recording process could handle a 100 MHz band). Since we will probably want higher resolution than 10 Hz we will need several analyzers for each IF signal. Using bandpass filters and mixers we can divide each 100 MHz IF band into m basebands each $100/m$ MHz wide. Thus, if we wish 1 Hz channels we choose $m = 10$ and use 10 analyzers each fed with a 10 MHz signal. For 0.1 Hz channels we choose $m = 100$ and use 100 analyzers.

Recording Materials

Any material whose optical transmission over its surface can be varied in magnitude or phase could, in principle, be used. However, just as phase or frequency modulation of a carrier produces higher order sidebands, so a sinusoidal phase grating diffracts light into several orders. If more than one frequency is present, intermodulation products are also generated. (These can be avoided by keeping the modulation index small, but this sacrifices dynamic range.) Since both phenomena represent spurious spectral responses, phase materials appear undesirable for Cyclops.

Several new recording media such as magneto-optic films, ceramics, and photochromics are being developed. So far as could be determined, all of these materials have serious shortcomings such as poor resolution, low diffraction efficiency, and low sensitivity. The major advantages to be gained from these materials are reusability and the elimination of chemical processing.

At present, ordinary silver halide photographic film seems to be best suited to our purposes. However, high resolution diazo films deserve further consideration. They are grainless, optically smooth, require only ammonia vapor processing, and are very inexpensive. Since they are ultraviolet sensitive no dark room handling is required, but a powerful UV laser is needed in the recorder.

Film Usage

The rate of film consumption is determined by the total bandwidth B to be analyzed and by the resolving

[2] Private communication from Robert V. Markevitch of the Ampex Corporation

power of the film and associated optical system. If the overall resolving power is k line pairs per centimeter, then the rate of film usage is

$$R = \frac{B}{k^2} \text{ cm}^2/\text{sec} \tag{10}$$

which is simply the film area per second needed to record B Hz and is independent of the frequency resolution desired. If we use m analyzers to increase the frequency resolution by a factor m the film runs $1/m$ times as fast through each analyzer.

Taking B = 200 MHz (to account for both polarizations) and k = 3000/cm (which seems to be possible with modern lenses and slow film), we find $R \approx 22$ cm^2/sec. The usable width of 35-mm film is about 24-mm so the consumption rate would be about 9 cm/sec. At 4.5 cents/ft this amounts to 1.3 cents/sec or $47/hr.

While this is not at all an unreasonable cost, the economics of reusing the film stock should be investigated. In any event the silver should be totally reclaimed to conserve our resources of this truly precious metal.

Doppler Rate Compensation

The minimum channel width that can be used in the analyzer is fixed by the frequency drift rate ν of the received signal and is given by $\Delta f_{\min} \approx \sqrt{\dot{\nu}}$. If the local oscillators used to beat the IF signals down to the baseband signals for the spectrum analyzers have a drift rate that matches that of the received signal, the difference frequency will be drift free. Since we do not know the Doppler drift rate, we must anticipate a range of drift rates by providing banks of oscillators having drift rates separated by $2\Delta f$ over the interval $2\dot{\nu}_{\max}$. Each bank of mixers would feed a corresponding bank of spectrum analyzers. In this way the Doppler rate could be reduced (in one of the banks) by the ratio $r^2 = \dot{\nu}_{\max}/\Delta f$. This reduces the channel width by r, and requires r times as many analyzers per bank. Since there are r^2 banks a grand total of r^3 times as many analyzers is needed. Thus, the equipment cost mounts at a staggering rate.

The film usage per bank is unchanged but with r^2 banks the total usage increases by this factor. To try to reduce the channel to one tenth its initial value by this method would require 1000 times as many analyzers (20,000 in all) and would require a film usage of $4700/hour.

Because of the complexity and cost of achieving a significant bandwidth reduction through Doppler rate compensation, this approach is *not recommended* for Cyclops.

System Cost

Assuming no Doppler rate compensation, the number N_a, of optical recorder-spectrum analyzer combinations required to handle two IF signals each B Hertz wide is

$$N_a = \frac{2B}{N\Delta f} \tag{11}$$

where N is the time-bandwidth product of each recorder analyzer and Δf is the channel width, or frequency resolution. Taking B = 100 MHz, we obtain the cost data given in Table 11-1

TABLE 11-1

	$N = 10^6$		$N = 10^7$	
Δf	N_a	$ Million	N_a	$ Million
10 Hz	20	2	2	0.24
1 Hz	200	20	20	2.4
0.1 Hz	2000	200	200	24

The above cost figures are based on an estimated unit cost of $200 K for an $N = 10^6$ system and $240,000 for an $N = 10^7$ system.

Although the equipment cost is appreciable it is only 1/250 to 1/2000 as much as would be needed to do the job with hard-wired Cooley-Tukey transformers. Although the equipment cost of $200 million for 2000 processors having an $N = 10^6$ is still small compared with the antenna cost, as a practical matter the care and feeding of this many machines would require a large crew and represent a substantial operating cost. The analysis of 100 MHz bands into 0.1 Hz channels using $N = 10^7$ analyzers appears reasonable.

The cost of a film processor capable of processing six films in parallel with a time delay of 5 min dry-to-dry is about $18,000 or $3000 per spectrum analyzer. Since this cost is less than the uncertainty in the other figures, it has been ignored.

POWER SPECTRUM PROCESSING

The optical spectrum analyzer described in the last section provides a power density spectrum of the entire received signal on a (delayed) real-time basis. We now

need to process this spectral information in some automated system in a manner that will assure detection of any faint coherent signals that may be present and yet not give too many false alarms because of noise alone. In this section we will describe some possible ways of doing this processing.

The power spectra will show a Boltzmann distributed background intensity due to noise. This background will differ in detail from sample to sample but the statistics are stationary and the long-term average at any point will approach a constant value. If there is a strong coherent signal present, there will be a bright spot in the spectrum at the frequency of the signal. If the coherent signal is weak—comparable in power to the noise power in a resolvable interval—the intensity in successive samples will fluctuate, being sometimes stronger, sometimes weaker than the average noise brightness; but the average intensity over many samples will be greater than that produced by noise alone. Thus, if the coherent signals showed no frequency drift, a simple detection scheme would be to integrate the power spectrum for a sufficient time and then scan it for points having a greater intensity than that likely to be produced by noise alone.

Unfortunately, because of Doppler rates and inherent source instability, we must expect any coherent signal spot in the power spectrum to drift in position with time. This means that if we are to detect weak drifting signals we must accumulate the power spectra with successive samples shifted by a constant amount frame to frame corresponding to an assumed drift rate. Since we do not know the drift rate *a priori* we must do this accumulation for a range of assumed drift rates, positive and negative, up to the maximum rate expected (or up to the maximum rate that will allow full response in the spectrum analyzer). Because Doppler drift rates change slowly with time and because interstellar sources, particularly beacons, may be assumed to be inherently rather stable, we probably need only allow for *constant* drift rates during any observation period.

Our problem is illustrated nicely by Figure 11-7, which is actually a photograph of a pulsar pulse. Each scanning line in the picture is the power output of a receiver as a function of time. The receivers in adjacent scanning lines are tuned to adjacent frequency bands. Because of dispersion in the interstellar medium, the lower the frequency to which the receiver is tuned, the later the pulsar pulse arrives. If we cover all but one line at a time, it becomes evident that the pulsar would not be visible in the output of any single receiver. Yet its signature stands out clearly in raster of traces produced by all the receivers.

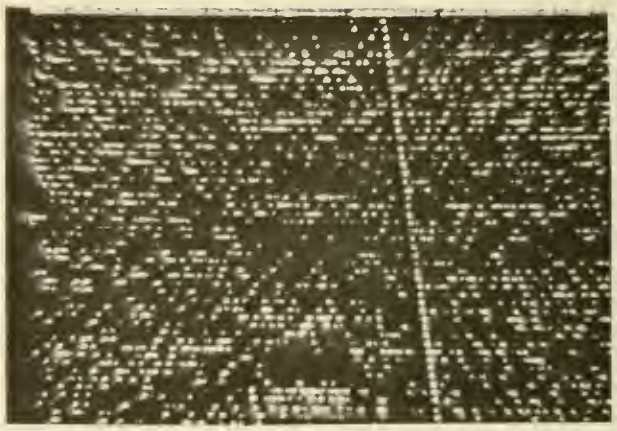


Figure 11-7. *Signature of a pulsar produced by simultaneous observation on adjacent frequency channels. (Photograph courtesy of Martin Ewing, Calif. Inst. of Technology)*

For our purposes we can take the photo to represent the intensity versus distance along a particular raster line, in the output of an optical spectrum analyzer, for successive samples of the power spectrum. The noise background has the same significance as before, but the "pulsar" is now a coherent signal drifting at a constant rate. Again it is evident that this signal could not be detected in any single sample of the power spectrum—that is, any single line of the photo—but the drifting trace stands out clearly in the raster of lines.

If the photo were to be scanned with a vertical slit, which moved horizontally across the picture, and the average brightness measured through this slit were plotted versus displacement, the resulting curve would show a small fluctuation about a mean value. So long as part of the signal trace were in the slit, the fluctuations might appear to be about a slightly higher mean value, but it is doubtful if this difference would be visible. If the slit were now tilted to be parallel to the signal trace and again scanned horizontally, the same sort of fluctuations about a mean value would occur until the trace lay entirely within the slit. At this point, there would be a large pulse in brightness clearly in excess of the normal noise fluctuations. By scanning the picture in this fashion we have converted the pattern, clearly visible to the eye, into a waveform clearly "visible" to threshold decision circuits. This is the principle we propose for use for the automatic detection of coherent signals.

Optical Processing

If one were simply to photograph successive samples of the power spectrum and arrange to project n such

samples simultaneously onto the same screen, any nondrifting coherent signal would stand out exactly as it would if a single frame had been exposed for the entire time. To detect drifting signals it is now necessary to displace the successive frames by a constant increment, corresponding to the drift rate, along the raster lines of the power spectrum. If all increments between appropriate negative and positive limits are tested any signal having a drift rate less than the limiting value will be detected.

The optical approach is conceptually simple and direct and makes use of the enormous data storage capability of photographic film. The film usage would be exactly equal to that of the optical spectrum analyzers. Unfortunately, we have not been able to find an elegant, nor even a completely practical embodiment.

Conceivably one might construct a column of, say, 100 projectors that would project 100 successive frames of a film into exact register on a screen. But to arrange to twist this column of projectors by various amounts so as to displace the images laterally along the raster lines with the accuracy needed seems very difficult.

A better method might be to pass the film with its 100 frames of spectra from a target star through a succession of perhaps 200 optical printers, in each of which a cumulative exposure is made on a single frame of film. In one of the printers, this single frame would be stationary during its exposure to the original 100 frames of power spectra, and the result would be the same as an integrated exposure at the spectrum analyzer. In all the other printers the single frame would be moved by different, constant increments along the raster lines between exposures to each of the 100 frames.

While the total film usage *rate* is now only three times as great as for the spectrum analyzer alone, we now have an enormous number of very slowly progressing films in process. Even if this could be done economically, a delay of several hours or even days is introduced in the detection process.

Film or processing defects (pinholes) in the final films would be a source of false alarms and, because of the delay, checking on false alarms would involve repositioning the array to the source.

Although we have discarded optical processing for these reasons, we concede that some clever technique, which may have eluded us, may yet be found.

Electrical Processing

If we are to process the power spectra electrically we must first convert them to electrical form. This can be done by letting one of the two power spectra produced by each analyzer fall on the target of a vidicon, or other type of TV camera tube. The stored image of the complete raster, produced by each frame of film in the optical analyzer gate, is then scanned and converted into a video signal. To avoid losing data from drifting signals at the ends of the power spectrum raster lines the scan should be extended slightly along the raster lines. This merely duplicates some information in the video signal.

The time-bandwidth product that can be used in the optical analyzer is more likely to be limited by the capabilities of the camera tube used than by anything else. An optical analyzer having $N = 10^6$ would produce a power spectrum having 1000 lines in the raster and 1000 resolvable elements per line. It is imperative that the vidicon have at least this high a resolving power to avoid degrading the images. Thus, tubes with a good deal more resolution than those used for commercial TV are needed.

The scanning circuits and waveforms must also be very precise, if the scanning lines are parallel to (and must therefore coincide with) the raster lines of the power spectrum. Probably a digital vertical sweep circuit or one that is servoed to the raster lines is needed. We have not explored these problems.

Data Storage System

Assuming the power spectrum has been converted successfully into a video signal, we could, in principle, record each sample of the complete power spectrum on one track of a magnetic tape loop. The next frame could be recorded on an adjacent track, and so on, until the samples for one observation time have been recorded. Now a single playback head having a gap long enough to span all the tracks would deliver a signal representing the integrated sum of all the tracks. By slowly rotating this playback head through a range of angles about the line perpendicular to the tracks, we could sum the power spectra with the relative displacements required to detect drifting signals.

If only one playback head were used, it would take at least twice as long to analyze the data as to record it. Thus a large number of heads, set at different angles, each feeding its own amplifier and threshold circuitry, should be used.

Several problems prevent us from seriously proposing the above system. First, it would be very difficult to lay down on tape the successive power spectra so that the same point in the spectrum occurred at exactly the same point along the tape on each track. Second, as the playback gap is turned at an oblique angle to the tracks a high-frequency loss occurs. Finally, the tape loop wear would be rapid and would present a serious maintenance problem.

We propose instead that the data be stored on magnetic disks equipped with flying (noncontacting) heads. With absolute filtering of the air and with constant operation (no shutdowns) disk wear is virtually absent, since the heads do not contact the disk surfaces. Such disks can record data in either analog or digital form. The former type is often used for instant replay of TV pictures while the latter is widely used for data storage in the computer field. The usual disk is 14 in. in diameter and can store on the order of 250 adjacent tracks of data with densities as high as 4000 bits/in. for digital data and 4000 Nyquist intervals per inch of analog data. Thus, each disk surface has a storage capacity of some 40 million bits of digital data or samples of analog data. Disks are built with moving heads that can be positioned in 10 to 20 msec to pick up or record a particular track, or with fixed heads, one per track. In the larger sizes the cost of moving head systems is about 0.007 cents per bit while for fixed head systems the cost is about 0.03 cents/bit. Head per track systems are thus about four times as expensive per data bit, but they offer the advantages of providing instantaneous switching between tracks and of having fewer moving parts. In a head-per-track system there is essentially only one moving part: a simple spindle with its payload of disks rigidly attached. Maintenance costs should be very low.

In some respects our application places *less* severe requirements on the recording system than other applications do. If we used digital recording we would need no more than 4 bits to describe the power spectrum amplitude. Similarly for analog recording we need no more than about a 20 dB signal-to-noise ratio. (Our signal is mostly noise anyway, and adding 1% to the noise power raises the system temperature only that amount.) Since we will be adding 100 or more tracks together to get the final signals, a defect in a particular track is of little importance. This is in contrast to computer applications where a dropped bit can be disastrous.

The various alternatives should be studied carefully before a decision is made among them for the final design. For the present, in order to get a rough cost estimate we shall assume analog recording on head-per-track disks.

Our data rate is fixed by the two 100 MHz IF channels, which produce 4×10^8 samples of analog data per second. Assuming this is recorded with fixed head disks at a cost of 0.03 cents per sample the cost of the data storage system will be

$$C = 4\times10^8 \times 3\times10^{-4} = \$120{,}000$$

per second of observing time. This figure is independent of how narrow our resolvable frequency intervals are or how many optical analyzers are needed. Thus, for 1000 seconds of observing time per star the cost of the data storage system will be about $120 million.

Digital Storage Systems

If the optical spectrum analyzers have rectangular gates, the video signal to be stored will be sharply band limited at about the bandwidth of the IF signal supplied to the spectrum analyzer. Thus, only slightly more than 2 samples/sec of spectrum must be taken per Hz of IF bandwidth. If 4 bits per sample are used to encode the signal, we will need about five times as many bits to store the data digitally as we need samples to store it in analog form. Thus, it appears that digital storage would be much more expensive at present.

The economics could change overnight, however. Magnetic "bubble" memories are being actively pursued in several laboratories, notably at Bell Telephone Labs. Conceivably these memories could ultimately bring the cost of digital storage down to 10^{-4} cents/bit or less. Bubble memories are suitable for *digital* storage only. They lend themselves most naturally to the construction of exceedingly long shift registers: registers with over one million bits per square inch of magnetic material. About 2×10^6 wafers could then store 1000 sec of data from both IF channels. At a cost of $10/wafer this would represent $20 million worth of storage capacity. Since they have no mechanical moving parts, bubble memories should have low maintenance and would be ideal for Cyclops.

Forming the Composite Signals

To process the stored data we propose that all samples of the power spectrum be played back simultaneously using the same heads that were used for the recording process. This eliminates all the major sources of timing errors and assures that the data corresponding to a given frequency in the power spectrum will appear simultaneously in each playback amplifier output. To permit summing the spectra with different relative displacements it is proposed to send the reproduced signals down video delay lines equipped with appropriate taps. Figure 11-8 illustrates the delay lines and tap arrangement required to synthesize 17 composite signals from 5 samples of the power spectrum, each composite signal representing an assumed drift rate. The pickup lines are merely indicated symbolically; actually, they would be coaxial or strip lines driven by current sources at each tap. At the center where all lines are driven by the same tap, a bridging amplifier with 17 independent

outputs is assumed.

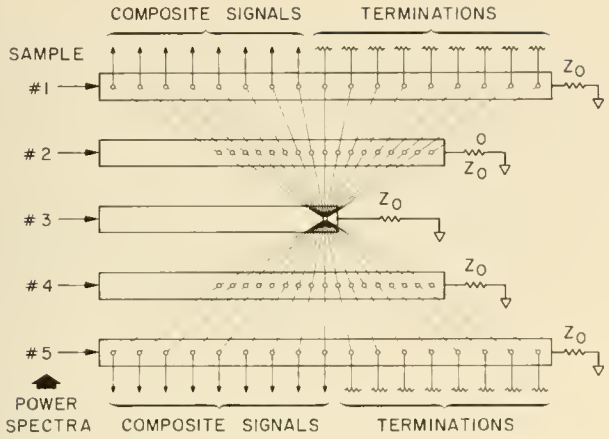


Figure 11-8. *Forming the composite signals.*

The delay line synthesis of the composite signals is believed to be more practical than attempting to pick these signals directly off the disks using an array of skewed heads. For one thing, the gap required in a skewed head on a disk cannot be straight but must be a short section of an Archemedean spiral. For another, the gap position where it crosses each track must be correct to within a few microinches. It seems unrealistic to expect such close tolerances to be maintained for years, even if they could be realized initially.

Lumped constant video delay lines are well suited to being tapped. The input capacitance of the bridging amplifier can be compensated for by reducing the shunt line capacitor at the tap, while the input conductance of the amplifier can be negligible. An analysis of lumped delay lines is given in Appendix P, where it is shown that with a properly chosen coefficient of coupling between adjacent coils on the line the delay can be held constant to ±0.03% up to one-fourth the line cutoff frequency. Over the same range, it is easy to terminate the line so that the reflection coefficient does not exceed 0.3%. At the cutoff frequency there is one half cycle (or one Nyquist interval) per section. We conclude that very satisfactory operation can be realized if there are four sections per Nyquist interval.

Number of Composite Signals Required

Assuming a rectangular gate in the optical spectrum analyzer, the video signals representing the power spectra will have a sharp upper cutoff frequency at B_0, where B_0 is the IF bandwidth handled by each analyzer. (Actually, if we overscan the raster lines to avoid information loss at the ends, the video bandwidth will be slightly greater than B_0). A coherent signal will produce a brightness distribution along a raster line given by

$$\frac{I}{I_0} = \frac{\sin^2 \pi x}{(\pi x)^2} \tag{12}$$

where x is the distance along the line measured in units of the resolvable frequency interval Δf, I is the intensity at x, and I_0 is the peak intensity. After scanning, x is time measured in Nyquist intervals.

Consider now the composite signal whose slope across the delay lines most closely matches the signal drift rate, and assume that *with respect to this composite signal* the received signal drifts by an amount x_0 between the first and last spectrum samples. The maximum output on the composite signal lines will occur when the signal pulse is centered on the tap on the middle delay line. At this instant, the signal pulses on the first and last delay lines are displaced by $\pm x_0/2$. Thus, the response, for a large number of delay lines (large number, n, of power spectra being summed) will be

$$a = k \int_{-x_0/2}^{x_0/2} \frac{\sin^2 \pi x}{(\pi x)^2}\, dx \tag{13}$$

where k is an arbitrary constant, as compared with a response

$$a_0 = k \int_{-x_0/2}^{x_0/2} dx = k x_0 \tag{14}$$

if the composite signal in question matched the drift rate exactly. The detection efficiency is therefore

$$\eta_0 = \frac{a}{a_0} = 2\,\frac{\mathrm{Si}(\pi x_0)}{\pi x_0} - \frac{\sin^2(\pi x_0/2)}{(\pi x_0/2)^2} \tag{15}$$

Figure 11-9 is a plot of η_0 versus x_0. We see that the response is down 1 dB ($\eta = 0.89$) for $x_0 = 0.66$. When we consider the cost of another decibel of antenna gain we quickly conclude we should not waste a decibel of signal at this stage, if we can avoid it. Thus, we will probably want the maximum value of x_0 to be no greater than 0.25, which gives $\eta_0 = 0.98$, and this requires the

composite signals to be formed at intervals such that $\Delta x = 1/2$ between the first and last spectrum.

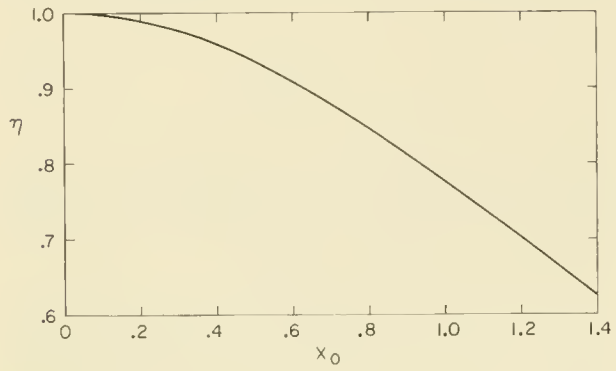


Figure 11-9. *Loss of sensitivity from rate mismatch.*

If we now assume that the maximum drift rate we will look for is one that causes a one Nyquist interval drift between successive spectra (this is one resolvable frequency interval Δf per scan, and causes some signal loss in the analyzers) and that we are adding n spectra to form the composite signals, then the number of composite signals needed is

$$n_c = 4n - 3 \tag{16}$$

The total number of taps, and hence the total number of independent bridging amplifier outputs, needed is

$$n_T = nn_c = 4n^2 - 3n \tag{17}$$

Number of Delay Line Sections

Under our assumption of four Nyquist intervals per section, the number of sections between the first and last taps on the first and last delay lines is $4(n-1)$. Including the terminating section on each end, the total length of the longest lines is $4n-2$ sections. The total number of sections for all lines (including terminating sections) is

$$\left.\begin{aligned} n_s &= 3n^2 \quad , n \text{ even.} \\ &= 3n^2 - 1 \quad , n \text{ odd.} \end{aligned}\right\} \tag{18}$$

Since n will be large we will take $n_s = 3n^2$ for all n. We note that for $n > 2$ the number of taps exceeds the number of sections, which simply means that for the central lines there are many taps supplying more than one pickup line.

Under our assumptions, the outermost lines will require a tap every other section. As we approach the central lines, there is an initial portion without taps and then a portion with a tap every section. Since the delays available at the taps are quantized, the rule is simply to connect each pickup line to the tap having the most nearly correct delay. There will thus be a delay error of $\pm 1/8$ Nyquist interval from this source and this will degrade η_0 as given by (15) to approximately

$$\eta \approx 1 - \sqrt{2}\,(1 - \eta_0) \approx 0.97 \tag{19}$$

Threshold Circuits

Associated with each composite signal line is an amplifier and threshold circuit. To compensate for gain variations that may exist in the system it is probably desirable to make this threshold "float"—that is, to make it a fixed multiple of the average noise level as determined at the amplifier in question. This refinement has not been investigated but no serious problems are foreseen.

Multiplexing

The power spectrum processing system outlined above, may, if analysis time permits, be multiplexed to analyze the outputs of m spectrum analyzers as shown in Figure 11-10. Each replay of the n tracks recorded in the data store takes a time of $1/\Delta f$ sec. If m analyzers are multiplexed as shown into a single spectrum processor, the total analysis time T_a per observation will be

$$T_a = \frac{m}{\Delta f} \tag{20}$$

Since no analysis can go on until the observing time T_0 is over and all tracks have been recorded, the analysis time adds to the total time per star, unless we go ahead and reposition the antennas in anticipation of a negative result. In this case we would like T_a to be less than the repositioning time. If the latter is, say, 10 sec then for 0.1 Hz channels ($\Delta f = 0.1$) we find $m = 1$, while for 1 Hz channels $m = 10$.

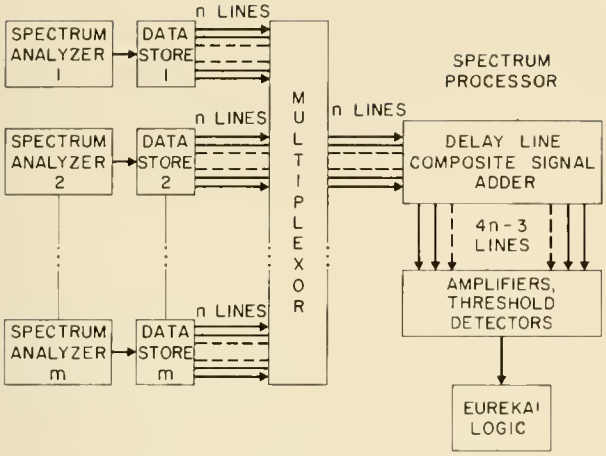


Figure 11-10. *Multiplexed spectrum processing.*

The cost of multiplexing is rather small. If the switching of the nm input lines to the multiplexor is done with a tree of transfer switches, the number required per analyzer – data store combination is less than $n(1 + 1/2 + 1/4 + 1/8 + \cdots) = 2n$. The cost is so much less than the uncertainty in the analyzer-data store cost that we will neglect it.

Cost and Cost Optimization

From the foregoing sections we can write the cost of the data processing system proposed for the detection of narrow band signals as

$$C = 4BT_0\gamma_b + N_a\left[\gamma_a + \frac{3n^2}{m}\gamma_s + \frac{4n^2 - 3n}{m}\gamma_t + \frac{4n-3}{m}\gamma_d\right] \qquad (21)$$

where

B = bandwidth of each IF channel

T_0 = observing time

N_a = number of optical analyzers

n = number of power spectra summed

m = multiplexing level

γ_a = cost per optical analyzer

γ_b = cost per bit or data sample stored

γ_d = cost per detector

γ_s = cost per section of delay line

γ_t = cost per tap (bridging amplifier)

Making the substitution $N_a = 2BT_0/Nn$, where N is the time-bandwidth product of each optical analyzer, we have

$$C = 2BT_0\left[2\gamma_b + \frac{1}{N}\left(\frac{\gamma_a}{n} + \frac{3n}{m}\gamma_s + \frac{4n-3}{m}\gamma_t + (4 - \frac{3}{n})\frac{\gamma_d}{m}\right)\right] \qquad (22)$$

As might be expected, the data processing cost is proportional to the IF bandwidth and to the observing time—that is, to the amount of data per batch. (We are happy that it is not some higher power of this number!) Making the following estimate of unit costs:

γ_a = \$100,000 ($N = 10^6$) and \$125,000 ($N = 10^7$)

γ_b = $\$2\times10^{-4}$

γ_d = \$50

γ_s = \$2/3

γ_t = \$1

we find from equation (22) that for T_0 = 1000 sec and B = 100 MHz:

1. $N = 10^6$ analyzers, 1 Hz Channels, n =1000, m =10

$C = \$2\times10^{11}[6\times10^{-4} + 10^{-6}(100 + 200 + 400 + 20)]$

$= \$[120 + 20 + 40 + 80 + 4]$ million

Data Storage	Optical Analyzer	Delay Lines	Delay Taps	Detector Circuits

= \$264 million

2. $N = 10^6$ analyzers, 0.1 Hz channels, n = 100, m = 1

$C = \$2\times10^{11}[6\times10^{-4} + 10^{-6}(1000 + 200 + 400 + 200)]$

$= \$[120 + 200 + 40 + 80 + 40]$ million

Data Storage	Optical Analyzer	Delay Lines	Delay Taps	Detectors

= \$480 million

3. $N = 10^7$ analyzers, 1 Hz channels, $n = 1000$, $m = 10$
$C = \$2\times10^{11}[6\times10^{-4} + 10^{-7}(125 + 200 + 400 + 20)]$
$= \$[120 + 2.5 + 4 + 8 + .5]$ million

Data Storage	Optical Analyzer	Delay Lines	Delay Taps	Detectors

$= \$135$ million

4. $N = 10^7$ analyzer, 0.1 Hz channel, $n = 100$, $m = 1$
$C = \$2\times10^{11}[6\times10^{-4} + 10^{-7}(1250 + 200 + 400 + 200)]$
$= \$[120 + 25 + 4 + 8 + .5]$ million

Data Storage	Optical Analyzers	Delay Lines	Delay Taps	Detectors

$= \$158$ million

Thus, in round numbers, the primary narrow band coherent signal detection system for Cyclops may cost anywhere from \$125 million to \$500 million, depending upon the time-bandwidth capacity of the analyzers and the channel width attempted.

We may not wish to let the channel width Δf be a floating variable to be fixed by cost minimization. Nevertheless, it is of some interest to minimize C as given by equation (22) by varying n. Setting $dC/dn = 0$ we find

$$n^2_{\text{opt}} = \frac{\gamma_a/m - 3\gamma d}{3\gamma_s + 4\gamma_t} = \frac{\gamma_a/\Delta f - 3\gamma d}{3\gamma_s + 4\gamma_t} \qquad (23)$$

For the same cost factors we have been using we find for an analysis time of 10 sec

1. 1 Hz channels, $N = 10^6$

$$n^2_{\text{opt}} = \frac{10^4 - 150}{2 + 4} = 1640$$

$$n_{\text{opt}} \approx 40 \quad \text{giving } T_0 = 40 \text{ sec}$$

2. 0.1 Hz channels, $N = 10^7$

$$n^2_{\text{opt}} = \frac{1.25\times10^5 - 150}{2 + 4} \approx 20{,}800$$

$$n_{\text{opt}} \approx 144 \quad \text{giving } T_0 = 1440 \text{ sec}$$

Thus it would appear that averaging somewhere from 50 to 150 samples of the power spectrum is appropriate so far as keeping the data processing costs in balance is concerned. Of course, case (2), above, which leads to a much longer observing time, also yields much greater sensitivity and is much more costly than case (1).

Possible Cost Savings

It is probably true that in no other area of the Cyclops system, except perhaps in the antenna structures, are there so many alternatives that deserve evaluation as in the data processing system. Data processing is a fast developing field and new techniques could obsolete the system we have described. Bubble memories are such a technique. But even in present technology there are other alternatives to consider.

For example, suppose that vidicons having 1000 sec or longer integration times could be made (or obtained by cooling present tubes). We could then split the output of each spectrum analyzer onto the targets of n vidicons. If each vidicon were translated along the power spectrum raster lines at the proper rate during the observation time, a simple scan of all these tubes would reveal the coherent signal we are seeking. No magnetic data storage, no intricate delay line. But a lot of vidicons!

Or suppose we simply make the pickup lines in the proposed delay line structure have the proper delay so that a signal taken from one end has a delay slope that differs from the signal taken from the other end by the amount between two adjacent pickup lines. We can then use the signals from both ends rather than wasting the signal at one end in a termination (See Figure 11-8). This reduces the number of taps and bridging amplifiers by a factor of two.

We mention these two examples—one fundamentally different, one a minor modification—to show how ephemeral our cost estimates are and how little we are prepared to defend the specific embodiment of the data processing system we have described. We felt compelled, in the time available, to present one feasible system and to cost it out, as a sort of existence proof. We are fully aware that further study may reveal ways to do the job better, or more cheaply, or both.

STATISTICS OF THE DETECTION PROCESS

Regardless of the details of its physical embodiment, and its cost, it is essential to know the detection capabilities of the proposed data processing system. The paramount question is: What received signal-to-noise ratio must we have to assure detection of a coherent

signal and yet have a tolerable false alarm probability?

The intensity of the light at each point in the output power spectrum raster of the optical analyzer is a measure of the noise power (in the signal sample in the gate) in a frequency band defined by

$$|K(f)|^2 = \left[\frac{\sin \pi(f-f_0)\Delta f}{\pi(f-f_0)/\Delta f}\right]^2 \tag{24}$$

where f_0 is the frequency associated with the point in question and Δf is the resolving power of the analyzer. The brightness out of the optical spectrum analyzer has exactly the same statistics as the amplitude out of a square law detector supplied with the same bandlimited signal. Thus, the results of Appendix D, in particular equations (D30) through (D41), apply to our detection process.

In Appendix D it is shown that if y is a normalized variable representing the average of n samples from a square-law detector, then the probability density function for y with noise alone as the input is

$$p_{no}(y) = n\frac{(ny)^{n-1}e^{-ny}}{(n-1)!} \tag{25}$$

The normalization of y is such that $\bar{y} = 1$; that is, the first moment of equation (25) is unity. By integrating this expression from y_T to ∞ the probability that y exceeds some threshold y_T is found to be

$$q_{no}(y_T) = e^{-ny_T}\sum_{k=0}^{(n-1)}\frac{(ny_T)^k}{k!} \tag{26}$$

This gives the probability *per datum* that noise alone will exceed a given threshold and thus produce a false alarm. Figure 11-11 shows $q_{no}(y_T)$ versus $10\log y_T$ for various values of the number n, of samples averaged.

In observing a single star we will have $2B/\Delta f$ independent data points per spectrum. Then we propose to average n spectrum samples in $4n-3$ ways to allow for drifting signals. We chose this number to avoid more than a 2 to 3% loss from improper drift rate match, but not all the composite signals so obtained are statistically independent. The effective number of independent composite signals is only n that is, is equal to the number of independent spectra that are averaged. Thus the number of independent data points per observation is

$$N_d = \frac{2B}{\Delta f}n = 2BT_0 \tag{27}$$

For B = 200 MHz and an observing time of T_0 = 1000 sec $N_d = 4\times10^{11}$.

The overall probability, p_{fa}, of a false alarm is given by

$$p_{fa} = 1 - e^{-N_d q_{no}} \tag{28}$$

and its cost is an increase in the observing time by a factor

$$k_{fa} = 1 + p_{fa} + p^2{}_{fa} + \cdots = \frac{1}{1-p_{fa}} \tag{29}$$

Combining equations (28) and (29) we find

$$q_{no} = \frac{\ell n(k_{fa})}{N_d} \tag{30}$$

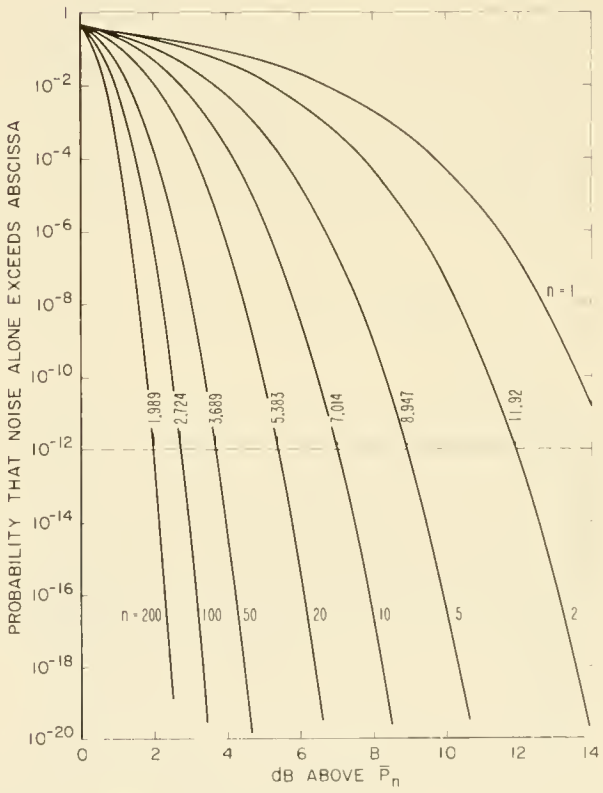


Figure 11-11. *False alarm probabilities.*

Assuming we can tolerate a 25% increase in search time because of false alarms, $k_{fa} = 1.25$ and we find $q_{no}(y_T) = 0.55 \times 10^{-12}$. We have used the value 10^{-12} for our calculations, but q_{no} is such a steep function of y_T that the error in our results is very small.

From Figure 11-11 we can now determine by how many decibels (referred to the input) the thresold y_T must exceed the average noise level to achieve the required false alarm immunity per datum point, for each value of n. For example, if $n = 10$, the thresold must be set at 10 log $y_T = 7.014$. We note that with increasing n the threshold may be set closer and closer to the average noise level $y = 1$ (or 10 log $y = 0$), and that the probability of a false alarm decreases more rapidly with increasing value of the abscissa.

When both a coherent signal and noise are present, the probability density function for y is given by

$$p_n(y) = n\left(\frac{y}{r}\right)^{\frac{n-1}{2}} e^{-n(r+y)} I_{n-1}(2n\sqrt{ry}) \tag{31}$$

where, as before, n is the number of independent samples that have been averaged, and r is the ratio of coherent signal power to noise power in the acceptance band (Δf) of the analyzer and I_{n-1} is the modified Bessel function of the first kind and of order $n-1$.

The probability that y lies *below* the threshold y_T and hence that a signal is not detected is given by

$$p_m(y_T) = \int_0^{y_T} p_n(y)\, dy \tag{32}$$

Since the integral is not known in closed form p_m was found by numerically integrating equation (31) for several values of r and for each value of n shown in Figure 11-11. The integration was performed on a Hewlett-Packard 9100B calculator and the curves were plotted on the associated 9125B plotter. Figure 11-12 is a sample plot for $n = 10$. The ordinate is p_m, the abscissa is 10 log y and the numbers on the curves are the input signal-to-noise power ratio expressed in decibels (that is, 10 log r).

Next the threshold value of 10 log y required for the assumed false alarm probability *per datum* of 10^{-12} (in this case for $n = 10$, 10 log $y_T = 7.014$) was determined from Figure 11-11 and drawn in. The values of p_m at

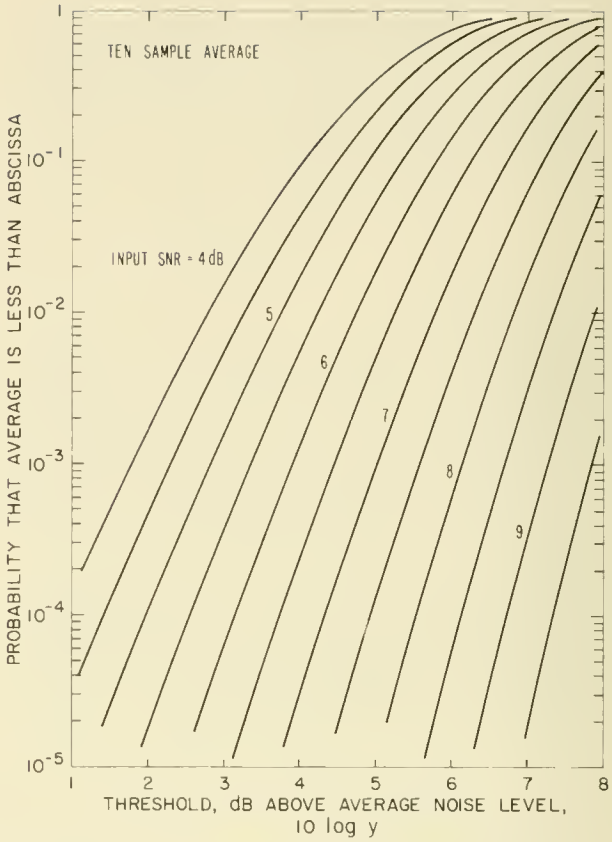


Figure 11-12. *Ten sample average of signal plus noise.*

this abscissa were then noted for each value of 10 log r, and a new curve (Figure 11-13) was then constructed showing p_m as a function of 10 log r at the threshold. Such plots were made for each value of n. From these latter plots the input signal to noise ratio required for any desired value of p_m can be found.

Finally, using the plots of p_m versus 10 log r, curves can be drawn showing the input signal to noise ratio required to assure a given value of p_m as a function of n, the number of samples averaged. Figure 11-14 shows two such curves for $p_m = 0.5$ and $p_m = 0.01$. The complete set of working curves used in preparing Figure 11-13 is included as Appendix Q.

From Figure 11-14 we see that, if we require a false alarm probability of 10^{-12}, then after 80 integrations there is a 50% chance of missing a signal when the received signal-to-noise ratio is unity (0 dB); after 150 integrations the probability of missing the signal is only 1%. Similarly, we see that for $n > 10$ an increase of only 2 dB or less in the received signal-to-noise ratio signal reduces the probability of missing the signal from 50% to 1%.

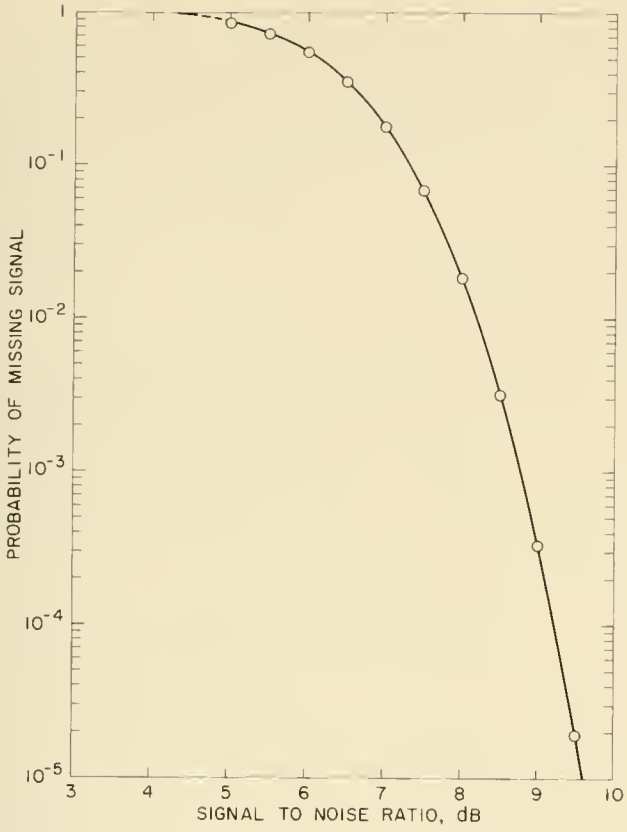


Figure 11-13. *Signal detection statistic, 10 sample average.*

In assessing these performance curves, we must remember that the signal-to-noise ratios shown are the ratio of received signal power to the noise power in one resolution bandwidth Δf of the analyzer. Thus, with $\Delta f = 0.1$ and $n \approx 100$, *we can detect a coherent signal that is 90 dB below the total noise power in the 100-MHz band.* That is the size of the needle we can find in any of the thousand to million haystacks we may have to examine. It should be noted that to decrease the false alarm probability from 20 to 2% requires, for $n = 100$, that the threshold be raised only 0.1 dB. (See Figure 11-11.)

For large values of n equation (25) approaches a gaussian distribution with the mean value $\bar{y} = 1$ and a standard deviation $\sigma_0 = 1/\sqrt{n}$. Likewise, equation (31) approaches a gaussian distribution with a mean value $r + 1$ and a standard deviation $\sqrt{(2r + 1)/n}$ (see Appendix D). If the distribution were gaussian for all n, the required signal-to-noise ratios would be as shown by the dashed lines in Figure 11-14. While the actual curves do approach the gaussian approximations in the limit as $n \to \infty$ the difference is larger than might be expected for large values of n. The reason is that in requiring a p_{fa} of 10^{-12} per datum we are far out on the tail of distribution equation (25), where the convergence is much slower than near the mean. When we require $p_m = 0.01$ rather than 0.5 we are also on the tail of distribution equation (31), which has a partial compensating effect.

If we wish the greatest smoothing for a given number, n, of spectrum samples, the samples should be *independent*, as we have been assuming. On the other hand, if we wish the greatest smoothing for a given observation time, T_0, and can let n increase without limit, we should average all possible samples. This means letting the film move continuously through the analyzer gate and integrating the power spectrum (with an appropriate set of drift rates) for the time T_0. As shown in Appendix D the variance about the mean with noise alone is reduced by the factor 2/3 as compared with averaging discrete adjacent samples for the same time. This is equivalent to an input noise power reduction of $\sqrt{2/3}$ or 0.88 dB. The variance with signal and noise present is also reduced, but by a smaller amount, which we have not determined.

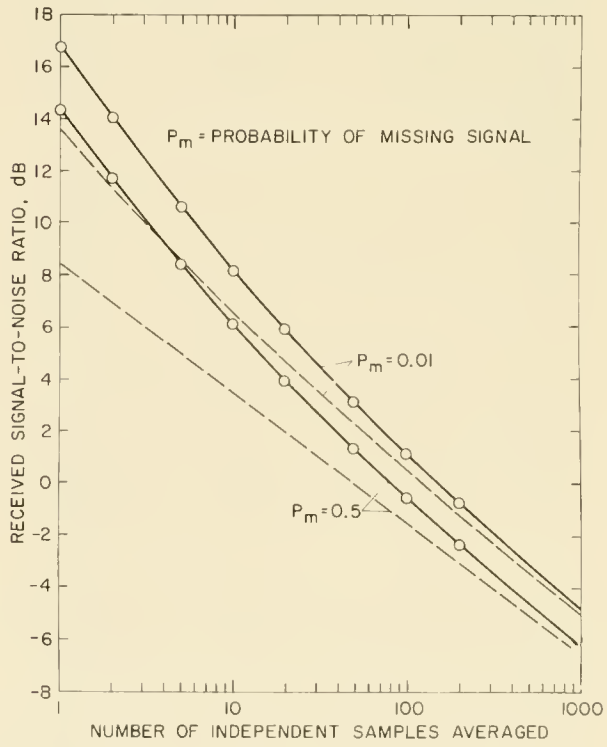


Figure 11-14. *Signal-to-noise ratios required for detection (false alarm probability per datum = 10^{-12}).*

Thus, with a rectangular gate in the optical analyzer, continuous averaging, if it could be implemented, would give about 1 dB improvement over the performance of the proposed system shown in Figure 11-14. Most of the improvement could be realized by doubling n and pulling the signal recording through the gate only

one-half frame between samples. Even though this method would substantially increase the cost of data processing, it might be a less expensive way of gaining almost 1 dB of performance than an increase in antenna area would be, and should be considered.

