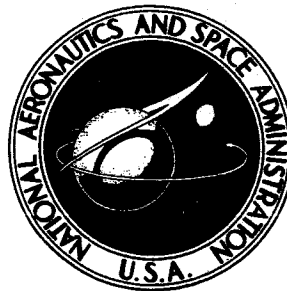


**NASA CONTRACTOR
REPORT**



N71-19643 to

N71-19650

NASA CR-1688

NASA CR-1688

**PARAMETRIC ANALYSIS OF
MICROWAVE AND LASER SYSTEMS
FOR COMMUNICATION AND TRACKING**

**Volume III - Reference Data for Advanced Space
Communication and Tracking Systems**

Prepared by

HUGHES AIRCRAFT COMPANY

Culver City, Calif. 90230

for Goddard Space Flight Center

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • FEBRUARY 1971

1. Report No. NASA GR-1688		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle Parametric Analysis of Microwave and Laser Systems for Communication and Tracking; Volume III - Reference Data for Advanced Space Communication and Tracking Systems				5. Report Date February 1971	
				6. Performing Organization Code	
7. Author(s)				8. Performing Organization Report No.	
9. Performing Organization Name and Address Hughes Aircraft Company Culver City, California				10. Work Unit No.	
				11. Contract or Grant No. NAS 5-9637	
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D. C. 20546				13. Type of Report and Period Covered Contractor Report	
				14. Sponsoring Agency Code	
15. Supplementary Notes Prepared in cooperation with all the available experts at the Hughes Aircraft Company and edited jointly by L. S. Stokes the Program Manager, K. L. Brinkman, the Associate Program Manager, and Dr. F. Kalil, the NASA-GSFC Technical Monitor, with L. S. Stokes being the primary contributing editor.					
16. Abstract Present and future space programs are requiring progressively higher communication rates. For instance, the Earth Resources Technology Satellite-A requires about 70 MHz total bandwidth in its S-Band downlink spectra, and it appears likely that future earth observation satellites will require more bandwidth because of the larger number of sensors and higher sensor resolutions. On the other hand, the frequency bands allocated via international agreements for space use are limited, and hence, the r-f spectrum is becoming crowded. However, the advent of the C-W laser systems offered a "new" and wide electromagnetic spectrum for use in space telecommunications. Although the laser systems offered this "new" capability, their technological development was also new. Therefore, this study was undertaken to make a comparative analysis of microwave and laser space telecommunication systems. A fundamental objective of the study was to provide the mission planner and designer with reference data (weight, volume, reliability, and costs), supplementary material, and a trade-off methodology for selecting the system (microwave or optical) which best suits his requirements. This report is the final report of that study. Because of the large amount of material, the report is presented in four volumes. This volume, Volume III is the "Reference Data for Advanced Space Communication and Tracking Systems." It contains theory and state-of-the-art performance for the following technology areas for both microwave and laser frequencies: transmitting power sources, modulators, detectors, transmitting and receiving apertures, acquisition and tracking, and prime power and heat rejection systems.					
17. Key Words (Selected by Author(s)) μ-wave and Laser Power Sources Optical Detectors and Modulators R-F Antennas Optical Apertures Acquisition and Tracking Prime Power Sources Heat Ejection				18. Distribution Statement Unclassified - Unlimited	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 589	
				22. Price* \$10.00	

PARAMETRIC ANALYSIS OF MICROWAVE AND LASER SYSTEMS
FOR COMMUNICATION AND TRACKING

VOLUME I	SUMMARY	.
VOLUME II	SYSTEM SELECTION	.
VOLUME III	REFERENCE DATA FOR ADVANCED SPACE COMMUNICATION AND TRACKING SYSTEMS	
VOLUME IV	OPERATIONAL ENVIRONMENT AND SYSTEM IMPLEMENTATION	

REFERENCE DATA FOR ADVANCED SPACE COMMUNICATION
AND TRACKING SYSTEMS

BRIEF INDEX OF VOLUME III

PART 1 — TRANSMITTING POWER SOURCES .

Section	Page
Radio Frequency Sources	2
Radio Frequency Tube Sources	10
Radio Frequency Solid State Sources	28
Optical Frequency Source Characteristics	46
Laser Mode Coupling and Frequency Stabilization	78
Laser Oscillators and Amplifiers	100

PART 2 — MODULATORS

Radio Frequency Modulators	120
OPTICAL MODULATORS	128
Electro-Optic Modulation	134
Elasto-Optic Modulation	152
Internal Modulation	168
Modulator Performance	176

PART 3 — DETECTORS

Radio Frequency Detectors	190
Optical Frequency Detectors	200

BRIEF INDEX OF VOLUME III (Continued)

PART 4 – TRANSMITTING AND RECEIVING APERTURES

Section	Page
Radio Frequency Antennas	234
Optical Frequency Apertures	282
Optical Apertures – Optical Configurations	296
Optical Frequency Apertures – Weight and Cost Relationships . .	318

PART 5 – ACQUISITION AND TRACKING

GENERAL ACQUISITION AND TRACKING SYSTEM CONSIDERATIONS

Mission Associated Considerations	334
Receiver Location Considerations	346

ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS

The Tracking Subsystem – Introduction	358
Acquisition	380
Detection Theory	386
Angle Noise Error in Optical Tracking Systems	396

COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS

Attitude and Tracking Sensors	424
Attitude Control Techniques	440
Passive Attitude Control Techniques	444

BRIEF INDEX OF VOLUME III (Continued)

Section	Page
Active Attitude Control Devices	452
Burden Relationships	466
PART 6 - PRIME POWER SYSTEMS	
Solar Power Systems	482
Nuclear Power Systems	490
Chemical Power Systems	508
Power Summary	516
PART 7 - HEAT EJECTION SYSTEMS	
Heat Ejection Elements	532
Heat Ejection Elements - Radiators	540
Weight and Cost Burdens	550

REFERENCE DATA FOR ADVANCED SPACE COMMUNICATION AND TRACKING SYSTEM

DETAILED INDEX OF VOLUME III

PART 1 — TRANSMITTING POWER SOURCES

	Page
Radio Frequency Sources	2
Introduction — Types of Transmitting Sources	2
Summary of Radio Frequency Transmitting Sources	4
Summary of Gas Lasers as Transmitting Sources	6
Radio Frequency Tube Sources	10
Fundamentals of UHF Sources	10
Fundamentals of Microwave Sources	12
Hughes 394H TWT Performance	16
Fundamentals of Millimeter and Submillimeter Sources	18
Weighting Factors	20
Performance of Vacuum Tube Sources	22
Radio Frequency Solid State Sources	28
Introduction to Solid State Sources	28
Theory of Operation for Impatt Oscillators	30
Theory of Operation for Gunn Oscillators	32
LSA (Limited Space Charge Accumulation) Power Sources	36
State of the Art for LSA, Impatt, and Gunn Microwave Sources	38
Radio Frequency Burden Relationships	42
Optical Frequency Source Characteristics	46
Introduction	46
Laser Operating Fundamentals	48
The Argon Laser, Excitation Process	50
Laser Amplifier Gain	54
Laser Power Output and Efficiency	58
CO ₂ Laser Excitation Process	64
CO ₂ Laser Frequency Spectrum	70
CO ₂ Laser Scaling Laws	72
Laser Mode Coupling and Frequency Stabilization	78
Laser Mode Coupling	78

DETAILED INDEX OF VOLUME III (Continued)

	Page
Laser Frequency Stabilization Considerations	84
Laser Stabilization by Mechanical and Thermal Methods	88
Laser Frequency Stabilization Using Feedback Systems	90
Optical AFC Systems Using Passive Cavity Discriminants.	92
Results of AFC for Lasers	98
Laser Oscillators and Amplifiers	100
CW Laser Performance	100
Pulsed Laser Oscillators	104
Laser Amplifiers	108
Gas Laser Selection for Space Communications	110
Laser Burden Values	114
PART 2 – MODULATORS	
Radio Frequency Modulators	120
Modulation Methods	120
Radio Frequency Modulations Implementation	124
Digital and Compound Modulation Implementation	126
Optical Modulators	
Introduction	128
Summary of Optical Modulators	130
Electro-Optic Modulation	134
Theory of Electro-Optic Modulation	134
Bias Point and Driver Level Considerations	136
Design of Optical Intensity Modulators	138
Design of Electro-Optic Frequency Translators	142
Bandwidth Limiting Factors in Lumped-Element Modulators	146
Traveling Wave Electro-Optic Modulators	148
Elasto-Optic Modulation	152
Theory of the Elasto-Optic Effect	152

DETAILED INDEX OF VOLUME III (Continued)

	Page
Theory of Ultrasonic Diffraction of Light	156
Acoustic Modulation Techniques	158
Properties of Ultrasonic Modulators	160
Summary and Tabulation of Elasto-Optic Modulators . .	164
Internal Modulation	168
Intensity Modulation	168
Frequency Modulation and Translation	172
Modulator Performance	176
Modulator Burden Considerations	176
Modulator Burden Relationships	178
Nomenclature Summary	180
PART 3 – DETECTORS	
Introduction	184
Summary of Detection Methods	186
Radio Frequency Detectors	190
Sensitivity of RF Detectors	190
Radio Frequency Amplifiers	194
Optical Frequency Detectors	200
Introduction	200
Characterization of Optical Detectors	202
Theory of Photomultiplier Detectors	204
Solid State Detectors Operating Concepts	210
Detection Limits of Solid State Detectors	214
Performance of Photoemissive Detectors	222
Detectors for 10.6 Microns	224
Detector Performance Summary	228
Burdens for Optical Frequency Detectors	230
PART 4 – TRANSMITTING AND RECEIVING APERTURES	
Radio Frequency Antennas	234
Introduction	234
Summary of RF Transmitting and Receiving Apertures	236

DETAILED INDEX OF VOLUME III (Continued)

	Page
Antenna Gain and Aperture Relationships	240
Gain Degradation Due to Predictable Systematic (Non Random) Phase Errors	244
Gain Degradation Due to Random Errors	248
Effects of Random Errors on Antenna Parameter Values	252
Antennas for Space Communication – Design Considerations	258
Antennas for Space Communication – Antenna Types . .	260
Antennas for Space Communication – Deployable Paraboloids	262
Antennas for Space Communication – Weight Burdens for Paraboloids	264
Special Purpose Multibeam and Self Steering Antennas	266
High Gain, Self-Steering Antenna System for Satellite-Earth Communications	268
Antennas for Space Communication – Surface Station Cost Burdens	274
Antennas for Space Communication – Spacecraft Cost Burdens	276
Optical Frequency Apertures	282
Introduction	282
Optical Configurations	284
Optical Configurations – Optics with a Large Field of View	286
Optical Configurations – Use of a Tracking Mirror . . .	290
Optical Configurations – Reflectance and Attenuation in Optical Systems	292
Optical Apertures – Optical Configurations	296
Effect of Surface Tolerances	296
Alignment Tolerances	298
Speed of Optical Configuration	310
Thermal Effects on Optical Configurations	312
Low Temperature Coefficient Material, CER-VIT	314

DETAILED INDEX OF VOLUME III (Continued)

	Page
Optical Frequency Apertures — Weight and Cost Relationships	318
Weight and Cost of Beryllium Mirrors	318
Weight and Cost for Fused Silica Mirrors	322
Weight and Cost Burden Relationships	324
PART 5 — ACQUISITION AND TRACKING	
Introduction and Summary	330
GENERAL ACQUISITION AND TRACKING SYSTEM CONSIDERATIONS	
Mission Associated Considerations	334
Signal Propagation Delays	334
Relative Motion Between Transmitter and Receiver . . .	338
Coordinate Reference Frame Error	340
Manned Versus Unmanned Vehicles	342
Receiver Location Considerations	346
Earth Based Receiver Atmospheric Considerations . . .	346
Earth-Satellite Based Receiver and Receiver Site Comparison Summary	350
ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS	
Introduction	353
The Acquisition Subsystem Operational Sequence	354
The Tracking Subsystem	358
The Tracking Subsystem — Introduction	358
DSV Tracking Subsystem — Pointing Error	360
DSV Tracking Subsystem — Description	364
Stabilization Subsystems	368
Earth Station — Pointing Error Budget	372
Signal to Noise Ratios for Star Trackers	374

DETAILED INDEX OF VOLUME III (Continued)

	Page
Acquisition	380
Mean Time to Acquisition	380
Acquisition Probabilities	382
Detection Theory	386
The Probability of Detection and False Alarm (Gaussian Case)	386
Probability of Detection and False Alarm (Poisson Case)	392
Angle Noise Error in Optical Tracking Systems	396
Introduction	396
Monopulse Quadrant Tracking System	400
Angle Noise Analysis of Monopulse Quadrant Tracking System	406
Beam Lobing PPM Tracking System	412
AM Reticle Tracking System	416
FM Reticle Tracking System	420
COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS	
Attitude and Tracking Sensors	424
Sun Sensors	424
Conversion Chart for Angular Measure	428
Star Sensors	430
Star Tracker Detectors	434
Planet Sensors	436
Attitude Control Techniques	440
Introduction	440
Passive Attitude Control Techniques	444
Solar Radiation Pressure	444
Gravity Gradient Forces	446
Magnetic Forces	448

DETAILED INDEX OF VOLUME III (Continued)

	Page
Active Attitude Control Devices	452
Reaction Wheels	452
Momentum Spheres and Fluid Flywheels	454
Momentum Wheels	456
Reaction Jets	458
Burden Relationships	466
Weight Burdens	466
Cost Burdens	468
Power Burdens	472
PART 6 – PRIME POWER SYSTEMS	
Introduction	476
Summary	478
Solar Power Systems	482
Solar Voltaic Systems	482
Solar Cell Degradation in a Space Environment	484
Solar Thermal Systems	486
Nuclear Power Systems	490
Introduction to Nuclear Power Systems	490
Thermoelectric Reactor Power Systems	492
Thermionic Reactor Systems	496
Reactor Dynamic Power Systems – Brayton Cycle	498
Reactor Dynamic Power Systems – Rankine Cycle	500
Radioisotope Thermoelectric Systems	502
Radioisotope Dynamic Systems	506
Chemical Power Systems	508
Fuel Cells	508
Batteries	510
Power Summary	516
Cost, Volume, and Weight	516
Prime Power Burdens	522

DETAILED INDEX OF VOLUME III (Continued)

	Page
PART 7 — HEAT EJECTION SYSTEMS	
Introduction	526
General Heat Ejection Considerations	526
Transmitter Source Characteristics	528
Summary	530
Heat Ejection Elements	532
Types of Heat Ejection Systems	532
Heat Pipe	534
Useful Heat Pipe Properties	536
Heat Ejection Elements — Radiators	540
Radiant Heat Transfer from a Flat Surface	540
Radiator Fin Effectiveness	542
Radiator Area Requirements — Condensing Systems . .	544
Radiator Area Requirements — Non Condensing Systems	546
Weight and Cost Burdens	550
Radiator Weight and Cost Variations	550
Weight and Cost Burden Constants	552

ACKNOWLEDGEMENTS

The material of this volume was prepared by several contributors. The sections and the corresponding contributors are given below.

Transmitting Power Sources — Principle contributors were Dr. D. C. Forster, Dr. W. B. Bridges and Mr. J. Burnsweig.

Modulators — The principle contributor to this part was Dr. J. F. Lotspeich.

Detector — Dr. G. S. Picus was the principle contributor to optical detectors while Dr. D. C. Forster and Mr. B. L. Walsh contributed to the radio detector portion.

Transmitting and Receiving Apertures — Dr. W. H. Kummer, Dr. T. S. Fong, Mr. S. S. Shapiro, and Mr. B. Golvin contributed to the radio aperture portion while Mr. J. Bagby was the principle contributor to the optical portion. Mr. Bagby, formerly with Hughes Aircraft Company is currently with Bell and Howell.

Acquisition and Tracking — Mr. J. W. Callis, Mr. J. R. Sullivan and Dr. E. J. Vourgourakis contributed to this part. Mr. Sullivan, formerly with Hughes Aircraft Company is presently with Systems Associates.

Prime Power Sources and Heat Ejection Systems — These parts were primarily contributed by Mr. J. R. Sullivan, formerly of Hughes Aircraft Company and presently with Systems Associates.

PART 1 – TRANSMITTING POWER SOURCES

Section	Page
Radio Frequency Sources	2
Radio Frequency Tube Sources	10
Radio Frequency Solid State Sources	28
Optical Frequency Source Characteristics	46
Laser Mode Coupling and Frequency Stabilization	78
Laser Oscillators and Amplifiers	100

INTRODUCTION - TYPES OF TRANSMITTING SOURCES

An understanding of the types, characteristics and operating parameters of known rf and optical transmitting power sources is required to design space tracking and communications systems.

In this section, radio frequency (rf) sources and optical frequency sources are described in terms of operating fundamentals and performance characteristics. The discussion is limited to factors useful in predicting the principal electrical and mechanical characteristics of space tracking and communications systems.

The division of the spectrum between rf and optical frequency source types was arbitrarily chosen at approximately 300 microns.

Radio Frequency Sources include both oscillators and amplifiers, for either may be used as the basic transmitter device. The rf discussion is limited to continuous wave devices as cw sources generally are significantly more efficient than non-continuous power sources. A natural classification of rf sources results as a function of frequency because of the fundamental mechanisms involved in the generation of rf energy. For this discussion, the frequency radio range is divided as follows:

VHF/UHF	100 MHz to 1000 MHz
Microwave	1 GHz to 30 GHz
Millimeter	30 GHz to 300 GHz
Submillimeter	300 GHz to 1000 GHz

The Optical Frequency Sources include ultraviolet, visible and infrared laser sources (to 300 microns). The discussion sets forth the general relationships between applicable operating characteristics (optical gain, saturation, noise) and applicable physical and atomic parameters (length, transition probabilities, line width, etc.).

Following the general introduction, a more detailed theory of selected gas laser sources is presented, including considerations unique to each particular laser. While selection of the lasers described was made on the basis of practicality, the individual discussions are generally representative of a class of lasers. For example, the theory of the blue and green transitions in singly-ionized argon will generally apply to all singly-ionized laser transitions.

In addition, an attempt is made to indicate the confidence with which the theories are held.

Transmitting Power Source Discussion Covers Both
Theory and Performance of RF and Optical
Frequency Sources

Radio Frequency Sources	Optical Frequency Sources
<ul style="list-style-type: none"> ● Theory of Radio Frequency Sources <ul style="list-style-type: none"> ● VHF/UHF Sources ● Microwave Sources <ul style="list-style-type: none"> Klystrons Traveling-wave tubes Cross-field Devices ● Millimeter Sources ● Submillimeter Sources ● Performance of Vacuum Tube Sources <ul style="list-style-type: none"> ● UHF Sources ● Microwave Sources ● Millimeter Sources ● Solid State Microwave Sources ● Weighting Factors 	<ul style="list-style-type: none"> General Theory of Laser Sources Argon Ion Laser CO₂ Laser Laser Mode-Coupling Laser Stabilization Laser Oscillators Laser Amplifiers Evaluation of Gas Laser Sources

Transmitting Power Sources
Radio Frequency Sources

SUMMARY OF RADIO FREQUENCY TRANSMITTING SOURCES

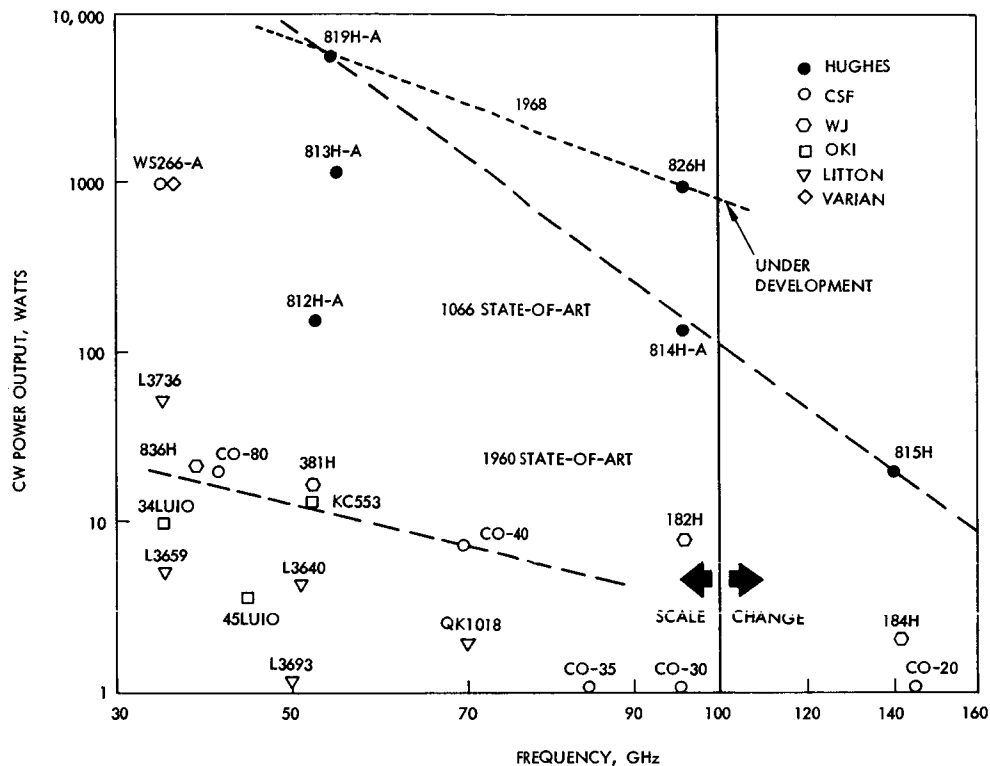
Frequency bands are indicated for radio frequency power sources.

Radio frequency sources are described in terms of operating theory and performance characteristics.

Radio frequency sources include both oscillators and amplifiers, for either may be used as the basic transmitter device. The rf sources are limited to continuous wave devices as cw sources generally are significantly more efficient than non-continuous power sources. A natural classification of rf sources results as a function of frequency because of the fundamental mechanisms involved in the generation of rf energy. The frequency range is divided as follows:

UHF	100 MHz to 1000 MHz
Microwave	1 GHz to 30 GHz
Millimeter	30 GHz to 300 GHz
Submillimeter	300 GHz to 1000 GHz

The figure indicates, in summary form, available power levels available from rf sources.



Power Characteristics of Available High Power CW Sources

Most significant is 35 percent efficiency achieved in the 813H and marked increase in available power which has occurred since 1960. The letter -A stands for amplifier and the letter -O signifies oscillator.

SUMMARY OF GAS LASERS AS TRANSMITTING SOURCES

A CO₂ transmitter for a spaceborne communication link and an Argon laser for an up link beacon appear to be the best choice for laser space communication.

The Table summarizes the characteristics of six wavelengths produced by gas lasers. Hundreds of other wavelengths are available, but these six have been selected as representative of each type (ion, molecular, and neutral gas). The reported output power, length and input power are given for the lasers selected.

Notes

1. This is a Hughes airborne quartz laser with a 46 cm bore length ~1 meter overall package length. It requires a magnetic field of ~1000 gauss, which implies a heavy structure and possibly more power.
2. This laser was reported by Raytheon in Electronic News; it is a quartz tube. The power out is 18 watts, provided the beam in the cavity was chopped to prevent damage to the mirrors.
3. This was produced under carefully controlled conditions at Bell Labs.
4. This is a commercially available Spectra-Physics model 125. 50 mw is guaranteed, but selected tubes produce 100 mw.
5. This is the Spectra-Physics model 125. It may be possible to double the power in a tube this size, but drastic improvements are quite unlikely at this wavelength.
6. This is an Hughes Research Laboratories (HRL) Laboratory-type tube. The output may be doubled, but more power than this is doubtful in a tube this size.
7. This is a TRG Laboratory-type tube and represents approximately two years of effort in developing a high power Xe laser. It is probably close to the ultimate for a tube this size.
8. This is an HRL Laboratory-type tube using flowing CO₂-N₂, mirrors were not optimized; more output power can be expected from this same tube (~20 watts). Seven watts were obtained with the tube sealed.
9. This is the Bell Telephone Laboratories work (C. K. N. Patel, Appl. Phys. Lett., 1 July 1965). A flowing gas system was used with a mixture of CO₂, N₂, O₂, H₂O.
10. This is a BTL result with a tube 4 inches in diameter and 12 feet long. Helium was used. It is hard to estimate how much power will eventually be obtained from a tube of this size. (C. K. N. Patel).

11. This is a small, non flow tube, with external mirrors; suitable for spacecraft. This work is due to T. J. Bridges and is rather preliminary. An account of similar tubes appears in the 1 November 1965 Appl. Phys. Letters.

In comparing the various lasers listed in the table, the suitability of the output signal for the communications task at hand must be kept in mind. All of the lasers listed can be made to operate in the lowest order spatial mode (TEM_{00}) alone with more or less difficulty. The task is easier at the shorter wavelengths where the laser output is visible and the characteristic beam size is small (proportional to the square root of the product of wavelength and a cavity parameter related to mirror radius). Mode selection at infrared wavelengths may be done with an image-converter or by using a heterodyne detector. Production of a single-frequency output is still quite difficult because of the longitudinal mode structure of the long Fabry-Perot cavities used. Only the 10.6μ CO_2 and 3.5μ Xenon lines are narrow enough to produce reasonable output by keeping the Fabry-Perot resonator short enough so that only one longitudinal mode oscillates. This is done to the narrow doppler-broadened line widths of these two transitions (50 MHz for 10.6μ CO_2 and 120 MHz for 3.5μ Xe). Even these two transitions will require further mode selection techniques if longer, higher power tubes are considered. Because of the broad doppler line widths of the Ar and He-Ne lasers, single-frequency operation through the use of a sufficiently short Fabry-Perot resonator entails a drastic loss in output power. Techniques involving 3 mirror resonators allow the use of longer tubes at the expense of added complexity both mechanical and electronic (servo-controlled mirror positioning), but still sacrifice output power because the entire line is not used. The most promising technique developed to date is that of intracavity mode locking¹ with a subsequent coherent recombination² or selective output coupling³. This technique has been demonstrated in the laboratory, but practical power levels at a single frequency are yet to be obtained. In any case the additional complexity will contribute to the weight, length and inefficiency of the laser, although perhaps not to a significant extent.

It appears that, at present, the best laser for optical space communications at present would be a small, efficient, light weight 10.6μ CO_2 laser in the spacecraft with coherent detection (superheterodyne) on the ground, employing a cooled Hg:Ge detector. The up-link would be best handled by a high-power multimode argon laser on the ground, employing pulse amplitude or pulse polarization modulation, and a simple ruggedized photomultiplier video receiver in the spacecraft. These conclusions are, of course, subject to revision as the state of the laser (and detector) art progresses.

¹Harris, S. E., and McDuff, O. P., Appl. Phys. Letts., 5, pp. 205-206, November 15, 1964.

²Massey, G. A., Ashman, M. K., and Taig, R., Appl. Phys. Letts., 6, p. 10, 1965.

³Hanes, S. E., and McMurtry, B. J., (to be published).

Transmitting Power Sources
Radio Frequency Sources

SUMMARY OF GAS LASERS AS TRANSMITTING SOURCES

Gas Laser Performance

Gas	Wavelength, microns	Manu- facturer	Output Power, watts	Length, meters	Input Power, watts	Efficiency	Note
Ar II	0.5	HRL	4.0	0.46	4,000	1×10^{-3}	1
Ar II	0.5	RAY	8.0	1.6	20,000	4×10^{-4}	2
He-Ne	0.63	BTL	1.0	5	500	2×10^{-3}	3
He-Ne	0.63	S-P	0.1	1.7	~200	5×10^{-4}	4
He-Ne	1.15	S-P	0.03	1.7	~200	1.5×10^{-4}	5
He-Ne	3.39	HRL	0.01	1.7	80	1.8×10^{-4}	6
He-Xe	3.51	TRG	0.08	2.0	~200	4×10^{-4}	7
CO ₂	10.6	HRL	10	2.0	150	6.7×10^{-2}	8
	10.6	BRL	12	2.0	—	3×10^{-2}	9
	10.6	BTL	130	4.0	~1,000	$\sim 1.3 \times 10^{-1}$	10
	10.6	BTL	0.1	0.5	30	3.3×10^{-3}	11

TRANSMITTING POWER SOURCES

Radio Frequency Tube Sources

	Page
Fundamentals of UHF Sources	10
Fundamentals of Microwave Sources	12
Hughes 394H TWT Performance	16
Fundamentals of Millimeter and Submillimeter Sources	18
Weighting Factors	20
Performance of Vacuum Tube Sources	22

Transmitting Power Sources
Radio Frequency Tube Sources

FUNDAMENTALS OF UHF SOURCES

Negative grid tubes dominate the rf power sources in the VHF/UHF region (100 MHz to 1000 MHz).

Conventional tubes such as triodes and tetrodes are used in external resonant circuits for Class A or B operation. At the higher frequencies, these resonant circuits take the form of coaxial lines. Electron transit time effects limit the extension of these techniques to higher frequencies, and, at these higher frequencies, power limitations occur due to the thermal capability of the collector and the envelope seals.

The Table gives relative advantages and disadvantages for this type of source.

Relative Advantages and Disadvantages of Negative Grid Tube Characteristics Summary

Advantages	Disadvantages
<ol style="list-style-type: none"> 1. Good phase stability and tracking characteristics, because of short transit time. Typical sensitivity to change in voltages is: Screen - 1 degree per 1 percent E_{sg} at 1 GHz Plate - 0.5 degree per 1 percent E_p Phase variations due to drive level and filament voltage changes are negligible. Because the gridded tube has short electrical length, its phase tracking characteristics are determined primarily by the associated circuitry. Mass-produced double-tuned circuits show typical phase track deviations of 5 degrees over the usable bandwidth. A suitable secondary coupling can be chosen to give linear phase-frequency characteristics. 2. High Efficiency. At 500 MHz, typical plate efficiencies approach 70 percent for class C operation. Overall efficiencies of 50 percent, including filaments, are common. 3. Economy. Most coaxial tubes have a simple structure and are easily mass-produced. It is usually possible to select a tube already in mass production. This affords additional economy. 4. Inherent Filtering Action. For a double-tuned reentrant cavity, in which a tetrode normally operates, the roll-off is at a rate of 12 db per bandwidth octave. For a 2 percent bandwidth device at 500 MHz, typical second harmonic suppression is more than 80 db for class B operation. 5. Much more experience exists in design of vacuum tubes than in other types. 6. No focusing magnets or field are required, therefore weight and volume are reduced. 7. Tetrodes can operate from a dc supply on the plate. Because the control or screen grid can act as a switching element, efficiency is greater and noise is lower. 8. Tetrodes are constant-current generators and tolerate mismatches better than devices depending upon traveling- or standing-wave operation. 9. They are relatively insensitive to temperature up to the rating of the envelope seals. 	<ol style="list-style-type: none"> 1. Limited gain-bandwidth product. 1000 to 2500 MHz for class A operation and 500 to 1000 MHz for class B are typical. These values hold for operations to about 500 MHz; above this frequency, the product drops off rapidly. 2. Operating frequency is limited to less than 1000 MHz. Above this, transit time is long enough to create current wave-form distortion in the plate current. The result is decreased gain-bandwidth product and lower efficiency. 3. Higher-power tubes are disproportionately expensive. This is attributed to the fact that higher-power tubes have not been manufactured in such large quantities as the smaller ones. 4. Reliability and life expectancy. The cathode of gridded tubes must deliver a greater current density than that required by other devices. In addition, maximum r-f current must be instantaneously available for the cathode, thus requiring that the cathode operate at a comparatively high temperature. As a result, normal cathode life expectancy is generally 2000 to 5000 hours, though some tubes are rated at 5000 to 15,000 hours. However, arcing and other failure mechanisms reduce the MTBF of most types to about 3000 hours. 5. The tetrode is a filamentary device requiring a variety of electrode voltages. Distribution of these voltages within a complex system can be troublesome. 6. Arcing problems are generally greater than in other devices due to close spacing of elements. 7. High output capacitance limits bandwidth.

FUNDAMENTALS OF MICROWAVE SOURCES

At microwave frequencies, where electron transit time effects are significant (1 GHz to 30 GHz), some form of velocity modulation tube is employed to overcome these limitations.

Basically, the velocity modulation technique provides a means for a bunch of electrons containing the input dc power to remain in approximate time or space synchronism with a component of the ac wave so that power can be transferred continuously. Transit time effects are no longer of consequence because an electron continues to see the same phase of the ac field in spite of the axial motion of the bunch of electrons. The electrons give up either kinetic energy in the case of a klystron or traveling-wave tube, or both kinetic and potential energy in a cross-field tube. The interaction region normally takes the form of a slow-wave circuit in the vicinity of the electron bunches such that the electrons see the rf fields associated with the wave propagating structure. The various types of these tubes are classified in Table A and described below.

Klystrons

The klystron amplifier is a well developed, reliable, and in many cases a long-life device. If conservatively designed, it should be capable of meeting space requirements. The requirement for a high voltage and a high power modulation technique reduces the overall efficiency of a transmitter chain using a klystron as the final power amplifier. Table B summarizes the significant advantages and disadvantages of klystrons.

TWT's

The traveling wave tube is inherently a high average power amplifier so there exists no problem in achieving the parameters necessary in a space system. Furthermore, the TWT can be made with large gains without sacrificing any of its outstanding electrical characteristics and without increasing prohibitively the overall package size and weight. Periodic focusing of TWT's has resulted in a very lightweight, compact structure ideally suited to spaceborne applications, especially in the frequency range from 2 - 14 GHz. Herein lies one of the significant advantages of the TWT compared to its counterpart in the klystron field. At present, klystrons are often focused with heavy, bulky permanent magnets of the horseshoe variety. These large magnets create an extensive leakage field which affects all of the surrounding electronics, and furthermore, there is no easy method of shielding these fields without degrading the klystron performance.

By appropriately designing the electron gun optics so that the emission density at the cathode surface is quite low, TWT's can be made to yield an arbitrarily long life. This has been established, for example, with the Hughes Aircraft TWT's used in communications satellites and other space programs without a single failure attributable to the tube design. An expected life greater than 40,000 hours has been established beyond all reasonable doubts for these tubes.

Table C summarizes the significant advantages and disadvantages of traveling wave tubes.

Table A. The Types of Microwave Velocity Modulation Tubes are Classified in Terms of Oscillators and Amplifiers

			Interaction Circuit					
			Backward Wave		Forward Wave		Standing Wave, cavity	
			Oscillator	Amplifier	Oscillator	Amplifier	Oscillator	Amplifier
Electric-Magnetic Fields	Crossed Field (M-type)	Injected Beam	M-BWO M-Carcinotron	M-BWA Ritermitron	TPOM	CFA TPOM Bimatron		
		Continuous Cathode	Stabilotron	Amplitron CFA	VTM	FWA-CFA Dematron	Magnetron	Circlotron
	Linear (Single or Parallel) Field (O-type)	Injected Beam	O-BWO O-Carcinotron	O-BWO		TWT TPO	Klystron Reflex Klystron Monofier Monotron	Klystron

Table B. Klystron Amplifier Characteristics

Advantages	Disadvantages
<ol style="list-style-type: none"> 1. Single envelope is practical for gains up to 30 db, depending on output power levels. 2. Much development experience exists both at high power (megawatts) and lower drive levels (tens of kws). 3. Bandwidths to 10 percent have been achieved in high-power units; 5 percent or less is more realistic at low power. 4. Klystrons can be made to operate with dc beam supplies by employing switching or modulating anodes. 5. Klystrons still offer higher power than any other tube type. 6. Focusing is normally performed by solenoids. 7. More design and development experience exists on this type of beam device than on any other. 8. Higher perveances have been achieved than in TWTs. This permits lower beam voltages for equivalent output power. 9. System reliability is enhanced in cases where 30 db gain is sufficient, since a single amplifier stage is adequate. 10. Cooling techniques are known and optional; air and liquid are common. 11. Since the klystron is a unidirectional device, operation into a mismatched load is possible without isolation, depending on system limits of power and phase shift. 12. More recently developed electrostatically focused klystrons offer reduced size and weight. 13. Shorter electrical length per unit gain than the TWT, thus suffering less voltage-phase sensitivity. 	<ol style="list-style-type: none"> 1. Longer electrical length per db of gain as compared to cross-field devices. 2. Gains are about half those obtained with TWTs. 3. Klystrons require a filamentary cathode and gun for operation, thus more electrode voltages are needed. 4. Higher beam voltages required for a given output power than in crossed-field devices. 5. Bandwidth limitations are severe at low power (tens kws); 2 to 5 percent is typical. 6. Good voltage regulation may be required for acceptable phase stability. 7. Efficiencies of 25 to 35 percent are typical. 8. Structure offers moderate filtering; may require separate high power filters at an additional cost to the system. 9. Reliability not reasoned to be as good as cold cathode devices operated without gun.

Transmitting Power Sources

Radio Frequency Tube Sources

FUNDAMENTALS OF MICROWAVE SOURCES

Crossed-Field Devices

This general category of tubes includes the conventional magnetron oscillator, the amplitron amplifier, and the many linear beam-type magnetron amplifiers. These devices are, in general, more efficient, lighter and smaller for a given output power than any other tube device. The basic problem with these tubes is their relatively short life. Recent advances, such as the coaxial magnetron oscillator, show promise of increasing the life; however, extensive data on these increases is not yet available. Table D summarizes the significant advantages and disadvantages of crossed-field devices.

Table C. Traveling Wave Tube Characteristics

Advantages	Disadvantages
<ol style="list-style-type: none"> 1. Single envelope is practical for gains to about 60 db. 2. Much development experience exists both at high power levels (megawatts) and lower drive levels. 3. Bandwidths beyond 10 percent are common in today's TWT's. 4. Can operate with dc beam supplies by employing switching or modulating anodes. Grid control is possible at power levels below 10 kw. 5. Long length, small-diameter form factor is suitable for phased arrays limited to tube diameters of less than $\lambda/2$. 6. Several types of slow-wave structures can be cascaded in a single envelope to optimize the rf coupling design. 7. Focusing can be accomplished by electromagnets, permanent magnets, or electrostatically, depending on system requirements resulting in low external magnetic fields. 8. Reliability is enhanced by single-stage operation. 9. Depressed-collector techniques ease regulation requirements on high-current beam supply. 10. Cooling techniques are known and operational; air and liquid are common. 11. Severed or attenuated slow-wave structures enable operation without circulators into a mismatched load, depending on the limits of phase tolerance. 12. No doubt exists about development and mass production of TWT's. 13. Light weight compared to alternate tube implementations. 14. Several lower power tubes can be operated in parallel to achieve higher power, higher reliability, and higher heat dissipation over larger areas. 	<ol style="list-style-type: none"> 1. Long electrical length per db of gain in comparison to other tubes. 2. Phase sensitivity of rf output is very dependent on beam voltage, because the device operates on the principle of synchronism between electron velocity and rf velocity on the slow-wave structure. 3. TWT's require a filamentary cathode and gun for operation — more electrode voltages are thus needed. 4. Low perveance of TWT's requires higher beam voltages for a given power output compared to other devices. High perveances involve hollow beams, whose increased current density may cause focusing difficulties and cathode loading problems. 5. High gains usually require solenoid focusing (or a recently proposed magnetic matrix focusing technique). Solenoids for high power tubes are large, require substantial power, and create packaging and distribution problems. 6. Acceptable phase stability calls for good beam voltage regulation. 7. Efficiency is ≈ 25 percent without depressed collectors. Depressed collectors raise efficiency by ≈ 10 percent. 8. Electrical structure offers very little filtering. Separate high-power filters may be needed.

Table D. Crossed-Field Tube Characteristics

Advantages	Disadvantages
<ol style="list-style-type: none"> 1. Gains to 45 db possible with injected beam variety; 10 percent bandwidth typical. 2. Good phase stability; fairly independent of anode voltages, since velocity synchronism is not an operating requisite. 3. Distributed emission type has shortest electrical length per unit gain of all tubes. 4. Can operate with a cold cathode; electron gun or filaments are avoided. 5. Reasoned to have long life compared to filamentary devices. 6. Small in size and relatively light in weight. 7. Distributed-emission type provides the lowest anode voltage for a given power output. This simplifies the power distribution system. 8. Air or liquid cooling is acceptable. 9. Highest microwave tube efficiencies: Injected beam — 35 to 50 percent Distributed emission — 45 to 65 percent 10. Operated at saturation to give a constant output independent of input power variation; makes a good limiter. 11. Can be used in a duplexed system with receiver and duplexer on the low input side. 12. Structure is ideally suited for mass production under tight mechanical tolerance control. 	<ol style="list-style-type: none"> 1. Limited gain of 10 to 15 db for distributed emission and types. 2. Has very limited dynamic range of output power for a particular anode voltage and varying input power. 3. Crossed-field tubes are bidirectional. Any reflections into the output will return directly to the driver or input circuit. A circulator or isolator is thus necessary. 4. Injected beam types require very good regulation of sole voltage for good phase stability. Typical values are 20 degrees for a 1 percent change in sole voltage. 5. In distributed emission types it is difficult to initiate emission. 6. Perveance of injected beam types is comparable to that of TWT's, thus requiring higher beam voltages for a given output power. 7. Relatively short life.

Transmitting Power Sources
Radio Frequency Tube Sources

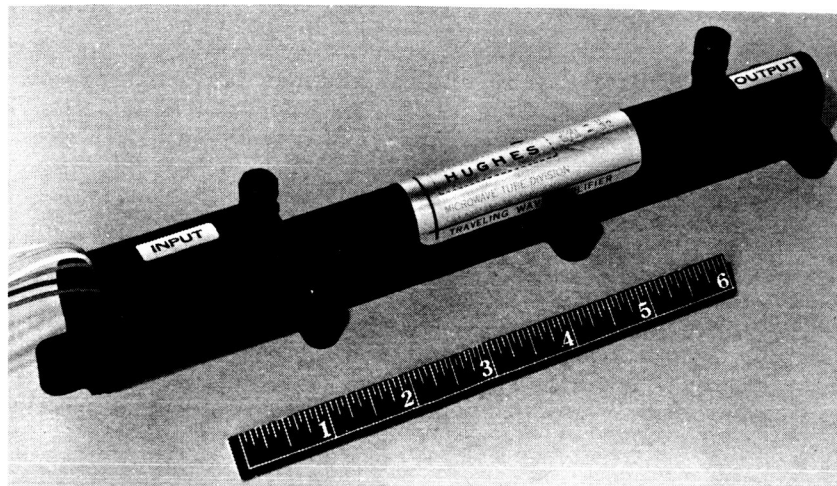
HUGHES 394H TWT PERFORMANCE

A space qualified traveling wave tube is described in some detail as representative of spacecraft power amplifiers.

Highly reliable traveling wave tube manufacturing techniques were developed at Hughes Aircraft Company during the 1960 to 1965 period. As a result several space qualified TWT have been provided for various United States spacecraft. These spacecraft include the Syncom series of spacecraft, Surveyor, Mariner, Lunar Orbiter, the ATS spacecraft, Apollo, and Intelsat I, II, IV. This series of tubes has an output power range of 2 to 20 watts.

Typical of these tubes in many respects is the 394H, the tube used in the Apollo spacecraft. One feature that is not typical of most of these tubes is the ability to change the power output. The 394H has two power outputs, 5 watts and 20 watts. This was incorporated in order to consume less d. c. power when a high power output was not needed.

The figure illustrates the 394H and the table gives typical specifications.



Hughes 394H Traveling Wave Tube

Specifications for 394H TWT

	5 Watt Mode	20 Watt Mode
Frequency	1.8-2.6 GHz	1.8-2.6 GHz
Power Output	5 W	20 W
Duty	CW	CW
Gain	20 dB	26 dB
Efficiency*	25%	33%
Beam Voltage (Cathode) E_b	-1175 V	-1425 V
Beam Current I_b	26 mA	60 mA
Helix Current I_w	0-6 mA	0-10 mA
Anode 1 Voltage E_a	-350 V	0 V
Anode 1 Current I_a	0.1 mA	0.1 mA
Anode 2 Voltage E_a	+100 V	+100 V
Anode 2 Current I_a	0.1 mA	0.1 mA
Collector Voltage E_c	-550 V	-450 V
Collector Current I_c	20-26 mA	50-60 mA
Heater Voltage	5.2 V	5.2 V
Heater Current	.3 A	.3 A
Expected Life	90,000 hrs.	25,000 hrs.
Cooling	Conduction	Conduction
Focusing	PPM	PPM
Weight	20 oz.	20 oz.
Length	9.5 in.	9.5 in.

* Midband efficiency including heater power.

FUNDAMENTALS OF MILLIMETER AND SUBMILLIMETER SOURCES

Linear beam and traveling-wave tubes are the types best suited for use as millimeter sources (30 GHz to 300 GHz). Submillimeter sources are still in the research and development stages and use harmonic generation to obtain the desired frequencies.

Millimeter Sources

Most of the successful attempts to develop sources in the millimeter wave region of the spectrum have involved extensions and extrapolations of the microwave device techniques. However, since the slow-wave circuit must have dimensions comparable to a wavelength, the problems of electron control and circuit thermal dissipations become formidable. The linear beam, or O-type distributed interaction device overcomes these difficulties better than the klystron or crossed-field tube for rather fundamental reasons.

In the traveling-wave tube the electron stream need not touch the rf circuit, and the beam collection function is accomplished by a separate and easily cooled electrode. A distributed interaction device, such as a traveling-wave tube, has an additional advantage over a single output gap tube, such as a klystron, for when all the power must be transferred to the output circuit via a single gap, high Q circuits are involved with the problems of multipactor effects and voltage breakdown.

Surprisingly, the efficiency of the overall device in millimeter wavelengths does not suffer appreciably when compared to microwave tubes. Because the millimeter interaction efficiency is necessarily low, the beam has a more uniform distribution of velocities. Accordingly, the collector electrode can be operated closer to cathode potential without danger of returning slow electrons into the circuit region. This operation results in a recovery of the overall conversion efficiency.

Submillimeter Sources

The conventional techniques of rf device design become impractical at frequencies over 300 GHz. Barring a technological breakthrough, one of the better means of producing power at these frequencies involves harmonic generation. Because large powers are available in the millimeter wave region, comparatively low harmonics are needed with reasonable conversion losses. The problem resolves itself to the development of a non-linear device that can accept large input powers. One of the likely candidates for this application is high pressure gas discharge plasmas. Notable success has been achieved in this area and there appears to be considerable promise for further development.

WEIGHTING FACTORS

Rule of thumb relationships are given to relate frequency, power and weight. Bounds on power output as a function of frequency is also given.

Because of the complexity of the tube design problem, it is difficult to reduce the available tradeoffs to simple scaling laws for all parameters. The general rule is to indicate the difficulty in achieving a particular performance level relating any two designs according to the factor of power times the square of the frequency. While this guide represents a gross oversimplification, it is still probably the best first approximation to a general scaling factor.

Some of the scaling laws pertaining to particular tube parameters can be given crude approximations. Since power, weight, voltage, bandwidth and frequency are the most important considerations for system application, the following guide is intended to relate these parameters.

For Constant:	then:	
Frequency	Power	Proportional to (Voltage) ^{1/2}
Power	Frequency	Proportional to Voltage
Frequency	Weight	Proportional to (% Bandwidth)

Figure A indicates the frequency power relationship based on an rf heating limitation, while Figure B indicates the fabrication tolerances required to achieve satisfactory TWT performance.