LOW FREQUENCY TIME ANOMALY DETECTOR

The spectrum analyzer discussed in the previous section is designed to detect only narrow-band signals, which change frequency very slowly if at all. If the signal is somewhat broadened (perhaps due to modulation) or sweeps in frequency, or pulsates with time, it may not be detected by the system. An example of such a signal is that generated by a pulsar. The pulsar radiates very wideband pulses typically lasting for 0.01 sec, repeated every 1 sec. Because of interstellar dispersion the Earth arrival time of these pulses varies with frequency, so we observe them as pulses whose frequency sweeps rapidly downward with time at a rate of a few MHz per second.

To detect this class of signals, one should examine as wide a frequency range as possible, with moderate channel bandwidth, short integration times, and search for time variations. Following is a possible method for doing this.

The entire frequency range available at any given time for analysis is limited by the Cyclops IF bandwidth to 100 MHz. An optical spectrum analyzer identical to those described in the previous section can be operated to cover the entire 100 MHz band with a channel bandwidth of 100 Hz and an integration time of 0.01 second. Channel bandwidth can be increased and integration time decreased as long as their product remains unity.

Although only one spectrum analyzer is needed per IF band for this coarse spectral analysis, the total data rate is not reduced. The recorder associated with this analyzer would have to handle 100 MHz of bandwidth and the film usage rate would be as great for this single unit as for the entire bank of analyzers in the narrow-band analyzing system. Also the data storage required would be the same, if the entire observation period were to be processed.

On the other hand, we could simply scan the output of the analyzer at 100 frames/sec with a vidicon and look for pulses or sweeping signals visually or with a set of threshold detectors. Or we might simply wish to photograph the power spectrum. Such a detector might prove very helpful in searching for pulsars or in spectral line work in radio astronomy.

If we do not try to store all the data magnetically, the cost is rather small: perhaps $300,000 for two analyzers and the associated video cameras, display tubes, and film cameras.

DRAKE ENSEMBLE SIGNAL DETECTION

The Drake technique (ref.10) is a means of detecting a large number of narrow band signals (an "ensemble"), where any individual signal alone may be too weak to detect. It makes use of the fact that the cross-correlation function of independent gaussian random signals has zero expectation for all values of shift. If, however, there is a small common component in the two signals (not necessarily detectable in either one by itself), the cross-correlatiion may show a peak at zero shift.

To use the method, the power spectrum for a frequency band suspected to contain signals is measured at two successive times. The two power spectra are then cross correlated. If a peak appears at zero shift, then there is a common component in the two spectra, and the detection of an ensemble of signals is established. This method does not yield the number of individual signals, or their exact frequencies.

The leakage radiation from a planet may well consist of a large number of fairly narrow-band signals. It may not be possible to detect any of them singly, but with this method they can be detected collectively or at a greater distance than any one signal.

Drake gives the detection condition for the ensemble method as

$$\frac{M}{\sqrt{N}} > \frac{1}{R^2} \tag{33}$$

R = signal-to-noise ratio

N = total number of receiver channels

M = number of channels in which signals are present

If we let $F = M/N$ be the fraction of channels occupied by a signal, the detection condition becomes

$$R > F^{-1/2} N^{-1/4} \tag{34}$$

Since the corresponding detection criterion for conventional detectors is $R > 1$, the detection threshold improvement factor for the Drake detector relative to a conventional detector is given by

$$S = \frac{R_{\text{conventional}}}{R_{\text{Drake}}} = F^{1/2} N^{1/4} \tag{35}$$

Stated another way, this is the sensitivity increase of the Drake detector over a conventional detector. It is

plotted in Figure 11-15 for various values of P and N. Note that when the number of channels is large, such as in Cyclops (perhaps 10^9), very substantial gains can be achieved. However, the leakage signal density must also be relatively high. In Figure 11-15 all signals are assumed to be of the same strength.

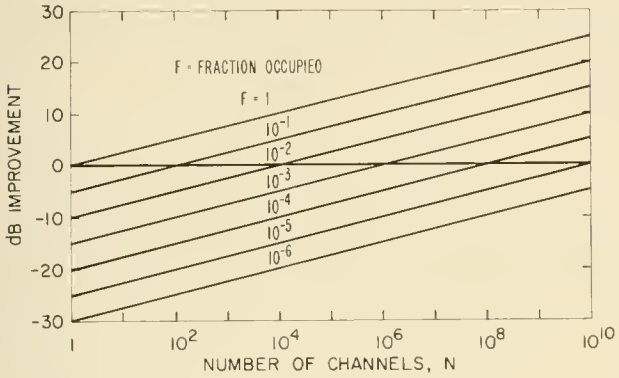


Figure 11-15. *Performance of Drake ensemble detector.*

It is of interest to determine if the Drake method would help to detect Earth leakage at great distances. Consider first the case of our UHF television band. The frequency range is 470 to 820 MHz, for a total width of 420 MHz. Each channel occupies 6 MHz and there are two signals per channel (one sound and one video). Both signals are wide-band, nevertheless much of the power is contained in relatively narrow-band carriers. Allowing for frequency tolerance of various stations on the same channel, assume a receiver channel bandwidth of 4 kHz. The number of channels is then

$$N = \frac{420 \times 10^6}{5 \times 10^3} = 8.4 \times 10^4$$

The number of signals (assuming all channels are occupied, which is not unreasonable in the eastern United States) is

$$M = \frac{420}{6} \times 2 = 140$$

which leads to

$$F = \frac{140}{8.4 \times 10^4} = 0.0017$$

Therefore, the sensitivity increase is

$$S = (0.0017)^{1/2} \times (8.4 \times 10^4)^{1/4} = 0.70$$

Thus, given the assumptions made, Drake's method is only 0.70 as sensitive as conventional methods.

A second example is our FM broadcast band. Making similar assumptions and calculations leads to an S factor of 1.25, indicating a superiority of Drake's method for that case.

Drake has investigated cross correlating more than two power spectra, and concludes that no advantage is obtained. If more than two spectra are available, they should be averaged in two groups, and the final pair cross correlated. The two groups should be interleaved in time, to reduce nonstationarity effects. This could readily be done by integrating on two photographic films alternate frames of the power spectrum from an optical analyzer having the appropriate resolution.

Since the additional hardware, software, and time required are small, Drake's method should be used in addition to the proposed narrow-band spectrum analysis method described in the last section. After the *entire* spectrum of a given target has been measured with the spectrum analyzer for individual signals, this *entire* spectrum should then be cross correlated for an ensemble of weak signals. In general, use of the entire spectrum, rather than each segment, will increase the sensitivity. All channel bandwidths need not be equal nor be equally spaced for the method to work. In the presence of irremovable terrestrial interference, Drake's method will always indicate a signal. This problem can be avoided by notching out known terrestrial signals and cross correlating the remaining spectra.

WIDE BAND IMAGING OF THE RADIO SKY

Conventional radio telescopes do not produce an image of the radio sky in the same sense that an optical telescope does; they merely measure the total radiation intensity received from all sources within the beam. The term *radio telescope* is a misnomer; *tele-radiometer* would be more accurate. Maps of the radio sky are presently made in one of two ways. Either a single radio telescope beam is scanned over an area of the sky in a raster and the image is synthesized from the elemental brightness readings as in television, or, for higher resolution pictures, two or more radio telescopes are used as an interferometer. By cross-correlating the telescope outputs, taken in pairs, we can find the amplitudes and phases of the Fourier components of the

brightness distribution over the common field of view of the antennas. Finally, the two-dimensional Fourier transform of the measured data is taken to get the desired map. The proposed VLA system will use this technique, called *aperture synthesis*. It is obvious that neither of these methods produces a real-time image of the sky.

The greater the resolution desired, the greater is the time required to map a given area of the sky. With the resolution available from the Cyclops array the times are very long indeed. Assume that we have 1000 dishes spread over an area 10 km in diameter, and operating at a wavelength of 10 cm. If we allow 3 dB gain falloff at the edges of each elemental area scanned, the number of resolvable directions is about $1.5\ (\pi d/\lambda)^2 = 1.5\times10^{11}$ (see Chap. 6). With an integration time of only 1 sec per elemental area scanned, it would take 1.5×10^{11} sec or about 5000 years to map the entire sky. Even to map an object such as M31 (the Andromeda galaxy) would take about 126 eight-hour observing days, or 4 months. But with 1000 elements in the array, we can form 1000 independent beams simultaneously, and thus map the whole sky in 5 years, or M31 in an hour.

Even with an imaging system, the field of view is limited by the beamwidth of the antenna elements. Moreover, in an array, there are additional grating lobes within this field that confuse the picture and reduce the usable field area by the filling factor. Thus, for the Cyclops array with a filling factor of 1/10, ten times as many fields of view would be needed to synthesize a picture of a given region as would be needed with a single 100-m dish. The advantage of the array is that the final picture contains 10,000 times the detail. Thus, the array with its 1000 dishes is 1000 times more powerful for mapping, as one should expect. In fact, if blurred areas in the map made with the single antenna are resolved into sharp points by the array, the integration time can be reduced, making the array even more powerful.

General Principles

To simplify our thinking let us initially assume that the Cyclops array is pointed at the zenith. A signal received from the zenith will then produce the same IF output signal at the central station from every antenna element.

Let us take Y to be the north-south axis of the array and X to be the east-west axis as shown in Figure 11-16a. A signal received at a small angle δ from the zenith and at an azimuth α will be received by an antenna at x,y with a delay.

$$\tau_r(x,y) = -\frac{x \sin\alpha + y \cos\alpha}{c} \sin\delta \qquad (36)$$

where c is the velocity of light. At the received frequency ω_r this delay causes a phase shift

$$\phi_r = \tau_r \omega_r \qquad (37)$$

The delay and, if the spectrum has not been inverted, the phase shift are preserved by all the heterodyning operations and thus are present at the IF outputs.

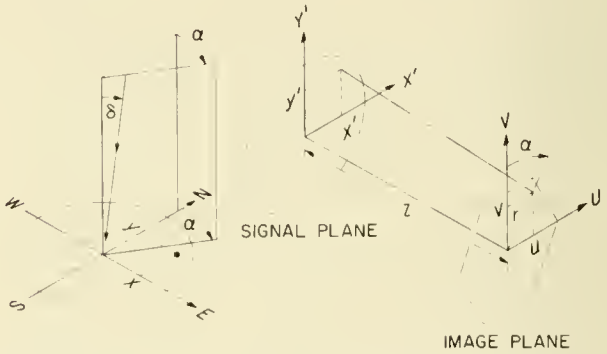


Figure 11-16. *(a) Antenna and source coordinate. (b) Imaging coordinates.*

Let us imagine the IF outputs to be arranged physically in a plane, forming a small scale map of the array, so that in Figure 11-16b $x' = \sigma x$ and $y' = \sigma y$ where σ is the scale factor. Now let us arrange a second lattice of points with coordinates u, v in a second plane, which we will call the image plane, and let us connect each of the IF outputs in the signal plane to every lattice point in the image plane via a transmission path having a phase shift

$$\phi_i = -k(ux + vy) \qquad (38)$$

At the image plane point at radius r and at angle α, the phase will be $\phi_i = -k_r\ (u \sin\alpha + v \cos\alpha)$ while at the angle $\pi + \alpha$ the phase will be

$$\phi_i = k_r\ (u \sin\alpha + v \cos\alpha) \qquad (39)$$

All the IF signals will arrive in phase if $\phi_r + \phi_i = 0$, that is, if

$$k_r = \frac{\omega_r}{c} \sin \delta \tag{40}$$

or, since $\delta << 1$, when

$$r = \frac{\omega_r}{kc} \delta = F\delta \tag{41}$$

The factor $F \equiv \omega_r/kc$ thus determines the scale of our map and is the effective focal length of the imaging system. If k is positive, as we have assumed, then F is positive and the image is *inverted;* that is, points at azimuth α are mapped at $\pi + \alpha$.

How large a field can we map? If the antenna elements are in a square lattice (in rows and columns separated a distance s), then we can substitute sa for x and sb for y in equation (38), where a and b are integers. It is then clear that if u or v is incremented by $2\pi/ks$, ϕ_i is unchanged. The imager thus produces a lattice of responses separated by

$$\Delta u = \Delta v = \frac{2\pi}{ks} \tag{42}$$

Since $\Delta u = \Delta(r \sin \alpha)$ and $\Delta v = \Delta(r \cos \alpha)$, and since equation (41) can be written as

$$(\delta \sin \alpha) = \frac{kc}{\omega_r} (r \sin \alpha) \tag{43}$$

or

$$(\delta \cos \alpha) = \frac{kc}{\omega_r} (r \cos \alpha) \tag{44}$$

these spurious responses correspond to

$$\Delta(\delta \sin \alpha) = \frac{2\pi c}{s\omega_r} = \frac{\lambda_r}{s} \tag{45}$$

and

$$\Delta(\delta \cos \alpha) = \frac{2\pi c}{s\omega_r} = \frac{\lambda_r}{s} \tag{46}$$

which are the grating responses of the original array; λ_r is the wavelength of the received radiation.

Thus, the image that will be formed is the convolution of the true brightness distribution of the radio sky with a lattice of δ functions representing the grating lobes of the antenna array. The side lobes are suppressed by the directivity of the antenna element, but since $s \approx 3d$ this factor is not very small for the first few lobes. Later we shall discuss how the image may be processed to remove or reduce this confusion. For the present we merely note that, since the recovered image is periodic in u and v, we need only image the central square $|u| < \pi/ks$, $|v| < \pi/ks$. If the outline of the array were square, the resolvable points within the image would be separated at a distance $2\pi/kS$ where S is the side of the array. Thus, the number would be $(S/s)^2 = n$, the number of elements. With a circular outline the shape of the figure of confusion changes, but we may still take n to be the number of independent image points.

For an array with a hexagonal lattice, the grating lobes also form a hexagonal array and the unit field becomes a hexagon whose sides are $\pi/\sqrt{3}ks$ and whose area is thus $2/\sqrt{3}$ times as great as for the square lattice. But since the array area is only $\sqrt{3}/2$ times as great in the hexagonal cases, the figure of confusion is $2/\sqrt{3}$ times as large in area so the number of image points in the field remains n. The hexagonal array has the advantage or more nearly isotropic mapping.

To avoid the spurious pattern in the image caused by sampling the intensity distribution at discrete points, we may well wish to increase the sampling density in both the U and V directions. If, for example, we double this density in both directions, we will have $4n$ points in the image with considerable correlation between adjacent points. Then by blurring this image only slightly we can eliminate the discrete structure without loss of real detail.

If $A(x_i,y_j)$ is the complex amplitude of the IF signal from the antenna whose coordinates are x_i,y_j we see from equation (38) that the complex amplitude $a(u,v)$ at a point u,v in the image plane is

$$a(u,v) = \sum_{i,j} A(x_i,y_j)\, e^{-ik(ux_i + vy_j)} \tag{47}$$

and is the (discrete) two-dimensional Fourier transform of $A(x_i,y_j)$. Because y is constant during the summation over i (or, alternatively, x is constant during the summation over j), equation (47) can be written as a two-step process

$$a(u,v) = \sum_j e^{-ikvy_j} \sum_i A(x_i,y_j)e^{-ikux_i} \quad (48)$$

Radiative Imaging

One obvious way to implement equation (47) is to *radiate* the IF signals from the signal plane and to receive this radiation in the image plane, in exact analogy to the optical spectrum analyzer. The waves radiated may be acoustic, as suggested by Oliver (ref. 11) and McLean and Wild (ref. 12), or electromagnetic. If the signal and image planes are separated a distance ℓ, several alternatives exist for bringing the waves to focus in the image plane:

1. We may interpose a lens of focal length $\ell/2$ midway between the planes.
2. We may interpose two lenses one of focal length ℓ (the focussing lens) in front of the signal plane and the other of focal length $\ell/2$ (the field flattener) in front of the image plane.
3. We may curve the signal and image planes into appropriate spherical surfaces.
4. We may delay the signals to the central elements so as to radiate a spherical wavefront from a plane array.

With any of these alternatives, the phase shift produced by the propagation delay between points on the two surfaces will be a constant ($-2\pi\ell/\lambda_i$ for alternative 3) plus a term

$$\phi_i = -\frac{2\pi}{\ell\lambda_i}(ux' + vy')$$

$$= -\frac{2\pi\sigma}{\ell\lambda_i}(ux + vy) \quad (49)$$

where λ_i is the wavelength of the radiation used.

Comparing equations (49) with (38) we see that

$$k = \frac{2\pi\sigma}{\ell\lambda_i} \quad (50)$$

so that the effective focal length equation (50) can now be written

$$F = \frac{f_r\lambda_i\ell}{\sigma c} = \frac{\lambda_i}{\lambda r}\frac{\ell}{\sigma} = \frac{f_r}{f_i}\frac{v}{c}\frac{\ell}{\sigma} \quad (51)$$

where v is the velocity of propagation of the waves used.

We now wish to point out a fundamental limitation of all radiative imaging processes. If f_i is obtained by heterodyning f_r, then $f_i = f_r - f_0$ where f_0 is a constant frequency. Then equation (51) becomes

$$F = \frac{v}{c}\frac{\ell}{\sigma}\frac{f_r}{f_r - f_0} \quad (52)$$

As f_r varies from one end of the IF band to the other, the fractional variation in $f_r - f_0$ is greater and the effective focal length F changes with frequency. Low IF frequencies are imaged with more magnification than high IF frequencies. Thus, a wide-band source will be imaged not as a point but as a radial line whose intensity profile is the power spectrum (versus wavelength) of the source. This may have its uses, but it does not produce good images. By analogy with a similar defect in lenses, we shall call this phenomenon *lateral chromatic aberration*. There appears to be no way to avoid lateral chromatic aberration other than to make f_0 zero. This means doing the radiative imaging at the original RF frequency.

We then have yet another problem: Unless adequate shielding is provided, the electrical signals generated for imaging could be picked up by the antennas, thereby producing serious feedback. Unless we are careful we might end up with the world's most expensive oscillator. Because the antennas need not be aimed at the imager and because the array is completely dephased for any nearby signal, we are really concerned only with the far out side lobe response of the nearest elements. Hopefully, this can be 20 dB or more below that of an isotropic antenna. Nevertheless, careful shielding is essential.

If we are forced by the feedback problem to use a frequency offset, some chromatic aberration will remain. To be effective in suppressing feedback, f_0 must be at least equal to the system bandwidth B. If f_C is the center frequency of the RF band and $f_0 = -B$ (that is, we actually use an upward offset), and if we let $x = B/f_0$ then from equation (52) the fractional change in F is

$$\frac{\Delta F}{F} = \frac{x^2(1+x)}{\left(1+\frac{3x}{2}\right)\left(1+\frac{x}{2}\right)} \approx \frac{x^2}{1+x} \qquad (53)$$

If we wish to hold $\Delta F/F$ less than 1%, then $x < 0.105$ which requires $f_0 > 9.5B$. For a 100 MHz band the operation would be satisfactory above 1 GHz.

Delay Line Imaging

Instead of radiating the IF signals, we can transmit each one to all image points via a set of cables. If we simply make the cable delays equal to the path length delays in a radiation imager (plus an arbitrary constant delay), the operation of the delay line imager will be the same as a radiative imager. In producing the required phase shifts, we will be delaying the IF signals by f_r/f_i times the RF delay and thus be producing a delay error

$$\Delta\tau = \left(\frac{f_r}{f_i} - 1\right)\tau_r(x,y) \qquad (54)$$

This can also be written

$$\Delta n_i = \left(1 - \frac{f_i}{f_r}\right) n_r \qquad (55)$$

where n_r is the number of cycles of delay at the received frequency and Δn_i is the delay error expressed in number of cycles of the imaging frequency, f_i.

At the edge of the useful field of view $n_r = 1/2$ between adjacent elements in a square lattice, and for a circular array, there are $\sqrt{4n/\pi}$ rows or columns across the array. The number of cycles of delay from the center to the edge is therefore

$$n_r = \frac{1}{2}\left(\sqrt{\frac{n}{\pi}} - \frac{1}{2}\right) \qquad (56)$$

For $n = 1000$, $n_r \approx 8.6$, and from equation (55) we see that if $f_i \ll f_r$, there will be about eight cycles of delay error at the imaging frequency.

When the signal is wide band noise, this delay error seriously decorrelates the signals received from the various antennas, so that the addition of *amplitudes* no longer occurs with full effectiveness. Because the delay error is systematic, the decorrelation manifests as the lateral chromatic aberration already discussed. The two effects are equivalent in the following sense. When a field of mutually incoherent sources is being viewed, the lateral chromatic aberration spreads the images of all points radially; thus, the signal at any one image point is the sum of the uncorrelated signals from all the radially adjacent object points that have a part of their power spectrum imaged at the point in question. If there is a single object point off axis in a dark field, its image, being spread into a radial line, is dimmer at all points by spreading ratio.

In imaging systems employing delay lines there are several ways of decreasing the delay error.

1. We can take advantage of the fact that equation (40) need only be satisfied modulo 2π. This allows the delay lines to match the corresponding RF delays to $\pm s/2$ cycles of the imaging frequency, where s is the number of steps taken in the transformation. If $s \leqslant 2$ this will cause only a small decorrelation loss if $B/f_i < 1$ (see Appendix D). The trouble with this method is that the delay corrections (the delay departures from the RF delays) needed to obtain proper phasing are frequency dependent. Thus, all the IF delay lines would need to be readjusted whenever the receiver band was changed. This is hardly practical.
2. We can set the IF delays precisely equal to the corresponding RF delays and add a phase shifter per cable. The phase shift could be introduced in the pilot signal used to demodulate IF signals to baseband before imaging, so the phase shifter need not be a broadband device. Nevertheless, the cost per unit would probably exceed $250 and with over 200,000 cables involved, the total would be on the order of a half billion dollars. Also the system is very complicated and therefore requires considerable maintenance.
3. We can use an analog fast Fourier transform (Butler matrix) system. The FFT can be implemented using delay lines to accomplish the multiplication by $e^{i\phi}$ and is applicable to the two-dimensional case. It offers the practical advantage of reducing the number of delay lines from a number on the order of $2n^{3/2}$ (setting $m = n$ in the results of the next section) to $2n \log_3 n$. The principal drawbacks of the approach are that (a) in achieving the right phase shifts along each path, no account is taken of the RF delay, (b) for large n the IF delays are different for different paths having the same phase shift and by an amount that is difficult to predict, and (c) like alternative (1), the delay lines must be readjusted when the RF band is changed. Also the pattern of interconnections is extremely complex and difficult to map in an orderly layout.

We conclude that none of the above alternatives is attractive, and the best way to avoid chromatic aberration in a delay line imager is to image at the original band, or near it.

If we wish to have m image points and if the n antennas in our array were scattered at random points, we would require mn cables to make all the connections in the imager. For $n = 1000$ and $m = 4000$ we need 4 million cables. However, if both the antennas and the image points are in regular lattice patterns, this number can be reduced by performing the transformation in two steps as allowed by equation (38).

The simplest case is that of a square antenna array of n elements arranged in $\sqrt{n}$ rows and columns, and a square image field of m image points arranged in $\sqrt{m}$ rows and columns. Between the signal and image plane we place an intermediate plane having $\sqrt{mn}$ junction points. If we transform first by rows and then by columns, the intermediate plane will have $\sqrt{n}$ rows and $\sqrt{m}$ columns. Each of the n signal plane points connects to the $\sqrt{m}$ intermediate plane points on the same row, while each of the n image plane points connects to the $\sqrt{n}$ intermediate plane points in the same column, for a total of

$$N = n\sqrt{m} + m\sqrt{n} \tag{57}$$

interconnections.

If the antenna array is circular with a square lattice there will be $\sqrt{4n/\pi}$ rows and therefore

$$n_j = \sqrt{\frac{4nm}{\pi}} \tag{58}$$

junction points in the intermediate plane. We then find

$$N = n\sqrt{m} + m\sqrt{\frac{4n}{\pi}} \tag{59}$$

The connections are illustrated in Figure 11-17 for the top row and left-hand column.

If we use hexagonal lattices, the image plane lattice is rotated 90° with respect to the signal plane lattice. Assuming the antenna array is circular and that in the image plane one of the three sets of rows, characteristic of a hexagonal lattice, is horizontal, there will be $\sqrt{8n/\pi\sqrt{3}}$ rows. The image field will be hexagonal in outline with two opposite sides of the hexagon forming the top and bottom, and will contain $(1 + 2\sqrt{4m-3})/3$ columns. There will therefore be

$$n_j = \frac{1 + 2\sqrt{4m-3}}{3}\sqrt{\frac{8n}{\pi\sqrt{3}}} \tag{60}$$

junction points and

$$N = n\,\frac{1 + 2\sqrt{4m-3}}{3} + m\sqrt{\frac{8n}{\pi\sqrt{3}}} \tag{61}$$

connections.

The values of n_j and N for the various configurations with $n = 1000$ and $m = 4000$ are shown in Table 11-2

TABLE 11-2

Antenna Array:		Junctions:		Connections:	
Shape	Lattice	Number	Ratio	Number	Ratio
Square	Square	2000	1.00	190,000	1.00
Circular	Square	2257	1.13	206,000	1.09
Circular	Hexagonal	3246	1.62	238,000	1.25

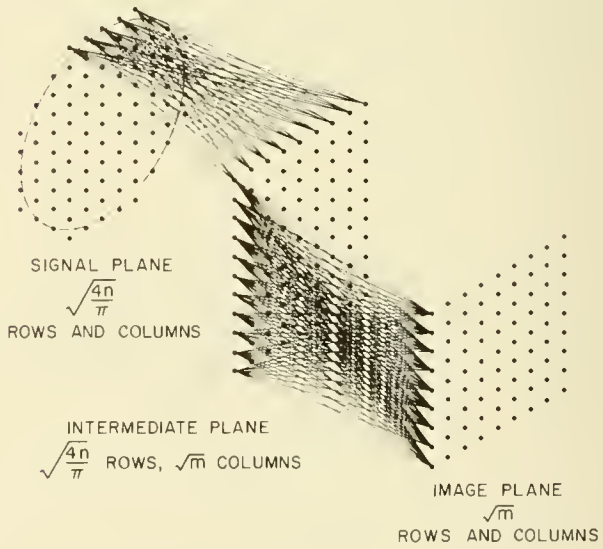


Figure 11-17. *Wild Cat's cradle.*

The cost of a delay line imager is

$$C = \gamma_s n + \gamma_j n_j + \gamma_i m + \gamma_c N \tag{62}$$

where

γ_s = unit cost of signal plane electronics

γ_j = unit cost of junction plane electronics

γ_i = unit cost of image plane electronics

γ_c = unit cost per cable

n_j = number of junctions

N = number of cables

For a 1000 element circular array in a hexagonal lattice, the maximum delay across the array at the corner of the image field is about 22 cycles. At an IF frequency of 1 GHz this is about 4.4 m of cable. Each receiver in the image plane must accept $\sqrt{4n}/\pi$ or about 34 inputs, which, because of the necessary connectors, requires a unit about 6-1/2 in. wide by 7-1/2 in. high. There are 88 such units across the field, so the image plane will be about 15 m across and 13-1/2 m high. If we separate the planes by 10 m, the longest cables will be 17 m and the shortest 12.6 m. Let us take the average length to be 15 m. We estimate the cost of cable at 15 cents/m, the cost of connectors at $1.50, and the cost of cutting to length and installing at $2.25 for an average cost γ_c = $6.

Taking the cost of the signal plane units at γ_i = $800 the cost of the repeaters at γ_j = $800 and the cost of the image plane receiver detectors at $500, we derive the costs given in Table 11-3.

TABLE 11-3

Antenna Array		Signal	Junction	Image		
Shape	Lattice	Plane	Plane	Plane	Cabling	Total
				$ Million		
Square	Square	0.8	1.6	2.	1.14	5.54
Circular	Square	0.8	1.8	2.	1.24	5.84
Circular	Hexagonal	0.8	2.6	2.	1.43	6.83

In spite of the maze of interconnections, the cabling is not the major cost. To get really accurate costs we need to refine our estimates of the electronics costs, which can only be done by designing the units.

The major operational disadvantage of the delay line imager is that the size of the useful field is inversely proportional to the RF frequency. If the field is proper at 1 GHz, only the central third (one-ninth of the area) is useful at 3 GHz. Unless we raise the surface density of the image points in the center of the field we will lose 90% of the detail.

Optical Imaging

A radiative imager using light could be made if there were some neat way to modulate the amplitude and phase of the coherent light transmitted at a small spot in a plate. We could them map the array as an array of such spots and modulate each in accordance with the amplitude and phase of the RF signal received by each antenna. But in addition to requiring techniques unavailable to us, such a method requires close tolerances and is afflicted with lateral chromatic aberration. Using light as the imaging radiation is equivalent to making f_0 in equation (52) very large and negative. The focal length is then proportional to f_r and the relative variation over the band is

$$\frac{\Delta F}{F} = \frac{B}{f_c} \tag{63}$$

To keep $\Delta F/F < 0.01$ with B = 100 MHz requires $f_C > 10$ GHz.

In a radiative imaging system, where the path delays change continuously with position across the signal and image planes, we can avoid the chromatic aberration only by using radiation at the original frequency. This leaves us with acoustic waves and microwaves to consider.

Acoustic Imaging

Piezoelectric transducers having only a few decibels of conversion loss over the frequency band from 0.6 to 1.8 GHz have been built using thin films of zinc oxide. Units to cover the band from 1 to 3 GHz appear possible with further development. Let us therefore examine the feasibility of acoustic imaging in this frequency region.

Since the waves are generated in a solid medium such as crystalline lithium niobate or sapphire, focusing lenses are probably not realizable. Instead we must either curve the signal and image array surfaces, or use RF delay to generate coverging wavefronts, or both. Figure 11-18 shows two concave arrays whose vertices are separated by a distance ℓ. With a solid medium this distance ℓ is fixed, and the most economical solution is not to use RF delays but simply to make the radius of curvature of the signal array r_s equal to ℓ. This places the center of curvature C_s at the vertex of the image array V_i. On-axis radiation will therefore focus at V_i. Off-axis radiation will focus on a spherical surface whose radius of curvature $r_i = \ell/2$. (That is, C_i will be located midway between V_s and V_i.) The image surface is not plane but is part of a Rowland sphere: a Rowland circle in both directions.

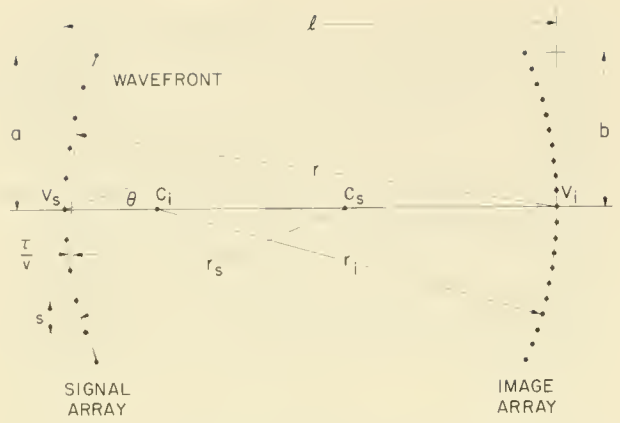


Figure 11-18. *Geometry of radiative imaging.*

For simplicity, we assume that the antenna array is circular and that the elements are in a square lattice. The transducers in the signal array are then also in a square lattice with a separation s. The results for a hexagonal array are not very different, but the algebra is more complicated. The radius a of the signal array is then

$$a \approx \left(\sqrt{\frac{n}{\pi}} - \frac{1}{2}\right) s \tag{64}$$

where n is the number of antenna or transducers. The image array we assume to be a square of sides $2b$. At the limits of the desired field

$$\sin\theta = \frac{\lambda}{2s} = \frac{b}{\ell} \tag{65}$$

where λ is the wavelength of the imaging radiation. Since this angle can exist in both dimensions simultaneously, the maximum angle θ_0 at the corners of the image is $\sqrt{2}$ times as large (for a square image array). From equation (65) we find

$$\ell = \frac{2bs}{\lambda} = \frac{b}{a}\left(\sqrt{\frac{4n}{\pi}} - 1\right)\frac{s^2}{\lambda} \tag{66}$$

The characteristics of some possible materials are given in Table 11-4.

Sapphire looks attractive because of its high velocity and its low loss, and because it can be made in long rods. Let us choose sapphire and assume an operating frequency of 1 GHz. Then the wavelength $\lambda = 11\mu$. If we

TABLE 11-4

Material	v(m/sec)	α(dB/m)(@ 1 GHz)
Fused silica	5,968	>2000
Quartz (X axis)	5,740	500
Sapphire (A axis)	11,100	20
Lithium niobate	7,480	30
Yttrium aluminum garnet	8,600	25

chose $s = 1/2$ mm, then $\theta_0 \approx 1.56\times10^{-3}$ radians, which is small enough to allow aberration free imaging. Then from equation (64) $2a = 1.75$ cm. If we assume $4n$ image points and make their spacing also be s, then $b = \sqrt{ns}$, and from equation (66) we find $\ell = 1.44$ m. The one-way loss is therefore about 29 dB.

If we double s to achieve a 1 mm transducer separation, ℓ will increase by a factor of four, making the loss about 115 dB. Thus, to achieve acoustic imaging at 1 GHz we must be able somehow to feed 1000 independent 1 GHz signals into 1000 transducers in a 2 cm diameter array without disturbing the acoustic match of these transducers to the medium on their front and back sides. We know of no way to accomplish this with present techniques. At higher frequencies the problems are even worse.

There are a host of other problems as well. For example, the size of the usable field decreases as we go up in frequency; the tolerances on the array surfaces are on the order of 0.3μ; and wall reflections must be absorbed. Microwave acoustics is a rapidly developing field and new discoveries or techniques may make acoustic imaging practical in the gigahertz region. At present we cannot recommend it for Cyclops.

Microwave Imaging

The one remaining alternative is to use microwaves. From equation (51) we see that the angular magnification of a radiative imager is

$$M \equiv \frac{F}{\ell} = \left(\frac{f_r}{f_i}\right)\frac{v}{c}\,\frac{1}{\sigma} \tag{67}$$

For electromagnetic waves in free spaces at the original frequency $f_r/f_i = 1$ and $v/c = 1$; thus,

$$M = \frac{1}{\sigma} = \frac{\theta_{\max}}{\delta_{\max}} \tag{68}$$

where θ_{max} is the maximum value of θ we can allow (see Figure 11-18) and δ_{max} is the corresponding value of δ (see Figure 11-16). We can consider using electromagnetic waves because θ_{max} can be much greater than δ_{max} and therefore the scale factor σ can be very small.

Let us assume we wish to image over the tuning range from 1 to 3 GHz. Then $\lambda_{max} = 0.3$ m. If we choose $s = 2\lambda_{max} = 0.6$ m, then for a 1000-element array $a \approx 10$ m. We can easily space the receiving antennas at $s/2$ so we choose $b/a = 1$ in equation (66) and find $\ell \approx 40$ m. Whereas the array dimensions in acoustic imaging were very small, the dimensions involved in microwave imaging are larger than we would like, but not impractically large.

A great advantage of microwave imaging over delay line or acoustic imaging is that we can vary the image size by changing ℓ, thus allowing us to match the useful field size to the image array size as we vary the operating frequency. At 3 GHz, for example, we can use the same arrays as in the above example but increase ℓ to 120 m. But to realize this advantage we must be able to keep the image focused as ℓ is changed.

One way to focus a microwave imager is with an artificial dielectric lens in front of the signal plane and a similar lens in front of the image plane to flatten the field. This allows both arrays to be plane, but requires several pairs of lenses to cover the frequency range. The cost and practicality of artificial dielectric lenses 20 m or more in diameter have not been evaluated.

The use of concave arrays causes great mechanical complications since the radii of curvature must be changed as the separation is changed. The following possibilities were explored

Case	r_s	r_i	Signal Delay?
1	∞	$\ell/3$	yes
2	ℓ	$\ell/2$	no
3	$\ell/2$	∞	yes

The radii r_s and r_i are as shown in Figure 11-18. The signal delays are those needed to make the wave front for an on-axis source be spherical with its center at V_i. Although cases 1 and 3 permit one array to be plane, they require delays and the other array must be concave and adjustable.

Probably the most satisfactory solution is to make *both* arrays plane, and obtain focusing by the use of delay alone. If we accept a small amount of spherical aberration on axis (by making the radiated wavefront paraboloidal rather than spherical), we can keep the aberrations within tolerable limits over a plane image surface.

Figure 11-19 shows two plane arrays separated a distance ℓ. The signal delay is shown symbolically as a curved surface a distance τ/c behind the signal array. Let us assume this surface is a paraboloid given by

$$\frac{\tau(\rho)}{c} = k\left(1 - \frac{\rho^2}{a^2}\right)\left(\sqrt{\ell^2 + a^2} - \ell\right) \qquad (69)$$

Then the total distance d from a point on this surface to V_i is $d = \tau(\rho)/c + r$, and the change in d with respect to the axial value d_0 measured in wavelengths is

$$\epsilon = \frac{d - d_0}{\lambda} = \frac{\ell}{\lambda}\left(\sqrt{1 + \frac{\rho^2}{\ell^2}} - 1\right) - k\frac{\rho^2}{a^2}\left(\sqrt{1 + \frac{a^2}{\ell^2}} - 1\right) \qquad (70)$$

Plots of ϵ versus ρ/a are shown in Figure 11-20. We see that for all cases the spherical aberration for $0 < \rho/a < 1$ is small if $k = 1$, and is tolerable for the other values of k.

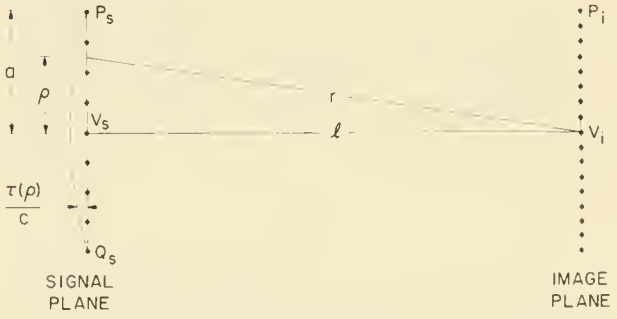


Figure 11-19. *Radiative imaging with plane arrays.*

Now an important property of a paraboloid is that adding a constant slope everywhere merely shifts the vertex; its shape remains unchanged. Thus, an off-axis signal, which is to be imaged at P_i, adds a delay slope that shifts the vertex of the paraboloidal radiated wavefront to P_s rather than V_s. We can therefore use the curves of Figure 11-20 to estimate the off-axis aberrations by simply noting that at P_s the abscissa ρ/a now is zero, at V_s the abscissa ρ/a is 1 and at Q_s the abscissa ρ/a is 2.

At 1 GHz and with $k = 0.97$ we see that, if the image plane is moved to increase ℓ by about 0.07λ as indicated by the dashed line, the peak error in the wavefront is

only ±0.07λ up to ρ/a = 1.77. Beyond this the error increases to about λ/4 at ρ/a = 2, but there is very little signal plane area at this value of the abscissa. Thus, the image quality should be quite good. The situation rapidly improves as we go to higher frequencies.

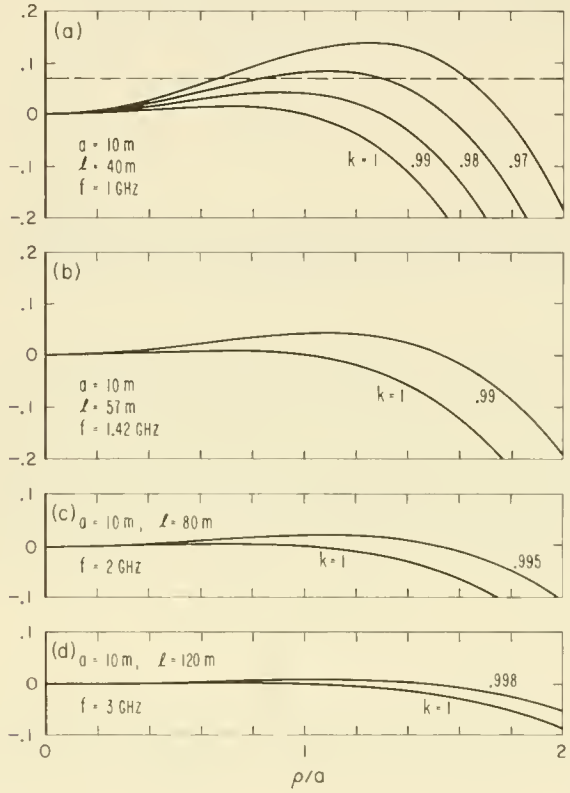


Figure 11-20. *Spherical aberrations with parabolic delay.*

Further studies might reveal that some other system, such as using a fixed curved signal array plus variable RF delay, would give improved results, but the system described seems entirely adequate. We are, after all, not attempting high-resolution imaging by optical standards; there are only 32 independent picture elements per line.

Another advantage of microwave imaging is that both polarizations can be imaged simultaneously. It is a simple matter to mix the IF signals (if necessary) to obtain signals representing vertical and horizontal polarization. They can then (after heterodyning to the imaging frequency) be radiated as such by crossed dipoles or loops. Good broadband performances can be realized with the crossed Vee antennas shown in Figure 11-21. If the height of the vanes is one-third their separation the characteristic impedance is 300 Ω. The length of the sides should be λ/2 greater than the altitude of the pyramid at the highest frequency used. Alternatively, the radiative imager could use short helical antennas. These also permit broadband operation and the two polarizations used could then be left and right handed circular. An advantage of helical antennas is that the phase of the radiation can be changed simply by rotating the helix about its axis. This phase changing capability could well simplify the construction of the imager and reduce the phase tolerances needed since it would be easy to trim each element.

With both polarizations imaged we can either add the intensities to decrease integration times, or subtract them to detect polarized sources. For the latter purposes we should be able to rotate the polarizations displayed, by using goniometers in the IF or RF lines.

The microwave imager just described requires a building with a clear volume about 80 ft high by 80 ft wide by 550 ft long. This is a volume of about 3 million cubic feet and a floor area of 41,250 square feet. The total surface area is 176,000 square feet. The building must be *completely* shielded and lined with microwave absorbing material to provide an anechoic chamber.

Very thin copper sheet will provide adequate shielding, provided there are *no cracks.* The skin depth of copper at 1 GHz is about 0.08 mil, so 5-mil thick sheet will provide about 62 nepers or 500 dB of attenuation. All doors must be sealed with contacting fingers and probably should be double with absorbing material between them. All signal and power leads entering and leaving must be shielded and equipped with low pass filters, as is done in screen rooms. Air vents must have a section, containing an egg-crate structure of waveguides below cutoff, that is soldered all around to the building shield.

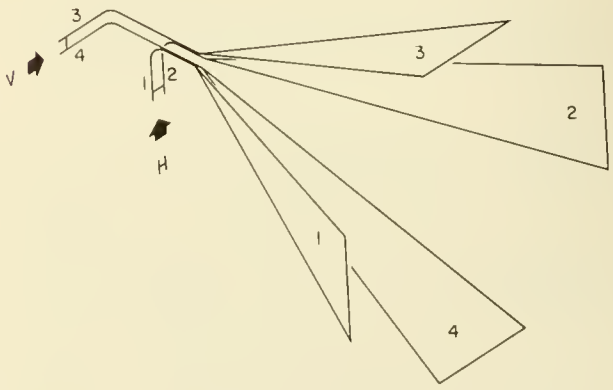


Figure 11-21. *Dual polarization broadband radiator.*

Rough estimates of the costs of building, shielding and imager are listed below:

Item	Cost ($ million)
Building structure	3.5
Shielding	1.0
Anechoic lining	1.4
Signal array	1.0
Image array	2.0
Mechanisms for varying separation	0.3
Total	9.2

The total shown is somewhat greater than our estimate for the delay line imager. The microwave imager, however, offers the advantages of (1) imaging both polarizations, (2) less complexity, (3) greater flexibility (any array pattern can be used), and (4) constant performance over a wide tuning range. These advantages are believed to outweigh the increased cost.

The large size of the microwave imager can be reduced if we accept a higher low-frequency limit. The linear dimensions scale with wavelength, and the length is proportional to the ratio of f_{max}/f_{min}. If we are content to image down to say 1.4 MHz (which allows us to map hydrogen in our own and nearby galaxies), the building needed would be 60 ft high by 60 ft wide by 220 ft long, and the total cost would be around $6 million. On the other hand, if we increase the array size to 2000 elements, we are back to our original building size.

Refining the Image

As noted earlier, the image will contain not only the detail from the major lobe of the array but will also be overlain by images from the grating lobes of the array. When an area is mapped, the mosaic of fields obtained can be computer processed to remove or greatly reduce these spurious images.

The picture produced by the imager is the convolution of the true brightness distribution of the sky with the lattice of grating lobes of the antenna array, all of which have the same shape as the main lobe but are decreased in amplitude by the directivity of the antenna elements. What we wish is the convolution of the true brightness distribution with the main lobe only. Thus, we can regard the main and grating lobes as a set of δ-functions of size $G_{i,j}$ and express the brightness, B', of the picture we get as

$$B'_{m,n}(u,v) = \sum_{i,j} B_{m+i,n+j}(u,v)\, G_{i,j} \qquad (71)$$

where m and n are the rank and file of a particular field and the mosaic of fields making up the whole picture, u and v are coordinate values within each field, and i and j are indices. To convert B' back to the desired distribution B, we need to convolve B' with a new array of δ-functions of size $H_{i,j}$ whose two-dimensional transform is the reciprocal of the transform of the original array of functions. That is, we compute the coefficients $H_{i,j}$ such that they satisfy the relation

$$\left[H_{i,j}\delta(iu_0, jv_0)\right] * \left[G_{i,j}\delta(iu_0 jv_0)\right] = \delta(0,0) \qquad (72)$$

and then perform the operation

$$B_{m,n}(u,v) = \sum_{i,j} B'_{m+i,n+j} H_{i,j} \qquad (73)$$

In equation (72) the quantities u_0 and v_0 are the dimensions of the field and the spacings of the grating lobes in the u and v directions.

Actually, because the off-axis grating lobes are smeared by chromatic aberration and, if the exposure is long, by the rotation of the earth, B' in equation (73) should be a weighted average over the appropriate area in each adjacent field, rather than the intensity at a single point.

Image Distortion and Rotation

So far we have considered only zenith imaging. At a zenith angle θ, the scale of the picture in the altitude direction changes by cos θ so that a greater range of altitude is mapped with correspondingly decreased resolution. This is a routine problem in radio astronomy mapping. It merely means that the map is not free of distortion as recorded and must be corrected if an orthographic projection is desired. Over small mapping areas the whole picture simply has a different scale in two perpendicular directions with the result that lines of constant right ascension and declination may no longer be orthogonal. The map may be correcting by projecting the individual fields through an anisomorphic lens system.

A more troublesome defect is that as an object is tracked for some time its image rotates, as in any alt-azimuth telescope. (This defect cannot be corrected

by equatorially mounting the antenna elements; it is the *array* that is alt-azimuth.) If the field is being photographed we may wish to rotate the camera at the appropriate rate. If the image is being stored in a signal averager we may wish to "rotate" the addresses. We have not had time to study this problem in detail.

UNDESIRED SIGNALS

Terrestrial Transmitters

Most radio astronomy today uses several frequency bands that have been set aside specifically for that purpose by governmental agencies. No terrestrial transmitters are permitted within these bands, which eliminates most interference to radio astronomy. Nevertheless, radio astronomers encounter interference from a variety of transmitters. Radars often radiate small but troublesome signals far outside their nominal pass bands. Harmonics or unexpected mixing products appear in radio-astronomy bands from stations operating nominally legally. Thus, even though radio astronomy is done largely in "protected" bands, interference from terrestrial transmitters is a problem.

Cyclops, on the other hand, will not always operate in protected bands, but over a wide frequency range occupied by thousands of terrestrial transmitters. Calculations have been made which indicate that UHF TV stations will not only be detectable, but will overload any reasonable receiver. Lower powered stations will be detectable for several hundred miles. Any flying object will be detectable out to lunar distances.

A certain degree of interference reduction can be achieved by generating a secondary beam that is larger than, and concentric with, the main beam. Subtraction of these two signals then removes most of the interferences.

It will probably be necessary to catalog all known transmitters and program the central computer to ignore them. Unfortunately, this creates many "blind" frequencies for Cyclops, and still does not include flying or time-varying transmitters. Alternatively, a *quiet zone* could be created for some reasonable distance surrounding the Cyclops array.

REFERENCES

1. Golay, M.J.E.: Note on Coherence vs. Narrow-Bandedness in Regeneration Oscillators, Masers, Lasers, etc. *Proc. IRE*, vol. 49, no. 5, May 1961, pp. 958-959.
2. Golay, M.J.E.: Note on the Probable Character of Intelligent Radio Signals From Other Planetary Systems. *Proc. IRE*, vol. 49, no. 5, 1961, p. 959.
3. Bracewell, R.N.: Defining the Coherence of a Signal. *Proc. IRE*, vol. 50, no. 2, 1962, p. 214.
4. Bracewell, R.N.: Radio Signals From Other Planets. *Proc. IRE*, vol. 50, no. 2, 1962, p. 214.
5. Hodara, H.: Statistics of Thermal and Laser Radiation. *Proc. IRE*, vol. 53, no. 7, pp. 696-704.
6. Cooley, J.W.; and Tukey, J.W.: An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comp.* (USA), vol. 19, Apr. 1965, pp. 297-301.
7. Goodman, J.W.: Introduction to Fourier Optics. McGraw Hill, 1968.
8. Thomas, C.E.: Optical Spectrum Analysis of Large Space Bandwidth Signals. *Applied Optics*, vol. 5, 1966, p. 1782.
9. Markevitch, R.V.: Optical Processing of Wideband Signals. Third Annual Wideband Recording Symposium, Rome Air Development Center, Apr. 1969.
10. Mamikumian, G.; and Briggs, M.H., eds.: Current Aspects of Exobiology. Ch. 9, 1965.
11. Oliver, B.M.: Acoustic Image Synthesis, unpublished memorandum.
12. McLean, D.J.; and Wild, J.P.: Systems for Simultaneous Image Formation with Radio Telescopes. *Australian J. Phys.*, vol. 14, no. 4, 1961, pp. 489-496.

12. CYCLOPS AS A BEACON

Although the Cyclops array was conceived as a large aperture receiving system, nothing inherent in the design prevents its use in the transmitting mode as well. This would allow radar astronomy to be performed to greater precision and over greatly extended ranges as discussed in Chapter 14. It also opens the possibility of using Cyclops as an interstellar beacon. Of course, transmission and reception could not go on simultaneously. Nevertheless, if our first search of the nearest 1000 target stars produced negative results we might wish to transmit beacons to these stars for a year or more before carrying the search deeper into space. We could then re-examine these stars at the appropriate later times looking for possible responses.

There are two natural ways to use Cyclops as a beacon: to focus the entire array on a single star, and to train each element on a different target star. The second method would require a different coding method for the antenna positioning information than the one described in Chapter 10, which permits only a few subarrays. However, the added cost of providing completely independent positioning for each element would be very small. It is therefore of interest to determine Cyclops's beacon capabilities in both modes.

Let us assume that the nominal Cyclops array of a thousand 100-m dishes is equipped with a 100-kW transmitter at each antenna. The total transmitted power would then be 100 MW. (To achieve this total would probably require on the order of 300 MW of power to be fed to the array over a somewhat heavier distribution system than that specified in Chapter 10.) At an operating frequency of 1.5 GHz (λ = 20 cm), the effective radiated power with the entire array aimed at the same star would then be 2.5×10^{17} watts. With the elements used individually as beacons the effective radiated power for each would be 2.5×10^{11} W.

Of course, we do not know all the characteristics of the receiver that might pick up our beacon signals. However, if it is intended for radio astronomy or deep space communication, and the other race is technologically advanced, it is reasonable to assume a noise temperature of 20° K or less. Using a value of 20° K we can then plot the reference range limit as a function of the receiver antenna diameter and bandwith.

Figure 12-1 shows the ranges for various antenna diameters when the entire array is focused on a single target star. We see that a small receiver with a 10-m diameter antenna and a 10-kHz bandwidth could detect the Cyclops beacon at a range of about 80 light-years. A 100-m antenna would extend the range to 800 light-years or permit detection at 80 light-years with a 1 MHz bandwidth. We also see that with a 1 Hz receiver bandwidth, the Cyclops array beacon could be picked up at 50 light-years with no reflector at all, merely on an isotropic antenna or dipole. The significance of this is that the signal might be detected by a search system at this range even if the latter were not pointed toward the Earth.

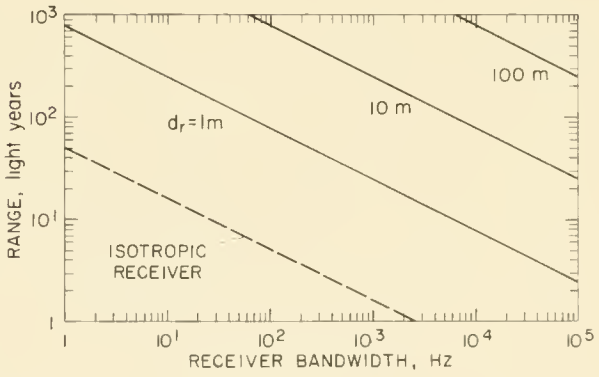


Figure 12-1. *Range limits of 100-MW, 3-km Cyclops array used as a beacon.*

When we spread the elements to point at 1000 different stars, the range is 1000th as great for the same receiver. However, as shown in Figure 12-2 we would be detectable at 80 light-years on a search receiver having a 1 Hz bandwidth and a 100-m antenna. Another Cyclops array could detect us at the same range if the receiver bandwidth were 1 kHz.

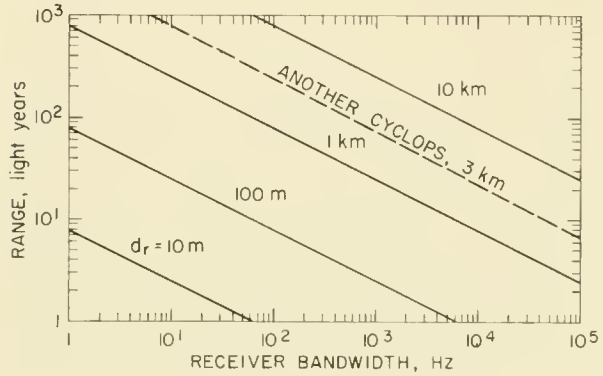


Figure 12-2. *Range limits of 100-kW Cyclops, 100-m antenna element as a beacon.*

The possibility of picking up with a 100-m antenna a signal that was sent by another 100-m antenna at a range of 80 light-years suggests a search mode in which the array is alternately employed as a search receiver and as a beacon for say, the nearest 2000 likely stars (only half of which are visible at any one time). The problem with this approach is that a separate data processing system must be devoted to each antenna in the receiving mode. Although the cost of the data processing system required to comb 100 MHz of RF spectrum (in two polarizations) is small compared with the cost of the entire array (see Chap. 7), it is large compared with the cost of a single antenna element. Unless the data processing cost can be reduced by about two orders of magnitude, parallel processing with a star per array element will be prohibitively expensive. For the present we are left with only the possibility of serial search star by star.

Of course, we could provide, say, ten data processing systems (at a total cost of $1 to $2 billion) and do parallel processing on ten stars at a time. But since this would reduce the array area for each star by a factor of ten, the number of stars that could be searched this way is only 3% of the total accessible with the full array. Thus, the time saved by parallel processing of nearer stars is such a small fraction of the total time that funds spent for parallel processing would be better spent for antenna area, where an increase reduces the necessary observation time for every star.

We conclude that Cyclops can, and probably should, be used as a beacon to illuminate stars closer than 100 light-years, during certain portions of the search phase. The transmission capability of the array provides a built-in powerful response system for any signal that might be detected at *any* range.

13. SEARCH STRATEGY

The high directivities that inevitably accompany coherent collecting areas many hundreds of wavelengths in diameter, together with the cost of replicating the necessary data processing equipment, force us into a serial search mode, in which we examine one star at a time. The fundamental objective of the search strategy is to organize the serial search process in such a way as to achieve contact in the least time. More precisely, we wish to maximize the probability of having made contact after any given length of search time.

If the Cyclops system were to materialize overnight in its full size, if we had a complete catalog of all stars by spectral type within 1000 light-years of the sun, if we knew the relative probabilities of the occurrence of advanced life on planets belonging to stars of different spectral types, and, finally, if we knew how the probability of detecting a signal decreased with increasing range, the search strategy would be quite straightforward. We could then compute an *a priori* probability p for the existence of a detectable signal from any given star. This probability would be the product of a function f of the spectral class S and a function g of the range R. That is, we would have

$$p = kf(S)\,g(R) \qquad (1)$$

where k is a constant that absorbs such factors as the longevity of the communicative phase. The optimum search strategy would then be to list all stars in order of decreasing p, begin at the top of the list, and work our way down.

Because the Cyclops system will grow in size with time, this simple procedure is complicated by the fact that g will be a function of both R and t; that is, equation (1) will be

$$p = kf(S)g(R,t) \qquad (2)$$

This merely means revising our tables of p as we go along and working always with the most probable stars on the most up-to-date list.

At present we do not know the exact forms of the functions f and g. Throughout this study we have assumed, on the basis of available knowledge, that f is a unimodal distribution centered on main sequence stars of spectral class G. This may be an anthropocentric view, but for the reasons given in Chapter 2 we find the assumption quite a logical one. For equally compelling reasons, the function g may be taken to decrease monotonically with range. Thus, the model that guides our thinking may be represented schematically as shown in Figure 13-1, where we have plotted p vertically as a

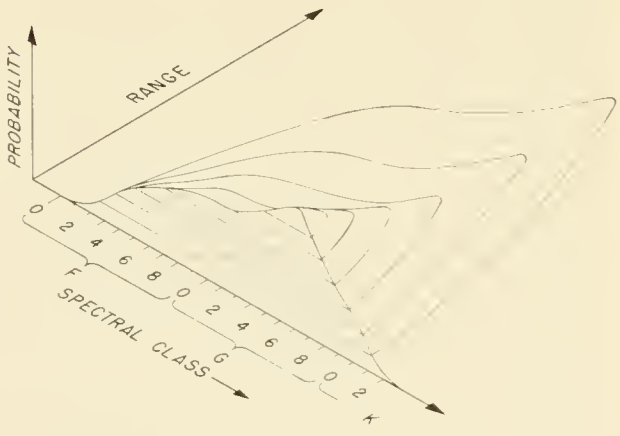


Figure 13-1. *Contours of equiprobable detectability.*

function of the spectral class and range coordinates. We imagine the search as commencing at the mountain top and successively picking up the stars around the surface at lower and lower altitudes. It is like examining the emerging shoreline of an island in a subsiding sea.

During the early construction years while the range is severely limited, the emphasis is naturally shifted to the closer stars. Because these are relatively few in number, we may spend a relatively long time per star searching for leakage as well as for beacons. However, since leakage signals are expected to be weak compared with beacons, the leakage search should be repeated for the stars within, say, 100 light-years after the array has reached its full size.

It is important to realize that the search strategy is affected very little, if at all, by changes in the expected mean distance between communicative races. This distance determines the probable *magnitude* of the search effort, but not the *order* in which target stars are examined.

Naturally we would like to know the shapes of the functions f and g more accurately. The shape of f affects the importance of knowing g. If, for example, f were constant between, say, spectral classes $F5$ and $K5$ and were zero outside this range, all we would need to know about g is that it decreased monotonically with range. The search strategy would then simply be to search all $F5$ through $K5$ stars in order of increasing range. If, on the other hand, as is more likely, f peaks at some spectral class and drops off smoothly on either side, we will want to examine a number of distant stars in the optimum part of the spectral range before we examine other nearer stars at the extremes of the likely spectral class range.

The shape of f depends in an unknown way on the selectivity and stellar longevity factors discussed in Chapter 2. Dole (ref. 1) has studied the factors that might make a planet habitable by man, and has published a table of probabilities of finding a habitable planet about stars of various spectral classes. To give a general picture of what f might look like, we have made the ordinates in Figure 13-1 proportional to Dole's values. Further study of this whole matter, taking into account more recent findings with respect to atmospheric evolution as a function of UV and X-ray flux from the star, for example, is certainly in order. It is important not to ignore stars that might support life and desirable not to waste time on those that cannot, or for which the probability is vanishingly small.

The shape of g depends on the probable distribution of beacon (and leakage) powers. If the economic reason given in Chapter 6 is valid, we may be able to refine our estimate of the shape of this function.

While these uncertainties prevent us from being able to specify an optimum search sequence, they are not the principal problem. At present we know the distances of only a few hundred target stars; the list of F, G, and K stars is fairly complete only out to a few tens of light-years. This is less than one-thousandth part of the list of target stars we need to do a thorough search out to 1000 light-years. If we do not have such a list and merely point our search system at all stars down to, say, magnitude 15 we will waste over 90 percent of the search time examining very distant giants and supergiants as well as a large number of nearer main sequence stars of the wrong spectral class. Thus, the first step in the search strategy is to develop a comprehensive list, by spectral class, of target stars in order of increasing distance out to, say, 1000 light-years. This is not an easy undertaking.

DISTANCE MEASUREMENT BY PARALLAX

The distances of the nearest stars can be measured by the parallax they display as a result of the Earth's orbital motion. In fact, 1 parsec (= 3.26 light-years) is the distance at which a star will show a parallax of 1 arc-sec. Parallax measurements are reliable only for distances up to 50 parsecs (i.e., for parallaxes greater than 0.″02). At present, only about 760 stars are known to have parallaxes greater than 0.″05, and these account for only one-fifth of the stars within 20 parsecs of the Sun. Beyond about 50 parsecs, other less direct methods must be used to determine stellar distances accurately.

DISTANCE INFERRED FROM PROPER MOTION

The closer a star the greater will be the proper motion produced by a given velocity of the star across our line of sight. Hence nearer stars will tend to show larger proper motions, and we might consider using proper motion as a measure of distance.

We can identify three generic causes for stellar velocities normal to our line of sight:

1. Peculiar galactic orbits
2. The general circulation of stars in the disk around the galactic center
3. Random velocity components over and above the general circulation.

A relatively few stars have highly elliptic galactic orbits producing large observed velocities relative to the Sun. These are generally Population II stars and are probably of little interest. Their high velocities would cause many of these to be included in our target list if proper motion is our criterion.

If all stars had the same period around the galactic center, their apparent proper motion, with respect to the

distant galaxies (or an inertial reference frame) would be proportional to the cosine of the galactic latitude, and would be independent of range. Since the period of stars increases with their distance from the galactic center, their velocities are not proportional to this central distance; in fact, in the solar neighborhood, the velocity decreases slightly with distance from the galactic center. The difference between these two functions is proportional to radial distance from the Sun's orbit and hence causes a proper motion that is a function of galactic longitude, but is independent of distance from the Sun.

Thus, the only class of stellar motions that can be of any use to us are those caused by random velocities. If we base our distance estimates on these:

1. We will fail to include a large number of valid target stars whose velocities are either low or along our line of sight. This fraction will increase with range.
2. We will include a substantial number of stars beyond range whose velocities are higher than normal.
3. Unless some additional technique is used, no selection by spectral or luminosity class is provided.

Clearly, proper motion studies, by themselves, are inadequate to prepare a good target list.

DISTANCE DETERMINATION FROM ABSOLUTE MAGNITUDE

As we saw in Chapter 2, the vast majority of stars fall into well-defined groups, called luminosity classes,[1] on the H-R diagram. The luminosity class of a star can be determined from certain features of its spectrum such as the strength of the Balmer absorption band in the ultraviolet, the strength and number of the absorption lines of metals and heavy elements, and the relative heights of emission lines (core reversals) particularly of the calcium line. Once the luminosity class is known, along with the spectral class, the luminosity and therefore the absolute magnitude of the star can be determined within rather narrow limits. This is particularly true for main sequence stars.

Since the absolute magnitude M is the magnitude the star would have at a distance of 10 parsecs, we can determine the distance from the inverse square law, knowing the apparent magnitude m:

$$R = 10^{\left(\frac{m-M}{5} + 1\right)} \text{ pcs} \qquad (3)$$

This method of distance determination is reliable and has been used to find the distances of a great many stars.

Since in any case we wish to know the spectral class of each star and to know that it is a main sequence star, the determination of its distance is a trivial additional step involving only the apparent magnitude. The crux of our problem is therefore to select or develop a rapid (and preferably automated) method of spectral classification that will exclude from consideration stars that are not on the main sequence. For our purposes, we do not need to determine the luminosity class of stars not on the main sequence, but this information would be of interest to astronomers in refining their statistics of stellar distributions.

UBV PHOTOMETRY

A widely used method of star classification involves the measurement of the apparent magnitudes of the star in three wavelength ranges, one in the ultraviolet, one in the blue, and one in the "visible" part of the spectrum, as listed in Table 13-1. These measurements have been made photoelectrically on single stars and, more recently, by densitometric measurement of star images on plates taken in the three spectral regions of fields of stars. The ultraviolet, blue, and visible magnitudes are designated U, B, and V, respectively.

TABLE 13-1

Measurement of	λ (center) (nanometers)	Δλ (nanometers)
U	365	70
B	440	90
V	548	70

With care, U, B, and V can be measured from photographic plates to ±0.05 magnitude. If the magnitude difference U-B is plotted against the difference B-V for a large number of stars including both supergiants and main sequence stars, the points fall into two loci as shown in Figure 13-2. Such a plot is known as a two-color diagram. The bands are drawn with a width of 0.1 magnitude to show the effects of measurement error.

Because the shape of the black-body radiation curve is temperature dependent, both U-B and B-V will change with stellar surface temperature, but U-B is additionally affected by the amount of ultraviolet (Balmer) absorption, which is related to the luminosity. Thus B-V is primarily a measure of spectral class, and the spectral

[1] These are commonly designated by roman numerals: I, supergiant; II, bright giant; III, giant; IV, subgiant; V, main sequence (dwarf); VI, sub or white dwarf.

class boundaries have been noted along the abscissa in Figure 13-2. We see that the U-B value separates the supergiants from the main sequence stars rather well in the spectral range *F*0 through *G*5, but that confusion exists for the *K* and *M* stars. Giant stars lie on a curve intermediate to the two curves shown, and the confusion of these with main sequence stars is correspondingly worse.

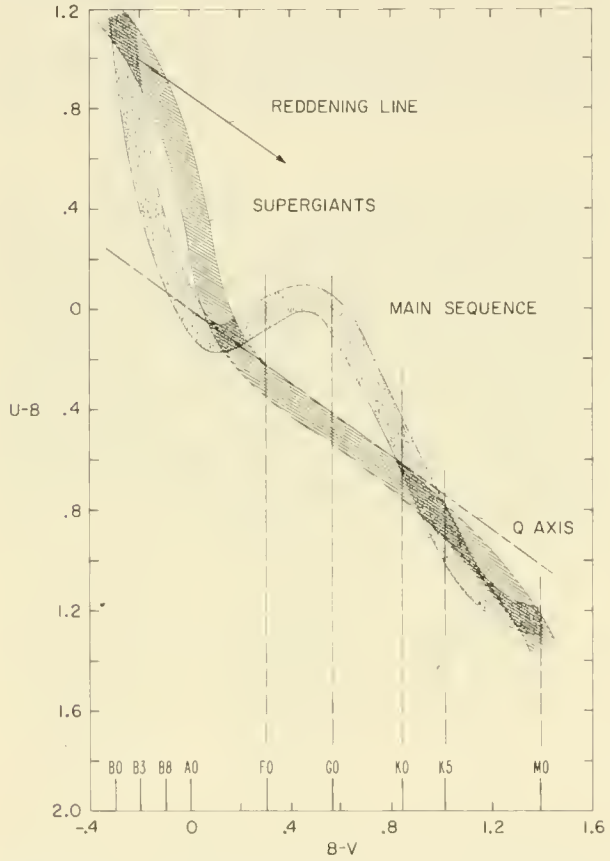


Figure 13-2. *Two-color relation for main sequence stars and supergiants.*

A further source of confusion arises from interstellar absorption, which increases with decreasing wavelength and thus reddens the light of distant stars seen in the galactic plane. The reddening decreases both U-B and B-V, the former by about 0.72 times as much as the latter, and thus shifts the positions of stars on the two-color diagram along a line having a slope of −0.72 as shown by the "reddening line" in Figure 13-2. The "Q-axis," also shown in the figure, is a line of slope −0.72 drawn through the point U − B = 0, B − V = 0. The Q-value of a star, defined as

$$Q = (U - B) - 0.72\,(B - V) \qquad (4)$$

is simply its distance above or below the Q-axis and is independent of the amount of reddening. On a plot of Q versus B-V, reddening would produce a horizontal shift in the star's position, but the confusion of spectral and luminosity classes produced by reddening would be just as great as on the standard two-color diagram.

We see that reddening will shift *F*0 and later type supergiants along the existing locus and will have little effect on the total confusion in the *K* region. Type *A* supergiants will be shifted into the *F* and *G* part of the main sequence curve, but these stars are very rare. More serious is the reddening of *B*3 through *B*8 main sequence stars into the *F*0 through *K*0 part of the main sequence curve.

It is of interest to see if adding a fourth wavelength in the red (*R*) or infrared (*I*) would enable the amount of reddening to be determined and thus eliminate reddened stars masquerading as *F* through *K* main sequence stars. Interstellar absorption is usually assumed to be proportional to $1/\lambda$. Figure 13-3 compares the curves of black-body radiation from a 6000° K (*G*0) star with the black-body radiation from a 10,000° K source that has had enough $1/\lambda$ absorption to give it the same B-V value. We see that the difference between the curves is slight, amounting to less than 0.1 magnitude in the 1μ region. On this basis, the addition of a fourth wavelength would hardly seem worthwhile. On the other hand, work by

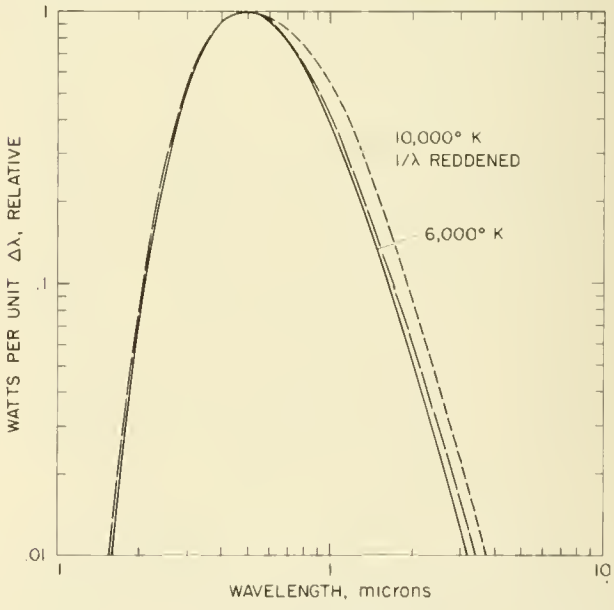


Figure 13-3. *Effect of reddening on black-body radiation.*

Johnson (ref. 2) seems to indicate that the absorption falls more rapidly than $1/\lambda$ beyond 0.6μ. According to his published curves, the reddened 10,000° K radiation

would follow the uppermost line in Figure 13-3 at long wavelengths, resulting in a difference of about 1/2 magnitude at 1μ. If this is true then the inclusion of a fourth wavelength at, say, 1μ would be of some help.

Even if we could completely eliminate the confusion produced by reddening we would be left with the basic confusion of giants and supergiants with main sequence stars in the late *G* and *K* region. To reduce this confusion we must reduce the measurement errors and thus narrow the loci in the two-color diagram. This probably means eliminating the photographic steps completely and going to direct photoelectric measurement, where the measurement errors can be reduced to about ±0.01 magnitude.

OBJECTIVE PRISM SPECTROSCOPY

Another well-known technique for the spectral classification of fields of stars is objective prism spectroscopy. A large prism is placed in front of the telescope objective causing the images of all the stars in the field to be dispersed into their spectra. A single plate thus records the spectra of a few dozen to several thousand stars, depending on the direction, the field of view, and the exposure time. This technique substantially reduces the telescope time needed to do a survey, but not the analysis time, since (with conventional methods) the spectroscopist must now examine the spectra one by one for the features that identify the spectral and luminosity class.

In principle, this method could be used to prepare the Cyclops target list; all we would need is enough trained spectroscopists to do the job in a reasonable time. At a limiting magnitude of 16, some 20 million spectra would have to be examined. Assuming a trained spectroscopist could make one classification every 8 min, about 1000 man-years would be required. Thus, 200 trained people could complete the job in 5 years, or 500 people in 2 years. We regard this solution as tedious and inelegant. In addition to the drudgery, mistakes would inevitably be made both in classification and coordinate specification. Also, several plates per field would be required to cover the magnitude ranges of stars in the field, so the above times are probably optimistic.

Objective prism spectrometry is difficult to automate. One reason for this is that the dispersed images of the stars tend to overlap and thus confuse an automatic scanner. Even a spectroscopist cannot always distinguish the pertinent features in overlapped spectra and may require plates to be taken with the objective prism set at different angles. For plates taken in the galactic plane to magnitude 16 the overlap problem is rather serious. Here there would be a few thousand star images per square degree; let us assume 3600. Assume the star images (undispersed) cover 3 square arc sec, and that we need 2 to 3 Å resolution over the traditional spectrum range of 3900 to 4400 Å. Using a filter to limit the spectrum to this range, each dispersed image would cover about 600 square arc sec on the plate for a total of 600 × 3600 square arc sec out of the $(3600)^2$ available. Thus, the probability of a star spectrum, placed at random on the plate, overlapping other images is 1/6 and, because each overlap confuses two images, we could expect something like 1/4 to 1/3 of the images to be confused.

The overlap problem can be reduced by exposing *n* plates, each to an *n*th part of the total spectral range. This would make the analysis problem more difficult for a spectroscopist but not for an automatic scanner. As *n* is increased the overlap decreases and, in the limit, when only one resolvable portion of the spectrum is recorded per plate, the overlap is no more serious than on an undispersed plate.

PHOTOELECTRIC SPECTROMETRY

The development of TV camera tubes has made possible the direct integration of spectra photoelectrically and the direct transfer of spectral information into computer memory. One instrument using this technique is being designed at the New Mexico Institute of Mining and Technology. This instrument uses a 30-in. telescope to image single stars into a spectroscope capable of producing either a low dispersion spectrum using a prism or a high dispersion spectrum using the prism and a crossed echelle grating. The spectra are integrated by an orthicon camera tube having a raster scan of 128 lines and 128 elements per line for a total of 16,384 picture elements. Without dispersion the camera tube can detect magnitude 15 stars in 2.5 sec. In the low dispersion mode, this time is increased to about 10 min and in the high dispersion mode to about 2 hr. The repositioning time of the telescope is about 1 sec if the coordinates of the next star are known and close to those of the last star.

Two factors make this instrument unsuitable for compiling the original Cyclops target list. First, the coordinates of almost all the stars we must examine are not known but must be found by the survey itself. Second, an analysis time of 10 min or more per star is too long. About a thousand years would be spent in examining 20 million stars with a single instrument. However, we could consider using several instruments of this type to refine a target list of a million stars prepared by some other faster method.

Both the unknown position problem and the speed limitation of the above instrument as a survey tool for

Cyclops would be solved by designing an instrument to work not on single stars but rather on star *fields* as in UBV photometry and objective prism spectroscopy. Such an instrument would simply be positioned to a succession of fields chosen to cover the entire sky visible from the Cyclops site. The coordinates of any star would then be determined from the known direction of sighting (i.e., the direction of the center of the field) and the position of the star image in the field. The positions of known stars could be checked to eliminate systemic errors. Because the light from as many as a few thousand stars would be integrated simultaneously, the total integration time would be correspondingly reduced

If dichroic mirrors were used to split the received light, different spectral regions could be integrated simultaneously in a bank of camera tubes, each having a photosurface optimized for the spectral ranges involved. Interference filters could be used to select narrow regions of the spectrum, and prisms could be used to disperse these regions. Thus, a wide variety of spectral analysis methods is at our disposal. It is not clear without extensive further study just what combination of techniques and what combination of spectral regions or spectral lines would be most effective in selecting our desired target stars and discriminating against giants and reddened interlopers.

One definite possibility is to use four (or more) spectral regions in a direct photoelectric photometry mode to make an initial screening of the stars in each field. If, through suitable calibration procedures the measurement errors in the various spectral bands can be held to ±0.01 magnitude (≈ ±1%), good correction for reddening should be possible and the size of the confused region for late *G* and *K* stars should be greatly reduced. This should permit rapid classification of the stars in a given field into three categories—target, doubtful, and nontarget—with only a small fraction falling in the doubtful category. These could then be examined spectroscopically using several telescopes of the type described earlier, while the next star field is being classified photometrically.

Another possibility is that several appropriately chosen wavelengths would permit the unambiguous classification of all stars in the field on one pass. If so, such a procedure might be preferable even if longer integration times are needed. For the present we can only conclude that:

1. No rapid method of preparing a clean target list for Cyclops exists at the present time.
2. Promising techniques do exist, and it appears likely that with adequate study and development, a satisfactory automated system could be designed.
3. Classification of all stars within 300 pcs of the Sun would be of considerable value in refining our knowledge of stellar evolution and is therefore of merit in itself.
4. Consideration should be given to funding a program to develop an automated system of rapid accurate stellar classification irrespective of the imminence of Cyclops.

THE OPTICAL-ELECTRONIC INTERFACE

Assuming that a suitable optical spectrum analysis technique can be developed, a few problems remain in transforming the optical information into digital form. Once the information is stored digitally the analysis can proceed rapidly and reliably with well-known data processing techniques; but first we must get the information into a proper digital format without significant loss of accuracy.

The first problem is one of dynamic range. If we are examining stars down to magnitude 16 we must accommodate a brightness range of 2.5 million to 1. This is beyond the capabilities of any camera tube. If we can assume that all stars down to magnitude 6 are already known and classified, we are left with a magnitude range of 10 or a brightness range of 10,000 to 1. Good noise-free linear integration can be obtained over at least a 10 to 1 brightness range. Thus, we will need to take at most four exposures of the same field differing in successive exposure times by at least 10 to 1. Assuming the readout time is negligible, this procedure increases the observing time by 11.1 percent over that required for the longest exposure alone and so causes no real difficulty. However, the software logic must be designed to ignore images that have already been analyzed, or that represent inadequate or overload amounts of integrated signal.

The second problem is to convert the analog information stored in the camera tubes into digital information in computer memory in such a way that precise positional and amplitude information is retained in spite of the discrete nature of a raster scan. To do this, the spacing between successive scanning lines must be small compared with the image of any star as limited by atmospheric blurring or diffraction—that is, the scanning line spacing and the spacing of data samples taken along each line should be small compared with 1 sec of arc.

A complete field would then be scanned, sampled, converted to digital form, and stored in a temporary memory. Assuming 1/4 sec of arc spacing between samples and 8 bit (256 level) quantization, 1.6×10^9 bits

per square degree of field would be required. At 0.03 cents/bit in head-per-track disks, this is about a half million dollars worth of memory. A software program would then determine the position of the centroid of each image as well as the amplitude of the image, store these data in core and erase the image from the temporary memory. After all images are erased, the temporary memory is freed for the next scan. If this part of the data processing takes a short time ($\approx$ 1 sec or less), the same temporary memory could be used for all spectral regions being examined; otherwise, several temporary memories would be needed.

The final step is to intercompare the amplitudes of the different spectral samples for each star in the field, and to select the spectral and luminosity class that best fits the observed data. Then from the now known absolute magnitude and from the observed magnitude (amplitude in one or more spectral bands) the distance is determined. If the star is a main sequence *F*, *G*, or *K* star within 300 pcs, it is entered into a master file and into a master target list, both on magnetic tape. If not, only the first entry is made for later astronomical use. Although the analysis operation is complicated, it is a straightforward process and could proceed at high speed. The time required should easily be less than 0.1 sec per star.

After the survey is completed, the data file can be sorted by well-known techniques. Probably the stars should be separated by spectral class and recorded on separate tapes in the original scanning process, although a single master tape could be so separated in a first sorting operation. The final sort consists of ordering all entries of each spectral class in order of increasing range (increasing magnitude).

As the target list is being compiled, the accuracy of the system operation should be checked frequently. Sample stars should be selected from each plate and tested by spectrographic measurement to determine how well the automatic system is performing. The accuracy of spectral typing, distance, and coordinate data should all be checked. A few weeks spent in validating the list could save years of wasted search effort.

The monitoring program would also provide experimental verification of the capabilities of the classification system, and give us an idea of the extent to which the initial target list is diluted with false entries or is lacking true target stars. False omissions should be considered a more serious defect than false inclusions.

REFINING THE TARGET LIST

Since we do not expect the target list to be perfect, we may want to refine it as time goes on. Although we wish to develop the initial list in a relatively short time so that the search can be started, the search itself will, as we have noted, probably require years or decades. Thus, while the search is in progress, the companion optical system can be used to verify listings in advance of the first (or certainly the second) search. We have not had time to study how this might best be done, and can therefore only suggest what seems to be a reasonable approach.

The system we visualize consists of a battery of telescopes each supplied with automatic positioning drives, a computer, tape reader, and a copy of part of the target list. Each telescope would go down its portion of the target list looking successively at the stars visible at that time of year. The star's spectrum under low or medium dispersion would be imaged onto the target of a long time-constant (cooled) vidicon for the time needed to obtain a good record of the spectrum. The exposure time could be controlled by letting part of the light of the undispersed star image fall on a photomultiplier and integrating the output current.

When the spectrum recording was completed it would then be scanned by the vidicon and stored in the computer memory. There it would be cross correlated with standard spectra taken by the same telescope system from stars of known spectral type. The spectrum showing the highest cross correlation would then be used to classify the star being examined. If the star were indeed a main sequence *F*, *G*, or *K* star, the distance would be computed from the exposure time required and the star would be entered on the new master tape listing. If not, the star would be rejected.

If the refining of the list is to keep pace with the search and if the average exposure time is equal to the search time, three telescopes would be needed, even if there were no false entries on the original list, because Cyclops can work all day and all night, while the telescopes can only work on clear nights. If, in addition, the refining process rejects half the original entries, then the telescopes must examine twice as many stars as Cyclops, so six telescopes would be needed.

It may also be possible to use Cyclops itself to refine the target list as discussed in the last section of this chapter.

THE FOUR SEARCH PHASES

The search for intelligent extraterrestrial life, like any voyage of discovery, may produce may surprises, and may be serendipitous. Columbus did not make the voyage he planned; he made a far shorter and more important one, though it took the world some time to realize this. In the same way, Cyclops itself may discover

facts about the Galaxy and the universe that could profoundly affect our ideas of how and where best to search for life. What we present here are only our present preliminary thoughts as to how the search might proceed as long as no such discoveries are made. It is, one might say, the worst-case plan. Obviously if any interesting signals are found at any time, the whole plan from then on is changed.

The search rather naturally divides into four phases, as outlined below.

Phase I: Preoperational phase

During the planning stages of Cyclops, before the first antennas are built and before the data processing and other systems are operational, the compilation of the target list should be going on. If the optical survey system is constructed first, then by the time Cyclops goes on the air, a large target list should be available, perhaps a complete though unrefined one.

Phase II: The Construction Years

We visualize the construction of Cyclops as taking place at the rate of perhaps 100 antennas per year over a 10- to 25-year period. As soon as the first few antennas have been built and the rest of the system is complete

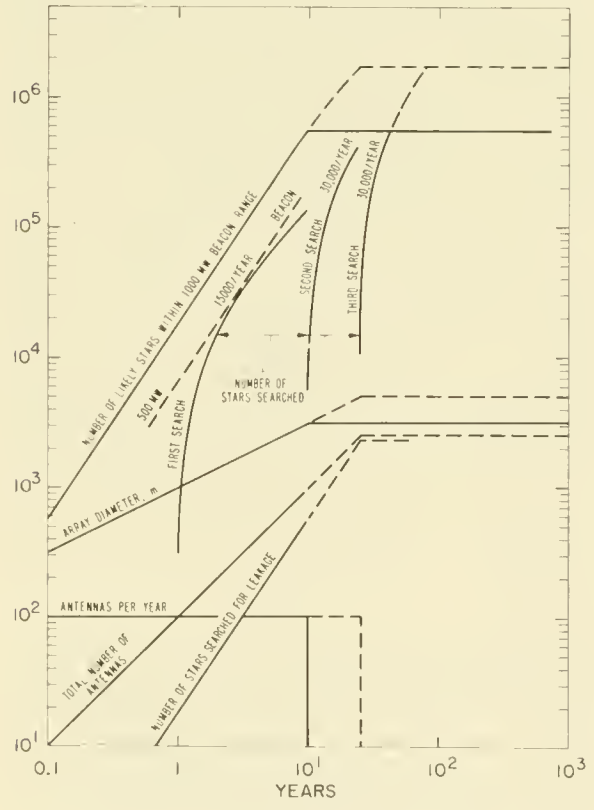


Figure 13-4. *Growth curves for Cyclops.*

and operational, the search may begin. During the first year, the nearest stars would be searched for both leakage and beacons, and the techniques for doing this efficiently would be developed and tested. Then, during the remaining construction years, the search would be carried farther and farther into space. Since the antenna area would be increasing linearly with time, the range would be increasing as the square root of time, and the number of stars accessible as the 3/2 power.

This buildup is shown in Figure 13-4. The first search is shown taking place at the rate of 15,000 stars per year, or an average observation time of 2000 sec per star. This allows about half the time to be devoted to leakage scans and to radio astronomy uses and the other half to the fully automated beacon search. At the end of 10 years we would probably wish to reexamine the stars already searched. By now, the system sensitivity would be an order of magnitude greater than it was the first time many of these nearer stars were first searched.

Phase III: Total Search with the Complete System

The first one or two scans may not have taken us beyond 500 to 700 light-years. Now with the complete system and a completed and refined target list available, the first complete scan out to 1000 light-years can commence. Prior to this third search, the decision may have been made to use the Cyclops array as a beacon for a year or more. If so, then the final search should repeatedly look at the 1000 or more nearer stars at appropriate times to see if any responses can be detected.

Phase IV: The Long-Term Effort

Assuming no signals have been found by the end of the third phase, the entire Cyclops concept should be reevaluated. If, as we expect, advancing knowledge has left the basic premises valid, the decision to build a long-range beacon must be faced. If this is done, the long-term search phase is begun. By now the target list will have been completely refined and updated, and both the Cyclops search and beacon systems can be almost totally automated. Cyclops would then revert to the search mode when not in use for other purposes.

It is to be hoped that, because other races have entered a phase IV, this phase may never be reached in our search!

STELLAR CORONA AND PLANETARY ARCHITECTURE STUDIES

Anything we can do to increase our knowledge of the shape of $f(S)$ (see equation (1)) will improve the search efficiency. If, for example, we could be sure that only

stars from *F*8 to *G*5 could support advanced life, our target list would be reduced to about one-third the length we have been assuming. We expect the cutoff of $f(S)$ at the high temperature end of the main sequence to occur at early type *F* stars since larger (and hotter) stars have lifetimes less than 3 billion years we assume are needed for the genesis and evolution of advanced life. At the lower end, the tidal braking of planetary rotation (because of the closeness to the primary star) may be the limiting factor and, according to Dole (ref. 1), would set the lower limit at about type *K*1 or *K*2 stars. We feel this lower limit is somewhat uncertain because tidal braking is enormously dependent on the extent, shape, and depth of ocean basins. It is extremely important to establish this low end cutoff more precisely, for 90% of all main sequence stars are type G5 and later. Since the frequency of stars by mass goes as $M^{-2.3}$ the position of the low end cutoff will greatly affect the length of our target list.

Another factor that may influence $f(S)$ considerably is the variation of coronal activity with spectral class. As noted in Chapter 2, the X-ray and UV flux of a star, which depends on coronal activity, is an important factor in atmospheric evolution. At present we have good knowledge of the coronal activity of only one *G* type star: our Sun. The visibility of normal stars with the Cyclops array is examined in Appendix R, where it is concluded that, if the microwave radiation of other stars is increased over their black-body radiation in the same ratio as for the *quiet* Sun, with a 3-km effective antenna diameter, the bandwidth and noise temperature of Cyclops, and a 1-min integration time, about a thousand stars could be detected at $\lambda = 10$ cm and about four times this many at $\lambda = 3$ cm. Thus, Cyclops should permit observations of the coronal activity of several thousand nearby stars on a time scale short enough to follow bursts and flares with good fidelity. This information would be a valuable contribution to astronomy and especially to the understanding of the evolution of planetary atmospheres.

There is also a possibility that with two large arrays, one the size of Cyclops, separated by several thousand kilometers, very long baseline interferometry could be done on normal stars and their positions determined to within 10^{-3} sec of arc or less. It is not clear that the absolute positions could be determined to this accuracy, but it might be possible to measure the relative positions of several stars this closely.

If so, Cyclops and its companion array would provide a tool for studying the architecture of planetary systems. At present this has been done optically only for Barnard's star. (See Chap. 2.) It would be of tremendous interest and importance if Cyclops could be used to obtain the data for enough other stars to begin to form a picture of the statistics of the architecture of planetary systems. Such a study would require decades of observing, because many planetary revolutions must be observed and the periods of major planets are apt to be quite long. Thus, we might make contact with intelligent life before such a study could be completed. This would give us far more direct and detailed information about at least one other planetary system than we could hope to get in any other way!

THE GALACTIC CENTER

If, for the reasons given in Chapter 2, interstellar communication is already a reality, the question arises: Is there any particularly likely direction in which to search for beacons? In this connection, Joshua Lederberg (private communication) has made an interesting suggestion. In a way it is much like Cocconi and Morrison's suggestion to use the hydrogen line in the spectrum. Lederberg asks: Is there any single point in the galaxy that is unique—that might be a natural cynosure for all races in the galaxy wherever they may be? He suggests the galactic center is that point.

Following this clue we might conjecture that either

1. The galactic superculture has constructed a powerful beacon at the galactic center (if indeed the center is accessible), or
2. Participating members of the galactic community are expected to radiate beacons not omnidirectionally but rather in beams directed away from (or toward) the galactic center. This greatly reduces the required beacon power or increases the range, or both.

We consider this suggestion to have merit and, while we have ruled out a blind search of the entire sky, we do recommend an area search over perhaps a one or two degree cone about the galactic center. (Such a search might also yield unexpected astronomical information.) We also recommend a similar search about the antipodal direction in case the strategy is to radiate toward the center so that the receiver will not be bothered by the high sky noise near the galactic center.

The difficulty with 2 is that it would tend to confine contact to a spoke or spokes radiating from the center since contact in a tangential direction would be difficult. Nevertheless it would be an appropriate and cheap strategy to attract new races, once interstellar communication had already spread around the galaxy through the use of omnidirectional beacons.

REFERENCES

1. Dole, Stephen H.: *Habitable Planets for Man*, Blaisdell (Ginn & Co.), New York.
2. Johnson, Harold L.: Interstellar Extinction in the Galaxy. *Astrophysical Journal, 141*, 1965, pp. 923-942.

14. CYCLOPS AS A RESEARCH TOOL

Throughout this first cut at the Cyclops system design, the sole intent has been to optimize its performance as an initial detector of intelligent transmissions. Nevertheless, it has characteristics that may make it a particularly desirable terminal for other deep space studies. Just as the JPL 210 ft Deep Space Facility has proved very useful in certain geodetic, radio and radar astronomical studies, there are research areas where Cyclops, once built, should have such special capability that it may be worth devoting a fraction of its operating time to these secondary research possibilities.

For example, consider a "3 km equivalent Cyclops."

Collecting area $= 7 \times 10^6$ m^2
$= 2200 \times$ (area of JPL – 210 ft)
$= 6700 \times$ (area of Haystack – 120 ft)

Angular resolution $\cong$ 1 arc sec at $\lambda = 3$ cm

Instantaneous frequency bandwidths from 0.1 Hz to 100 Mz

Rapid frequency-band change in the range 0.5 – 10 GHz

Ability to operate in various subarray formats

Three research fields where these properties should be particularly useful are discussed below.