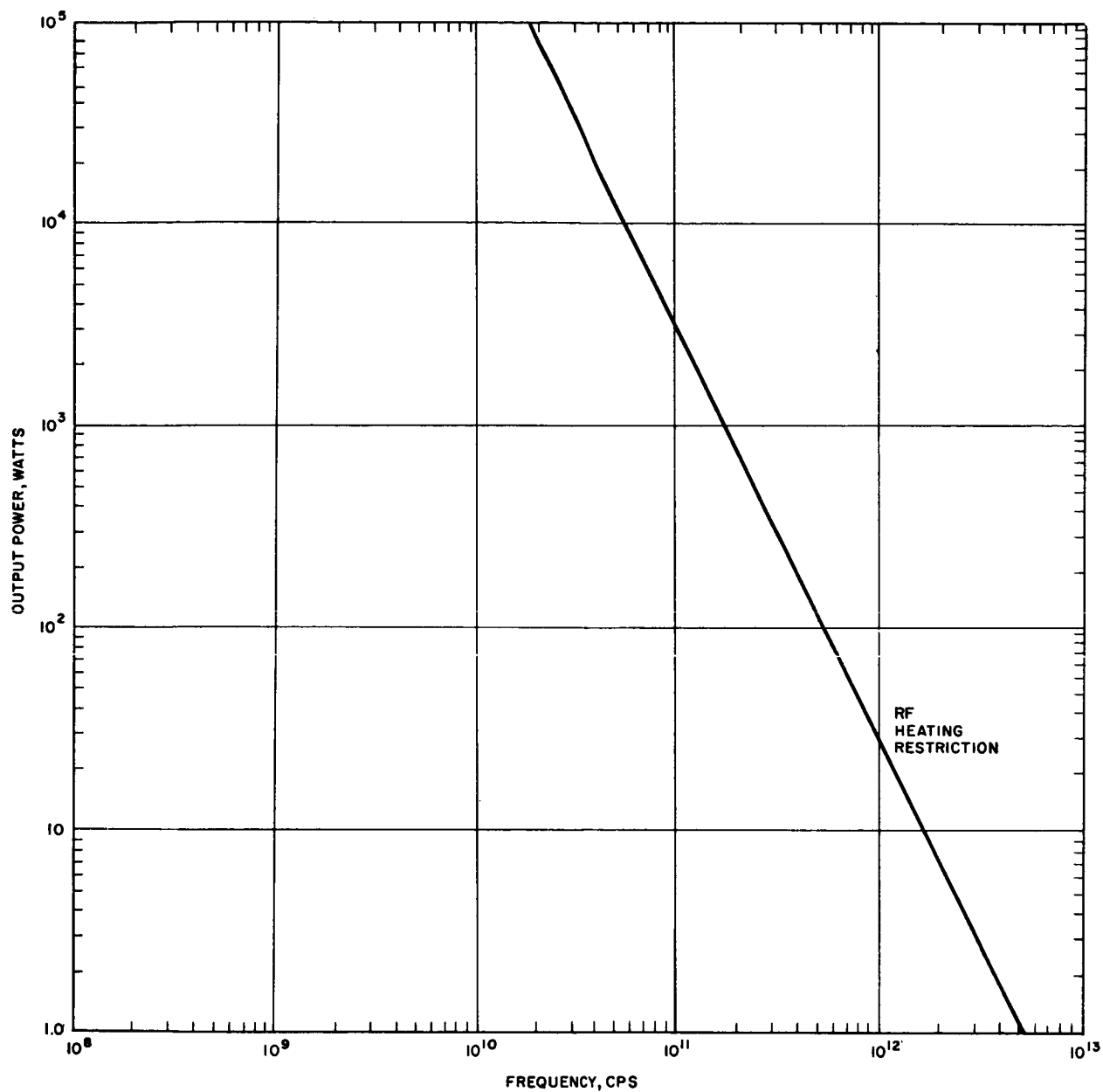


Figure A. Power Limitations for a Single TWT, as a Function of Frequency

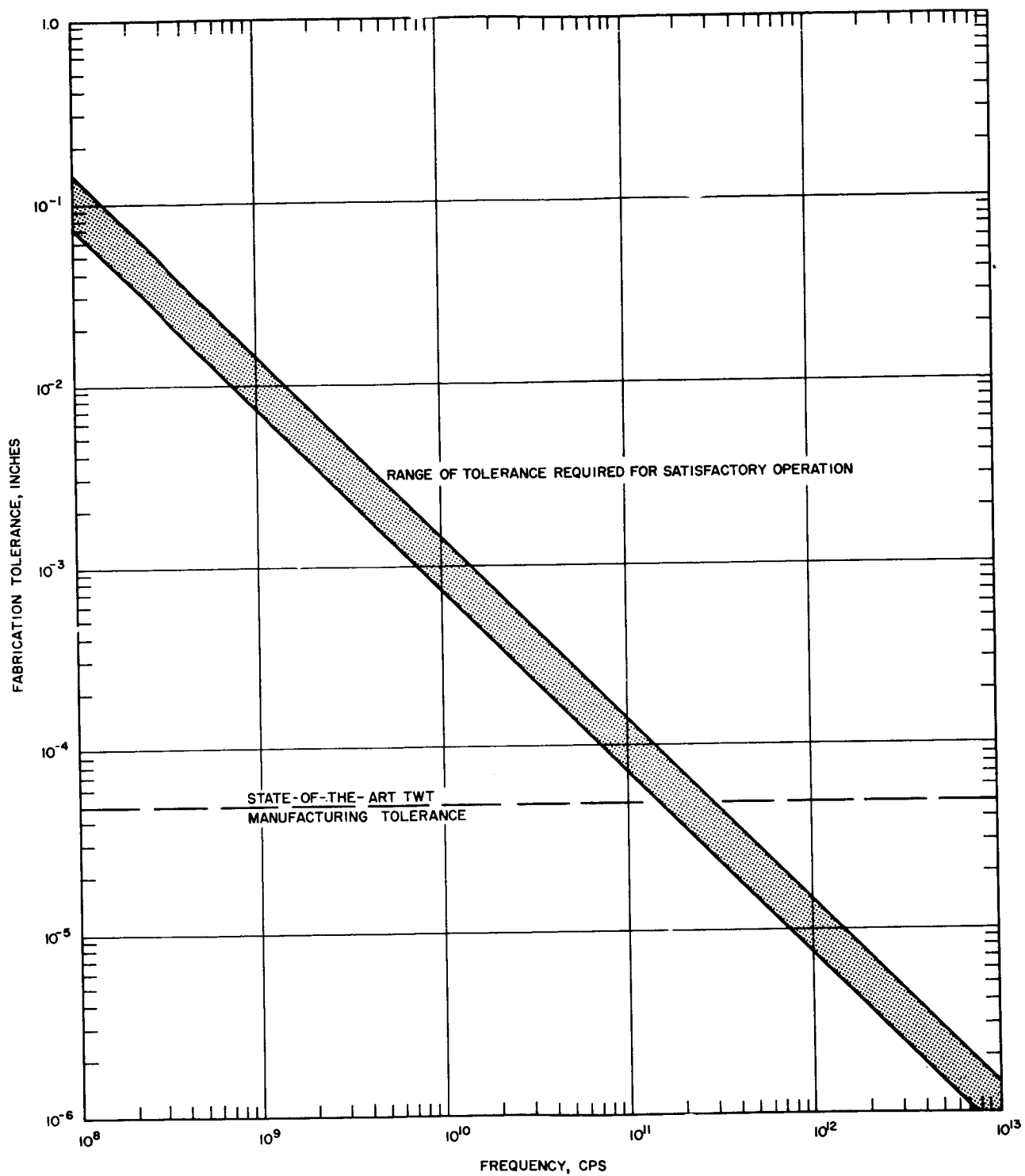


Figure B. Required Manufacturing Tolerances in TWT's as a Function of Frequency

PERFORMANCE OF VACUUM TUBE SOURCES

UHF, microwave, and millimeter wave sources provide a wide selection of unique performance parameters.

UHF Sources

The present state-of-the-art in UHF power grid tubes varies from a 100-MHz tube that supplies 100 kilowatts and weighs 100 pounds, to a 1000-MHz tube that supplies 100 watts and weighs 5 pounds.

Microwave Sources

Microwave sources fall into two general classifications. First are the relatively low power tubes that are very light and provide long and reliable life operation. Several such tubes have been used in current spacecraft with excellent results. In general, they provide from 2 to 20 watts in the range of 1 to 10 GHz and weigh about one pound. However, these tubes cannot compete on a power/weight basis with the brute force power tubes that have been built without primary concern for total system weight. The latter tubes demonstrate power in excess of 100 kw, although the weight of the tube and magnet system can easily exceed 1000 pounds. Indeed, development is in progress on a tube to produce one megawatt of cw power at X-band. While the weight and voltage penalty for these tubes is high, they are without competition for transmitting maximum power levels.

Millimeter Wave Sources

Available power levels for millimeter waves between 30 and 100 GHz have increased by three orders of magnitude since 1960, and the efficiency of millimeter sources has increased to be competitive with microwave sources. One kilowatt cw sources are now available at 35 and 55 GHz with efficiencies up to 35 percent. These levels are being achieved in devices using reasonable voltages and having operating lives of many thousands of hours. Methods are being developed to permit economical manufacture. New techniques are being exploited to realize lightweight sources suitable for airborne and space use.

It is convenient to separate the cw millimeter sources into low and high power categories. Since there seems to be a relative abundance of sources delivering tens of milliwatts, but very few delivering over 1 watt of cw power, a division at the 1 watt level has been chosen to separate "low" power from "high" power millimeter sources.

All of the commercially available low power prime sources are either backward-wave oscillators or klystrons. The power available from the prominent tube lines supplied by various manufacturers is shown in Figures A and B for the millimeter portion of the spectrum. Extensive lines of low power backward-wave oscillators are marketed by four companies. The most complete line is the Bendix TWO series, which covers the entire range from 40 to 140 GHz. Most of these tubes can be procured with either solenoids or permanent magnets. They use Karp structures and can be electronically tuned over 15 percent ranges. Maximum operating voltages are 3500 volts or less.

Sperry has developed a similar line of tubes covering the range up to about 90 GHz. Designations are SBM 421 and SBE 402. These tubes are packaged in permanent magnets and weigh only 7 pounds. Both tubes use the Karp structure and maximum operating voltage is 3200 volts.

Siemens-Halske markets the RWO 40, 60, and 80 covering the frequencies from 26.5 to 90 GHz. Power output varies from 60 to 5 mw. The RWO 60 weighs about 17 pounds. These tubes require several variable voltages for various focusing electrodes. They use a form of interdigital slow-wave structure and convergent electron guns.

CSF of France dominates the very high frequency range beyond that shown in Figure A. The COE-20 will deliver 500 mw over a 10 GHz range at about 140 GHz, and will deliver about 1 watt over a few gigahertz. This is the first tube discussed which uses the Millman structure and is more typical of the high power designs. The COE 10 delivers 10 to 20 mw over a 10 percent range of 300 GHz. The COS-09 delivers 30 to 50 mw in the 350 GHz range. The COS-07 delivers 5 to 10 mw in the 400 GHz range, while the COS-06 delivers 5 mw in the vicinity of 485 GHz. This company (CSF) has demonstrated an oscillator which delivers 1 mw at 708 GHz, the highest frequency oscillator ever generated by this means. All of these tubes use the same general structure together with highly convergent electron guns. For the most part, operating voltage is kept to 7000 volts or less.

Many companies market extensive low power klystron lines operating up to 170 GHz. Most extensive coverage is achieved by Varian, which lists tubes capable of delivering 100 mw or more to 140 GHz, and 50 mw to 170 GHz. Maximum voltage is 2500 volts. The tubes are extremely light in weight and are air cooled. Oki Electric of Japan lists a series of reflex klystrons covering the range from 30 to 100 GHz. Available performance data indicate power levels of over 100 mw in the 30 GHz range to about 60 mw in the 75 GHz range. Amperex (Philips) lists the DX 184, 151, 242, and 237 which operate at 8 mm, 3.2 mm, and 2.5 mm, respectively. They deliver several tens of milliwatts up to 100 mw. Raytheon markets an extensive line with frequency coverage to about 120 GHz in its QKK series. Power levels vary from 100 mw at the lower frequencies to 20 mw at the higher frequencies. Litton Industries also markets several relatively high powered reflex klystrons in the 35, 50, and 70 GHz ranges.

The high power CW tubes shown in Figure C delivers 1 watt or more. All of the tubes which meet this requirement, except CMO8X, are linear beam, O-type devices. The type number of each is shown, followed by the letters O or A to designate it as an oscillator or amplifier. The oscillators shown in Figure C are of either the floating drift tube klystron type or the Millman backward-wave oscillator type, or are closely related to these types of tube.

Hughes markets backward-wave oscillators delivering 15, 8 and 2 watts at 5.5, 3.2, and 2 mm, respectively. These tubes are air cooled and operate at efficiencies up to 15 percent. All use depressed collectors to achieve this relatively high efficiency and low power dissipation in the tube.

PERFORMANCE OF VACUUM TUBE SOURCES

Litton markets the Elliott line of floating drift tube klystrons which operate in the 35 and 50 GHz ranges. The highest power available is 50 watts at 35 GHz from the L3736. All of the tubes are liquid cooled and mounted in permanent magnet packages weighing about 10 pounds.

Oki lists the 35F10 and 50F10 laddertrons which deliver 10 and 2 watts, respectively. Their principle of operation is similar to the floating drift tube klystron except that a distributed interaction is provided in the cavity. CSF lists another Millman type of oscillator which delivers 8 watts over an extensive range at 70 GHz. Watkins-Johnson advertises the W-J 266 amplifier which includes an internal feedback mechanism to make it oscillate.

Hughes lists the 812H, 813H, and 814H traveling-wave amplifiers which deliver 150 and 1200 watts at 5.5 mm, and 150 watts at 3.2 mm. These tubes use coupled-cavity slow-wave structures and highly convergent electron guns. The 812H and 814H are air cooled while the 813H requires liquid cooling. All of the tubes use depressed collectors (operating at about 30 percent of beam potential) to maximize efficiency. The 812H and 813H operate at 30 to 35 percent efficiency, while efficiency of the 814H is 20 percent. Small signal gain of the 813H is 26 db, with saturated gain over 20 db.

Two new tubes have been developed which have significantly advanced the state of the art of high power Millimeter Wave Source generation. The Hughes 819H has demonstrated over 5 kw of average power output at 55 GHz with a large signal gain of more than 20 db. Collector depression results in efficiencies of about 30 percent. The Hughes 826H tube has shown output power of almost 600 watts at 94 GHz. This tube has been operated in excess of 10 percent duty cycle with such very long pulse widths that the design is believed to be fully capable of cw operation. With reference to Figure C, these results obviously describe a new boundary to the state-of-the-art. In six years, the available output power has been increased by approximately two orders of magnitude. The W-J 226 also uses the coupled cavity circuit and delivers 1 kw at 35 GHz. Gain is 13 db and operating efficiency is about 10 percent. This tube is also liquid cooled, but does not incorporate a depressed collector.

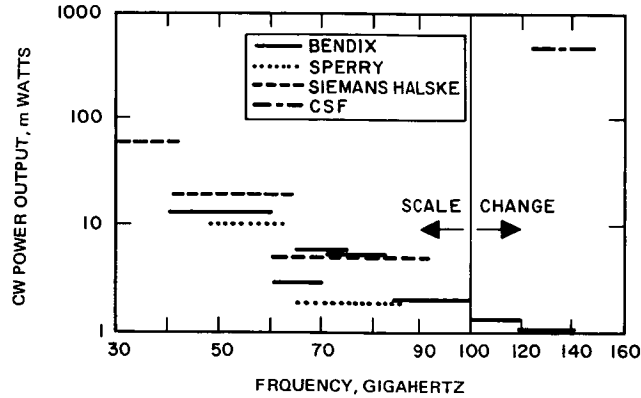


Figure A. Power Characteristics of Available Low-Power Backward-Wave Oscillators

Sperry and Bendix use Karp structure, and CSF used the vane line. In general, these tubes use low operating voltages and are suited for local oscillator and laboratory source use.

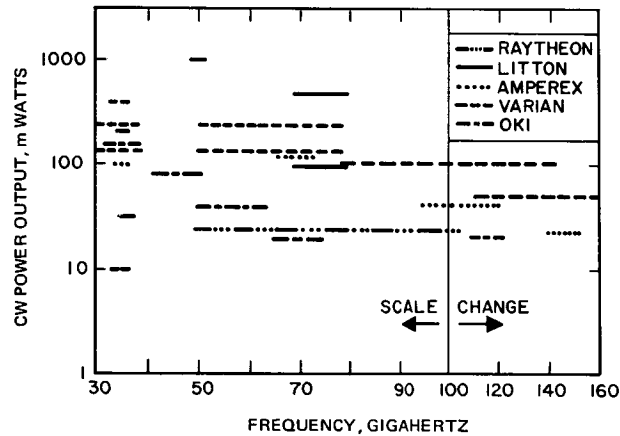


Figure B. Power Characteristics of Available Low Power Klystrons

All tubes are reflex klystrons and are light in weight since they use no magnetic focusing fields. They are particularly suited as pump, local oscillator, and laboratory signal sources.

Transmitting Power Sources Radio Frequency Tube Sources

PERFORMANCE OF VACUUM TUBE SOURCES

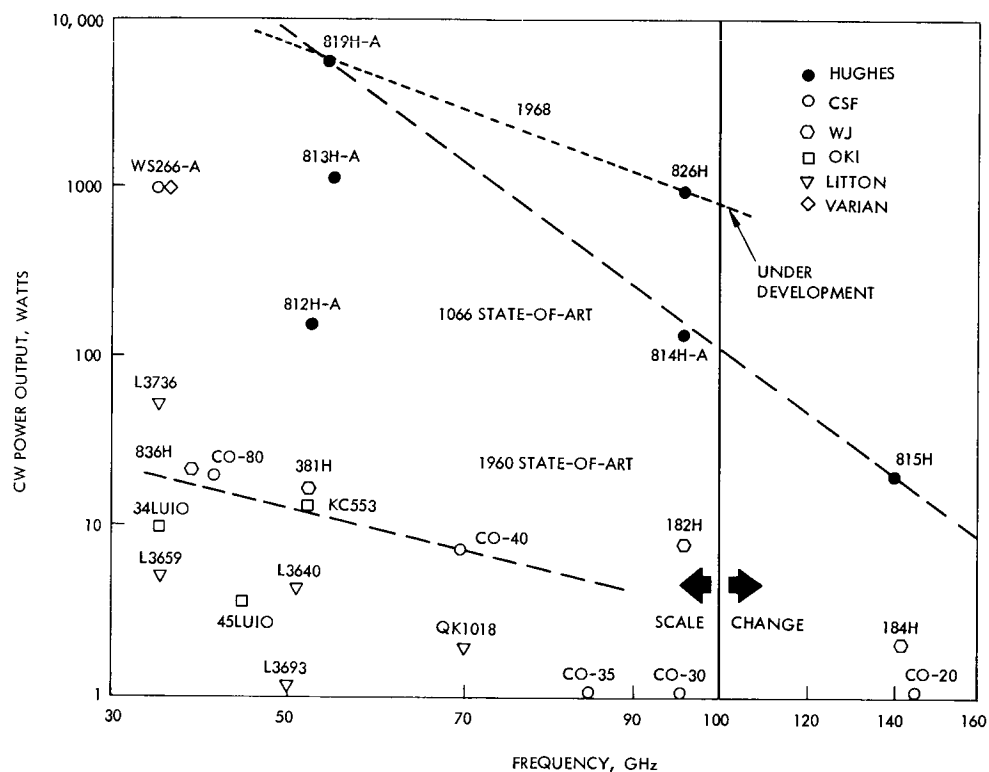


Figure C. Power Characteristics of Available High Power CW Sources

Most significant is 35 percent efficiency achieved in the 813H and marked increase in available power which has occurred since 1960. The letter -A stands for amplifier and the letter -O signifies oscillator.

TRANSMITTING POWER SOURCES

Radio Frequency Solid State Sources

	Page
Introduction to Solid State Sources	28
Theory of Operation for Impatt Oscillators	30
Theory of Operation for Gunn Oscillators	32
LSA (Limited Space Charge Accumulation) Power Sources	36
State of the Art for LSA, Impatt, and Gunn Microwave Sources	38
Radio Frequency Burden Relationships	42

INTRODUCTION TO SOLID STATE SOURCES

IMPATT and Gunn solid state microwave sources hold promise of producing, for their extremely small size, relatively high power levels at high efficiencies. Transistor power output is given as a function of frequency.

Most active microwave solid state devices can be used either as oscillators or amplifiers, depending on the associated external circuitry and choice of operating point on the device characteristic. This discussion is limited to the oscillator mode of operation, since, to date, significant data is available only for this mode. The various microwave solid state oscillator sources are classified as follows:

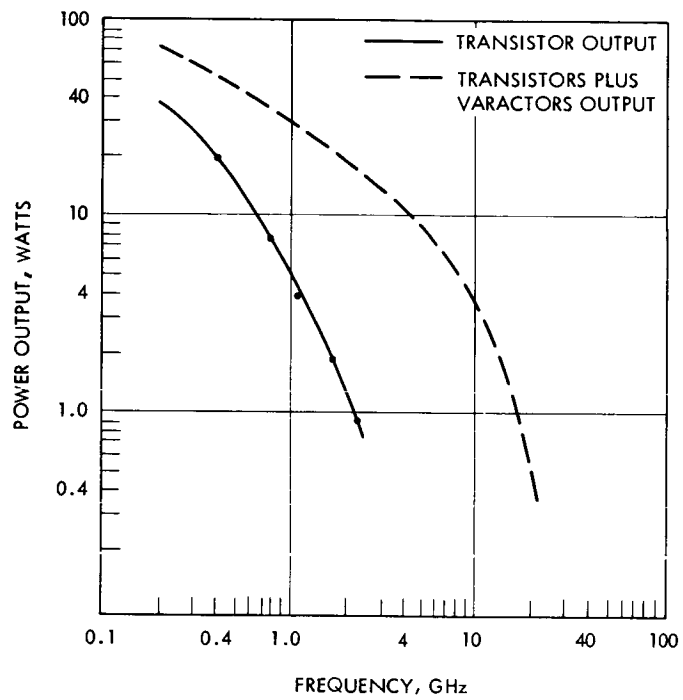
- Parametric oscillators
- Tunnel diode oscillators
- Transistor oscillators
- Transistor oscillator - varactor multiplier combination
- Impact avalanche transit time (IMPATT) oscillators
- Gunn effect oscillators

Parametric oscillators are not of prime interest since they require auxiliary high frequency pump oscillators. Tunnel diode oscillators operate at high frequencies, but appear to be power limited. Thus, these two types are excluded from further consideration in this discussion. The power generation capabilities of transistor oscillators and associated varactor multipliers are summarized briefly. The bulk of the discussion is devoted to the IMPATT (IMPact Avalanche Transit Time) and Gunn oscillators, which appear at this time to offer potential for high power levels at high efficiency.

Transistor Oscillators

Extensive efforts have been underway for many years to improve the power-frequency characteristics of transistor oscillators. Recent progress has been achieved through use of paralleled transistors, monolithic integrated circuit techniques, and incorporation of improved varactor multipliers. The current state of the art as summarized by Matthei¹ is shown in the figure. The stimulus for study of the IMPATT and Gunn oscillators is provided by the fact that the power frequency performance of the figure is already exceeded in many instances with rudimentary versions of these more recent devices.

¹Matthei, W. G., "Recent Developments in Solid State Microwave Devices," Microwave Journal, 9, No. 3, p. 39, March 1966.



State of the Art of the Power Output of
Primary Transistor Oscillators and
Transistor Varactor-Multipliers
Versus Output Frequency

THEORY OF OPERATION FOR IMPATT OSCILLATORS

Physical constants of IMPATT oscillators are used to describe the operation of IMPATT oscillators and the frequency of oscillation.

In general, IMPATT oscillators involve the use of a semiconductor biased in the reverse direction and mounted in a microwave cavity. Bias is applied so that the operating point occurs in the avalanche breakdown region of the diode characteristic. Negative resistance and oscillations have been obtained in simple reverse biased p-n junctions as well as in the more complex n^+pip^+ configuration as first proposed by Read.

Gilden and Hines¹ have depicted the typical p-n junction in reverse bias for avalanche conditions as shown in Figure A. The active zone is composed of the thin avalanche region followed by the depletion zone through which carries drift at the saturation velocity, V_D . The realization of a negative resistance is associated with a 90-degree phase lag between current and applied ac voltage which occurs in the avalanche process, followed by a further 90-degree lag during the transit time of the carriers through the depletion zone.

Hines has shown that the equivalent impedance of the diode can be expressed as

$$Z = R_s + \frac{l_D^2}{V_D \epsilon A} \frac{1}{\frac{1-\omega^2}{\omega_a^2}} + \frac{1}{j\omega C} \frac{1}{\frac{1-\omega_a^2}{\omega^2}} \quad (1)$$

under the condition that the transit time through the depletion layer (expressed in radians)

$$\theta = \frac{\omega l_D}{V_D} = \omega \tau_D$$

is less than $\pi/4$. R_s is the bulk resistance of the region outside the depletion zone, l_D is the length of the depletion zone, A is the junction area, and C is the capacitance of the depletion and avalanche zones, i. e.,

$$C = \frac{\epsilon A}{l_a + l_D}$$

¹Gilden, M. and Hines, M. E., "Electronic Tuning Effects in Read Microwave Avalanche Diodes," IEEE Trans on Electron Devices, ED-13, No. 1, p. 169, January 1966.

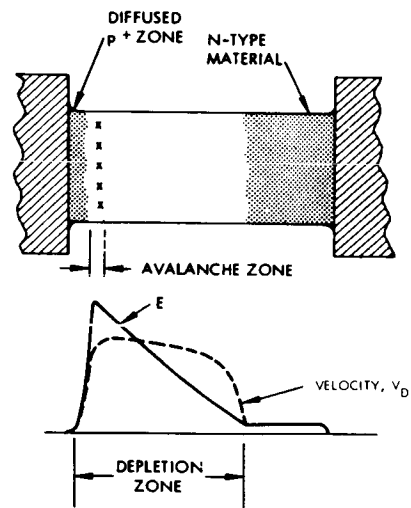


Figure A. Typical P-N Junction
in Reverse Bias (Conditions at
Avalanche)

THEORY OF OPERATION FOR IMPATT OSCILLATORS

The parameter ω_a is the avalanche frequency and is given by

$$\omega_a^2 = \frac{2a' V_D I_0}{\epsilon A}$$

where a is the ionization coefficient and the prime denotes the derivative with respect to electric field, and I_0 is the bias current. It is important to note that ω_a is a function only of the materials, junction geometry, and bias current. The equivalent circuit suggested by equation (1) is shown in Figure B, where the variable resistance R_D represents the second term and the impedance of the tuned circuit the third term. Note that for $\omega < \omega_a$, R_D is positive and the tuned circuit presents an inductive reactance. For $\omega > \omega_a$, R_D becomes negative and the reactance becomes capacitive. This dependence is sketched in Figure C. The oscillator is formed by supplying an inductance L in parallel through microwave circuitry; the negative of this inductive reactance is plotted in Figure C. The oscillator will operate at frequency ω_r where the diode capacitive reactance is equal to the inductive reactance; ω_0 is the resonant frequency of the system at zero bias current. Tuning of the oscillator is accomplished either by varying I_0 (and thus ω_a) or by varying L through mechanical adjustment of the microwave cavity. The operating frequency of the IMPATT oscillator will be given approximately by

$$f \approx \frac{V_D}{2\ell_D}$$

Since V_D is approximately twice as high for gallium arsenide as compared with silicon, the former will be more desirable for higher frequency operation. On the other hand, silicon technology is more advanced, and the higher defect density and higher thermal resistivity of gallium arsenide may limit its power capability.

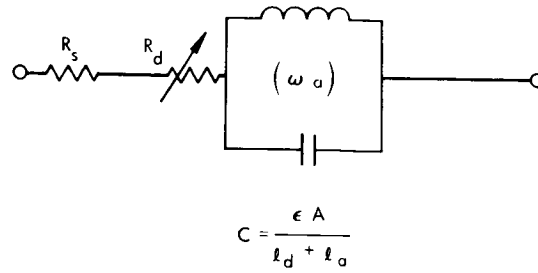


Figure B. Equivalent Circuit
for Avalanche Diode at
Small Transit Angle

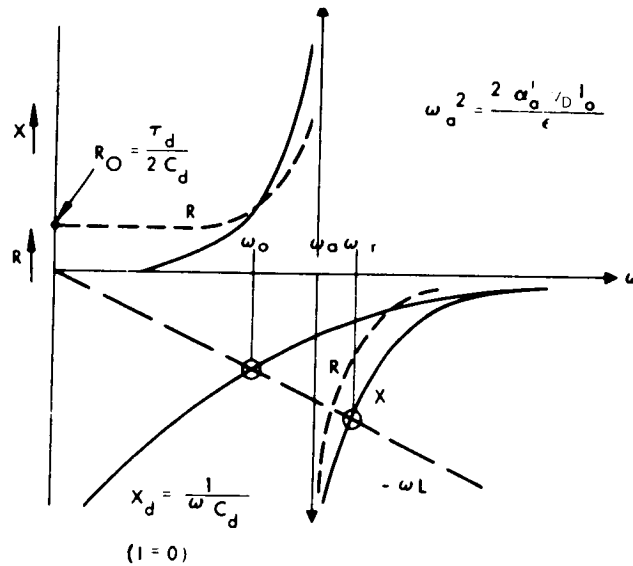


Figure C. Frequency Dependence of
the Real and Imaginary Parts of
the Diode Impedance

THEORY OF OPERATION FOR GUNN OSCILLATORS

The theory of the Gunn oscillator is described and the oscillation frequency is related to the operating constants.

Unlike the IMPATT oscillator, the Gunn effect oscillators do not require the formation of a junction. The effect takes place in a sample of the bulk material which is supplied with ohmic contacts on each side. Typical length of the samples is 25 to 200 μ . Gallium arsenide is the most common material; however, the effect has been observed in cadmium telluride and indium arsenide. Typical doping densities for gallium arsenide are 10^{13} to $10^{14}/\text{cm}^3$.

Two modes of operation associated with the general Gunn effect concept have been observed. The first, as originally observed by Gunn,¹ takes place in the parameter range

$$\rho L > 2 \times 10^{12} \text{ cm}^2$$

where ρ is the doping density and L the length of the sample. A second mode of operation as observed by Thim,² et al., occurs in the range

$$2 \times 10^{12} > \rho L > 10^{10} \text{ to } 10^{11} \text{ cm}^2$$

The Gunn effect is essentially a current instability which occurs when an electric field of 3 to 5 kv/cm is applied across the bulk semiconductor. Associated with this field is saturation of the drift velocity of the carriers, thus removing the response out of the ohmic range. Increasing the field slightly beyond the onset of current saturation results in an instability which is manifested by large periodic spikes in the current.

This effect is explained for gallium arsenide by a decrease in average mobility of the carriers as field is increased beyond the threshold. In the energy band structure of gallium arsenide, in addition to the low lying principal conduction band, there are six satellite bands lying approximately 0.36 eV above the principal band. The mobility of electrons in the principal band is much higher than for that of the satellite bands. When the electric field reaches the threshold value, many electrons acquire sufficient energy to transfer to the satellite conduction bands where their mobility is low. Thus, on the average, a decrease in mobility occurs, and consequently a reduction in current, or, equivalent, a negative resistance characteristic evolves.

¹Gunn, J. B., "Instabilities of Current in III-V Semiconductors," IBM Journal Res. Develop., 18, No. 2, pp. 141-159, April 1964.

²Thim, H. W., et al., Microwave Amplification in a dc Biased Bulk Semiconductor, "Appl. Phys. Letters, 7, p. 167, September 1965.

When the applied field is above threshold, then it can be shown that a high field region (domain) with two low field regions are created within the sample. These domains then travel the length of the sample, L , in a transit time, τ .

$$\tau = \frac{L}{V_D}$$

Where V_D is the saturation velocity. The domains are created at the negative electrode, travel across the sample to the anode where they disappear simultaneously with the formation of the next domain. The fundamental frequency of the modulated current which results is given by

$$f = \frac{1}{\tau} = \frac{V_D}{L}$$

Since the frequency is dependent upon transit time, it is essential to have a planar surface at the negative electrode. Curved surfaces can lead to erratic domain formation in space and thus erratic output frequency.

LSA (LIMITED SPACE CHARGE ACCUMULATION) POWER SOURCES

LSA solid state devices avoid some of the limitations of IMPATT and Gunn oscillators to present promise of much higher rf power.

LSA power sources do not operate in a transit time limited mode as do the IMPATT and Gunn oscillators. Thus, a power-frequency-impedance restriction does not dictate the performance of the devices and allows the use of a bulk sample of the material many times larger than the transit length for the frequency of interest. Also, since the space charge is immediately quenched after traveling only a small portion of the total length of the active region of the device, the field remains in the negative resistivity range throughout most of the sample. This means that a greater volume of the sample contributes to microwave power generation than in the other solid state devices and a comparatively large piece of active material may be used to obtain the high power levels predicted.

Because LSA devices are not transit time controlled, the frequency of oscillation is determined by the tuning of the microwave circuit. The nature of the LSA operation and the state of the present devices is such that a fast rise-time bias pulse must be applied to power the generation of the microwaves and yet not destroy the device. LSA modes occur when the applied potential is several times the threshold for the formation of space-charge domains (approximately 4 or 5 times this threshold for best efficiency). If such domains are allowed to form and are not quenched after traveling a small portion of the device length, the heat generated by the large internal currents destroys the present devices. A radio-frequency voltage large enough to swing the field in the material below the threshold, must therefore be applied to the device as early as possible; at least within one transit time for a domain ($1/2$ to 1 nanoseconds for 350 to 700 micron thick devices now being studied). The large radio frequency voltage is started by ringing the microwave tuning cavity with a fast rise-time pulse. The start or turn-on time of the microwave output is therefore determined largely by the characteristics of the microwave resonator. The applied pulses presently have rise-times of less than 1 nanosecond.

The maximum theoretical power output from an LSA device is determined largely by the active bulk volume. The thermal conductivity of the present LSA material, GaAs, is $0.81 \text{ W/cm}^2 \text{ } ^\circ\text{C}^{-1}$ for good material, which is about $1/2$ that of silicon. However the heat in an LSA device is not generated in a small volume junction which can be located less than 1 micron from a good heat sink, such as in the case of the IMPATT diodes. GaAs LSA devices therefore have the problem of adequately removing heat from the bulk of the device and present devices have been limited to short pulse length operation. By physically forming the devices to provide better heat transfer to a heat sink, longer pulse lengths and, consequently, greater average power can be obtained.

It has been predicted that LSA devices will deliver several hundred kilowatts peak power within several years. A main limitation is one of materials technology. In order to create and quench the space-charge near the cathode of the device, the dielectric relaxation times of the carriers for fields, below and above the threshold, must be adjusted for the desired frequency of operation. The relaxation times are a function

of the doping of the material and a limitation on the n/f ratio arises. (Here, n is the doping/cm³, and f is the desired frequency.) It can be shown that $2 \times 10^4 < (n/f) < 2 \times 10^5$, in order for LSA modes to occur, meaning that the doping for an X-band device must be in the neighborhood of 10^{14} cm⁻³. With the present state of the art of GaAs technology it is difficult to obtain large pieces of material with these doping levels and the carrier mobility required for LSA devices. When they become available, it will be possible to make devices in a physical shape with the thermal characteristics required for production of higher peak powers and much longer pulse lengths.

Preliminary noise investigations on LSA devices have shown that FM and AM noise figures are better by 10 to 20 db than those available from IMPATT devices.

STATE OF THE ART FOR LSA, IMPATT, AND GUNN MICROWAVE SOURCES

LSA, IMPATT, and Gunn oscillators have all demonstrated considerable power generation capability, especially for pulsed operation.

Current state of the art for solid-state devices is shown in the Table.

IMPATT Oscillators

Ultimate performance of IMPATT oscillators remains in doubt; many of the current results are being obtained from off-the-shelf varactor diodes which are not optimized for IMPATT operation.

An IMPATT oscillator circuit built at Hughes in microwave circuitry exhibited the characteristics shown in Figure A.

An analysis of the anticipated noise performance when used in amplifiers predicts noise figures of 37 to 40 db. Accordingly, it is anticipated that IMPATT oscillators will be relatively noisy. Brand,¹ et al., have observed a relatively high noise level in gallium arsenide oscillators. They note a significant increase in the noise figure of a single-ended mixer using the IMPATT oscillator as a local oscillator when compared with results using a klystron. However, in a balanced mixer configuration, they are able to duplicate the performance obtained with the klystron. Also by injection locking IMPATT oscillators considerable reduction in noise can be achieved as is shown in Figure B.

In summary, IMPATT oscillator characteristics are as follows:

1. Relatively high power-frequency characteristic
2. Frequency tuned by external circuit
3. Electronically tuned by bias current
4. Relatively high noise content in output

LSA Oscillators

The performance evaluation of the LSA solid state sources includes their promise for very high power but the large difficulty in material and cooling.

Gunn Oscillators

The state-of-the-art for power generation by Gunn oscillators is shown in the Table. Again it is difficult to anticipate ultimate performance from these relatively new devices. However, speculative estimates of pulsed powers range from 10 kw at L-band to 100 watts at X-band.

¹ Brand, F. A., et al., "Performance Characteristics of CW Silicon and GaAs Avalanche Diode Oscillators," G-MTT Symposium, Palo Alto, California, May 16-19, 1966.

Published Performance of Solid State Microwave Devices

LSA		GUNN		IMPATT	
Pulse	CW	Pulse	CW	Pulse	CW
<ul style="list-style-type: none"> 615 W. peak 2.3% efficiency 100 ns, 60 cps 7.7 GHz (Cornell) 	<ul style="list-style-type: none"> 20 mw 44-88 GHz 2% efficiency (BTL) 	<ul style="list-style-type: none"> 143 W. peak 19% efficiency 2.2 GHz 55-200 ns 60 cps (RCA) 	<ul style="list-style-type: none"> 110 mw 11 GHz 3% efficiency (BTL) 	<ul style="list-style-type: none"> 16 W. peak 0.1% duty 4.5% efficiency (Hughes) 	<ul style="list-style-type: none"> 1.2 w. (9.6% effective) 10.7 GHz (Hughes)
<ul style="list-style-type: none"> 3 W. peak 10 s 60 cps 50 GHz (Cornell) 		<ul style="list-style-type: none"> 1 GHz 20% efficiency (Varian) 	<ul style="list-style-type: none"> 43 mw 9-11 GHz 1.8% efficiency (RCA) 	<ul style="list-style-type: none"> 2 W. peak X-band 1.0% efficiency (Hughes microstrip) 	<ul style="list-style-type: none"> 544 mw 3% efficiency 10.2 GHz (Hughes microstrip)
<ul style="list-style-type: none"> 85 W. peak 100 s 60 cps (Gayaga diode, Hughes) 				<ul style="list-style-type: none"> 28 W. peak 3% efficiency 8 GHz 1 s 1 KHz (M. A.) 	<ul style="list-style-type: none"> 1.1 W. 7.7% efficiency 12 GHz (BTL)
<ul style="list-style-type: none"> 50 W. peak 6% efficiency 100 ns 60 cps 8 GHz (Varian) 				<ul style="list-style-type: none"> 435 W. peak* 22% efficiency 425 MHz 10% duty (RCA) 	
<ul style="list-style-type: none"> 200-100 mw peak 9% efficiency 50 GHz (BTL) 				<ul style="list-style-type: none"> 180 W. peak* 60% efficiency 775 MHz 10% (RCA) 	

* Characterized as an "Anomalous IMPATT," Theory and predicted performance are not well understood. It is not known whether or not this type of diode can be scaled or will perform at X-band.

STATE OF THE ART FOR LSA, IMPATT, AND GUNN MICROWAVE SOURCES

Evaluation

Currently, the consensus of opinion seems to favor the IMPATT oscillator as the most useful on the basis that its power frequency characteristic is higher, its frequency can be adjusted by external circuitry and bias current, and the width of a depletion zone can be made extremely small with relative ease as compared to the thickness of bulk samples.

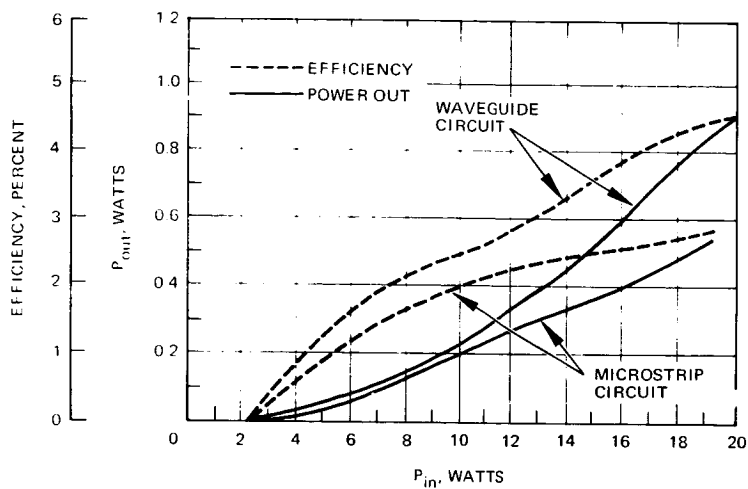


Figure A. CW Power Output and Efficiency Versus Power Input

For comparison the performance is shown for diodes from the same batch used in a waveguide circuit and a microstrip circuit operated at X-band.

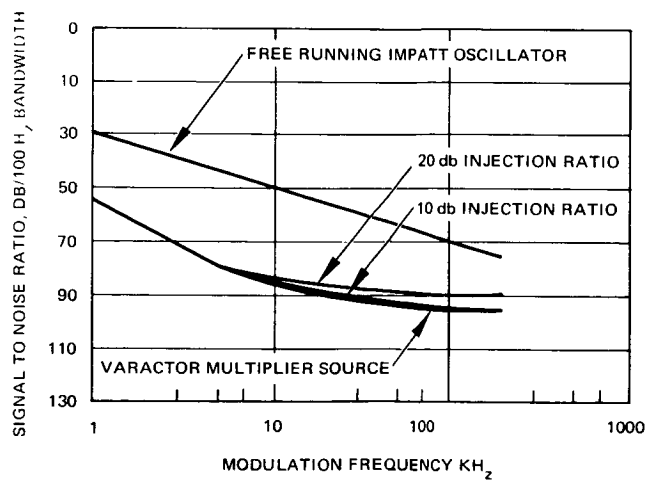


Figure B. FM Noise Characteristics of Injection Locked Microstrip IMPATT Power Source

RADIO FREQUENCY BURDEN RELATIONSHIPS

Cost and weight burden values are given for 2.3 GHz carrier frequency

A main purpose of this contract (NAS 5-9637) is to compare the performance of several communications systems operating at different wavelengths. In order to do this an extensive modeling was undertaken which expressed parameters in the communication link equation in terms of cost or weight. (See Appendix A of this Volume.)

From the material given in this Part; Part 1, Transmitting Power Sources; and from other investigations, constants have been chosen which relate the transmitted power, P_T , to the weight of the transmitter, WP_T , and to the cost of the transmitter, CP_T .

The values used in these relationships are the best that could be determined at the date of this final report and are certainly subject to change. This is especially true in the cost relationships which represent estimates of fabrication costs only and do not include development costs.

Figure A gives the expected weight of a 2.3 GHz transmitter and Figure B gives the expected cost, both given as a function of transmitted power, P_T .

The efficiency of a 2.3 GHz source is taken as 25 percent.

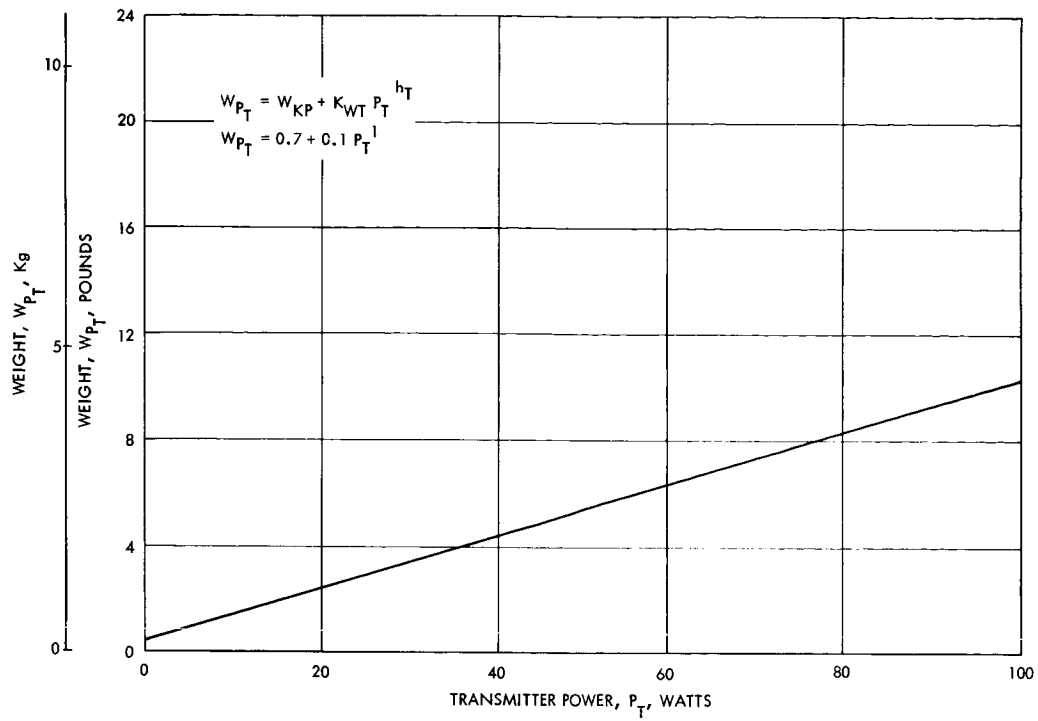


Figure A. Weight of S-Band Transmitting Power Source

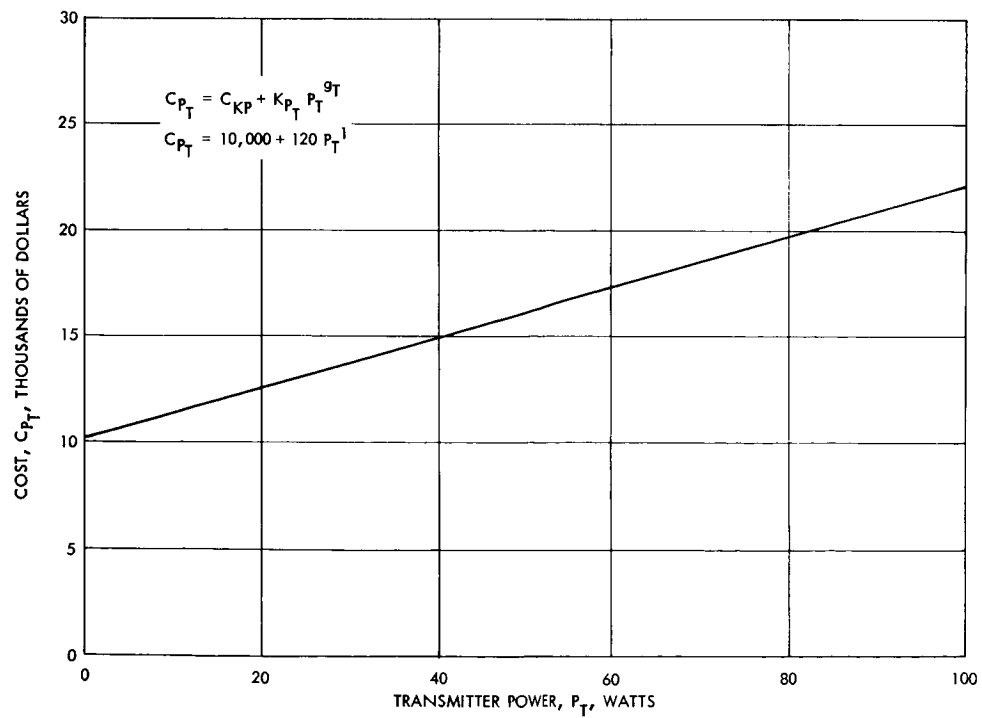


Figure B. Cost of S-Band Transmitting Power Source

TRANSMITTING POWER SOURCES

Optical Frequency Source Characteristics

	Page
Introduction	46
Laser Operating Fundamentals	48
The Argon Laser, Excitation Process	50
Laser Amplifier Gain	54
Laser Power Output and Efficiency	58
CO ₂ Laser Excitation Process	64
CO ₂ Laser Frequency Spectrum	70
CO ₂ Laser Scaling Laws	72

INTRODUCTION

The laser sections describing Theory, scaling laws, mode coupling, AFC, and state of the art are introduced.

Optical frequency sources, in the form of gaseous and solid state lasers, are discussed in the following three sections. The purpose of this discussion is to provide the designer with an understanding of laser sources, some of the problems that must be solved for application to communications, means for solving these problems and finally the state of the art for several laser implementations. A brief summary of each of the major areas of interest is given below.

Fundamentals

Indicates the basic construction and operation of lasers.

Theory

The theory of laser operation is known with varying degrees of completeness. Some phenomena can be described quite well while others require further study and experimentation.

Scaling Laws

These have resulted from the theory and provide general guidelines which relate laser physical parameters with output power.

Mode Coupling

Lasers tend to operate in a number of oscillation modes simultaneously. Since some communication applications require a single oscillation, means for obtaining such performance is discussed.

AFC

Frequency stabilization is important in lasers operating in a communication system using heterodyne or homodyne detection. The stabilization of the laser frequency is extremely high compared to normal microwave requirements. Methods of stabilizing and effects of environment or stability are given.

State-of-the-Art

The state-of-the-laser-art is given in terms of power art, stability obtained and empirical relationships derived for the methodology which relate laser weight and cost to power out.

LASER OPERATING FUNDAMENTALS

Basic laser configurations and components are illustrated.

The basic requirements for a laser are twofold: A suitable material and a source of energy. The materials that have been used include solids, liquids and gases. Forms of energy that have been used include dc energy, r. f. energy, heat energy, and light energy.

A laser material is excited by the externally applied energy on a molecular or atomic level. When the excited atoms return to a lower energy level a discrete amount of energy is emitted. This energy is in the form electromagnetic energy of a fixed wavelength.

Many means have been developed to increase the amount of energy from the laser. A prominently used method is to put mirrors at each end of the laser material. The mirrors reflect the laser light through the laser material several times. In this way there is built up in the laser cavity of a relative high energy at the laser wavelength. One of the two mirrors is partially reflecting to allow the output energy to be coupled out of the laser. Figure B illustrates a typical configuration.

The cavity formed by the two mirrors has a ratio of stored energy to output energy or Q . Generally if the Q is too low no laser action results. It is possible then to change the Q to stop and start the laser action. Two common ways of varying the Q are to insert an electroptic switch (Kerr cell) in the laser cavity or to rotate one of the mirrors. This is often called "Q-switching". And can be used to generate very narrow pulses which start at a known time. Figure C illustrates this.

Laser modulation may be achieved either within or without the laser cavity as is indicated in Figures D and E. Internal modulation has the advantage of the laser energy passing through the modulator more than once and thus multiplying its effectiveness but has the disadvantage of having a larger amount of energy passing through the modulator (the circulating energy is Q times the output energy) and thus heating problems must be solved.

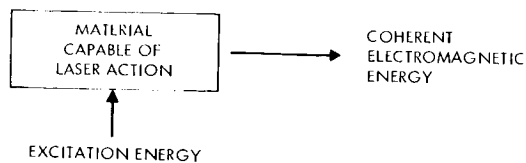


Figure A. Basic Requirements for Laser Action

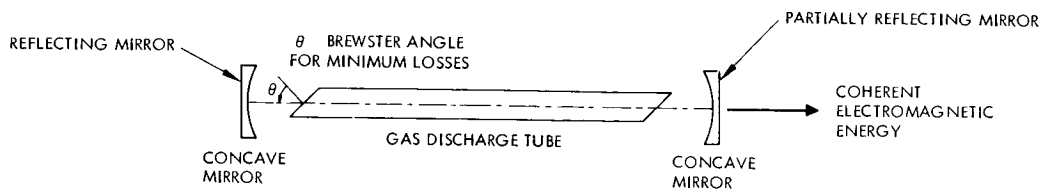


Figure B. External Confocal Mirror Laser Employing Brewster - Angle Windows

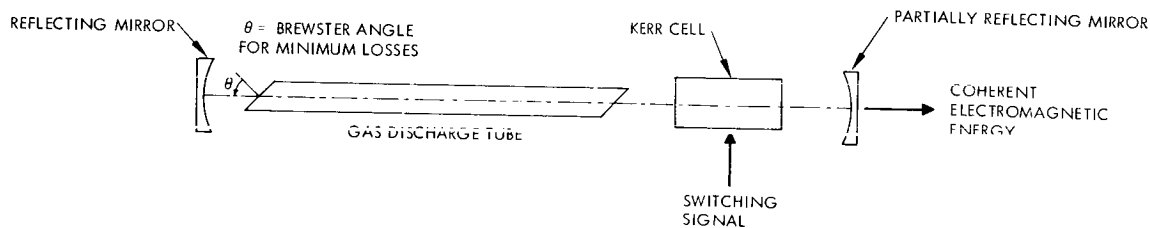


Figure C. Laser Employing Kerr cell for "Q-Switching"

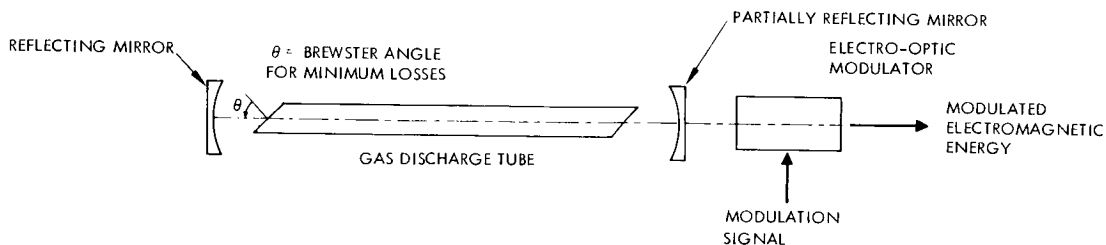


Figure D. External Laser Modulation

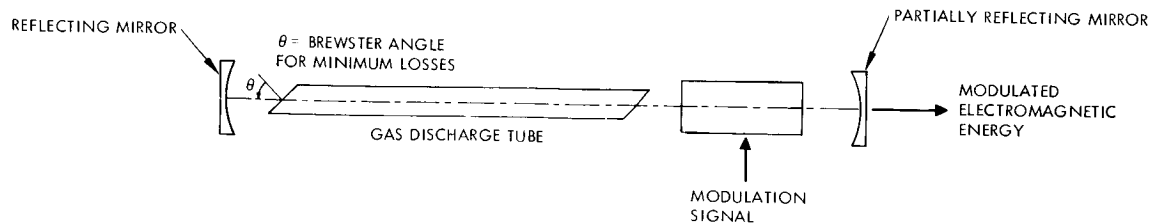


Figure E. Internal Laser Modulation

THE ARGON LASER, EXCITATION PROCESS

Ion lasers theory may be described in terms of excitation processes.

The model which seems to fit best the observed behavior of the argon ion laser is a three-level system originally proposed by Gordon, Labuda, and Miller. In this model, the upper laser level is populated by two successive electron collisions; the first produces an unexcited ion from a neutral atom, and the second excites the ion to the upper laser level. This two step process is consistent with the observed I^2 dependence of spontaneous emission from the singly ionized laser upper levels and the I^4 dependence of power output. The depopulation of the lower laser level then occurs by vacuum ultraviolet radiation to the ion ground state; this suggests a lifetime for the lower level roughly $(\lambda_{\text{vac}} U F / \lambda_{\text{laser}})^3$ shorter than the upper level.

The population processes are shown schematically in Figure A. It is not known directly whether the atom makes a "round trip" to the neutral ground state for each laser photon emitted. (This is an important factor in the laser's efficiency, since the electron energy that goes into the creation of the ion is essentially wasted.) However, the observed I^2 and I^4 dependences seem to indicate that it does. There is also some evidence that ionic metastable levels may play a role in the second electron collision, so that the picture of the three level system shown may be oversimplified.

The ultraviolet radiation which depopulates the lower level can be increasingly trapped as the ion density builds up. The "dead" region which develops at higher currents is a region of high attenuation rather than gain and is caused by radiation trapping. The gain is quite sensitive to the trapping coefficient* which, in turn, depends on the gas temperature. As the current pulse persists, the gas temperature rises and the trapping decreases, producing an inversion and oscillation for the remainder of the pulse, even if it extends to continuous operation. For lower currents or shorter pulse lengths this effect is not observed and the laser pulse follows the current pulse exactly.

Bennett, et al., have proposed that direct electron excitation from the neutral atom ground state to the upper laser level is the dominant populating mechanism. The preponderance of laser lines with p upper states follows from the selection rules of the "sudden perturbation" process referred to by Bennett. The short radiative lifetimes of the s and d lower levels then guarantee an inversion. From the evidence available at the present time, it seems likely that this population mechanism is dominant only in pulsed discharges with high E/p and short pulse

* Briefly, when trapping is included in the rate equations, the current-independent part of the expression for the population inversion appears as the small difference of two large terms, one of which contains the gas temperature. Changes in the gas temperature can then swing the population difference from attenuation to amplification.

duration; these are not conditions conducive to high efficiency or high average power. If very short pulse lasers (nanoseconds) are desired, then this mode of excitation may prove interesting.

In general, we may say that the excitation and de-excitation mechanisms are reasonably well identified for cw argon ion lasers. However, quantitative measurements of the relative importance of the various processes and construction of a numerically accurate mathematical model will require further detailed experimental study.

A 5 watt Argon laser is shown partially disassembled in Figure B, with the laser cavity removed from the solenoid. Figure C shows a similar laser mounted in a fixture with end mirrors.

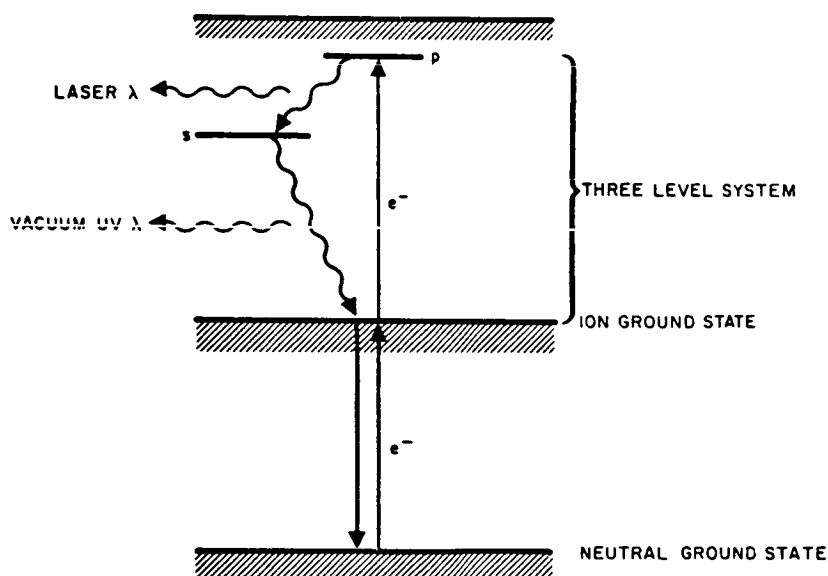


Figure A. Schematic Energy Level Diagram and Processes for Singly Ionized Atoms

Transmitting Power Sources
Optical Frequency Source Characteristics

THE ARGON LASER, EXCITATION PROCESS

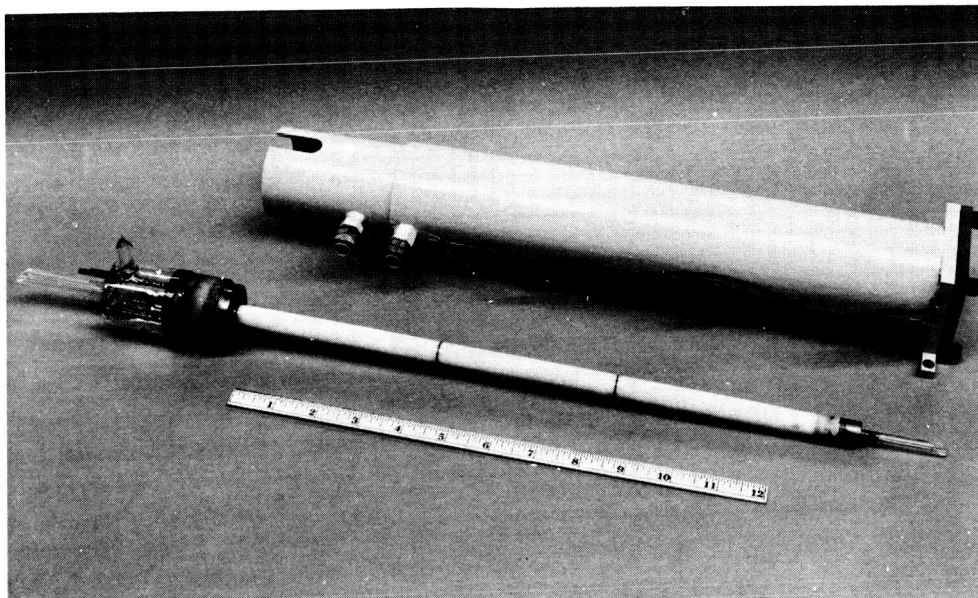


Figure B. Five Watt Argon Laser With its Solenoid

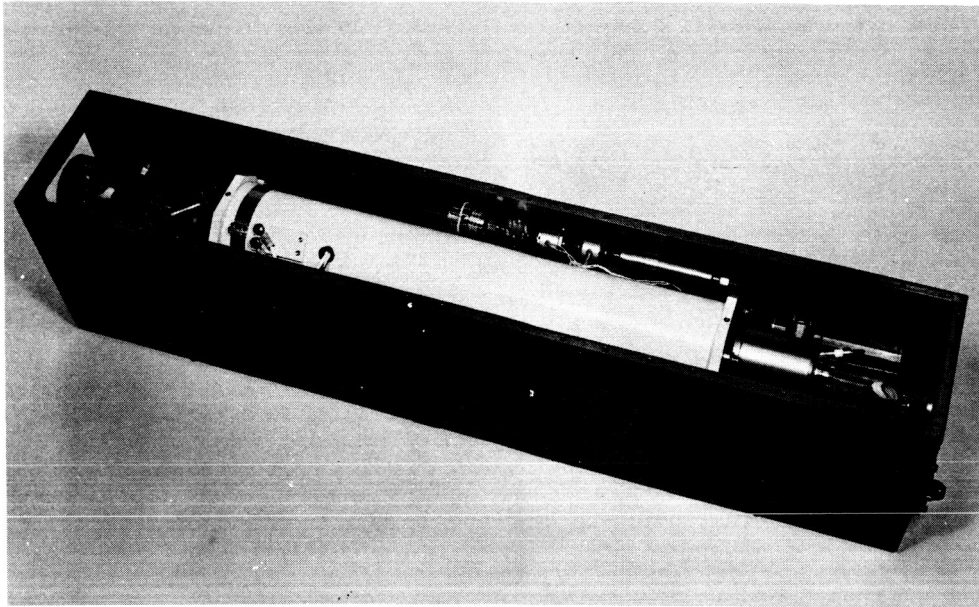


Figure C. Five-Watt Argon Laser Mounted Between Mirrors

LASER AMPLIFIER GAIN

The gain of a laser amplifier is given and it is to the Doppler broadened laser line is discussed.

The gain of a laser amplifier may be expressed as

$$G(S) = 1 + \frac{g_o L}{(S/S_{3dB})^n} \quad (1)$$

where $G(S)$ is the actual gain in a single pass of a signal of intensity $S(\text{w/m}^2)$ through a laser of small signal gain coefficient $g_o (\text{m}^{-1})$ and length L . The power level S_{3dB} is the signal power density at which the small signal gain has been reduced by 3 dB (in the case $n = 1$). The exponent n equals 1 for homogeneous interaction (that is, interaction in which the signal may, with equal probability, interact with any atom in the laser), and is equal to 1/2 for inhomogeneous interaction (that is, the interaction may take place between the wave and only a small fraction of the excited atoms).

Equation (1) is valid of $S \gg S_{3dB}$ and if $G(S)$ is not too large (i. e., if the approximation

$$\exp [G(S)] \approx 1 + G(S) \quad (2)$$

is valid). Note this does not require that $g_o L$ be close to unity.

Low power gas lasers usually fall into the category of inhomogeneous interactions; since the Doppler line width is so much greater than the natural line width, radiation at a single frequency can interact only with a small fraction of the atoms, e. g., those whose Doppler-shifted frequency falls within one natural line width of the incident radiation. We say this radiation "burns a hole" in the Doppler line when it depletes the excited atoms available to it without affecting the remainder.

Figure A contrasts the cases of (a) a hole burned by inhomogeneous interaction and (b) the depletion of the entire line by true homogeneous interaction. This description in terms of inhomogeneous interaction remains valid for the gas laser even when many frequencies are present (multimode operation), provided the holes do not "overlap" sufficiently (Figure B). With sufficient overlapping, the gain line is effectively "burned off" (Figure C) and the laser behaves as if the interaction were homogeneous. Essentially, every atom can interact with some radiation in the closely spaced multimode case.* Even if the modes are not closely spaced, the line will be burned off and behave as if the interaction were

*Note that this is a physically different effect from the case of a homogeneously broadened line that typically occurs in solid state lasers (ruby, for example). In the case of a homogeneously broadened line, every atom can interact with a single frequency signal. The net effect on the gain saturation is the same, however.

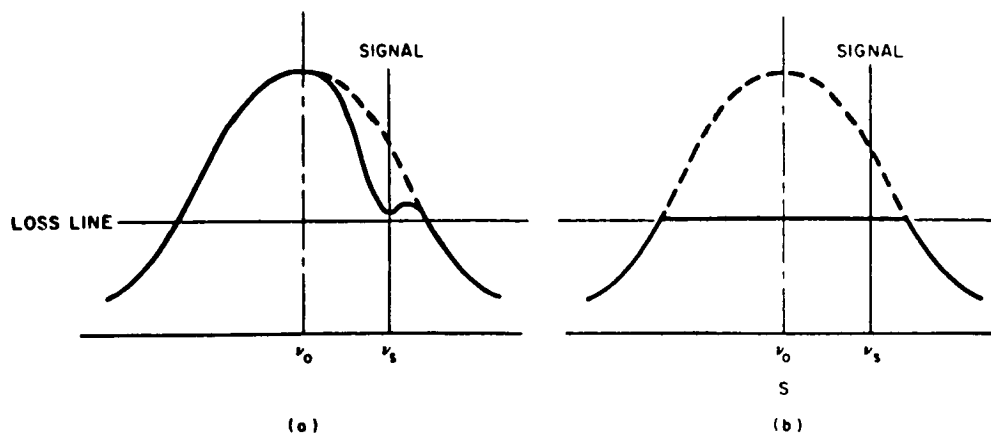


Figure A. Comparison of (a) Inhomogeneous Interaction (Hole Burning) and (b) Homogeneous Interaction With a Single Frequency

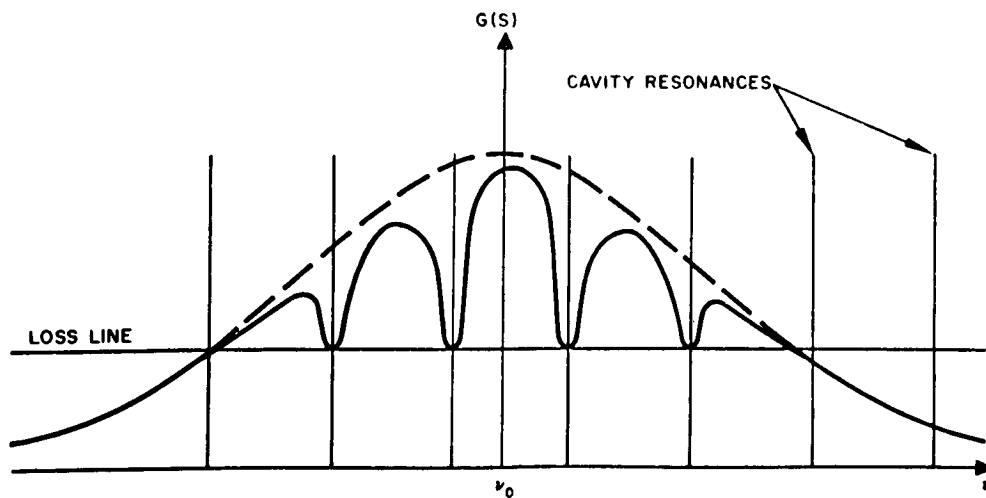


Figure B. Nonoverlapping Holes Burned in the Gain Line by a Multimode Laser

LASER AMPLIFIER GAIN

homogeneous, provided the signal intensity becomes sufficiently high. This is a result of the Lorentzian shape of the hole: it has broad "wings" which do not fall off rapidly from the hole center (not nearly so fast as the Gaussian, for example). Thus a gas laser capable of high power operation under multimode conditions will exhibit inhomogeneous interaction at low signal levels and homogeneous interaction at high signal levels.

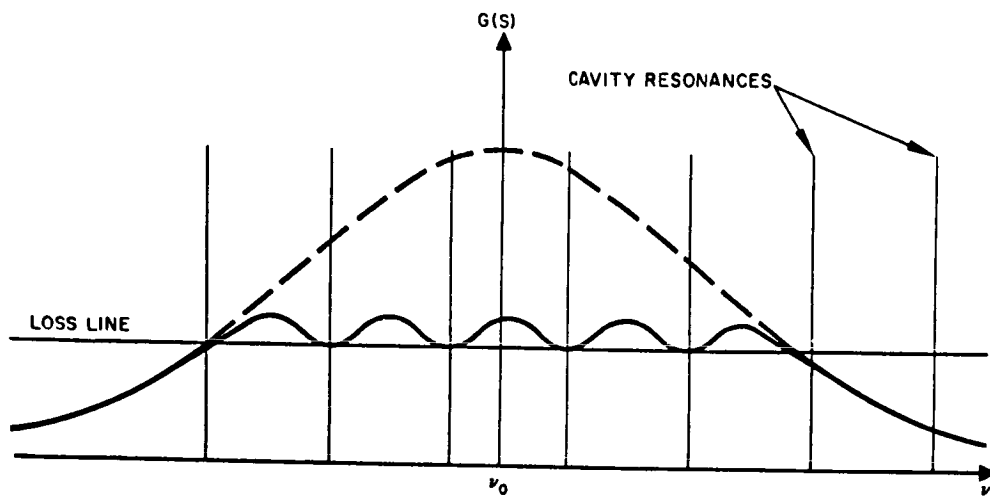


Figure C. Gain Line "Burned Off" by a Multimode Laser

LASER POWER OUTPUT AND EFFICIENCY

Qualitative expressions are derived for laser output power and laser efficiency.

Power

The condition for steady state oscillation is

$$R G(S) = 1$$

where

R is the fraction of the total laser power that is reflected back into the laser cavity

$G(S)$ is the amplifier gain in a single pass of a signal with intensity S .

From the previous topic,

$$S = S_{3dB} \left(\frac{R g_o L}{1 - R} \right)^{1/n} \quad (1)$$

Assuming further that $R g_o L / (1 - R)$ is $\gg 1$ (which will be true for the important ion laser lines), equation (1) simplifies to

$$S \sim \begin{cases} g_o L & \text{homogeneous interaction} \\ g_o^2 L^2 & \text{inhomogeneous interaction} \end{cases} \quad (2)$$

where S is the power circulating in the cavity. The output power will be $S(1 - R - \text{losses})$. R and the losses are assumed to be constant.

To relate equation (2) to the discharge conditions, it is necessary to know the dependence of gain on the current. Spontaneous emission measurements show that the number density in both upper and lower levels N_2 and N_1 (m^{-3}) vary as J^2 (J is the current density in amperes per square meter)

$$N_2 - N_1 \sim J^2 \quad (3)$$

The "constant" of proportionality actually contains some implicit current dependence (viz., the effect of gas heating on the Doppler line width and the effects of radiation trapping on N_1 caused by the current dependence

of the ion density N_0). However, if this variation may be neglected for the moment, equation (3) gives the major variation of inversion with current. The gain coefficient of a laser discharge is approximately proportional to the inversion

$$g_0 \sim N_2 - N_1 \quad (4)$$

except very near threshold, where it varies more as the $3/2$ power

$$g_0 \sim (N_2 - N_1)^{3/2} \quad (5)$$

(Actually, this is very approximate; the actual dependence is not expressible as a simple exponent.) For a discharge tube of diameter D , the power output (watts) and current (amperes) is given by

$$P = \frac{\pi D^2 S}{4} \quad (6)$$

$$I = \frac{\pi D^2 J}{4} \quad (7)$$

The above equations assume uniformity of S and J in the radial direction. This is probably a good assumption for J but not for S ; however, to first order only the numerical coefficient (the cavity mode filling factor) will change and $P \sim D^2 S$ for a given cavity. Combining (2) through (7) yields:

$$P \sim \left\{ \begin{array}{l} D^2 g_0 L \sim D^2 (N_2 - N_1) L \sim D^2 J^2 L \sim \frac{I^2 L}{D^2} \text{ saturated} \\ \\ D^2 g_0^2 L^2 \sim \left\{ \begin{array}{l} D^2 (N_2 - N_1)^2 L^2 \sim D^2 J^4 L^2 \sim \frac{I^4 L^2}{D^6} \text{ moderate levels} \\ \\ D^2 (N_2 - N_1)^3 L^2 \sim D^2 J^6 L^2 \sim \frac{I^6 L^2}{D^{10}} \text{ near threshold} \end{array} \right. \end{array} \right. \quad (8)$$

LASER POWER OUTPUT AND EFFICIENCY

The existence of $I^6 \rightarrow I^4 \rightarrow I^2$ behavior has been well verified in this laboratory; however, the exact limits of each regime are not well known as a function of the other parameters of the laser discharge (that is, those parameters which make up the constant of proportionality). The Figure indicates, in a highly schematic way, the expected variation of power output with current as the discharge length L and the cavity length ℓ are varied. The assumptions made are, of course, that $c/2\ell > \Delta\nu_N$ and $\ell > L$.

When a magnetic field is used to confine the plasma, the scaling is slightly different. Since the rate at which ions leave the discharge by diffusion to the walls is reduced by the magnetic field, the rate at which plasma is generated must also decrease in order to maintain an equilibrium state. The plasma generation rate is proportional to the electron density and the average electron energy, and the new equilibrium is maintained by a decrease in the average electron energy. The longitudinal electric field is roughly proportional to the electron energy, and it is the decrease in the longitudinal electric field (and the corresponding decrease in tube voltage) which accounts for the improvement in the efficiency of the discharge.

However, since the excitation of ions to the upper laser level is by electron-ion impact, the excitation rate is proportional to the plasma density squared and the average electron energy. The laser performance should be enhanced in a magnetic field because of the increased plasma density and efficiency of the discharge, this improvement should eventually be balanced or reversed by the decrease in the average energy of the plasma electrons.

Magneto-optical effects can further complicate the dependence of laser performance on the magnetic field.

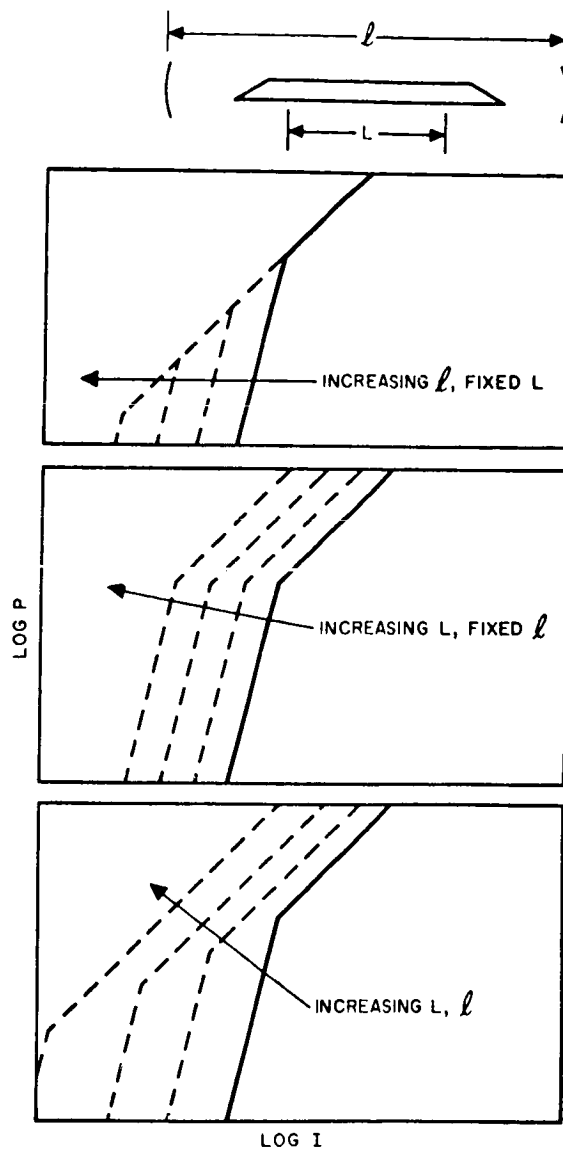
It is quite clear that one of the major first steps toward improving the power output and efficiency of the ion laser is to make a thorough and careful set of measurements of gain, gain saturation, power output, and efficiency with the physical parameters I , L , ℓ , D , B , and pressure varied over as wide a range as is practicable. Such measurements will confirm, explore the limits of validity of the theory discussed here and provide a set of design criteria for future laser development.

Efficiency

The variation of efficiency with the major parameters may also be estimated. The power input to the discharge is approximately

$$P_{in} = IEL + IV_{cath}$$

where E is the average longitudinal electric field in the positive column and is a function of the diameter and the magnetic field. V_{cath} is the cathode fall, a function only of the cathode type and material. (For hot cathodes, such as those used in our cw tubes, $V_{cath}=20$ to 40 V; for cold cathodes, V_{cath} may be several thousand volts.) The exact dependence



Schematic Representation of the Change in Power With Different Discharge and Cavity Lengths

LASER POWER OUTPUT AND EFFICIENCY

of E on diameter and magnetic field is not known. For the helium-neon laser the scaling relation $pD = \text{constant}$ results in the relation $ED = \text{constant}$. Observations to date indicate that in the Ar II laser (and in the neutral xenon 3.5μ laser) the variation of E is faster than $1/D$. The efficiencies corresponding to the three regions of operation may then be written

$$\eta \sim \left\{ \begin{array}{l} \frac{I}{ED^2 + \frac{V_{\text{cath}} D^2}{L}} \text{ saturated} \\ \frac{I^3 L}{ED^6 + \frac{V_{\text{cath}} D^6}{L}} \text{ moderate level} \\ \frac{I^5 L}{ED^{10} + \frac{V_{\text{cath}} D^{10}}{L}} \text{ near threshold} \end{array} \right.$$

The saturated region will give the region of highest efficiency. For tubes long enough that V_{cath}/L is $\ll E$, the saturated efficiency is independent of length, V_{cath} is typically 20 to 40 V, and E ranges from 0.5 to 20 V/cm, depending on the diameter and the magnetic field. With these numbers, it is apparent that the cathode fall will significantly degrade the efficiency in short tubes (20 cm or so).

CO₂ LASER EXCITATION PROCESS

The effects of various gas additives upon CO₂ laser performance is described.

This topic contains a discussion of the excitation processes for several types of CO₂ lasers. The types are distinguished by the gas additives to the basic CO₂ gas.

CO₂

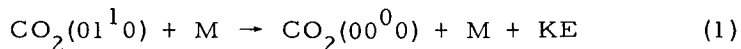
Laser action in CO₂ has been observed on both the P- and R-branches of the 00⁰₁ - 10⁰₀ and 00⁰₁ - 02⁰₀ vibrational bands and on the P-branch of the 01¹₁ to 03¹₀ vibrational band. The following discussion will be primarily concerned with the high power P-branch rotational transitions of the 00⁰₁ - 10⁰₀ vibrational band which occur in a narrow wavelength interval near 10.6μ. The Figure illustrates the pertinent energy levels for CO₂ and CO₂-N₂ lasers. For simplicity the rotational levels are not shown for each vibrational state. The vibrational levels at the left of the CO₂ energy level diagram are those of the symmetric stretching mode (ν₁), at the right are those of the asymmetric stretching mode (ν₃), and in the center are those of the doubly degenerate bending mode (ν₂).

Excitation of the upper laser level in the CO₂ laser may occur via recombination and/or cascade from higher levels, or by electron impact, either directly or via a compound. Experiments indicate that the dominant mechanism may vary as conditions of the discharge are changed.

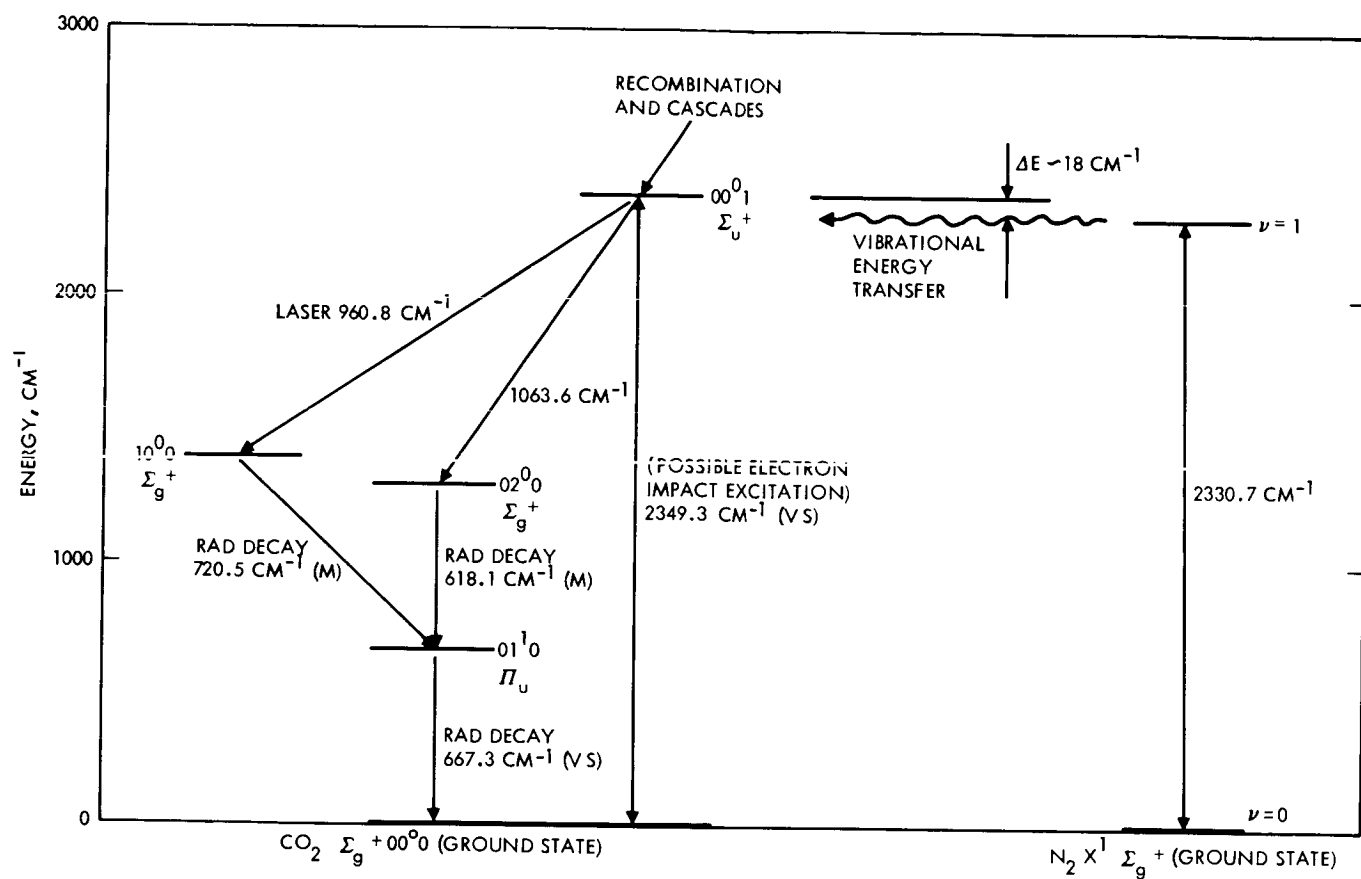
Decay of vibrationally excited CO₂ molecules is by spontaneous emission and/or collisional de-excitation. The calculated spontaneous radiative transition probabilities between the lower vibrational levels in CO₂ are listed in the Table.¹

It is evident that for transitions other than 00⁰₃ - 00⁰₍₃₋₁₎, spontaneous radiation is negligible. The lifetime of the 00⁰₁ level, in the presence of radiation trapping, is ~2 x 10⁻² seconds.

The amount of power produced by CO₂ lasers requires that the lifetime of the lower laser levels be 10⁻³ seconds. Consequently, de-excitation by CO₂-CO₂ collisions involving the conversion of vibrational energy to translational energy is the dominant relaxation mechanism. The levels of each vibrational mode reach internal thermal equilibrium via vibration-vibration exchange, then cross-relax to the fastest relaxing mode. In CO₂ the most probable vibration-translation relaxation mechanism involves vibrational quanta of the lowest frequency mode ν₂, i. e.,



¹Statz, H., Tang, C. L., and Koster, G. F., "Probabilities for Radiative Transitions in the CO₂ Laser System," presented at 1966 International Quantum Electronics Conference, Phoenix, Arizona, April 1966.



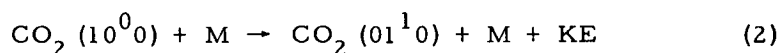
Energy Level Diagram Showing Pertinent Levels in CO_2 and N_2

CO₂ LASER EXCITATION PROCESS

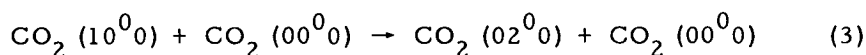
Spontaneous Radiative Transition Probabilities in CO₂

Transition	Dominant Branch	Spont. Trans. Prob.
00 ⁰ ₁ - 10 ⁰ ₀	P	0.34 sec ⁻¹
00 ⁰ ₁ - 02 ⁰ ₀	P	0.20 sec ⁻¹
10 ⁰ ₀ - 01 ¹ ₀	Q	0.53 sec ⁻¹
02 ⁰ ₀ - 01 ¹ ₀	Q	0.48 sec ⁻¹
01 ¹ ₀ - 00 ⁰ ₀	Q	1.07 sec ⁻¹
00 ⁰ ₁ - 00 ⁰ ₀	P	2 x 10 ² sec ⁻¹

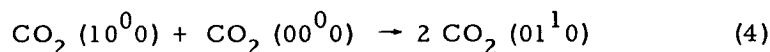
where M is CO₂ or another species. The corresponding relaxation rates to the ground state for molecules in the lowest vibrational levels of the ν_1 and ν_3 modes, 10⁰₀ the lower laser level, and 00⁰₁ the upper laser level, respectively, are several orders of magnitude slower. The lower laser level (10⁰₀) relaxes to 01¹₀ by



or, since the 10⁰₀ and 02⁰₀ levels are in Fermi resonance, by



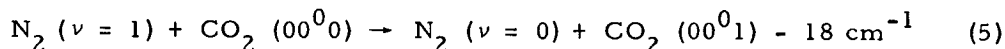
or



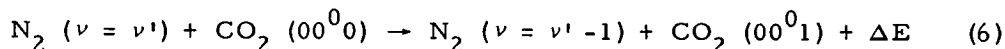
In each case, the final relaxation process is governed by (1). Calculations indicate that the rates for processes (1) and (2) are of the same order of magnitude, although the latter is somewhat slower, while the rate for 02⁰₀ - 01¹₀ vibration-translation relaxation is faster. Thus, the de-excitation of the lower laser level 10⁰₀ is by vibration-vibration exchange with the bending mode which then relaxes via vibration-translation exchange to the ground state. The rate of this latter process controls the relaxation of the lower laser level.

CO₂ - N₂

The increase in output power from the CO₂-N₂ laser over that from the pure CO₂ discharge may be attributed to population of the upper laser level in CO₂ by transfer of vibrational energy from N₂ molecules in the $\nu = 1$ vibrational level of their electronic ground state. The selective excitation of the CO₂ molecule from its ground state to the 00⁰1 state takes place during a two-body collision involving a CO₂ ground-state molecule and a vibrationally excited N₂ molecule. From the Figure it is evident that the $\nu = 1$ vibrational level of N₂ at 2330.7 cm⁻¹ is in very close coincidence with the 00⁰1 vibrational level in CO₂ at 2349.16 cm⁻¹ ($\Delta E = 18.46 \text{ cm}^{-1} < kT_{\text{rotational}} \sim kT_{\text{translational}} \sim 210 \text{ cm}^{-1}$.) In addition, since N₂ has a zero permanent dipole moment, vibrational relaxation times for excited N₂ molecules are long and are determined by collisions with other molecules and walls. The addition of CO₂ allows excited N₂ molecules to relax via vibration-vibration exchange with 00⁰n₃ levels of CO₂. The process with the largest cross section is



For N₂ molecules in higher energy vibrational states



Because of the large energy difference ($\sim 950 \text{ cm}^{-1}$) between the lower laser level 10⁰0 in CO₂ and the $\nu=1$ vibrational level in N₂, the cross-section for excitation of ground state CO₂ molecules to the lower laser level through collisions with N₂ ($\nu=1$) is much smaller than that for the reaction described in equation (5). In addition, the excitation of CO₂ (00⁰0) molecules to the 10⁰0 energy level involves a reaction in which both transitions; i. e., $\text{N}_2 (\nu=1) \rightarrow \text{N}_2 (\nu=0)$ and $\text{CO}_2 (00^0 0) \rightarrow \text{CO}_2 (10^0 0)$, are optically forbidden. It has been shown² that, for reactions involving collisions of the second kind, the cross-section for a reaction involving two optically forbidden transitions is smaller than that involving one forbidden transition for the same energy level discrepancy.

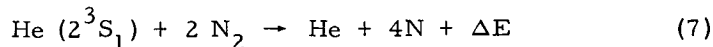
CO₂-N₂-He

A significant increase in power over that produced by CO₂ and CO₂-N₂ lasers is achieved by the addition of relatively large amounts of He to either discharge. At this time the role of the He has not been clearly defined although it has been determined experimentally that the He affects

²Bates, D. R., "Collision Processes not Involving Chemical Reactions," Discussions Faraday Soc., 33, p. 7, 1962.

CO₂ LASER EXCITATION PROCESS

the excitation of the upper laser level as well as the relaxation of the lower laser level. The former effect may result from the creation of plasma conditions (e. g., electron temperature) which favor excitation of the 00⁰1 level in CO₂. In addition there is the possibility of a three-body collision involving the transfer of energy from metastable He(2³S₁) atoms to N₂ molecules, viz



where $\Delta\text{E} < 0.1$ eV. The atomic nitrogen, thus produced, recombines to form vibrationally excited nitrogen in high vibrational levels which excite CO₂ molecules to the 00⁰1 vibrational level by the processes described in Equations (5) and (6).

The addition of He has little effect on the decay of the upper laser level 00⁰1. However, the rate for vibration-translation relaxation of CO₂ via the bending mode is dependent on the reduced mass of the colliding particles and is thus greater for CO₂-He collisions than for those involving two CO₂ molecules. The former collision process is one to two orders of magnitude more efficiency for relaxation.³ Thus, it appears that improved laser performance due to the addition of He may be due in part to an increased relaxation rate of the terminal laser level.

An entirely different role may be played by the He in that it is able to reduce the gas temperature because of its high thermal conductivity. A low gas temperature leads to a reduction in thermal excitation of the lower laser level, which may be appreciable since the energy is only 0.17 eV, as well as an increase in the gain coefficient for the laser transitions.

Other Gas Additives

The addition of other gases (air, CO, H₂, water vapor, etc.) influences the output from the CO₂ laser in varying degrees; however, none is as effective as He. In the case of H₂ and water vapor it is felt that the primary effect is to enhance the relaxation rate of the terminal laser level by collisions. Air is thought to combine the roles of N₂ and water vapor. CO, which has a long relaxation time, may selectively excite CO₂ by vibration-vibration exchange with asymmetric stretching mode ν_3 .

Effect of Gas Temperature

The lower level of 10⁰ of the laser transition lies sufficiently close to the ground state (~ 0.17 eV) that a significant fraction of the CO₂ molecules may be thermally excited to that state if the gas temperature becomes too high. This effect may be quite appreciable in large diameter tubes.

³Patel, C.K.N., Tien, P.K., and McFee, J.H., "Cw High-Power CO₂-N₂-He Laser," Appl. Phys. Letters, 7, p. 290, 1965

Under the assumption that laser action will cease when 10 percent of the ground state molecules are thermally excited to the lower laser level, the cut-off temperature is 662°C . With a power input of 0.1 W/cm^3 and in the absence of any cooling, this temperature can be reached and laser action cease in times of the order of 0.1 second. Under steady-state conditions in which cooling is provided at the discharge tube walls, thermal excitation of the lower laser level will present oscillation along the laser axis for diameters greater than approximately 5 inches.

With the assumptions that the laser transition is predominantly Doppler-broadened and that the rotational level populations are described by a Boltzmann distribution at a temperature $T_{\text{rotational}} \approx T_{\text{translational}}$, it can be shown that the gain coefficient of the laser transitions is inversely proportional to the molecular temperature, i. e.,

$$g_o \propto (T_{\text{translational}})^{-3/2} \quad (8)$$

The validity of the latter assumption is supported by the fact that rotational thermalization times are of the order of microseconds while in CO_2 the vibrational and radiative lifetimes are of the order of milliseconds and seconds, respectively. In addition, increased laser gain and output power have been observed experimentally by cooling the walls of the discharge tube.

CO₂ LASER FREQUENCY SPECTRUM

The transitional frequencies for a CO₂ laser are tabulated.

Vibrational-Rotational Transitions

The transitions of the P and R branches of the $00^0_1 - 10^0_0$ vibrational band of CO₂ on which cw laser oscillation has been observed are listed in the Table. All the lines do not oscillate simultaneously but are obtained by placing a dispersive element, such as a prism or grating, in the optical cavity. In the absence of such a wavelength selective resonator, several of the strongest P-branch transitions will oscillate at the same time.

Single Wavelength Operation

The gain of the CO₂ laser is sufficiently high that several rotational transitions may oscillate simultaneously. For single wave-length operation, wavelength selective cavities utilizing prisms or gratings will be required. Because of the rapid rotational thermalization ($\tau \sim 10^{-6}$ sec), molecules in the other rotational levels of the 00^0_1 level may cross-relax into the upper laser level of the rotational vibrational transition which is selected by the resonator configuration. Similarly, molecules in the lower laser level may cross-relax into other rotational levels of the 10^0_0 level. Consequently, the output power obtained under single wave-length operation may be close to the total power produced when several transitions oscillate simultaneously.

Single Mode Operation

The narrow Doppler width of the CO₂ laser transition (50-75 MHz) allows operation in a single axial mode with resonators 2 and 3 meters long. Single transverse mode operation in low power lasers (≤ 10 W) may be achieved by proper resonator design. However in very high power CO₂ lasers, transverse mode control may be complicated by the presence of self focusing and the production of a filamentary structure in the laser output.

CO₂ Continuous Wave Laser Oscillation Wavelengths
in the 00⁰1 to 10⁰0 Band of CO₂

Transition	Measured Frequency (cm ⁻¹)	Transition	Measured Frequency (cm ⁻¹)
P ₂	959.43	R ₂	963.33
P ₄	957.76	R ₄	964.74
P ₆	956.16	R ₆	966.18
P ₈	954.52	R ₈	967.73
P ₁₀	952.88	R ₁₀	969.09
P ₁₂	951.16	R ₁₂	970.50
P ₁₄	949.44	R ₁₄	971.91
P ₁₆	947.73	R ₁₆	973.24
P ₁₈ *	945.94	R ₁₈ *	974.61
P ₂₀ *	944.15	R ₂₀ *	975.90
P ₂₂ *	942.37	R ₂₂ *	977.18
P ₂₄ *	940.51	R ₂₄ *	978.47
P ₂₆	938.66	R ₂₆	979.67
P ₂₈	936.77	R ₂₈	980.87
P ₃₀	934.88	R ₃₀	982.08
P ₃₂	932.92	R ₃₂	983.19
P ₃₄	930.97	R ₃₄	984.35
P ₃₆	928.94	R ₃₆	985.42
P ₃₈	926.96	R ₃₈	986.49
P ₄₀	924.90	R ₄₀	987.56
P ₄₂	922.85	R ₄₂	988.63
P ₄₄	920.77	R ₄₄	989.61
P ₄₆	918.65	R ₄₆	990.54
P ₄₈	916.51	R ₄₈	991.47
P ₅₀	914.41	R ₅₀	992.46
P ₅₂	912.16	R ₅₂	993.34
P ₅₄	909.92	R ₅₄	994.18
P ₅₆	907.73		

* Strongest transitions in the group

CO₂ LASER SCALING LAWS

Scaling laws for power as a function of the laser tube diameter are given. Some early CO₂ laser developments are also documented.

Variation of Output Power

The variation of output power per unit volume and per unit length as a function of discharge tube diameter is shown in Figure A. The data points which are shown represent actual reported performance of laser oscillators as achieved by Hughes, Raytheon, Perkin-Elmer, Bell Telephone Laboratories, and C. G. E. in France. The shaded areas represent an estimate of the uncertainty implied with the dark curve through the center representing a best estimate for use in establishing design criteria. Because of the scatter of data points in the 6 to 9 cm range, extrapolation of the power per unit length parameter is difficult. The power per unit volume estimates are more consistent and may indicate the existence of some asymptotic value.

Variation of Input Power

In Figure B, the axial electric field along the discharge tube for optimum laser performance is plotted as a function of tube diameter. As expected, it decreases rapidly with diameter until the loss of ions to the walls become negligible. It is apparent that the field is approaching an asymptote of approximately 30 volts/cm. In Figure C, the optimum discharge current is plotted as a function of tube diameter. Although difficult to prove conclusively from the data it is expected that this current is increasing with a dependence very close to (diameter)². These conditions lead to an asymptotic value for input power of approximately 0.1 watt per cm³ as the diameter is increased beyond 10 cm. Since high efficiency operation (i.e., approximately 18 percent) has been claimed for diameter of 6 centimeters it is reasonable to assume that an output power of 10 to 20 milliwatts per cm³ is a possibility. This is consistent with the graphic extrapolation shown in Figure A.

Modes of Operation

CW

As illustrated in Figures B and C, the CO₂-N₂-He laser is a relatively low current, high voltage discharge. These features, plus the fact that heated oxide coated cathodes are extremely susceptible to poisoning by this type of discharge, suggest the use of cold cathode discharges. Both dc and ac excitation can be used. Although the latter method produces slightly higher power, it has several disadvantages. The light output is amplitude modulated at twice the ac excitation frequency, and, because the discharge is self-striking, the current, and hence the output, exhibits random fluctuations.