DEEP-SPACE PROBES

The several-thousandfold increase in sensitivity and/or bit rate capacity over present facilities could alter the format of deep space exploration by probes. For instance, with no appreciable change in probe design, the same total information and bit rate would obtain for observations of Uranus (mean orbital radius = 19.2 AU) as are now possible for Mars (mean orbital radius = 1.52 AU). Alternatively, lighter and faster probes, using the same launch facilities, could examine the remote planets after appreciably shorter travel times than present facilities permit (5 to 10 years, or more). This is not the place to discuss possible tradeoffs in detail, but it would seem likely that if a major consideration in probe design were to change by a factor on the order of 10^3 a major change in probe strategy might result.

As another example, an important problem in solar system physics concerns the nature and position of the boundary between the solar wind and the interstellar plasma (ref.1). Pioneer F will be launched early in 1972 and should reach Jupiter in about 650 days. After Jupiter flyby, the probe will have a velocity sufficient to escape the solar system, drifting outward radially at about 2 AU/year. If the boundary of the heliosphere is 30-60 AU as some suspect, it will be a long time (in human terms) before we know the answers to our questions. Again, if Cyclops were available, it would seem worthwhile to consider a light, fast probe bearing minimal instrumentation, that could reach the heliospheric boundary in a much shorter time and relay back information for as long a time as it continued on out into the adjacent interstellar medium. Any experiment possible using the JPL 210 ft facility could be carried on at a distance $(2.21 \times 10^3)^{1/2} = 47$ times as great with Cyclops.

RADAR ASTRONOMY

Over the past decade, radar astronomy, using modest equipment and highly developed analytical techniques, has provided striking new data and insights in the field of solar system physics. Because of power and antenna limitations, its studies have been restricted in the main to nearby objects: the Moon, Venus, and Mars. With our 3-km assumption, using Cyclops only as a receiving terminal would extend the present radar capabilities by a factor of 7 to 9 in range, depending on whether the comparison is with the 210 ft or the 120 ft antennas. If the same total transmitter power presently used is installed in each element of the Cyclops array the range of current experiments could be extended by factors of

47 or 81. This would let us map Venus and Mars with a precision approaching that now obtained in studying the lunar surface. Also, we could map Pluto in the useful fashion now employed on Mars problems. At present, Pluto is beyond our reach. The high directivity of Cyclops would remove for all but the outermost planets the north-south hemisphere ambiguity of radar mapping.

Radar observations of the planets utilize little telescope time, so a byproduct of Cyclops could well be a complete radar study of the solar system out to Pluto.

RADIO ASTRONOMY

Throughout the brief history of radio astronomy, the principal instrumental efforts have been to increase power sensitivity and angular resolution of radio telescopes. Radiation from cosmic sources is weak at wavelengths accessible through the microwave window (see Figure 5-2), and the sources often exhibit a fine spatial structure down to milliseconds of arc and less. Therefore, large antennas are required to collect significant information about cosmic sources, and radio astronomers have struggled vigorously with the fiscal and technological problems involved. It is a fact that, despite exemplary ingenuity, radio astronomy is presently strongly inhibited by the lack of sufficiently large antenna systems.

An antenna array large enough to perform the Cyclops task will certainly be capable of important observations in radio astronomy. Even the minimum usable Cyclops array would represent an order-of-magnitude increase in sensitivity and resolution over available telescopes; and of seriously proposed radio telescopes, only the VLA (Very Large Array) of the National Radio Astronomy Observatory (NRAO) would be competitive with respect to resolution.

Some areas of astronomy in which Cyclops could be expected to yield valuable new data are outlined below, together with some discussion.

Discrete Source Studies

One of the oldest applications of radio astronomy techniques to fundamental astronomical studies is in cosmology, the study of the properties and history of the universe. It became apparent early in the history of the subject that some of the stronger discrete, extragalactic radio sources are at very great distances so that the radiation presently observed was emitted early in the history of the Universe. By studying the way in which the number of sources of a given flux varies with flux, it is possible to test various cosmological hypotheses. At present it is believed that the most distant detectable discrete radio sources represent violent events that occurred shortly after the apparent birth of the universe, so that essentially all of the past of the universe could be observed by sufficiently advanced radio telescopes. These intrinsically strong sources probably represent the strongest of all types of radio sources. Cyclops would probably be able to see other classes of sources out to the cosmological horizon; for example, normal galaxies of low radio luminosity and "radio-quiet" quasi-stellar objects. A broader roster of objects observable at the cosmological distance would greatly aid the selection among competing cosmologies. Further, with virtually all objects of cosmological significance observable at all distances, it might even become possible to measure distances directly in some way.

Other properties of discrete sources including continuum spectra, time variability, spatial structure, and polarization, could also be studied with a Cyclops system. The system should be able to improve upon the performance of existing telescopes in all these areas except spatial structure. Large correlator arrays, either existing or proposed, will have equal or better high-resolution mapping capability and are better optimized for radio astronomy. However, the existence of a single array with the sensitivity of Cyclops would immediately suggest its incorporation into a synthetic-aperture correlator array, in conjunction with a number of smaller antennas. A 3-km diameter Cyclops array, for example, combined with a distant antenna of, say, 30-m diameter would yield an interferometer with the equivalent area of two 300-m dishes. In other words, it would be possible to produce the world's most powerful interferometer by using a small, cheap dish in conjunction with the Cyclops array. If the two elements were situated at opposite ends of the continental United States and operated at $\lambda = 30$ cm, the resolution would be 10^{-2} sec of arc. If a series of 30-m dishes were established at intervals across the country, a source whose angular extent is one second of arc could be mapped completely with a resolution of 10^{-2} sec. The enormous sensitivity of such a system would permit mapping sources at least two orders of magnitude weaker than those accessible to present-day synthetic-aperture systems.

Spectroscopy

One of the most active areas of radio astronomy at present is the spectroscopy of molecular and atomic line radiations from the interstellar medium of the Galaxy. The most widely useful such spectral line is that of neutral atomic hydrogen at the 21 cm wavelength. The study of this line has revolutionized our understanding of the structure of our Galaxy and holds promise of much information about the structure of other galaxies.

Only a few such galaxies have been studied at this wavelength because existing telescopes lack the necessary sensitivity and resolution. Cyclops, however, would permit detailed analyses of the hydrogen distribution in galaxies, and perhaps in intergalactic space, out to distances of cosmological significance. The data processing system proposed for Cyclops would permit high resolution (1 Hz) spectroscopy over a 100 MHz band on a real-time basis, greatly speeding data acquisition and ensuring the discovery of weak lines. Similarly, the distribution of hydroxyl, and some of the more complex interstellar molecules in other galaxies could be determined out to cosmological distances. It is difficult to assess the impact on astronomy of this wealth of new data, but it would surely revolutionize astronomical thinking.

In the past the detection of interstellar ions, atoms, and molecules by radio techniques has been the result primarily of a process that began with theoretical estimates based on laboratory data. There have been a few accidental discoveries, but usually the radio astronomer has been provided with a careful estimate of just what narrow frequency range to search. This nearly complete reliance on prediction is due to the absence of truly broadband, high spectral resolving power equipment. As a result, many observable emission lines have probably gone undetected because of either a lack of understanding or imagination, or a lack of sufficiently precise laboratory data. It is often difficult or impossible to substitute laboratory procedures for the conditions existing in interstellar space. Because of its powerful spectral sensitivity, and remembering the parallel optical situations, Cyclops should be able to complement the predictive procedure by first observing the interstellar absorption and emission lines. There seems little reason to doubt the scientific value of employing such balanced procedures.

Pulsars

It seems well established that pulsars are neutron stars remaining after supernova explosions. Of the few dozen known in our Galaxy, most were discovered by their pulsed emissions, which are visible on telescope records above the receiver and sky background noise. The rate of discovery of pulsars has slowed almost to zero. Searching for weak, unknown pulsars without prior knowledge of their pulse periods is an excessively tedious process. It is necessary to perform an autocorrelation analysis or spectrum analysis of a long sample of signals over the possible range of pulse periods (or pulse repetition frequencies), for each point in the sky and each radio frequency band to be examined. An extremely sensitive antenna such as Cyclops, with the capability of multiple beams and sophisticated spectral analysis, would advance the search for pulsars dramatically. It has been estimated that there are 10^4 to 10^5 pulsars in our Galaxy and Cyclops could probably detect most if not all of them.

Pulsars undoubtedly exist in other galaxies but none has so far been detected. Studies of galactic and stellar evolution would benefit greatly if these extragalactic pulsars could be observed. Cyclops could detect most of the pulsars in the Virgo cluster of galaxies, for example, at a distance of 14 megaparsecs, assuming they have the characteristics of the well-known Crab pulsar.

Stars

In radio astronomy, only a few individual normal stars besides the Sun have been observed so far, and then just barely. Cyclops, because of its sensitivity, should be able to observe many of the nearer stars and add enormously to our knowledge of the physics of stellar coronas as a function of stellar type. Present knowledge is crude, relying as it does almost solely on extrapolation of solar data. Further, with the addition of suitable very long base-line interferometric (VLBI) capability, it should be possible to determine stellar diameters on the order of 10^{-3} arc-second or better. Furthermore, if a satisfactory background reference source exists in the field of view of the array element, it should be possible to study stellar motion and planetary perturbations to equal or better precision. Besides the fundamental astronomical interest in such observations, the planetary data would be of particular use in sharpening the Cyclops search strategy. At present, we have only crude estimates of the probability of occurrence of planets in the life zone of stars.

Equipment

There is another way in which radio astronomy could benefit from the prosecution of the Cyclops plan argued here, and that is in the matter of technological improvements. Radio astronomers have been prominent in the development of antenna and electronic circuit techniques, as well as in the art of data analysis. By and large, however, their efforts have been severely limited by insufficient engineering funds. Straightforward developments of the state of the art usually have had to wait upon the somewhat haphazard and occasional support provided by other space, military, or industrial projects. But, in principle, Cyclops requires the efficient and economic production of, and development of, just the type of equipment every radio observatory has or would very much like to have. (There are other fields where this is also true.) Therefore, a major spinoff of Cyclops

would be the relatively low cost availability of much more efficient antennas, receivers, spectrum analyzers, circuit components, and so forth. Not only astronomy but many other research areas and communication technology in general would profit enormously from the expenditure of Cyclops development funds.

SUMMARY

An antenna array with the characteristics of Cyclops could make possible many important advances in radio astronomy and could thus permit a significant increase in man's understanding of the universe. Observations now possible over interstellar distances would then be possible over intergalactic distances. Observations at cosmological distances, now limited to galaxies that are intense radio emitters, would be possible for a much wider class of sources.

Many of the researches enumerated above could be accomplished with far less expensive systems than Cyclops, and some of them no doubt will be. However, many of the observations, while theoretically *possible* with lesser instruments, would require a thousand to a million times more observing time and so cannot be considered to be practical. One of the merits of a Cyclops system, properly designed for flexibility, is that an enormous amount of astronomical data could be gathered with relatively little telescope time. Thus a great boost could be given to radio astronomy without subtracting more than 10% of the time from Cyclops's primary mission.

In spite of its potential value to radio astronomy, *proposals for a Cyclops program should not be allowed to interfere with the orderly development of instruments for radio astronomical instrumentation.* Should the Cyclops system ever emerge as a definite funded scheduled program, it would then be wise to plan other radio astronomy instrumentation around it, but not until then. It must be remembered that many radio astronomical instruments that may have been built in the interim will have their capabilities augmented by Cyclops, and those that will not are independently justifiable in any event.

REFERENCE

1. Dessler, A.J.; and Park, R.A.: The First Step Beyond the Solar System. American Astronautical Society reprint No. AAS-71-165.

15. CONCLUSIONS AND RECOMMENDATIONS

In this final chapter we attempt to summarize some of the major findings of the Cyclops study. Based on these, we suggest a course of action that could, over the next few years, bring about a general awareness in scientific circles as to the value of the search and of the specific techniques to be used, firm up all aspects of the system design, reduce the system cost, win popular support for the search, and thus pave the way for mankind to embark on one of the most exciting scientific adventures of all time.

Before listing our conclusions, it is perhaps not amiss to restate here some of the premises behind them. These are beliefs, based on present knowledge, that have evolved among scientifically trained people who have been studying the question of the prevalance of life on other worlds.

1. Planetary systems are the rule, not the exception. Most stars are now believed to possess planetary systems. These are a byproduct of the mechanism of stellar formation whenever the gas cloud out of which the star condenses is slowly rotating. Since the galaxy as a whole is rotating, most interstellar gas clouds are also.
2. Many planetary systems will contain at least one planet in the stellar ecosphere, where it will be kept warm by its star, but not too hot, and where its atmosphere can evolve from the primitive reducing one to an oxidizing Earth-like atmosphere.
3. Organic precursors of life are formed in abundance from the ingredients of a primordial or a tectonic atmosphere, but in any case are also found in space and on certain meteorites (the carbonaceous chondrites).
4. Main sequence stars cooler than *F*5 stars have lifetimes sufficiently long for Darwin-Wallace evolution to be effective. The ecospheres of stars cooler than spectral class *K*7, or perhaps *K*5, are so close to the star that tidal friction will either stop or greatly diminish the planet's rotation.
5. Intelligent life will have existed at some time during the life of the star on favorably situated planets circling a substantial fraction of the middle class main sequence stars in the universe.
6. The longevity of technologically advanced civilizations is unknown within orders of magnitudes. Other than assessing (while struggling to improve) our *own* chances for long-term survival, we know of no way to resolve this question, short of actually making contact with other races.
7. The establishment of interstellar contact may greatly prolong the life expectancy of the race that does so. Those races that have solved their ecological and sociological problems and are therefore very long lived may already be in mutual contact sharing an inconceivably vast pool of knowledge. Access to this "galactic heritage" may well prove to be the salvation of any race whose technological prowess qualifies it.

In view of the diminishing hope for finding other life in the solar system as our space probes gather fresh evidence of the barrenness of the Sun's other planets, it seems appropriate to ask how the search for life might best be carried to the other stars. The question is not a new one, but is timely. Many of our answers are not new, but represent confirmation of views already held by many. For these reasons, and because further study may modify many of our findings, we hesitate to call these findings conclusions. Nevertheless, the Cyclops study has gone into several aspects of the interstellar communication problem in greater detail than previous

studies, so we feel the following statements can be made with greater conviction than ever before.

CONCLUSIONS

1. It is vastly less expensive to look for and to send signals than to attempt contact by spaceship or by probes. This conclusion is based not on the present state of our technological prowess but on our present knowledge of physical law.
2. The order-of-magnitude uncertainty in the average distance between communicative civilizations in the galaxy strongly argues for an expandable search system. The search can be begun with the minimum system that would be effective for nearby stars. The system is then expanded and the search carried farther into space until success is achieved or a new search strategy is initiated.
3. Of all the communication means at our disposal, microwaves are the best. They are also the best for other races and for the same reasons. The energy required at these wavelengths is least and the necessary stabilities and collecting areas are fundamentally easier to realize and cheaper than at shorter wavelengths.
4. The best part of the microwave region is the low frequency end of the "microwave window"—frequencies from about 1 to 2 or 3 GHz. Again, this is because greater absolute frequency stability is possible there, the Doppler rates are lower, beamwidths are broader for a given gain, and collecting area is cheaper than at the high end of the window.
5. Nature has provided us with a rather narrow quiet band in this best part of the spectrum that seems especially marked for interstellar contact. It lies between the spectral lines of hydrogen (1420 MHz) and the hydroxyl radical (1662 MHz). Standing like the Om and the Um on either side of a gate, these two emissions of the disassociation products of water beckon all water-based life to search for its kind at the age-old meeting place of all species: the water hole.
6. It is technologically feasible today to build phased antenna arrays operable in the 1- to 3-GHz region with total collecting areas of 100 or more square kilometers. The Cyclops system is not nearly this large, but we see no *technological* limits that would prevent its expansion to such a size.
7. With antenna arrays equivalent to a single antenna a few kilometers in diameter at both the transmitting and receiving end, microwave communication is possible over *intergalactic* distances, and high-speed communication is possible over large interstellar distances. Thus rapid information transmission can occur once contact has been confirmed between two civilizations.
8. In the search phase we cannot count on receiving signals beamed at us by directive antennas. Neither can we afford to overlook this possibility. Beamed signals *may* be radiated at relatively low powers by communicative races to as many as a thousand nearby likely stars and for very long times. Long range beacons, intended to be detectable at any of the million or so likely stars within 1000 light-years, will probably be omnidirectional and very high powered ($> 10^9$ W).
9. Beacons will very likely be circularly polarized and will surely be highly monochromatic. Spectral widths of 1 Hz or less are probable. They will convey information at a slow rate and in a manner that does not seriously degrade their detectability. How best to respond will be contained in this information.
10. The efficient detection of beacons involves searching in the frequency domain with very high resolution (1 Hz or less). One of the major contributions of the Cyclops study is a data processing method that permits a 100 MHz frequency band to be searched simultaneously with a resolution of 0.1 Hz. The Cyclops system provides a receiver with a billion simultaneous narrow channel outputs. Although the Cyclops system bandwidth is 100 MHz, no very great technological barriers prevent widening it to 200 MHz. This would permit searching the entire "water hole" simultaneously. *If our conclusion as to the appropriateness of this band is correct, the problem posed by the frequency dimension of the search can be considered solved.*
11. The cost of a system capable of making an effective search, using the techniques we have considered, is on the order of 6 to 10 billion dollars, and this sum would be spent over a period of 10 to 15 years. If contact were achieved early in this period, we might either stop expanding the system or be encouraged to go on to make further contacts. The principal cost in the Cyclops design is in the antenna structures. Adopting an upper frequency limit of 3 GHz rather than 10 GHz could reduce the antenna cost by a factor of two.
12. The search will almost certainly take years,

perhaps decades and possibly centuries. To undertake so enduring a program requires not only that the search be highly automated, it requires a long term funding commitment. This in turn requires faith. Faith that the quest is worth the effort, faith that man will survive to reap the benefits of success, and faith that other races are, and have been, equally curious and determined to expand their horizons. We are almost certainly not the first intelligent species to undertake the search. The first races to do so undoubtedly followed their listening phase with long transmission epochs, and so have later races to enter the search. Their perseverance will be our greatest asset in our beginning listening phase.

13. The search for extraterrestrial intelligent life is a legitimate scientific undertaking and should be included as part of a comprehensive and balanced space program. We believe that the exploration of the solar system was and is a proper initial step in the space program but should not be considered its only ultimate goal. The quest for other intelligent life fires the popular imagination and might receive support from those critics who now question the value of landings on "dead" planets and moons.
14. A great deal more study of the problem and of the optimum system design should precede the commitment to fund the search program. However, *it is not too early to fund these studies.* Out of such studies would undoubtedly emerge a system with greater a capability-to-cost ratio than the first Cyclops design we have proposed.
15. The existence of more than one Cyclops-like system has such great value in providing complete sky coverage, continuous reception of detected signals, and in long base-line studies, that international cooperation should be solicited and encouraged by complete dissemination of information. The search should, after all, represent an effort of all mankind, not just of one country.

The above conclusions are the consensus of the Cyclops system group after studying the reports of the Antenna, Receiver, Transmission and Control, and Data Processing system design groups. While many in these groups would agree with the conclusions stated above, others might disagree with certain points and it would be unfair to represent these points as a unanimous consensus of the entire team. the same comments apply to the recommendations, which follow.

RECOMMENDATIONS:

1. Establish thc search for extraterrestrial intelligent life as an ongoing part of the total NASA space program, with its own budget and funding.
2. Establish an office of research and development in techniques for communication with extraterrestrial intelligence. Appoint a director and a small initial staff.
3. Take steps to protect the "water hole." Through the FCC and corresponding international agencies, use of the spectrum from 1.4 to 1.7 GHz should be limited to interstellar communication purposes. The hydrogen line is already protected. All that is needed is to extend this protection upward in frequency to include the hydroxyl line.
4. Establish, perhaps through the National Academies of Science and Engineering, an advisory committee consisting of interested astronomers, radio astronomers, engineers, physicists, exobiologists, and appropriate specialists. The advisory committee should have the initial responsibility for reviewing the available material on the subject, including this report, and of recommending an appropriate course of action. Assuming the committee concurs that further investigations should be undertaken, it should have the responsibility to see that the necessary preliminary scientific studies and engineering design and development are carried out in an orderly manner over a 3 to 5 year period.
5. Make use of outside study contracts initially, but gradually build up internal design and system analysis teams to provide competent contract review and creative in-house (NASA) leadership.
6. As the various systematic and strategic problems of the search yield to continued study and the overall feasibility approaches general acceptability, begin a series of releases to the scientific community and to the general public to stimulate interest in, and appreciation of, the value of the search.
7. Establish at the outset a policy of open liaison with comparable groups in other countries, that there be no classified information and that all reports be publicly available.
8. When all systemic and strategic problems have been solved, a go-no-go decision must be made. If "go", then political support must be obtained for the funding. The funding must be on a long term basis so that construction, once started, is not interrupted and can proceed in an orderly way.
9. Make it clear that the system will be available for

a certain fraction of the time during the search phase for radio astronomy research and for other space programs.

10. Establish the policy of reporting publicly all advances made through the use of the facility. Produce educational films on its capabilities and its mission, and conduct tours of the facility for the public, to sustain interest and develop a popular sense of participation in, and identification with, the search.

APPENDIX A

ASTRONOMICAL DATA

DISTANCE UNITS

Astronomical unit (AU)	=	1.496×10^{11} m
(= semimajor axis of Earth's orbit)	=	499 light-sec.
Parsec (pc) = 206,265 AU	=	3.086×10^{16} m
Light-year (LY)	=	9.46×10^{15} m

STELLAR BRIGHTNESS AND LUMINOSITY

The *magnitude scale* of stellar brightness dates from antiquity. In modern usage, five magnitude steps correspond to a brightness ratio of 100:1; thus, one visual magnitude step is a brightness ratio of $(100)^{1/5} = 2.512$. The higher the magnitude, the dimmer the star.

The *absolute visual magnitude* (M_v) of a star is the apparent magnitude it would have at a distance of 10 pc.

The *luminosity* of a star (L_*) is its total power output. The luminosity of the sun is 3.9×10^{26} watts.

The *bolometric magnitude* of a star is a logarithmic measure of its luminosity. The bolometric and visual absolute magnitude are related by a *bolometric correction* (M_v-M_{bol}), which depends on spectral type and has the values shown in table A-1 for main sequence stars. The bolometric magnitude of the Sun is 4.7; thus, the luminosity of a star and its bolometric magnitude are related by

$$L_* = (3.9 \times 10^{26})\,(100)^{(4.7 - M_{bol})/5} \text{ watts}$$

$$= (3 \times 10^{28})\, 10^{-0.4 M_{bol}} \text{ watts}$$

$$M_{bol} = 2.5 \log \frac{3 \times 10^{28}}{L_*}$$

One step in bolometric magnitude is −4 dB in luminosity.

TABLE A-1

STELLAR TEMPERATURE, LUMINOSITIES, AND MAGNITUDE VS. SPECTRAL TYPE

Main Sequence Spectral Type	Te (°K)	L/L_0	M_{bol}	M_v	M_v-M_{bol}
O5	35,000	1.3×10^6	−10.6	− 6	4.6
B0	21,000	3.6×10^4	− 6.7	− 3.7	3.
B5	13,500	760.	− 2.5	− 0.9	1.6
A0	9,700	76.	0	0.7	0.68
A5	8,100	16.	1.7	2.0	0.30
F0	7,200	6.3	2.7	2.8	0.10
F5	6,500	2.3	3.8	3.8	0.00
G0	6,000	1.1	4.6	4.6	0.03
G5	5,400	.7	5.1	5.2	0.10
K0	4,700	.36	5.8	6.0	0.20
K5	4,000	.14	6.8	7.4	0.58
M0	3,300	.07	7.6	8.9	1.20
M5	2,600	.009	9.8	12.0	2.1

SUN

Radius ($R_\odot$)	$= 6.96\times10^{8}$ m	Angular velocity (16° latitude)	$= 2.86\times10^{-6}$ rad/s
Area	$= 6.09\times10^{18}$ m^2	Moment of inertia	$= 6\times10^{46}$ kg m^2
Luminosity ($L_\odot$)	$= 3.9\times10^{26}$ W	Angular momentum ($H_\odot$)	$\approx 1.7\times10^{41}$ kg m^2/s*
Radiation per unit area	$= 6.4\times10^{7}$ W/m^2	Rotational energy	$= 2.5\times10^{35}$ J
Effective surface temperature	$= 5800°$ K	Inclination of axis	$= 7\frac{1}{4}°$
Mass ($M_\odot$) $= 3.32\times10^{5}\ M_\oplus$	$= 1.99\times10^{30}$ kg		

*Uncertain because rotation rate of *interior* of sun may be greater than surface.

EARTH

Equatorial radius (a)	$= 6.378\times10^{6}$ m	Thermal fluxes:	
Polar radius (c)	$= 6.357\times10^{6}$ m	Solar	$\approx 10^{17}$ W
Mean radius $(a^2c)^{1/3}$	$= 6.371\times10^{6}$ m	Geothermal	$\approx 2.5\times10^{13}$ W
Surface area	$= 5.1\times10^{14}$ m^2	Tidal friction	$\approx 3.4\times10^{12}$ W
Volume	$= 1.083\times10^{21}$ m^3	Coal burning	$\approx 2\times10^{12}$ W
Mass ($M_\oplus$)	$= 5.977\times10^{24}$ kg	Oil burning	$\approx 3\times10^{12}$ W
Angular velocity	$= 7.29\times10^{-5}$ rad/s	Natural gas burning	$\approx 1.4\times10^{12}$ W
Angular momentum	$= 5.86\times10^{33}$ kg m^2/s	Nuclear power (planned)	$\approx 0.3\times10^{12}$ W
Rotational energy	$= 2.138\times10^{29}$ J		
Mean orbital velocity	$= 2.978\times10^{4}$ m/s	Total artificial	$\approx 6.7\times10^{12}$ W
Solar constant (above atmosphere)	$= 1.388\times10^{3}$ W/m^2	Temperature increase due to "thermal pollution"	$< 0.005°$C

SOLAR SYSTEM

Total mass of planets	$= 447.9\ M_\oplus$	Total angular momentum of planetary system $= 185\ H_\odot$	$= 3.15\times10^{43}$ kg m^2/s
Total mass of satellites	$= 0.12\ M_\oplus$	Total planetary rotational energy	$= 0.7\times10^{35}$ J
Total mass of asteroids	$= 3\times10^{-4}\ M_\oplus$	Total orbital kinetic energy	$= 2\times10^{35}$ J
Total mass of meteoric matter	$= 5\times10^{-10}\ M_\oplus$	$M_\odot/\Sigma M$	$= 0.9986$
Total mass of planetary system	$= 448\ M_\oplus$	$H_\odot/\Sigma H$	≈ 0.0054

PLANETS

Planet	Radius ($R_\oplus = 1$)	Mass ($M_\oplus = 1$)	Density	Period of Rotation	Inclination of Axis
Mercury	0.38	0.054	5.4	58.6d	?
Venus	0.96	0.815	5.1	242.9d	157°
Earth	1.0	1.0	5.52	$23^h56^m4^s$	23°27′
Mars	0.53	0.108	3.97	$24^h37^m23^s$	23°59′
Jupiter	11.19	317.8	1.334	9^h50^m	3°05′
Saturn	9.47	95.2	0.684	10^h14^m	26°44′
Uranus	3.73	14.5	1.60	10^h49^m	97°55′
Neptune	3.49	17.2	2.25	15^h	28°48′
Pluto	0.47	0.8 (?)	?	6^h4	
Moon	0.273	0.0123	3.34	27.322^d	5°8′43″*

*Inclination of orbit to ecliptic.

PLANETARY ORBITS

Planet	Semimajor Axis (AU)	Sidereal Period (yr)	Orbital Velocity (km/s)	Eccentricity	Inclination to Ecliptic
Mercury	0.387	0.241	47.9	0.2056	7°0′11″
Venus	0.723	0.615	35.05	0.0068	3°23′37″
Earth	1.0	1.0	29.8	0.0167	
Mars	1.524	1.88	24.14	0.0933	1°51′1″
Jupiter	5.203	11.86	13.06	0.0483	1°18′31″
Saturn	9.54	29.46	9.65	0.0559	2°29′33″
Uranus	19.18	84.01	6.08	0.0471	0°46′21″
Neptune	30.07	164.8	5.43	0.0085	1°46′45″
Pluto	39.44	248.4	4.74	0.2494	17°10′

STAR DENSITIES IN THE SOLAR NEIGHBORHOOD

	Giants and Supergiants		Main Sequence		White Dwarfs	
Class	Number/pc^3	Percent	Number/pc^3	Percent	Number/pc^3	Percent
O			2.5×10^{-8}	0.00003		
B			1×10^{-4}	0.13	1.26×10^{-3}	1.6
A			5×10^{-4}	0.66	2×10^{-3}	2.6
F	5×10^{-5}	0.07	2.5×10^{-3}	3.3	1.26×10^{-3}	1.6
G	1.6×10^{-4}	0.21	6.3×10^{-3}	8.3	6.3×10^{-4}	0.8
K	4×10^{-4}	0.52	1×10^{-2}	13.1	1×10^{-3}	1.3
M	3×10^{-5}	0.04	5×10^{-2}	65.6		
Total	6.4×10^{-4}	0.84	7×10^{-2}	91.	6.1×10^{-3}	8.0
Total *F*, *G*, *K* Main Sequence			1.9×10^{-2}	25.		

NOTE: Individual values are accurate to ±10%, at best. Percentages do not add exactly because of rounding errors.

GALAXY

Diameter of disk	≈30 kpc = 100,000 LY	Absolute magnitude (as seen from outside Galaxy, normal to disk)	≈ −20.5
Thickness at center	≈ 4 kpc = 13,000 LY		
Total mass	= $1.1\times10^{11}\ M_\odot$	Luminosity (same direction)	≈ $1.3\times10^{10}\ L_\odot$
Sun's distance from galactic center	= 30,000 LY		
Sun's distance from galactic plane	= 26 ± 40 LY (north)	Density limit of all matter in solar neighborhood (Oort limit)	= $0.14\ M_\odot$/pc^3
Rotational velocity in solar neighborhood	≈ 215 km/s		
Rotational period at solar neighborhood	≈ 236×10^6 years	Density due to stars in solar neighborhood	≈ $0.057\ M_\odot$/pc^3
Age of galaxy	≈ 12×10^9 years	Total number of stars	≈ 5×10^{11}

Rate of recession of galaxies: Values of the Hubble constant have been reported ranging from 50 to 120 km/sec Mpc. The primary uncertainty is not the determination of the redshift, but the scale of cosmological distance. The most recent value reported by Sandage is

$$H_0 = 55 \pm 7 \text{ km/sec Mpc}$$

$$= (1.78 \pm 0.22) \times 10^{-18}/\text{sec}$$

$$= (5.62 \pm 0.72) \times 10^{-11}/\text{yr}$$

If the recession speed is assumed constant, the age of the universe would be $t_0 = H_0^{-1}$ as measured from the initial singularity. The value of H_0 given above yields

$$15.8 \times 10^9 \leqslant t_0 \leqslant 20.4 \times 10^9 \text{ years}$$

Radius of observable universe $= c/H_0$

$= 16$ to 20×10^9 LY.

Galaxies in observable universe $\gtrsim 3 \times 10^9$.

APPENDIX B

SUPERCIVILIZATIONS AND CULTURAL LONGEVITY

The Cyclops system, described in the body of this report, is based on what Freeman Dyson has called the "orthodox view" of interstellar communication. This he succinctly describes in these words:

> Life is common in the universe. There are many habitable planets, each sheltering its brood of living creatures. Many of the inhabited worlds develop intelligence and an interest in communicating with other intelligent creatures. It makes sense then to listen for radio messages from out there, and to transmit messages in return. It makes no sense to think of visiting alien societies beyond the solar system, nor to think of being visited by them. The maximum contact between alien societies is a slow[1] and benign exchange of messages, an exchange carrying only information and wisdom around the galaxy, not conflict and turmoil.

While this view is probably the dominant one, there are competing views. In this appendix we mention some of these and examine their premises and consequences.

The arguments presented in Chapter 2 for the prevalence of life in the universe are based on our emerging knowledge of the evolution of the physical universe, particularly of stars and their planetary systems, and of the factors leading to the origin of life on suitable planets. The latter knowledge concerning life and particularly that concerning the evolution of intelligence is totally geocentric for the simple reason that we have no direct knowledge of any other life. Nevertheless, one example is better than none, and the evolution of intelligence on Earth is considered to be indicative of what would very likely happen elsewhere.

It is when we consider not the past but the future course of evolution that the narrative turns into almost pure speculation and major differences of opinion are apt to arise. For here we have no hieroglyphs nor fossils to guide us and can only extrapolate existing technological and sociological trends as we perceive them. In this attempt the physical scientist, except that he is less apt to violate natural law in his predictions, has little advantage over the science fiction writer, and both may be poorer prophets than the sociologist or anthropologist.

If we admit the future worlds of science fiction into consideration (and we see no reason to exclude some of them as less likely than others seriously proposed by scientists) we are confronted with a hopeless task of assessment. We can do no more here than mention a few categories that have commanded serious attention.

KARDASHEV'S CIVILIZATION TYPES

Noting that historically the technological advances of our civilization have been accompanied by (or made possible by) a greater per capita energy budget, the Soviet astrophysicist N.S. Kardashev (ref. 1) has suggested that societies might be classified according to the amount of energy they are capable of harnessing for their purposes. Kardashev's object was to arrive at the power levels advanced cultures might use for communication, but his three classifications apply to total energy usage:

Type I Civilizations have mastered the energy resources of their planets. Our present civilization with its power consumption of about 6.6×10^{12} watts, or our near-fu-

[1] Slow in the sense that the mails are slow, that is the transit time is long. As shown in Chapter 5 the *information rate*, even with our present technology, can be enormous.

ture society with its mastery of controlled nuclear fusion, would fall in this category.

Type II Civilizations are capable of utilizing a substantial fraction of the radiation of their parent star so have powers on the order of 10^{26} watts at their disposal.

Type III Civilizations are extended communities with the ability to control powers comparable to the radiation of an entire galaxy—that is, powers on the order of 10^{37} watts.

Kardashev reasons that any civilization could afford to devote a small fraction, say one percent, of its energy resources to interstellar communication, and this leads him to postulate extremely powerful radiations from Type II and III civilizations. This has the advantage of making detection very easy for mere Type I civilizations, such as ourselves, and obviates the need for expensive receiving arrays of antennas. Alas, so far no such powerful radiations of artificial origin have been detected, so Type II and III civilizations remain hypothetical. Nevertheless, Kardashev's classifications are useful reference terms in discussing supercivilizations.

DYSON CIVILIZATIONS

Freeman Dyson (refs. 2,3) has suggested that the pressure of population growth will have forced many advanced societies to create more living space in their planetary systems by disassembling unfavorably situated planets and redistributing their matter in various ways about the parent star. Dyson points out that the mass of Jupiter, if distributed in a spherical shell at 2 AU from the Sun, would have a surface density of about 200 gm/cm^2 (actually 168 gm/cm^2) and, depending on the density, would be from 2 to 3 m thick. He goes on to say: "A shell of this thickness could be made comfortably habitable and could contain all the machinery required for exploiting the solar radiation falling onto it from the inside." When it was pointed out that such a shell would be dynamically unstable[2] he replied that what he really had in mind was a swarm of independent objects orbiting the star. In a subsequent paper (ref. 3), he proposes that these objects be lightweight structures up to 10^6 km in diameter, the limit being set by solar tide raising forces, and notes that at 1 AU from the Sun, 200,000 of these (actually 360,000) would be needed to intercept and thus utilize all the Sun's radiation.

One consequence of this would be that a substantial fraction of the Sun's luminosity would be reradiated from an extended source at about 300°K rather than a much smaller source at 5800°K. On this basis, Dyson feels that we are more apt to detect advanced civilizations because of the excess infrared radiation they produce in pursuit of their own survival than as a result of intentional beacon signals they might radiate.

Although Dyson describes an entertaining mechanism for the disassembly of planets to obtain the material for lightweight orbiting structures, no details are given as to how these structures are to be made habitable. Presumably, since these lightweight structures would not have enough gravity to hold an external atmosphere, the advanced beings are to live' inside. To fill 200,000 spheres each 10^6 km in diameter with air at normal Earth atmospheric pressure would require about 1.36×10^{32} kg of air, or about 50,000 times the total mass of the Sun's planets. The air would have to be supported against contraction under its own gravity; otherwise, only the central region would be habitable, or (with enough air added to fill the sphere) the object would become a massive star. Since, even with support, the total air mass per sphere is about 100 Earth masses, the supporting structure could hardly be the lightweight affair Dyson describes.

We conclude that the size limit of Dyson's spheres is more apt to be set by the amount of air available and by the self-gravity effects it produces than by tidal forces. If all the mass of the Sun's planets were in the form of air at atmospheric pressure, this air would fill a spherical shell 1 AU in radius and 7.3 km thick. Thus, instead of 360,000 spheres each 10^6 km in diameter, we would need more than 4×10^{15} spheres each less than 10 km in diameter to catch all the Sun's light at 1 AU. These considerations cause us to be skeptical of Dyson's latest model and to base our calculations of excess IR radiation (given in Chap. 4) on the redistribution of the heavy element mass of the solar system into several new earths rather than 10^{15} "mobile homes" in orbit at 1AU from the Sun.[3] Even this seems to us a formidable enough undertaking. We note, however, that Dyson appears so convinced of the detectability and inevitability of the kind of astroengineering he describes that he construes our failure to detect any such activities as evidence for the absence of advanced intelligent life!

[2] The total *heavy element* content of the Sun's planets would allow a sphere at 2 AU radius around the sun to be only about 1 cm thick. If rotating, the sphere would flatten and collapse; if stationary, the Sun's gravity would cause compressive stresses of about 300,000 lb/in^2 in the shell, ensuring buckling and collapse. With only one solar gravity at 2 AU (1.48×10^{-3} m/s^2) no atmosphere would remain on the eternally dark outside, while anything on the inside would gently fall into the Sun. It is hard to see how Dyson finds these conditions "comfortably habitable."

[3] The excess IR radiation from this vast number of spheres would be indistinguishable from that produced by a lot of dust around a star. Gaps in the orbital pattern would cause some direct starlight to filter through, not in occasional flashes as with fewer 10^6 km diameter spheres, but in a fairly steady amount.

ABIOLOGICAL CIVILIZATIONS

Many writers have imagined the end result of biological evolution to be the creation of and ultimate dominance of inorganic intelligence: super robotic computers capable of self replication and of design improvement generation to generation. Hoyle and Elliot hint at this in their fiction *A for Andromeda* and *Andromeda Breakthrough.* Another example is Harry Bates' *Farewell to the Master.*

It is very difficult to assess the probability and consequences of this kind of evolution. Perhaps the most pertinent comment is that, given the premise of an artificial intelligence at least equal to its biological precursor, it would be rather immaterial whether that intelligence were artificial or not. Also it seems very unlikely that an advanced natural species would voluntarily abdicate its dominant position to an artificial species. Artificial intelligent civilizations would therefore seem more likely to represent extensions in time or space, or both, of their biological forebears, complete with their historical perspectives and many of their motivations. Such extensions might provide a vicarious form of space travel or a way of reproducing the original biological species on suitable planets of other stars.

SOCIALLY AND AESTHETICALLY ADVANCED CIVILIZATIONS

In all the "advanced" civilizations discussed above, the state of advancement was measured by technological prowess; in fact, the assumption was tacit that the two were synonymous. A very different school of thought holds this to be an absurdity, pointing out that some very primitive tribes are quite civilized in their customs while technologically advanced cultures are often wantonly cruel toward one another. These authors picture advanced cultures as races of beings who have long ago mastered all the secrets of the universe and have a tremendous technology at their disposal, but who subordinate this technology and instead emphasize the spiritual and artistic aspects of life. Their communities are deceptively simple, indeed, almost pastoral; their lives are full of love and beauty, but woe betide the barbarous invader who considers them weak or naive.

Although this theme is common in future fiction it is, in a sense, underrepresented in our literature. To the vast majority of people a society in peace, a society full of trust and security, a society without hunger or misery is a far more advanced society in terms of human values than the most powerful technologically advanced society imaginable. The writers who see progress as measured only in terms of kilowatts consumed per person, or planets conquered, are not reflecting the concerns of most people on this earth.

Yet underneath these apparently idyllic societies that cherish all living things and protect the weak lurks the specter of stasis: of arrested evolution. As H.G. Wells observed in *The War of the Worlds,* man has won his birthright to Earth through a billion deaths, the deaths of those unfit to survive and breed. How to shoulder nature's responsibility for natural selection in a compassionate society is a problem mankind has not yet solved, but must solve if genetic evolution is to continue. If truly socially advanced societies have existed for thousands of millenia, the techniques by which they have maintained genetic evolution would be priceless information that could ensure our own survival and development.

THE REAL NATURE OF PROGRESS

The above types of advanced societies by no means exhaust the list of those that have been conjectured but are sufficient to illustrate one point. A plausible case can be made for almost any kind of advanced society by emphasizing one, or a few, aspects of life and ignoring or underrating other aspects. Thus, the kind of prediction one makes tends to reflect extrapolations of certain trends one sees and considers most important. Kardashev, in his classifications, equates advancement with mastery over energy resources. Dyson appears haunted by Malthusian principles and apparently considers astroengineering a simpler solution than birth control, albeit a temporary one, since no space colonization program whatever can outstrip an *exponential* growth rate. The forecasters of robotic worlds are enchanted with the current explosive advance in computers and in particular with the possibilities of artificial intelligence and carry this theme to various "logical" conclusions. Finally, the protagonists for utopian societies ask what purpose would be served by spreading human misery across the galaxy and suggest instead that the kingdom of heaven is within us.

About all that can be said for sure about such prophesies is that, however stimulating they may be, they are almost certainly all wrong. To become convinced of this we need merely observe how unpredictable our own progress has been over the last two millenia. What ancient Greek or Alexandrian would have predicted the dark ages, the discovery of the new world, or the atomic era? Who among the ancients, wise as they were in many ways, would have forecast the automobile, television, antibiotics, or the modern computer? To Aristotle, the fact that men could do sums was proof that they had souls. Yet here we sit attempting

predictions about worlds not two thousand but hundreds of thousands or even millions of years beyond us, and of independent origin at that!

While uncannily accurate prophesies have been made over decades by extrapolation, this technique is of no avail over very long times. Current trends grow and then vanish to be replaced by others. Old questions may never be answered but instead be discovered to be meaningless. An age of religion dissolves into an age of science, but science alone may not answer all the problems of the world and the forefront of progress may shift again. Cultural values change. It is only within the last century and a half at the most that man has made a serious attempt to discover and preserve the relics of his past, and only within the last decade that he has begun to show serious concern for preserving his future by preserving Earth itself. This growing societal consciousness may mark a turning point, may herald a decreased emphasis on physical science and technology for their own sake and an increasing emphasis on using our knowledge to assure the longevity of the human race. We may be entering an era of research and development in better social forms and progress toward more rational mores. We may find ourselves grappling with the potentialities and dangers of genetic engineering.

Thus we see human progress (or change, at least) occurring first in one area then another with the emphasis shifting as new needs and new possibilities become apparent. In the short term one aspect of human knowledge may jump ahead, but in the long run an overall balance tends to be struck. For this reason, and because totally unforeseeable discoveries appear as apt to alter drastically the future of the world as they have the past, we must assign any *particular* model of an advanced society a very low probability of being a typical end result.

PASSIVE VS ACTIVE SEARCH STRATEGIES

One consequence of infatuation with a particular model of an advanced society is that it may lead to an ineffective strategy for interstellar contact. If we seriously accept Kardashev's Type II civilizations, for example, we need do nothing but wait for the day when very powerful radiations are detected by existing astronomical instruments. If we believe Dyson's sphere builders to be the rule, then, as he suggests, anomalous IR radiation might provide a clue to the whereabouts of supercultures. If we regard robotic extensions of advanced societies as likely perhaps we should expect hordes of self replicating probes à la Bracewell to be expanding across the galaxy from various centers. Or if we consider a self-centered utopian culture to be the ultimate end of life we may conclude that all electromagnetic search would be fruitless.

It seems very evident that if we decide on a passive strategy, if we wait to be discovered or wait for evidence of life on a stellar (Type II) scale, we are placing *all* the burden of contact on extraterrestrial life and not doing our share of the job. We are forcing on other life a far more formidable and expensive task than the operation of a few comparatively modest beacons. Further we are betting on a model, or models, of advanced societies that may not, in fact, exist.

Rather than base our hopes on models that represent extreme extrapolations into the future, it would appear far more productive to *apply the assumption of mediocrity to our known present capabilities.* Although others probably have already advanced far beyond our present state, as we hopefully also will, it seems reasonable to assume that:

1. Most civilizations *at some point in their development* perceive the likelihood of other life in the universe, as we do, and find themselves technically able to search for and signal that life, as we are.
2. Many civilizations decide on an active search strategy.
3. Many of the earliest to do so followed their search phase with a radiative phase and were subsequently detected.
4. Success in one contact led to
    a. Great intellectual excitement and social benefits.
    b. A sustained effort at further contacts.
5. As a late entry, we are heirs to the fruits of all successful past efforts, including beacons to attract our attention.

This line of reasoning leads to an active strategy in which we qualify for contact by investing our share in the large receiving system needed for a *balanced* search effort (see Chap. 6).

Arguments that lead to a passive search strategy are cheap in the sense that they require no funding, or very little. No matter how outrageous the premises, they are apt to be acceptable because of this comforting fact. On the other hand, premises that lead to an active search program costing billions of dollars are immediately subject to close scrutiny point by point. This is very understandable, but if we are to proceed rationally *we should be careful not to use one set of credibility criteria for do-nothing arguments and another more severe set for arguments that solicit action.* Furthermore, we must not only weigh the cost of action against the possible benefits to be realized, we must also give some thought to the cost of inaction.

INTERSTELLAR CONTACT AND CULTURAL LONGEVITY

The age-old struggle of each species against rival species and natural enemies has had its counterpart in wars of territorial and cultural expansion. Today man faces a new threat as the dominant species of this planet. In Pogo's words: "We have met the enemy and he is us." Man must either sublimate the basic drive of uninhibited growth, converting it into a quest for quality rather than quantity, and assume responsibility for the husbandry of the planet he dominates, or die a cultural failure killing his own kind.