By using rf excitation, the electrodes can be removed from the discharge altogether. With small diameter discharge tubes (<2.5 cm), rf excited lasers have exhibited essentially the same output power, efficiency and pressure dependence as comparable dc excited tubes. However, the latter have two distinct advantages: (1) no rf interference in auxiliary

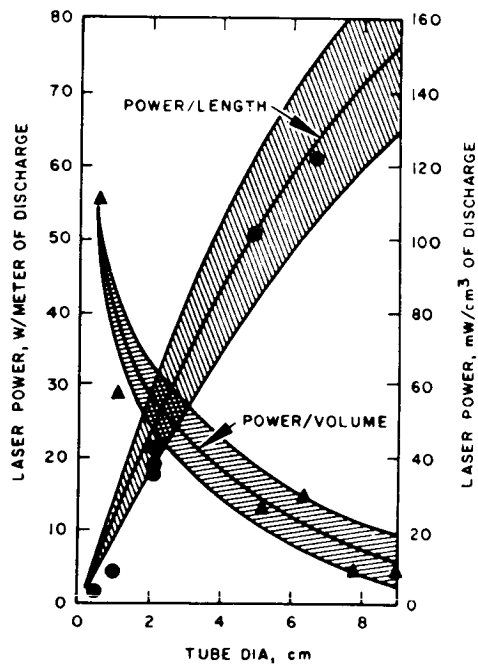


Figure A. Variation of CO₂ Laser Power per Unit Volume of Discharge and Laser Power per Unit Length of Discharge With Discharge Diameter

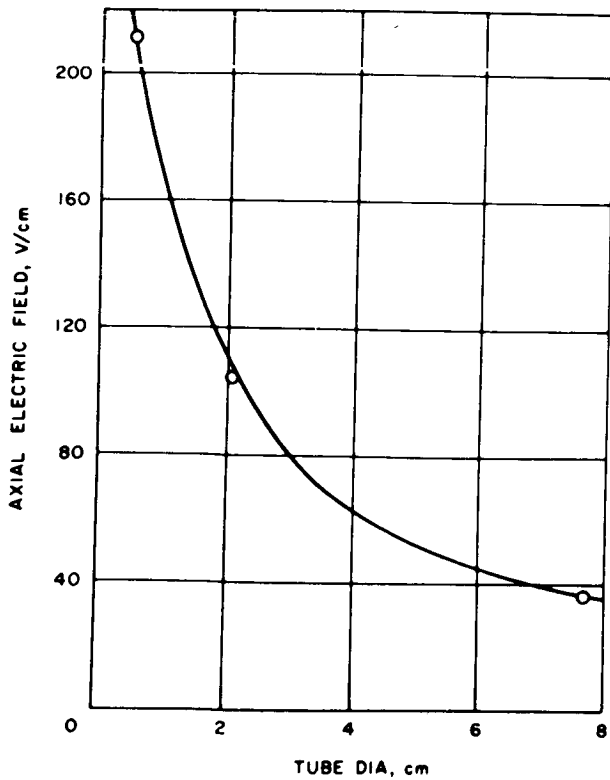


Figure B. Variation of Axial Electric Field With Discharge Tube Diameter, CO₂ Laser

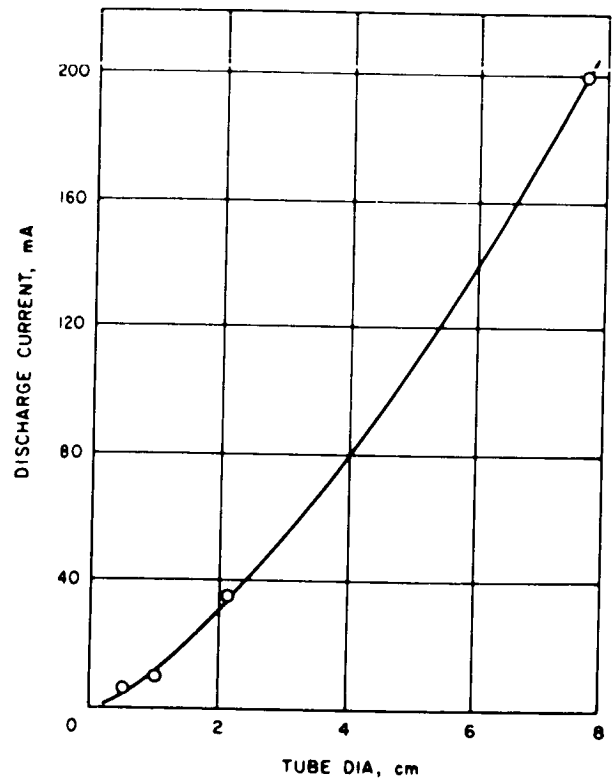


Figure C. Variation of Optimum Discharge Current With Discharge Tube Diameter, CO₂ Laser

CO₂ LASER SCALING LAWS

equipment, and (2) higher power dc supplies are easier to build and operate. No data are presently available regarding rf excitation of large diameter (≥ 5 cm) lasers.

Pulsed

High peak power can be obtained by exciting the CO₂-N₂-He laser with dc pulses (e. g., peak power of 825 w from a laser producing 17 w cw, current pulse 100 μ sec, light pulse 150 μ sec¹). No delay is observed between the current pulse and the light output.

Q-Switched

Because of the long lifetime ($\sim 2 \times 10^{-2}$ sec) and small spontaneous emission transition probability (~ 0.34 sec⁻¹) of the upper laser level 00⁰₁, the CO₂-N₂-He laser may be Q-switched.² Lasers whose cw output is a few watts will produce giant pulses of the order of 10 kw with a pulse length of approximately 100 nsec when operated in the Q-switched mode. The maximum switching speed is determined by the rate at which the upper laser level is excited and is in the neighborhood of several KHz.

State-of-the-Art in CO₂ Lasers

Since the first report of laser action in CO₂ in April 1964, CO₂ laser technology has developed at a much faster rate than that of any other laser device. In approximately two years, the output power has been increased from milliwatts to hundreds of watts, the latter obtained at efficiencies of 10 to 20 percent. The Table lists the characteristics of some early CO₂-N₂-He lasers.

¹Frapard, C., "Vibrational Excitation of CO₂ by Electronic Impact in Pure CO₂ Laser," presented at 1966 International Quantum Electronics Conference, Phoenix, Arizona, April 1966.

²Kovacs, M.A., Flynn, G.W., and Javan, A., "Q-Switching of Molecular Laser Transitions," Appl. Phys. Letters, 8, p. 62, 1966.

State-of-the-Art CO₂ Lasers (May 1966)

P _{out} (w)	Length (m)	Diam (cm)	Eff. (%)	Comments	Organization
500	10	5	15.5	Multimode output Flowing gas system	Raytheon
10	0.5	1	10	Single TEM mode output ^{oo} Sealed-off tube	HRL
150	2.4	6	18	Multimode output Flowing gas system	CGE (France)
130	4	7.5	13	Multimode output Flowing gas system	BTL

TRANSMITTING POWER SOURCES

Laser Mode Coupling and Frequency Stabilization

	Page
Laser Mode Coupling	78
Laser Frequency Stabilization Considerations	84
Laser Stabilization by Mechanical and Thermal Methods	88
Laser Frequency Stabilization Using Feedback Systems	90
Optical AFC Systems Using Passive Cavity Discriminants	92
Results of AFC for Lasers	98

LASER MODE COUPLING

Lasers typically operate simultaneous and more than one frequency or mode. Means are suggested which reduce the number of modes by using AM, PM, and FM techniques.

Introduction

The atomic populations of most gas lasers are sufficiently inhomogeneously broadened to allow the simultaneous oscillation of a large number of axial modes. Each of these modes is driven by spontaneous emission from atoms in different regions of the atomic fluorescence line; to a first approximation, these modes are uncoupled and oscillate independently. Under normal conditions, and especially for closely spaced modes (long optical cavities), the amplitudes and phases of the individual modes will fluctuate in a random manner. The output of a multimode laser is therefore amplitude modulated and has a frequency coherence much less than that of any single axial mode.

By introducing an optical modulator within the laser cavity, the previously uncoupled modes may be coupled together so that their relative amplitudes and phases are constant in time. In gas lasers, suitable modulators vary either the optical path length of the cavity (phase modulation), or the cavity losses (loss modulation). The frequency spectrum and the time domain output of a mode-coupled laser vary with the type, frequency, and amplitude of the modulation.

The effect of mode-coupling within the laser cavity is to change the form of the laser signal. The laser output may be considered as an optical carrier. By mode coupling, the carrier can be converted to a more useful form, such as an unmodulated pulse train or a single frequency. These internal modulation techniques should not be confused with schemes which use a modulator within the cavity to impose information on a direct or scattered beam.

A general discussion of the characteristics of loss and phase modulated, mode-coupled lasers is given in the paragraphs below.

Internal Loss Modulation of Multimode Lasers

The theory of internal loss modulation has been presented by several authors,^{1, 2, 3} and successful mode-locked operation has been reported for He-Ne at 6328 Å^{3, 4} and for Al II at 4880 Å.³ To introduce a time-varying loss, a suitable modulator is placed within the optical cavity as near as possible to an end mirror. The coupling effect is independent of the manner in which the losses arise.

¹Yariv, A., J. Appl. Phys., 36, pp. 388-391, February 1965.

²DiDomenico, M. Jr., J. Appl. Phys., 35, pp. 2870-2876, October 1964.

³Crowell, M. H., IEEE J. of Quantum Electronics, QE-1, pp. 12-20, April 1965.

⁴Hargrove, L. E., Fork, R. L., and Pollack, M. A., Appl. Phys. Letts., 5, pp. 4-5, 1 July 1964.

When the frequency of the time-varying loss, ν_m , is equal to or very close to the mode spacing $c/2L$, coupling of the modes results from the nonlinear polarization of the medium. The relative amplitudes of the modes becomes approximately Gaussian, centered about the line center, and all modes oscillate with equal phase. For a mode-coupled laser with a spontaneous emission linewidth $\Delta\nu$ and n axial modes oscillating in phase (normally $n \sim \Delta\nu / (c/2L)$) the laser output (as predicted by simple Fourier analysis) is a pulse train of ν_m pps with the following characteristics:

1. The pulse width is roughly $(n \nu_m)^{-1} \sim (\Delta\nu)^{-1}$.
2. The average power is approximately equal to the free-running laser power.
3. The peak pulse power is n times the average power.

The effect of the internal loss modulation on a multimode laser is then to convert the output to a pulse modulated signal. Experimental results³ obtained for two types of lasers with $\nu_m = c/2L \sim 100$ MHz are:

<u>Laser</u>	<u>λ</u>	<u>$\Delta\nu$</u>	<u>Pulse Length</u>	<u>Peak Power/ Average Power</u>
He-Ne	6328 Å	1.5 GHz	0.5 nsec	17
A II	4800 Å	4.5 GHz	0.25 nsec	20-30

Such a high-intensity pulse modulated source may have applications in PCM systems where an external optical shutter is provided to modulate the pulse height.

Internal Phase Modulation of Multimode Lasers

A mode coupling of a gas laser may also be accomplished by placing a phase modulator (KDP crystals have been used to date) inside the laser cavity. When an rf field is applied to the electro-optic crystal, its index of refraction is changed, causing a change in the length of the optical path of the cavity. This approach is equivalent to vibrating one end of the cavity mirrors at the modulation frequency, and gives rise to the generation of fm sidebands for each oscillating mode.

Two different types of mode-coupling can result depending on the frequency of the modulation. When the crystal is driven at a frequency which very nearly, but not exactly, corresponds to the frequency separation of the free running modes, the laser operates in the fm region. If the driving frequency is tuned to exactly the frequency separation of the modes, then "phase locked" operation occurs which has different characteristics from the fm region. These two modes of operation will be

LASER MODE COUPLING

discussed separately below. A detailed discussion of the limits of operation of the two regions as predicted by theory^{5, 6} and observed in experiments⁷ is contained in the literature.

Mode Coupling in the FM Region

Since the modulation frequency is roughly the frequency separation of the free running modes, the fm sidebands of each mode very nearly coincide with the frequencies of other axial modes. While each mode and its sidebands are originally a distinct fm oscillation, a parametric energy exchange takes place which allows the previously independent modes to be coupled to other modes through the overlapping sidebands. There is then a competition between the fm oscillations for the gain of the active material, and under appropriate conditions it is possible for one fm oscillation to quench or extinguish other oscillations. The result is that the total output can be made up of only one fm carrier and its sidebands. The output signal then appears as a single frequency which is swept back and forth about line center at the modulation frequency. The fm signal $E(t)$ can be represented as

$$E(t) = E_0 \cos(\omega_c t + \Gamma \cos(\omega_m t))$$

where ω_c is the carrier frequency (the laser frequency) Γ is the modulation index, or ratio of the peak phase deviation to the modulation frequency, and ω_m is the modulation frequency (approx. the axial mode spacing). The relative amplitudes of the output laser modes have Bessel function relationships to each other; the central mode will have an amplitude given by $J_0(\Gamma)$, the first sidebands $J_1(\Gamma)$, and so on.

The modulation index can be written as

$$\Gamma = \frac{1}{\pi} \frac{\Delta\Omega}{\Delta\nu} \delta \quad (1)$$

where $\Delta\Omega$ is the frequency separation of the modes, $\Delta\nu$ is the difference in frequency between $\Delta\Omega$ and the modulation frequency ν_m , and δ is the single pass phase retardation of the phase modulator, in radians. Typical values for a He-Ne laser with $\Delta\Omega$ equal to 150 MHz are $\Delta\nu = 150$ kHz, $\delta = 0.01$ and $\Gamma = 3.19$.

⁵Harris, S.E., and McDuff, O.P., Appl. Phys. Letts., 5, pp. 205-206, November 15, 1964.

⁶Harris, E.E., and McDuff, O.P., IEEE J. of Quantum Electronics, QE-1, pp. 245-262, September 1965.

⁷Amman, E.O., McMurtry, B.J., and Oshman, M.K., IEEE J. of Quantum Electronics, QE-1, pp. 263-272, September 1965.

The above equation makes it clear why fm operation is not obtained when the modulation frequency exactly equals the mode spacing. In this case, $\Delta\nu$ would be zero and Γ would be infinite. At that point the output no longer resembles an fm signal, and a new phase-locked coupled mode solution occurs.

The fm region of operation offers promise as a means to stabilize the amplitude of a cw laser. In addition, by employing the "supermode" technique or selective output coupling as discussed below, the fm laser may yield single frequency operation.

Mode Coupling in the Phase-Locked Region. When the modulation frequency is equal to the separation between axial modes ($\Delta\nu' = \Delta\Omega$), the modulation index describing fm operation becomes infinite. Under these conditions, all the modes oscillate in phase and the amplitudes of the modes have a Gaussian distribution about the center frequency. The output in this case is the same as is observed for loss modulation; that is, a train of narrow pulses with a repetition rate corresponding to the frequency separation of the modes.

Single Frequency Operation of Mode-Coupled Lasers

The "Supermode" Laser

The so called "supermode" laser⁸ approach uses the controlled spectral output of the fm laser to produce a single frequency. To accomplish this, the output of the fm laser is passed through a second phase modulator located outside of the cavity which is driven at the same frequency as the internal modulator. If Γ' is the modulation index of the external modulator and ϕ is the difference in phase between the two modulators, then the output of the external modulator will be

$$E(t) = E_0 \cos \left[\omega_c t + \Gamma \cos \omega_m t + \Gamma' \cos(\omega_m t + \phi) \right]$$

when Γ' is made equal to Γ and $\phi = 180^\circ$, then $E = E_0 \cos \omega_c t$. This is a monochromatic signal at a frequency near the center of the original free-running spectrum. Briefly then, the supermode laser produces a single-frequency output by first controlling the free-running modes in a specific manner through the f-m laser technique, and then converting this controlled signal to a single frequency.

The major limitation of this approach is the difficulty in building a practical external modulator for which Γ' can be equal to Γ . Typical values of Γ for an fm laser lie in the range of from 1 to 7. Because of the multiple pass characteristic of the laser cavity (as reflected in the term $\Delta\Omega/\Delta\nu'$ in Equation (1), Γ may be obtained using values of δ less than

⁸Massey, G.A., Oshman, M.K., and Targ, R., Appl. Phys. Letts., 6, pp 10-11, 1 January 1965.

LASER MODE COUPLING

0.3. For an external modulator, $\Gamma' = \delta'/2\pi$ so that the required values of δ' are in the range of 5 to 40 radians. It is not practical at this time to construct modulators having such large phase retardations. Future development of multipass modulators using new material such as KTN or LiNbO_3 may make it practical to use the supermode technique to obtain a single output frequency.

Frequency Selective Output Coupling to the Mode-Coupled Laser

The important property of a mode-coupled laser utilized in this approach to obtaining single frequency operation is that all the modes are coupled together in amplitude and phase. If the gain or loss of any one of the coupled modes is changed, the relative amplitudes of the modes will still be very nearly maintained and the oscillation level will adjust such that the net average power absorbed or dissipated by all modes remains zero. The method of single frequency operation discussed here is based on the fact that there is an optimum output coupling (mirror transmission) which allows the maximum power to be taken from the mode-coupled laser - and that whether this coupling is provided as a sum of equal increments to all modes, or instead is provided entirely to one mode, is not of significance. Theory⁹ has shown that if it is desired to extract the entire output as a single frequency from the q^{th} mode from line center, the ratio of necessary coupling to that mode, as compared to the coupling which should optimally be seen by all modes, is $1/J_q^2(\Gamma)$.

The method which has been used thus far,^{9, 10, 11} to obtain the selective output coupling to a single mode of a phase modulated laser is to replace one end mirror of the laser with a Fabry-Perot etalon having a free spectral range greater than the line-width of the transition. Fine tuning of the etalon allows the narrow transmission curve of the etalon to be tuned so as to couple to just one mode. Ideally it should be possible to extract the total power which was originally available from all the modes to just one mode.

Single frequency operation of a multimode laser using internal modulation to achieve mode coupling and an output etalon to select a single frequency has been demonstrated for He-Ne¹⁰ and Ar lasers.¹¹ A He-Ne laser, which as a free-running oscillator produced 68mW, yielded 53mW at a single frequency using these techniques. For argon, 45mW of single frequency power has been obtained from a 100mW free-running laser. This approach is still being refined, and higher powers can be expected, especially from argon ion lasers.

⁹Harris, S.E., and McMurtry, B.J., Appl. Phys. Letts., 7, pp. 265-267, 15 November 1965.

¹⁰Targ, R., and McMurtry, B.J., paper presented at 1966 International Quantum Electronics Conference, Phoenix, Arizona, April 12-15, 1966.

¹¹Osterink, L., Byers, R., and Harris, S.E., paper presented at 1966 International Quantum Electronics Conference, Phoenix, Arizona, April 12-15, 1966.

LASER FREQUENCY STABILIZATION CONSIDERATIONS

The stabilization of lasers is difficult due to the small fractional bandwidths available. Fractional bandwidth is expressed in terms of the dimensions of the laser.

Introduction

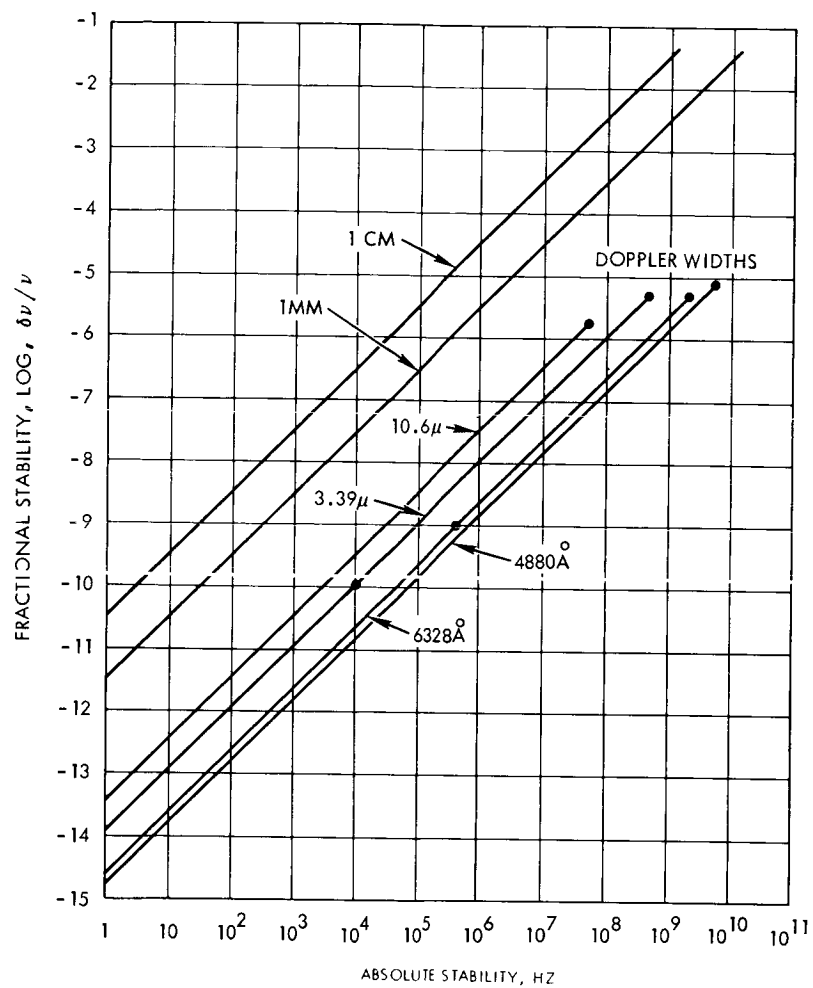
For all but the most simple communications systems, the laboratory laser does not possess the frequency stability, amplitude stability, or spectral purity equivalent to conventional microwave sources. The following discussion assumes that the amplitude is stable and the spectrum is pure, and is directed to the problems of obtaining improved frequency stability.

To first order, the exact oscillation frequency of a laser is determined by the mechanical properties of the optical cavity rather than the parameters of the atomic line. Thus, a brief review of optical cavity properties and materials will be given. Following this topic will be other topics containing stabilization methods, and the most recent results.

System Stability Requirements

Consider the simplest microwave heterodyne receiver with an operating frequency of 10 GHz and an i-f bandwidth of 1 MHz, the required local oscillator (and transmitter oscillator) stability must be better than 10^6 in 10^{10} or 1 part in 10^4 to keep the received signal within the receiver passband. For most practical communications systems the actual stability required might be two orders of magnitude greater than this, or 1 part in 10^6 . For the present, however, consider the crudest system which requires the signal remain within the passband. The same bandwidth used in an optical heterodyne receiver, with operating frequency of 500 THz (Terahertz = 10^{12} Hz), corresponding to a visible wavelength of 6000 Å, requires an oscillator stability of 10^6 in 5×10^{14} or 2 parts in 10^9 . By translating the same information bandwidth from the microwave to the optical region, the stability requirements have changed from that of everyday hardware to that of a primary frequency standard. The Figure shows the fractional stability required as a function of the desired absolute stability (equal to the information bandwidth in the example considered here) for the most commonly used gas laser transitions. For a practical heterodyne system, the required stability would be an order of magnitude tighter. The curves end at the points marked "Doppler widths" since at least this stability is required to keep a given cavity mode within the gain line half-power points and to keep the laser oscillating. * Also shown for comparison with conventional sources are curves for 1-mm and 1-cm wavelengths.

* Typical laboratory lasers have stability much worse than this; however, because of the multimode nature of the cavities, the dropping in and out of oscillation of a given mode is not ordinarily seen. For the optical heterodyne consideration must be given only to a single mode (i. e., single frequency) oscillators, so that the stabilization must be at least good enough to keep this mode within the gain line width.



Fractional Stability Required as a Function of
Desired Absolute Stability

LASER FREQUENCY STABILIZATION CONSIDERATIONS

The best results obtained to date¹ are shown as dots on the various curves. No stabilization has been reported for those wavelengths with no dot shown.

Properties of the Optical Cavity

The resonant frequency, ν , of an optical cavity formed by mirrors of radii b_1 and b_2 , spaced a distance, L , apart is²

$$\nu = \frac{c}{2L} \left[q + \frac{1}{\pi} (1+m+n) \cos^{-1} \left\{ \sqrt{\left(1 - \frac{L}{b_1}\right) \left(1 - \frac{L}{b_2}\right)} \right\} \right] \quad (1)$$

where c is the velocity of light and m , n , q are three integers that describe the particular optical mode (field distribution); q is the number of half wavelengths between mirrors and m and n are number of zeros (or phase reversals) the electric field contains in the transverse direction. Typically q is of the order of 10^6 , while m , n are small ($0, 1, 2, \dots$). Modes are usually designated TEM_{mnq} ; the most useful, i. e., the ones providing the closest to a uniform illumination and thus possessing the lowest diffraction spread for a given diameter, are the TEM_{00q} . In all that follows it is assumed that only these modes are present, selected by intracavity apertures, cavity dimensions, etc. In addition, it is assumed that by proper choice of L compared to the gain line width, only one particular value of q will oscillate at a time.

For TEM_{00q} modes equation (1) reduces to

$$\nu = \frac{c}{2L} \left[q + \frac{1}{\pi} \cos^{-1} \left\{ \sqrt{\left(1 - \frac{L}{b_1}\right) \left(1 - \frac{L}{b_2}\right)} \right\} \right] \quad (2)$$

Since $q \gg 1$, this formula, for first order effects, reduces further to

$$\nu \approx \frac{cq}{2L} \quad (3)$$

¹White, A.D., "Frequency Stabilization of Gas Lasers," IEEE J. of Quantum Electronics, QE-1, pp. 349-357, November 1965.

²Boyd, G.D., and Kogelnik, H., "Generalized Confocal Resonator Theory," Bell Sys. Tech. J., 41, pp. 1347-1370, July 1962.

Then, to first order, the cavity frequency depends only on the spacing L . The variation in frequency $\delta \nu$ caused by a change in cavity length δL is

$$\delta \nu = - \frac{c q}{2L^2} \delta L = - \frac{\nu}{L} \delta L \quad (4)$$

or,

$$\frac{\delta \nu}{\nu} = - \frac{\delta L}{L} \quad (5)$$

Second order corrections arise because of the occurrence of L in the second term in equation (2), but equation (5) will be accurate enough for most purposes.

If, in addition to changes in the physical length, δL , there is a change in the effective length due to a change, δn , in the index of refraction of the intracavity medium, then

$$\frac{\delta \nu}{\nu} = - \left(\frac{\delta L}{L} + \frac{\delta n}{n} \right) \quad (6)$$

Equation (6) is most applicable to solid-state lasers where the mirrors are coated directly on a dielectric rod of index n . For gas lasers in which the active medium has index ℓ , but the atmosphere fills the space from the end windows to the mirrors, a fraction $L-\ell/L$, where ℓ is the length of the discharge tube, then

$$\frac{\delta \nu}{\nu} = - \left(\frac{\delta L}{L} + \frac{L-\ell}{L} \frac{\delta n}{n} \right) \quad (7)$$

LASER STABILIZATION BY MECHANICAL AND THERMAL METHODS

Lasers may be stabilized by using materials which are mechanically "stiff" and have low coefficients of expansion with temperature.

The most direct method of laser stabilization is to control the environment of the optical cavity sufficiently well so that the resonant frequency variation falls within the prescribed limits. This requires accurate control of the cavity temperature (often in the presence of a strong heat source, for example, the laser discharge tube), control of or isolation from mechanical stresses and vibrations, and isolation from air currents (or other fluctuations in the intracavity medium).

The change in frequency with change in cavity length caused by thermal temperature variation is given by equation (1) and considering the definition of thermal expansion coefficient, μ .

$$\frac{\delta \nu}{\nu} = - \frac{\delta L}{L} = -\mu \delta T \quad (1)$$

The Figure shows the fractional stability resulting from a given temperature fluctuation, δT , for several common materials suitable for cavity fabrication. This figure may be combined with the figure of the previous topic to see the resulting absolute frequency change for the different laser wavelengths. To achieve a long term stability of 1 part in 10^9 using the lowest expansion coefficient material, fused quartz, a temperature control of 10^{-3} degrees Centigrade is required. By choosing a proper geometry and composite material for the package laser cavity, one may temperature-compensate to achieve perhaps an order of magnitude improvement over simple cavity expansion. Also, there are now some experimental glasses and ceramics that are internally compensated in their composition to achieve lower expansion coefficients.

The change in frequency produced by a change in air pressure depends on the fraction $L-\ell/L$ of the optical path that is exposed. A value of 0.1 for this fraction is typical for Brewster-angle gas lasers. The fractional frequency change produced by a change in air pressure, δp , is then

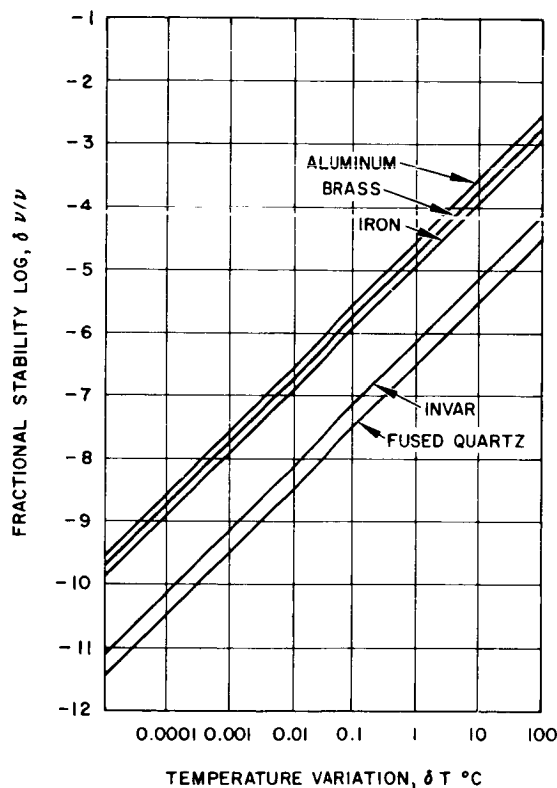
$$\begin{aligned} \frac{\delta \nu}{\nu} &= - \frac{L-\ell}{L} \frac{\delta n}{n} \\ &= - \frac{L-\ell}{L} (0.00029) \frac{\delta p}{760} \\ &\cong - \frac{L-\ell}{L} \times 3.8 \times 10^{-7} \delta p \end{aligned}$$

or,

$$\cong 4 \times 10^{-8} \delta p$$

for $L-l/L = 0.1$, and δp in Torr*. The long-term variation in pressure due to weather changes may be of the order of 20 to 30 Torr or more. Thus, $\delta\nu/\nu > 10^{-6}$, determined by atmospheric variations. Launching a gas laser into space ($\delta p = 760$ Torr) would produce a $\delta\nu/\nu$ of approximately 3×10^{-5} , which is greater than the maximum variation allowed to maintain oscillation within the doppler linewidth for all the gas lasers listed in the Figure of the previous topic. Of course, internal mirror lasers do not suffer from pressure effects, except through possible mechanical stresses set up by changes in pressure.

Perhaps the most difficult source of laser instability to describe is that resulting from room microphonics and mechanical vibration. Needless to say, every precaution must be taken to isolate the laser from sources of vibration. (This may be easier to do in a space vehicle than it is on the earth's surface.) Even when electronic stabilization is used with the laser, the microphonics determine the loop gain required and the trade-off between capture-range and sensitivity of the frequency discriminator to be used.



Fractional Stability Resulting From
a Given Temperature Fluctuation,
 δT , for Several Common
Materials Suitable for
Cavity Fabrication

* 1 Torr \equiv 1 mm of Mercury

LASER FREQUENCY STABILIZATION USING FEEDBACK SYSTEMS

The types of discriminants which could be used in an optical AFC are listed and a typical block diagram discussed.

It is possible to generate a discriminant or discriminator characteristic as shown schematically in Figure A in several ways for use in a feedback frequency control system. These include: passive cavity discriminants such as the Fabry-Perot discriminator and the two-beam interference discriminator, atomic line discriminants such as the laser gain curve or Zeeman effect discriminator and other discriminants such as a "frequency pushing" discriminator. Of these, the Fabry-Perot and the two-beam interference discriminators will be given in the next topic to show typical implementation. (For implementation of the other methods, the reader is referred to the Third Quarterly Report of this contract published June 1966.) Before these examples are given, general characteristics of optical frequency control systems will be discussed.

A discriminator must (1) measure the amount of frequency deviation from some desired center frequency (it is desirable but not necessary that the output amplitude be a linear function of the frequency deviation), and (2) indicate the sense of the deviation, usually by a positive or negative output. The discriminator output should also be independent of the signal amplitude. If it is not, the independence may be achieved as it is usually done in the radio-frequency case by limiting; unfortunately, optical frequency limiters are rare, occurring only for those transitions which possess a high, saturable gain characteristic, such as the $\text{Xe I } 3.508\mu$ transition. Otherwise, the signal source must be stabilized to obtain amplitude independence for those amplitude-dependent discriminants.

Having obtained a suitable optical discriminator, the output is then fed to a frequency-controlling element in the laser--for example, to an electro-mechanical element controlling the mirror spacing, or to an electro-optic element controlling the effective index of a portion of the cavity. This servo loop may become quite sophisticated in its detail. Integrators to obtain zero-error characteristics may be included; multiple loops derived from different discriminants may be used to stabilize for different "terms" i.e., short-term, long-term*), or different loops may be added to correct for mirror tilt in addition to mirror spacing, a second order effect. A typical system block diagram is shown in Figure B.

It is useful to recall a few general facts about frequency discriminator systems. If a discriminant is derived from a resonance phenomena, then it is quite generally true that the discriminator curve will resemble the derivative of the amplitude function of the resonance. If the width of the resonance is $\Delta\nu$ then the width between the peaks of the discriminator curve will be about $\Delta\nu$; if the resonance is asymmetrical, the discriminator will be asymmetrical, etc. The finer details depend on the exact method

* Short term means the low audio range to perhaps 1 kHz; long term means hours to years. Optical cavity discriminants are suitable for short term stability, while atomic discriminants are better suited for long term stability.

of obtaining one from the other. It is also true that the steeper the slope of the discriminant (or the higher the gain of the feedback loop) the tighter the frequency lock will be. On the other hand, the capture range is proportional to the discriminant width.

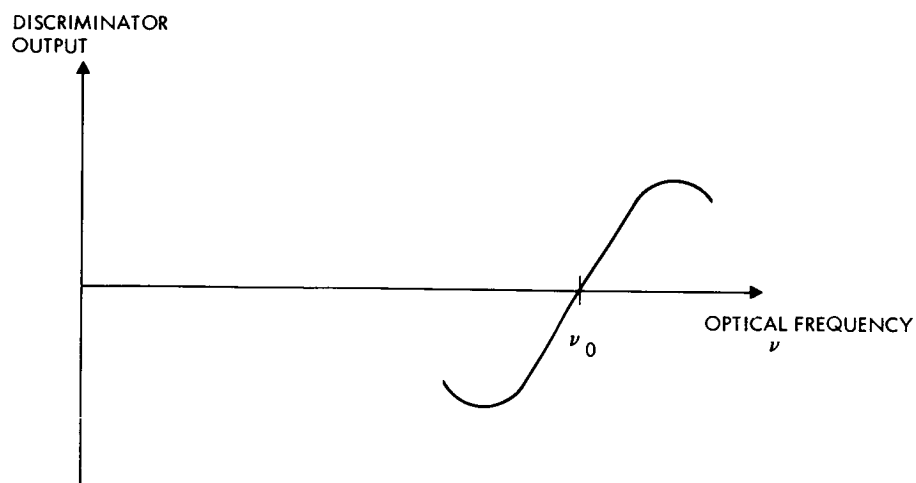


Figure A. An Optical Discriminant for Use in Feedback Frequency Control System, Schematic

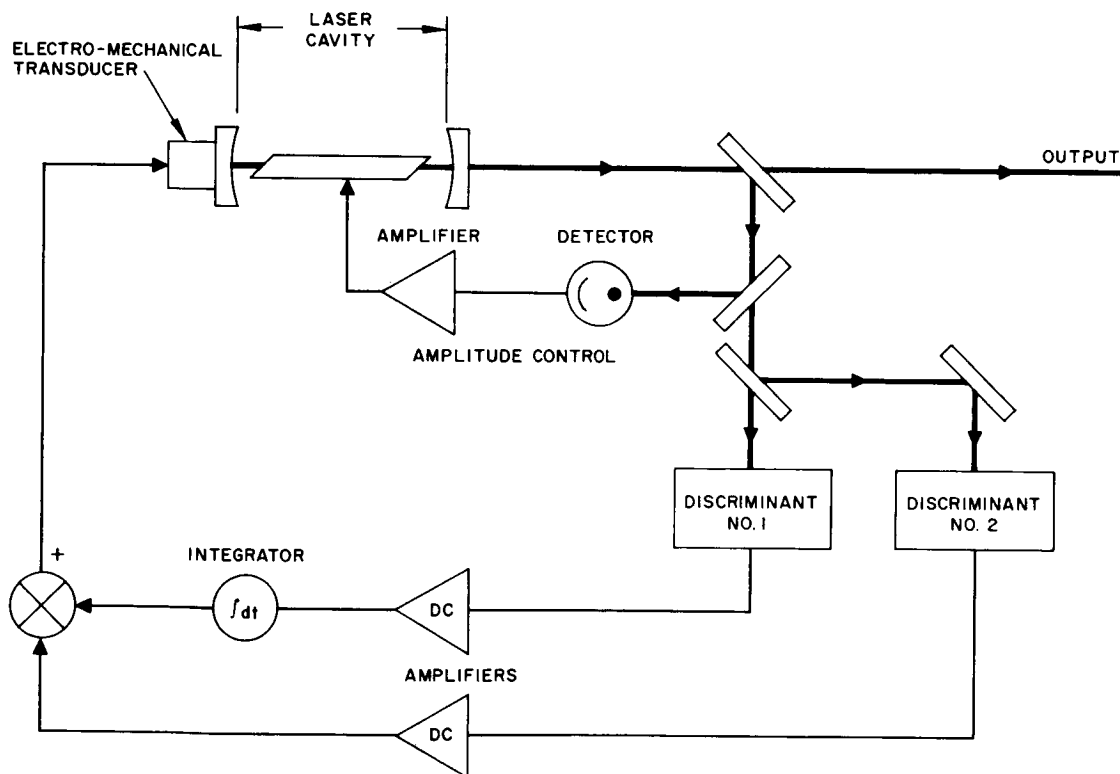


Figure B. Typical Feedback Laser Frequency Control System

OPTICAL AFC SYSTEMS USING PASSIVE CAVITY DISCRIMINANTS

Typical implementation of two passive cavity discriminators are shown for frequency controlling a laser.

It is possible to generate an optical discriminant by using interference in optical cavities. Such discriminants are probably best used for short-term stabilization, since the small cavities used can be made quite rugged and free from microphonics. However, there is no guarantee of long-term stability because of temperature drift, mechanical drift due to strains, etc.

Interference can may be classified as two-beam or multiple beam. The Michelson and Mach-Zehnder configurations are perhaps the best known examples of two-beam instruments, while the Fabry-Perot is the best known multiple beam device.

a. Fabry-Perot Interferometer Discriminator

The transmission characteristic of a typical Fabry-Perot (F-P) interferometer is shown in Figure A. Typical numerical values are given in parentheses. The bandwidth $\Delta\nu$ may be only a few MHz, much less than the Doppler width of a typical visible or near-IR laser transition. The simplest method of obtaining a discriminant from this resonator is shown schematically in Figure B. The laser output is passed through the F-P to an amplitude detector; a bias level is subtracted from the detector output and fed in the proper phase through a dc amplifier to the mirror transducer. The resulting discriminant is shown in Figure C. The output amplitude of the laser itself must be stabilized by an independent loop, since fluctuations in laser amplitude would be sensed the same as fluctuations in frequency, as shown in Figure C.

A better method is shown in Figure D. Here a small modulation ("dither"), typically at an audio rate is applied to the mirror transducer. The modulated beam is passed through the F-P. A sample is envelope-detected and the modulated output of the detector is compared with the modulating source in a phase-sensitive detector. The filtered (dc) output is then amplified and fed to the mirror transducer. That the output of the phase detector produces the desired discriminant is easily seen from Figure E. The amplitude of the modulation on the light increases as the slope of the transmission curve increases. The maximum modulation occurs at the inflection points ν_1 and ν_2 . The modulation amplitude is zero at ν_0 . (Only the second harmonic of the modulation frequency occurs at ν_0 , and it is a maximum here, zero at the inflection points--it may also be used as an error signal). The phase of the modulation on the light with respect to the modulating signal also changes as the frequency passes through ν_0 .

By adding a bias voltage to the dc amplifier output, the frequency may be stabilized at ν_3 rather than ν_0 .

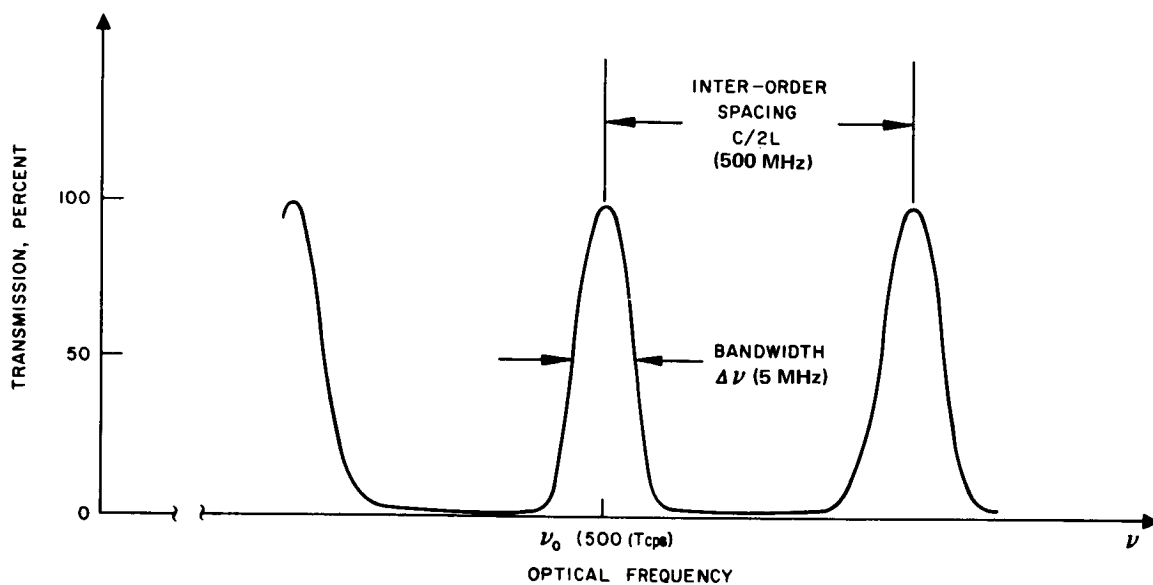


Figure A. Typical Fabry-Perot Transmission Characteristic
Numerical values typical of gas lasers are given in parentheses.

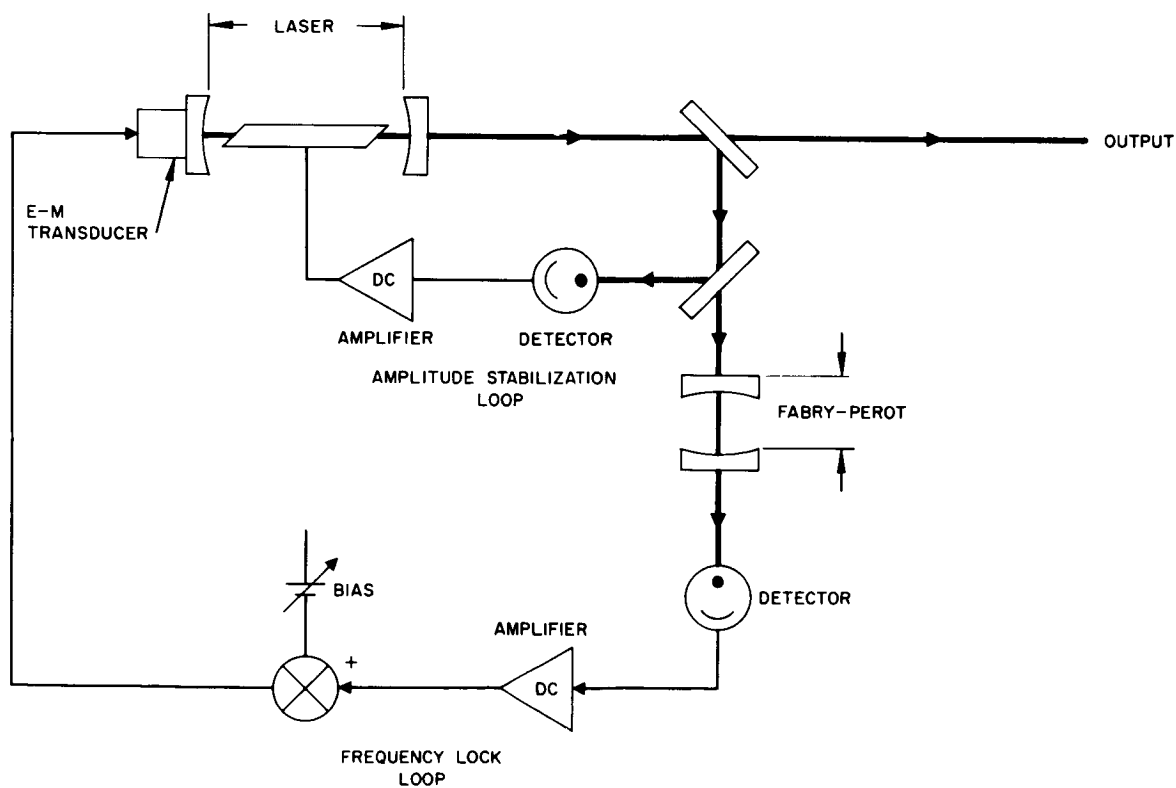


Figure B. Simple Stabilized Laser Using a Fabry-Perot
Cavity as a Reference

OPTICAL AFC SYSTEMS USING PASSIVE CAVITY DISCRIMINANTS

One of the disadvantages of the system shown in Figure D is that the laser output is frequency- and amplitude-modulated at the "dither" frequency. To eliminate this, the F-P mirrors may be dithered instead of the laser mirrors; the dc error signal is fed back to the laser as in the first system. The operation is otherwise the same.

b. Two-beam Interferometer Discriminator

Consider the simple arrangement shown in Figure F, this system may be thought of as an unequal-arm Mach-Zehnder interferometer. It is easily shown¹ that the output of photodetectors 1 and 2 is

$$\dot{Z}_1 \sim \sin^2 \left[\frac{\pi \nu}{C} (L_a - L_b) \right]$$

$$\dot{Z}_2 \sim \cos^2 \left[\frac{\pi \nu}{C} (L_a - L_b) \right]$$

The photo currents may be subtracted to form the discriminant shown in Figure G. The lengths L_a and L_b are adjusted so that the zero occurs at the desired optical frequency. The width of the discriminant between peaks is

$$\Delta \nu = \frac{C}{2 (L_a - L_b)}$$

and can be made smaller (more sensitive) by increasing the inequality in the arm lengths. Of course, the mechanical stability becomes worse the larger $L_a - L_b$ is made, so that there will be an optimum design for a given set of construction techniques and environment.

Instead of using two arms of physically different lengths, it is possible to use a birefringent crystal to obtain the necessary two paths of different optical length. A discriminator based on this idea has been proposed by Harris.²

¹Kaminow, I. P., "Balanced Optical Discriminator," Appl. Optics, 3, pp. 507-510, April 1964.

²Harris, S. E., "Demodulation of Phase-Modulated Light Using Birefringent Crystals," Proc. IEEE, 52, pp. 823-831, July 1964.

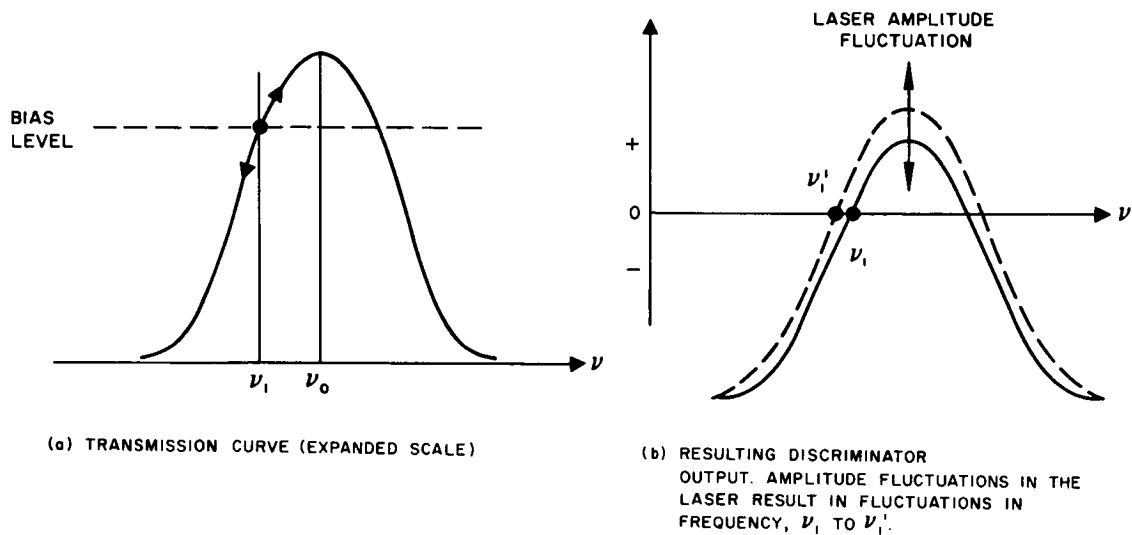


Figure C. Discriminant Characteristics

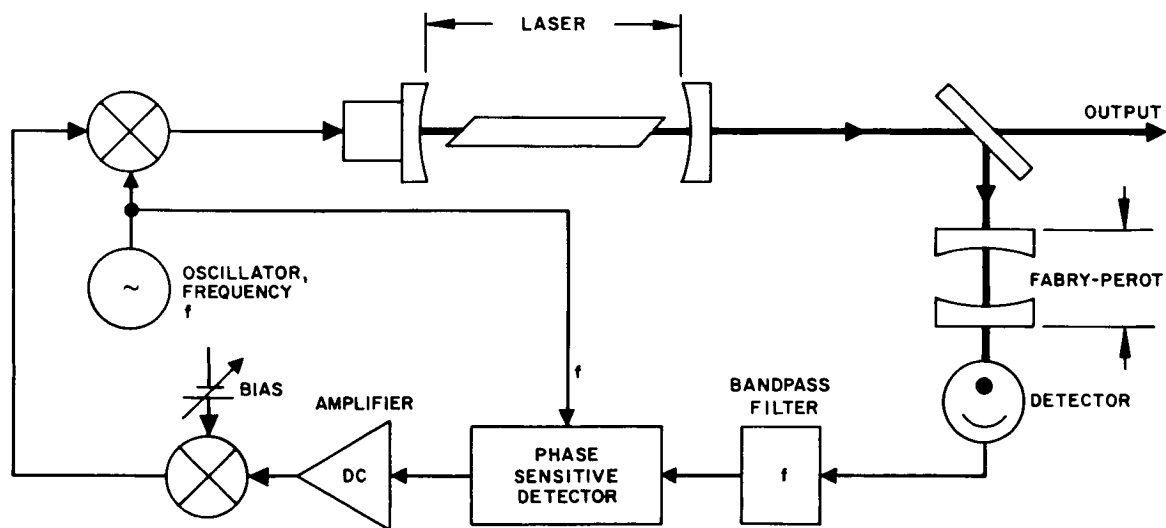
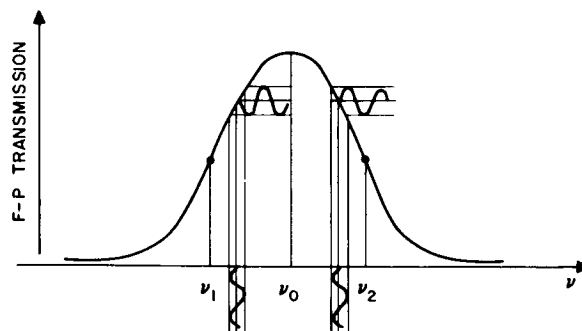
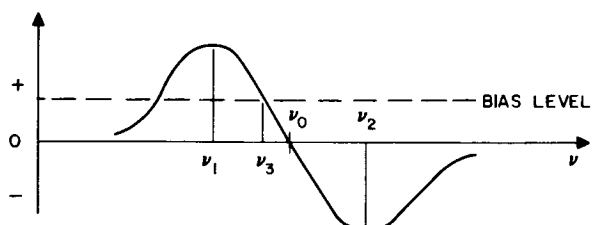


Figure D. Dither Method of Locking a Laser to a Fabry-Perot Cavity Resonance

OPTICAL AFC SYSTEMS USING PASSIVE CAVITY DISCRIMINANTS



(a) TRANSMISSION OF AN F-P RESONANCE. FREQUENCY MODULATION ABOUT A POSITION OFF LINE CENTER PRODUCES AMPLITUDE MODULATION OF THE TRANSMITTED SIGNAL. RELATIVE PHASE SHIFTS 180° GOING THROUGH LINE CENTER. MAXIMUM MODULATION OCCURS AT INFLECTION POINTS ν_1, ν_2 .