We are almost surely not the first culture in the Galaxy to face this problem; in fact, it would appear that our present situation is the direct result of technological mastery over our environment and would therefore confront all life forms at our stage of development. If our own case is typical the problems of overpopulation arise and require urgent attention long before the culture is able to engage in Dyson's feats of planetary remodeling, so other solutions must be the rule and Dyson's mobile homes must be rare exceptions.

We are, however, at the point where we could attempt interstellar contact. Such contact would most likely occur with cultures that have already faced and solved our immediate problem. Those we detect will probably have achieved stability and avoided their own extinction. If we can learn through such communication the ways that have proved effective in assuring long-term social stability, it is likely that our own chances would be significantly increased. For this reason alone, the cost of establishing interstellar contact would be more than justified.

If indeed a galactic community of cultures exists, it is this community that might be expected to have individual cultural life expectancies measured in aeons rather than millennia. The long time delays of the communication would allow (in fact would force) continued individual innovation by each culture, but the accumulated wisdom of the group would serve to guide each member society. Finally, the pride of identification with this supersociety and of contributing to its long-term purposes would add new dimensions to our own lives on Earth that no man can imagine.

REFERENCES

1. Kardashev, N. S., Soviet Astronomy–A.J. (Astronomicheskhi Zhurnal) Vol. 8, Page 217 (1964).
2. Interstellar Communication, A.G.W. Cameron (ed) W.A. Benjamin, Inc., New York (1963), pages 111-114.
3. Dyson, F.J., The Search for Extraterrestrial Technology, Perspectives in Modern Physics, R. E. Marshak (ed), John Wiley & Sons (1966), page 641.

APPENDIX C

OPTIMUM DETECTION AND FILTERING

In communication systems we are frequently faced with the problem of obtaining the best estimate of the amplitude of a signal having a known shape, but which has had noise added to it. Historically, this problem was first analyzed[1] in connection with the detection of the amplitudes and time of occurrence of radar pulses. We begin by considering pulse detection. The results, however, will be quite general, and applicable to a wide variety of signals.

Assume that a pulse $f(t)$ recurs periodically and that on each recurrence we wish to measure its amplitude as accurately as possible in spite of added noise. One of the most straightforward ways to do this is to multiply the signal by a "gate" that lets the signal through, but excludes noise before and after the pulse, and then integrate the gated signal. Thus if $n_k(t)$ is the particular noise wave on the kth cycle, and $g(t)$ is the gate, we form the gated signal

$$s(t) = g(t)\ [f(t) + n_k(t)] \tag{C1}$$

and integrate the result to obtain a charge

$$q(k) = \int_{-\infty}^{\infty} g(t)[f(t) + n_k(t)]\ dt \tag{C2}$$

This charge may be regarded as the sum of a component

$$q_s = \int_{-\infty}^{\infty} g(t)f(t)\ dt \tag{C3}$$

due to the signal, and a component

$$q_n(k) = \int_{-\infty}^{\infty} g(t)n_k(t)\ dt \tag{C4}$$

due to the noise. The integration actually occurs only over the time for which $g(t) \neq 0$, but the limits may be taken as infinite without affecting the result. We assume that the various $n_k(t)$ are uncorrelated and are samples of a statistically stationary process. The mean square value of $q_n(k)$ is then

$$\overline{q_n^2} = \frac{Av}{k}\left[\int_{-\infty}^{\infty} g(t)n_k(t)\ dt\right]^2 \tag{C5}$$

where Av/k means the average over the index k.

Using Parseval's theorem, we may rewrite equation (C4) as

$$q_s = \int_{-\infty}^{\infty} \overline{G(\nu)}F(\nu)d\nu \tag{C6}$$

Equation (C5) may be written as the product of two independent integrals

$$\overline{q_n^2} = \frac{Av}{k}\left[\int_{-\infty}^{\infty} g(t)n_k(t)dt \int_{-\infty}^{\infty} g(\tau)n_k(\tau)d\tau\right]$$

$$= \frac{Av}{k}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t)g(\tau)n_k(t)n_k(\tau)dt\ d\tau\right] \tag{C7}$$

The change of variable $\tau = t + x$, $d\tau = dx$ may now be

[1] By Claude Shannon, informal communication in 1941

introduced and the order of the integrating and averaging operations exchanged to give:

$$\overline{q_n^2} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(t)\, g(t+x)\, \frac{Av}{k}\left[n_k(t)\, n_k(t+x)\right] dt\, dx \tag{C8}$$

The average over k of $n_k(t)\, n_k(t + x)$ is the autocorrelation, $A(x)$, of the noise and is independent of t. Thus,

$$\overline{q_n^2} = \int_{-\infty}^{\infty} A(x) \left[\int_{-\infty}^{\infty} g(t)g(t+x)dt\right] dx \tag{C9}$$

$$= \int_{-\infty}^{\infty} A(x)\left[\int_{-\infty}^{\infty} |G(v)|^2\, e^{i2\pi vx}\, dv\right] dx \tag{10}$$

where equation (C10) follows from equation (9) by applying Parseval's theorem to the second integral. Now $A(x)$ is the transform of the power spectrum of the noise, while the quantity in brackets in equation (C10) is the inverse transform of $|G(v)|^2$. Applying Parseval's theorem once more equation (C10) becomes

$$\overline{q_n^2} = \int_{-\infty}^{\infty} \frac{\psi(v)}{2}\, |G(v)|^2\, dv \tag{C11}$$

where $\psi(v)$ is the one-sided power spectral density of the noise. Since our integrals extend from $-\infty$ to $+\infty$ in frequency, we must use the two sided density $\psi(v)/2$.

The problem may now be stated as follows: Find the $G(v)$ that minimizes $\overline{q_n^2}$ as given by equation (C11) under the constraint that q_s as given by equation (C6) is held constant. This is a straightforward problem in the calculus of variations and yields the result

$$G(v) = m\, \frac{F(v)}{\psi(v)} \tag{C12}$$

where m is an arbitrary real constant. If the noise is "white"–that is, if $\psi(v)$ = constant, then $G(v)$ = constant $\times$ $F(v)$ and

$$g(t) = \text{constant} \times f(t) \tag{C13}$$

In the presence of white noise the best gate is the signal itself. From equation (C2) we see that what we should do is cross correlate the received signal with the expected waveform of the noise-free signal. In function space this amounts to finding the projection of $s(t)$ on $f(t)$.

From equations (C6) and (C11) and using the optimum gate equation (C12), we obtain the output signal-to-noise power ratio

$$\left(\frac{S}{N}\right)_{\text{opt}} = \frac{q_s^2}{\overline{q}_n^2} = 2\int_{-\infty}^{\infty} \frac{|F(v)|^2}{\psi(v)}\, dv \tag{C14}$$

If $\psi(v)$ is constant over the signal spectrum, this ratio is expressed as

$$\left(\frac{S}{N}\right)_{\text{opt}} = \frac{2E}{\psi} \tag{C15}$$

where E is the pulse energy, and, for thermal noise, $\psi = kT$.

If, prior to gating, the received signal and noise are first passed through a filter having the transmission $K(v)$, then in all subsequent expressions $F(v)$ is replaced by $F(v)\, K(v)$ and $\psi(v)$ by $\psi(v)\, |K(v)|^2$. Equation (C12) in particular then becomes

$$G(v)\, \overline{K(v)} = m\, \frac{F(v)}{\psi(v)} \tag{C16}$$

and we observe that it is only the *product* of the gate spectrum and the conjugate of the filter transmission that is specified. Deficiencies in either can be compensated for by the the other. A case of particular interest occurs if we set $G(v)$ = constant. This corresponds to a gate that is a δ–function at $t = 0$. We then have

$$K(v) = \text{constant}\, \frac{\overline{F(v)}}{\psi(v)} \tag{C17}$$

and for white noise

$$K(v) = \text{constant} \times \overline{F(v)} \tag{C18}$$

As given by equation (C17) or (C18) $K(v)$ is called a *matched filter* and gives the highest ratio of peak signal

to rms noise. We note that the matched filter achieves this result by (1) removing all phase dispersion (lining up all frequencies in the signal so that their peaks occur together at $t = 0$), and (2) weighting each component directly in proportion to the signal component amplitude to noise power spectral density at that frequency.

When $f(t)$ is an RF or IF signal rather than a baseband signal, the relations (C12) and (C16) require that $g(t)$ also be a RF or IF signal. The gating then amounts to a kind of generalized homodyning operation which excludes the noise power that is in quadrature with the carrier. One cannot use a matched filter at RF or IF as specified by equation (C17) or (C18) followed by, say, a linear or square law detector and achieve the S/N ratios given by equations (C14) and (C15). The gate at $t = 0$ is needed to exclude the quadrature noise power. The most practical method of realizing full matched filter performance is usually to homodyne the received signals to baseband and do all gating and filtering in that frequency range. In this way, the use of nonlinear detectors is avoided. (See Appendix D on square-law detection theory.)

APPENDIX D

SQUARE LAW DETECTION THEORY

A square-law detector is a square-law *device* followed by a *filter* that transmits a desired range of frequencies as the output. Thus the output of a square-law detector is a filtered measure of the instantaneous power incident on the square-law device. When the incident power is in the radio frequency band, the output filter usually excludes harmonics of the radio frequencies, and thereby averages the incident power over a time comparable to one RF cycle. At optical frequencies, the square-law device is a photon detector whose output is generally taken to be proportional to the product of the amplitude of the incident wave with its complex conjugate. This product contains no double frequency component, and output filtering may be omitted. In either the radio frequency or optical detector, the postdetection filtering may be much narrower in bandwidth than any predetection filtering so that the output is averaged over times long compared with the reciprocal of the input bandwidth, let alone the center frequency of this band.

Shot noise is usually negligible in radio-frequency detectors, but may be dominant in photodetectors. Fluctuation noise may or may not be important in photodetection. We shall include both sources of noise in order to derive more generally applicable results. Although we shall use a photodetector as a model of a square-law device, the analysis applies equally well to radio frequency devices, provided only one polarization mode is assumed.

Assume that a coherent monochromatic power P_r of frequency ν_0 in a particular polarization falls on a photodetector of quantum efficiency η. Assume that a total incoherent power P_0 in a band of frequencies centered at ν_0 also falls on the detector. The total power P_0 may comprise a part P_S representing an incoherent signal received from space, which we wish to measure, and a part P_n representing receiver noise or dark current—so that $P_0 = P_S + P_n$. We represent the *amplitude* of the coherent wave as $\sqrt{2}\,Ae^{i2\pi\nu_0 t}$ and the *amplitudes* of the components of the incoherent wave as $\sqrt{2}\,[a_1(t) + a_2(t) + ib_1(t) + ib_2(t)]\; e^{i2\pi\nu_0 t}$, where the subscript 1 indicates a component of the same polarization as the coherent wave and the subscript 2 indicates a component of orthogonal polarization. The a's and b's are gaussian variables of zero mean and zero cross correlation. The instantaneous power falling on the photodetector is half the sum of the products of the total amplitudes for each polarization, and each time phase, with their complex conjugates, and is given by

$$P = [A + a_1(t)]^2 + a_2^2(t) + b_1^2(t) + b_2^2(t)$$

$$= A^2 + 2Aa_1(t) + \left[a_1^2(t) + a_2^2(t) + b_1^2(t) + b_2^2(t)\right]$$

$$= A^2 + 2Aa_1(t) + P_i \qquad \text{(D1)}$$

where P_i is the instantaneous incoherent noise power; that is, $\overline{P_i} = P_0$.

We take the probability per unit time of the emission of a photoelectron to be proportional to P. The photo current will thus contain shot noise, which we will include later. For the present, while we are considering fluctuation noise, we will deal in *expected* values and write $i = \alpha P$, where $\alpha = \eta q/h\nu$. The average photocurrent is thus from equation (D1)

$$\overline{i} = I_r + I_0 = \alpha(P_r + P_0) = \alpha(P_r + P_S + P_n) \qquad \text{(D2)}$$

If we are trying to detect the coherent power, only

the component I_r constitutes the signal. If, on the other hand, we are measuring starlight (or incoherent radio sources), we might have $P_r = 0$ and take $I_s = \alpha P_s$ to be the signal. The accuracy with which either I_r or I_0 can be measured depends on the magnitude of the fluctuations.

The fluctuation noise power of the current i is given by:

$$\begin{aligned}\overline{\Delta i_f^2} &\equiv \overline{i^2} - (\bar{i})^2 \\ &= \alpha^2 \left[\overline{[A^2 + 2Aa_1(t) + P_i]^2} - [A^2 + P_0]^2\right] \\ &= \alpha^2 \left[4A^2\overline{a_1^2(t)} + (\overline{P_i^2} - P_0^2)\right] \qquad \text{(D3)}\end{aligned}$$

As will be shown later, P_i is Boltzmann distributed; that is,

$$p(P_i) = \frac{e^{-P_i/P_0}}{P_0} \qquad \text{(D4)}$$

and therefore $\overline{P_i^2} - P_0^2 = P_0^2$. If, as we have assumed, P_0 is randomly polarized, then $\overline{a_1^2(t)} = P_0/4$. At radio frequencies, or at optical frequencies when a polarizing filter is used, only the polarization corresponding to the coherent signal would strike the detector. Thus $a_2(t)$ and $b_2(t)$ would be zero and $\overline{a_1^2(t)} = P_0/2$. In general then, $\overline{a_1^2(t)} = P_0/2m$ where $m = 1,2$ is the number of orthogonal polarizations reaching the detector. Making these substitutions, we find

$$\overline{\Delta i_f^2} = \alpha^2 \left[\frac{2}{m} P_r P_0 + P_0^2\right] \qquad \text{(D5)}$$

To determine the effect on the noise power of filtering the output, we need to know the power spectrum of $\overline{\Delta i_f^2}$. The first term in the brackets arises from components of the incoherent wave at frequencies $\nu_0 - f$ and $\nu_0 + f$ mixing with the coherent wave to produce outputs at frequency f. If $\psi(\nu)$ is the power spectrum of the incoherent source, and $K(\nu) = |F(\nu)|^2$ is the *power* transmission of the predetection filter, then the power spectrum of the first term will be proportional to $[\psi(\nu_0 - f)K(\nu_0 - f) + \psi(\nu_0 + f)K(\nu_0 + f)]$. If the filter is symmetrical about ν_0 then $K(\nu_0 + f) = K(\nu_0 - f) \equiv H(f)$, and if $\psi(\nu)$ is either constant or varies linearly with ν over the band, this expression becomes simply $H(f)\ [\psi(\nu_0 - f) + \psi(\nu_0 + f)] = 2\ \psi(\nu_0)H(f)$. Thus the power spectrum of the first term is proportional to the low pass equivalent of the predetection filter.

The second term in the brackets in equation (D5) arises from noise components mixing with one another. All pairs of components separated by f beat to form an output at frequency f. Assuming again that $K(\nu)$ is symmetrical about ν_0 and that $\psi(\nu)$ is linear in ν, the power spectrum of this term will be proportional to the autocorrelation of $H(f)$, which we represent by $H(f) * H(f)$ or simply $H*H$.

To these two components of equation (D5) we may now add the shot noise power of the current $\bar{i}$, which on a two-sided spectrum basis has the spectral density $q\bar{i} = \alpha(h\nu/\eta)$. Thus we find for the total (two-sided) power spectrum

$$\gamma(f) = \alpha^2 \left[\frac{h\nu}{\eta}(P_r + P_0) + \frac{2}{m} P_r P_0 \frac{H(f)}{\int H\,df} + P_0^2 \frac{H*H}{\int H*H\,df}\right] \qquad \text{(D6)}$$

If $G(f)$ is the (power) transmission of the postdetection filter, then the total noise power in the output is simply $N = \int G(f)\gamma(f)\,df$, while the signal power is either $S = \alpha^2 G(0)\ P_s^2$ or $S = \alpha^2 G(0)\ P_r^2$ depending upon whether we wish to measure P_s or detect P_r.

For the measurement of incoherent radiation in the absence of coherent radiation, we set $P_r = 0$ in (6) and obtain for the output signal-to-noise power ratio

$$\frac{S}{N} = \frac{G(0)\,P_s^2}{P_0 \dfrac{h\nu}{\eta} \int G\,df + P_0 \dfrac{\int G(H*H)df}{\int H*H\,df}} \qquad \text{(D7)}$$

For the detection of the coherent signal in the incoherent background:

$$\frac{S}{N} = \frac{G(0)\,P_r^2}{(P_r + P_0)\dfrac{h\nu}{\eta}\int G\,df + \dfrac{2}{m} P_r P_0 \dfrac{\int G\,H\,df}{\int H\,df} + P_0^2 \dfrac{\int G(H*H)df}{\int H*H\,df}} \qquad \text{(D8)}$$

If the output filter is very wide compared with $H(f)$ and $H*H$ so that $G(f) \approx G(0)$ over the significant range of integration, then for the incoherent case equation (D7) becomes

$$\frac{S}{N} = \frac{P_S^2}{P_0 \frac{h\nu}{\eta} \frac{\int G\, df}{G(0)} + P_0^2} \qquad \text{(D9)}$$

while for the coherent case, equation (D8) becomes

$$\frac{S}{N} = \frac{P_r^2}{(P_r + P_0) \frac{h\nu}{\eta} \frac{\int G\, df}{G(0)} + \frac{2}{m} P_r P_0 + P_0^2} \qquad \text{(D10)}$$

If the output filter is very narrow compared with $H(f)$ and $H*H$ so that $H(f) \approx H(0)$ and $H*H = [H*H]_{f=0}$ over the significant range of integration, then for the incoherent case equation (D7) becomes

$$\frac{S}{N} = \frac{G(0) P_S^2}{\left[P_0 \frac{h\nu}{\eta} + P_0^2 \frac{(H*H)_{f=0}}{\int H*H\, df}\right] \int G\, df} \qquad \text{(D11)}$$

while for the coherent case, equation (D8) becomes:

$$\frac{S}{N} = \frac{G(0) P_r^2}{\left[(P_r + P_0) \frac{h\nu}{\eta} + 2P_r P_0 \frac{H(0)}{\int H\, df} + P_0^2 \frac{(H*H)_{f=0}}{\int H*H\, df}\right] \int G\, df} \qquad \text{(D12)}$$

S/N RATIOS IN THE DETECTION OF INCOHERENT RADIATION

We will now use the general expressions derived in the last section to find the signal-to-noise ratios in certain specific cases of interest. In the usual radio-frequency situation P_r and P_0 will have been amplified to substantial power levels before detection and shot noise may be ignored; we ignore it in some cases but include it in those expressions that may apply to photon detection as well.

No Postdetection filtering

If shot noise is negligible; expression (D5) becomes simply

$$\frac{S}{N} = \frac{P_S^2}{P_0^2} = \left(\frac{P_S}{P_S + P_n}\right)^2 \qquad \text{(D13)}$$

which is always less than unity unless there is no receiver noise. The receiver bandwidth and band shape do not affect the S/N ratio.

Ideal RF Bandpass Filter of Width B, Ideal Low Pass Postdetection Filter of Cutoff Frequency W

In this case:

$$H(f) = \begin{cases} 1, |f| < B/2 \\ 0, |f| > B/2 \end{cases} ; \quad H*H = \begin{cases} B - |f|, |f| < B \\ 0\ , |f| > B \end{cases}$$

If $W > B$, the case of no postdetection filtering applies, since the output filter is too wide to have any effect.

If $W < B$, and we neglect shot noise we find from equation (D7)

$$\frac{S}{N} = \left(\frac{P_S}{P_0}\right)^2 \frac{B}{W} \frac{1}{2 - (W/B)} \qquad \text{(D14)}$$

For $W \ll B$ the signal-to-noise ratio is improved over case of no filtering by the factor $B/2W$ which is the ratio of the bandwidth of the low pass equivalent of the RF filter to the bandwidth of the postdetection filter.

Ideal RF Bandpass Filter of Width B, Output Averaged for a Time τ

If an output filter is used and n *independent* samples of the output taken $1/B$ sec apart are averaged then equation (D13) applies and

$$\frac{S}{N} = n\left(\frac{P_S}{P_0}\right)^2 = (B\tau)\left(\frac{P_S}{P_0}\right)^2 \qquad \text{(D15)}$$

If instead we use an output filter whose *power* transmission is $G(f) = [(\sin \pi\tau f)/\pi\tau f]^2$, the filtered output will at all times be the average of the unfiltered output for the last τ seconds. We then find from equation (D7) ignoring shot noise,

$$\frac{S}{N} = \frac{B^2 P_S^{\,2}}{2P_0^{\,2}\int_0^B (B-f)\,\frac{\sin^2 \pi\tau f}{(\pi\tau f)^2}\,df} \qquad \text{(D16)}$$

In the limit when $\tau >> 1/B$, this expression approaches (D15) so the continuous averaging offers no improvement in this case over the discrete average.

Sin x/x RF Filter, Output Averaged for a Time τ

Here we take $H(f) = [(\sin \pi Tf)/\pi Tf]^2$ so that the frequency interval between the first nulls of the RF filter is $2/T$. This RF filter takes a running average of the input signal that is T sec long.

If no output filter is used and n *independent* samples of the output taken T seconds apart are averaged so that $\tau = nT$, then equation (D13) again applies and

$$\frac{S}{N} = n\left(\frac{P_S}{P_0}\right)^2 = \frac{\tau}{T}\left(\frac{P_S}{P_0}\right)^2 \qquad \text{(D17)}$$

If instead we take a *running* average of the output with the same filter used in the preceding case, we now have

$$H*H = \frac{\sin^2(\pi Tf)}{(\pi Tf)^2} * \frac{\sin^2(\pi Tf)}{(\pi Tf)^2}$$

$$= \frac{1}{(\pi Tf)^2}\,\frac{1}{T}\left[1 - \frac{\sin(2\pi Tf)}{2\pi Tf}\right]$$

$$[H*H]_{f=0} = \frac{2}{3T}$$

$$\int H*H\,df = \frac{1}{T^2}$$

and we obtain from equation (D7) neglecting shot noise:

$$\frac{S}{N} = \frac{P_S^{\,2}}{P_0^{\,2} T \int \frac{\sin^2 \pi\tau f}{(\pi\tau f)^2}\left(1 - \frac{\sin 2\pi Tf}{2\pi Tf}\right)\frac{df}{(\pi Tf)^2}}$$

If $\tau >> T$ we can use equation (D11) and obtain

$$\frac{S}{N} = \frac{3\tau}{2T}\left(\frac{P_S}{P_0}\right)^2 \qquad \text{(D18)}$$

In this case there *is* an improvement from the continuous averaging. Although the discrete samples averaged were statistically independent in the last two cases, the intermediate samples in the third case are so highly correlated with those already taken that no improvement results from their inclusion, whereas in the fourth case the intermediate samples are less correlated and their inclusion improves the S/N ratio.

S/N RATIOS IN THE DETECTION OF COHERENT RADIATION

The specific combinations of pre- and postdetection filters treated in the last section are presented here once again, but the object now is the detection of the coherent radiation in the presence of the incoherent background.

No Postdetection Filtering

Again neglecting shot noise, and taking $m = 1$ equation (D10) becomes

$$\frac{S}{N} = \frac{P_r^{\,2}}{2P_rP_0 + P_0^{\,2}}$$

$$= \frac{(P_r/P_0)^2}{1 + 2(P_r/P_0)} \qquad \text{(D19)}$$

Since P_r/P_0 is the *input* signal-to-noise ratio, we see that for input signal-to-noise ratios much greater than unity, the square-law detector degrades the input ratio by a factor of two (3 dB). At very low input signal-to-noise ratios we have $(S/N) \approx (P_r/P_0)^2$, and the unfiltered square-law detector doubles the (negative) signal-to-noise ratio expressed in decibels. Signals rapidly become undetectable as they drop below unity input signal-to-noise ratio.

Ideal RF Bandpass Filter of Width B, Ideal Low Pass Postdetection Filter of Cutoff Frequency W

If $W \geq B$ the output filter has no effect except to limit shot noise, and we have from equation (D10):

$$\frac{S}{N} = \frac{P_r^{\,2}}{2(P_r + P_0)\,\frac{h\nu W}{\eta} + \frac{2}{m}\,P_rP_0 + P_0^{\,2}} \qquad \text{(D20)}$$

If $B/2 \leqslant W < B$, the last term in the denominator equation (D8) is also affected and we find

$$\frac{S}{N} = \frac{P_r^2}{2(P_r + P_0)\frac{h\nu W}{\eta} + \frac{2}{m}P_r P_0 + P_0^2\left(2\frac{W}{B} - \frac{W^2}{B^2}\right)} \tag{D21}$$

Finally, if $W < B/2$, all terms in the denominator of equation (D8) are affected and

$$\frac{S}{N} = \frac{B}{2W}\frac{P_r^2}{(P_r + P_0)\frac{h\nu W}{\eta} + \frac{2P_0P_r}{m} + P_0^2\left(1 - \frac{W}{2B}\right)} \tag{D22}$$

The factor $B/2W$ is essentially the number of independent samples that are averaged at any time.

Ideal RF Bandpass Filter of Width B, Output Averaged for a Time τ

If no output filter is used and n independent samples are taken at intervals of $1/B$ sec then equation (D19) applies and

$$\frac{S}{N} = n\frac{P_r^2}{\frac{2}{m}P_rP_0 + P_0^2} = (B\tau)\frac{P_r^2}{\frac{2}{m}P_rP_0 + P_0^2} \tag{D23}$$

If instead we use the output filter

$$G(f) = [(\sin \pi\tau f)/\pi\tau f]^2$$

which performs a running average over the last τ sec, then equation (D8) becomes

$$\frac{S}{N} = P_r^2\left[(P_r + P_0)\frac{h\nu}{\eta\tau} + \frac{4P_rP_0}{mB}\int_0^{B/2}\frac{\sin^2\pi\tau f}{(\pi\tau f)^2}\,df + \frac{2P_0^2}{B^2}\int_0^B (B-f)\frac{\sin^2\pi\tau f}{(\pi\tau f)^2}\,df\right]^{-1} \tag{D24}$$

For $\tau >> 1/B$, we use equation (D12) and find

$$\frac{S}{N} = (B\tau)\frac{P_r^2}{(P_0 + P_r)\frac{h\nu B}{\eta} + \frac{2}{m}P_0P_r + P_0^2} \tag{D25}$$

which is the same as (D23) except for the inclusion of the shot noise term. This expression may also be written as

$$\frac{S}{N} = \frac{\tau P_r^2}{\frac{h\nu}{\eta}(P_0 + P_r) + \frac{P_0}{B}\left(\frac{2}{m}P_r + P_0\right)} \tag{D26}$$

in which form it is clear that the ratio of $h\nu/\eta$ to the spectral density P_0/B of the incoherent power determines whether shot noise or fluctuation noise dominates.

Sin x/x RF filter, Output Averaged for a Time τ

As in the coherent case we take $H(f) = [(\sin \pi Tf)/\pi Tf]^2$. If we use no output filter and average n independent samples taken T seconds apart we again find from equation (D19).

$$\frac{S}{N} = n\frac{P_r^2}{\frac{2}{m}P_rP_0 + P_0^2} = \frac{\tau}{T}\frac{P_r^2}{\frac{2}{m}P_rP_0 + P_0^2} \tag{D27}$$

as in (D23) but with $n = \tau/T$ rather than $(B\tau)$.

If instead we use the output filter of the preceding case to obtain a running average, we find from equation (D8) that:

$$\frac{S}{N} = P_r^2\left[(P_r + P_0)\frac{h\nu}{\eta\tau} + \frac{2P_rP_0T}{m}\int_{-\infty}^{\infty}\frac{\sin^2\pi Tf}{(\pi Tf)^2}\frac{\sin^2\pi\tau f}{(\pi\tau f)^2}\,df + P_0^2T\int_{-\infty}^{\infty}\frac{\sin^2\pi\tau f}{(\pi\tau f)^2}\left(1 - \frac{\sin^2\pi}{2\pi Tf}\right)\frac{df}{(\pi Tf)^2}\right]^{-1} \tag{D28}$$

Assuming that $\tau >> T$ we can start instead with equation (D12) and obtain

$$\frac{S}{N} = \frac{\tau}{T}\frac{P_r^2}{(P_r + P_0)\frac{h\nu}{\eta} + \frac{2}{m}P_rP_0 + \frac{2}{3}P_0^2} \tag{D29}$$

Comparing this result with (D27) we see that, as in the incoherent detection case, going to continuous averaging has reduced the P_0^2 term in the denominator by the factor 2/3.

STATISTICS OF THE SQUARE LAW DETECTOR OUTPUT (CLASSICAL CASE)

So far we have considered only the signal-to-noise power ratios of the square-law detector output. These are useful in computing the accuracy with which a received power can be measured; but to determine the false alarm probability p_{fa} (of detecting a signal when none is present) or the probability of missing a signal p_{ms} that is actually present, we need to know the actual statistical distribution functions.

Again we assume a coherent power P_r and an incoherent background power P_0. We assume from here on that the latter is present in only one polarization, since this is true for the radio case and can be true for the optical case with the use of a polarizing filter. Then the instantaneous power at the detector input is

$$P = [A + a(t)]^2 + b^2(t) \qquad \text{(D30)}$$

where as before $P_r = A^2$, $P_0 = \overline{a^2}(t) + \overline{b^2}(t)$, and $a(t)$ and $b(t)$ are gaussian variables with zero mean and zero cross correlation. The vector relations between $a(t)$, $b(t)$, their resultant $c(t)$, the coherent signal amplitude A, and the total signal amplitude $s(t)$ are shown in the Figure D-1.

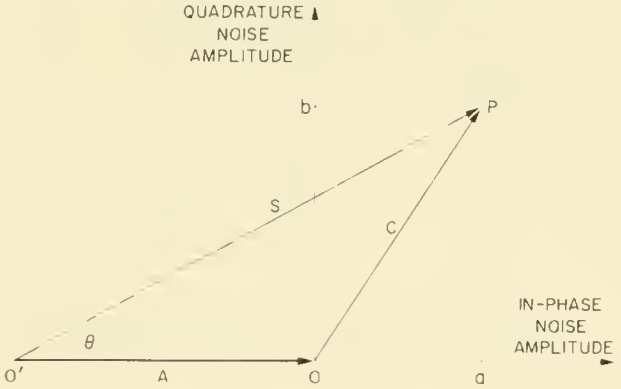


Figure D-1. *Noise and signal vectors. P has a two-dimensional Gaussian distribution about O, which makes C have a Rayleigh distribution in amplitude, and S have the distribution given by equation (D34).*

The probability density function for a is

$$p(a) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-a^2/2\sigma^2} \qquad \text{(D31)}$$

Similarly,

$$p(b) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-b^2/2\sigma^2} \qquad \text{(D32)}$$

where $\sigma^2 = \overline{a^2}(t) = \overline{b^2}(t) = P_0/2$. Since a and b are independent, the joint probability density distribution for a and b is

$$p(a,b) = p(a)p(b) = \frac{e^{-(a^2 + b^2)/2\sigma^2}}{2\pi\sigma^2} \qquad \text{(D33)}$$

Since $a^2 + b^2 = c^2 = A^2 + s^2 - 2sA\cos\theta$, the probability density function for s is found by integrating equation (D33) over all θ, that is:

$$p(s) = \frac{1}{\pi P_0}\int_0^{2\pi} s\, e^{-\frac{(A^2 + s^2 - 2sA\cos\theta)}{P_0}}\, d\theta$$

$$= \frac{2s}{P_0}\, e^{-\frac{A^2 + s^2}{P_0}}\, \frac{1}{\pi}\int_0^{\pi} e^{\frac{2sA}{P_0}\cos\theta}\, d\theta$$

$$= \frac{2s}{P_0}\, e^{-\frac{A^2 + s^2}{P_0}}\, I_0\left(\frac{2sA}{P_0}\right) \qquad \text{(D34)}$$

where I_0 is the zero order modified Bessel function of the first kind. We note in passing that $p(s)$ is the probability density function for the output of a linear detector and that $p(s)/2\pi s$ is the probability density versus radius distribution for the Golay detector. (See Chap. 11).

If $q(P)$ is the probability density function for the instantaneous power P, we require that

$$q(P)\, dP = p(s)ds$$

with $dP = 2s\, ds$ and therefore find

$$q(P) = \frac{1}{P_0}\, e^{-\frac{P_r + P}{P_0}}\, I_0\left(2\,\frac{\sqrt{P_r P}}{P_0}\right) \qquad \text{(D35)}$$

Let us now introduce the normalized variables $x = P/P_0$ and $r = P_r/P_0$. Then equation (D35) becomes

$$p(x) = e^{-(r+x)} I_0(2\sqrt{rx}) \qquad \text{(D36)}$$

If we add n *independent* samples of the output the resulting distribution will be the n-fold convolution of $p(x)$. The Fourier transform of $p(x)$ is given by Campbell and Foster (ref. 1, pair 655.1) as

$$C(\omega) = \frac{e^{-r}}{1 + i\omega} \exp\left(\frac{r}{1 + i\omega}\right)$$

Thus

$$[C(\omega)]^n = \frac{e^{-nr}}{(1 + i\omega)^n} \exp\left(\frac{nr}{1 + i\omega}\right)$$

and by pair 650, op. cit., we find the inverse transform to be

$$p_n(x) = e^{-nr} \left(\frac{x}{nr}\right)^{\frac{n-1}{2}} e^{-x} I_{n-1}(2\sqrt{nrx}) \qquad \text{(D37)}$$

If now we replace the sum x by the arithmetic mean $y = x/n$, we obtain

$$p_n(y) = n\left(\frac{y}{r}\right)^{\frac{n-1}{2}} e^{-n(r+y)} I_{n-1}(2n\sqrt{ry}) \qquad \text{(D38)}$$

Now

$$I_{n-1}(2n\sqrt{ry}) = (n\sqrt{ry})^{n-1} \sum_{k=0}^{\infty} \frac{(n^2ry)^k}{k!(n-1+k)!}$$

so

$$p_n(y) = \frac{n^n y^{n-1} e^{-n(r+y)}}{(n-1)!} \left[1 + \frac{n^2ry}{1 \cdot n} \times \left(1 + \frac{n^2ry}{2(n+1)} \left(1 + \frac{n^2ry}{3(n+2)} \left(1 + \ldots\right.\right.\right.\right.$$

$$\left.\left.\left.\left.\right)\right)\right)\right] \qquad \text{(D39)}$$

For large n, we may replace $(n-1)!$ by

$$\frac{n!}{n} = \sqrt{\frac{2\pi}{n}}\, n^n\, e^{-n + \frac{1}{12n}}$$

giving

$$p_n(y) = \frac{n}{2\pi}\, y^{-n-1}\, e^{-n\left(r - 1 + \frac{1}{12n^2} + y\right)} \times \left[1 + \frac{n^2ry}{1 \cdot n} \left(1 + \frac{n^2ry}{2(n+1)} \left(1 + \text{etc.}\right)\right)\right] \qquad \text{(D39a)}$$

Actually equation (D39a) is in error by only 0.2% for $n = 1$ and so may be used to compute $p_n(y)$ for any n.

The probability that y is less than a certain threshold y_T may be found by numerically integrating equation (D39a) from $y = 0$ to $y = y_T$. This integral gives the probability that signal plus noise fails to exceed the threshold y_T and hence that the signal is not detected.

In the absence of a signal, $r = 0$, and equation (D39) becomes

$$p_n(y) = n\frac{(ny)^{n-1} e^{-ny}}{(n-1)!} \qquad \text{(D40)}$$

This may now be integrated from y_T to infinity to give the probability $q_n(y_T)$ that noise alone exceeds the threshold y_T. We find

$$q_n(y_T) = e^{-ny_T} \sum_{k=0}^{n-1} \frac{(ny_T)^k}{k!} \qquad \text{(D41)}$$

The threshold level required to give a specified p_{fa} is found from equation (D41). The signal-to-noise ratio, r, needed to give a permissible probability of missing the signal is then found by integrating equation (D39a) from $y = 0$ to $y = y_T$ with trial values of r.

GAUSSIAN APPROXIMATIONS

From the central limit theorem we know that as $n \to \infty$ both (D38) and (D40) will approach gaussian distributions. Reverting to unnormalized variables, the distribution approached by the signal plus noise will be centered at $P = P_0 + P_r$ and will have a variance

$$\sigma = \frac{\sqrt{2P_0P_r + \mu P_0^2}}{\sqrt{n}} \qquad \text{(D42)}$$

while that approached by the noise alone will be centered at $P = P_0$ and will have the variance

$$\sigma_0 = \frac{\sqrt{\mu P_0}}{\sqrt{n}} \qquad \text{(D43)}$$

where $\mu = 1$ if n discrete independent samples are averaged and $\mu = 2/3$ if sin x/x filters are used with $n = \tau/T$. See eqs. (D18) and (D29).

The threshold must exceed the mean, P_0, of the noise distribution by some multiple m_0 times σ_0 where m_0 is chosen to give the desired p_{fa}, and is determined by

$$p_{fa} = \frac{1}{\sqrt{2\pi}} \int_{m_0}^{\infty} e^{-z^2/2} dz \qquad \text{(D44)}$$

The mean, $P_0 + P_r$, of the distribution with signal present must be some multiple m times σ above the threshold, where m is chosen to give the allowable probability of missing the signal, and is determined by

$$p_{ms} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{m} e^{-z^2/2} dz \qquad \text{(D45)}$$

The limit of satisfactory operation is thus given by

$$P_r = m_0\sigma_0 + m\sigma = m_0 \frac{\sqrt{\mu} P_0}{\sqrt{n}} + m \frac{\sqrt{2P_0P_r + \mu P_0^2}}{\sqrt{n}}$$

or:

$$\frac{P_r}{P_0} = \frac{(m^2 + \sqrt{\mu n}\, m_0) + m\sqrt{m^2 + 2\sqrt{\mu n}\, m_0 + \mu n}}{n} \qquad \text{(D46)}$$

If $p_{fa} = p_{ms}$, then $m_0 = m$ and

$$\frac{P_r}{P_0} = 2\left(\frac{m^2}{n} + m\sqrt{\frac{\mu}{n}}\right) \qquad \text{(D47)}$$

PHOTON-COUNTING STATISTICS

At optical frequencies we may elect to use wide band circuits following the photocell and simply count the number of photons received. The statistical distribution of this count will depend on the signal and background signal strengths and on the number of Nyquist intervals over which the count is taken — that is, on the value of $n = (B\tau)$ where B is the bandwidth of the (ideal) optical filter ahead of the photodetector and τ is the integration time.

Incoherent Radiation: Short Integration Time

If the integration time is short compared with $1/B$, the probability $p(n)$ of receiving n photons is the exponential distribution

$$p(n) = \frac{1}{1 + \bar{n}_0}\left(\frac{\bar{n}_0}{1 + \bar{n}_0}\right)^n \qquad \text{(D48)}$$

where $\bar{n}_0$ = expectation count. The first moment of this distribution is $M_1 = \bar{n}_0$, while the second moment is $M_2 = \bar{n}_0\,(1 + 2\bar{n}_0)$. The mean square fluctuation in count is therefore

$$\sigma_0^2 = M_2 - M_1^2 = \bar{n}_0(1 + \bar{n}_0) \qquad \text{(D49)}$$

If the count $\bar{n}_0 = \bar{n}_b + \bar{n}_S$, where $\bar{n}_b$ is a background count and n_S is the count due to the received radiation, the signal-to-noise power ratio is $\bar{n}_S^2/\sigma^2$. Letting $\bar{n}_0 = \eta(P_0/h\nu)\,\tau$, $\bar{n}_S = \eta(P_S/h\nu)\,\tau$, we find from equation (D49) that

$$\frac{S}{N} = \frac{\bar{n}_S^2}{\bar{n}_0(1 + \bar{n}_0)} = \frac{P_S^2}{(h\nu/\eta)P_0\tau + P_0^2} \qquad \text{(D50)}$$

which is exactly the result we would expect from (D9) with

$$G(f) = [\sin \pi\tau f)/(\pi\tau f)]^2$$

Coherent Radiation

If only coherent radiation falls on the detector the probability per unit time of the emission of a photon is constant and the count will have the Poisson distribution

$$p(n) = \frac{\bar{n}^n e^{-\bar{n}}}{n!} \qquad \text{(D51)}$$

where $\bar{n}$ = expectation count of signal photons = $\eta(P_r\tau/h\nu)$. The first moment of this distribution is $M_1 = \bar{n}$; the second moment is $M_2 = \bar{n}\ (1 + \bar{n})$, so the mean square fluctuation in count is

$$\sigma_r^2 = M_2 - M_1^2 = \bar{n} \qquad \text{(D52)}$$

Thus the signal-to-noise power ratio is

$$\frac{S}{N} = \bar{n} = \frac{\eta P_r}{h\nu}\,\tau \qquad \text{(D53)}$$

Coherent and Incoherent Radiation:
Short Integration Time

With both coherent and incoherent radiation incident on the detector we might expect the total noise power to be given by $\sigma_0^2 + \sigma_r^2$. Adding the denominator of equation (D50) to the denominator of (D53) and comparing the result with (D10) we see that the cross-power term $2\,P_rP_0$ is not represented in the result. We therefore infer that the mean square fluctuation in count for short averaging times is given by

$$\sigma^2 = \bar{n} + \bar{n}_0 + 2\bar{n}_0\bar{n} + \bar{n}_0^2 \qquad \text{(D54)}$$

and that the signal-to-noise power is

$$\frac{S}{N} = \frac{\bar{n}^2}{(\bar{n} + \bar{n}_0) + \bar{n}_0\,(2\bar{n} + \bar{n}_0)} \qquad \text{(D55)}$$

The exact statistical distribution when both coherent and incoherent power are present and the averaging time is short is very complicated. It is reasonable to expect that this distribution would very closely be given by the convolution of equation (D48) (representing the second and fourth terms in equation (D54) the result then convolved with a discrete distribution resembling a normal law with $\sigma = \sqrt{2\bar{n}_0\bar{n}}$ (representing the third term in (D54). Because none of the cases treated in this report require a solution of this general case, we have not pursued the matter further.

Since the first term in parenthesis in the denominator of (D55) represents "pure" shot noise—the noise that would exist if both the incoherent and coherent counts were independent and Poisson distributed—we can identify the remaining term with fluctuation noise and conclude that shot noise will dominate if $\bar{n}_0 << 1$ in the time $\tau \approx 1/B$, while fluctuation noise will dominate if $\bar{n}_0 >> 1$ in the same integration time.

Coherent and Incoherent Radiation:
Long Averaging Time

When the integration time is long, and in particular when $(B\tau) >> (\bar{n} + \bar{n}_0)$, the number of "coherent" and "incoherent" photons received per Nyquist interval $1/B$ will be small and we can compute the statistics quite accurately assuming that a Poisson distribution (D51) with $\bar{n}$ replaced by $\bar{n}_0$ applies when only incoherent radiation is present and that (D51) with $\bar{n}$ replaced by $\bar{n} + \bar{n}_0$ applies when both radiations are present.

Further, if $\bar{n}_0$ and $\bar{n}$ are both large we may use gaussian approximations for both. Since the distribution with incoherent radiation alone will have the mean value $\bar{n}_0$ and $\sigma_0 = \sqrt{\bar{n}_0}$, while that with the both radiations present will have the mean value $\bar{n} + \bar{n}_0$ and $\sigma = \sqrt{\bar{n}_0 + \bar{n}}$, we require that $\bar{n} = m_0\sigma_0 + m\sigma$ and obtain for the minimum signal count

$$\bar{n} = \frac{(m^2 + 2m_0\sqrt{\bar{n}_0}) + m\sqrt{m^2 + 4m_0\sqrt{\bar{n}_0} + 4\bar{n}_0}}{2} \qquad \text{(D56)}$$

when $m_0 = m$

$$\bar{n} = m^2 + 2m\sqrt{\bar{n}_0} \qquad \text{(D57)}$$

If the incoherent count is the sum of a background count and star noise we will have $\bar{n}_0 = \bar{n}_b + \bar{n}_* = n_b + b_*\bar{n}$, where b_* is the fixed ratio of $\bar{n}/\bar{n}_*$ (independent of range). For $m_0 = m$, the range limit will then occur when

$$\bar{n} = m^2\left[(1 + 2b_*) + 2\sqrt{b_*(1 + b_*) + \bar{n}_b/m^2}\right] \qquad \text{(D58)}$$

SIGNAL AVERAGING WITH SPATIAL COHERENCE

In the derivation of equations (D46) and (D47) it was assumed that noise was uncorrelated in the samples that were averaged. When the samples to be averaged are taken from the closely spaced elements of an antenna array, the star noise is apt to be correlated. If we let $P_0 = P_n + P_* = P_n + b_*P_r$ where P_n is the receiver noise, P_* is the star noise, and $b_* = P_r/P_n$, then the σ of the

distributions after averaging n_r receivers and n_s time samples are

$$\sigma_0 = \frac{\sqrt{P_n^2 + 2b_* P_n P_r + n_r b_*^2 P_r^2}}{\sqrt{n}} \qquad \text{(D59)}$$

for the noise alone, and

$$\sigma = \frac{\sqrt{2P_n P_r + 2n_r b_* P_r^2 + P_n^2 + 2b_* P_n P_r + n_r b_*^2 P_r^2}}{\sqrt{n}} \qquad \text{(D60)}$$

for the signal plus noise, where $n = n_r n_s$. The means of the two distributions are at P_0 and $P_0 + P_r$, respectively, and so we require $P_r = m(\sigma_0 + \sigma)$. After considerable algebra, we find

$$\frac{P_r}{P_n} = 2\,\frac{\frac{m^2}{n}\left[\left(1-2\frac{m}{n_s}b_*\right)+2b_*\right] + \frac{m}{\sqrt{n}}\sqrt{\left(1-2\frac{m^2}{n_s}b_*\right)^2 - 4\frac{m^2}{n}b_*\left(1+2\frac{m^2}{n_s}b_*\right) - 4\frac{m^2}{n}b_*^2(1+n_r)}}{\left(1-2\frac{m^2}{n_s}b_*\right)^2 - 4\frac{m^2}{n_s}b_*} \qquad \text{(D61)}$$

If $m^2/n_s b_* \ll 1$, this rather formidable expression reduces to

$$\frac{P_r}{P_n} \approx 2\left[\frac{m^2}{n}(1 + 2b_*) + \frac{m}{\sqrt{n}}\right] \qquad \text{(D62)}$$

REFERENCE

1. Campbell, George A.; and Foster, Ronald M.: Fourier Integrals for Practical Application. D. Van Nostrand Co., Inc., Princeton, N.J.