(b) DISCRIMINANT PRODUCED BY THIS SYSTEM. OSCILLATION IS LOCKED AT ν_0 WITH NO BIAS, AT ν_3 FOR BIAS SHOWN.

Figure E. Discriminant Characteristics

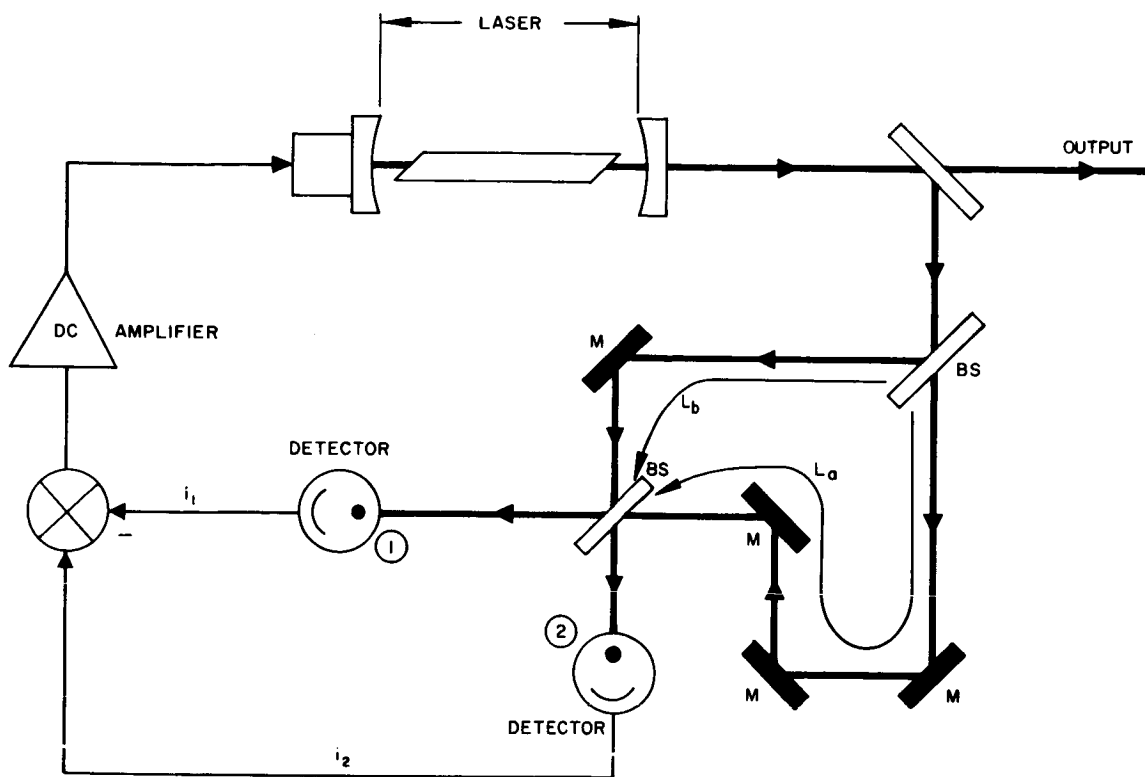


Figure F. Two-beam Interferometer Balanced Discriminator Frequency Stabilizing System (After Kaminow⁽¹⁾)

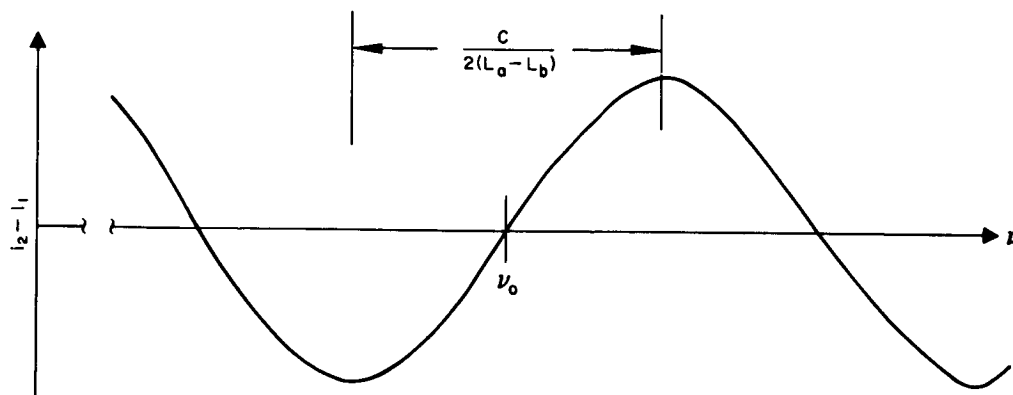


Figure G. Optical Discriminant Obtained From a Two-beam Interferometer With Path Length Difference of $L_a - L_b$

RESULTS OF AFC FOR LASERS

AFC stabilization developmens have achieved stabilization of lasers as good as one part in 10^{10} for a duration of 8 hours.

For a good review and summary of results obtained to date, the reader is referred to the excellent review article by White.¹ The best fractional stability for any laser system quoted is one part in 10^{10} for a duration of 8 hours, obtained by Bennett, et al.,² using the hole-repulsion effect with a 3.39μ He-Ne laser. For certain favorable periods of 1-minute duration or so, one part in 10^{12} was obtained. The best stability obtained to date for a free-running laser in a controlled mechanical and thermal environment is two parts in 10^8 obtained by Collinson³ with an r-f excited 6328 \AA He-Ne laser in a fused quartz cavity. Using a Brewster's angle window laser and a dither scheme, Rowley and Wilson⁴ have achieved one part in 10^8 at 6328 \AA . Shimoda and Javan⁵ have used a somewhat more elaborate dither system with a 1.15μ internal mirror He-Ne laser to achieve two parts in 10^9 for many days. White⁶ has used a combination of the Zeeman effect system for long-term stability and a passive cavity loop for short-term stability to achieve one part in 10^9 at 6328 \AA .

No results of frequency stabilization experiments on the ionized argon transition (e.g., Ar II 4880 \AA or Ar II 5145 \AA) have been reported. This is no doubt due to the difficulty in obtaining single-frequency operation with such a large Doppler linewidth ($\sim 5\text{ GHz}$). Intracavity methods of obtaining single-frequency operation show promise of self-stabilization and may well be applied to the argon ion laser.

¹White, A.D., "Frequency Stabilization of Gas Lasers," IEEE J. of Quantum Electronics, QE-1, pp. 349-357, November 1965.

²Bennett, W.R., Jr., Jacobs, S.F., Latourett, J.T., and Rabinowitz, P., "Dispersion Characteristics and Frequency Stabilization of a Gas Laser," Appl. Phys. Letts., 5, pp. 56-58, August 1964.

³Collinson, J.A., "A Stable, Single-Frequency r-f Excited Gas Laser at 6328 \AA ," Bell Sys. Tech. J., 44, pp. 1511-1519, September 1965.

⁴Rowley, W.R.C., and Wilson, D.C., "Wavelength Stabilization of an Optical Maser," Nature, p. 745, November 1963.

⁵Shimoda, K., and Javan, A., "Stabilization of the He-Ne Maser on the Atomic Line Center," (pt. 1), J. Appl. Phys., 36, March 1965.

⁶White, A.D., "A Two-Channel Laser Frequency Control System," IEEE J. of Quantum Electronics, QE-1, pp. 322-333, October 1965.

TRANSMITTING POWER SOURCES

Laser Oscillators and Amplifiers

	Page
CW Laser Performance	100
Pulsed Laser Oscillators	104
Laser Amplifiers	109
Gas Laser Selection for Space Communications	110
Laser Burden Values	114

CW LASER PERFORMANCE

CW laser sources have been demonstrated from a number of laser types producing powers as high as 8000 watts.

This section provides an initial survey of lasers for space communications and tracking. This survey list is far from complete but does exhibit those lasers which have promising performance with respect to relevant communications system parameters.

Laser oscillators have been classified according to their mode of operation as: cw, pulsed-high energy, pulsed-high power, pulsed-high repetition rate. CW operation for gas lasers is discussed in this topic. Other classifications are discussed in subsequent topics.

The following operating characteristics are also of interest: average power, pulse energy, peak pulse power, pulse repetition frequency. Other aspects of laser oscillators information needed include: wavelength, frequency stability, beam divergency/lateral coherence, noise, efficiency, temperature, size, and weight.

The cw laser oscillator will be discussed first. The output signal is a highly monochromatic beam of light with nearly constant output power.

The active materials which are used in cw laser oscillators include gases, solids, and semiconductors. While laser action has been demonstrated in materials other than those listed in the Table, only the most commonly used or most promising materials are included here.

The wavelengths available for operation comprise a set of discrete spectral lines associated with the various active materials. Generally more than one line is excited in a given laser; however, techniques exist for suppressing oscillation on all but the desired lines. Laser frequency tuning may also be accomplished, but only over a very limited range (usually less than one part in 10^4).

The output powers of existing cw laser oscillators vary from a few milliwatts for single mode operation to hundreds of watts for multifrequency operation in gas lasers. Recently 40 watts cw has been realized in a Nd:YAG laser pumped by an experimental 20 kw argon lamp. The most important factor limiting the output power is the very low energy conversion efficiency of most cw lasers. It usually ranges from 0.01 to 1.0 percent; however, a few exceptions are the CO₂ molecular laser (15 percent) and the GaAs injection laser (approximately 25 percent at cryogenic temperatures). Scattering losses in the laser material and in the cavity mirrors must be overcome in order to obtain high power output in practical devices. In addition, cryogenic temperatures are needed in present injection lasers for efficient operation.

The spectral bandwidth of the laser output, which is an indication of the frequency stability, may range from a few Hertz to several tens of gigahertz. Thermal fluctuations and mechanical vibrations are chiefly responsible for line broadening in single mode operation. For multimode operation, the spectral width of the output is closely related to the fluorescence linewidth of the atomic transition. Gas lasers exhibit the most

CW Laser Oscillators

Active Material	Wavelength, μ	Output Power	Dimensions of Active Material	Comments	References
1. He-Ne	0.6118 0.6328 1.084 1.152	5 mW 50 mW 5 mW 20 mW	6 mm x 1.8 m	Single mode, commercially available	1
2. He-Ne	0.6328	900 mW	10 mm x 5.5 m	Research devices	2
3. He-Ne	0.6328	100 mW	5 mm x 1.2 m		
4. Xe	3.5 9.0	0.1 mW 0.5 mW	2.6 mm x 50 cm	Research device	3
5. Ar ⁺	0.4579 (0.05) 0.4765 (0.1) 0.4880 (0.25) 0.4965 (0.1) 0.5107 (0.1) 0.5145 (0.4)	10 W	6 mm x 60 cm	Research devices, 0.1 - 0.2% efficiency	4
6. Ar ⁺	(as in 5)	16 W	4 mm x 2.6 m		
7. Ar ⁺	0.4880	1 W	3 mm x 45 cm	Airborne development device	6
8. CO ₂	10.57 (0.75) 10.59 (0.25) 10.59 10.59 10.59	16 W 155 W 2000 8000	25 mm x 2.0 m	1.0% efficiency, single mode for each line, 15% efficiency	7
9. Cr ⁺³	0.6943	70 mW	2 mm x 2.54 cm	Water cooled	10
10. Nd ⁺³ (CaWO ₄)	1.06	1 W	3 mm x 3.5 cm	Methyl alcohol cooling (approximately 300°K)	11
11. Nd ⁺³ (YAG)	1.06	1.5 W	2.5 mm x 3.0 cm	Water cooled, commercially available, portable	12
12. Nd ⁺³ (YAG)	1.06	0.5 W	---		13
13. Dy ⁺² (CaF ₂)	2.36	0.75 W	4.8 mm x 2.54 cm	Liquid neon (27°K) bath	14
14. GaAs	0.84	12 W	0.5 mm x 0.4 cm (diode dimensions)	Liquid He (4°K) bath, 23% efficiency	15
15. Ruby	0.6943	2.4 W	2 mm diam. x 7.5 mm rod	Water cooled, 0.11 percent efficiency	15

CW LASER PERFORMANCE

monochromatic output since their fluorescence linewidths are typically one to two orders of magnitude less than for solids or semiconductors.

The spectral characteristics of laser oscillators depend primarily upon the material used. As mentioned previously, gas lasers display a narrower fluorescence linewidth than solid or semiconductor lasers. The linewidths of these latter oscillators vary from between 0.001\AA for ruby at cryogenic temperatures to several hundred \AA for Nd:glass. It should be pointed out that the linewidth of ruby oscillators depends heavily on the operating temperature.

The beam divergence for axial mode operation is given in the diffraction limit by λ/D , where λ is the operating wavelength and D is the diameter of the beam at the output aperture. For most laser oscillators, the output beam diameter is of the order of a few millimeters, so beam divergences of the order of 0.1 milliradian can be obtained for visible light. (The semiconductor laser oscillator is somewhat unique since the active region has dimensions of the order of a few microns. The resulting beam divergence, even in the diffraction limit, is of the order of a few degrees.) For maximum output power, however, nonaxial modes are excited in the oscillator and the beam divergence increases to as much as 10 milliradians.

The beam divergence of the output from a laser oscillator can be made smaller by passing the beam through an optical system. The price one must pay for smaller beam divergence is a larger beam diameter and a small loss in signal strength due to reflection and absorption in the lens system.

Laser oscillators can have three types of noise: (1) spontaneous emission noise, (2) gain fluctuations, and (3) mode-interference noise. Except for operation near threshold, spontaneous emission noise can be neglected. Gain fluctuations due to pump power modulation can generally be reduced to an insignificant level by careful design of the pump source and associated power supplies. Mode-interference noise is not so easily eliminated. It does not occur in lasers operating in a single mode; however, if two or more modes are excited, then beat frequencies between the various modes will be produced. These may range from several kilohertz to hundreds of megahertz. Schemes for phase locking the various modes by means of an intracavity modulator are currently under development and promise to make it possible to obtain high power output with little or no mode-interference noise.

Laser operation is frequently compromised by thermal problems. One of these is associated with the low energy conversion efficiency which leads to excessive heating of the laser components. Another problem is the reduction of optical quality of the laser material due to thermally-induced distortion, or stress birefringence. The effect of a non-linear temperature distribution across the laser rod, such as that which arises during the pump cycle, is to cause depolarization of the laser output. This depolarization is not constant but displays a radial dependence. As a result, the output from the laser has a seemingly random polarization and, if one requires a polarized signal, the beam must be repolarized, introducing a loss in signal energy.

REFERENCES

1. Spectra-Physics Model 125 Gas Laser, Spectra-Physics, Inc., Mountain View, California.
2. Perry, D.L., "Mirror Coating Procedures for High Power Gas Lasers," 1964 NEREM (November) paper, p. 3.5.
3. Hughes Research Laboratories Final Report, JPL Contract No. 950803, April 1965.
4. Gordon, E.I., private communication.
5. Paananen, R., "Progress in High Power Ionized Argon Lasers," Proceedings of the Second Conference on Laser Technology, Chicago, Ill., 6-8 April 1965.
6. Hughes Research Laboratories Model 50 Argon Laser, W.B. Bridges, private communication.
7. Patel, C.K.N., Bell Laboratories Record, July/August 1965.
8. Tien, P.K., Bell Laboratories, private communication.
9. Forster, D.C., Hughes Research Laboratories, private communication.
10. Evtuhov, V. Appl. Phys. Lett. 6, 75, 1965.
11. Aldag, H.R., Appl. Optics 4, 559, 1965.
12. Geusic, J.E., Appl. Phys. Lett. 6, 175, 1965.
13. KORAD Model KY-1; Korad Corp., Santa Monica, California.
14. RCA Interim Engineering Report, No. 2, "Solid State Laser Explorations, May 1965.
15. Evtuhov, V., and Neeland, J.K., "Power Output and Efficiency of Continuous Ruby Lasers," J. of Appl. Phys. 38, No. 10, September 1967, pp. 4051-4056.

PULSED LASER OSCILLATORS

Typical pulse energies duration prf are given for high energy pulsed lasers, high peak power pulsed lasers and high prf pulsed lasers.

Pulsed Laser Oscillators - High Energy

High energy pulsed laser oscillators produce a short burst of intense monochromatic radiation in a narrow output beam. These lasers operate in a quasi-cw conversion mode - converting incoherent flashtube illumination into laser radiation. Pulse length is limited only by thermal considerations. The pulse may be smooth, as in the case of the super-fluorescent radiator, or it may consist of a series of closely spaced, random or regular spikes. Pulse repetition frequencies are very low (about one per minute for energies measured in tens of joules) due to the time required for cooling of the laser components.

Data for some high energy laser oscillators are given in Table A. Indicative of the energy available, an off the shelf pulsed ruby oscillator with a guaranteed energy of 140 joules is available. These materials are used because of their relatively high ruggedness and conversion efficiency when used to convert flashtube pumping radiation to laser emission. Large size rods of good optical quality are available to match large pumping flashtubes.

Pulse durations of a few milliseconds are usual. Peak powers may be several hundreds of kilowatts.

The energy conversion efficiency is about 1.0 percent for ruby and as high as 6.0 percent for glass. Spectral linewidths vary. For multimode glass it could be as high as 100\AA - much less for ruby.

The beam divergence of high energy laser oscillators is typically of the order of 100 milliradians.

Pulsed Laser Oscillators - High Peak Power

High peak powers are obtained by using the "Q-switched" mode of operation. This is an energy storage mode rather than a conversion mode. The energy is stored in the laser rod and released impulsively. The output is a smooth, very short pulse of monochromatic radiation which is contained in a very small solid angle.

The operating characteristics of a few of these devices are summarized in Table B.

Peak powers as high as 700 megawatts from a single ruby oscillator have been achieved. The pulse duration is usually of the order of a few nanoseconds with pulse energies of a few joules being typical.

Table A. Typical High Energy Laser Oscillators

Material	Wave-length (μ)	Pulse Energy (J)	Pulse Duration	Rod Size	Comments	References
1. Ruby	0.6943	50	0.5 msec	12 mm x 15 cm	Commercially Available; 7 mr beam divergence	1
2. Ruby	0.6943	140	0.5 msec	12 mm x 15 cm		2

Table B. High Peak Power Laser Oscillators

Material	Wavelength (microns)	Peak Power (megawatts)	Pulse Energy (Joules)	Rod Size	Comments	Reference
1. Ruby	0.6943	50	1.0	---	Commercially available, beam divergence = 7 mr	3
2. Ruby	0.6943	80	2.8	10 mm x 7.6 cm	beam divergence = 0.2 mr	4
3. Ruby	0.6943	500	4.0	---	Commercially available; beam divergence = 7 mr	1
4. Ruby	0.6943	700	--	13 mm x 15.2 cm		5
5. Nd:glass	1.06	30	1.5	6 mm x 46 cm		6
6. Nd:glass	1.06	100	--	---	U. of Moscow	5

Table C. High Repetition Frequency Laser Oscillators

Material	Wavelength (Microns)	Peak Power	Pulse Energy	Repetition Rate	Comments	Reference
He-Ne	1.118	300 w	0.25 mj	2.6 kcps	ave. power 0.6 w	7
Ar ⁺	several lines 0.4579- 0.5145	50 w	0.05 mj	200 cps	ave. power 0.1 w	8
Nd:YAG	1.06	250 w	0.05 mj	100 cps	cw pump, output power not optimized	9
Nd:YAG	1.06	2 kw	0.2 mj	5 kcps		
Nd:CaWO ₄	1.06	6 mw	140 mj	100 cps	ave. power 14 w	10
GaAs	0.84	2 w	2 μ j	10 kcps	77°K, 7 Å, efficiency = 1%	11
GaAs	0.90	4 w	0.2 μ j	1 kcps	300°K, 14 Å, efficiency = 0.04%	12

PULSED LASER OSCILLATORS

The energy conversion efficiency of this type of operation is lower than in the high energy mode - typically about 0.1 percent. Spectral line-widths can vary from 0.001 Å for single mode operation in ruby to 100 Å for multimode operation in Nd:glass. Correspondingly, the beam divergence for single mode operation can be as small as approximately 0.2 milliradians and for multimode operation may be as large as approximately 10 milliradians.

At the present time, these devices can operate at a pulse repetition frequency of about one pulse every second. Cooling of the laser components is the major problem to be overcome in achieving higher repetition frequencies.

Pulsed Laser Oscillators - High Pulse Repetition Frequency (PRF)

Table C contains information on a number of pulsed laser oscillators with a PRF of 100 pps or higher. Pulse repetition frequencies as high as 12 kcps have been achieved in a He-Ne laser and 5 kcps in Nd:YAG. Peak powers vary from a few watts at very high repetition rates to several megawatts with Q-switched operation at lower rep rates.

Pulse energies are quite small, usually of the order of a millijoule, with the result that the average power ranges from a few milliwatts to a maximum of about 15 watts.

When the repetition rate is comparable to build up time the efficiency of a high prf system can be severely degraded if repetition is obtained by pump modulation. Conversely if the prf is comparable to relaxation time, no loss in efficiency is obtained by using cw pumping and modulating the regeneration cavity, either by Q modulation or other intracavity modulation techniques.

REFERENCES

1. KORAD Model K-2Q; Korad Corp., Santa Monica, California.
2. KORAD Model K2; Korad Corp., Santa Monica, California.
3. KORAD Model K-1Q; Korad Corp., Santa Monica, California.
4. Peressini, E. R., Appl. Phys. Letts., 3, p. 203, 1963.
5. Sooy, W. R., private communication.
6. Snitzer, E., "Neodymium Glass Laser," Proceedings of the Third International Conference on Quantum Electronics, Paris, France, 11-15 February 1963.
7. Goldsmith, J., "Measurement of High Power Output from a He-Ne Pulse Gas Laser Employing an Exit Mirror of Optimum Reflectivity," MIL-E-CON Conference, Washington, D.C., 14-16 September, 1964.
8. Bridges, W. B., private communication.
9. Young, C. G., Appl Phys. Letts., 2, p. 151, 1963.
10. Gilmer, A., private communication.
11. KORAD Model KS-3; Korad Corp., Santa Monica, California.
12. KORAD Model KR-2; Korad Corp., Santa Monica, California.

LASER AMPLIFIERS

Laser amplifiers characteristics for CW and pulsed operation are described.

Elaborate laser transmitters may employ one or more laser amplifiers to increase the output power and energy. Another use of laser amplifiers is to amplify single mode oscillators and thus have high spectral radiance.

Laser amplifiers may be operated on a cw-pump or pulse basis. The operating characteristics of interest are

1. Small signal power gain
2. Saturation power
3. Small signal energy gain
4. Saturation energy
5. Bandwidth
6. Operating wavelength
7. Distortion
8. Noise
9. Efficiency.

CW Laser Amplifiers. The small signal power gain of a gas laser can be quite high (as much as 70 db/m for xenon). High power amplifiers have been constructed for 10.6 microns which provide output powers of several kilowatts.

The 3-db bandwidth of a laser amplifier is related to the amplifier gain, G , and to the fluorescent linewidth of the laser transition by (for large gain):

$$(3\text{-db bandwidth}) \cong (\text{fluor. linewidth}) / \sqrt{\ln G}$$

where

ℓ = length between reflectors

n = mode number

Fluorescent linewidths vary from 100 MHz to 5000 MHz for gas lasers. Thus, for a gain of 20 db, the 3-db bandwidth of the laser amplifier will be narrowed by a factor of approximately two.

The relatively narrow passband of a laser amplifier requires that the operating wavelength of the amplifier and oscillator be closely matched. The usual way to accomplish this is to use the same laser material for both oscillator and amplifier. Even in this case, however, a temperature differential may be sufficient to put the oscillator frequency outside the passband of the amplifier. The oscillator-amplifier frequency matching problem can be alleviated somewhat by the use of laser frequency tuning techniques. These allow one to change the frequency by small amounts, generally by less than one part in 10^4 . Nonlinear devices may also be used to achieve frequency diversification. The distortion or spreading of the laser beam in passing through an amplifier depends on

the optical homogeneity of the laser amplifier medium. For gas laser amplifiers this is not a problem. A diffraction limited input beam yields a diffraction limited output beam.

Amplifier noise is due to spontaneous emission. It increases with the gain of the amplifier according to the relation

$$P_s = h\nu\Delta f(G-1)$$

where $h\nu$ is the photo energy, Δf is the passband of the amplifier, and G is the gain. At 10 db gain, gas laser amplifiers produce noise powers ranging typically from 10^{-10} to 10^{-8} watts depending on the passband.

Pulsed Laser Amplifiers. For pulsed laser amplifiers, the input pulse power is always sufficient to cause power gain saturation. For very short pulses, only the leading edge of the input pulse sees the gain of the fully pumped amplifier. Subsequent portions of the input pulse see a smaller gain due to depopulation by the leading edge of the pulse. The result of this saturation phenomena is a "pulse-shaping" or distortion of the input pulse envelope. Small signal power gains of 30 db in a 15-cm ruby rod have been achieved. Due to the finite rise time of the input pulse generated by a laser oscillator, the peak power gain is more like 10 to 15 db for the same rod.

Since power gain varies during the pulse, it is sometimes convenient to define an energy gain as (output pulse energy)/(input pulse energy).

Small signal energy gains of 20 db in a 15-cm ruby rod have been achieved. Saturation occurs when the input pulse energy density (joules/cm²) is sufficient to completely depopulate portions of the amplifier rod. For ruby and Nd:glass this occurs at approximately 7 j/cm²; for Nd:YAG at 0.06 j/cm². Solid state materials are used for high power or high energy laser amplifiers due to their energy storage capability. Ruby and Nd:glass can store typically from 1 to 5 j per cc. Consequently, the available wavelengths are 0.6943 micron in ruby and 1.06 microns in Nd⁺⁺⁺. Pulsed solid state laser amplifiers have larger passbands than cw gas laser amplifiers; the passband for ruby is about 1 Å and for Nd:glass, it is several tens of Å.

Distortion in pulsed amplifiers in the form of increased beam divergency may exist due to transient thermal gradients in the amplifier material. However, operation of ruby amplifiers with no spreading of the input beam in passing through the amplifier has been demonstrated.

Two factors tend to limit oscillator-amplifier performance. These are damage thresholds for power density and for energy density. If the power density in a dielectric becomes too high, the resultant electric field can cause dielectric breakdown of the material. If the energy density in a pulse, whose duration is short compared to some thermal relaxation time, becomes large, energy will be absorbed by the material faster than it can be carried away and the material will melt. This seems to be the most important factor limiting the performance of high energy glass laser amplifiers. The efficiency of ruby and Nd:glass laser amplifiers approximates the values obtained in high energy quasi-cw conversion mode oscillators.

GAS LASER SELECTION FOR SPACE COMMUNICATIONS

A CO₂ transmitter for a spaceborne communication link and an Argon laser for an up link beacon appear to be the best choice for laser space communication.

The Table summarizes the characteristics of six wavelengths produced by gas lasers. Hundreds of other wavelengths are available, but these six have been selected as representative of each type (ion, molecular, and neutral gas). The reported output power, length and input power are given for the lasers selected.

Notes

1. This is a Hughes airborne quartz laser with a 46 cm bore length, ~1 meter overall package length. It requires a magnetic field of ~1000 gauss, which implies a heavy structure and possibly more power.
2. This laser was reported by Raytheon in Electronic News; it is a quartz tube. The power out is 18 watts, provided the beam in the cavity was chopped to prevent damage to the mirrors.
3. This was produced under carefully controlled conditions at Bell Labs.
4. This is a commercially available Spectra-Physics model 125. 50 mw is guaranteed, but selected tubes produce 100 mw.
5. This is the Spectra-Physics model 125. It may be possible to double the power in a tube this size, but drastic improvements are quite unlikely at this wavelength.
6. This is an Hughes Research Laboratories (HRL) Laboratory-type tube. The output may be doubled, but more power than this is doubtful in a tube this size.
7. This is a TRG Laboratory-type tube and represents approximately two years of effort in developing a high power Xe laser. It is probably close to the ultimate for a tube this size.
8. This is an HRL Laboratory-type tube using flowing CO₂-N₂, mirrors were not optimized; more output power can be expected from this same tube (~20 watts). Seven watts were obtained with the tube sealed.
9. This is the Bell Telephone Laboratories work (C. K. N. Patel, Appl. Phys. Lett., 1 July 1965). A flowing gas system was used with a mixture of CO₂, N₂, O₂, H₂O.
10. This is a BTL result with a tube 4 inches in diameter and 12 feet long. Helium was used. It is hard to estimate how much power will eventually be obtained from a tube of this size. (C. K. N. Patel).

Gas Laser Performance

Gas	λ (μ)	Manu- facturer	P out (W)	L (M)	P in (W)	η	Note
Ar II	0.5	HRL	4.0	0.46	4,000	1×10^{-3}	1
Ar II	0.5	RAY	8.0	1.6	20,000	4×10^{-4}	2
He-Ne	0.63	BTL	1.0	5	500	2×10^{-3}	3
He-Ne	0.63	S-P	0.1	1.7	~200	5×10^{-4}	4
He-Ne	1.15	S-P	0.03	1.7	~200	1.5×10^{-4}	5
He-Ne	3.39	HRL	0.01	1.7	80	1.8×10^{-4}	6
He-Xe	3.51	TRG	0.08	2.0	~200	4×10^{-4}	7
CO ₂	10.6	HRL	10	2.0	150	6.7×10^{-2}	8
	10.6	BTL	12	2.0	-	3×10^{-2}	9
	10.6	BTL	130	4.0	~1,000	$\sim 1.3 \times 10^{-1}$	10
	10.6	BTL	0.1	0.5	30	3.3×10^{-3}	11

GAS LASER SELECTION FOR SPACE COMMUNICATIONS

11. This is a small, non flow tube, with external mirrors; suitable for spacecraft. This work is due to T. J. Bridges and is rather preliminary. An account of similar tubes appears in the 1 November 1965 Appl. Phys. Letters.

In comparing the various lasers listed in the Table, the suitability of the output signal for the communications task at hand must be kept in mind. All of the lasers listed can be made to operate in the lowest order spatial mode (TEM_{00}) alone with more or less difficulty. The task is easier at the shorter wavelengths where the laser output is visible and the characteristic beam size is small (proportional to the square root of the product of wavelength and a cavity parameter related to mirror radius). Mode selection at infrared wavelengths may be done with an image-converter or by listening to self-beats in a heterodyne detector. Production of a single-frequency output is still quite difficult because of the longitudinal mode structure of the long Fabry-Perot cavities used. Only the 10.6μ CO_2 and 3.5μ xenon lines are narrow enough to produce reasonable output by keeping the Fabry-Perot resonator short enough so that only one longitudinal mode oscillates. This is done to the narrow doppler-broadened line widths of these two transitions (≈ 50 MHz for 10.6μ CO_2 and ≈ 120 MHz for 3.5μ Xe). Even these two transitions will require further mode selection techniques if longer, higher power tubes are considered. Because of the broad doppler line widths of the Ar and He-Ne lasers, single-frequency operation through the use of a sufficiently short Fabry-Perot resonator entails a drastic loss in output power. Techniques involving 3 mirror resonators allow the use of longer tubes at the expense of added complexity both mechanical and electronic (servo-controlled mirror positioning), but still sacrifice output power because the entire line is not used. The most promising technique developed to date is that of intracavity mode locking¹ with a subsequent coherent recombination² or selective output coupling³. This technique has been demonstrated in the laboratory, but practical power levels at a single frequency are yet to be obtained. In any case the additional complexity will contribute to the weight, length and inefficiency of the laser, although perhaps not to a significant extent.

It appears that, at present, the best laser for optical space communications at present would be a small, efficient, light weight 10.6μ CO_2 laser in the spacecraft with coherent detection (superheterodyne) on the ground, employing a cooled Hg:Ge detector. The up-link would be best handled by a high-power multimode argon laser on the ground, employing pulse amplitude or pulse polarization modulation, and a simple ruggedized photomultiplier video receiver in the spacecraft. These conclusions are, of course, subject to revision as the state of the laser (and detector) art progresses.

¹Harris, S.E., and McDuff, O.P., Appl. Phys. Letts., 5, pp. 205-206, November 15, 1964.

²Massey, G.A., Ashman, M.K., and Taig, R., Appl. Phys. Letts., 6, p. 10, 1965.

³Hanes, S.E., and McMurtry, B.J., (to be published).

LASER BURDEN VALUES

Cost and weight burden values are given for the CO₂ laser ($\lambda = 10.6\mu$) and the Argon laser ($\lambda = 0.51\mu$).

A main purpose of this contract (NAS 5-9637) is to compare the performance of several communications systems operating at different wavelengths. In order to do this an extensive modeling was undertaken which expressed parameters in the communications link equation in terms of cost or weight. (See Appendix A of this Volume.)

From the material given in this Part; Part 1, Transmitting Power Sources; and from other investigations, constants have been chosen which relate the laser transmitted power, P_T , to the cost of obtaining this power, C_{P_T} , and to the weight of a transmitter supplying this power, W_{P_T} . This has been done for a CO₂ laser ($\lambda = 10.6\mu$) and for an Argon laser ($\lambda = 0.51\mu$).

The values used in these relationships are the best that could be determined at the date of this final report and are certainly subject to change. This is especially true of the cost relationships which represent estimates of fabrication cost only and do not include development costs.

Figures A and B give the expected weight for a CO₂ laser and an Argon laser as a function of output power. Figures C and D give the cost for these two lasers as a function of output power. The efficiency of CO₂ and Argon lasers are taken as 10 percent and 0.1 percent respectively.

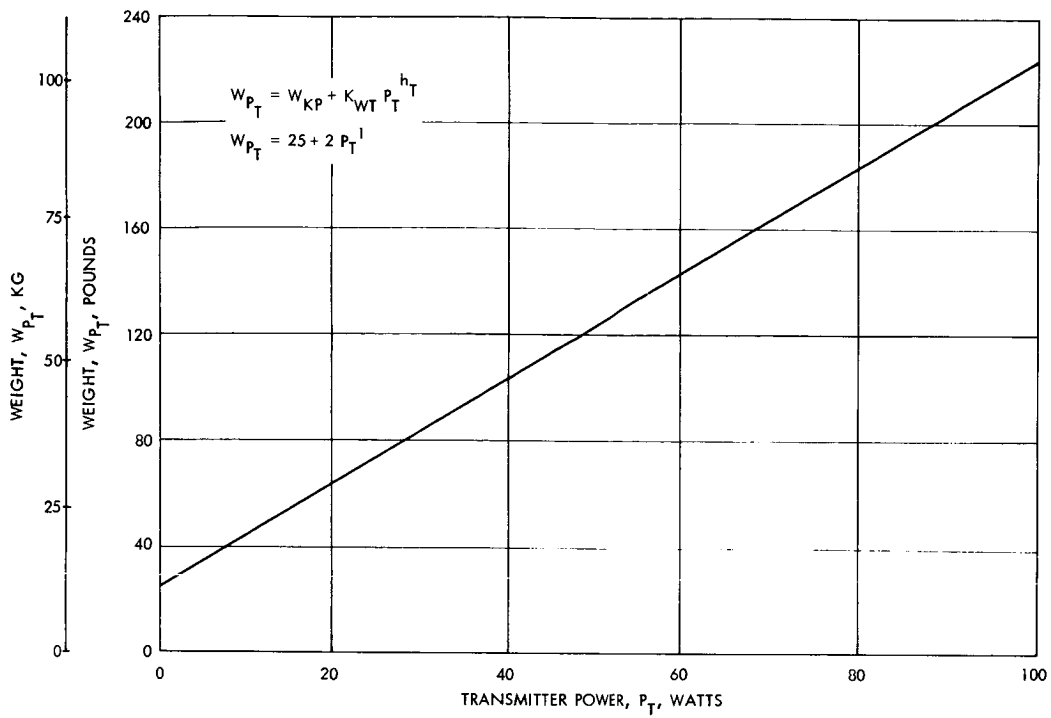


Figure A. Weight of a CO₂ Laser ($\lambda = 10.6\mu$)

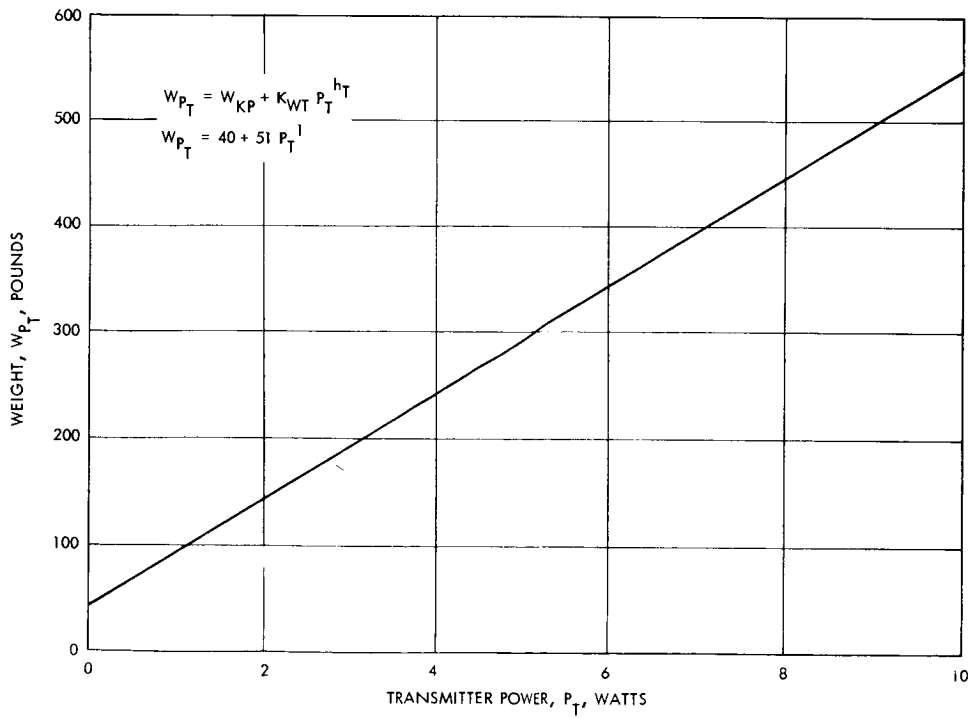


Figure B. Weight of an Argon Laser ($\lambda = 0.5\mu$)

Transmitting Power Sources Laser Oscillators and Amplifiers

LASER BURDEN VALUES

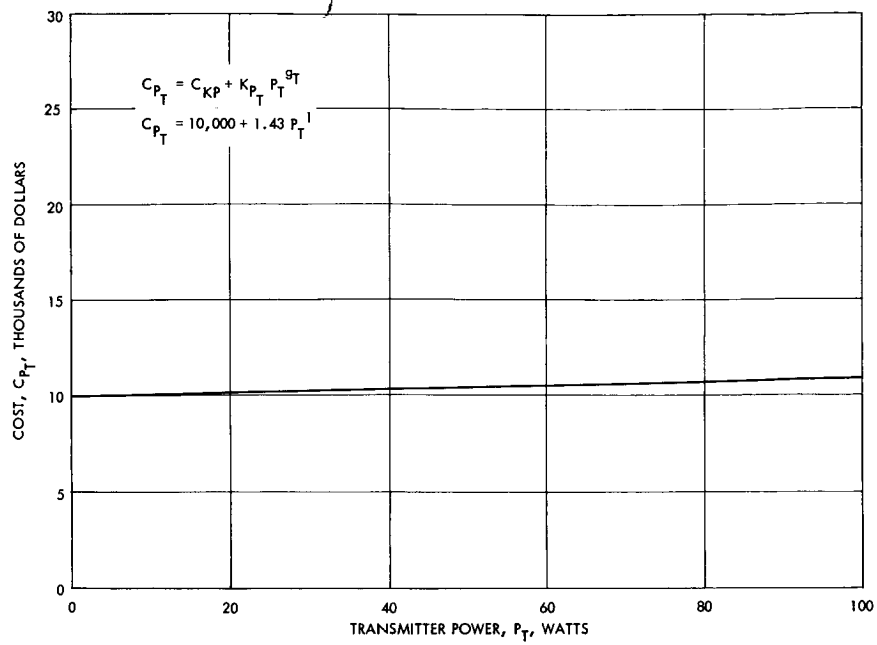


Figure C. Cost of a CO₂ Laser ($\lambda = 10.6\mu$)

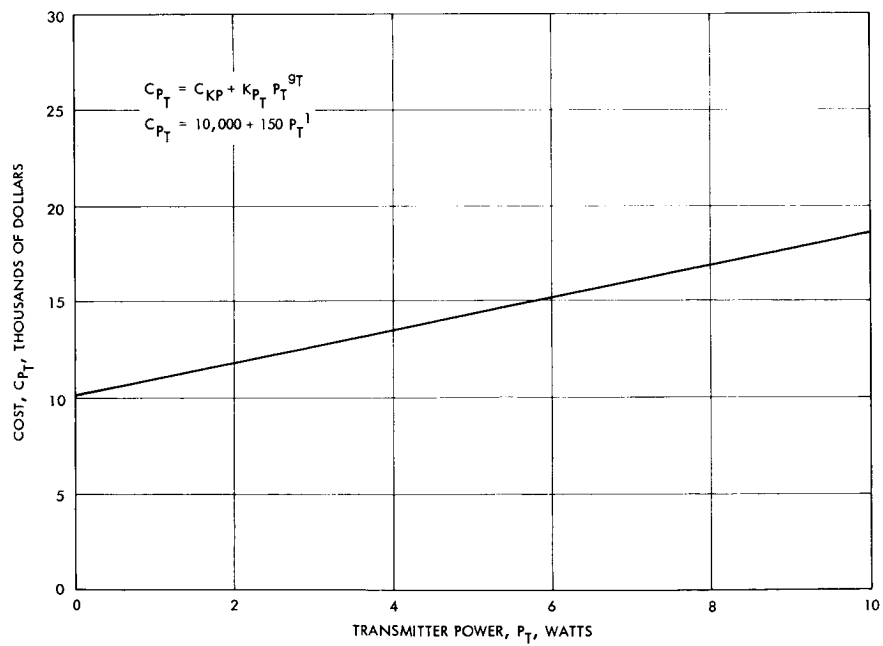


Figure D. Cost of an Argon Laser ($\lambda = 0.5\mu$)

N71-19645

PART 2 – MODULATORS

Section	Page
Radio Frequency Modulators	120
Optical Modulators	128
Electro-Optic Modulation	134
Elasto-Optic Modulation	152
Internal Modulation	168
Modulator Performance	176

MODULATORS

Radio Frequency Modulators

	Page
Modulation Methods	120
Radio Frequency Modulations Implementation	124
Digital and Compound Modulation Implementation	126

MODULATION METHODS

Amplitude, frequency and phase modulation are discussed.

A signal to be modulated may be described by equation 1.

$$A(t) = E_{mo} \sin (\omega t + \phi) \quad (1)$$

where: E_{mo} is the peak amplitude of the signal
 ω is the angular frequency
 ϕ is the phase

It is possible to modulate this signal by changing the constants: E_{mo} , ω , or ϕ . This produces amplitude, frequency, and phase modulation respectively.

Amplitude Modulation

Amplitude modulation may be analyzed by letting

$$E_{mo} = E_m + mE_m \sin qt \quad (2)$$

$$\phi = 0$$

$$\omega = \omega_o$$

$$0 \leq m \leq 1$$

then

$$A(t) = E_m \left[1 + m \sin qt \right] \sin \omega_o t \quad (3)$$

by trigonometric expansion this equation may be written

$$A(t) = E_m \sin \omega t - \frac{mE_m}{2} \cos (\omega_o + q)t + \frac{mE_m}{2} \cos (\omega_o - q)t \quad (4)$$

Equation (4) shows the carrier term and the two sidebands.

The bandwidth required for an AM system is twice the modulating frequency. However, it is possible by suitable filtering to suppress the carrier and one sideband and transmit a single sideband only. This is usually noted as single sideband amplitude modulation (SSB-AM). As may be seen from equation 4 the modulation frequency and amplitude may

be recovered from SSB-AM if the carrier frequency is known at the receiver. SSB-AM transmission is approximately 5 db more efficient than normal AM.

Frequency Modulation

Frequency modulation may be analyzed by letting

$$\omega = \omega_o + 2\pi k f_o \cos qt \quad (5)$$

$$\phi = 0$$

$$E_{mo} = E_{mo}$$

Equation 1 becomes then

$$A(t) = E_{mo} \sin \left[(\omega_o + 2 k f_o \cos qt)t \right] \text{ or by expansion} \quad (6)$$

$$\begin{aligned} A(t) = E_{mo} \left\{ J_0(m_f) \sin \omega_o t + J_1(m_f) \left[\sin (\omega_o + q)t - \sin (\omega_o - q)t \right] \right. \\ \left. + J_2(m_f) \left[\sin (\omega_o + 2q)t - \sin (\omega_o - 2q)t \right] \right. \\ \left. + J_3(m_f) \left[\sin (\omega_o + 3q)t - \sin (\omega_o - 3q)t \right] \right. \\ \left. + \dots \right\} \quad (7) \end{aligned}$$

where

$$m_f = \frac{\Delta f}{f_o} = \frac{2\pi \Delta f}{\omega_o}$$

and $J_m(m)$ is the Bessel function.

The bandwidth of a frequency modulated signal is infinite. However, higher order Bessel functions have small values and may be neglected. A practical bandwidth requirement for a frequency modulated wave is

$$BW_{IF} = 2f_m (m_f + 1) \quad (8)$$

$$\text{where } f_m = \frac{q}{2\pi}$$

MODULATION METHODS

A frequency modulated wave realizes an improvement in signal to noise ratio (SNR) when detected. The relationship of the modulated SNR to the detected SNR is

$$\frac{(\text{SNR})_{\text{detected}}}{(\text{SNR})_{\text{modulated}}} = \frac{3}{2} \frac{\Delta f^2 B W_{\text{IF}} \delta}{f_m^3} \quad (9)$$

where δ is a loss due to imperfect filtering.

Phase Modulation

Phase modulation may be analyzed by letting

$$\begin{aligned} \phi &= m_p \sin qt \\ E_{mo} &= E_{mo} \\ \omega &= \omega_o \end{aligned} \quad (10)$$

Then equation 1 may be written as

$$A(t) = E_{mo} \sin (\omega_o t + m_p \sin qt) \quad (11)$$

Equation 11 may be expanded to yield:

$$\begin{aligned} A(t) = E_{mo} \left\{ J_0(m_p) \sin \omega_o t + J_1(m_p) \left[\sin (\omega_o + q)t - \sin (\omega_o - q)t \right] \right. \\ + J_2(m_p) \left[\sin (\omega_o + 2q)t + \sin (\omega_o - 2q)t \right] \\ + J_3(m_p) \left[\sin (\omega_o + 3q)t - \sin (\omega_o - 3q)t \right] \\ \left. + \dots \right\} \end{aligned} \quad (12)$$

The bandwidth of a phase modulated signal may be obtained from equation 8 by substituting m_p for m_f . Similarly, the detection improvement obtained at demodulation may be obtained from equation 9 by a slight rearranging of terms to yield

$$\frac{(\text{SNR})_{\text{detected}}}{(\text{SNR})_{\text{modulated}}} = \frac{3}{2} \frac{m_p^2 BW_{IF} \delta}{f_m} \quad (13)$$

RADIO FREQUENCY MODULATION IMPLEMENTATION

Typical circuits are given for amplitude, phase and frequency modulation.

Implementations for amplitude, phase, and frequency modulation can take on a great variety of forms. Three single circuits are discussed below which can accomplish these modulation forms.

Amplitude Modulation

Amplitude modulation has, as a basic equation:

$$A(t) = E_m \left[1 + m \sin qt \right] \sin \omega_o t \quad (1)$$

Thus there is a basic operation of multiplying the modulation waveform and the carrier waveform. This can be done using the circuit of Figure A. Figure B may be used to generate amplitude modulation with suppressed carrier. If this signal is passed through a suitable bandpass filter single sideband amplitude modulation results.

Frequency Modulation

The equation describing frequency modulation is:

$$A(t) = E_{mo} \sin \left[(\omega_o + 2 k f \sin qt)t \right]$$

here a frequency is added to the carrier frequency. This is most commonly done by using an electronically controllable reactive element in the frequency determining circuit of an oscillator. Figure C illustrates such an implementation.

Phase Modulation

The equation describing phase modulation is:

$$A(t) = E_{mo} \sin (\omega_o t + m_p \sin qt)$$

Here a phase addition is indicated. Such a modulation is often accomplished using a phase shifter which can be controlled electronically. Figure D illustrates such an implementation.

It should be noted that frequency and phase modulators are often followed by frequency multipliers. When this occurs the phase and frequency modulation indices are also multiplied by the multiplying factor.