APPENDIX E

RESPONSE OF A GAUSSIAN FILTER TO A SWEPT SINUSOID

We may represent a "sinusoid" whose frequency changes linearly with time by the real part of

$$f(t) = e^{i\dot{\omega}t^2/2} \tag{E1}$$

The Fourier transform of this function, which may be considered to be the limit of the transform of $e^{-(\alpha + i\dot{\omega})t^2/2}$ as $\alpha \to 0$, is

$$F(\omega) = \sqrt{\frac{2\pi i}{\dot{\omega}}}\, e^{-i\omega^2/2\dot{\omega}} \tag{E2}$$

If $f(t)$ is passed through a filter whose transmission is,

$$K(\omega) = e^{-\omega^2/2\sigma^2} \tag{E3}$$

the spectrum of the output will be $G(\omega) = F(\omega)K(\omega)$, which corresponds to the time function

$$g(t) = \frac{1}{m}\, e^{-\frac{\dot{\omega}^2 t^2}{2} + i\left(\frac{\dot{\omega}t^2}{2m^2} + \phi\right)} \tag{E4}$$

where

$$m = \sqrt{1 + \dot{\omega}^2/\sigma^4} \tag{E5}$$

$$\phi = \frac{1}{2}\tan^{-1}(\dot{\omega}/\sigma^2) \tag{E6}$$

By superposition, and use of the frequency translation property of multiplication by $e^{-i\omega_0 t}$, we see that the response of a filter whose transmission is

$$K(\omega) = e^{-\frac{(\omega - \omega_0)^2}{2\sigma^2}} \tag{E7}$$

to an input consisting of the swept sinusoid

$$f(t) = a\cos\left(\theta + \omega_0 t + \frac{\dot{\omega}t^2}{2}\right) \tag{E8}$$

will be the time function

$$g(t) = \frac{a}{m}\, e^{-\frac{\dot{\omega}^2 t^2}{2m^2\sigma^2}} \cos\left(\theta + \phi + \omega_0 t + \frac{\dot{\omega}t^2}{2}\right) \tag{E9}$$

If we square-law detect this wave, we will get the pulse

$$k(t) = \frac{a^2}{2m^2}\, e^{-\frac{\dot{\omega}t^2}{m^2\sigma^2}} \tag{E10}$$

The input wave has an average power $P = a^2/2$. The peak output power is

$$S = \frac{P}{m^2} \tag{E11}$$

The filter $K(\omega)$ passes a noise power

$$N = \frac{kt}{2}\int_0^\infty [K(\omega) + K(-\omega)]^2\, df = \frac{kT}{2\sqrt{\pi\sigma}} \tag{E12}$$

Thus the ratio of the peak output pulse power to the noise power is

$$\frac{S}{N} = 2\sqrt{\pi}\,\frac{P}{kT\sigma m^2} \tag{E13}$$

Now $\sigma m^2 = \sigma + \omega^2/\sigma^3$ and has a minimum when

$$\sigma = (3)^{1/4}\dot{\omega}^{1/2} \tag{E14}$$

At this value, $m^2 = 4/3$ so we find from (E13)

$$\left.\frac{S}{N}\right|_{\text{opt}} = \frac{P}{kT\left(\dfrac{2\sqrt[4]{3}}{3\sqrt{\pi}}\,\dot{\omega}^{1/2}\right)} = \frac{P}{kT\left(\dfrac{2^{3/2}}{3^{3/4}}\,\dot{\nu}^{1/2}\right)} \tag{E15}$$

The factor

$$\beta = \frac{2^{3/2}}{3^{3/4}} = 1.24\ldots \tag{E16}$$

degrades the output signal-to-noise ratio to the value that would exist with the same filter, and a nonsweeping signal of 0.937 dB less power.

APPENDIX F

BASE STRUCTURES

AZ-EL BASE STRUCTURES

The conventional az-el mount affords maximum sky coverage, but this coverage is expensive due to increased structural weight and expensive gearing and/or track and wheel assemblies. Elevation rotation is generally provided by gear or chain drives and generally requires counterweighting of the dish to locate the dish center of gravity on the elevation axis. The gearing on the elevation drive must provide the necessary resistance to wind torques. If the basic dish weight (including backup structure) is W, then a weight penalty of roughly $0.3W$ may be incurred in positioning of the center of gravity.

Dish attachments to the base structure are either near the center (king-post design) or near the dish extremity (Manchester mount). Azimuth positioning and dish support are provided either by a hydrostatic bearing coupled with a gear drive or wheel and track drive, or solely by wheel and track support. Typical support configurations are shown in Figure F-1.

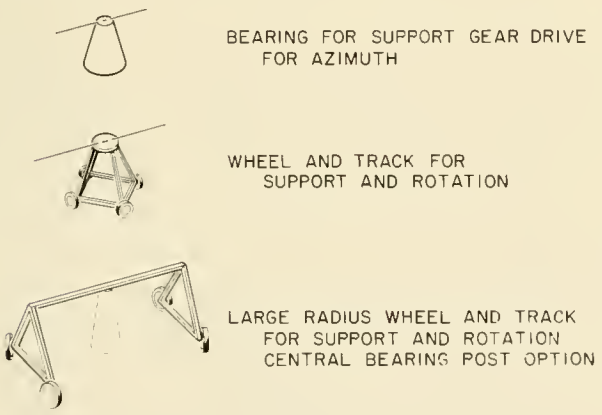


Figure F-1. *Typical configurations.*

MODIFIED AZ-EL BASE STRUCTURE

Available cost figures for contemporary designs indicate that the base structure, which includes all drive elements, is a significant fraction of the cost of the telescope structure. Here we consider a support structure that lends itself to mass production fabrication. Three-point support of the dish is provided, and the need for counterweighting of the dish is eliminated. The assembly uses a single piston for elevation control and a wheel and track arrangement for azimuth positioning. The use of the piston for elevation control eliminates the need for rigid support on the backup structure to support the bull gear segment required on conventional designs. This rigid support structure and the bull gear segment are comparatively expensive items. A novel feature of the base structure is the use of a central bearing to absorb all side loads; the wheel and track arrangement need only provide vertical support and therefore does not require close control of the track radius.

Preliminary Sizing

Figure F-2 shows two views (to scale) of the proposed mount designed for a 100 *foot* dish with a maximum zenith angle of 65°. A preliminary calculation has been made to obtain a weight estimate assuming the dimension $L_1 \doteq 30$ ft. With L_1 and θ given, the maximum piston extension p is computed to be 42 ft, which implies that a piston casing of about 45 ft is required. Since the piston casing sits vertically when the piston is fully collapsed (dish looking at zenith), the total height of the base structure h must be at least 45 ft. The height (and eventually the weight) depends strongly on the desired angle θ. Once h is found, the lengths of the truss members can be easily determined by simple geometry.

There are 24 members in the base structure (not including the piston).

For the geometry chosen, the lengths of the members are as follows:

No. of Members	Length (ft)
3	61.5
3	52.0
3	42.0
3	45.0
6	54.5
6	58.7

The structure configuration is well suited for mass production. The basic element consists of three identical truss elements, which can be factory assembled and erected with a minimum of field welding.

Weight and Cost

A preliminary weight estimate of the structure may be obtained without a detailed structural design effort by requiring that all members have L/r ratio less than 200 (r = radius of gyration of the member). Rough calculations indicate that 8 in SC 40 pipe will meet this requirement and should be strong enough. Based on the use of this pipe (28.56 lb/ft), a total weight of base structure W is easily found. We obtain W = 36,600 lb. Allowing an additional 10% for fittings we obtain a base weight on the order of 41,000 lb. Note that this weight does not include bearings, wheel and track, or pistons. At a material cost of $0.20/lb, we find that the basic steelwork will cost $8,200 per base structure. It seems likely (although no figures can be provided at this time) that the assembled structure (again excluding piston, bearing, and wheel and track) would cost $16,000 to $24,000 if mass produced. No calculations were made for a 100 *meter* dish. If the design is practical for this larger size the material cost alone would be on the order of $300,000.

Design Considerations

If we require that the center of the dish at zenith be located directly over the central bearing of the base structure, we find that the maximum dish radius that can be accommodated is R_{max} = 68 ft. The limiting factor is backup structure interference with ground at maximum angle from zenith.

The structure is statically determinate, carries its loading by tension or compression in all members, and is completely triangularized; no stability problems should occur. Horizontal loads are carried by the central bearing, while vertical loading is carried by the three wheel supports. Drive for the azimuth positioning is provided by powering the wheels. Because side loads are carried by the central bearing, the wheel and track tolerances in the horizontal plane are not critical. In fact, it may not be necessary to provide more than a reasonably level roadway since final pointing adjustments in elevation can be made by the elevation piston if compensating feedback is provided.

A better estimate of the cost cannot be obtained until strength calculations are complete for the design loadings. Such strength calculations will yield requirements on piston, bearing, and wheel and track assemblies; only after this is completed can a cost estimate of these members be obtained.

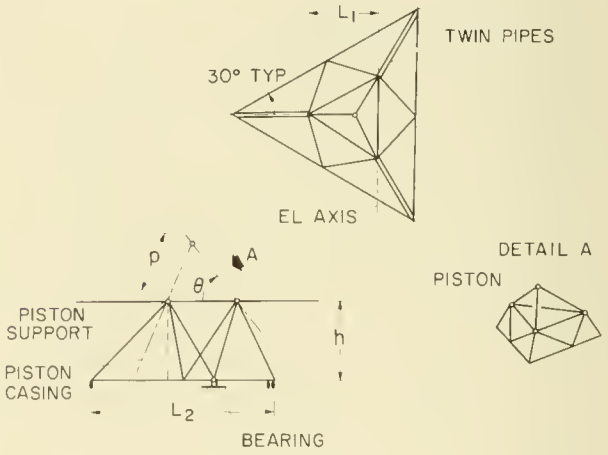


Figure F-2. *Mass produced Az-El mount.*

ACTIVE ELEMENT BASE STRUCTURE

The active element base structure design is an attempt to incorporate the mechanical positioning devices and the structural support into a single unit. It comprises three main pistons and two stabilizing pistons (see Fig. F-3). The dish is supported at three points. The piston connected to point a_1 moves only vertically, whereas the pistons connected to a_2 and a_3 rotate in ball joints at both ends. The pistons P_4 and P_5 provide only tension and are needed only for stability. Tracking and declination angles are achieved by adjusting the relative lengths of all three pistons (P_1, P_2, P_3), with the constraint that the extension of piston P_1 should be kept to a minimum since this piston carries all horizontal loads. A simple rack, connected to the upper ball of the piston and parallel to the piston, drives an encoder, which provides the position feedback required for pointing. Since deformations in the structure below do not influence the

rack, it should not be necessary to provide an external independent position sensor except possibly for initial calibration purposes.

Preliminary Sizing

Based on preliminary calculations the angle θ, was selected as 45°. For a 100-ft dish with a support circle radius at half the dish radius it was necessary to have the support structure approximately 30 ft above the ground to obtain enough travel in pistons P_2 and P_3 to achieve a south zenith angle of 70°. In order to obtain minimum extension of piston P_1, the north zenith angle was limited to approximately 45°. If it is assumed that a 20° elevation above the horizon is the limiting condition due to atmospheric noise then maximum area sky coverage available is:

$$C_{\max} = 2\pi[1 + \sin(70 - 32)] = 10.1515 \text{ steradians}$$

for a location at 32° north latitude. The area sky coverage for the active element base structure with 12.5 ft extension of P_1 and an extension of $P_2 = P_3 = 41.3$ ft is:

$$C_{AE} = 2\pi[\sin(45 + 32) + \sin(70 - 32)]$$

$$= 9.99 \text{ steradians}$$

The percent coverage is:

$$\% \text{ coverage} = \frac{C_{AE}}{C_{\max}} \times 100 = 98.4\%$$

which is a comparatively minor loss when compared with the potential cost savings, and is due to the constraint of minimizing the extension of P_1 since this piston reacts against all of the shear loads. Comparison of various configurations is listed below in Tables F-1 and F-2.

The overall height above ground can be significantly reduced by excavating for pistons P_2 and P_3. This would have to be justified on the basis of excavation cost, which would be a function of the angular motion of the piston casings versus truss costs for piston P_1 and the associated reduction in survival wind load. The limiting factor in this case would be interference of the dish with ground.

TABLE F-1

DISH SIZE VS. ZENITH & "TRACKING ANGLES" FOR DISH HT. = 29.5′
Extension of Pistons = P_1 = 12.5′; $P_2 = P_3$ = 41.3′

South Zenith Angle °	North Zenith Angle °	"East-West" Angle °	Dish Diam. Ft
		Support Rad. = 0.5 Radius of Dish	
70	45	±70	100
45	40	±60	155
		Support Rad. = 0.25 Radius of Dish	
70	45	±70	200
45	40	±60	310

TABLE F-2

PISTON LENGTH & DISH HT. VS. ZENITH & TRACKING ANGLES FOR 100 FT. DISH DIAM.
Support Rad. = 0.5 Radius of Dish

South Zenith Angle °	North Zenith Angle °	"East West" Angle °	Dish Height Ft	Extension of P_1 Ft	Extension of $P_2 = P_3$ Ft
70	45	±70	29.5	12.5	41.3
45	40	±60	19.0	8.1	26.8

Main Piston Design

The pistons are of a differential type (Figure F-4) with the high pressure side of the three pistons interconnected. In essence, this supports the dead weight load of the dish. Control is achieved by varying the pressure in the upper chamber, which is at a lower pressure. The pressure in the upper chamber produces a tensile load on the piston casing that aids its stability. Thrust loads may be taken by linear roller bearings mounted on top of the piston casing.

Assuming a 100,000-lb load on pistons P_2 and P_3, a light weight 12-in. diameter pipe appears adequate from the point of view of column stability.

Stabilizing Piston Design

The stabilizing pistons are attached to the top of the

casings of piston P_2 and P_3, respectively. Their total elongation is comparatively small and is defined by the angular motion of the casing of pistons P_2 and P_3. Since hydraulic power is already available, minimal additional cost should be required for these pistons. The design ensures that only tensile loads exist in these pistons, and therefore long rods or cable can be used to attach the pistons to the casings.

Weight and Cost Considerations

An initial estimate of the total weight, including end attachments, is 15,000 to 25,000 lb. The total base structure including all mechanical drives comprises five elements, all of which could be manufactured in a shop and shipped as complete assemblies ready for attachment to the dish and foundation. The repetitive operations in the shop should reduce the cost of each element. The machining of moderate length pistons and casings is well within the state of the art.

A significant reduction in the weight of the backup structure is achieved by the elimination of the elevation bull gear segment and its rigid support structure. In addition, the need for a counterweight as part of the backup structure is eliminated.

This base structure may have the potential for significant cost savings, but further study is needed to assess its lateral stability, and to practicality. The flexural stiffness of the pistons when extended may pose problems, in view of the need to know the phase center of the element within 1 or 2 mm. As in the case of the previous base structure, no analysis was made for a 100 *meter* dish.

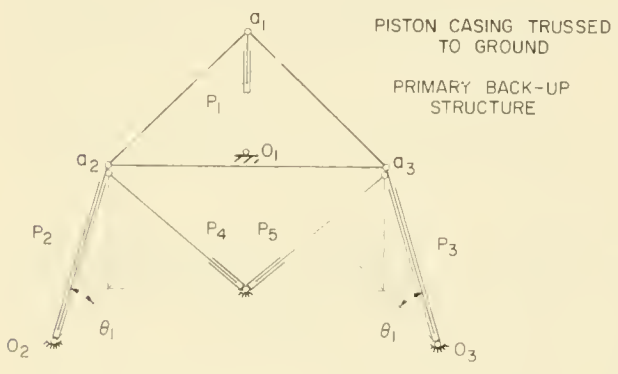


Figure F-3. *Machine structure schematic.*

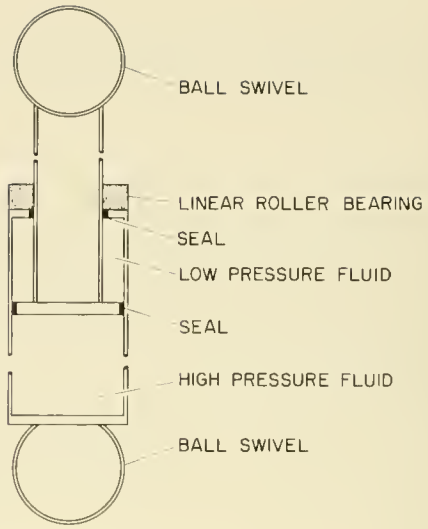


Figure F-4. *Piston cross section.*

EQUATORIAL MOUNT

Adequate consideration has not been given to this design for limited tracking angles. If the tracking time is limited to 2 to 3 hours, which is equivalent to ±15° to ±22-1/2°, many of the major drawbacks associated with this design are reduced. The counterweight becomes comparatively small and the antenna deformations associated with the "flopover" problem are decreased. Declination angles of ±45° should be achievable and, with this limitation, the pedestal structure can be kept reasonably close to the ground. Study of conventional designs suggests that a very lightweight base structure can be designed under these restricted conditions.

The ease of tracking, requiring only single axis rotation, should be considered even though for this design it is not of primary importance. The error introduced in the aiming vector by gravity deformations while tracking, may be eliminated by a feedback loop in the elevation drive system. This may eliminate the necessity for a sensor independent of the structure.

TETHERED BUOYANT TELESCOPE

This concept provides for a minimum weight base structure, having no bearings or tracks, and without a structural members in compression. The concept is to build a buoyant dish, immerse it in a fluid, and tie it down with three cables. The design is attractive if sky

coverage is limited to at most a 30° zenith angle. The proposed concept is illustrated in Figure F-5.

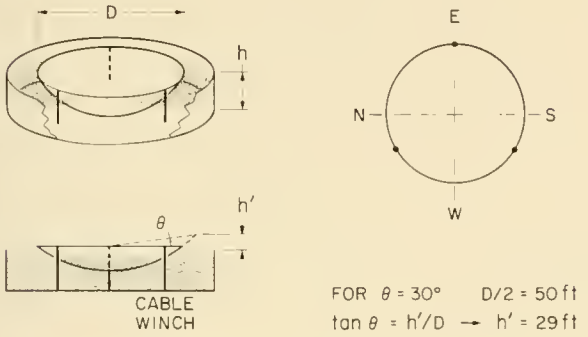


Figure F-5. *Tethered buoyant telescope.*

The three control cables are always in tension. The direction is controlled by selective winching of cables. The buoyancy provides some stiffness against rotation of the antenna in the horizontal plane but lateral guying may be needed. The structure is self-leveling, gravity deformation is minimum and wind loading is relatively small.

The total structural height when the element is at zenith is approximately h + 29 ft from ground level. For a 100-ft diameter dish, it is estimated that $h \approx 40$ ft. It should be noted that most of the element is protected from direct exposure to wind; therefore, operation should be possible in high winds assuming a way is found to protect the fluid from wind effects. The fluid basin could be above or below ground level; if it is below ground, holes could be made by precision blasting.

No computations have been made on this proposal except to consider some of the factors in the pointing accuracy. Consider a cable under load $\Delta\Upsilon$ as shown in Figure F-6, where $\Delta\Upsilon$ = 5000 lb and cable area = 2 sq in. with $E = 15 \times 10^6$ psi, then

$$\delta = \frac{\Delta L}{AE} = \frac{5000 \times (40 \times 12)}{2 \times (15 \times 10^6)} = 0.008 \text{ in.}$$

CABLE UNDER LOAD

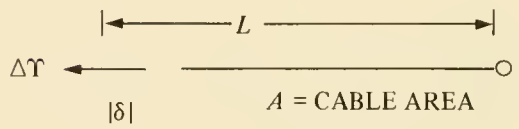


Figure F-6.

It can be seen that fluctuation in the vertical plane can be kept very small. Oscillations in the horizontal plane can be minimized by keeping large tensions in the cable, and by additional guys.

The structure is inherently failsafe in that if a cable breaks, the dish goes up, not down. It seems likely that the dish costs would be less than those for present dishes, since more uniform support is provided by the fluid.

APPENDIX G

BACK-UP STRUCTURES

There are four primary methods for limiting the deflections of an antenna structure that appear to be usable for the Cyclops design.

MAXIMUM STIFFNESS APPROACH

This is the time-honored procedure that has been used by structural engineers for designing buildings, bridges, etc. It has the distinct disadvantage of requiring more material compared with other methods, to achieve the desired result. This method is useful to the Cyclops design only if structural members can be incorporated for multiple use—for example, incorporating the reflector surface and backup structure into some sort of semimonocoque type of arrangement.

ENVIRONMENTAL SHIELDING

A radome may be placed over the antenna to shield it from wind and thermal effects. According to reference 1, this approach gives a total structural cost that may be appreciably below the typical cost curves for all current existing antenna designs. (Although our study here does not necessarily confirm this finding.)

BEST-FIT PROCEDURE

It is possible to reduce the rms deviation by fitting a paraboloid of revolution to the distorted surface for the various angles of tilt of the dish. The so-called "homologous design" as discussed in reference 2 is a further exploration of this method. Figure 8-2 from reference 2 shows the so-called "natural" limits that exist for a steerable antenna. Assuming that the rms deviation for the Cyclops antenna elements is between 1 mm and 3 mm, it can be observed from Figure 8-2 that it is possible to construct dishes with a diameter of 40 to 70 m without violating the gravitational limit. For dishes in excess of these dimensions, it is necessary to employ a refinement in structural analysis or design to meet the requirements of rms deviation. Regardless of the size of the dish to be selected, a best fit procedure should be used to minimize the rms error and total structural weight.

THE USE OF MECHANICALLY ACTIVATED ELEMENTS

If the antenna element is fitted with force or deformation compensating devices so that excessive deflections can be removed, then the structural elements can be made much lighter. These compensation devices can consist of hydraulic jacks, that are properly arranged counterweights. According to reference 1 it is possible to accomplish proper compensation with as few as three opposing force systems. This approach seems to lack appeal for the Cyclops array because, with the large number of elements, the problem of maintenance appears to outweigh any savings represented by initial cost savings.

REFERENCES

1. *A Large Radio–Radar Telescope Proposal for a Research Facility.* (4 vol.) Northeast Radio Observatory Corp., June 1970.
2. *A 300 Foot High High-Precision Radio Telescope.* National Radio Astronomy Observatory, Green Bank, West Virginia, May 1969.

APPENDIX H

POLARIZATION CONSIDERATIONS

The sensitivity of the Cyclops array depends on the alignment of the polarization of the transmitter with the polarization of the receiver. The polarization of the transmitter is unknown, however, and to avoid loss of sensitivity, steps must be taken to ensure that the receiver receives most of the incident polarization. The discussion is facilitated if the Poincaré sphere (ref. 1) shown in Figure H-1 is used. Any arbitrary polarization is described by the radius vector from the origin to a point on the sphere. If the vector representing the incident polarization makes an angle θ with the vector representing the receiver polarization, the received field strength is:

$$E = E_0 \cos \frac{\theta}{2}$$

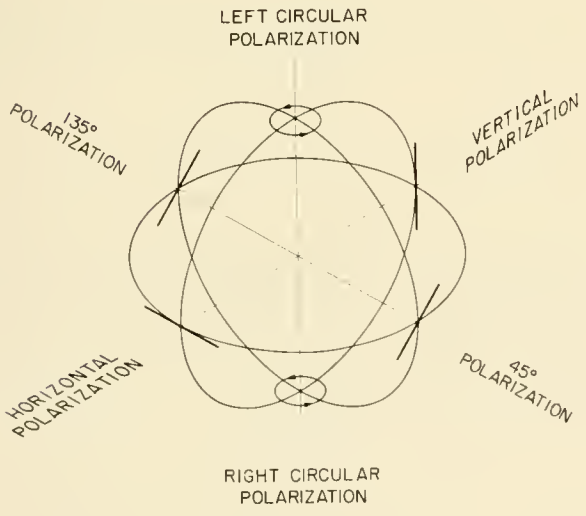


Figure H-1. *The Poincaré sphere.*

If horizontally and vertically polarized receivers are provided, then the angle θ will be no greater than $\pi/2$ from one or the other of the receivers. The maximum loss would then be 3 dB. An expected value of loss may also be calculated if the incident polarization is assumed to be randomly distributed. This number is 1.3 dB.

Suppose now that the horizontal and vertical channels, designated by *H* and *V*, are received, and that after amplification the processing scheme shown in Figure H-2a is used. This processing results in an equal spacing of the receivers around the equator of the Poincaré sphere. Again the maximum loss is 3 dB but the expected value of loss is now 0.7 dB.

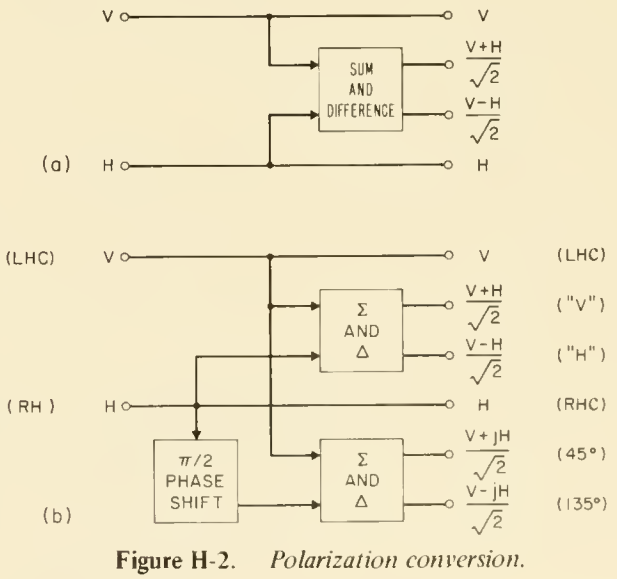


Figure H-2. *Polarization conversion.*

To carry this operation to its logical conclusion, receivers would be placed at the cardinal points of the Poincaré sphere. The processing scheme shown in Figure

H-2b accomplishes this. The maximum loss in this case is 1 dB, and the average, or expected value is 0.39 dB. With vertical and horizontal polarizations supplied as inputs, the outputs are as shown in the first right-hand column. If, on the other hand, the inputs are right and left circular polarization, as shown in parentheses, the outputs will be the polarizations shown in parentheses in the last column on the right. Regardless of which pair of orthogonal inputs is supplied, the outputs will always correspond to six polarizations equally spaced around the Poincaré sphere.

Leakage signals will probably have random polarizations and will be detected with an average loss of 0.39 dB if all six outputs are processed. Beacons, for the reasons given in Chapter 6, are expected to be circularly polarized. The Cyclops proposal is to search both circular polarizations simultaneously while looking for beacons. By using the same data processing system on the outputs of the polarization converter of Figure H-2b taken in pairs, all polarizations provided can be searched in three times the nominal search time.

REFERENCE

1. Kraus, J.D., Radio Astronomy. McGraw-Hill, New York, 1966.

APPENDIX I

CASSEGRAINIAN GEOMETRY

Consider the paraboloidal primary mirror with its vertex at V and its focus at F as shown in Figure I-1. The equation of its surface is

$$r = \frac{2f}{1 + \cos\theta_1} = \frac{f}{\cos^2(\theta_1/2)} \tag{11}$$

where f is the focal length FV. The diameter of the paraboloid included by the cone of rays of half-angle θ_1 is $d = 2r \sin\theta_1$, so

$$\frac{d}{4f} = \frac{\sin\theta_1}{1 + \cos\theta_1} = \tan\frac{\theta_1}{2} \tag{12}$$

If an isotropic radiator radiating I W/sr is placed at F, the flux reflected off the parabola will be:

$$\phi = \frac{I}{r^2} = \frac{I}{f^2}\left(\frac{1 + \cos\theta_1}{2}\right)^2 = \frac{I}{f^2}\cos^2\frac{\theta_1}{2} \tag{13}$$

Now assume that a hyperbolic secondary mirror is introduced with its vertex at V_1. Received energy will now be brought to focus at F'. If we let the distance $OV = OV' = a$, then $OF = OF' = \epsilon a$ where ϵ is the eccentricity of the hyperboloid. Thus the focal distance $FV_1 = f_1 = (\epsilon - 1)a$ and the focal distance $F'V_1 = f_2 = (\epsilon + 1)a$. The magnification is

$$m = \frac{f_2}{f_1} = \frac{\epsilon + 1}{\epsilon - 1} \tag{14}$$

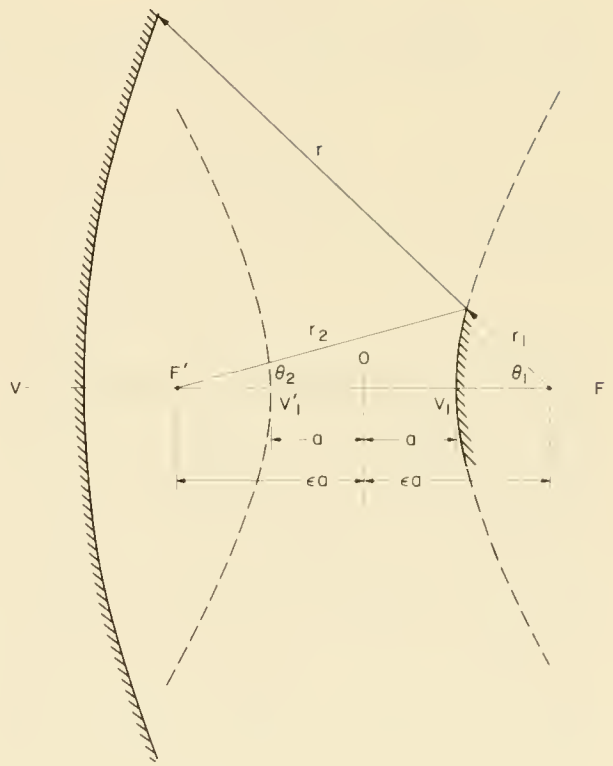


Figure I-1. *A Cassegrainian telescope.*

or, conversely,

$$\epsilon = \frac{m + 1}{m - 1} \tag{15}$$

The polar equation of the secondary mirror is

$$r_1 = \frac{f_1(\epsilon + 1)}{1 + \epsilon \cos\theta_1} = \frac{a(\epsilon^2 - 1)}{1 + \epsilon \cos\theta_1} \tag{16}$$

We see from Figure 1-1 that

$$\tan\theta_2 = \frac{r_1 \sin\theta_1}{2\epsilon a - r_1 \cos\theta_1} = \frac{(\epsilon^2 - 1)\sin\theta_1}{2\epsilon + (\epsilon^2 + 1)\cos\theta_1} \tag{17}$$

and therefore that

$$\sin\theta_2 = \frac{(\epsilon^2 - 1)\sin\theta_1}{(\epsilon^2 + 1) + 2\epsilon \cos\theta_1} \tag{18}$$

$$\cos\theta_2 = \frac{2\epsilon + (\epsilon^2 + 1)\cos\theta_1}{(\epsilon^2 + 1) + 2\epsilon \cos\theta_1} \tag{19}$$

From equations (8) and (9) we then have

$$\frac{\sin\theta_2}{1 + \cos\theta_2} = \frac{1}{m}\,\frac{\sin\theta_1}{1 + \cos\theta_1}$$

$$\tan\frac{\theta_2}{2} = \frac{1}{m}\tan\frac{\theta_1}{2} = \frac{1}{m}\,\frac{d}{4f} \tag{110}$$

Thus the Cassegrainian system behaves exactly like a simple paraboloid having m times as great an f/d ratio as the primary mirror, at least as far as paraxial rays are concerned. For either case the illumination of the primary mirror will be given by

$$\frac{\phi}{\phi_0} = \left(\frac{1 + \cos\theta}{2}\right)^2 = \cos^2\frac{\theta}{2} \tag{111}$$

where θ is the angle subtended at the feed. Thus for a given uniformity of illumination we can use a primary mirror having an f/d ratio $1/m$ times as large.

APPENDIX J

PHASING OF THE LOCAL OSCILLATOR

The phase of the local oscillator signal at each antenna site must have a value proportional to the distance from the antenna to one of the arriving plane wavefronts. Note that this is a phase correction, modulo 2π—and not a time delay correction—since the local oscillator is a monochromatic signal. Due to the sidereal rotation of the earth, this phase setting will continuously vary with time, depending on the antenna coordinates, the source coordinates and the reference phase. Thus, it is necessary to incorporate a variable (modulo 2π) phase shifter into each local oscillator. This phase shifter may also be used to compensate for phase errors in the RF system provided these errors are known.

Consider a phase shifter with an initial setting of ϕ_O, which is chosen to put the local oscillator in the correct phase for the arriving wavefront. As the antenna tracks the source, this plane wavefront tilts and the phase setting must be varied also. When the phase setting reaches 2π, the phase shifter is reset to zero phase as shown in Figure J-1. This phase change is equivalent to a frequency offset in the local oscillator signal of

$$\Delta f = \frac{1}{T}$$

To obtain negative offset frequencies it is required that the phase shifter to be driven from 2π to 0° thus obtaining a negative slope.

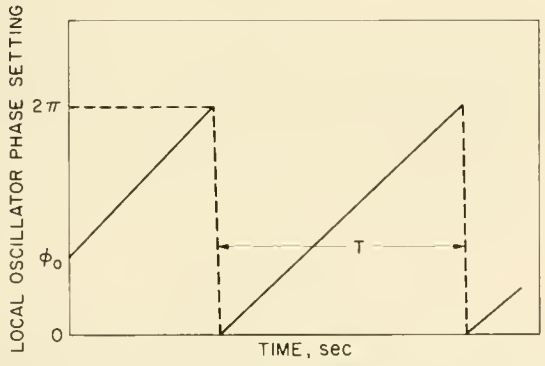


Figure J-1. *Required phase shift versus time.*

Microwave phase shifters may be either digital or analog type devices. Digital phase shift devices essentially operate by switching different lengths of transmission line in and out of the signal transmission path. Analog phase shift devices operate by varying the effective permittivity or permeability of the transmission path. Consider the phase ramp approximation of Figure J-2. A system like this would be very easy to control because of its digital nature. Klein and Dubrowsky (ref. 1) have built a frequency translator of this type. In general, their results were good. However, because of the digital nature of the device, sidebands are created at frequencies related to the time per bit. The amplitudes of the peak sidebands decayed with bit size as shown in Figure J-3. From this it is seen that many bits must be used to obtain low sideband amplitudes. However, as the number of bits increase, so does the complexity of the driver of the phase shifter. It is doubtful if a phase shifter of this type would have any significant advantage over an analog phase shifter, and since the analog phase shifter will not have any sidebands, it is preferable for this application.

The phase shift may be controlled in an analog manner either mechanically with a rotating vane phase shifter or electrically with a ferrite phase shifter. Both

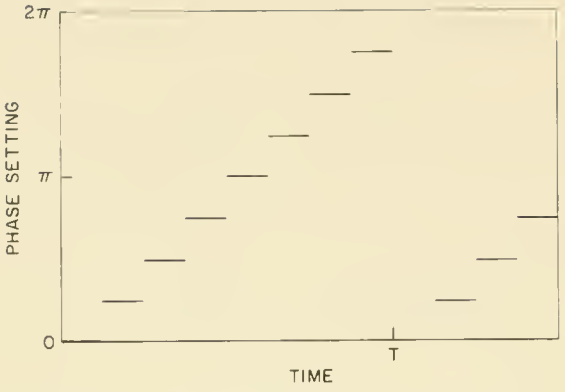


Figure J-2. *Step approximation to ramp.*

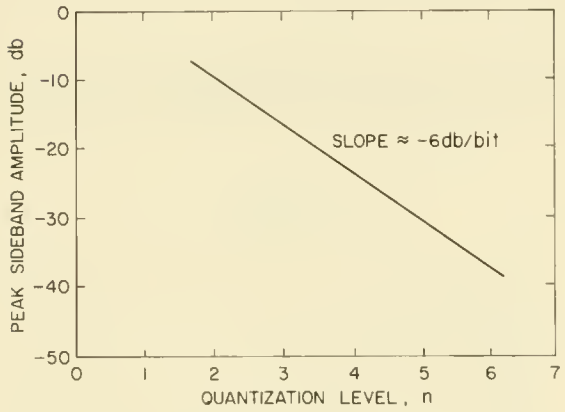


Figure J-3. *Sidebands due to approximation.*

types cost roughly the same so that the ferrite phase shifter is to be preferred on the basis of improved reliability. A typical phase shift versus control current curve of a ferrite phase shifter is shown in Figure J-4. Note that the phase shifter is nonreciprocal—that is, different phase shifts are obtained when the polarity of

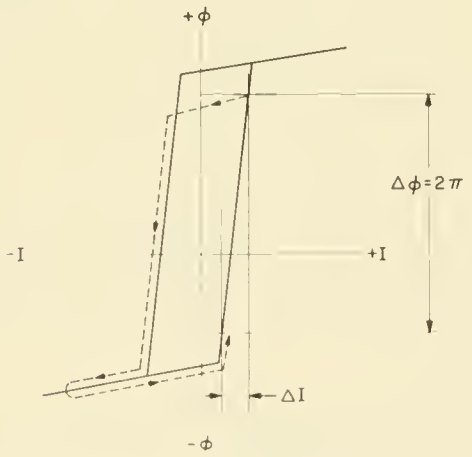


Figure J-4. *Ferrite phase shifter transfer characteristic.*

the coil current is reversed—and that the device exhibits hysteresis. For a positive frequency offset, the phase shifter would be driven into saturation with a large negative current to ensure a stable phase condition, and then to a predetermined preset point. This would take place in approximately a microsecond.

The phase would then be advanced at the appropriate rate by increasing the positive current until a total excursion of 2π had occurred. The cycle would then be repeated by resetting the phase shifter as indicated by the dotted lines in Figure J-4. For negative frequency offsets, the stable reset condition is achieved by saturating the phase shifter with a positive current.

For accurate phase control it is desirable to be able to set the initial phase and then specify the rate of change (the frequency offset). Microwave phase shifters can be set to an accuracy of about 2°. This is adequate for Cyclops. Rate control may be accomplished by frequently updating the phase shift desired, with this phase shift stored in a digital numeric register as suggested in Chapter 10, or by a true rate system (Figure J-5) in which the input and output signals at the phase shifter are sampled by directional couplers and then mixed. The difference frequency may then be compared with the commanded frequency offset, and if the two are not the same, an error signal may be generated that can be used to modify the sweep rate of the phase shifter so, as to drive the error to zero.

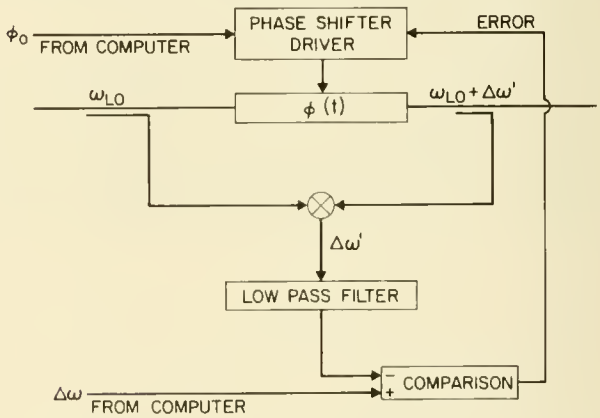


Figure J-5. *Rate controlled phase shifter.*

REFERENCE

1. Klein and Dubrowsky, The Digilator, a New Broadband Microwave Frequency Translator, IEEE Trans. on Microwave Theory and Techniques, *MTT-15*, 3, March 1967, pp. 172-179.

APPENDIX K

EFFECT OF DISPERSION IN COAXIALS

The propagation constant of a coaxial line in which the dielectric loss of the insulating medium may be neglected is $\gamma = \alpha + i\beta$, with

$$\alpha = \sqrt{\frac{\epsilon}{8\sigma}} \frac{(1/a) + (1/b)}{\ell n\,(b/a)} \omega^{1/2} \qquad \text{(K1)}$$

$$\beta = \omega\sqrt{\mu\epsilon} + \alpha \qquad \text{(K2)}$$

where

a = outer radius of inner conductor

b = inner radius of outer conductor

ϵ = dielectric constant of dielectric medium

μ = permeability of dielectric medium and conductors

σ = conductivity of conductors

Equations (1) and (2) cannot be obtained from the equivalent circuit representation of a transmission line, which yields $\gamma = \sqrt{(R + i\omega L)(G + i\omega C)}$, but must be derived directly from Maxwell's equations (ref. 1).

If the line length is ℓ, then the total phase shift θ is

$$\theta = -\beta\ell = -(\omega\sqrt{\mu\epsilon}\ \ell + \alpha\ell) \qquad \text{(K3)}$$

Since the first term in (3) is the phase shift we would expect from the delay $\tau_0 = \sqrt{\mu\epsilon}\ \ell$ of a dispersionless line, we see that the loss contributes an excess phase of $\alpha\ell$. But $\alpha\ell$ is the line loss in nepers, so *the excess phase is one radian per neper of line loss.* Since this phase is proportional to $\omega^{1/2}$ rather than ω, it does not represent a constant delay.

The analysis of the phase compensation schemes of Figure 9-14 and 9-15 shows that with zero offset frequency ($\delta = 0$) and constant delay lines there is zero phase error. To determine the errors caused by the second term of equation (K3) we need merely repeat the analysis considering this term to be the only phase present.

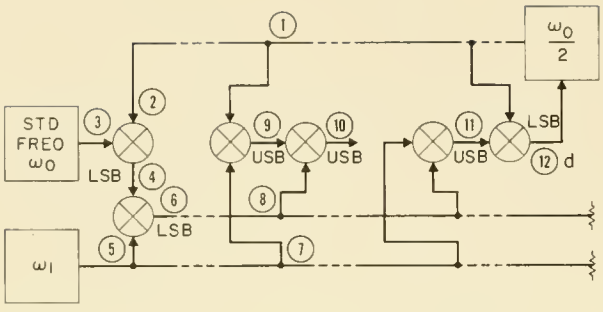
Let N_0 be the line loss in nepers at $\omega = \omega_0$. Then the loss at any frequency is

$$N = N_0 \left(\frac{\omega}{\omega_0}\right)^{1/2} \qquad \text{(K4)}$$

For convenience the circuits of Figures 9-14 and 9-15 are reproduced here as Figures K-1 and K-2, with δ assumed to be zero as a result of phase locking the remote oscillator. The line losses are from the central station to the point in question along the line.

We see that for the system of Figure K-1 the phase error is

$$\Delta\theta_1 = \left[\frac{1}{\sqrt{2}} - \left(\frac{1}{2} - \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0 \qquad \text{(K5)}$$



POINT	FREQUENCY	EXCESS PHASE
①	$\frac{\omega_0}{2}$	θ_0
②	$\frac{\omega_0}{2}$	$\theta_0 - \frac{1}{\sqrt{2}}\ N_0$
③	ω_0	0
④	$\frac{\omega_0}{2}$	$-\theta_0 + \frac{1}{\sqrt{2}}\ N_0$
⑤	ω_1	θ_1
⑥	$\frac{\omega_0}{2} - \omega_1$	$-\theta_0 - \theta_1 + \frac{1}{\sqrt{2}}\ N_0$
⑦	ω_1	$\theta_1 - \left(\frac{\omega_1}{\omega_0}\right)^{1/2}\ N_0$
⑧	$\frac{\omega_0}{2} - \omega_1$	$-\theta_0 - \theta_1 + \left[\frac{1}{\sqrt{2}} - \left(\frac{1}{2} - \frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0$
⑨	$\frac{\omega_0}{2} + \omega_1$	$\theta_0 + \theta_1 - \left(\frac{\omega_1}{\omega_0}\right)^{1/2}\ N$
⑩	ω_0	$\Delta\theta = \left[\frac{1}{\sqrt{2}} - \left(\frac{1}{2} - \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0$
⑪	$\frac{\omega_0}{2}$	$-\theta_0 + \Delta\theta$
⑫ d	0	$2\theta_0 \rightarrow \Delta\theta$

Figure K-1. *Dispersion phase error in Cyclops proposal I.*

while for the system of Figure K-2 we find

$$\Delta\theta_2 = \left[1 - \left(\frac{1}{4} + \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{1}{4} - \frac{\omega_1}{\omega_0}\right)^{1/2}\right] N_0$$

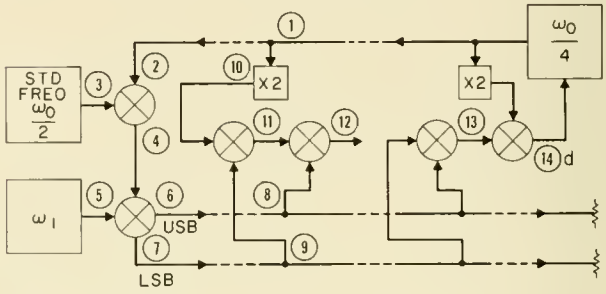
$$\approx \left[2\left(\frac{\omega_1}{\omega_0}\right)^2 + 10\left(\frac{\omega_1}{\omega_0}\right)^4 - \ldots\right] N_0 \qquad \text{(K6)}$$

Figure 9-16 shows the curves of $\Delta\theta$ vs. (ω_1/ω_0).

The attenuation of 1-5/8 in. coaxial at 1 GHz is about 24 dB/km or 120 dB for a 5-km run. Thus $N_0 \approx 14$ nepers. If $\omega_1 = 10$ MHz, then with $\omega_0 = 1$ GHz.

$$\Delta\theta_2 = 2.8\times10^{-3} \text{ radians} = 0.114°$$

For $\omega_1 = 25$ MHz, $\Delta\theta_2 \approx 0.72°$ which represents a 7.2° degree error at 10 GHz. But since we are only concerned with *variation* in $\Delta\theta$ with time and temperature, we can easily stand this much absolute error.