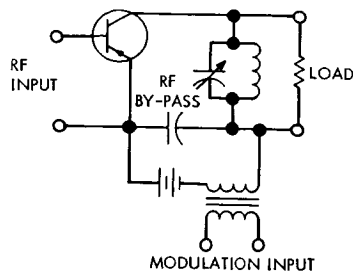


Figure A. Amplitude Modulator

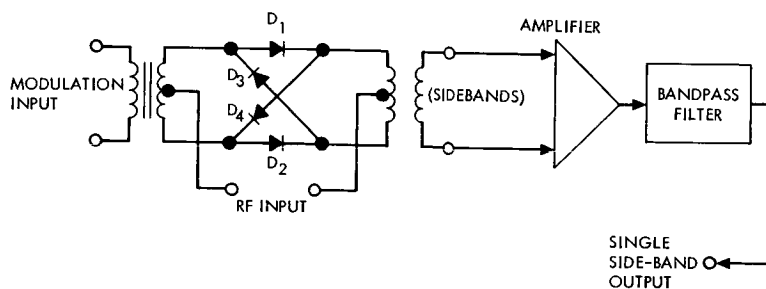


Figure B. Single Sideband Modulator

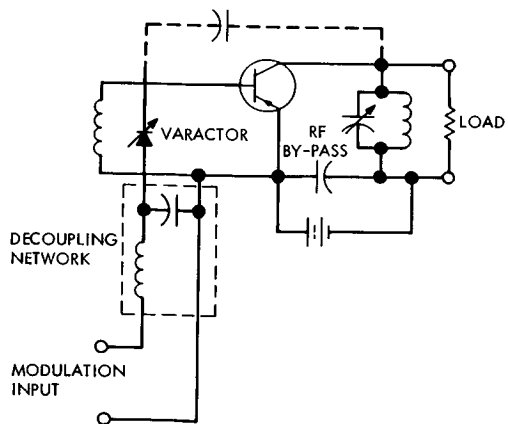


Figure C. Frequency Modulator

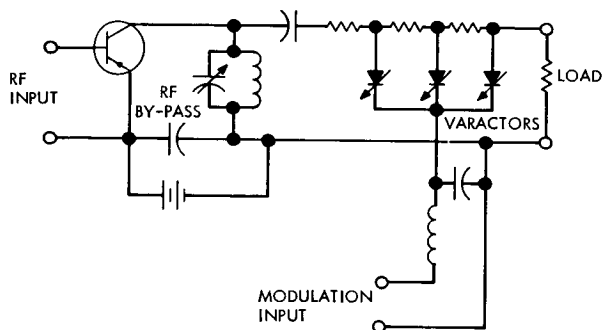


Figure D. Phase Modulator

DIGITAL AND COMPOUND MODULATION IMPLEMENTATION

Compound modulation combines several modulation steps in series, e.g. PCM/FM/PM.

The previous topic discussed amplitude frequency and phase modulation as methods where one time function is impressed on a second.

Digital Modulation

Before describing compound modulation one additional modulation method should be mentioned, namely pulse code modulation or PCM. PCM is a means of representing a given time waveform to an arbitrarily small error using a digital code word. The conversion of a time waveform to a PCM waveform consists of two basic steps, sampling and coding. In the sampling step the analog waveform is sampled at a rate higher than twice its highest frequency. The amplitudes of the samples are then represented as a code word. Clearly the bandwidth of a PCM signal is higher than the bandwidth of the input time waveform. An approximate* bandwidth is given by:

$$BW_{PCM} = (K_1) (f_m) (n) K_2$$

where K_1 is a constant greater than 2

f_m is the highest frequency in the time waveform

n is an exponent of 2** (e.g. 2^3)

K_2 is a constant to account for synchronizing pulses,
normally $1.00 < K_2 < 1.10$

Compound Modulation

In compound modulation several modulation processes are cascaded in series. An example is PCM/FM/PM. Here the original waveform is represented by a PCM binary code. This code in turn frequency modulates a sine wave. The frequency modulated sine wave then phase modulates a second sine wave. This type of modulation is often used in radio space applications where scientific data is converted to PCM using an analog-to-digital (A/D) converter. The PCM pulse stream then modulates a relatively low frequency subcarrier oscillator (SCO). The SCO then phase modulates the carrier. By using several subcarrier oscillators several channels may be imposed on a carrier. Similarly several channels may be incorporated in the PCM bit stream by using a commutator before the A/D converter.

*Varies with various filtering criteria

**For binary modulation

Many forms of compound modulation can be implemented, a few of the more common types are listed below.

PCM/FM	PCM/PM
PCM/FM/PM	FM/FM
PCM/AM	AM/FM

INTRODUCTION

The interaction of optical and electric or acoustic fields in certain optical materials is the basis for achieving a variety of types of laser beam modulation for useful applications in space communication systems.

Laser modulation is accomplished by passing the beam through an optically transmissive medium in which one or more of the optical transmission parameters is varied by the application of a modulating field. The interaction of the laser beam and the modulating field within the medium makes it possible to achieve a wide variety of types of optical modulation, including intensity, frequency, phase, and polarization.

The advent of lasers has motivated extensive research and development in optical modulation. Laboratory and commercial devices presently are available for obtaining all forms of modulation of visible and infrared lasers. Most of these devices are based on the use of the electro-optic effect or elasto-optic effect in crystals and liquids. Up until two years ago, the best immediately available techniques for modulating infrared lasers (beyond about 1.5 microns) involved the use of elasto-optic, or acoustic effects. However, advances in the technology of growing single-domain ferro-electric crystals, some of which exhibit a strong first-order electro-optic effect, promise a brighter future for modulation up to 5 microns by electro-optic means; and the recent advent of high resistivity gallium arsenide opens the way to practical modulation systems at 10.6 microns.

A detailed assessment of the performance and burdens of existing visible and infrared optical modulators may be somewhat misleading to the designer of future laser systems for space applications because the relevant technology is in a constant state of flux and probably will continue to be so for at least several years. The discovery and perfection of non-linear optical materials and the invention and perfection of modulation interaction structures and devices will lead to continuing improvement in performance and reduction of burdens.

At present, electro-optic modulators for lasers which operate in the wavelength range between about 0.4 and 10.6 microns can provide information bandwidths larger than the maximum believed to be required by space optical communication and tracking systems (i. e., 100 MHz of bandwidth). The burdens of weight and modulator driving power at the longer wavelengths, however, appear excessive. The problems associated with handling very high laser beam powers also have not yet been studied in depth, and may present some additional difficulties. However, advances both in the synthesis of better electro-optic materials, and in the design of modulator structures will provide optical modulators suitable for space communications systems within the next few years.

In the meantime, it would seem advisable to pursue further research in acoustic and other means for modulating infrared lasers above 1.5 microns. Present acoustic materials and techniques have serious limitations in achieving the larger information bandwidths that may be required (i. e., in extending present bandwidths of 1 to 5 MHz, up to 100 MHz).

Other physical processes which may be useful for optical modulation, such as controllable photon absorption in solid-state materials, and

magneto-optics, either are not very promising, or are in too early a state of research to evaluate accurately their potential usefulness.

These introductory remarks on optical modulation have been addressed exclusively to the final modulation process of impressing the information on the optical carrier. It should be pointed out that there may be very good reasons for impressing the signal information on a radio-frequency subcarrier, prior to performing the final optical modulation. All of the conventional modulation techniques may be utilized in the preliminary step. This will eliminate the necessity for impressing video or very low-frequency modulation components on the laser beam. The elimination of low-frequency components on the optical signal may provide important advantages both in the ease of design and driving of optical modulators, and in achieving satisfactory transmission through the earth's atmosphere. The latter factor is mentioned even though it is not yet certain that it will be either feasible or desirable to include the earth's atmosphere in the optical part of a space-earth link.

Finally, it should be pointed out that research and development remain to be done before the technology will exist to provide optical modulators with fully acceptable performance and burdens for space communication and tracking systems. The programs now in progress, if continued, may accomplish the desired results. This is in an area, however, which needs additional direction and support in order to assure that the results will be forthcoming on a suitable schedule.

SUMMARY OF OPTICAL MODULATORS

Advances in the technology of electro-optic modulator materials and design techniques permit a reasonably accurate assessment of expected performance.

It is difficult at this time to make accurate assessment of the performance characteristics of optical modulators which will have continuing value to the designer of laser systems for space communications and tracking. However, great progress has been made in the past few years in the development of optical modulation techniques and materials. This work has demonstrated that all forms of modulation can be impressed on optical carriers in the band between 0.4 and 10.6 μ . At wavelengths longer than 1.5 μ , optical modulation technology is in a somewhat more primitive state; but the advent of high grade semi-insulating GaAs as a useful electro-optic modulator material in the infrared region of 2 to 12 μ has constituted a notable breakthrough for CO₂ laser applications. Moreover, the technology of synthesizing superior grade LiNbO₃ and LiTaO₃ has made possible useful and efficient modulators in very broadband applications over the wavelength range of 0.4 to 5 μ . Research and development of acoustic modulation techniques is progressing at a moderate pace, and acoustic modulators in some instances offer a significant power advantage; but on the other hand they provide modulation bandwidths far short of those believed to be needed in optical space communication systems.

In considering communication and tracking systems using infrared lasers, it must be realized that there is an inherent inverse dependence of modulation efficiency on wavelength at a given driver power level. Thus, for example, a specific modulator element requires 10 times as much voltage at 5 μ as it does at 0.5 μ to achieve the same depth of modulation. In order to keep the power within reasonable limits, it is necessary to (1) extend the interaction length and/or (2) settle for less modulation scheme which can provide a high effective modulation index for intensity modulation and reasonable frequency deviations in optical FM systems with very modest levels of driver power, when the attendant bandwidth limitations are acceptable.

The table presents, in more or less chronological order, a series of electro-optic modulators and their operating characteristics. This listing is by no means complete, but an attempt has been made to itemize those modulators which represent significant advances in their particular regime of applications. Except where otherwise noted, the performance characteristics are applicable for an optical wavelength of 6328Å. The commercial device produced by Isomet Corporation reflects the advanced technology now reached in the field of electro-optic modulator design and production.

Characteristics of Some Electro-Optic Modulators - December 1968

Parameter	TM ₀₀ Mode Cavity Modulator	Traveling Wave Intensity Modulator	Polarization Modulator (NASA Contract)	Multi-Element Intensity Modulator	Resonant SSBSC Modulator	Traveling Wave SSBSC Modulator	Single Element PCM Modulator	10-6 μ Intensity Modulator	Commercial Modulator
Development Status	Built at BTL	Built at Pennsylvania	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at BTL	Built at RCA	Product of Isomet Corp.
Material	KDP	KDP	KDP	KDP	KDP	KDP	LiTaO ₃	GaAs	y-cut ADP
Crystal Dimensions	1 cm long x 2.5 mm dia.	100 cm long x 2 mm wide	50 x 0.4 x 0.4 cm	16 each 1/4" x 1/4" x 1/2"	2 each 1 1/2" long x 1/2" dia.	80 x 0.4 x 0.4 cm	1 x 0.025 x 0.025 cm	6.7 x 0.3 x 0.3 cm	20 x 0.4 x 0.4 cm
Optical Attenuation	0.1 dB	6 dB	1.5 dB	0.3 dB	1.0 dB	2.0 dB	1.5 dB	0.4 dB	0.4 dB
Useful Optical Range	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 to 5 μ	2 to 12 μ	0.25 to 0.75 μ
Modulation Frequency	3 Gc	Baseband	Baseband	Baseband	850 Mc	200 Mc	Baseband	Baseband	Baseband
3 dB Bandwidth	4 MHz	3 GHz	30 MHz*	10 MHz*	5 MHz	100 MHz	220 MHz*	100 MHz*	>100 MHz
Modulating Power	1.5 W	12 W	20 W	12 W	3 W	10 W	250 mW	70 W	50 W
Modulation Index	0.13	~1.0	0.5	0.5	0.01	0.3	0.4	0.1	0.5
Extinction Ratio	Unknown	Unknown	8:1	250:1	NA	NA	80:1	~80:1	70:1
Weight	Unknown	Unknown	20 lbs	2 lbs	15 lbs	25 lbs	Unknown	Unknown	Unknown
Size	Unknown	~100 in ³	144 in ³	30 in ³	850 in ³	216 in ³	Unknown	12 in ³	80 in ³

* 3dB Bandwidth limited by modulator driver and not by the electro-optic structure.

OPTICAL MODULATORS

Electro-Optic Modulation

	Page
Theory of Electro-Optic Modulation	134
Bias Point and Driver Level Considerations	136
Design of Optical Intensity Modulators	138
Design of Electro-Optic Frequency Translators	142
Bandwidth Limiting Factors in Lumped-Element Modulators	146
Traveling Wave Electro-Optic Modulators	148

THEORY OF ELECTRO-OPTIC MODULATION

Optical modulation in various forms may be achieved using materials which exhibit the electro-optic effect.

The phenomenon of electrically-induced birefringence exhibited by most optical materials provides the mechanism for electro-optic modulation of optical beams. Application of a properly-directed electric field within the material gives rise to a perturbation of its refractive properties, characterized by two unique orthogonal directions, called the "fast" and "slow" axes. Optical beams plane-polarized along these two axes and directed normal to their plane, will travel at different velocities, determined by their respective indices of refraction. After traversing a length L of the material, a single optical beam initially plane-polarized at 45 degrees to these principal axes will emerge in general elliptically polarized, since its principal components will have suffered a relative optical phase shift, Γ , of

$$\Gamma = \frac{2\pi L}{\lambda} \Delta N \text{ radians,}$$

where λ is the free-space optical wavelength and ΔN is the difference between the fast and slow refractive indices. If the emerging beam is passed through a linear polarizer (analyzer) whose preferred direction is perpendicular to that of the incident beam polarization, some of the elliptically-polarized beam will be transmitted. Its intensity I , relative to the incident intensity I_0 , is given by

$$I = I_0 \sin^2 \left(\frac{\Gamma}{2} \right),$$

assuming no losses in the material. Variations in Γ resulting from electro-optic changes in the birefringence ΔN thus give rise to intensity modulation of the incident beam. If the analyzer is omitted, a receiver equipped with suitable polarization isolators and matched detectors can process the beam in either a phase-difference modulation or polarization modulation mode. Finally, if the modulator is designed to provide a real or simulated rotation of the principle axes of induced birefringence, optical frequency translation may be achieved.

Depending upon the nature of the material, the electro-optic effect may take different functional forms. In liquids and some solids, the changes in refractive indices exhibit a quadratic dependence upon the field; this is designated the Kerr effect.¹ Crystalline solids which are piezoelectric display a linear relationship known as the Pockels effect.² Explicit relations for some common classes of electro-optic materials are shown in the Table.

¹Jenkins, F. A. and White, H. E., Fundamentals of Optics, Third Edition, McGraw-Hill, New York, pp. 604-605, 1957.

²West, C. B. and Carpenter, R. O'B., in American Institute of Physics Handbook, Section 6, McGraw-Hill, New York, pp. 94-97, 1957.

Properties of Some Typical Electro-Optic Materials

Crystal Class or Material	Retardation Equation	Examples	Indices of Refraction	Electro-Optic Coefficients
$\overline{4}2m$	$\Gamma = \frac{2\pi L}{\lambda} N_o^3 r_{63} E$	KDP	$N_o = 1.51$	$r_{63} = 1.03 \times 10^{-11} \text{ m/volt}$
		KD*P	$N_o = 1.51$	$r_{63} = 2.6 \times 10^{-11} \text{ m/volt}$
$\overline{4}3m$	$\Gamma = \frac{2\pi L}{\lambda} N_o^3 r_{41} E$	Cu Cl	$N_o = 1.93$	$r_{41} = 6.14 \times 10^{-12} \text{ m/volt}$
		GaAs	$N_o = 3.31$ (at 10 microns)	$r_{41} = 1.6 \times 10^{-12} \text{ m/volt}$
3m	$\Gamma = \frac{2\pi L}{\lambda} N_o^3 r_{22} E$ (field parallel to b-axis) $\Gamma = \frac{\pi L}{\lambda} N_o^3 \left[\left(\frac{N_e}{N_o} \right)^3 r_{33} - r_{13} \right] E$ (field parallel to c-axis)	LiNbO ₃	$N_o = 2.241$ $N_e = 2.158$ (at 1 micron)	$r_{22} = 6.7 \times 10^{-12} \text{ m/volt}$ $\left(\frac{N_e}{N_o} \right)^3 r_{33} - r_{13} = 1.73 \times 10^{-11} \text{ m/volt}$
		LiTaO ₃	$N_o = 2.139$ $N_e = 2.143$ (at 1 micron)	$r_{22} \approx 1 \times 10^{-12} \text{ m/volt}$ $\left(\frac{N_e}{N_o} \right)^3 r_{33} - r_{13} = 2.16 \times 10^{-11} \text{ m/volt}$
m3m	$\Gamma = \frac{\pi L}{\lambda} N_o^3 (g_{11} - g_{12}) \epsilon^2 E^2$	KTN	$N_o = 2.29$	$(g_{11} - g_{12}) \epsilon^2 = 1.36 \times 10^{-15} (\text{m/volt})^2$
Kerr Fluids	$\Gamma = 2\pi B L E^2$	Nitrobenzene	$N_o = 1.55$	$B = 2.5 \times 10^{-12} \text{ m/volt}^2$
<p>E is the electric field</p> <p>n_{ij} are the Pockels constants (linear electro-optic coefficients)</p> <p>N_o is the ordinary refractive index</p> <p>N_e is the extraordinary refractive index</p> <p>The electro-optic coefficients for the crystals are the "unclamped", zero stress values obtained with dc or low frequency fields.</p> <p>g_{11} and g_{12} are the quadratic electro-optic coefficients in the cubic axes system.</p> <p>ϵ is the permittivity</p> <p>B is the Kerr constant</p>				

BIAS POINT AND DRIVER LEVEL CONSIDERATIONS

The transfer characteristics of electro-optic modulators depend upon the quiescent or bias point and the modulation drive level.

It is important to be familiar with the characteristics of the modulated output beam of electro-optic modulators as a function of both the quiescent operating point and the level or depth of modulation. Consider in particular a beam intensity modulator whose output I as a function of the input I_0 and optical phase shift Γ is given by

$$I = I_0 \sin^2 \frac{\Gamma}{2}$$

In some applications such as optical printing, displays, or perhaps digital modulation, it is customary to modulate the beam from a condition of zero intensity ($\Gamma = 0$) to various levels of intensity. Practical systems will customarily operate with peak values of Γ less than about one radian, in which case it is valid to make the approximation $\sin (\Gamma/2) \cong \Gamma/2$, yielding

$$I = I_0 \frac{\Gamma^2}{4}$$

This shows that for modulators using linear electro-optic crystals, the output intensity is proportional to the square of the applied electric field; in other words, the output optical power is proportional to the input driver power.

It is clear that with zero bias, an rf sine wave driver signal producing $\Gamma = \Gamma_0 \sin \omega_m t$ would yield an intensity modulated optical output beam containing only the second harmonic of ω_m , since

$$\frac{I}{I_0} \cong \frac{1}{4} \Gamma_0^2 \sin^2 \omega_m t = \frac{1}{8} \Gamma_0^2 (1 - \cos 2\omega_m t).$$

In intensity modulation with sine wave rf drive, minimum harmonic distortion is achieved by biasing the incident beam with a quarter-wave plate to give circular polarization. In this case

$$\Gamma = \frac{\pi}{2} + \Gamma_0 \sin \omega_m t,$$

and

$$\frac{I}{I_0} = \frac{1}{2} \left[1 + \sin (\Gamma_0 \sin \omega_m t) \right] = \frac{1}{2} \left[1 + 2J_1 (\Gamma_0) \sin \omega_m t + 2J_3 (\Gamma_0) \sin 3\omega_m t + \dots \right],$$

where J_n is the n th order Bessel function. When $\Gamma \leq \pi/6 \cong 0.5$, this can be written in the approximate form

$$\frac{I}{I_0} = \frac{1}{2} \left[1 + \Gamma_0 \sin \omega_m t \right].$$

In this case $\Gamma_0 \equiv m$, the usual designation of modulation index for intensity-modulated waves. More generally, however, there is a modulation index $m_i = 2J_i(\Gamma_0)$ for the fundamental and higher harmonics. For convenience some values of m_i for a selected range of Γ_0 are listed below.

Γ_0	m_1	m_3
0.25	0.248	0.001
0.50	0.485	0.005
$\pi/4$	0.726	0.019
1.00	0.880	0.039
1.25	1.021	0.074
$\pi/2$	1.134	0.138

These figures show that if an attempt is made to make $\Gamma_0 = \pi/2$, the modulated wave contains almost 14 percent third-harmonic power. More significantly, since the driver power is proportional to Γ_0^2 , it takes four times as much power to achieve this modulation index as it does to achieve $\Gamma_0 = \pi/4$ and nine times as much as that for $\Gamma_0 = \pi/6$. Note also that this last depth of modulation introduces only about 1/2 percent third-harmonic distortion.

DESIGN OF OPTICAL INTENSITY MODULATORS

Special design schemes may be utilized to construct practical electro-optic modulators requiring only modest levels of driver voltage.

The optical modulators used as examples for the present discussion employ KDP as the active electro-optic element. The customary and only practical mode of optical modulation using KDP and its isomorphs is that which applies the modulating signal field along the crystalline c-axis (optic axis). If it is desired to avoid the natural birefringence of this class of material and at the same time realize maximum electro-optic phase retardation, the light beam must also travel parallel to the c-axis. Under the influence of a signal field, the principal axes of induced birefringence lie in the a-b plane at ± 45 degrees relative to the crystalline axes. The maximum field strength E_m in terms of the voltage V_m is $E_m = V_m/L$, where L is the length of crystal along the field (and optical beam). From the Table, given in the topic entitled "Theory of Electro-Optic Modulation,"

$$\Gamma_o = \frac{2\pi N_o^3 r_{63} V_m}{\lambda} \quad (1)$$

The output phase difference, Γ , of the principal optical components is therefore independent of the length of the crystal. To reach the maximum $\Gamma (= \pm \pi/2)$ in KDP, for light of wavelength $\lambda = 6328 \text{ \AA}$, requires a signal voltage amplitude $V_m = 5.5 \text{ kV}$, which is obviously high for practical operation.

Two types of modulators are discussed below which are designed to circumvent the above limitation. Both systems provide a long interaction length with modulating fields sufficiently strong to yield ample modulation depth at low levels of driving voltage.

The first modulator is a multielement device consisting of a linear chain of small KDP crystals separated by thin, flat-ring electrodes. This is shown schematically in Figure A. Figure B is a photograph of one such unit containing 16 elements. Alternate electrodes carry common signal voltage polarity; therefore, adjacent crystal elements have opposing fields. Since the fast and slow axes of induced birefringence exchange roles upon reversal of the field, alternate crystal elements are rotated 90 degrees about their common c-axis, so that the instantaneous fast and slow axes all have a common direction throughout the chain. The effect of this arrangement is to produce a cumulative phase shift between principal components of the light beam which is the sum of the separate elemental phase shifts. By the same token, the driving voltage required to produce a given total phase shift is $1/n$ times the voltage needed to produce the same phase shift in a single element, where n is the number of elements.

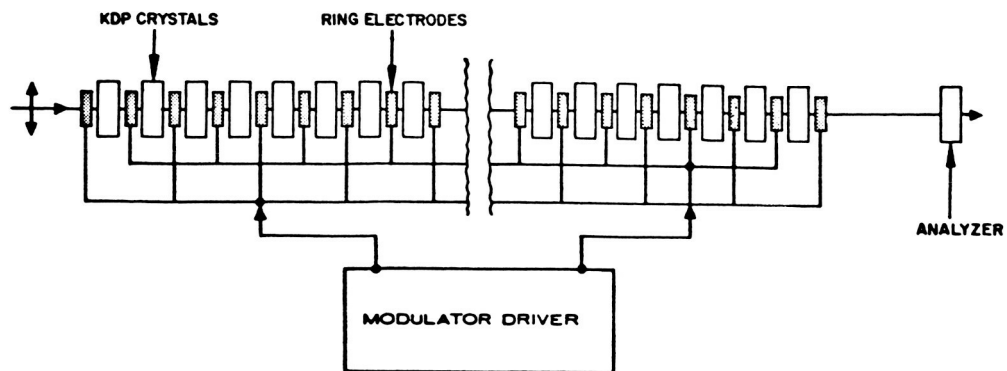


Figure A. Multielement Optical Modulator

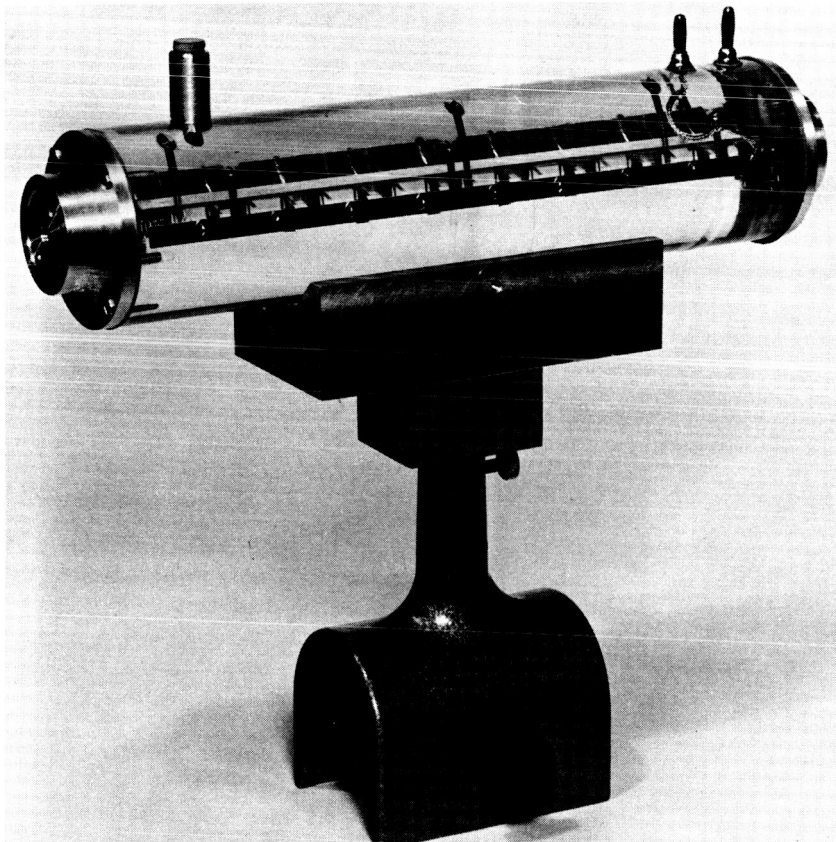


Figure B. Sixteen-Element KDP Crystal Modulator

DESIGN OF OPTICAL INTENSITY MODULATORS

The second, or alternative, modulator design has quite a different configuration.¹ It consists of two long, narrow, parallel-strip condensers filled with KDP. These are made as nearly identical as possible and are oriented in tandem along the optical beam axis. The electric fields are applied as before along the crystalline optic axis; however, this direction is now transverse to the beam, and natural birefringence must be taken into account. The purpose of having two similar units is to effect a cancellation of this natural birefringence by rotating the second unit 90 degrees relative to the first. Maximum optical phase shift is achieved if the fast axis (x-axis) of induced birefringence in the first unit is directed transverse to the beam and parallel to the plane of the electrodes, and if the slow axis (y-axis) in the second unit is similarly oriented. The incident light beam is plane polarized at 45 degrees to the optic axis (z-axis). Figure C illustrates schematically this modulator system. A simplification in the design of this configuration can be made which allows the use of a single pair of electrodes.² The two crystal elements are aligned with their z-axes parallel and are separated by a $\lambda/2$ plate, which is the optical equivalent of the original 90 degree rotation.

The optical phase shift Γ_o produced by this type of modulator is

$$\Gamma_o = \frac{2\pi}{\lambda} \left[LN_o^3 r_{63} E_m + \Delta L(N_e - N_o) \right]$$

in which L is the average value of the separate crystal lengths and ΔL is their length difference. An optical compensator,² as shown in Figure C, can be incorporated into the modulator and adjusted to cancel out the residual birefringence resulting from ΔL . The signal-dependent expression for Γ is:

$$\Gamma_o = \frac{2\pi LN_o^3 r_{63} V_m}{\lambda t} \quad (2)$$

where again V_m is the signal voltage amplitude and t is the electrode, or plate, separation. Equation (2) is the same as Equation (1) for a single element longitudinal field modulator except for the factor L/t . Such an enhancement of the field strength is a cogent reason for considering this form of modulator.

Since the birefringence of most uniaxial crystals is temperature dependent, the second term in the above expression for Γ_o will cause the operating bias point to vary as the ambient temperature changes. With the KDP value of temperature coefficient,² the result is

$$\frac{\Delta \Gamma}{\Delta T} = 1.09 \Delta L \text{ rad/}^\circ\text{C} \quad (\Delta L \text{ in cm.})$$

No technological difficulty is anticipated in achieving a ΔL considerably less than 0.10 mm.

$$\frac{\Delta \Gamma}{\Delta T} < 0.01 \text{ rad/}^\circ\text{C}$$

¹ A ruby laser Q-switch employing the same principle of operation has been described by J. L. Wentz, Proc. IEEE 52, p. 716, 1964.

² Peters, C. J., NEREM Record, p. 70, 1964.

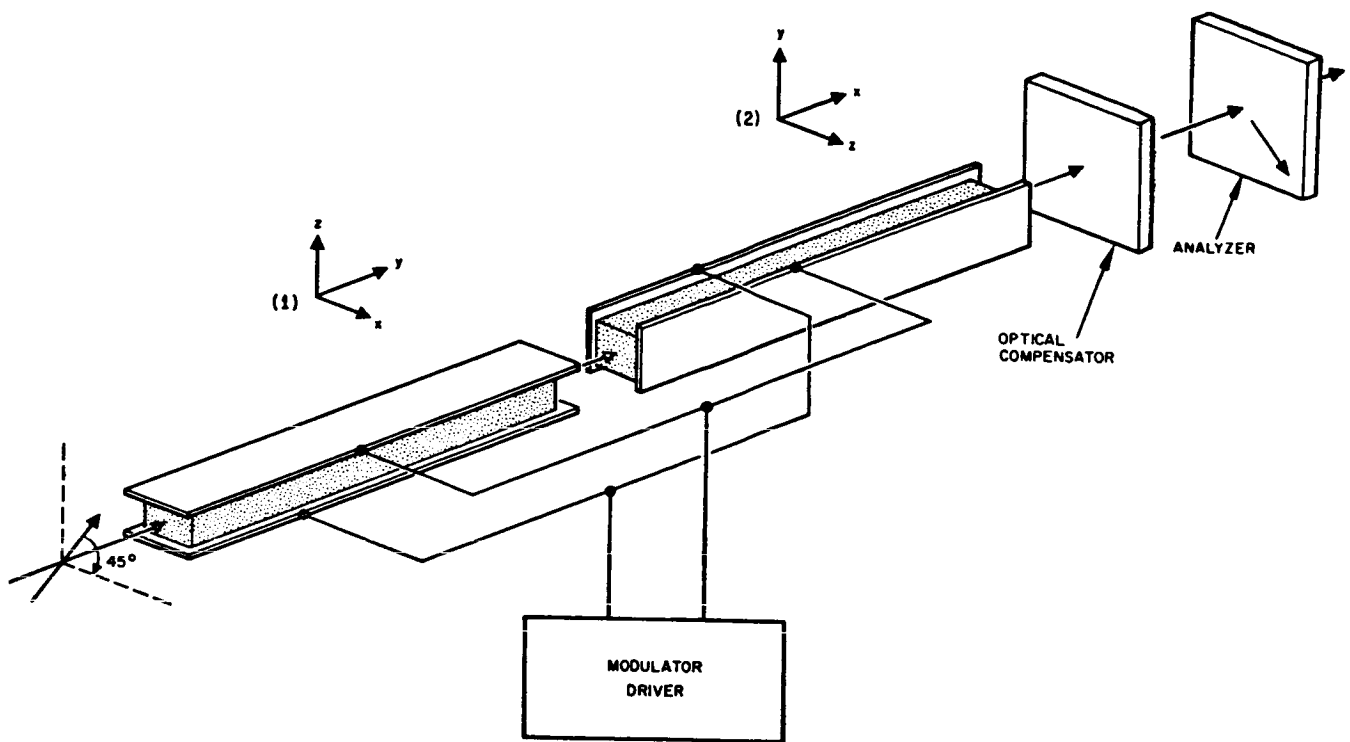


Figure C. Two-Element Parallel-Strip Optical Modulator

DESIGN OF ELECTRO-OPTIC FREQUENCY TRANSLATORS

Electro-optic frequency translation is effected by generating a real or simulated rotation of the principal axes of induced birefringence.

A laser heterodyne communication link normally employs two identical lasers, one being translated in frequency. Either could be used as the local oscillator, with the other serving as transmitter carrier. Optical frequency translation can be affected by several techniques. One method for Single Sidelobe Suppressed Carrier (SSBSC) operation is the electro-optic rotating birefringent plate.¹ In some forms, this instrument is capable of completely transforming a circularly polarized light beam into a beam of the opposite polarity and shifted in frequency by twice the frequency of rotation of the wave plate. More generally, however, only a portion of the original beam is thus shifted, and suppression of the carrier is affected by circular polarizer. Cubic crystals of a symmetry class $\bar{4}3m$ and $m\bar{3}m$, when driven by a rotating electric field in their (111) crystal plane, exhibit a true rotating birefringence in that plane, and thus produce frequency translation of a beam of light traveling along the (111) direction. The most popular and readily available electro-optic crystals, of class $\bar{4}2m$ such as KDP, do not have the above property; and thus an alternate technique is used with them.^{2,3,4} Two crystals are oriented in tandem along the optic beam and driven in phase quadrature. These elements are oriented with their principal axes of induced birefringence at an angle of 45 degrees with respect to each other. For incident circularly-polarized light this arrangement simulates a rotating birefringent plate, producing a component of circular polarization of the opposite sense and shifted by the signal frequency.

The one-element SSBSC modulator constitutes a true rotating birefringent element which serves the function of optical frequency translation by adding a constant rate of change of phase to a wave of given input frequency. It is thus the optical analog of the conventional microwave phase shifter which utilizes a mechanically rotated birefringent half-wave plate. To realize such an optical device requires in practice the use of electro-optic crystals of special symmetry classes: $\bar{4}3m$, such as CuCl or GaAs, $3m$, such as LiNbO_3 , or $m\bar{3}m$, such as $\text{KTa}_{0.65}\text{Nb}_{0.35}\text{O}_3$ (KTN) or other similar perovskites operating in their paraelectric phase. The first two classes possess a first order electro-optic (Pockels) effect, the third, a quadratic (Kerr) effect. For classes $\bar{4}3m$ and $m\bar{3}m$ the crystals are driven by an electric field rotating in the crystalline (111) plane, while the optical beam, circularly-polarized, travels through in the (111) direction. For class $3m$, the field is in the (001) plane; the beam direction is (001).

¹ Buhrer, C. F., Bloom, L. R., and Baird, D. H., *Applied Optics*, 2, p. 839, 1963.

² Buhrer, C. F., Fowler, V. J., and Bloom, L. R., *Proc. IRE*, 50, p. 1827, 1962

³ Buhrer, C. F., *Proc. IEEE*, 52, p. 969, 1964.

⁴ Targ, R., *Proc. IEEE*, 52, p. 303, 1964.

Under the influence of a signal frequency f_m rotating, say, clockwise, the induced axes of birefringence of the $3m\bar{3}$ type rotate with the field at frequency f_m , while those of the $\bar{4}3m$ and $3m$ type rotate counterclockwise at a frequency $(1/2)f_m$. Since, as noted previously, the optical beam undergoes a frequency translation equal to twice the rotation rate of the wave plate, this shift can equal either f_m or $2f_m$ depending upon the class of crystal employed. In either case, if the amplitude of the applied signal voltage is sufficient to produce a full half-wave of induced birefringence, the optical frequency translation is complete. Figures A and B illustrate two possible embodiments of single-element frequency translators.

The desirability of employing the two-element modulator for optical frequency translation is dictated mainly by the crystal characteristics of the more readily available and high quality materials of class $\bar{4}2m$, such as potassium dihydrogen phosphate (KDP) and its isomorphs. Because this type is naturally birefringent, the simplest arrangement is to orient the light beam along the optic axis, in order to avoid this unwanted birefringence in the absence of modulating signal. For optimum electro-optic effect it is also necessary to apply the signal field along the c-axis; thus the beam and signal field are parallel. With this configuration the induced axes of birefringence lie in the (001) plane at a fixed orientation of 45 degrees relative to the a and b axes, exchanging roles of "fast" and "slow" axis upon reversal of the signal field.

The action of a simulated rotating birefringent plate is provided by orienting two identical crystal elements in tandem along the beam axis and so oriented relative to each other that the corresponding a and b axes (or equivalently the induced principal axes) lie at 45 degrees with respect to each other. The modulating signal is applied to the two elements in phase quadrature. This configuration behaves like a single wave plate rotating at half the frequency of the applied signal, so that the optical frequency is translated by an amount equal to that of the signal. Figure C illustrates this particular system under low frequency operation. At microwave frequencies, the simple capacitive electrodes and associated circuit is normally replaced by suitable reentrant or cylindrical cavities. In some cases the crystals are driven in phase but separated by a distance equal to a quarter of a signal wavelength so that the optical photons automatically experience the quadrature driving fields by virtue of their transit time.

Complete disappearance of the carrier occurs not at half-wave voltage when the optical relative phase retardation Γ_O is $\pi/2$, as in the case of the single-element device, but rather when $J_0(\Gamma_O/\sqrt{2}) = 0$ or when $\Gamma_O = 3.40$. Here J_0 is the Bessel Function of zero order. Also, in addition to the first order single sideband, an upper and lower second-order sideband are generated; also a single third-order sideband, etc. The optical amplitude of these sidebands are proportional to their respective Bessel Functions of the argument $\Gamma_O/\sqrt{2}$.

DESIGN OF ELECTRO-OPTIC FREQUENCY TRANSLATORS

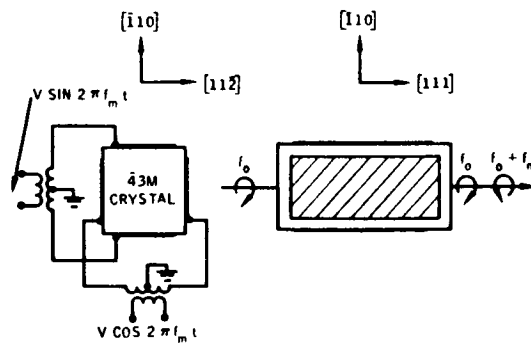


Figure A. Low Frequency One-Element Optical Frequency Translator

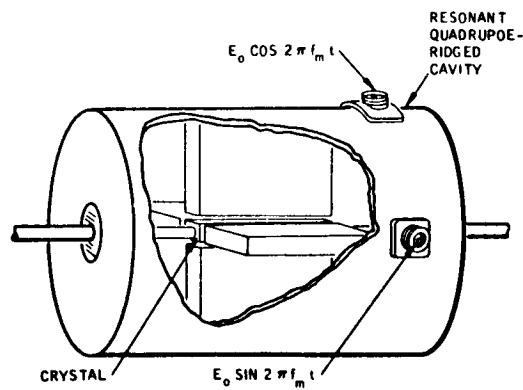


Figure B. Possible Microwave Optical Frequency Translator

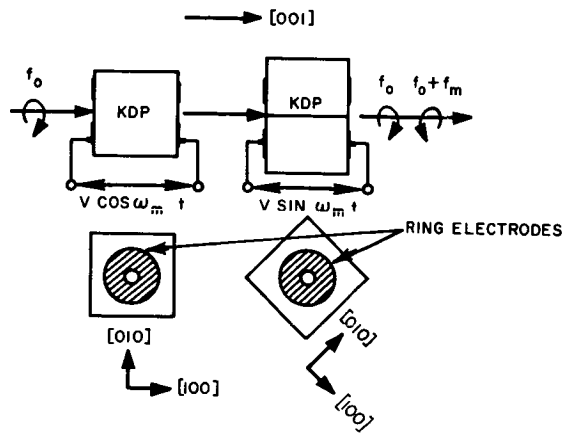


Figure C. Two-Element KDP
Optical Frequency Translator
For SSBSC Modulator

BANDWIDTH LIMITING FACTORS IN LUMPED-ELEMENT MODULATORS

Optical photon transit time and excessive load capacitance are the two basic factors which limit the operating bandwidth of lumped-element electro-optic modulators.

In order to realize a substantial optical phase shift, and thus a large modulation depth, one is tempted to relieve a voltage or power burden by resorting to long interaction lengths. In lumped-element modulators, however, there are two principal bandwidth-limiting factors which restrict the length of active material. These are optical transit time and load capacitance.

Optical transit time effects become appreciable when the signal frequency is sufficiently high to have a period comparable with the time required for an optical wavefront to traverse the electro-optic medium. For maximum phase shift at a given signal level, an optical wavefront should see a constant signal field strength during its transit of the crystal medium. Obviously, however, this cannot obtain if the signal frequency is so high that the field polarity actually changes sign during this transit time. It can be shown that the effective optical phase shift Γ_{eff} has a frequency dependence given by

$$\Gamma_{\text{eff}} = \Gamma_m \frac{\sin u}{u},$$

where

$$u \equiv \frac{\omega_m L}{2v_o} = \frac{\pi N_o L}{\lambda_m}, \quad (1)$$

and where

$\Gamma_m \equiv$ maximum (low frequency) phase shift

$\omega_m/2\pi \equiv$ signal frequency

$v_o \equiv$ optical wave velocity

$N_o \equiv$ optical refractive index

$\lambda_m \equiv$ free space signal wavelength.

As a specific example, when the optical transit time L/v_o is a quarter of a signal period - or equivalently, when the optical path length $N_o L$ is one quarter of a free space signal wavelength - then $u = \pi/4$,

$$\sin u/u = 0.707/0.785 = 0.90,$$

and the phase shift has been reduced by 10 percent. Consider, for example, a two-element transverse field modulator. For an active length of 50 cm, the frequency at which u becomes equal to $\pi/4$ is 100 MHz for KDP-type materials ($N_0 \approx 1.5$) and 64 MHz for LiNbO_3 ($N_0 \approx 2.3$).

Except for those crystals which have a relatively low dielectric permittivity, a more stringent bandwidth-limiting factor than optical transit time is the load capacitance presented by the lumped-element modulator to the signal driver. For example, a 50-cm KDP modulator having square cross-section has a capacitance of 90 pF, neglecting strays. To operate this up to 100 MHz would require a driver with maximum output impedance of only about 35 ohms, even with critical inductive video peaking. It would also, incidentally, require nearly 45 watts to modulate a 5000 Å beam 50 percent, assuming a 5 mm electrode separation. Indeed, more generally it can be shown that if optical transit time is chosen as the limiting criterion for either bandwidth or modulator length, then the driver output impedance depends only on optical velocity and crystal dielectric. Arbitrarily choosing $u = \pi/4$ as the maximum acceptable value, then from Equation (1), $\omega_2 L$ is:

$$\omega_2 L = \frac{\pi v_0}{2}, \quad (2)$$

where $\omega_2/2\pi = f_2$ is the upper frequency limit. A signal source capable of driving a purely capacitive load over a video bandwidth f_2 must have an output impedance less than approximately

$$R = \frac{2}{\omega_2 C} = \frac{2}{\omega_2 \epsilon L} \quad (3)$$

with optimized video peaking. Again a modulator of square cross-section is assumed having a capacitance $C = \epsilon L$, where ϵ is the permittivity of the electro-optic material. Combining Equations (2) and (3), the following is obtained:

$$R = \frac{4}{\pi \epsilon v_0},$$

which was to be proved. This result gives $R = 35$ ohms for KDP, 15 ohms for KD^*P , and 40 ohms for LiNbO_3 .

TRAVELING WAVE ELECTRO-OPTIC MODULATORS

The electrical characteristics of terminated traveling wave modulator structures afford a substantial increase in operating frequencies and bandwidths.

Limitations on bandwidth imposed by load capacitance and transit time effects are circumvented by resorting to a TEM parallel strip traveling wave structure, terminated by its characteristic impedance. Figure A shows the general form taken by a modulator of this type, using a KDP class of material. The dual-element configuration noted earlier can be adapted to a single electrode pair simply by incorporating a half-wave plate between the two modulator units. It causes the vertical and horizontal components of the beam to exchange roles, thereby simulating a 90-degree rotation of the second modulator section. With the positive optic axes (z-axes) of the two sections oppositely directed as shown, the axes of induced birefringence indicated by x are fast and slow axes, respectively.

By proper design, this type of modulator can be made to have a signal wave velocity which matches that of the beam. Thus any given optical wavefront sees a field strength which is constant throughout the entire modulator length, and transit time is no longer a controlling factor. The signal wave velocity can be adjusted by a suitable choice of the dielectric filling factor, that is, the ratio of crystal width to strip width.¹

An upper limit of useful bandwidth is ultimately reached when the frequency is so high that the mode of transmission is no longer TEM. This occurs when the signal wavelength is comparable with the width of the dielectric. Under this condition the electric field amplitude is no longer constant across the wave structure; rather the mode becomes compressed within the dielectric, and the wave velocity is determined solely by the crystal dielectric permittivity.¹ This means the wave is slowed down and the velocity matching is lost. Typically the upper frequency limit at which such a degeneration of the TEM mode occurs is 2 to 3 GHz² with KDP type materials.

A second example of a parallel strip traveling wave electro-optic modulator is illustrated in Figure B for the case of a cubic 43m class of crystal, such as GaAs. This particular configuration is stressed for two reasons: (1) GaAs is a recent development for practical modulators of the 10.6 μ CO₂ laser output and is available in good crystal sizes of acceptable optical quality and modest IR attenuation, and (2) the optical and r-f dielectric permittivities are so nearly equal that one is able to achieve a good velocity-matched condition with a fully-loaded wave structure, thus minimizing the power burden.

The indicated orientation of the crystalline axes represents the optimum configuration for the intensity-modulation mode of operation. The incident vertically-polarized laser beam is transformed by a quarter-wave plate into circular polarization in order to set the proper optical bias for best linear operation. By the action of a vertical modulation field along the [110] direction, the original circular polarization is rendered elliptical. That part of the beam which emerges from the vertically-oriented analyzer exhibits the desired intensity modulation according to the impressed signal.

¹Kaminow, I.P., and Liu, J., Proc. IEEE, 51, p. 132, 1963.

²Peters, C.J., Proc. IEEE, 51, p. 147, 1963.

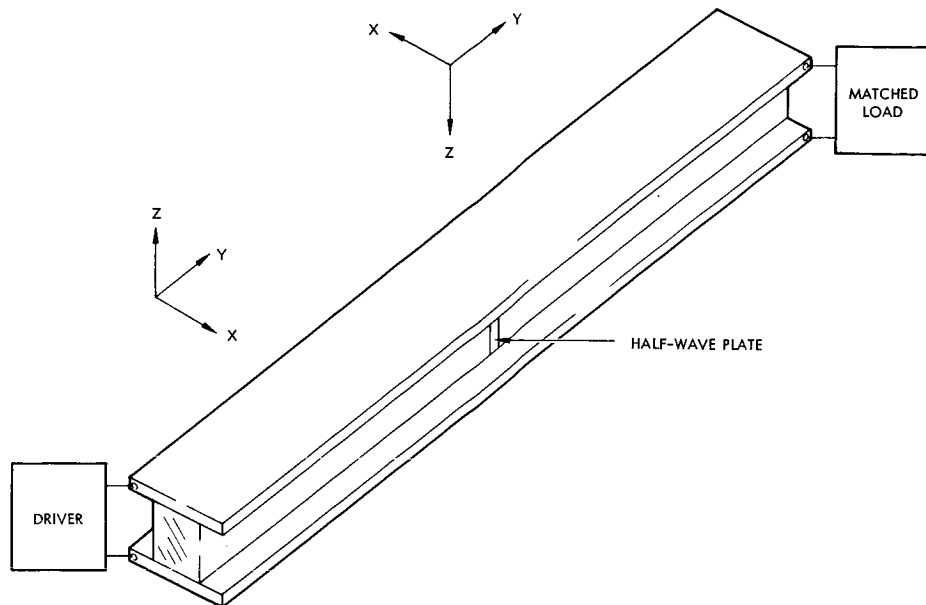


Figure A. Parallel Strip Traveling-Wave Electro-Optic Modulator Structure using KDP-Type Material

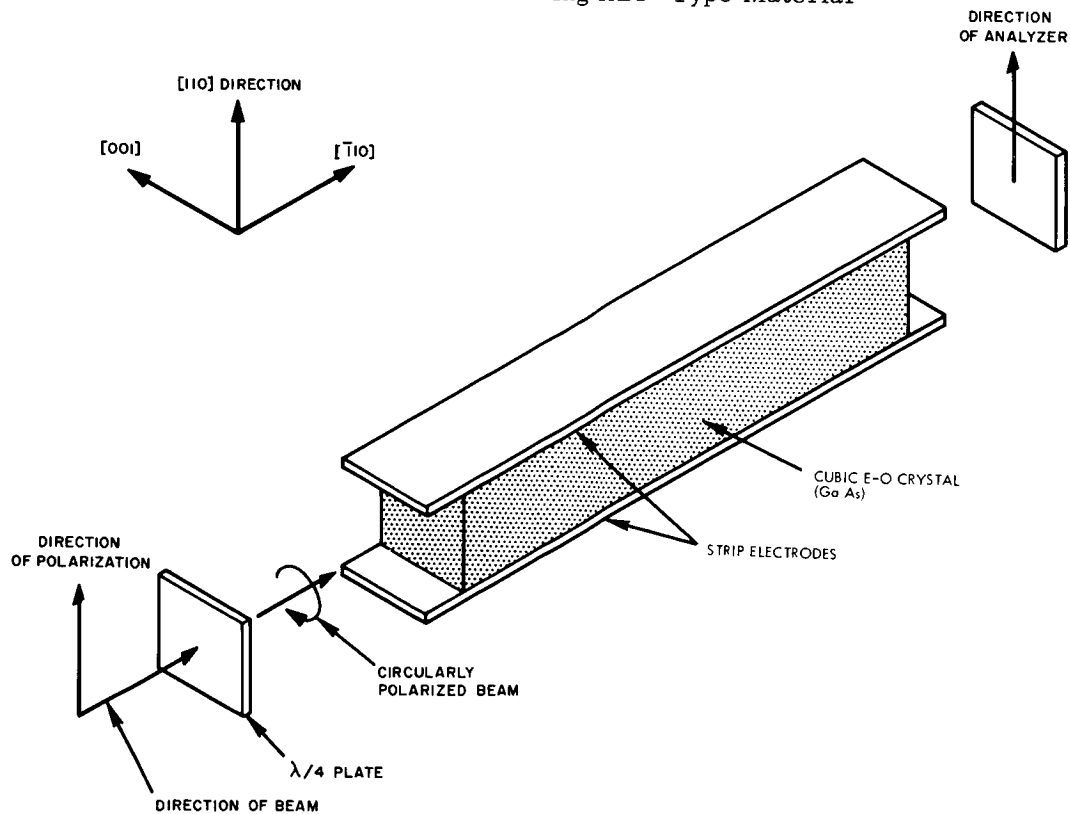


Figure B. Parallel Strip Traveling-Wave Modulator Configuration Using a Cubic 43m Crystal

OPTICAL MODULATORS

Elasto-Optic Modulation

	Page
Theory of the Elasto-Optic Effect	152
Theory of Ultrasonic Diffraction of Light	156
Acoustic Modulation Techniques	158
Properties of Ultrasonic Modulators	160
Summary and Tabulation of Elasto-Optic Modulators	164

THEORY OF THE ELASTO-OPTIC EFFECT

Variations of refractive index in optical materials may be strain-induced by propagation of acoustic waves.

The phenomenon of diffraction of light by the passage of a sound wave through the optical medium is largely governed by the change in refractive index induced in the medium by the passage of the sound wave and also by the radiation geometry of the sound and light wave sources. This section lists the equations that relate the change in refractive index to the stress-strain amplitude generated by the ultrasonic wave.

The elasto-optical coefficients for an anisotropic medium p_{ijkl} are defined* by^{1, 2}

$$\Delta B_{ij} = p_{ijkl} \eta_{kl}$$

where η is the strain and B_{ij} is the dielectric impermeability tensor, which is related to the dielectric constant κ_{ij} by

$$\sum_j \kappa_{ij} B_{jk} = \delta_{ik}$$

where δ_{ik} is the Kronecker delta. For the case of cubic crystals κ is isotropic and thus

$$\Delta \kappa_{ij} = -\kappa^2 \Delta B_{ij} = -\kappa^2 p_{ijkl} \eta_{kl} \dots$$

Furthermore, if one restricts the problem to the propagation of longitudinal sound waves (no shear components) along principal crystallographic directions only, the strain component can be represented by a single subscript and the change in dielectric constant can be simply represented by the following equation

$$\Delta \left(\frac{1}{\kappa} \right)_i = \sum_j p_{ij} s_j$$

¹Nye, J. F., Physical Properties of Crystals, Oxford, New York; 1960.

²Krishnan, R. S. and Viswanathan, S., Progress in Crystal Physics, Madras; Central Arts Press, 1958.

* Double summation over indices k and l is implied.

where s_j , the strain tensor, is a special case of η_{kl} , and p is the fourth rank elasto-optic tensor.

If one passes a longitudinal sound wave along the $[1, 0, 0]$ crystallographic axis of a simple cubic crystal such as silicon or germanium and this is labeled the $j = 1$ direction, the strain subscript will be (s_1) . For the case of light wave polarization parallel to the direction of sound propagation the elasto-optic tensor p_{11} will apply; for orthogonal polarization the tensor p_{21} will apply. Thus for parallel and orthogonal polarization relative to the direction of propagation of a pure longitudinal $[1, 0, 0]$ ultrasonic wave the following expressions are obtained for the change in refractive index.

$$\begin{aligned} \Delta\left(\frac{1}{\kappa}\right)_1 &= p_{11}s_1 && \text{(Parallel)} \\ \Delta\left(\frac{1}{\kappa}\right)_2 &= p_{21}s_1 && \text{(Perpendicular)} \end{aligned} \tag{1}$$

Due to the symmetry of a cubic crystal the directions $[0, 1, 0]$ and $[0, 0, 1]$ have the same elasto-optic behavior. However for the case of sound propagation along the $[1, 1, 1]$ crystallographic direction the elasto-optic constant is a linear combination of p_{11} , p_{21} and p_{44} , the later term being generated by shear stresses. For the case of a rotation of coordinates where the Z direction is the $[1, 1, 1]$ crystallographic direction represented by $j = 3'$, the Y direction lies in the $[1, 1, 0]$ plane and is represented by $j = 1'$, one finds for sound along $j = 3'$ and light polarization along $j = 1'$ and along $j = 3'$, that

$$\begin{aligned} p_{1'3'} &= \frac{1}{3} p_{11} + \frac{2}{3} p_{12} - \frac{1}{3} p_{44} && \text{perpendicular} \\ &&& \text{polarization} \\ p_{3'3'} &= \frac{1}{3} p_{11} + \frac{2}{3} p_{12} + \frac{2}{3} p_{44} && \text{parallel} \\ &&& \text{polarization} \end{aligned} \tag{2}$$

In more complicated crystals, such as uniaxial or biaxial, the unabridged notation applies and can be used in an analogous manner. The preceding relations indicate that the elasto-optic behavior of a single cubic crystal is completely described by three constants p_{11} , p_{12} , and p_{44} . These constants can only be measured by inducing strain in the crystal and determining the resultant change in index of refraction. Measurements have been made on the elasto-optic constants of selected crystals in the visible region of the spectrum and at infrared wavelengths. Work has been in progress at Hughes Research Laboratories and elsewhere to evaluate the elasto-optic behavior of various materials at 3.39 and 10.6 microns.

THEORY OF THE ELASTO-OPTIC EFFECT

Using the relationship between index of refraction and dielectric constant ($\kappa = N^2$) one finds the change in refractive index from Equations (1) and (2) to be

$$\Delta N_1 = -\frac{N^3}{2} p_{11} s_1 \quad [1, 0, 0] \text{ sound, parallel light}$$

$$\Delta N_2 = -\frac{N^3}{2} p_{21} s_1 \quad [1, 0, 0] \text{ sound, perpendicular light}$$

$$\Delta N_1^1 = -\frac{N^3}{2} \left(\frac{1}{3} p_{11} + \frac{2}{3} p_{12} - \frac{1}{3} p_{44} \right) s_3^1 \quad [1, 1, 1] \text{ sound, perpendicular light}$$

$$\Delta N_3^1 = -\frac{N^3}{2} \left(\frac{1}{3} p_{11} + \frac{2}{3} p_{12} + \frac{2}{3} p_{44} \right) s_3^1 \quad [1, 1, 1] \text{ sound, parallel light}$$

In the case of non-cubic crystals the tensor relationships become more complicated and terms such as p_{13} , p_{31} , p_{14} , etc., appear.

THEORY OF ULTRASONIC DIFFRACTION OF LIGHT

Acoustic waves in an optical medium form a diffraction grating for light beams, producing Bragg angle reflection.

It has been shown both theoretically and experimentally^{1, 2, 3} that if the angle of incidence of a light wave relative to the wavefront of a plane sound wave obeys the following equation then the intensity in the -1 diffraction order is stronger than that of the +1 order and that under optimum conditions the intensity of the +1 order may be reduced to zero.

$$\lambda_L = 2\lambda_s \sin \frac{\theta}{2}$$

where λ_L and λ_s are the optical and acoustic wavelengths, respectively, and θ is the angle between the incident and the diffracted beams. This phenomenon may be most easily understood if one considers the layered nature of the sound wave acting as a Bragg reflection grating (see the figure). At an appropriate angular condition the multiply-scattered beams are phased to produce a maximum in one direction and a minimum in all other directions. If the ultrasonic transducer is sufficiently long, there are more scattering centers, or in other terms the grating is large enough to yield a large scattering cross section, and the intensity of the -1 order beam is strengthened at the expense of the other beams. Under the Bragg condition the intensity of the -1 order beam relative to the zero beam is given by

$$\frac{I_{-1}}{I_0} = \sin^2 \left(\frac{v}{2} \right)$$

where

$$v = \frac{2\pi d \Delta N}{\lambda}$$

d is the transducer length along the optical path, ΔN is the change in index of refraction of the medium produced by the passage of the sound wave, and λ is the wavelength of light in vacuo. The change in index is related to the power in the ultrasonic wave by the equations

$$\Delta N = -\frac{1}{2} N_o^3 p_{ij} s_j$$

¹Quate, C.F., et al., Proc. IEEE, 53, pp. 1604, 1623, 1965.

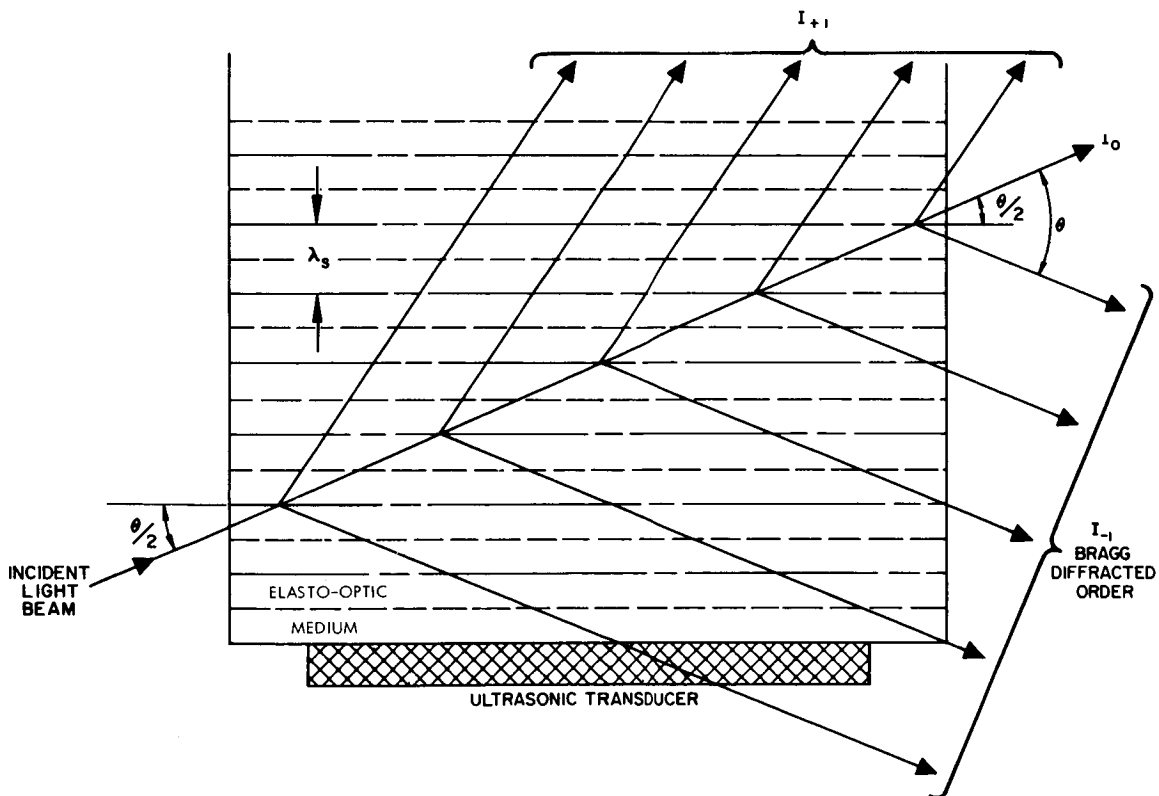
²Hance, H.V., and Parks, J.K., J. Acoust. Soc. Am., 38, p. 14, 1965.

³Cohen, M.G., and Gordon, E.I., Bell Syst. Tech. J., 44, p. 693, 1965.

$$s_j = \sqrt{\frac{2S_a}{\rho v_s^3}}$$

$$S_a = \frac{TS_{rf} \times 10^7}{A}$$

Where S_a is the acoustic power density, ρ is the density of the medium, v_s is the velocity of sound of the medium, T is the electromechanical transduction efficiency of the transducer, A is the transducer area, and S_{rf} is the incident r-f power. It has been shown experimentally that for sufficiently large values of v , the Raman-Nath parameter, approximately 95 percent of the energy incident on the modulator can be translated into the -1 diffraction order at Bragg angle. However, the choice of experimental conditions must be optimum and would seldom be encountered in practice. It is important to note that the modulation effect falls off with increasing wavelength so that for a modulator capable of producing 60 percent modulation at 0.6328μ , approximately 1/300 of this or 0.2 percent modulation would be obtained at 10.6μ .



Ultrasonic Diffraction Geometry

ACOUSTIC MODULATION TECHNIQUES

Acoustic modulation techniques include frequency translation, intensity modulation, and polarization modulation.

Frequency Translation

It has been shown previously that a traveling ultrasonic wave can diffract a light beam into an arbitrary number of diffraction orders depending on the transducer length and the radiation geometry. The diffracted beams will be frequency shifted by an amount

$$\Delta f = f_o \pm n f_s$$

where f_o is the frequency of the light wave, f_s is the frequency of the sound wave, and n is the diffraction order. By adjusting the modulator for the Bragg condition (see previous topic), the intensity in all orders except -1 may be substantially reduced to the point where the device produces a single frequency shifted sideband I_{-1} . This mode of operation is entitled frequency translation and is useful in generating local oscillator beams for heterodyne detection systems. A detailed tabulation of the modulation efficiencies expected from a number of materials is presented in a later topic.

Intensity Modulation

If the traveling ultrasonic wave is reflected at a boundary of the modulation medium so that it returns to the transducer undeviated, a reflected or backward ultrasonic wave is generated, which, if unattenuated, is capable of producing a diffracted beam system spatially coincident with that produced by the forward wave and of equal intensity. Moreover, if the forward wave produces a positive frequency translation in a given order, the backward wave produces a negative one and vice versa. Thus, the sideband spectra become double sidebands and act as suppressed carrier modulated beams. The central or zero order beam becomes intensity modulated at twice the ultrasonic wave frequency since one now has an ultrasonic standing wave which diffracts energy away from the main beam twice per cycle. Because of the large values of mechanical Q usually encountered in ultrasonic resonators, the amplitudes of the forward and backward ultrasonic waves may reach values not obtainable in a traveling wave situation, and may, therefore, obtain large values of modulation index at low r-f drive levels. However, this system is of limited usefulness where large amplitude modulation bandwidths are desired because of the high mechanical Q of the system. One can, of course, use mechanical damping techniques to broaden the bandwidth of the modulator at the expense of modulation index and efficiency.

By proper choice of modulator dimensions and other parameters it is possible to produce a single diffraction maximum consisting of two frequency-translated optical waves. This system may prove valuable as a mechanism for locking of laser modes at infrared wavelengths, if the diffracted beam is fed back into the laser system.

If the optical faces of the modulator are made optically flat, highly parallel, and highly reflective, one can orient the crystal to obtain maximum transmission of light through the crystal at a selected set of optical wavelengths. This is the well-known Fabry-Perot effect. Similarly, if one passes sound waves through the crystal, one finds that the multi-reflected light wave interacts many times with the sound wave and, in consequence, the interaction cross section or modulation is stronger. This effect seems to be of limited usefulness in the visible region of the spectrum because of the change in modulator dimension when ultrasonic energy is dissipated in the crystal. At longer wavelengths this technique may be practicable because the effect of expansion is reduced, since the light wavelength is longer. With the use of good temperature regulation of the crystal, a Fabry-Perot acoustic modulator may be useful at 10.6 microns, where the dimensional tolerances will be about 20 times less than in the center of the visible spectrum.

Polarization Modulation

By focusing the optical beam to a diameter small compared to the ultrasonic wavelength it is possible to modulate the polarization of the optical beam, i. e., change the polarization from linear to elliptical, because of the birefringence induced in the modulation medium by the ultrasonic strain wave. The birefringence is simply calculated by taking the difference between the elasto-optic constants p_{11} and p_{21} or p_{33} and p_{31} and multiplying by the appropriate variables to yield a value ΔN , the induced change in refractive index. Further discussion of the elasto-optic birefringence modulation will not be detailed here since the application of birefringence modulation is well described in the section of electro-optic modulators.

PROPERTIES OF ULTRASONIC MODULATORS

The properties of elasto-optic media require specialized mechanical and electronic techniques for designing efficient optical modulators.

Choice of Acoustic Media

From the foregoing treatment, it may be seen that when a traveling sound wave interacts with a traveling light wave in an elasto-optic medium (solid or liquid) a frequency translated diffracted beam can be obtained which may be used for system applications requiring frequency translation. At ultrasonic frequencies below 50 MHz optically transparent liquids such as water, alcohol, tetrachloroethylene, carbon tetrachloride, etc., may be used as the modulation medium, provided they are transmissive to the optical wave. At frequencies above 50 MHz the high ultrasonic attenuation in all liquids prevents satisfactory modulation indices and forces consideration of solids as a modulation medium, since they have intrinsically lower losses than liquids. In the next topic, a tabulation is presented which compares the relative performance of useful ultrasonic modulation media at one visible and two infrared wavelengths. A general result is that materials with large values of refractive index and low values of sound velocity are the most effective modulation media.

Anechoic Terminations

The development of single sideband frequency translators in elasto-optic media introduces a new problem, namely the absorption or termination of the ultrasonic wave. This operation is most conveniently done by shaping the modulator crystal so that the coherence of the ultrasonic wave is destroyed and it can no longer coherently diffract the light wave. Such terminations are termed anechoic chambers and are designed according to the same principles as optical black bodies. However, the expense of cutting and shaping the single crystal is easily avoided by placing the crystal in contact with a pool of mercury which is contained in an anechoically shaped container. In practice it is found that mercury is a good acoustical impedance match for a large number of solids since the product of its density and sound velocity (acoustic impedance) is within 30 percent of that of many crystals. Thus the power reflection coefficient at a quartz-mercury or a quartz-silicon interface is of the order of 3 percent or less. Therefore, most of the ultrasonic energy passes into the mercury and is then rendered incoherent at the rough boundaries of the mercury container.

Transducer Bonding

Another problem common to solids but not encountered in liquids is the production of a satisfactory bond between the transducer and the modulator crystal. The bond should be lossless and of uniform thickness and it should be small compared to an ultrasonic wavelength. If all of these conditions are met then the transducer bandwidth and electrical impedance can be calculated or predicted in a straightforward manner. If they are not met, the bond can act as a mechanical transformer a quarter wavelength long and thus preclude prediction of the frequency response and efficiency of the system.

In practice, it has been difficult to produce ideal ultrasonic transducer bonds without a great deal of care for transducers operating in the 5- to 100-MHz range. For frequencies above 100 MHz, a new transducer technique using evaporated thin films has some advantages not previously available; e.g., direct application of the transducer to the modulator without use of intermediate layers, fundamental frequency operation over a broad band, and fairly high transduction efficiencies.

Electrical Impedance Matching

Because of the high fundamental operation frequency, the transducers are very thin and thus present a high capacitance to the r-f generator or amplifier. It is usually difficult to inductively compensate this capacitance over a broad band of frequencies without using resistive damping and suffering an attendant increase in transduction loss. Theoretical calculations indicate that the use of a resonating inductance with a Q of 200 as an impedance matching transformer will yield an ultrasonic transduction loss of 10 dB, for a quartz transducer radiating into quartz, germanium, or silicon. In practice, transduction losses of the order of 12 dB are usually measured indicating the presence of other losses such as dielectric loss, skin effect losses in wires, and dissipation in the ultrasonic bonds. If transducers with higher electromechanical coupling than quartz were used, e.g., ZnO, CdS, etc., a factor 2 to 4 improvement in transduction efficiency could be obtained. Thus, materials such as CdS or ZnO are preferable for ultrasonic transduction since they enable the designer to obtain a larger electrical bandpass and at the same time a reduction in the transduction loss that is obtained with quartz. These materials are used for making thin film evaporated transducers since in single crystal form they are quite brittle and do not lend themselves to the production of large area transducers.

Transit Time Effects

In the Fabry-Perot modulator, as in the normal ultrasonic modulator, the transit time of the ultrasonic wave across the optical beam is a modulation bandwidth determining factor. For example, if the velocity of sound is 5×10^5 cm/sec and the laser beam diameter is 10^{-1} cm, then the transit time is approximately 0.2×10^{-6} sec or the bandwidth is confined to frequencies below 5 MHz. It is possible to use a long focal length lens to reduce the beam diameter to the diffraction limit, since the beam diameter at focus, a , is given by

$$a = \lambda \frac{f}{D}$$

where λ is the wavelength of light, ($\approx 10^{-5}$ - 10^{-3} cm), f is the focal length of the lens and D is the diameter of the laser beam incident on the lens. Thus with the aid of a 10 cm focal length lens, a light beam of diameter 10^{-1} cm at 6328 \AA can be focused to a spot of dimension 6.3×10^{-2} cm corresponding to a transit time $\approx 10^{-7}$ sec or a bandwidth of 10 MHz. Use of a shorter focal length lens can reduce the beam diameter even further if necessary to reduce the transit time of the ultrasonic wave across the light beam.

PROPERTIES OF ULTRASONIC MODULATORS

In effect, at the higher modulation frequencies,¹ the long transit time produces spatial modulation of the light wavefront and thus reduces the intensity of the modulation in a given diffraction order. Operating at frequencies low compared to the transit time will produce a percentage modulation in a given order approaching 100 percent; however, operation at the frequency equal to the reciprocal of the transit time will reduce the modulation in a given order to about 30 percent of the maximum.

These considerations are usually important when one is concerned with highly efficient modulator systems. However, ultrasonic modulator systems in their present state of development are not highly efficient since the physical nature of the transduction system limits one to at least 6 dB transduction loss with the most advanced experimental transducers.

¹Hance, H. V., and Parks, J. K., op. cit.

SUMMARY AND TABULATION OF ELASTO-OPTIC MODULATORS

An appropriate figure of merit for elasto-optic modulator materials is the factor $(N_o^6 p^2 / \rho v_s^3)$.

On the basis of the equations shown in preceding topics, the intensity in a modulated optical sideband produced by a traveling ultrasonic wave is given by

$$\frac{I_{-1}}{I_o} = \sin^2 \left[\frac{\pi d N_o^3}{2\lambda} \cdot p_{ij} \sqrt{\frac{2 T S_{rf} \times 10^7}{\rho v_s^3 A}} \right]$$

where $A = d \times h$, the transducer area, for small values of the argument or for optimum Bragg conditions, i. e., when the transducer is of sufficient length that a single Bragg maximum is produced. This can be approximated by

$$\frac{I_{-1}}{I_o} = \frac{\pi^2}{2\lambda^2} \left(\frac{N_o^6 p_{ij}^2}{\rho v_s^3} \right) \left[\left(\frac{d}{h} \right) (S_{rf} T \times 10^7) \right]$$

Practical values of transduction efficiency T are of the order of 10 to 15 dB, corresponding to a power ratio of 0.1 to 0.03. Experimentally, it is found that the p_{ij} values range from 0.01 to 0.30 in most transparent solids, so that it is advantageous to select materials with as large a value of N_o^6 and as small a value of ρv_s^3 as possible.

Assuming a d over h ratio of 5:1, an rf power level of 20 watts, and a transduction efficiency of 0.03, the term in brackets takes a value of 3×10^7 . The quantity $\pi^2 / 2\lambda^2$ takes a value of 1.23×10^9 at a wavelength of 0.633μ , 4.3×10^7 at 3.39μ and, 4.38×10^6 at 10.6μ . An approximate table of values for I_{-1}/I_o for various materials is given in the Table.

The Table indicates the potential of KRS-5 and CdS as candidates for producing 100 percent modulation at 6328 \AA with 20 watts of rf drive power.

In the infrared region of the spectrum, germanium and KRS-5 are the leading contenders for single sideband frequency translators, but the production of high percentage of modulation at infrared wavelengths will require prohibitively high rf power levels ($\approx 2\text{kW}$). It is important to realize that a good transduction system can change the efficiency by a factor of 3 so that the possibility of 50 percent modulation at 10.6 microns exists.

A high level of modulation at 10.6μ has been achieved¹ using tellurium. However, despite its very high figure of merit ($N_o^6 p^2 / \rho v_s^3$), it is an extremely fragile material, so its use in practical systems is recommended only with considerable reservation.

Elasto-Optic Performance Parameters for Selected Crystals

Material	Relative Elasto-Optic Coefficients			Modulation Ratio***		
	$\frac{N_o^6}{\rho v_s^3} (p_{ij})^2$			I_{-1}/I_o		
	0.63 μ	3.39 μ	10.6 μ	0.63 μ	3.39 μ	10.6 μ
Quartz	1.07×10^{-18}	2.66×10^{-19}		0.04	3.4×10^{-4}	
Cadmium Sulfide	24.4×10^{-18}	6.1×10^{-18}	4×10^{-18}	0.9	7.8×10^{-3}	5.3×10^{-4}
Zinc Oxide	10×10^{-18}	2.5×10^{-18} **	2×10^{-18} **	0.37	3.2×10^{-3}	2.6×10^{-4}
Silicon		8.9×10^{-18}	6×10^{-18} **		1.02×10^{-3}	7.9×10^{-4}
Germanium		5.3×10^{-17} **	4×10^{-17}		6.8×10^{-2}	5.3×10^{-3}
KRS5		2.7×10^{-17} **	2×10^{-17} **	$\approx 1.0^*$	3.5×10^{-2}	2.6×10^{-3}
LiNbO ₃				0.16		
<p>*When the argument of $\sin^2(v/2) \approx (v/2)^2$ exceeds the range of validity of the approximation, one must return to $\sin^2(v/2)$ to avoid values of $I_{-1}/I_o > 1$.</p> <p>**Most of the values of p_{ij} are estimated on the basis of the measured behavior of CdS in the visible and CdS and Si at 3.39μ.</p> <p>***Values of the modulation ratio computed for 20 watts of electrical driving power applied to the transducer and 15 dB transduction loss.</p>						

¹Dixon, R. W. and Chester, A. N., Appl. Phys. Letters, 9, p. 190, 1966.

OPTICAL MODULATORS

Internal Modulation

	Page
Intensity Modulation	168
Frequency Modulation and Translation	172

INTENSITY MODULATION

Intensity modulation by internal means is achieved by coupling out a variable amount of optical power from within the laser cavity.

The term, internal modulation, is used to describe modulation schemes performed within the optical resonator.¹ Intensity modulation is also referred to as coupling modulation.

Three such schemes have been demonstrated. The common feature of these schemes is to transfer a fraction of the internal laser power from its original form, for which the resonator has a high Q (i. e., a high reflection), to a form in which it escapes readily from the resonator. This new form can be due to a modification of: (1) the polarization direction, (2) the direction of propagation, (3) the frequency.

The amount of power coupled in this manner can be made proportional to the (instantaneous) modulation voltage so that a square-law detector will recover the modulation signal.

The main advantage of these schemes over external modulation is that by acting on the internal laser energy, which may exceed the external power by more than two orders of magnitude, a given modulation power can give rise to proportionately higher absolute modulation of the output beam.

One form of intensity modulation uses an electro-optic crystal inside the resonator. The crystal orientation and biasing with respect to the laser beam is identical to that obtained when the same crystal is used externally, i. e., the crystal couples a fraction

$$\frac{I}{I_0} = \sin^2 \frac{\Gamma}{2}$$

from the original laser polarization to one in which the optical E vector is rotated by 90 degrees about the direction of propagation. Γ , the optical retardation, is given as before by

$$\Gamma = \frac{2\pi L}{\lambda} N_o^3 r E$$

In order to have a useful modulation scheme, a way to couple out the modified radiation must be devised. This can be done by a polarizing prism (such as Rochon or Glan-Thompson prism) which separates the two incident polarizations, or, less efficiently, by Brewster windows which reflect a fraction, typically ~20 percent to 60 percent, of the second polarization.

As an example, consider the case when the applied field consists of a d-c bias plus a time dependent modulation field, i. e.,

¹Gürs, K., and Müller, R., Phys. Letters, 5, p. 129, 1963.

$$E = E_B + E_m(t)$$

letting $(2\pi L/\lambda)N_o^3 r = a$, yields:

$$\frac{I}{I_o} \approx \frac{a^2}{4} (E_B + E_m)^2$$

for

$$\Gamma = aE \ll 1;$$

and when

$$E_m \ll E_B,$$

$$\frac{I}{I_o} \approx \frac{a^2}{4} (E_B^2 + 2E_B E_m)$$

The radiation coupled out of the cavity is thus in the form

$$I \approx I_o \frac{a^2}{4} [E_B^2 + 2E_B E_m(t)]$$

and thus has a modulation index of $m = 2E_m/E_B$. A square law detector will yield an output current

$$I_{\text{detector}} \propto I_o E_B E_m(t)$$

thus recovering the original information signal $E_m(t)$.

The advantage of the internal scheme is now apparent. The radiation intensity inside the laser exceeds that outside by a factor equal to the inverse of the mirror transmittance. For 1 percent transmittance, to use a common value, $(I_o)_{\text{int}}/(I_o)_{\text{ext}} = 100$. It follows then that for a given modulation signal and the same E_B , the detector output is larger by 100.

A schematic diagram of a polarization modulation scheme is shown in Figure A.

Another mode of intensity modulation is based on the reflection of light from traveling sound waves.² A sound wave traveling in a medium (solid or liquid) placed inside the resonator intercepts the optical beam at the Bragg angle. The spatial modulation of the optical dielectric constant which is proportional to the local strain, acts as a diffraction grating and scatters a portion of the laser energy into the various diffraction orders.

²Siegman, A.E., et al., Appl. Phys. Letters, 5, p. 1, 1964.

INTENSITY MODULATION

The amount of scatter intensity is proportional to the modulation acoustic power. In addition, the scattered beam is Doppler shifted in frequency by the acoustic frequency. One scattered beam is shifted upwards while a second scattered beam, emerging in a direction opposite to the first, is shifted downwards in frequency.

A third scheme consists of shifting a fraction of the internal energy in frequency and then extracting the frequency-shifted power by an etalon mirror which has a passband at this frequency.³ The internal frequency shift is performed by low index phase modulation inside the laser resonator by an electro-optic crystal which has an induced optical axis parallel to the optical electric field. The amount of power coupled by this scheme and its dependence on the modulation voltages is similar to that previously described for intensity modulation. A schematic diagram of this scheme is shown in Figure B.

It should be pointed out that some internal modulation schemes result in a compromise in laser design which may deleteriously affect the communications system. The mode selection problem becomes more severe and the laser output power is necessarily reduced because the reflection at the output mirror must be increased to compensate for the losses introduced into the cavity by the modulator.

³Peterson, Don G., and Yariv, Amnon, Appl. Phys. Letters, 5, p. 184, 1964.

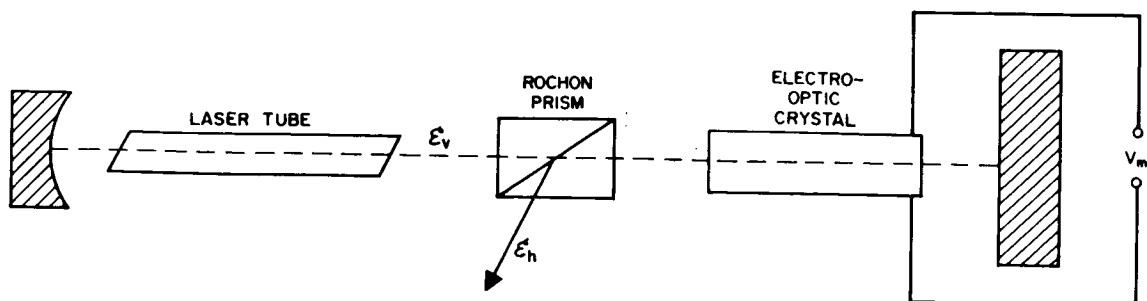


Figure A. Laser Oscillator with Internal Polarization Modulation

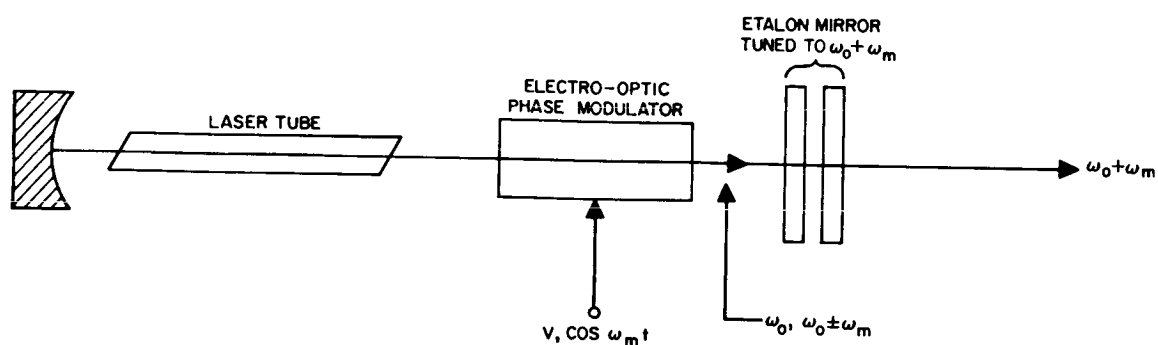


Figure B. Laser with Coupling Modulation by Internal Frequency Shifting

FREQUENCY MODULATION AND TRANSLATION

Optical frequency translation and frequency modulation may be effected by electro-optic elements properly oriented within the laser cavity resonator.

Several techniques of frequency modulation and translation of optical laser beams have been reported which employ electro-optic modulators mounted within the Fabry-Perot resonant structure of the laser. Three such devices which perform quite different operations will be discussed here briefly.

The first system,¹ illustrated in the figure is basically an SSBSC modulator and is related to the two-element scheme discussed in an earlier topic. The basic components of this system consist of a Nicol prism or its equivalent, an electro-optic modulator, a $\lambda/8$ wave plate, and a highly-reflective mirror. In the figure, the Brewster window of the laser serves as the Nicol prism.

Operation of this configuration is as follows: The vertically-polarized laser beam traverses the modulator whose induced axes of birefringence lie at 45 degrees to the vertical. The beam then passes through the eighth-wave plate and is reflected from the end mirror, located a distance $\lambda_m/8$ from the center of the modulator, where λ_m is the free-space wavelength of the modulating signal. The return beam after traversing the $\lambda/8$ plate and modulator now contains a horizontal component shifted in frequency by f_m , and part of this is reflected out of the laser by the Brewster window. The double pass through the eighth-wave plate has the effect of rotating the modulator through 45 degrees for the return beam, and the round-trip photon transit time between modulator and mirror provides the 90-degree phase shift of the modulating signal.

This system can be used outside the laser as well, provided a real Nicol prism and additional reflector are used. In either configuration, the amplitude of the extracted beam is proportional to $J_1(\Gamma_o/\sqrt{2})$, as with the previous two-element modulator.

The second type of internal device² is simply an electro-optic modulator, oriented for pure phase modulation; i.e., the principal induced birefringent axis is parallel to the vertical polarization of the laser mode. When the modulation frequency is adjusted to be approximately but not exactly equal to the laser axial mode spacing, the laser mode can be made to oscillate with FM phases and nearly Bessel Function amplitudes. Thus, the output is forced into the state of a frequency-modulated optical signal which is swept over the entire Doppler line-width at the modulation frequency.

This scheme promises to have valuable applications in some communication systems, but it appears to have a limited information bandwidth capability.

¹Targ, R., Massey, G.A., and Harris, S.E., Sylvania Electronic Systems, Mountain View, California, internal report.

²Harris, S.E., and Targ, R., Appl. Phys. Letters, 5, p. 202, 1964.

The third scheme³ is one which is especially useful as a direct baseband frequency modulator of the laser optical carrier. It involves varying the electrical length (or optical path length) of the laser cavity. An electro-optic modulator crystal is placed inside the cavity and oriented, as in the previous example, with a principal axis of induced birefringence parallel to the polarization direction. Under application of a modulator field E , the change Δf_o in the laser frequency is given by

$$\Delta f_o = \frac{c L N_o^3 r E}{2 \lambda L_o}$$

where

$c \equiv$ velocity of light

$L \equiv$ modulator length

$L_o \equiv$ effective optical length of cavity

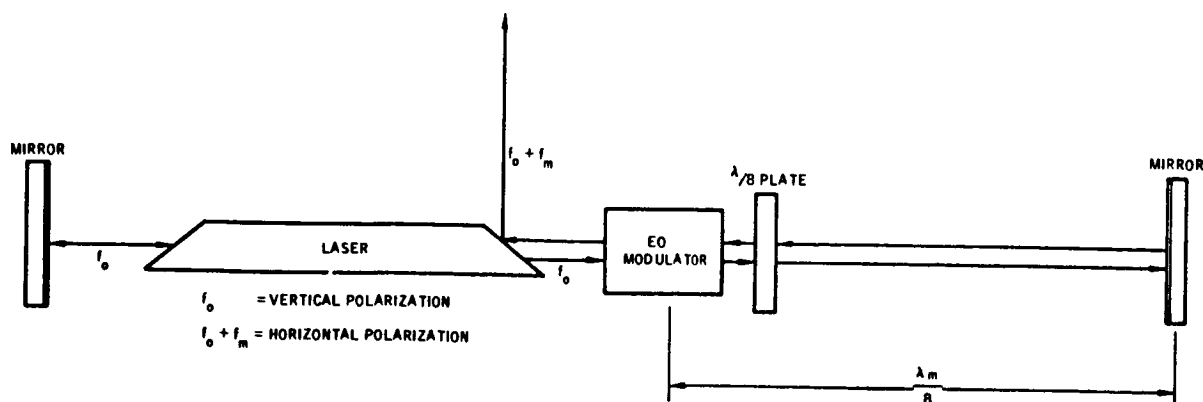
The other symbols have been defined previously. In terms of the customary expression for optical phase retardation

$$\Gamma = \frac{2 \pi L N_o^3 r E}{\lambda}$$

the frequency shift may be written

$$\Delta f = \frac{c \Gamma}{4 \pi L_o}$$

The deviation frequency attainable is limited by the gain profile of the laser atomic transition. For example, in a 3.39μ laser, this limit is about 300 MHz; for the CO_2 10.6μ laser, the limit is less than 25 MHz. An additional restriction is that the frequency cannot change faster than the cavity Q permits. It can be shown³ that the limiting FM bandwidth is then simply f_o/Q , where $f_o = c/\lambda$ is the laser resonant frequency.



Internal SSBSC Modulator

³Goodwin, F.E., IEEE J. Quant. Electronics, QE-3, p. 524, 1967.

MODULATORS

Modulator Performance

	Page
Modulator Burden Considerations	176
Modulator Burden Relationships	178
Nomenclature Summary	180

MODULATOR BURDEN CONSIDERATIONS

Advances in the technology of electro-optic modulator materials and design techniques permit a reasonably accurate assessment of performance and burden characteristics for the years immediately ahead.

It may still be premature at this time to make accurate assessment of the performance and burden characteristics of optical modulators which will have continuing value to the designer of laser systems for space communications and tracking. However, great progress has been made in the past few years in the development of optical modulation techniques and materials. This work has demonstrated that all forms of modulation can be impressed on optical carriers in the band between 0.4 and 10.6 μ . At wavelengths longer than 1.5 μ , optical modulation technology is in a somewhat more primitive state; but the advent of high grade semi-insulating GaAs as a useful electro-optic modulator material in the infrared region of 2 to 12 μ has constituted a notable breakthrough for CO₂ laser applications. Moreover, the technology of synthesizing superior grade LiNbO₃ and LiTaO₃ has made possible useful and efficient modulators in very broadband applications over the wavelength range of 0.4 to 5 μ . Research and development of acoustic modulation techniques is progressing at a moderate pace, and acoustic modulators in some instances offer a significant power advantage; but on the other hand they provide modulation bandwidths far short of those believed to be needed in optical space communication systems.

In considering communication and tracking systems using infrared lasers, it must be realized that there is an inherent inverse dependence of modulation efficiency on wavelength at a given driver power level. Thus, for example, a specific modulator element requires 10 times as much voltage at 5 μ as it does at 0.5 μ to achieve the same depth of modulation. In order to keep the power burden within reasonable limits, it is therefore necessary to (1) extend the interaction length and/or (2) settle for less modulation index. An exception to this prescription is the intracavity modulation scheme which can provide a high effective modulation index for intensity modulation and reasonable frequency deviations in optical FM systems with very modest levels of driver power, when the attendant bandwidth limitations are acceptable.

The table presents, in more or less chronological order, a series of electro-optic modulators and their operating characteristics. This listing is by no means complete, but an attempt has been made to itemize those modulators which represent significant advances in their particular regime of applications. Except where otherwise noted, the performance characteristics are applicable for an optical wavelength of 6328 Å. The commercial device produced by Isomet Corporation reflects the advanced technology now reached in the field of electro-optic modulator design and production.

Characteristics of Some Electro-Optic Modulators - December 1968

Parameter	T _{M00} Mode Cavity Modulator	Traveling Wave Intensity Modulator	Polarization Modulator (NASA Contract)	Multi-Element Intensity Modulator	Resonant SSBSC Modulator	Traveling Wave SSBSC Modulator	Single Element PCM Modulator	10-6 μ Intensity Modulator	Commercial Modulator
Development Status	Built at BTL	Built at Sylvania	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at Hughes Aircraft Company	Built at BTL	Built at RCA	Product of Isomet Corp.
Material	KDP	KDP	KDP	KDP	KDP	KDP	LiTaO ₃	GaAs	y-cut ADP
Crystal Dimensions	1 cm long x 2.5 mm dia.	100 cm long x 2 mm wide	50 x 0.4 x 0.4 cm	16 each 1/4" x 1/4" x 1/2"	2 each 1 1/2" long x 1/2" dia.	80 x 0.4 x 0.4 cm	1 x 0.025 x 0.025 cm	6.7 x 0.3 x 0.3 cm	20 x 0.1 x 0.4 cm
Optical Attenuation	0.1 dB	6 dB	1.5 dB	0.3 dB	1.0 dB	2.0 dB	1.5 dB	0.4 dB	0.4 dB
Useful Optical Range	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 - 1.5 μ	0.4 to 5 μ	2 to 12 μ	0.25 to 0.75 μ
Modulation Frequency	3 Gc	Baseband	Baseband	Baseband	850 Mc	200 Mc	Baseband	Baseband	Baseband
3 dB Bandwidth	4 MHz	3 GHz	30 MHz*	10 MHz*	5 MHz	100 MHz	220 MHz*	100 MHz*	>100 MHz
Modulating Power	1.5 W	12 W	20 W	12 W	3 W	10 W	250 mW	70 W	50 W
Modulation Index	0.13	~1.0	0.5	0.5	0.01	0.3	0.4	0.1	0.5
Extinction Ratio	Unknown	Unknown	8:1	250:1	NA	NA	80:1	~80:1	70:1
Weight	Unknown	Unknown	20 lbs	2 lbs	15 lbs	25 lbs	Unknown	Unknown	Unknown
Size	Unknown	~100 in ³	144 in ³	30 in ³	850 in ³	216 in ³	Unknown	12 in ³	80 in ³

* 3dB Bandwidth limited by modulator driver and not by the electro-optic structure.

MODULATOR BURDEN RELATIONSHIPS

Weight, cost, and power burden relationships are given for 10.6 microns and 0.5 microns.

In Figures A, B, and C are presented the modulator burden relationships, in which the burdens - weight, power, and cost - are plotted as functions of the information bit rate R_B . Two curves are presented in each case, one applicable to 0.5 microns and the other to a wavelength of 10.6 microns. A communication system employing polarization modulation is assumed. (It can be shown that a minimum of 3 dB gain in signal-to-noise ratio at the receiver is achieved with polarization modulation over intensity modulation.) The burden levels indicated are applicable for 50 percent modulation index at 0.5 μ using LiNbO_3 and 10 percent modulation index at 10.6 μ using GaAs.

The functional relationships between the burdens and bit rate are taken to be linear, and of the form

$$W_M = W_{KM} + K_M R_B \text{ for weight } W_M$$

$$P_M = K_{PM} W_M = K_{PM} (W_{KM} + K_M R_B) \text{ for power } P_M$$

$$C_M = C_{KM} + K_{FM} R_B \text{ for cost } C_M$$

where

W_M = the modulator weight

W_{KM} = the modulator weight independent of bit rate

K_M = a constant relating bit rate to weight

P_M = the power used by the modulator

K_{PM} = a constant relating the modulator weight to the power used by the modulator

C_M = the modulator cost

C_{KM} = the cost of the modulator independent of the bit rate

K_{FM} = a constant relating bit rate to cost

The Table tabulates the values of these constants.