POINT	FREQUENCY	EXCESS PHASE
①	$\frac{\omega_0}{4}$	θ_0
②	$\frac{\omega_0}{4}$	$\theta_0 - \frac{1}{2}$ NO
③	$\frac{\omega_0}{2}$	0
④	$\frac{\omega_0}{4}$	$-\theta_0 + \frac{1}{2}$ NO
⑤	ω_1	θ_1
⑥	$\frac{\omega_0}{4} + \omega_1$	$-\theta_0 + \theta_1 + \frac{1}{2}$ NO
⑦	$\frac{\omega_0}{4} - \omega_1$	$-\theta_0 - \theta_1 + \frac{1}{2}$ NO
⑧	$\frac{\omega_0}{4} + \omega_1$	$-\theta_0 + \theta_1 + \left[\frac{1}{2} - \left(\frac{1}{4} + \frac{\omega_1}{\omega_0}\right)^{1/2}\right]$ NO
⑨	$\frac{\omega_0}{4} - \omega_1$	$-\theta_0 - \theta_1 + \left[\frac{1}{2} - \left(\frac{1}{4} - \frac{\omega_1}{\omega_0}\right)^{1/2}\right]$ NO
⑩	$\frac{\omega_0}{2}$	$2\theta_0$
⑪	$\frac{3\omega_0}{4} - \omega_1$	$\theta_0 - \theta_1 + \left[\frac{1}{2} - \left(\frac{1}{4} - \frac{\omega_1}{\omega_0}\right)^{1/2}\right]$ NO
⑫	ω_0	$\Delta\theta = \left[1 - \left(\frac{1}{4} + \frac{\omega_1}{\omega_0}\right)^{1/2} - \left(\frac{1}{4} - \frac{\omega_1}{\omega_0}\right)^{1/2}\right]$ NO
⑬	$\frac{\omega_0}{2}$	$-2\theta_0 + \Delta\theta$
⑭ d	0	$4\theta_0 \rightarrow \Delta\theta$

Figure K-2. *Dispersion phase error in Cyclops proposal II.*

REFERENCE

1. Morgan, S.P.: Mathematical Theory of Laminated Transmission Lines, B.S.T.J. *31*, 5, Sept. 1952, p. 895.

APPENDIX L

TUNNEL AND CABLE LENGTHS

Exactly n-1 lines are needed to connect n points, so n-1 tunnels between adjacent antennas suffice to connect an array of n antennas. One more tunnel is then needed to tie the array to the central control headquarters. However, certain main tunnels may be larger than the branch tunnels and, depending on the array configuration, not all tunnels may be of the same length, so some analysis of this problem is in order.

To minimize the tunnel and cable lengths required, the array should be made as compact as possible and the control and processing center should indeed be at the physical center of the array. Also the outline of the array should be circular, although small departures from circularity cause only a second-order change in the total tunnel and cable length. Two obvious configurations come under consideration: a square lattice in which the antennas are placed in equally placed rows and columns, and a hexagonal lattice in which the antennas are at the centers of the cells of a honeycomb. These configurations correspond to tessellating the array area with squares and hexagons, respectively.

The ratio of the area of a circle to the area of a circumscribed square is $\pi/4$ while that of a circle and the circumscribed hexagon is $\pi/2\sqrt{3}$. However, the spacing s between antennas must be greater than their diameter d to prevent shadowing of the antennas by their neighbors at low elevation angles. If ϵ is the minimum elevation angle and $\theta_m = (\pi/2) - \epsilon$ is the maximum angle form the zenith for unobstructed reception then

$$s = \frac{d}{\cos \theta_m} = \frac{d}{\sin \epsilon} \qquad \text{(L1)}$$

The filling factor for a square lattice array is therefore

$$f_s = \frac{\pi}{4}\left(\frac{d}{s}\right)^2 = \frac{\pi}{4}\sin^2\epsilon \qquad \text{(L2)}$$

while that for a hexagonal array is

$$f_h = \frac{\pi}{2\sqrt{3}}\left(\frac{d}{s}\right)^2 = \frac{\pi}{2\sqrt{3}}\sin^2\epsilon \qquad \text{(L3)}$$

Equations (L2) and (L3) apply to circular dishes: If elliptical dishes are used, only the vertical dimension (minor axis) must be reduced to $d = s \sin \epsilon$; the horizontal dimension (major axis) need be only slightly less than s. Thus, for elliptical dishes the filling factors involve $\sin \epsilon$ rather than $\sin^2\epsilon$. Table L-1 shows the realizable filling factors without obstruction.

TABLE L-1

			Circular		Elliptical	
ϵ	θ_m	s/d	f_s	f_h	f_s	f_h
45°	45°	1.1414	.393	.453	.455	.641
30°	60°	2.000	.196	.227	.393	.453
20°	70°	2.924	.092	.106	.269	.310
10°	80°	5.760	.024	.027	.136	.157

The filling factor for a hexagonal lattice array is always $2/\sqrt{3} = 1.155$ times as great as for a square lattice array.

If we wish unobstructed operation down to a 20° elevation angle, the array diameter must be roughly three times as great with circular dishes, or $\sqrt{3}$ times as great with elliptical dishes, as would be required for zenith operation only. Clearly, the use of elliptical dishes would permit significant reduction in the array size and

a corresponding reduction in tunneling and cabling cost. However, the antenna structural cost might be adversely affected and certainly the feed horn design would be complicated. These considerations forced us to consider only circular dishes for Cyclops but the problem deserves further study.

ARRAY RADIUS

If we assume n antennas spaced a distance s apart in a square lattice, the radius a_s of the array (distance from array center to *center* of outer antennas) is given by

$$\pi \left(a_s + \frac{s}{2}\right)^2 \approx ns^2$$

$$a_s \approx s\left(\sqrt{\frac{n}{\pi}} - \frac{1}{2}\right) \qquad \text{(L4)}$$

while for a hexagonal lattice with antenna spacing s the radius a_h is given by

$$\pi \left(a_h + \frac{s}{2}\right)^2 \approx \frac{\sqrt{3}}{2} ns^2$$

$$a_h \approx s\left(\frac{\sqrt{3}\,n}{2\pi} - \frac{1}{2}\right) \qquad \text{(L5)}$$

Using these expressions and taking s = 300 m to permit clear operation to 20° elevation with 100-m diameter antennas, we obtain the values given in Table L-2.

TABLE L-2

n	a_s	a_h
100	1543 m	1425 m
200	2244 m	2080 m
500	3635 m	3370 m
1000	5200 m	4830 m
2000	7420 m	6890 m

In round numbers, a 1000-element array of 100-m dishes spaced 300 m apart will be 10 km in diameter and have an effective clear aperture diameter of 3.16 km.

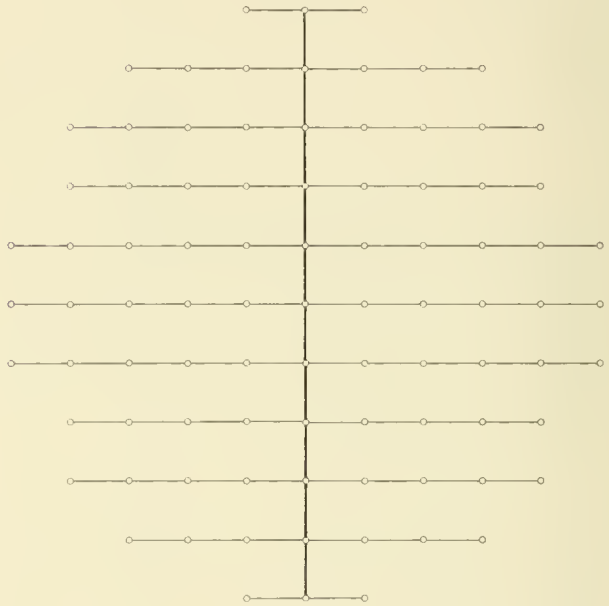
TUNNEL LENGTHS AND COSTS

The branch or side tunnels will, in general, connect adjacent antennas and so consist of $(n-1-N_m)$ sections of length s, where N_m is the number of sections of main tunnel. Let the length of each section of main tunnel be ks. If now the cost per unit length of these side and main tunnels is γ_s and γ_m, respectively, the total cost will be

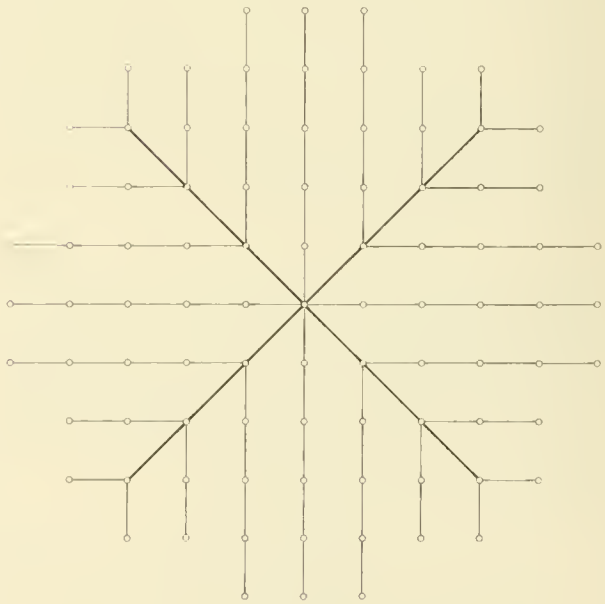
$$C = L_s\gamma_s + L_m\gamma_m$$

$$C = [\,(n - N_m)\; s\,\gamma_s + N_m\, ks\,\gamma_m]$$

$$C = s\,[\,n + (kr-1)\,N_m\,]\,\gamma_s \qquad \text{(L6)}$$



Pattern 1(a)



Pattern 1(b)

Figure L-1. *Tunnel patterns for square lattice array.*

where $r = \gamma_m/\gamma_s$. If we let

$$N_{\text{eff}} = n + (kr - 1)N_m \tag{L7}$$

be the *effective* number of side tunnels of length s, then we have simply

$$C = sN_{\text{eff}}\gamma_s \tag{L8}$$

For the tunnel pattern shown in Figure L-1a, $k = 1$, and N_m is *twice* the nearest integer to $a_s/s = \sqrt{n}/\pi - 1/2$. Thus,

$$N_m = 2\left(\text{int}\sqrt{\frac{n}{\pi}}\right) \tag{L9}$$

where "int" means "the integer part of." Substituting in equation (L7) we find

$$N_{1a} = n + 2(r-1)\,\text{int}\sqrt{\frac{n}{\pi}} \tag{L10}$$

For the tunnel pattern of Figure L-1b, $k = \sqrt{2}$ and N_m is *four* times the nearest integer to $(a_s/s\sqrt{2}) - 1$. Thus,

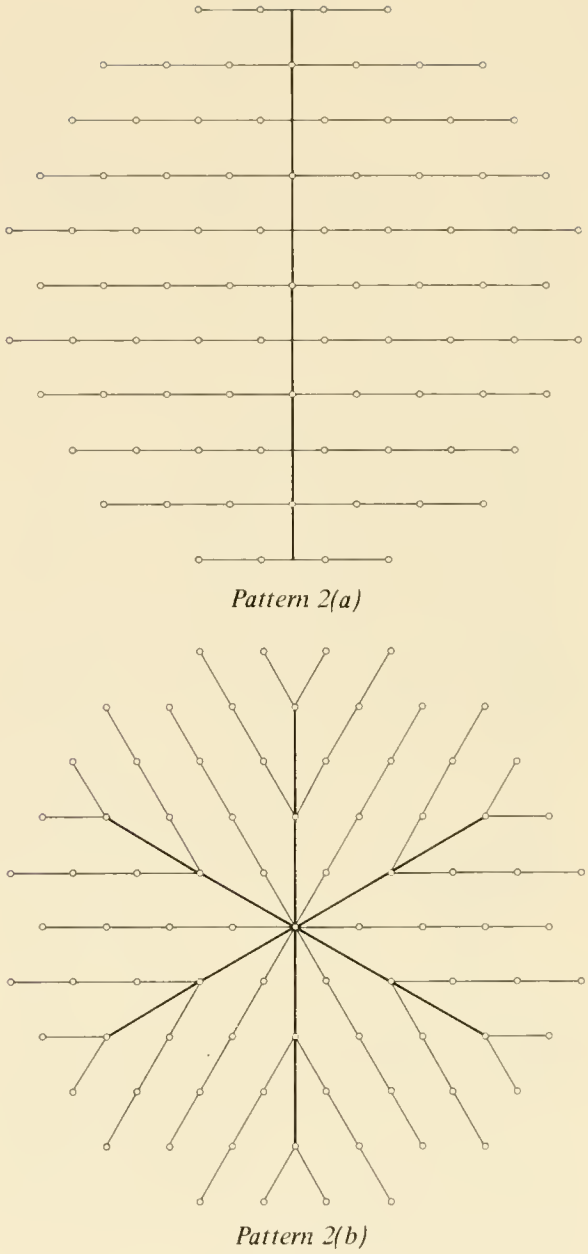
$$N_{1b} = n + 4(\sqrt{2}\,r - 1)\,\text{int}\left(\sqrt{\frac{n}{2\pi} - \frac{1}{2\sqrt{2}}}\right) \tag{L11}$$

For the tunnel pattern of Figure L-2a, $k = \sqrt{3}/2$ and n_m is *twice* the nearest integer to $2a_h/\sqrt{3}\,s$. Thus,

$$N_{2a} = n + (\sqrt{3}\,r - 2)\,\text{int}\left(\sqrt{\frac{2n}{\pi\sqrt{3}} - \frac{1}{\sqrt{3}} + \frac{1}{2}}\right) \tag{L12}$$

Finally, for the tunnel pattern of Figure L-2b, $k = 2$ and N_m is *six* times the nearest integer to $(a_h - s\sqrt{3}/2)/2s$.

$$N_{2b} = n + 6(2r - 1)\,\text{int}\frac{1}{4}\left(\sqrt{\frac{2\sqrt{3n}}{\pi} + 1 - \sqrt{3}}\right) \tag{L13}$$



Pattern 2(a)

Pattern 2(b)

Figure L-2. *Tunnel patterns for hexagonal lattice array.*

If we divide N by n to obtain the pattern factor $f = N/n$, we can express the tunneling cost simply as

$$C = fns\gamma_s = fnC_u \tag{L14}$$

where C_u = cost of a unit side tunnel. Table L-3 lists the pattern factors for the tunnel pattern of Figures L-1 and L-2 for various sizes of array, and for $r = 1,2$.

TABLE L-3

	$r = 1$				$r = 2$			
n	f_{1a}	f_{1b}	f_{2a}	f_{2b}	f_{1a}	f_{1b}	f_{2a}	f_{2b}
100	1	1.05	.987	1.12	1.10	1.22	1.07	1.26
200	1	1.04	.989	1.09	1.07	1.18	1.06	1.27
500	1	2.03	.993	1.06	1.05	1.12	1.04	1.18
1000	1	1.02	.995	1.05	1.03	1.09	1.03	1.14
2000	1	1.01	.997	1.03	1.02	1.06	1.02	1.10

If the side tunnels are 8 ft in diameter and the main tunnels are 10 ft in diameter, the value $r = 2$ is about right. However, the main tunnels do not have to be larger for their entire lengths and, depending on their number, may not need any larger sections. With the 8-ft tunnel cross section shown in Figure L-3, about 8 sq ft or more of area is available for the IF cables. This will accommodate 1700 cables 7/8 in. in diameter or about 500 cables 1-5/8 in. in diameter. The number of array elements that can be used without requiring larger main tunnels is given in the Table L-4.

TABLE L-4

IF CABLE DIAMETER	TUNNEL PATTERN 1*a*	1*b*	2*a*	2*b*
7/8 in.	3432	6840	3430	10,360
1-5/8 in.	1018	2050	1016	3,084

Clearly, if we use the *b* patterns, or 7/8-in. cable, or both, very large arrays can be built without oversize main tunnels. In this case $r = 1$ and the cost difference between patterns is only a few percent—much less than our uncertainty in the basic price. Further, any difference is partially offset by the saving in IF cable lengths when the *b* patterns are used.

At the present state of refinement we can therefore estimate the tunnel cost as

$$C = \$600 \times n \times s \qquad \text{(L15)}$$

where the \$600 is about 6% greater than the estimated cost of \$566/m for 8 ft tunneling.

However, in deciding among the various tunnel patterns, we must take account of the variations given in Table L-3. For convenience we compute the *incremental* cost over that represented by pattern 1*a*

$$\Delta C = \$566 \times n \times s \times (f - f_{1a})$$

$$= \$566\, ns\,(f - 1) \qquad \text{(L15a)}$$

Using $r = 1$ and $s = 300$ m we obtain the incremental cost figures given in Table L-5

TABLE L-5
Excess Tunnel Cost Over That of Pattern 1a

n	ΔC_{1b} (\$ million)	ΔC_{2a} (\$ million)	ΔC_{2b} (\$ million)
100	0.85	−0.22	2.04
200	1.39	−0.37	3.06
500	2.21	−0.59	5.09
1000	3.40	−0.85	8.15
2000	4.75	−1.02	11.21

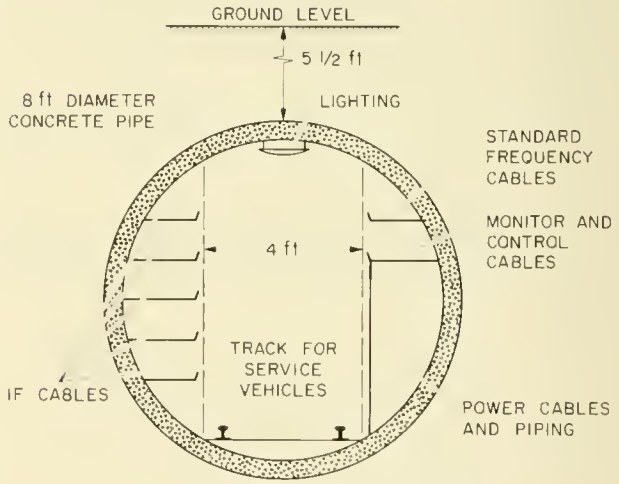


Figure L-3. *Tunnel cross section.*

CABLE LENGTHS

The power, standard frequency and control cable lengths are essentially equal to the tunnel length and will be slightly longer in the *b* patterns than in the *a* patterns. The total table cost of these distribution systems is on the order of \$20/m and so again is a small fraction of the tunnel cost. The IF distribution system on the other hand involves an individual cable to each antenna and is expensive. The total IF cable length for the *b* patterns is substantially less than for the *a* patterns.

If we let 2θ be the central angle of the sector of the array covered by any main tunnel, then $\theta = \pi/m$ where $n = 2,4,6$ is the number of main tunnels. From Figure

L-4 we see that the length of IF cable to an antenna whose coordinates are x,y is

$$\ell = \frac{y}{\sin\theta} + \left(x - \frac{y}{\tan\theta}\right)$$
$$= \frac{1-\cos\theta}{\sin\theta}\, y + x$$
$$= y \tan\frac{\theta}{2} + x \qquad \text{(L16)}$$

The maximum cable length is found by setting x equal to $\sqrt{a^2-y^2}$ in equation (L16) and differentiating. The result is

$$\ell_{max} = \frac{a}{\cos(\theta/2)} = \frac{a}{\cos(\pi/2m)} \qquad \text{(L17)}$$

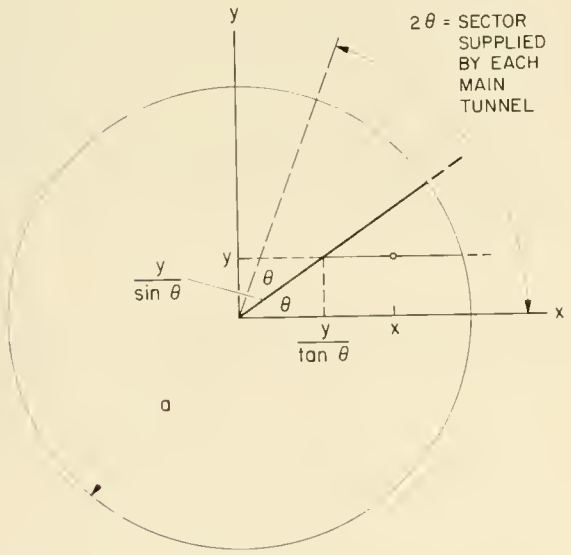


Figure L-4. *Distance through tunnels to an antenna*

The average cable length is very nearly that for a continuous distribution of end points (antennas); that is,

$$\bar{\ell} = \frac{\int \ell\, dA}{\int dA} \qquad \text{(L18)}$$

where dA is an element of the array area. Taking the integrals over the area between the main feeder tunnel and the x-axis (see Figure L-4), we have

$$\ell = \frac{2}{\theta a^2} \int_0^{a\sin\theta} dy \int_{y/\tan\theta}^{\sqrt{a^2-y^2}} \left[y \tan\frac{\theta}{2} + x\right] dx$$

$$= \frac{2a}{3}\, \frac{\tan(\theta/2)}{(\theta/2)} = \frac{4ma}{3\pi} \tan\frac{\pi}{2m} \qquad \text{(L19)}$$

For the tunnel patterns we have been considering the data given in Table L-6

TABLE L-6

θ	m	ℓ_{max}/a	$\bar{\ell}/a$	$\bar{\ell}/\bar{\ell}_0$
$\pi/2$	2	1.4142	0.8488	1.2732
$\pi/4$	4	1.0824	0.7032	1.0548
$\pi/6$	6	1.0353	0.6823	1.0235
0	∞	1.0000	0.6666	1.0000

The case $\theta = 0$ corresponds to having an individual straight tunnel for each IF line, and gives the minimum possible length. We see that the hexagonal lattice with 6 main tunnels (pattern 2*b*) has an average length only 2.35% greater, that the square lattice with 4 main tunnels (pattern 1b) has an average length only 5.5% greater, but that the patterns with only two main tunnels (the *a* patterns) increase the average length by over 27%.

The total length of IF cabling can now be expressed in terms of the number of antennas n, and their spacing s. For our purposes we can omit the term 1/2 in equations (L4) and (L5). Substitution into equation (L19) then gives for pattern 1*a*:

$$L_{1a} = \frac{8}{3}\, \frac{n^{3/2}}{\pi}\, s = 0.4789\, n^{3/2} s \qquad \text{(L20)}$$

for 1*b*:

$$L_{1b} = \frac{16}{3}\, (\sqrt{2}-1)\, \frac{n^{3/2}}{\pi}\, s = 0.3967\, n^{3/2} s \qquad \text{(L21)}$$

for 2*a*:

$$L_{2a} = \frac{8}{3} \sqrt{\frac{\sqrt{3}}{2}}\, \frac{n^{3/2}}{\pi}\, s = 0.4457\, n^{3/2} s \qquad \text{(L22)}$$

and for $2b$:

$$L_{2b} = 8\sqrt{\frac{\sqrt{3}}{2}}(2-\sqrt{3})\frac{n^{3/2}}{\pi}s = 0.3582\,n^{3/2}s \tag{L23}$$

Two cable sizes were shown to be attractive by the VLA study: 7/8-in. diameter solid outer conductor cable costing \$3/m and 1-5/8-in. diameter solid outer conductor cable costing \$6/m. For a low level of multiplexing the smaller cable should be adequate. Assuming a repeater every kilometer costing \$2000, the total cost of the IF transmission system would then be about \$5/m.

We now compute the excess cable cost of the various patterns over that of Figure L-1a; results are given in Table L-7

TABLE L-7
Excess Cable Cost Over That of Pattern 1*a*

n	ΔC_{1b} (\$ million)	ΔC_{2a} (\$ million)	ΔC_{2b} (\$ million)
100	−0.12	−0.05	−0.18
200	−0.35	−0.14	−0.51
500	−1.38	−0.56	−2.02
1000	−3.90	−1.57	−5.73
2000	−11.03	−4.45	−16.19

We may now add the entries in Tables L-5 and L-7 to get the excess total costs given in Table L-8.

TABLE L-8
Excess Total Cost of Cable and Tunnels Over Pattern 1*a*

n	C_{1b} (\$ million)	C_{2a} (\$ million)	C_{2b} (\$ million)
100	0.73	−0.27	1.86
200	1.04	−0.51	2.55
500	0.83	−1.15	3.07
1000	−0.50	−2.42	2.42
2000	−6.28	−5.47	−4.98

On the basis of the first cost alone, the choice among the four patterns is seen to depend on the size of the array. However, the differences are small compared with the total cost and favor the *b* patterns as n increases. Since the ultimate size of Cyclops is somewhat uncertain and could exceed 2000 elements, the *b* patterns are preferable, particularly in view of the greater freedom from delay and phase variations to be expected because of the shorter cable runs.

APPENDIX M

PHASING AND TIME DELAY

Assume that we are receiving a spatially coherent wide-band noise signal from a point source at an angle θ from the zenith as shown in Figure M-1. Let $s_p(t) \cos \omega_c t + s_q(t) \sin \omega_c t$ be the amplitude of this signal on the wavefront OP, where ω_c is the center frequency of the RF band. Then the effective input to receiver A is

$$f_a(t) = [s_p(t) + a_p(t)] \cos \omega_c t + [s_q(t) + a_q(t)] \sin \omega_c t \tag{M1}$$

where a_p and a_q are the in-phase and quadrature component amplitudes of the noise receiver A referred to the input. For receiver B we have

$$f_b(t) = [s_p(t-\tau) + b_p(t)] \cos \omega_c(t-\tau) + [s_q(t-\tau) + b_q(t)] \sin \omega_c(t-\tau) \tag{M2}$$

In these expressions s_p and s_q are statistically similar but uncorrelated; that is, $\overline{s_p s_q} = 0$. Likewise a_p, a_q, b_p, b_q are all statistically similar and uncorrelated; that is, the average value of any cross product is zero. So is the average of any cross product of an s and an a or b term.

Assume the local oscillator at receiver A has the amplitude $\cos \omega_0 t$ while at receiver B has the amplitude

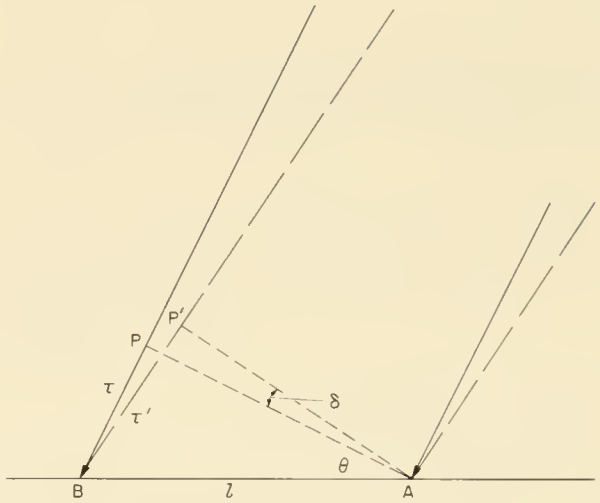


Figure M-1. *Delay and phase differences between two antennas.*

$\cos(\omega_0 t + \phi)$. The IF signals generated at the two receivers are then given by:

$$g_a(t) = [s_p(t) + a_p(t)] \cos \omega_i t + [s_q(t) + a_q(t)] \sin \omega_i t \tag{M3}$$

$$g_b(t) = [s_p(t-\tau) + b_p(t)] \cos [\omega_i t - \omega_c \tau - \phi] + [s_q(t-\tau) + b_q(t)] \sin [\omega_i t - \omega_c \tau - \phi] \tag{M4}$$

where $\omega_i = \omega_c - \omega_0$. Let us now consider various ways of combining these signals.

CASE I – PERFECT LO PHASING AND DELAY MATCHING

If we set

$$\phi = -\omega_0 \tau \text{ , mod } 2\pi \tag{M5}$$

then equation (M4) becomes

$$g_b(t) = [s_p(t-\tau) + b_p(t)] \cos \omega_i(t-\tau) + [s_q(t-\tau) + b_q(t)] \sin \omega_i(t-\tau) \tag{M6}$$

If we now delay $g_a(t)$ by the time τ and add $g_a(t-\tau)$ to equation (M6) we get for the combined signal

$$g(t) = [2s_p(t-\tau) + a_p(t-\tau) + b_p(t)] \cos \omega_i(t-\tau) + [2s_q(t-\tau) + a_q(t-\tau) + b_q(t)] \sin \omega_i(t-\tau) \tag{M7}$$

which represents a power

$$P = \frac{1}{2} \left\{ [2s_p(t-\tau) + a_p(t-\tau) + b_p(t)]^2 + [2s_q(t-\tau) + a_q(t-\tau) + b_q(t)]^2 \right\} \tag{M8}$$

$$= \frac{1}{2} \left\{ \overline{4s_p^2} + \overline{a_p^2} + \overline{b_p^2} + \overline{4s_q^2} + \overline{a_q^2} + \overline{b_q^2} \right\} \tag{M9}$$

If P_s is the received signal power per receiver and P_n is the noise power per receiver, then

$$P_s = \frac{1}{2} (s_p^2 + s_q^2) = \overline{s_p^2} = \overline{s_q^2} \tag{M10}$$

$$P_n = \frac{1}{2} (\overline{a_p^2} + \overline{a_q^2}) = \frac{1}{2} (\overline{b_p^2} + \overline{b_q^2})$$

$$P_n = \overline{a_p^2} = \overline{a_q^2} = \overline{b_p^2} = \overline{b_q^2} \tag{M11}$$

Thus

$$P = 2 [2P_s + P_n] \tag{M12}$$

The signal to noise ratio of the combined signal is thus twice that of the individual signals. For n signals added this way, the improvement is n fold.

CASE II – NO LO PHASING

Here we set $\phi = 0$ in equation(M4) and delay the IF signals of (M3) by an amount $\tau_a = \tau + \Delta\tau$ to obtain

$$g_a(t-\tau_a) = [s_p(t-\tau_a) + a_p(t-\tau_a)] \cos \omega_i(t-\tau_a) + [s_q(t-\tau_a) + a_q(t-\tau_a)] \sin \omega_i(t-\tau_a) \tag{M13}$$

$$g_b(t) = [s_p(t-\tau) + b_p(t)] \cos (\omega_i t - \omega_c \tau) + [s_q(t-\tau) + b_q(t)] \sin (\omega_i t - \omega_c \tau) \tag{M14}$$

If we now set

$$\omega_i \tau_a = \omega_c \tau \tag{M15}$$

the arguments of the trigonometric functions will agree. We note that if $\omega_i = \omega_c$, $\tau_a = \tau$ and the delays match. This merely says that if there is no heterodyning or if the IF channels are remodulated up to the original frequency before combining them (and the delay is introduced *after* this remodulation), no delay error is introduced.

Since ordinarily $\omega_i/\omega_c << 1$, condition (M15) would require that $\tau_a >> \tau$ and hence also that $\Delta\tau >> \tau$ were it not for the fact that (M15) need only be satisfied modulo 2π. We thus can set

$$\tau_a = \frac{\omega_c}{\omega_i} \tau - \frac{2\pi k}{\omega_i} \text{ , } \quad k = 0,1,2 \ldots \tag{M16}$$

and obtain

$$-\frac{\pi}{\omega_i} \leqslant \Delta\tau \leqslant \frac{\pi}{\omega_i} \qquad \text{(M17)}$$

as shown in Figure M-2. (The negative delay shown can be avoided by adding delay to receiver *B*.)

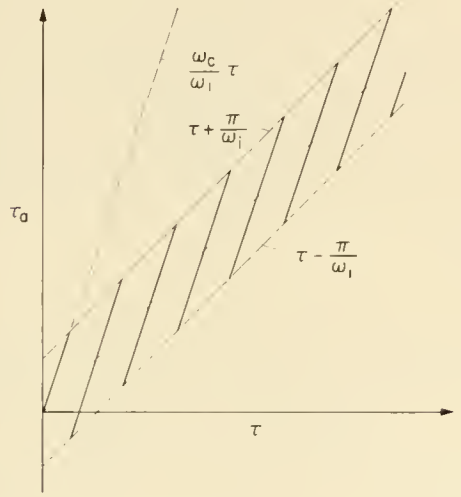


Figure M-2. *Delay required to phase array with no local oscillator phase shift.*

If we assume the arguments of the trigonometric functions have been set equal in this fashion, and the signals have been added, the power will be

$$P = \frac{1}{2}\left\{\overline{[s_p(t-\tau_a) + s_p(t-\tau) + a_p(t-\tau_a) + b_p(t)]^2} + \overline{[s_q(t-\tau_a) + s_q(t-\tau) + a_q(t-\tau_a) + b_q(t)]^2}\right\}$$

$$= \frac{1}{2}\left\{2\overline{s_p^2} + 2R_p(\Delta\tau) + \overline{a_p^2} + \overline{b_p^2} + 2\overline{s_q^2} + 2R_q(\Delta\tau) + \overline{a_q^2} + \overline{b_q^2}\right\} \qquad \text{(M18)}$$

where

$$R(\Delta t) \equiv \overline{s(t)s(t+\Delta\tau)}$$

If we let $\psi(\Delta\tau) \equiv R(\Delta\tau)/R(0)$ be the normalized autocorrelation of the received signal, then for the addition of two channels, we find

$$P = 2\left(P_s\ [1 + \psi(\Delta\tau)] + P_n\right) \qquad \text{(M19)}$$

If n channels are added,

$$P = n\left\{P_s\left[1 + \frac{1}{n}\sum_{i=1}^{\frac{n(n-1)}{2}}\psi(\Delta\tau_i)\right] + P_n\right\} \qquad \text{(M20)}$$

since there are $n(n-1)/2$ cross-product terms between pairs of antennas, each having (possibly) different delay errors.

Let us now define $\overline{\psi}$ as the average value of $\psi(\Delta\tau_i)$. Then equation (M20) becomes

$$P = n[P_s(1 + \frac{n-1}{2}\ \overline{\psi}) + P_n] \qquad \text{(M21)}$$

With no delay errors $\overline{\psi} = 1$ and

$$P \equiv P_0 = n\left(\frac{n+1}{2}\ P_s + P_n\right) \qquad \text{(M22)}$$

The loss in signal power (and *S/N* ratio) caused by the delay errors is therefore

$$L = 10\log\frac{2 + (n-1)\overline{\psi}}{n+1} \qquad \text{(M23)}$$

which for large n becomes simply

$$L \approx 10\log\overline{\psi} \qquad \text{(M24)}$$

The delay errors of the individual channels will tend to be uniformly distributed over the range $-(1/2f_i)$ to $(1/2f_i)$. Thus, the delay differences between pairs of channels will have the triangular distribution

$$p(\Delta\tau) = f_i(1 - f_i|\Delta\tau|), \quad |\Delta\tau| < \frac{1}{f_i}$$
$$= 0 \qquad , \quad |\Delta\tau| \geqslant \frac{1}{f_i} \qquad \text{(M25)}$$

If the source is "white," ψ will be entirely determined by the receiver selectivity characteristic. If $F(\omega)$

is the low pass equivalent of this selectivity characteristic, then ψ is the Fourier transform of $|F(\omega)|^2$. Let us assume the receiver has an ideal bandpass characteristic B Hz wide. Then

$$\psi(\Delta\tau) = \frac{\sin \pi B \Delta\tau}{\pi B \Delta\tau} \qquad \text{(M26)}$$

Since (M25) and (M26) are even functions, their product is also even and:

$$\bar{\psi} = 2f_i \int_0^{1/f_i} (1 - f_i \Delta\tau) \frac{\sin \pi B \Delta\tau}{\pi B \Delta\tau} \, d(\Delta\tau)$$

$$= \frac{2f_i}{\pi B} \left[\mathrm{Si}(\pi B/f_i) + \frac{f_i}{\pi B} (\cos \pi B/f_i - 1)\right] \qquad \text{(M27)}$$

Thus the loss due to the delay errors uniformly distributed from $-1/2f_i$ to $1/2f_i$ is

$$L = 10 \log 2 \left(\frac{\mathrm{Si}\, u}{u} + \frac{\cos u - 1}{u^2} \right)$$

$$u = \pi \frac{B}{f_i} \qquad \text{(M28)}$$

Figure M-3 is a plot of L as a function of B/f_i. We see that for a 100 MHz band centered at 125 MHz ($B/f_i = 0.8$) the loss is about 0.73 dB. This is the loss we would incur in observing wideband incoherent signals if we adjusted the delays of the entire IF signal (both bands plus the pilot signal) to get the proper phasing, and then used the pilot to modulate both IF bands down to the range 75 to 175 MHz.

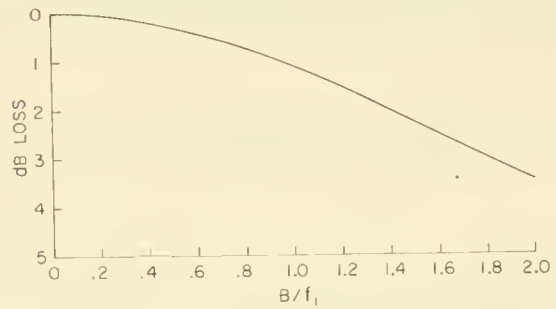


Figure M-3. *Decorrelation loss produced by delay phasing.*

On the other hand, if we use delay phasing on the pilots only, the loss will be much less since shorter delays are needed to get the required phase shift at the pilot frequency.

Pure delay phasing is thus a distinct possibility for Cyclops when observing wideband signals, and deserves further study.

CASE III – ARRAY PHASED FOR ANGLE θ, DELAYS ADJUSTED TO PRODUCE OFF AXIS BEAM

Assume the local oscillators are phased for a signal arriving at an angle θ from the zenith, and the signal from antenna A is delayed an amount $\tau_a = \tau + \Delta\tau$ to produce an off-axis beam making an angle $\theta + \delta$ with the zenith. The delay at receiver B for a signal from this direction is

$$\tau' = \frac{d}{c} \sin(\theta + \delta)$$

$$= \frac{d}{c} (\sin\theta \cos\delta + \cos\theta \sin\delta)$$

and for $\delta \ll 1$

$$\tau' \approx \frac{d}{c} \sin\theta + \frac{d \cos\theta}{c} \delta$$

$$\approx \tau + \frac{d \cos\theta}{c} \delta = \tau + \Delta\tau' \qquad \text{(M29)}$$

The signals from receiver A (delayed) and receiver B are

$$g_a(t - \tau_a) = [s_p(t - \tau_a) + a_p(t - \tau_a)] \cos \omega_i (t - \tau_a)$$

$$+ [s_q(t - \tau_a) + a_q(t - \tau_a)] \sin \omega_i (t - \tau_a)$$

$$\text{(M30)}$$

$$g_b(t - \tau') = [s_p(t - \tau') + b_p(t)] \cos [\omega_i (t - \tau)$$

$$- \omega_i (\tau' - \tau)]$$

$$+ [s_q(t - t') + b_q(t)] \sin [\omega_i (t - \tau)$$

$$- \omega_c (\tau' - \tau)] \qquad \text{(M31)}$$

For the arguments to be equal we require

$$\omega_i(t - \tau_a) = \omega_i(t - \tau) - \omega_c(\tau' - \tau)$$

$$\omega_i \Delta\tau = \omega_c \Delta\tau' \tag{M32}$$

$$\begin{aligned} \tau_a - \tau' &= \tau + \Delta\tau - \tau' \\ &= \left(\frac{\omega_c}{\omega_i} - 1\right) \Delta\tau' \\ &= \left(\frac{\omega_c}{\omega_i} - 1\right) \frac{\ell \cos\theta}{c} \delta \end{aligned} \tag{M33}$$

Assume that $\delta = \lambda c/2d$, that is, that we are attempting to image out to the half-power angle of each element. Taking $\theta = 0$, we find for the delay error

$$\begin{aligned} \tau_a - \tau' &= \left(\frac{\omega_c}{\omega_i} - 1\right) \frac{\ell}{d} \frac{\pi}{\omega_c} \\ &= \left(\frac{1}{\omega_i} - \frac{1}{\omega_c}\right) \pi \frac{\ell}{d} \end{aligned}$$

$$\omega_i(\tau_a - \tau') = \left(1 - \frac{\omega_i}{\omega_c}\right) \pi \frac{\ell}{d} \tag{M34}$$

Now ℓ/d may be on the order of 100, so on the order of 50 $(1 - (\omega_i/\omega_c))$ cycles of delay error will exist near the IF frequency. Thus to avoid a serious decorrelation loss we must either (1) Use a very narrow (<1% BW) IF, (2) remodulate to make $\omega_i \approx \omega_c$ and do the incremental delay near the original frequency, or (3) satisfy (M32) modulo 2π.

These alternatives are discussed in Chapter 9 where various imaging systems are described.

APPENDIX N

SYSTEM CALIBRATION

The phase of the signal from a given antenna, as it reaches the data processing center, depends on (1) the geometrical time delay caused by the orientation of the array with respect to the source, and (2) several other effects that may be regarded as errors, including unequal amounts of refraction in the atmosphere above the various antenna and unequal phase shifts in the various amplifiers, transmission lines, and other components. These errors must be eliminated, insofar as possible, which means that each antenna channel must be phase-calibrated at regular intervals.

The most practical method of phase calibration would appear to involve well-known cosmic radio sources. One of the several antennas in the array is chosen as the reference, with which all others are compared. A correlator, that is, a multiplier capable of accommodating the IF bandwidth, is connected between the reference antenna and each antenna to be calibrated. The output of this correlator is given by

$$F = \frac{2\pi D}{\lambda_0} \quad [\xi - \sin\delta \sin d - \cos\delta \cos d \cos(H - h)]$$

in which

D = distance between antennas

λ_0 = wavelength at the center of the RF passband

ξ = incidental phase shift in the electronic system, in fractional wavelength

δ = declination of the cosmic radio source

d = declination at which the baseline interests the celestial sphere

H = hour-angle of the cosmic source

h = hour-angle at which the baseline interests the celestial sphere

Presumably H and δ are known precisely; H increases uniformly with time; therefore, the "fringe function" F oscillates with time as the source moves with respect to the baseline between the antennas.

The calibration procedure would be to observe the source for several minutes and to store the output of the correlator. From assumed values of the parameters in the above formula a synthetic fringe function is computed and compared with the observed function. The unknown parameters, say ξ, h and d, are varied until the mean-square difference between the observed and the computed fringe function is minimized. The resulting values of the parameters constitute the calibration factors of the system. Atmospheric refraction effects will be included in the parameter ξ.

Five minutes of observation would probably be required to calibrate one channel. A number of channels could be calibrated simultaneously if a sufficient computing capacity were available. The permissible interval between calibrations will depend on the mechanical and electronic stability of the system, but will probably be on the order of days.

Atmospheric effects will probably vary much more rapidly than outlined here; except under unusual circumstances, however, it is not expected that atmospheric refraction will degrade the array performance seriously. Small-scale atmospheric inhomogeneity will increase the RMS phase error and will therefore decrease the gain of the array, while large-scale inhomogeneity will cause beam-steering errors. Experience with current radio-astronomical interferometers suggests that phase errors as large as 30° may occasionally occur over baselines of a few km at 10-cm wavelength. This effect varies strongly with weather and with season at a given site, and it is

apparent that some sites are considerably better than others. Proper site selection, therefore, is of great importance.

The calibration scheme discussed here should permit phase errors to be reduced to, say, 5° or less in 5 minutes of observation, exclusive of atmospheric errors.

APPENDIX O

THE OPTICAL SPECTRUM ANALYZER

If a transparency having the complex amplitude and transmittance $g(x,y)$ is placed in plane P_1 of Figure 11-5 and illuminated by a unit amplitude monochromatic plane wave of wavelength λ, then the distribution of complex amplitude in plane P_2 is approximately (ref. 1)

$$G(u,v) = \frac{1}{j\lambda f}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\exp\left[\frac{-j2\pi}{\lambda f}(ux+vy)\right]dx\,dy$$

where f is the focal length of the lens. Thomas (ref. 2) and Markevitch (ref. 3) have shown that the one-dimensional data handling capability of this basically two-dimensional operation can be greatly increased by converting the one-dimensional signal to be analyzed, $s(t)$, into a raster-type, two-dimensional format as shown in Figure 11-4. The following parameters are defined:

- b = width of spectrum analyzer input window
- h = length of spectrum analyzer input window
- c = scan line spacing
- N = scan lines within input window = h/c
- a = width of scan line
- B_0 = maximum signal frequency
- k = maximum spatial frequency of recorded signals
- V = recording scan velocity = B_0/k
- ρ = frequency resolution in spectrum
- T = time window temporal duration of signal within input window of analyzer

The signal to be analyzed $s(t)$, is suitably limited, added to a bias to make it non-negative, and then recorded by some sort of scanned recorder as explained in Chapter 11. Following Thomas, the nth line of recording has an amplitude transmittance

$$s_n(x) = s\left[\frac{x+(2n-1)\,b/2}{V}\right]\text{rect}\left(\frac{x}{b}\right),\quad 1\leqslant n\leqslant N$$

and the y variation

$$s_n(y) = \delta\left\{y-\left[\frac{h-(2n-1)\,c}{2}\right]\right\} * \text{rect}\left(\frac{y}{a}\right)$$

where

$$\text{rect}(\xi) = \begin{cases} 1, & |\xi| < 1/2 \\ 0, & \text{elsewhere} \end{cases}$$

and $*$ denotes convolution. The overall transmittance is

$$g(x,y) = \sum_{n=1}^{N} s\left[\frac{x+(2n-1)\,b/2}{V}\right]\text{rect}\left(\frac{x}{b}\right)\left[\delta\, y-\left[\frac{h-(2n-1)\,c}{2}\right] * \text{rect}\left(\frac{y}{a}\right)\right]$$

and, neglecting constants, the output plane complex amplitude is,

$$G(u,v) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\exp\left[\frac{-j2\pi}{\lambda f}(ux+vy)\right]dx\,dy$$

Define

$$Q_n(x) = s\left[\frac{x + (2n-1)\,b/2}{V}\right] \text{ rect } \left(\frac{x}{b}\right)$$

and

$$P_n(y) = \delta\left\{y - \left[\frac{h-(2n-1)c}{2}\right]\right\} * \text{rect}\left(\frac{y}{a}\right)$$

Then

$$G(u,v) = \sum_{n=1}^{N} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Q_n(x)\exp\frac{-j2\pi}{\lambda f} ux\ dx \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_n(y)\exp\frac{-j2\pi}{\lambda f} vy\ dy$$

The first integral yields

$$Vb\ \text{sinc}\left(\frac{bu}{\lambda f}\right) * \left\{\exp\left[\frac{j\pi(2n-1)bu}{\lambda f}\right] S\left(\frac{2\pi Vu}{\lambda f}\right)\right\}$$

and the second gives

$$a\ \text{sinc}\left(\frac{av}{\lambda f}\right)\exp\left\{\frac{-j\pi}{\lambda f}\,[h-(2n-1)c]\,v\right\}$$

where $\text{sinc}(\xi) = \sin \pi\xi/\pi\xi$ and $S(\omega)$ is the spectrum of $s(t)$.

Thus

$$G(u,v) = Vab\ \text{sinc}\left(\frac{av}{\lambda f}\right)\sum_{n=1}^{N}\left(\sin c\left(\frac{bu}{\lambda f}\right) * \left\{S\left(\frac{2\pi Vu}{\lambda f}\right)\exp\left[\frac{j\pi(2n-1)bu}{\lambda f}\right]\right\}\right)$$

$$\exp\left\{\frac{-j\pi}{\lambda f}\,[h-(2n-1)c]\,v\right\}$$

Now let $s(t)$ be a sinusoidal signal such that

$$S(\omega) = \delta(\omega - \omega_0).$$

Then

$$G(u,v) = Vab\ \text{sinc}\left(\frac{av}{\lambda f}\right)\text{sinc}\left(\frac{bu}{\lambda f} - \frac{b\omega_0}{2\pi V}\right) \times \exp\left(\frac{-j\pi hv}{\lambda f}\right)\sum_{n=1}^{N}\exp j(2n-1)\left(\frac{b\omega_0}{2V} + \frac{\pi cv}{\lambda f}\right)$$

The displayed intensity pattern is

$$I(u,v) = |G(u,v)|^2 = (Vab)^2\ \text{sinc}^2\left(\frac{av}{\lambda f}\right)\text{sinc}^2\left(\frac{bu}{\lambda f} - \frac{b\omega_0}{2\pi V}\right) \times \sum_{n=1}^{N}\exp[j(2n-1)\theta]\sum_{n=1}^{N}\exp[-j(2n-1)\theta]$$

where

$$\theta = \left(\frac{b\omega_0}{2V} + \frac{\pi cv}{\lambda f}\right)$$

The product of the two series is equal to $\sin^2 N\theta/\sin^2\theta$ which has narrow (for large n) peaks at $\theta = \ell\pi$, $\ell = 0, \pm 1, \ldots$ The u coordinate of the bright spots is $u = (\lambda f/2\pi V)\omega_0$ and the width of the spot between first zeros is $2\lambda f/b$. The u coordinate of these spots gives coarse frequency information. In the v direction a comb function with a sinc^2 envelope is obtained. The distance between zeros of the sinc^2 function is $2\lambda f/a$; the comb peaks occur at

$$v = \frac{\lambda f}{c}\left(n - \frac{b\omega_0}{2\pi V}\right)$$

and have a spacing of $\lambda f/c$. Since $a < c$, the envelope width is greater than the comb spacing. Thus, within $|v| \leqslant \lambda f/2c$ there is one comb spike at a time. The location

of this spike gives fine frequency resolution—it is a frequency vernier. The frequency resolution in the v direction is

$$\Delta v = \frac{-\lambda f b}{2\pi V c}\,\Delta\omega_0 = \frac{-\lambda f b \rho}{Vc}$$

where

$$\Delta\omega_0 = 2\pi\rho$$

For N signal scan lines, the total temporal duration of the signal within the input window is $T = Nb/V$. According to Thomas, there are at least N resolution elements along a frequency locus within $|v| \leqslant \lambda f/2c$. Thus $\Delta v = \lambda f/Nc$ which gives a frequency resolution of

$$\rho = \frac{Vc}{\lambda f b}\,\Delta v = \frac{V}{Nb}$$

or

$$\rho = \frac{1}{T}$$

We see that the frequency resolution achieved is as good as can theoretically be expected, the reciprocal of the integration time.

The output format consists of a family of frequency loci as shown in Figure 11-2. For a given signal component at ω_0 the output display is a set of bright spots spaced $\lambda f/c$ apart along the line $u = (\lambda f/2\pi V)\omega_0$. The row of spots moves downward with increasing ω_0. As one spot moves below $v = -(\lambda f/2c)$, another crosses downward through $v = (\lambda f/2c)$ on the next higher frequency locus. There is always one spot within $|v| \leqslant \lambda f/2c$ and its location determines ω_0.

The input scan line cross section was assumed to be rectangular in the above discussion; this introduced a sinc^2 weighting of the spectral display in the v direction. In practice, the scan lines will not have a rectangular cross section and the spectral weighting will not be sinc^2. In general, if the cross-section distribution is $\psi(y/a)$, then the spectral weighting will be $|\psi|(av/\lambda f)^2$ where $\psi(\eta)$ is the Fourier transform of $\psi(\xi)$.

REFERENCES

1. Goodman, J.W.: Introduction to Fourier Optics. Chap. 8, McGraw-Hill 1968.
2. Thomas, Carlton E.: Optical Spectrum Analysis of Large Space Bandwidth Signals. Applied Optics, 5, 1966, p. 1782.
3. Markevitch, Bob. V.: Optical Processing of Wideband Signals. Third Annual Wideband Recording Symposium, Rome Air Development Center, April 1969.

APPENDIX P

LUMPED CONSTANT DELAY LINES

Lumped constant low pass constant-k filters make reasonably good video delay lines, but their delay constancy with frequency can be greatly improved by adding mutual inductance between the coils as shown in Figure P-1a. The equivalent circuit is then as shown in Figure P-1b. If $\Gamma = \ell n(i_{n+1}/i_n)$ is the propagation constant, then as is well known for a ladder structure

$$\cosh \Gamma = 1 + \frac{ZY}{2} \tag{P1}$$

where Z is the series impedance and Y is the shunt admittance. In our case $Z = j\omega(L+2M)$ and $Y = [(1/j\omega C) - j\omega M]^{-1}$ so, letting $k = M/L$,

$$\cosh \Gamma = \frac{1 - \omega^2 LC/2}{1 + k\omega^2 LC} \tag{P2}$$

The cutoff frequency occurs when $\cosh \Gamma = 1$, or

$$\omega = \omega_c = \frac{2}{\sqrt{L(1-2k)C}} \tag{P3}$$

In the passband, $\omega < \omega_c$, $\cosh \Gamma = \cos \beta$, where β is the phase shift per section. The delay per section is $\tau = -d\beta/d\omega$. Thus

$$\tau = -\frac{d}{d\omega} \cos^{-1} \frac{1 - \omega^2 LC/2}{1 + k\omega^2 LC} \tag{P4}$$

$$= \frac{2}{\omega_c} \sqrt{\frac{1-2k}{1+2k}} \times \left[1 + \frac{4k}{1-2k} \left(\frac{\omega}{\omega_c}\right)^2 \sqrt{1 - \left(\frac{\omega}{\omega_c}\right)^2} \right]^{-1} \tag{P5}$$

If $k = 0$ we obtain

$$\tau_0 = \frac{2}{\omega_c} \frac{1}{\sqrt{1-(\omega/\omega_c)^2}} \tag{P6}$$

For delay flat to the fourth order in ω, we set or $k = 1/10$ and obtain

$$\tau = \frac{2}{\omega_c} \sqrt{\frac{2}{3}} \frac{1}{[1 + (\omega/\omega_0)^2/2] \sqrt{1-(\omega/\omega_c)^2}} \tag{P7}$$

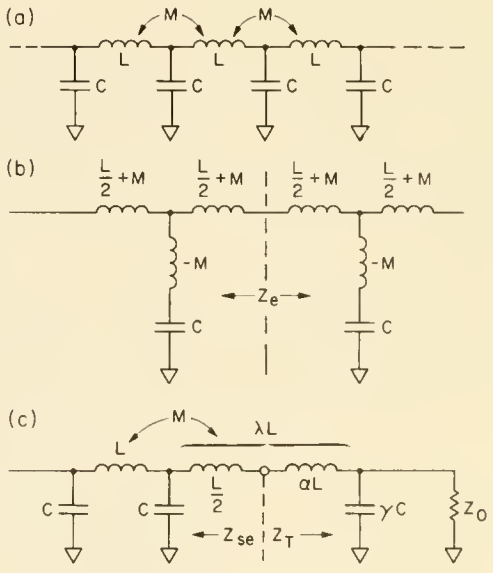


Figure P-1. *Delay line with inertial inductance.*

The delay relative to the delay at zero frequency is

$$\frac{\tau}{\tau(0)} = \left[\left[1 + \frac{4k}{1-2k}\left(\frac{\omega}{\omega_c}\right)\right]^2 \sqrt{1 - \frac{\omega}{\omega_c}^2}\right]^{-1} \tag{P8}$$

Figure P-2 is a plot of equation (P8) for $k = 0$ and $k = 1/10$. Figure P-3 is an expanded plot for various values of k from 0.1 to 0.11. We see, for example, that if the line is to be used only up to $f/f_c = 1/4$, a value of k of about 0.1033 would yield a delay constant with frequency to ±0.025%. Thus, for a total delay of 200 Nyquist intervals the delay error at any frequency would be less than ±0.05 Nyquist interval. At this level of accuracy second-order effects caused by coupling between alternate coils become important and must be included in the analysis.

To terminate the line we first notice that the midseries impedance of the line (i.e., the impedance seen in either direction on an infinitely long line from the center of a series inductance element) is

$$Z = Z_0\sqrt{1 - (\omega/\omega_0)^2} \tag{P9}$$

when

$$Z_0 = \sqrt{\frac{L(1+2k)}{C}} \tag{P10}$$

If we now terminate the line at such a point and make the termination consist of an added series inductance αL followed by a shunt capacitor γC and resistance of value Z_0, as shown in Figure P-1c, the impedance looking into this terminating half section will be

$$Z_T = j\omega\alpha L \frac{Z_0}{1 + j\omega C Z_0}$$

$$= Z_0 \frac{\left[1 - \frac{4\alpha\gamma}{1-2k}\left(\frac{\omega}{\omega_c}\right)^2\right] + j\frac{2\alpha}{\sqrt{1-4k^2}}\left(\frac{\omega}{\omega_c}\right)^2}{1 + j2\gamma\sqrt{\frac{1+2k}{1-2k}}\left(\frac{\omega}{\omega_c}\right)} \tag{P11}$$

and the reflection coefficient, $\rho = Z_T - Z/Z_T + Z$, will be given by

$$\rho = \frac{1 - \frac{4\alpha\gamma}{1-2k}\left(\frac{\omega}{\omega_c}\right)^2 - \sqrt{1-\left(\frac{\omega}{\omega_c}\right)^2}}{1 - \frac{4\alpha\gamma}{1-2k}\left(\frac{\omega}{\omega_c}\right)^2 + \sqrt{1-\left(\frac{\omega}{\omega_c}\right)^2}} \cdots$$

$$\cdots \frac{+j\frac{2\omega/\omega_c}{\sqrt{1-(\omega/\omega_c)^2}}\left[\alpha - \gamma(1+2k)\sqrt{1-\left(\frac{\omega}{\omega_c}\right)^2}\right]}{+j\frac{2\omega/\omega_c}{\sqrt{1-(\omega/\omega_c)^2}}\left[\alpha + \gamma(1+2k)\sqrt{1-\left(\frac{\omega}{\omega_c}\right)^2}\right]} \tag{P12}$$

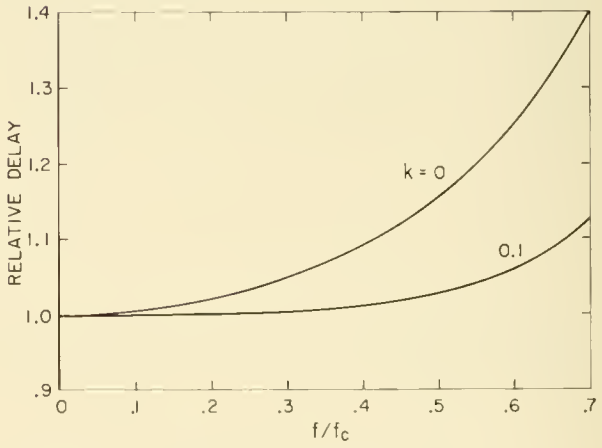


Figure P-2. *Delay versus frequency for* k = 0, 0.1.

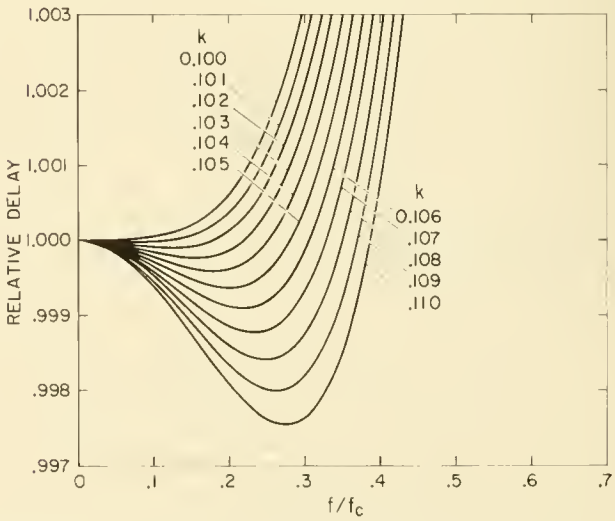


Figure P-3. *Delay versus frequency for* 0.1 ≤ k ≤ 0.11.

The value of the final coil on the line is

$$\lambda L = [(1/2) + \alpha] L.$$

If we simply make $\alpha = \gamma = 1/2$ (thus making $\lambda = 1$ and all coils identical and ending the line in a half-section capacitance), the magnitude of ρ will be shown by the top curve of Figure P-4. For $\omega/\omega_c \ll 1$, the reflection coefficient is approximately $k(\omega/\omega_c)$ and arises from the imaginary term in the numerator.

Some improvement occurs if we make the bracket in this term approach zero as $\omega/\omega_c \to 0$.

This requires that

$$\gamma(1 + 2k) = \alpha \qquad \text{(P13)}$$

For $\alpha = 1/2$ and $k = 1/10$ we find $\alpha = 5/12$, and this case is shown as the middle curve of Figure P-4. Now the reflection for $\omega/\omega_c \ll 1$ is principally due to the real part of the numerator of equation (P12) and can be further reduced if we set

$$\frac{4\alpha\gamma}{1 - 2k} = \frac{1}{2} \qquad \text{(P14)}$$

Solving equations (P13) and (P14) simultaneously, we obtain

$$\gamma = \frac{1}{2\sqrt{2}} \sqrt{\frac{1 - 2k}{1 + 2k}}$$

$$\alpha = \frac{1}{2\sqrt{2}} \sqrt{1 - 4k^2} \qquad \text{(P15)}$$

With these values, equation (P12) becomes

$$\rho = \frac{1 - (\omega/\omega_c)^2/2 - \sqrt{1 - (\omega/\omega_c)^2}}{1 - (\omega/\omega_c)^2/2 - \sqrt{1 - (\omega/\omega_c)^2}} \cdots$$

$$\cdots \frac{+ j(\omega/\omega_c)\left[1 - \sqrt{1 - (\omega/\omega_c)^2}\right]/\sqrt{2}}{+ j(\omega/\omega_c)\left[1 + \sqrt{1 + (\omega/\omega_c)^2}\right]/\sqrt{2}} \qquad \text{(P16)}$$

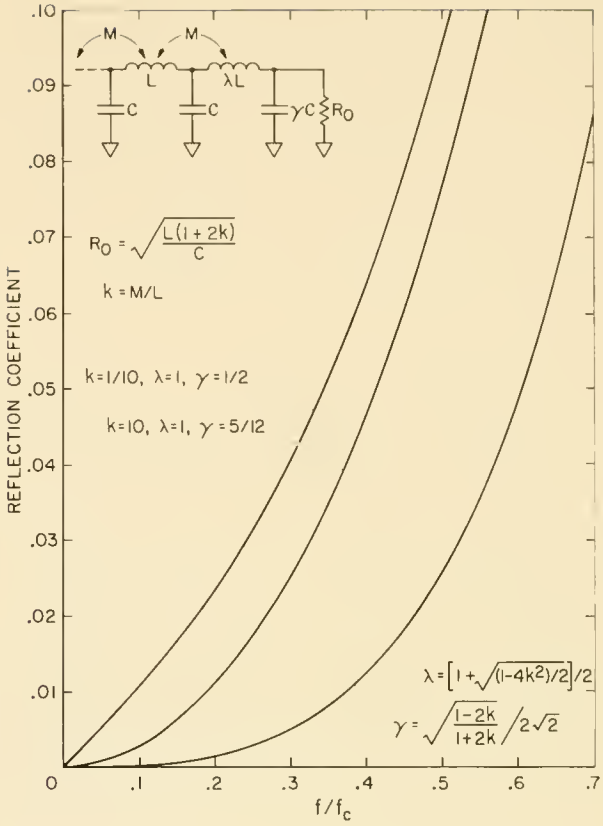


Figure P-4. *Reflection coefficient versus frequency.*

which is independent of k. This case is shown as the bottom curve of Figure P-4. Up to one third the cutoff frequency the reflection is only 0.7%; up to one fourth the cutoff frequency, only 0.3%. Since we will only be using the line up to about this frequency, this simple termination method seems entirely adequate. Thus, we find for $k = 0.1$:

$\alpha = 0.3464$

$\lambda = 0.8464$

$\gamma = 0.2887$

We note that the end coils have about 85% of the inductance of the others. This reduction can easily be achieved without materially affecting the mutual inductance simply by removing a few end turns.

In summary, if Z_0 is the desired characteristic impedance, and τ_0 is the desired delay per section, then

$$C = \tau_0/Z_0 \qquad \xrightarrow[k = 1/10]{} \tau_0/Z_0$$

$$L = \tau_0/Z_0/(1 + 2k) \qquad \to 0.8\overline{3}\,\tau_0 Z_0$$

Conversely

$$Z_0 = \sqrt{1+2k}\ \sqrt{L/C} \rightarrow 1.095\sqrt{L/C}$$

$$\tau_0 = \sqrt{1+2k}\ \sqrt{LC} \rightarrow 1.095\sqrt{L/C}$$

$$f_c = \sqrt{\frac{1+2k}{1-2k}} \Big/ \pi\tau_0 \rightarrow 0.39/\tau_0$$

For the end sections

$$\lambda = \frac{1}{2}\left(1+\sqrt{\frac{1}{2}-2k^2}\right) \rightarrow 0.8464$$

$$\gamma = \frac{1}{2\sqrt{2}}\sqrt{\frac{1-2k}{1+2k}} \rightarrow 0.2887$$

TAPS

When a line is tapped the input capacitance of the bridging amplifier should be compensated by reducing the shunt capacitance of that section by the same amount. Thus, in Figure P-5 we make $C' = C - C_i$. The input resistance of the amplifier should be made as high as possible. For the simple amplifier shown it is βR, which can easily exceed $1000\,R_0$.

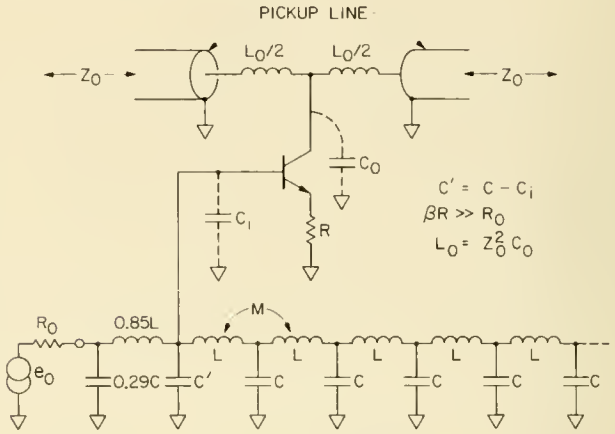


Figure P-5. *Tapping a lumped delay line.*

Finally, the output capacitance of the amplifier should be compensated so as to avoid reflections on the pickup line. This can either be done by adding series inductance $L_0 = Z_0{}^2 C_0$ as shown or by removing a capacitance C_0 from the pickup line the injection point.

APPENDIX Q

CURVES OF DETECTION STATISTICS

To calculate the range limits of the Cyclops system for given probabilities of a false alarm and of detection failure, it is necessary to compute expression (D41) and to integrate expression (D38) given in Appendix D. Expression (D41) gives the probability, in an average of n samples of the output of a square-law detector, that the noise alone will exceed a certain threshold, while (D38) gives the probability that the noise plus signal will not.

Because the curves obtained have rather general utility and do not seem to be available elsewhere, they are reproduced here. Figure Q-1 is a plot of (D41) for several values of n. Figures Q-2 through Q-9 are plots of the integral of (D38) for each of the values of n shown in Figure Q-1.

Figure Q-10 shows for various values of n, the probability of missing the signal as a function of the *received* signal-to-noise ratio, when the threshold is set to give a false alarm probability of 10^{-12}. These curves are derived from the previous graphs by determining the threshold for each value of n from Figure Q-1, drawing these thresholds on the corresponding Figures Q-2 through Q-9 and plotting the probability of missing the signal at the selected threshold versus the input signal to-noise ratio.

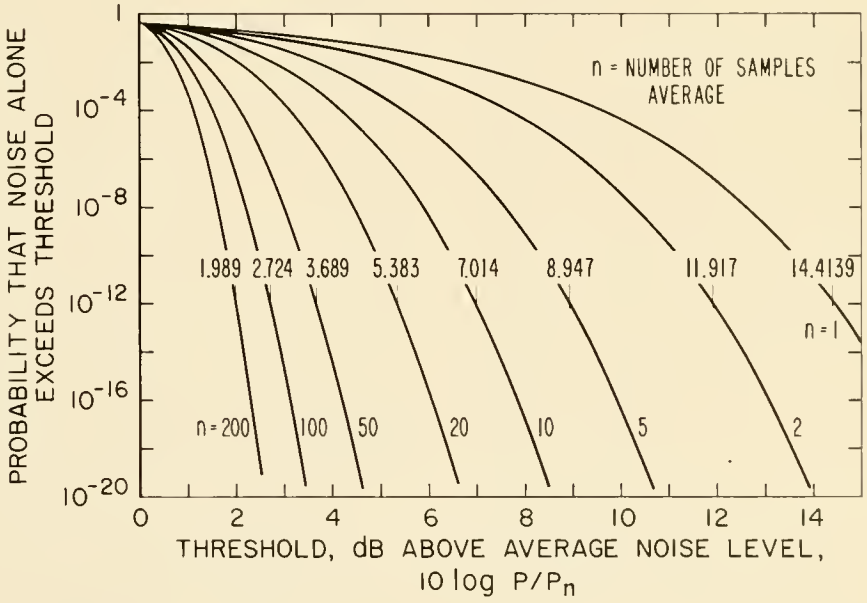


Figure Q-1. *False alarm probabilities.*

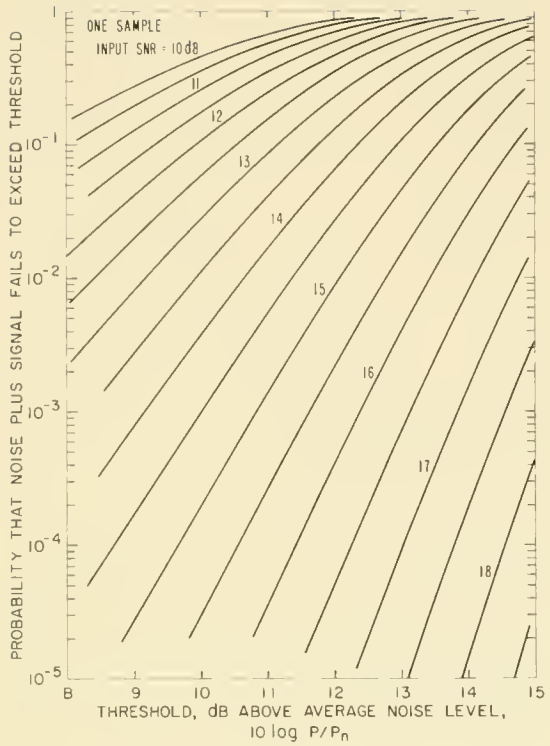


Figure Q-2. *One-sample average of signal plus noise.*

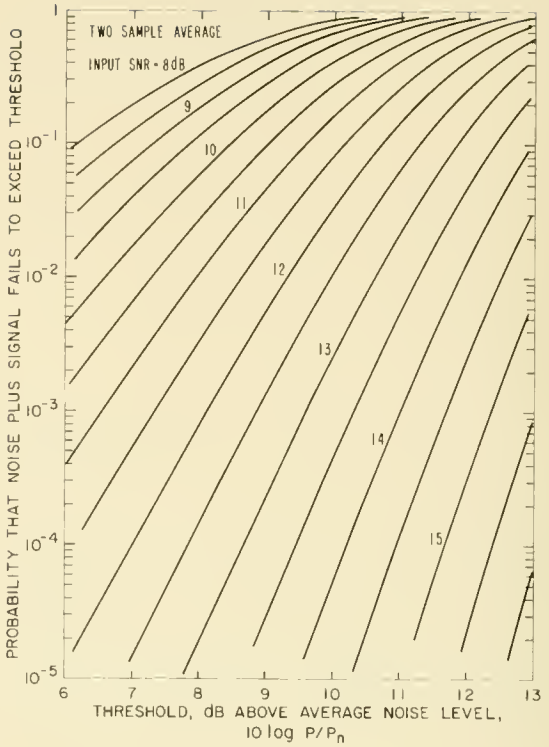


Figure Q-3. *Two-sample average of signal plus noise.*

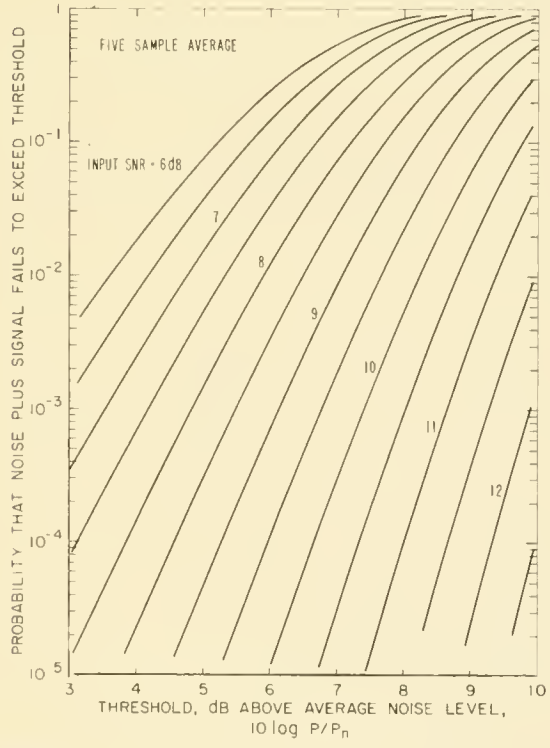


Figure Q-4. *Five-sample average of signal plus noise.*

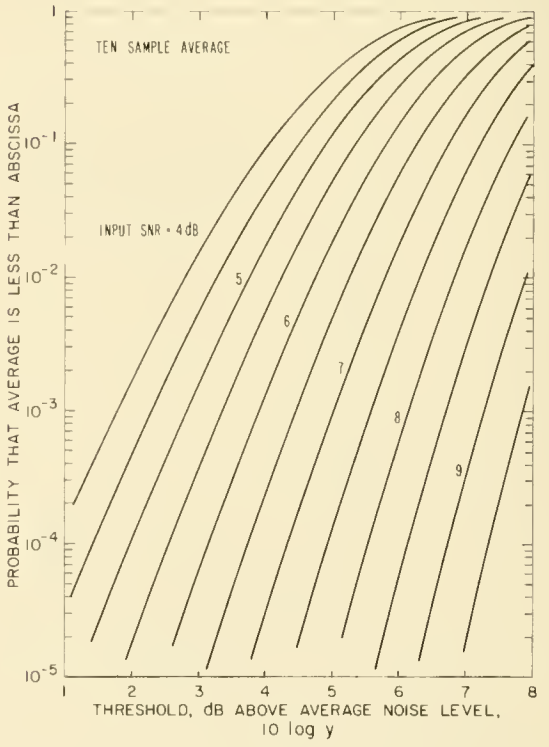


Figure Q-5. *Ten-sample average of signal plus noise.*

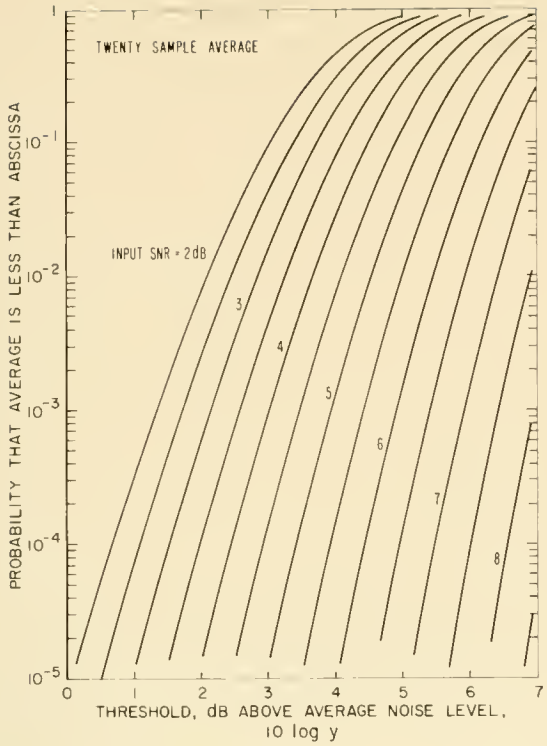


Figure Q-6. *Twenty-sample average of signal plus noise.*

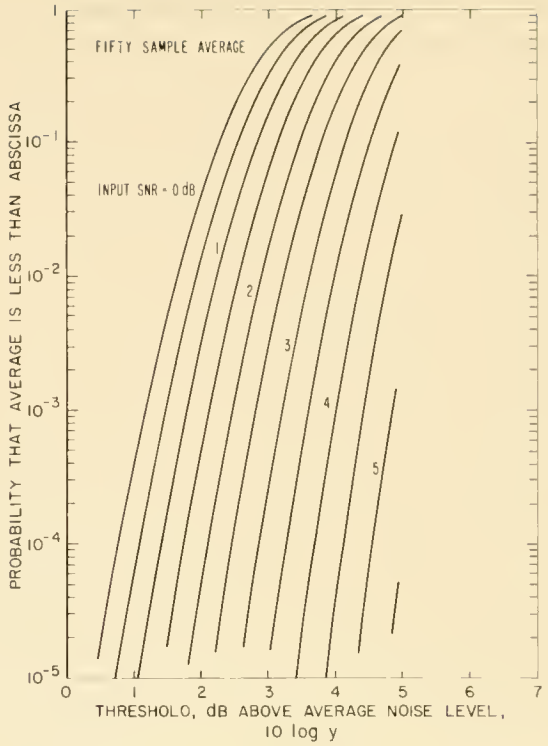


Figure Q-7. *Fifty-sample average of signal plus noise.*

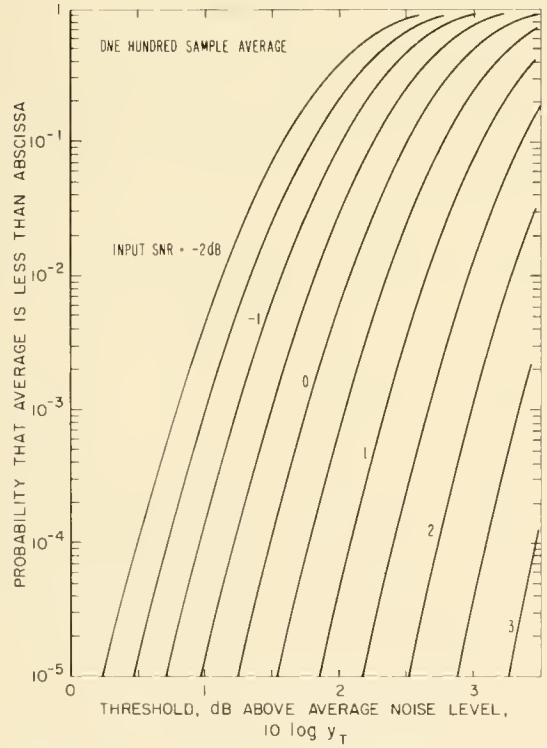


Figure Q-8. *One-hundred sample average of signal plus noise.*

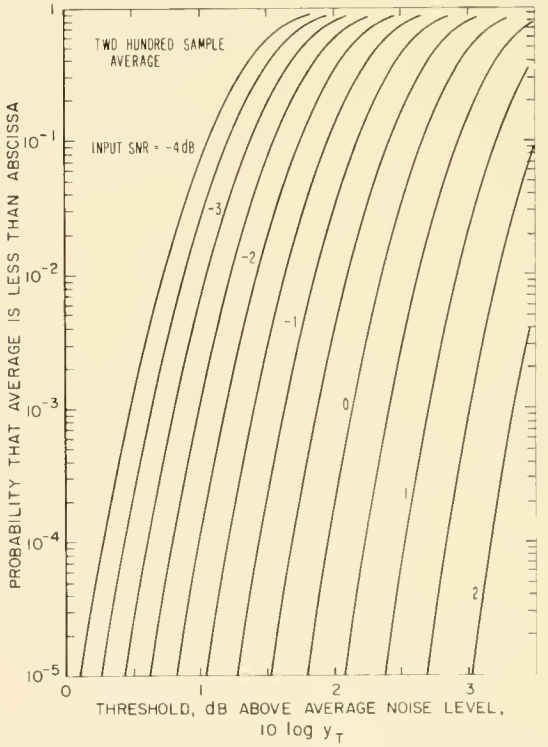


Figure Q-9. *Two-hundred sample average of signal plus noise.*

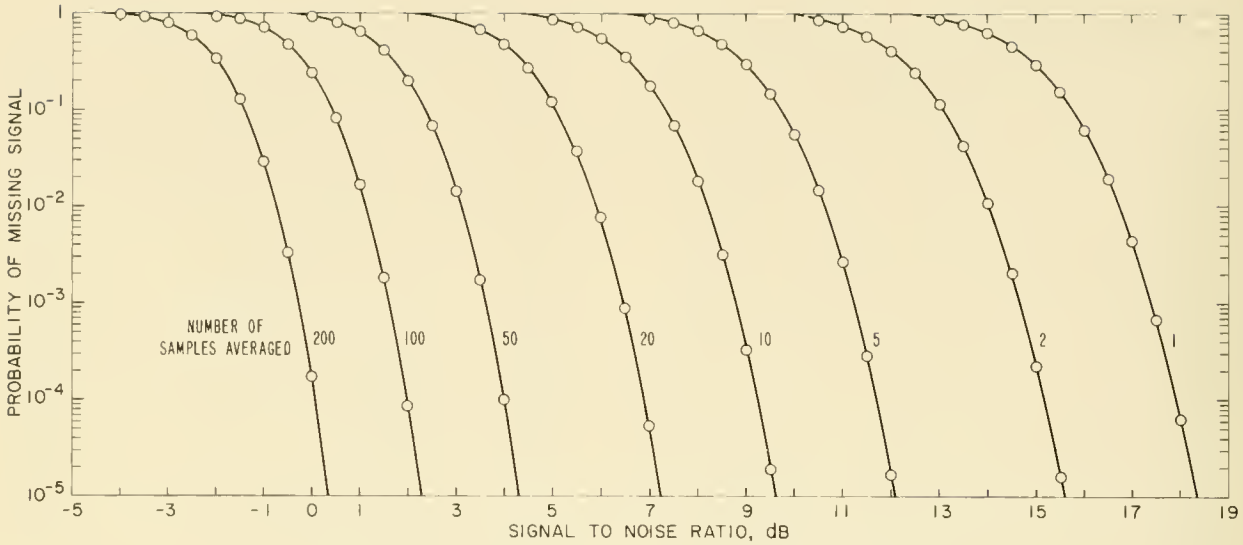


Figure Q-10. *Probability of missing signal (false alarm probability = 10^{-12})*

APPENDIX R

RADIO VISIBILITY OF NORMAL STARS WITH CYCLOPS

Broadband radiometric techniques permit the measurement of very small increases ΔT in the system noise temperature T. If the RF bandwidth is B and the output is integrated for τ sec the rms fluctuation in the readings so obtained will be given by

$$\sigma = \frac{T}{\sqrt{B\tau}} \qquad \text{(R1)}$$

The conventional limit of detectability is taken as $\Delta T = \sigma$. This corresponds to setting $S/N = 1$ in equation (D15). From equation (R1) and equation (22), Chapter 5, we find the range limit to be

$$R = \frac{\pi d_* d_r}{4\lambda}\left(\frac{T_*}{T}\right)^{1/2}(B\tau)^{1/4} \qquad \text{(R2)}$$

where:

d_r = receiving antenna aperture

d_* = diameter of star

T_* = effective temperature of star

λ = radio wavelength

If we take

$d_r = 10^4$ m

$T = 16 + T_{sky}$

$B = 2\times10^8$ Hz (both polarizations used)

$\tau = 60$ sec

then

$$R = 2.75\times10^{-10}\ \frac{d_*}{\lambda}\left(\frac{T_*}{16 + T_{sky}}\right)^{1/2} \text{ light-years} \qquad \text{(R3)}$$

RADIO RANGE OF THE SUN

Substituting $d_\odot = 1.392\times10^9$ meters for d_* in equation (R3) gives

$$R_\odot = \frac{0.3825}{\lambda}\left(\frac{T_\odot}{16 + T_{sky}}\right)^{1/2} \qquad \text{(R4)}$$

for the range at which the Sun could be detected. Using the values of $T_\odot(\lambda)$ and $T_{sky}(\lambda)$ given in Chapter 5, we obtain the ranges in light-years given in Table R-1. These data have been plotted in Figures R-1 and R-2 using antenna aperture as the abscissa.

TABLE R-1

λ(cm)	$R_\odot$ (quiet)	$R_\odot$ (active)
1	380	410
3.16	310	595
10	190	1047
31.6	107	1630
100	29	1170

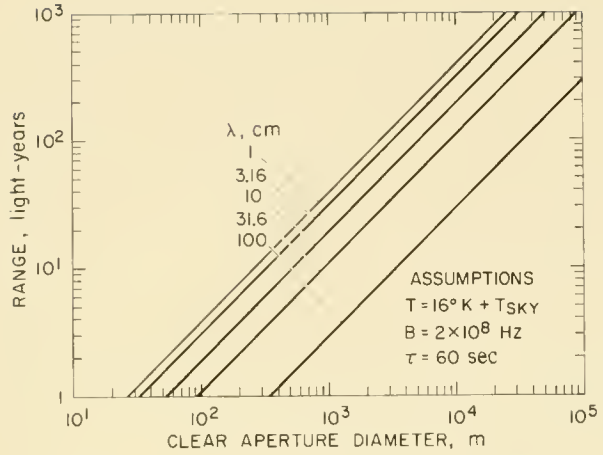


Figure R-1. *Radio visibility of the quiet sun.*

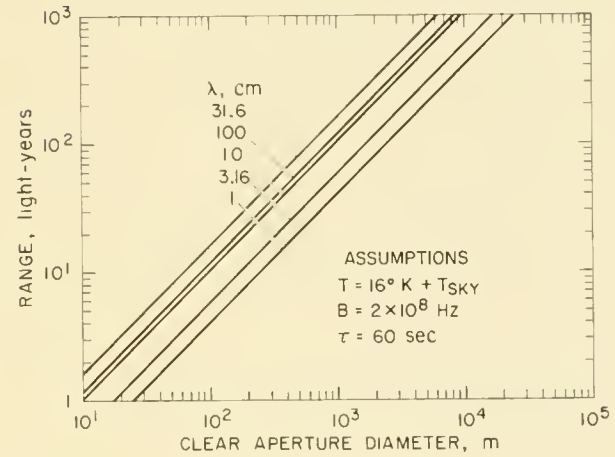


Figure R-2. *Radio visibility of the active sun.*

VISIBILITY OF OTHER STARS

We can get a crude estimate of the number of stars detectable by Cyclops by assuming that a star will be detectable if its apparent magnitude is equal to that of the Sun at the range limit. The apparent magnitude of the Sun at 10 pcs is about 5; at 100 pcs it would be a magnitude 10 star. From Table R-1 we see that for a 10-km array at $\lambda = 3$ cm, the range limit for the Sun is in fact about 100 pcs (326 light-years). If all magnitude 10 stars were detectable, the number would be about 350,000. For a 3-km array the number would drop to about 10,000 stars.

The above numbers are pessimistic for two reasons. First, as stellar temperature drops, the radiation falls faster in the visible than at longer wavelengths because the quantum cutoff shifts. The majority of stars are cooler than the Sun, and thus would be relatively more detectable at radio than at optical wavelengths. Second, a large number of the stars will be giants and supergiants at great distances in the galactic plane, where interstellar absorption dims their optical brightness but not their radio brightness. Thus, it would not be at all surprising to be able to detect over a million stars with a 10-km aperture or 30,000 with a 3-km aperture Cyclops system, with 1 min of integration time.

Most of the detectable stars will be giants and supergiants. Although these are in fact less numerous than main sequence stars, their large size makes them visible to 10 to 100 times as great a range. If we *assume* that the radiation of main sequence stars is enhanced over their black-body radiation in the same ratio as for the quiet sun, their range limit will be given by

$$R_* = R_\odot \left(\frac{T_*}{T_\odot}\right)^{1/2} \frac{d_*}{d_\odot} \qquad \text{(R5)}$$

Using the data for T_* and d_* given in Allen for main sequence stars of various spectral classes, we find the range and ratios given in Table R-2.

TABLE R-2

Type	$T_*/T_\odot$	$d_*/d_\odot$	$R_*/R_\odot$	$(R_*/R_\odot)^3$
O5	5.83	17.78	42.9	79,100
B0	3.50	7.58	14.18	2,850
B5	2.25	3.98	5.97	213
A0	1.62	2.63	3.35	37.5
A5	1.35	1.78	2.07	8.85
F0	1.20	1.35	1.48	3.23
F5	1.08	1.20	1.25	1.94
G0	1.02	1.05	1.06	1.16
G5	0.90	0.93	0.88	0.687
K0	0.78	0.85	0.75	0.423
K5	0.667	0.74	0.604	0.220
M0	0.55	0.63	0.467	0.102
M5	0.433	0.316	0.208	0.009

Figure R-3 was plotted using the value of $R_\odot = 190$ light-years given in Table R-1 for $\lambda = 10$ cm and the values of $R_*/R_\odot$ in Table R-2.

The volume of space over which stars will be detectable is

$$V = \frac{4\pi}{3} R_\odot{}^3 \left(\frac{R_*}{R_\odot}\right)^3 \qquad \text{(R6)}$$

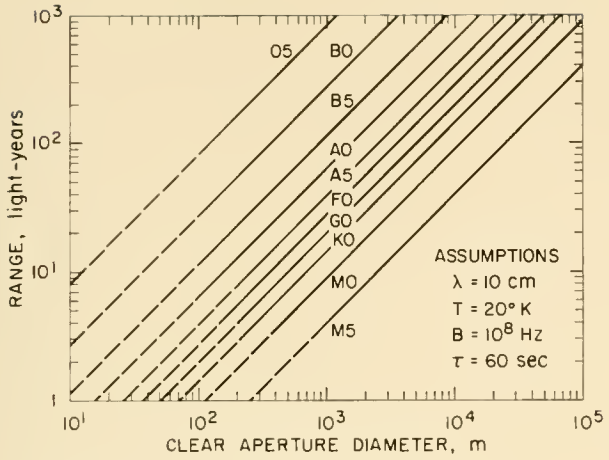


Figure R-3. *Radio visibility of normal stars.*

Assuming a 3-km array and expressing distances in parsecs we find

$$V = 100{,}000\,(R_*/R_\odot)^3\ pc^3 \text{ for } \lambda = 3.16 \text{ cm}$$

$$V = 23{,}500\,(R_*/R_\odot)^3\ pc^3 \text{ for } \lambda = 10.00 \text{ cm}$$

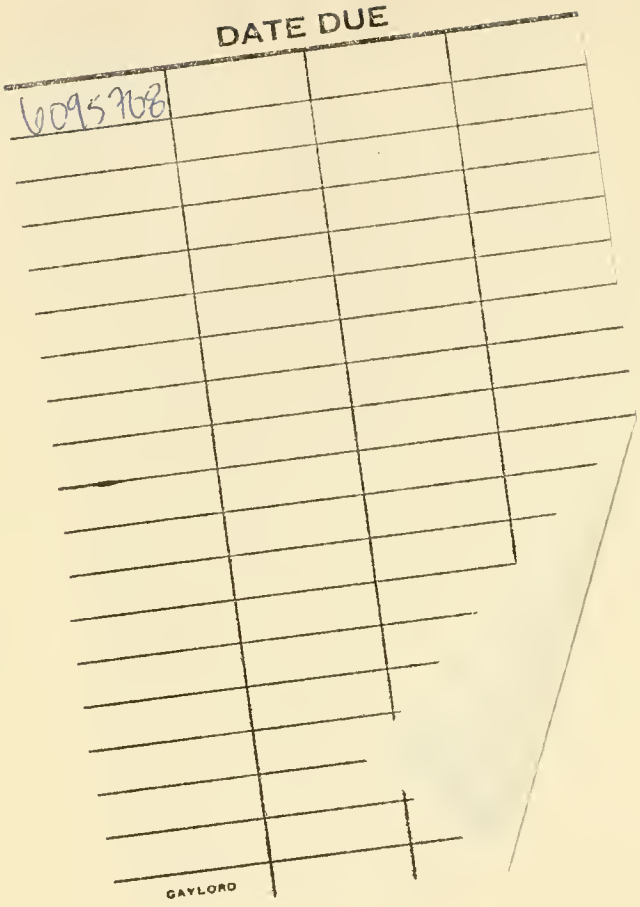
If we now multiply the densities for each spectral class by the volume over which the middle of each spectral class is detectable, we obtain a rough estimate of the number of each spectral type that can be detected.

TABLE R-3

		λ = 3.16 cm		λ = 10 cm	
Type	n/pc^3	V	N	V	N
O	2.5×10^{-8}	(6.5×10^{8})*	16	(2.44×10^{8})*	6
B	1×10^{-4}	2×10^{7}	2000	5×10^{6}	500
A	5×10^{-4}	8.9×10^{5}	445	2.1×10^{5}	105
F	2.5×10^{-3}	2×10^{5}	500	4.6×10^{4}	115
G	6.3×10^{-3}	6.8×10^{4}	428	1.6×10^{4}	101
K	1×10^{-2}	2.2×10^{3}	220	5.2×10^{3}	52
M	5×10^{-2}	9×10^{2}	45	2.1×10^{2}	11
Total			3654		890

*(corrected for galactic disk thickness)

Thus, depending on the array diameter and operating wavelength, we should expect to be able to see from several hundred to several thousand main sequence stars in their normal quiet phases. If the flare activity of these stars were comparable to that of the Sun, about 10 times as many could be seen at 3.16 cm and about 100 times as many at 10 cm when in their active state. We should therefore be able to monitor the coronal and flare activity of several thousand main sequence stars.


DATE DUE
6095708
GAYLORD

qQB 54 .055 1973

Project Cyclops