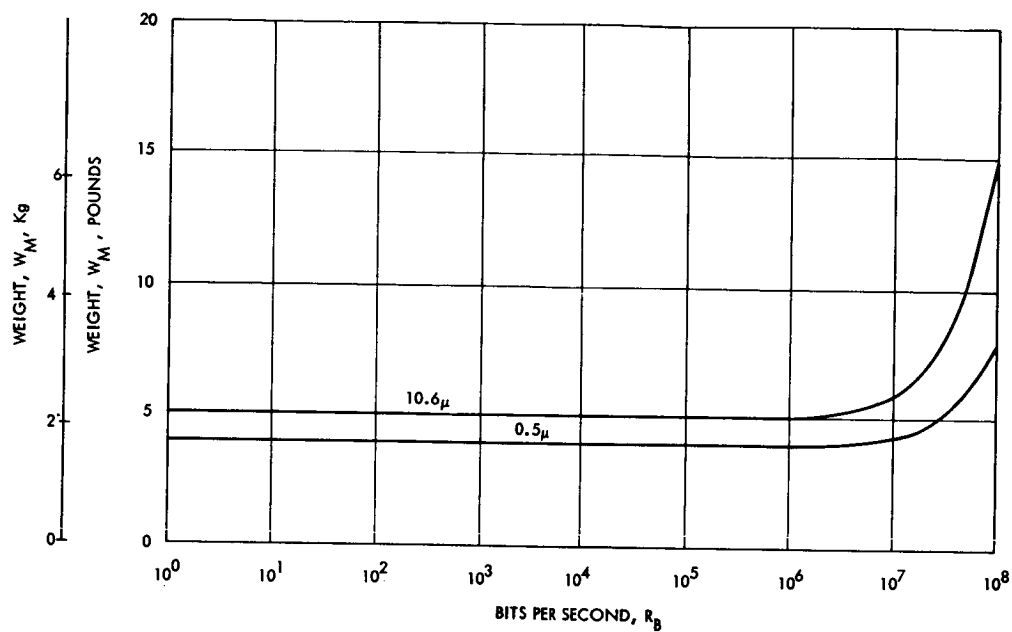


Figure A. Modulator Weight Burdens

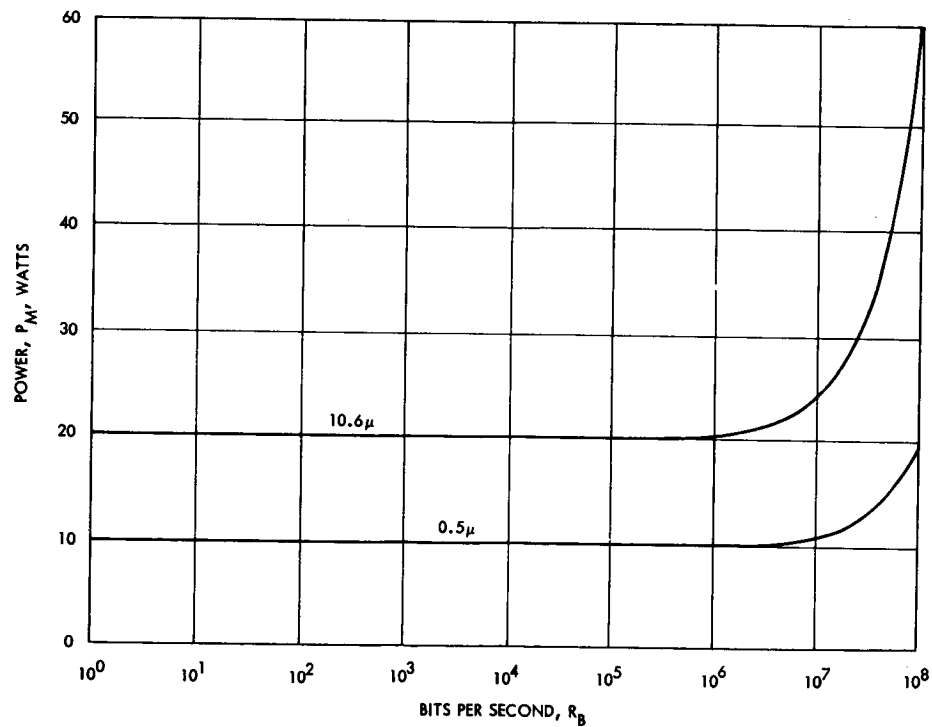


Figure B. Modulator Power Burdens

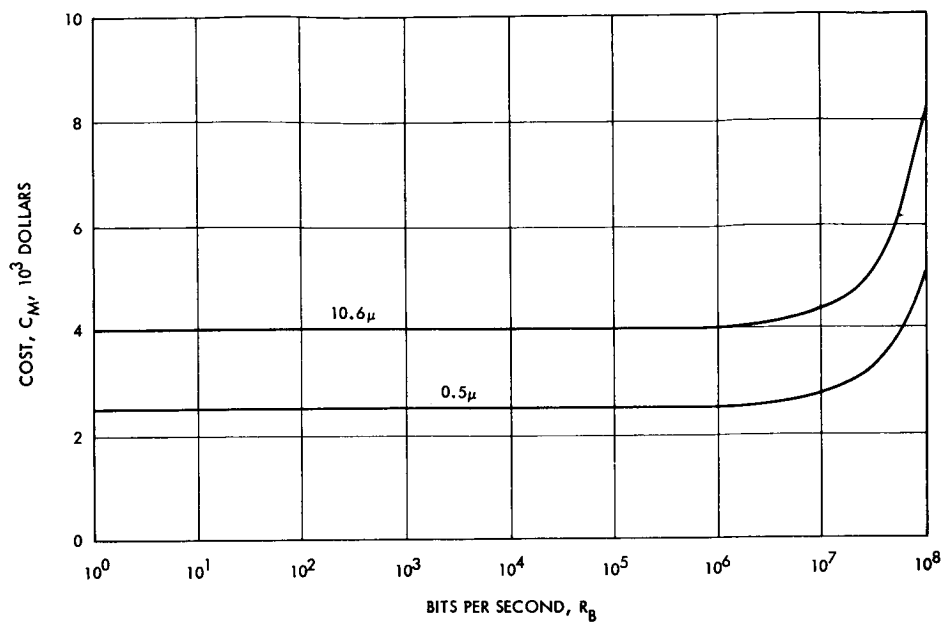


Figure C. Modulator Cost Burdens

Optical Modulator Burden Summary

Burden Constant	Wavelength	
	0.5 Micron	10.6 Micron
W_{KM}	4	5
K_M	4×10^{-8}	10^{-7}
K_{PM}	2.5	4
C_{KM}	2500	4000
K_{FM}	2.5×10^{-5}	4×10^{-5}

NOMENCLATURE SUMMARY

Since some nomenclature is repeated from topic to topic, a summary of the nomenclature used in Optical Modulators is given below.

a	beam diameter at focus; also a symbol for $(2\pi L/\lambda)N_o^3 r$
A	transducer area
B	Kerr constant
$B_{ij}, B_{jk}, \text{ etc.}$	dielectric impermeability tensor or indicatrix, a dimensionless quantity
c	velocity of light in vacuum
C	capacitance, e.g., of an optical modulator
C_{KM}	minimum modulator cost
C_M	modulator cost
d	transducer length along the light path
D	laser beam diameter
E	electric field, dc or ac
E_B	d-c bias electric field
E_m, E_o	electric field, amplitude, or peak value
f	focal length of a lens
f_m	modulation signal frequency
f_o	frequency of the optical wave
f_s	frequency of the ultrasonic wave
$g_{11}, g_{12}, \text{ etc.}$	quadratic electro-optic coefficients
h	transducer height
I	output intensity of a modulated light beam
I_o	beam intensity at input to modulator
I_{-1}, I_o, I_{+1}	intensities of the $n = -1$, $n = 0$, and $n = +1$ order diffracted beams
$J_n(x)$	n^{th} order Bessel function of an argument x
K_{FM}	rate of increase of modulator cost with bit rate

K_M	rate of increase of modulator weight with bit rate
K_{PM}	ratio of modulator power to weight
L	modulator length
L_o	electrical (or optical) length of a laser cavity
m	modulation index
m_i	effective modulation index of the i^{th} harmonic of a modulation signal
n	diffraction order number
N	general symbol for refractive index
N_o	ordinary refractive index of a birefringent material
N_e	extraordinary refractive index of a birefringent material
p_{ijkl}	elasto-optic coefficient of the medium, dimensionless
P_M	modulator power requirement
$r, r_{63}, r_{41}, \text{ etc.}$	linear electro-optic (Pockels) coefficients
R	resistance or output impedance
R_B	modulation bit rate
s_j	strain component of η_{kl} for a principal crystallographic direction
S_a	acoustic power density
S_{rf}	total radio frequency or microwave frequency power input to the transducer
t	electrode spacing or plate separation
T	transduction efficiency, dimensionless; also temperature
u	symbol for $\pi N_o L / \lambda_m$
v	dimensionless Raman-Nath parameter and is equal to $2\pi d \Delta N / \lambda$
v_o	optical wave velocity

NOMENCLATURE SUMMARY

v_s	velocity of sound in an optical medium
V	voltage, dc or ac
V_m, V_o	amplitude or peak value of an applied voltage
W_{KM}	minimum modulator weight
W_M	modulator weight
Γ	relative phase shift between principal components of a polarized light beam induced by a birefringent element; e.g., an optical modulator
Γ_{eff}	effective optical phase shift at high modulation frequencies taking into account optical transit time
Γ_o	peak value of Γ
δ_{ik}	Kronecker delta
ϵ	permittivity or inductive capacity of a dielectric medium
ϵ_h	optical electric field, horizontal polarization
ϵ_v	optical electric field, vertical polarization
η_{kl}	strain induced in the elasto-optic medium, dimensionless
θ	angle, specifically the angle between an incident and diffracted light beam
κ_{ij} , etc.	dielectric constant of the medium, a dimensionless quantity
λ, λ_L	optical wavelength in free space
λ_m	free space signal wavelength
λ_s	sound wavelength in the diffracting medium
ρ	density of an optical medium
ω	2π times an optical frequency
ω_m	2π times a modulation signal frequency

PART 3 – DETECTORS

Section		Page
Radio Frequency Detectors		190
Optical Frequency Detectors		200

INTRODUCTION

Theory and state-of-the-art performance are given for radio and optical detectors.

Radio and optical detectors are considered in two major sections.

Radio Detectors

The normal figure of merit for radio detectors is the noise figure or noise temperature. These are documented as a function of frequency. Means of reducing the noise temperature of detectors by use of a low noise preamplifier is also quantitatively described.

Optical Detectors

Optical frequency detection falls into two general categories: that where heterodyne detection is required to obtain efficient detection and that category where direct detection will provide efficient detection. The latter is essentially limited to the visible light spectrum.

Operating theory is given for both categories as is measured performance.

SUMMARY OF DETECTION METHODS

Heterodyne performance for radio and optical frequencies is given. Optical heterodyne detection will probably find its greatest application in a wide band data link from a space probe. Direct detection is practical only at optical frequencies. It will probably find its greatest application as a beacon link to a deep space vehicle.

Heterodyne Detection

The performance of optical heterodyne detectors is similar to that of radio heterodyne systems but still differs markedly due to the high operating frequency. This has the effect of changing the dominant noise contribution from a spectral density given by kT (k is Boltzmann's constant and T absolute temperature) to a spectral density given by $h\nu$ where h is Plank's constant and ν is the operating frequency. In general, the noise power spectral density N increases with frequency.

Ideally,

$$N = \frac{h\nu}{e^{h\nu/kT} - 1} + h\nu$$

This curve is plotted as a function of frequency in the figure, for various noise temperatures, T_{in} °K. The figure compares detector performance over the radio-to-optical spectrum and shows the projected capability of optical detectors (using the above equation) in comparison with known radio receiver performance.

More recently heterodyne detectors have been constructed at both 3.39 microns¹ and 10.6 microns². The performance of these detectors is much improved over that of the direct detector.

The radio frequency detectors indicated in the figure are of two types, a heterodyne mixer and a heterodyne mixer preceded by a low noise amplifier. When the gain of the amplifier is high its low noise characteristics become the dominant noise contribution of the detecting system.

Direct Detection

Direct detection is practical only for optical frequencies. In direct detection, the output signal is dependent only upon the input signal and background power. (No local oscillator power).

¹ Goodwin, F. E., "A 3.39 micron Infrared Optical Heterodyne Communication System," IEEE Journal of Quantum Electronics, QE-3, No. 11, pp. 524-531, November 1967.

² Goodwin, F. E., and Nussmeier, T. A., "Optical Heterodyne Communications Experiments at 10.6 μ ," Presented at International Quantum Electronics Conference, Miami, Florida, May 14-17, 1968.

Visible frequency detectors are largely of the direct detection type. These operate quite efficiently due to the relatively large photon energies at these frequencies.

The equation describing the signal-to-noise ratio for direct detection is:

$$\frac{S}{N} = \frac{\left(\frac{G\eta q}{hf} P_C \right)^2}{kTB_o G^2 \left(\frac{\eta q}{hf} P_C + \frac{\eta q}{hf} P_B + I_D \right) R_L}$$

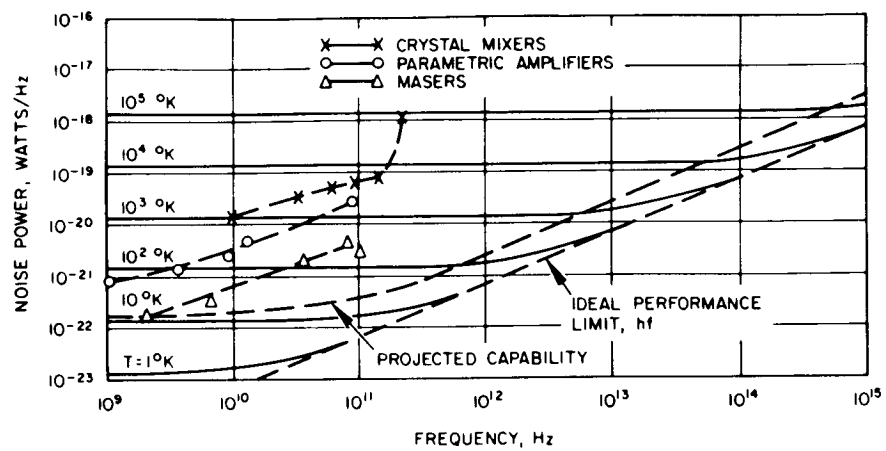
where:

- G = detector gain
- η = detector quantum efficiency
- q = electronic charge, 1.602×10^{-19} coulombs/electron
- h = Plank's constant 6.624×10^{-34} watt sec. sec
- f = light frequency, Hz
- P_C = received carrier power, watts
- R_L = load resistance, ohms
- k = Boltzmann's constant, 1.38×10^{-23} watts/Hz °K
- T = Amplifier noise temperature, °K
- B_o = Amplifier bandwidth, Hz
- P_B = Background received power, watts
- I_D = dark current, amps

Symbolically this is

$$\frac{S}{N} = \frac{\text{Signal Power}}{\text{Thermal noise power} + \text{shot noise power} + \text{background noise power} + \text{dark current noise power}}$$

In any detector application it is necessary to evaluate this equation and adjust the parameters such that an adequate signal to noise ratio is obtained.



Heterodyne Receiver Noise Performance

DETECTORS

Radio Frequency Detectors

	Page
Sensitivity of RF Detectors	190
Radio Frequency Amplifiers	194

SENSITIVITY OF RF DETECTORS

Noise figures for crystal mixers, parametric amplifiers, and masers are given for microwave and millimeter wavelengths.

Detector or receiver sensitivity is conventionally characterized by the effective noise temperature, which is a measure of the receiver noise referred back to its input terminals, or by the noise figure, defined as

$$NF = \frac{S_{in}/N_{in}}{S_{out}/N_{out}}$$

The receiver noise temperature*, T_e , is related to the noise figure when the receiver is connected to a load at 0°K by the expression

$$T_e = (NF-1) T_o$$

where T_o is a standard temperature taken to be 290°K .

The noise performance for several detector implementations is listed in this topic over a broad range of radio frequencies.

Micro/Millimeter-Wave Detectors

The more sensitive detection devices in this frequency range are the conventional crystal mixers, parametric amplifiers, and masers.

Crystal mixers are normally characterized by their conversion loss. The noise figure of the mixer is then

$$NF = L_C (NF_{IF} + N_R - 1)$$

where L_C is the conversion loss, NF_{IF} is the noise figure of the IF amplifier, and N_R is the crystal noise ratio. The conversion loss is defined as the ratio of RF input power to the measured IF output power at the mixer. The crystal noise ratio, N_R , is the ratio of noise power developed by the crystal to the thermal or Johnson noise of an equivalent resistance at 290°K and is typically about 2 in a well-designed system. Estimates of best available noise performance from crystal mixers are summarized in the Table.

* Other system noise considerations are given in Volume IV, Part 1, "Background Radiation and Atmospheric Propagation."

Noise Performance for Radio Detectors

Noise Performance Estimates for Crystal Mixers

Frequency	Conversion Loss*	Noise Figure	Noise Temperature
10 GHz		6 db	870° K
35 GHz	5.5 db	10.3 db	2,800° K
60 GHz	6.5 db	11.3 db	3,600° K
94 GHz	8.2 db	12 db	4,300° K
140 GHz	9 db	13.8 db	6,700° K
200 GHz	19-20 db	25 db	91,000° K
300 GHz		35 db	910,000° K
*Where only conversion losses are quoted, i-f noise figure and crystal noise ratio are taken to be 2.			

SENSITIVITY OF RF DETECTORS

Parametric amplifiers are generally available in the microwave range, but in the millimeter region availability is restricted to experimental models. Diode cutoff frequencies, associated with spreading resistances and junction capacitances, prevent the practical extension of operation beyond 100 GHz except by direct insertion of the diode into a cavity.

Masers are available in the microwave range. As is the case for parametric amplifiers, operation at millimeter wavelengths has been restricted to experimental models. Representative values of the best noise performance at several frequencies are also given in the Table.

Hot carrier detectors are based on the application of microwave power to a noninjecting point contact. As the majority carriers are excited, a temperature gradient between the point contact and broad contact is established, and a unidirectional thermoelectric voltage is generated with frequency following as high as 100 GHz. Work¹ on these detectors is in the experimental stage.

Submillimeter-Wave Detectors

The mechanisms for detection of submillimeter radiation are based on either a thermal effect or a photoelectric effect. Since thermal response times are generally long (in the millisecond range), applications to high-data-rate communication systems are limited to detectors utilizing photoelectric mechanisms. In general, these devices must be cooled to reduce thermal lattice vibrations so that only electrons absorbing the low electron energy (1.2×10^{-3} eV for $\lambda = 1$ mm) are excited to the conduction band. Detection has been demonstrated using several semiconductor materials, and a superconductive detector based on electron tunneling has been proposed.² In addition, detection of submillimeter and millimeter waves by down-conversion to the microwave region has recently been proposed and analyzed.³ At present, however, definitive data are not generally available on detectors in the submillimeter region.

¹Harrison, R. I. and Zucker, J., "Hot-Carrier Microwave Detectors," Proc. IEEE 54, pp. 588-595, 1966.

²Shapiro, S. and Jonus, A. R., "RF Detection by Electron Tunneling between Super Conductors," Proc. 8th International Conf. on Low Temperature Physics., pp. 321-323, Butterworth, London, 1962.

³Krumm, C. F. and Haddad, G. I., "Millimeter- and Submillimeter-Wave Quantum Detectors," Proc. IEEE 54, pp. 627-632, 1966.

Noise Performance for Radio Detectors

State-of-the-Art Performance of Parametric Amplifiers

<u>Frequency</u>	<u>Noise Figure</u>
1 GHz	0.8 db
3 GHz	1.3 db
9 GHz	2 db
14 GHz	3.5 db
94 GHz (estimated)	10 db

State-of-the-Art Performance of Masers

<u>Frequency</u>	<u>Noise Temperature</u>
2 GHz	10-15°K
8 GHz	20-25°K
35 GHz	130°K
81.3 GHz	300°K
94 GHz	200°K

RADIO FREQUENCY AMPLIFIERS

Masers and helium cooled paramps are the best candidates for a very low noise earth receiving station.

There are five low-noise microwave amplifiers which may be considered for use in a deep space communication system. These are

- Transistor Amplifier
- Tunnel Diode Amplifier (TDA)
- Traveling Wave Tube (TWT)
- Parametric Amplifier
- Maser

Brief discussions of the characteristics of these amplifiers are given in the following paragraphs:

Transistor Amplifiers

Microwave transistor amplifiers are relatively new devices which have promise of moderately low noise figures. At the present time noise temperatures of 200 to 625°K can be obtained at frequencies up to about 1 GHz with approximately 20 db gain. It is estimated that in ten years 120° to 170°K noise temperatures are likely at 2 GHz and feasible up to 15 GHz. Transistor amplifiers appear to have their most useful application as a second stage amplifier following an ultra low noise amplifier.

Tunnel Diode Amplifiers

The tunnel diode amplifier (TDA) is the simplest solid-state microwave amplifier and has moderate gain and noise characteristics. Noise temperatures range from about 360°K at 1 GHz to 520°K at 10 GHz using gallium antimonide diodes. Germanium diodes have about 1 db higher noise figures but are available for operation up to about 20 GHz. Single stage amplifiers normally provide about 17 db gain. However, stable gains as high as 30 db can be obtained by careful attention to temperature control and power supply stability. Bandwidths are more than adequate for space communication systems. Noise figures of room temperature TDA's are not expected to improve significantly. Such TDA's will, therefore, be of use mainly as second stage devices following ultra-low noise amplifiers when high noise performance is required.

The noise generated by a tunnel diode amplifier is largely caused by shot noise from the diode bias current. The magnitude of this noise contribution is determined by the characteristics of the material used and can be reduced by using low energy gap materials such as indium antimonide. Such materials must be operated at cryogenic temperatures. However, assuming that a suitable material can be found, a cryogenic TDA with a noise temperature of 20°K at 1 to 2 GHz would have a strong advantage over cooled paramps and masers in that no pump power would be required. A tunnel diode amplifier is shown in Figure A.

Traveling Wave Tubes

These devices offer high gain and moderately low noise figures. Noise temperatures in the 360 to 440°K range are presently obtainable over narrow bandwidths up to 2 GHz. It is unlikely that noise temperatures below about 225°K will be consistently obtained in the next decade. The two major noise sources in a TWT are beam shot noise and thermal noise from the attenuator. Present low noise TWT designs require complex anode structures to achieve space charge smoothing for shot noise reduction. Since no significant advances in space charge smoothing have been made since the late 1950's, it is not anticipated that major progress will be made within the next decade. Some improvement in noise temperatures can be expected by cooling the attenuator, but it does not appear that TWT's will be competitive with cooled paramps.

Parametric Amplifiers

Parametric amplifiers have demonstrated room temperature noise performance superior to that of transistor and tunnel diode amplifiers and cryogenic noise performance approaching that of the maser. In recent years the uncooled parametric amplifier has achieved a level of reliability that has permitted applications on a broad basis and in large numbers. Noise temperatures range from about 60°K at 1 GHz to 250°K at 10 GHz for well-designed narrow band amplifiers.¹ When cooled to 20°K, noise temperatures between 14°K at 1 GHz and 30°K at 10 GHz are possible with careful design using presently available components. Narrow band gains as high as 30 db are possible using extremely stable temperature and pump power control. Major advances for the cooled paramp are likely to be in cost reduction and reliability particularly in the associated cryogenic equipment. A paramp is shown in Figure B.

Masers

Masers find applications in special areas where the ultimate in low noise performance is either dictated by technical requirements or provides the most economical solution to the problem. The noise temperature of the maser itself is approximately that of its physical temperature, about 50°K. To this must be added the noise contribution of the section of input transmission line over which the temperature transition to room temperature is made. For frequencies in the range 1 to 20 GHz this contribution can be held to 5 to 10°K giving an overall maser noise temperature of 10 to 15°K. Gains of better than 30 db with bandwidths of 1 to 2 MHz or more are readily obtainable with a cavity maser. The major disadvantage of masers is that they must operate at a temperature of a few degrees Kelvin in order to provide sufficient gain. The complexity and cost of a cryogenic system increases rapidly as the temperature approaches 0°K.

Future improvements in the maser are likely to be in the area of reliability and cost reduction.

¹Matthei, W. G., "Recent Developments in Solid-State Microwave Devices," *The Microwave Journal*, 9, No. 3, pp. 39-47, March 1966.

RADIO FREQUENCY AMPLIFIERS

Summary

A summary of the important properties of low noise amplifiers is given in the Table. Of the devices surveyed, only the maser and helium cooled paramp can presently meet the low-noise temperature requirements for the down-link receiver in a deep-space communication system, and a system based on reasonable extensions of today's state of the art would require one of these two devices. A suitable maser requires a physical temperature of 4.2°K or less while the paramp can provide adequate noise performance at physical temperatures up to 20°K. It follows that the cryogenic system for the paramp would be considerably less complex and less expensive. Barring the discovery of new types of ultra low noise amplifiers, the helium cooled parametric amplifier operating at 20°K presently appears to provide the most economical solution. Transistor and tunnel diode amplifiers could be used as low noise second stage devices.

Low Noise Preamplifier Characteristics

Amplifier Type	Gain Per Stage, db	Coolant Temperature °K	Relative Cost (1975-1980)	Relative Reliability (1975-1980)	Present Noise Temperatures (Amplifier Only), °K			Estimated Best Noise Temperature (1975-1980) (Amplifier Only), °K		
					Frequency, GHz			Frequency, GHz		
					1	2	10	1	2	10
Traveling Wave Tube	35	290	Medium to High	Medium	380°	400°	1340°	225°	250°	350°
Tunnel Diode Amplifier	20	290	Low	High	360°	380°	625°	310°	330°	500°
	20	20	Medium to Low	Medium	225°*	225°*	525°	200°†	250°†	350°†
Transistor Amplifier	20	290	Low	High	200°	625°	---	75°	125°	170°
Parametric Amplifier	20	290	Medium	High	60°	80°	250°	30°	40°	130°
	20	20	High	Medium	18°	20°	50°	15°	15°	20°
Maser	30	4.3	Very High	Low	10-15°	10-15°	10-15°	12°	12°	12°
*Estimate based on use of indium antimonide tunnel diode at 77°K										
†Feasibility not established										

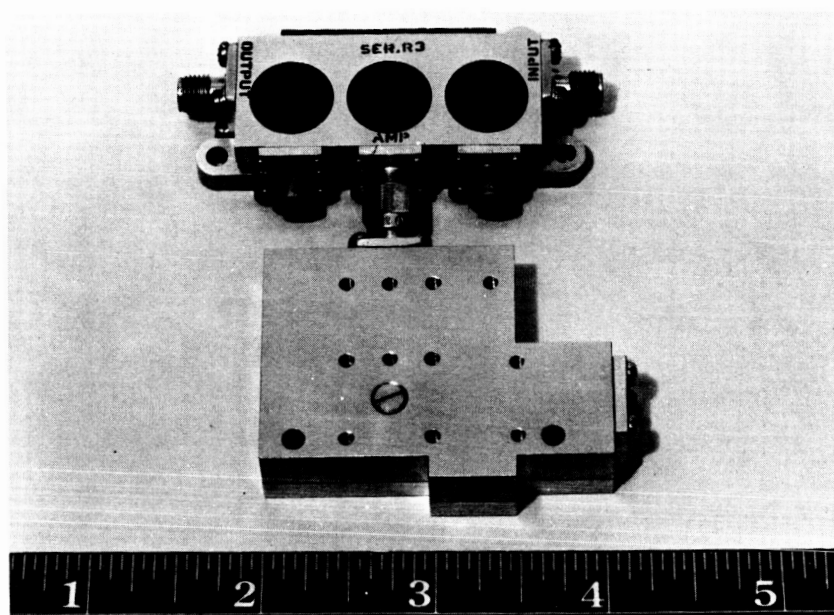


Figure A. Tunnel Diode Amplifier.

(The amplifier is encompassed by isolators — dark circles — and attached to a power supply)

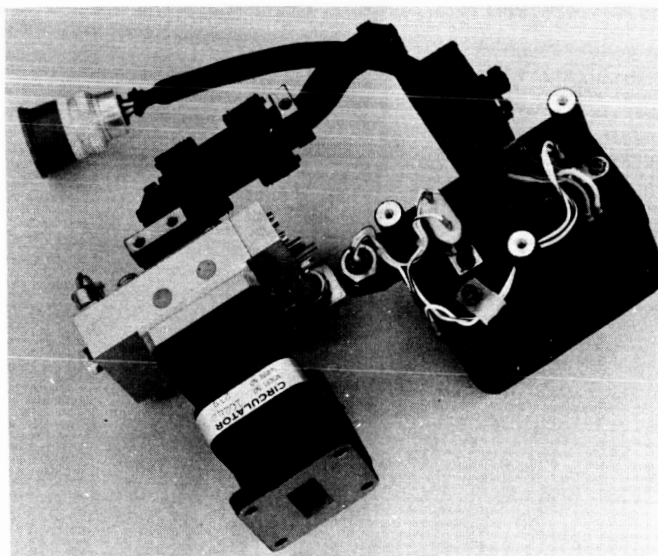


Figure B. Parametric Amplifier — Paramp.

(The amplifier is driven at a frequency higher than the one amplified. The source for this is the dark enclosure to the right of the picture. The paramp proper is housed in the light colored enclosure.)

DETECTORS

Optical Frequency Detectors

Introduction	Page 200
Characterization of Optical Detectors	202
Theory of Photomultiplier Detectorss	204
Solid State Detectors Operating Concepts	210
Detection Limits of Solid State Detectors	214
Performance of Photoemissive Detectors	222
Detectors for 10.6 Microns	224
Detector Performance Summary	228
Burdens for Optical Frequency Detectors	230

INTRODUCTION

Two general implementations for optical detectors are of concern in laser communications, photomultiplier detectors, used in the visible and near visible portion of the spectrum, and semiconductor devices used in the visible and in the infrared.

The material in this sub-section is concerned with two types of optical detection, that done using photomultipliers and that done using semiconductor photodiodes. Immediately following these topics, are topics dealing with the theory of operation of these detectors. Following the theory are topics which give present implementations of these types of detectors.

The final group of topics in this subsection deals with the relations of these detectors to their weight and cost. These burden relationships are to be used in the optimization of space communications systems described in Volume II, Part I of this final report.

Detectors may be used as direct detectors (those where the output is dependent only upon the signal input) or heterodyne detectors (where the output signal is dependent both upon the input signal and upon a local oscillator). Visible frequency detectors are largely of the direct detection type. These operate quite efficiently due to the relatively large photon energies at these frequencies.

More recently heterodyne detectors have been constructed at both 3.39 microns¹ and 10.6 microns². The performance of these detectors is much improved over that of the direct detector.

The performance of optical heterodyne detectors is similar to that of radio heterodyne systems but still differs markedly due to the high operating frequency. This has the effect of changing the dominant noise contribution from a spectral density given by kT (k is Boltzmann's constant and T absolute temperature) to a spectral density given by hf where h is Plank's constant and f is the operating frequency. In general, the noise power spectral density, N , increases with frequency.

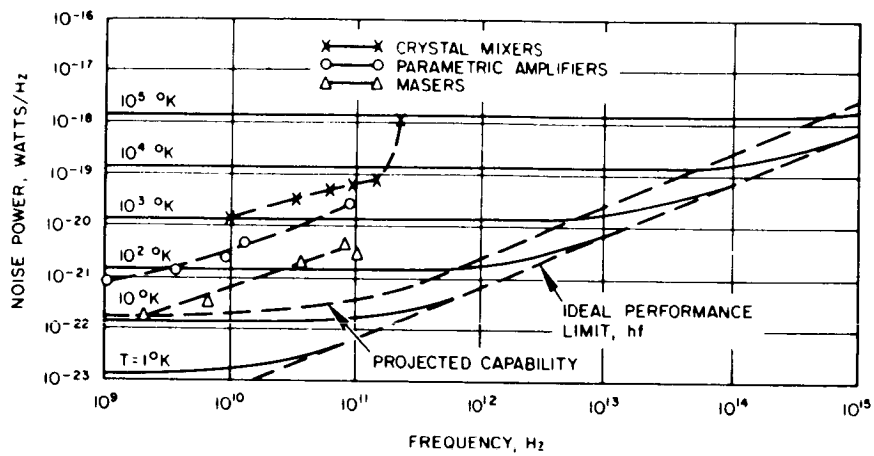
Ideally,

$$N = \frac{hf}{e^{hf/kT} - 1} + hf$$

This curve is plotted as a function of frequency in the figure, for various noise temperatures, $T_{in}^{\circ}K$. The figure compares detector performance over the radio-to-optical spectrum and shows the projected capability of optical detectors (using the above equation) in comparison with known radio receiver performance.

¹Goodwin, F. E., "A 3.39 micron Infrared Optical Heterodyne Communication System," IEEE Journal of Quantum Electronics, QE-3, No. 11, pp. 524-531, November 1967.

²Goodwin, F. E. and Nussmeier, T. A., "Optical Heterodyne Communications Experiments at 10.6 μ ," presented at International Quantum Electronics Conference, Miami, Florida, May 14-17, 1968.



Heterodyne Receiver Noise Performance

CHARACTERIZATION OF OPTICAL DETECTORS

Optical detector performance can be described by certain photon input-electrical output response characteristics such as Responsivity; Detectivity, D^* ; and NEP.

The detectivity, D , is a quantitative measure of the relative performance of a receiver in terms of the signal-to-noise ratio of the detector under a specified level of irradiation.

The responsivity as defined by

$$R(\lambda) = \frac{S(\lambda)}{P(\lambda)}$$

where

$S(\lambda)$ = electrical output of detector at wavelength λ

$P(\lambda)$ = rms value of the incident power at wavelength λ

If $N(\lambda)$ is the receiver noise power, the detectivity is given, in terms of the responsivity by

$$D(\lambda) = \frac{R(\lambda)}{N(\lambda)}$$

A common figure of merit for detector is the noise equivalent power, NEP, given by

$$NEP = 1/D$$

Semiconductor infrared detectors generally have detectivities that vary with bandwidth as $(\Delta f)^{-1/2}$ and with sensitive area as $A^{-1/2}$. A specific detectivity D^* is then defined by the relation

$$D^* \equiv D(\lambda, f) = D \sqrt{(A)(\Delta f)}$$

where λ is the optical wavelength where D^* is measured and f is the frequency at which D^* is measured.

THEORY OF PHOTOMULTIPLIER DETECTORS

Photomultiplier detectors operate best at visible frequencies where the incoming photon energy is large enough to overcome the work function of the surface and cause electron emission.

The most useful lasers for optical communications systems are found in the wavelength region from 0.4 to 10 μ ; i. e., from the visible through the near infrared region of the spectrum. A variety of detection mechanisms and devices have been used in this spectral region, but not all are useful for the contemplated systems applications. Thermal detectors, for example, are much too slow, having time constants no shorter than several milliseconds, and their use is generally restricted to spectroscopy and energy reference measurements. Detectors based on mechanisms in which the energy of the absorbed photon goes into direct electronic excitation of the material are much more useful where fast response is required. These include the photo-emissive detectors generally used in the visible and the semiconductor devices applied in the infrared and visible. This topic will be limited to photo emissive detectors while the following topic deals with semiconductor devices.

In the visible and near-infrared region of the spectrum, the photomultiplier is the most efficient and convenient detector of radiation. It is based on the external photoelectric effect and the subsequent amplification of the electron current by means of a number of secondary emitting stages termed "dynodes." The basic and determining process in the photomultiplier is the external photoelectric effect, which consists of two steps: absorption of light by a solid and emission of an electron.

Electron emission will take place only when a photon possesses sufficient energy to overcome the work function of the solid, i. e., when

$$h\nu \geq e\Phi$$

where e is the electronic charge and Φ is the work function. This relation sets a long wavelength limit for every photoemissive detector; it may be put in the form

$$\lambda_L = \frac{1.24}{e\Phi}$$

where λ_L is the wavelength limit in microns and $e\Phi$ is the work function of the photocathode in electron volts.

The element with the lowest work function is cesium. For this element $e\Phi = 1.9$ eV; therefore, the wavelength limit of a cesium photocathode is about 6500 Å. Composite photocathodes consisting of combinations of metals and oxides have lower work functions; they are capable of functioning up to about $\lambda = 1.2\mu$. Naturally, as the limit is approached the efficiency of the detector decreases. The variation of detector efficiency with wavelength is apparent from the responsivity curves or the curves of quantum efficiency, which in the case of photoemissive detectors are simply related to the responsivity curves.

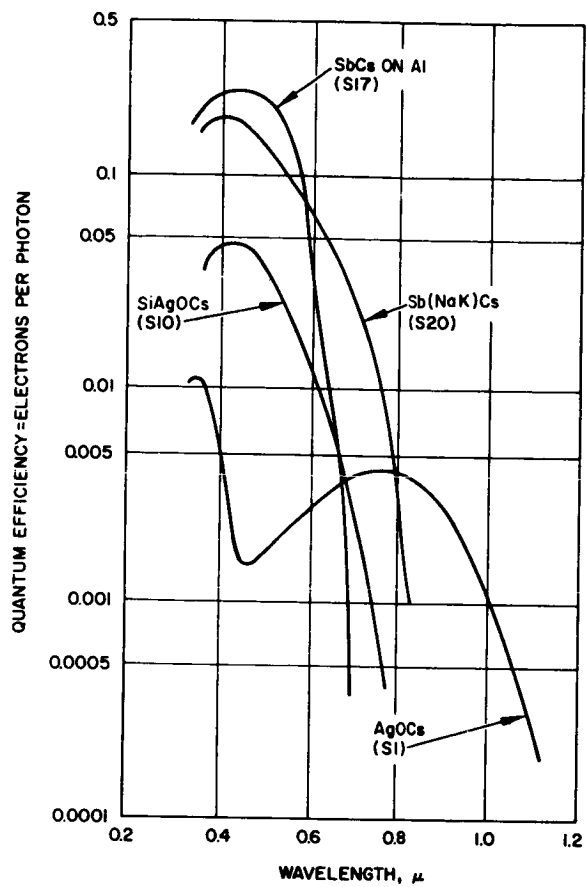


Figure A. Spectral Response of Various Photocathodes

THEORY OF PHOTOMULTIPLIER DETECTORS

Quantum efficiency, η , is the number of emitted electrons divided by the number of incident photons. The frequency, ν , quantum efficiency, η , incident power, P , and the photoelectric current, i , are related as follows: The number of incident quanta per second is $n = P/h\nu$, the electron current emitted from the photosurface is ηne ; therefore

$$i = \eta e P/h\nu$$

Responsivity $R(\nu)$ is proportional to i/P ; therefore

$$R(\nu) = \frac{Ge\eta(\nu)}{h\nu}$$

where G is the gain of the photomultiplier. The quantum efficiencies, $\eta(\nu)$, of the red-sensitive photocathodes are shown as a function of in Figure A.

As a circuit element, a photomultiplier appears simply as a capacitance in parallel with a current generator of magnitude $i = Ge\eta P/h\nu$, as in Figure B.

In the photomultiplier there is shot noise due to the fluctuations in the mean dc current. This current is the sum of dark currents that flow in the absence of any light input and the average photocurrent due to both signal and extraneous optical inputs. These sources together contribute a white noise power spectrum with a circuit representation as a current generator with a mean square amplitude (Figure C).

$$\overline{i_{nd}^2}(f) = \left(2eI_d + 2e^2\eta \frac{P_{av}}{h\nu} G \right) \Delta f$$

I_d is the dark current and P_{av} the average incident optical power. There will also be a noise contribution from the resistive part of the equivalent load R_L whose magnitude is

$$i_{nr}^2(f) = \frac{4kT_e}{R_L} \Delta f$$

T_e is an effective temperature used to characterize the noise performance of the receiver. Photomultiplier detectivity is given by:

$$D = \frac{Ge\eta}{h\nu} \frac{1}{\left[2eI_d + 2e^2\eta \left(P_{av}/h\nu \right) G + \left(4kT_e/R_L \right) \right]^{1/2} (\Delta f)^{1/2}}$$

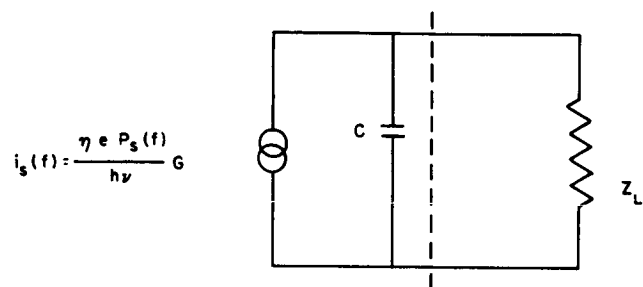
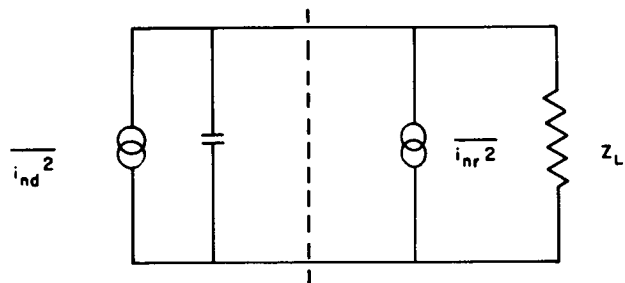


Figure B. Signal Equivalent Circuit
of Photomultiplier

THEORY OF PHOTOMULTIPLIER DETECTORS

The response times of ordinary photomultipliers are between 1 and 3 nsec. The responsivity is fairly uniform until the frequency $f = 1/2\pi\tau$ is reached. Thus the performance of the commercial photomultipliers begins to be degraded between 50 and 150 MHz modulation frequency. Special tubes are required for the detection or demodulation of signals varying faster rates.



$$\overline{i_{nd}^2} = (2eI_d + 2e^2 \eta \frac{P}{h\nu} G) \Delta f$$

$$\overline{i_{nr}^2} = \frac{4k T_e \Delta f}{R_L}$$

Figure C. Noise Equivalent Circuit of Photomultiplier

SOLID STATE DETECTORS OPERATING CONCEPTS

Photoconductive, photoelectromagnetic, and photovoltaic modes of detection are described.

The treatment of the various internal photoeffects in semiconductors can be presented in a unified fashion. Therefore both intrinsic and extrinsic mechanisms will be discussed in this topic.

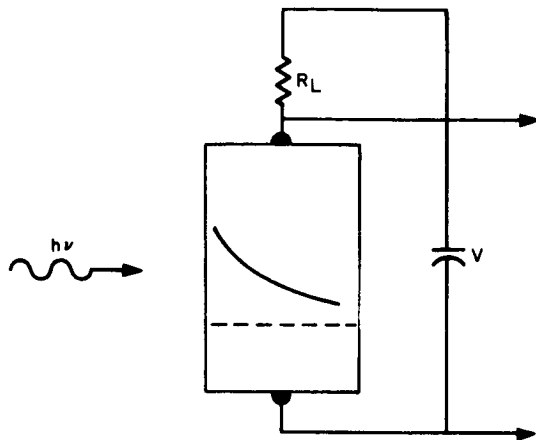
At wavelengths greater than 1.1μ , photoemissive devices no longer have sufficient sensitivity and detectors utilizing internal photoeffects must be employed. Solid state detectors with detectivities approaching the theoretical photon noise limit in the wavelength range 0.4μ to 10μ are available as a result of the considerable progress in infrared technology in the past decade.

There are two internal photoeffects which are the basis of solid state detectors operating in the near infrared. In both cases, absorption of photons leads to a change in the concentration of free, mobile charge carriers within the material. In the first class of detectors, called intrinsic, the energy of an absorbed photon creates an electron-hole pair, i. e., the excitation process raises an electron from a valance band state to a conduction band state and only photons with energies greater than the intrinsic band gap are effective. The excitation process in the second class of detectors, called extrinsic, is the ionization of an impurity center to produce a free carrier and a charged defect site. The optical absorption constant in intrinsic materials is large, ranging up to 10^5 cm^{-1} , whereas in extrinsic materials it is rarely greater than 10^2 cm^{-1} . Photogenerated carriers are therefore confined to much smaller regions of intrinsic detectors.

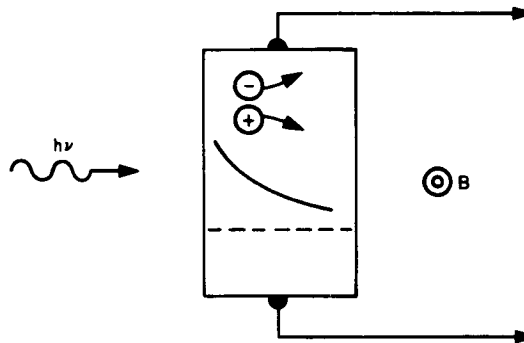
In intrinsic detectors there are three commonly used techniques for sensing the rate at which electron-hole pairs are generated by the incident light. These are the photoconductive, the photoelectromagnetic, and the photovoltaic modes of operation.

Figure A-1 represents a detector operated in the photoconductive mode. The dashed line represents the steady state electron-hole concentration in a uniform block of material. The current that flows in response to an electric field applied between the electrodes provides a measure of the free carrier concentration. When signal photons are incident on the detector, an additional concentration of carriers is set up which varies with position in the detector as shown by the solid curve. Because of the additional carriers a larger current flows in the external circuit.

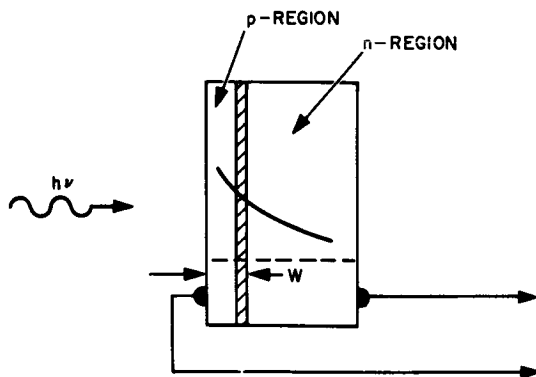
The photoelectromagnetic mode of operation is obtained if the electric field is removed from the photoconductor so that it is connected directly across a load and a magnetic field is applied perpendicular to the plane of the figure. Because of the concentration gradient of electron-hole pairs produced by the absorbed radiation in the material, carriers drift in the direction in which radiation is incident. The Lorentz force due to the magnetic field on the moving carriers deflects the electrons and holes to opposite electrodes and produces current in the external load as illustrated in Figure A-2.



1) Photoconductive Mode



2) Photoelectromagnetic Mode



3) Photovoltaic Mode

Figure A. Three Modes of Operation of Solid State Photodetectors

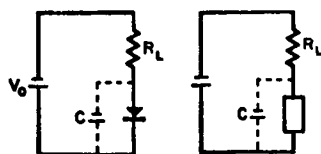


Figure B. Operating Circuits of Photodiode and Photoconductor Detectors

SOLID STATE DETECTORS OPERATING CONCEPTS

The third mode of operation is the photovoltaic mode in which a p-n junction is produced immediately behind the surface on which radiation is incident by diffusing a p-type dopant into n-type material (Figure A-3). The generated electron hole pairs diffuse to the junction under the influence of the concentration gradients set up in response to the incident radiation. The electric fields in the junction drive the electrons to the n-side and holes to the p-side of the detector. With zero applied bias, an external photocurrent can be observed or the detector may be operated with reverse bias in what is essentially a photoconductive mode.

By appropriate control of the fabrication process, it is possible in some semiconductors to make the transition region between the n and p regions large or even to arrange that the detector consist of three regions: a thick n-type base region on which there is first a moderately thick intrinsic region, and then a very thin p region at the surface on which radiation is incident. Such a p-i-n structure has somewhat different response time characteristics when biased in the reverse direction and is particularly useful for weakly absorbed radiation.

Typical circuits in which the reverse biased photodiode and the photoconductor are operated are shown in Figure B. The load resistor R_L represents the input resistance of the amplifier, and C includes all shunt capacities in the input circuit. The current-voltage characteristics of these two devices are represented schematically in Figure C for various incident photon fluxes. The direction of the arrow indicates increasing incident flux.

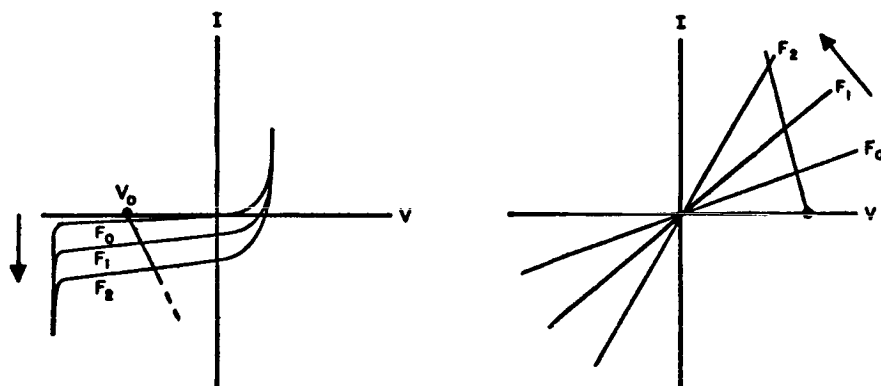


Figure C. Representative Current-Voltage Characteristics of a Photodiode (left) and a Photoconductor (right) for Various Values of Incident Flux

Arrows indicate directions of increasing flux,
i.e., $F_2 > F_1 > F_0$.

DETECTION LIMITS OF SOLID STATE DETECTORS

Signal to noise ratios are derived for solid state detectors considering the various types of noises which may be present.

The constant current equivalent input circuits for photodiode and photoconductor detectors are shown in Figure A. They are identical except for the inclusion of the series resistance R_s in the diode circuit. R_s is usually of the order of a few ohms for most well made diodes and so is negligible except for extremely high frequency operation.

Current flows in the detector circuits in response to the incident photon fluxes. In the diode each generated electron-hole pair results in the traversal of the external circuit by an elementary charge. For the photoconductor, the generated free carriers contribute to the external current only for a mean time, τ , their lifetime. Lifetime effects may reduce the response of a diode, but the fast diode can be designed to minimize this effect. The constant current generators in the equivalent circuits of Figure A accordingly provide currents given by

$$I = I_{dc} + i_s(\omega) = e\eta M(F_B + F_{Lo} + F_{so}) + e\eta M F_{so} S(\omega) e^{j\omega t}$$

When I_{dc} is the direct current

$i_s(\omega)$ is the frequency dependant current

e is the electronic charge

η is the quantum efficiency

M is the diode multiplication factor

F_B is the background photon flux

F_{Lo} is the laser local oscillator photon flux

F_{so} is the signal photon flux

$S(\omega)$ is the frequency degradation due to transit time and lifetime effects.

This expression is applicable to both the diode and the photoconductor. In the ordinary diode the factor M is always equal to unity; when an avalanche diode which provides internal gain through impact ionization is used, M is the multiplication factor of the diode. In the photoconductor, $M = \tau/\tau_R$, where τ is the lifetime of the free carriers and τ_R is their transit time (i.e., the time required for a free carrier to traverse the distance between the contacts on the device). The quantity τ_R is given by $L^2/\mu V$ where L is the length of the detector in the direction of current flow, μ is the mobility of the free carriers, and V is the bias voltage applied across the detector. Note that $V/L = E$, where E is the electric field intensity in the sample.

$S(\omega)$ is a frequency dependent term that gives the degradation of the detector response as a result of lifetime and carrier transport time effects. Its form for the diode can be quite complicated, depending on its detailed construction. However, it quite generally indicates that the response drops at frequencies so large that $\omega\tau_R \geq 1$. For the photoconductor, where it results from the finite lifetime of the carriers, it has the form $S(\omega) = (1 + j\omega\tau)^{-1}$.

In addition to the currents, which are caused by photons incident on the detector, internal currents may be present which would flow even if the detector were in a thermal equilibrium environment with no external incident radiation. In diodes there will be the usual reverse saturation currents and in some cases leakage currents around the junction region. For the photoconductor, leakage currents are possible under low background conditions where sample impedances would be large. Thermally generated currents will appear if the detector operating temperature is too high. Although such currents will not affect the signal currents at frequency, ω , they must be considered as possible sources of undesirable noise.

Noise at the signal frequency will arise from a number of sources in the detector and its associated circuitry. A basic noise source, over which there is no possibility of control through circuit design or detector fabrication techniques, is fluctuations in the photon-induced currents. In the diode this noise is referred to as shot noise and in the photoconductor as generation-recombination (g - r) noise. The mean power spectrum for this noise can be represented by

$$\overline{i^2(\omega)} = 2eM_n I_{dc}$$

for both structures. M_n is 1 for the diode and is $2(\tau/\tau_R)$ for the photoconductor. It should be noted that M_n does not have the same values for different types of detectors. I_{dc} includes both internally generated and photon generated currents.

A second type of noise whose effects can usually be minimized by appropriate circuit design is the thermal equilibrium Nyquist noise arising in the resistive components in the input circuits. When extremely wide bandwidth operation is required, Nyquist noise usually determines the performance limits of the detector.

In addition to these two unavoidable noises, practical detectors and amplifiers exhibit excess noises that are difficult to control and whose sources may not even be well understood. The $1/f$ noises found at low frequencies in almost all semiconductor devices are included in this group, as are the excess noises arising in preamplifiers. Since $1/f$ noise in the detectors can usually be made negligible above at least a few kilocycles by proper fabrication techniques, this noise source will be neglected in the following treatment. Excess noises arising in the preamplifier will also be neglected for the sake of clarity. Only shot, generation-recombination, and Nyquist noise will be considered of further.

DETECTION LIMITS OF SOLID STATE DETECTORS

The noise equivalent circuit for the diode and the photoconductor is shown in Figure B. For this circuit the noise current power spectrum is given simply by

$$\overline{i_n^2(\omega)} = 2eI_{dc} |S(\omega)|^2 + 4kT/R_p$$

or more explicitly by

$$\begin{aligned} \overline{i_n^2(\omega)} = & 2eI_i M_n |S(\omega)|^2 + 2e^2 \eta M M_n (F_B + F_{Lo} \\ & + F_{so}) |S(\omega)|^2 + 4kT/R_p \end{aligned}$$

where

k is Boltzmann's constant

T is absolute temperature

R_p is the parallel resistance of the detector and load

I_i is internal current

where the internal currents and the photon generated currents have been written explicitly. F_B represents the background photon flux. The significance of the M factor is:

$M = 1$ ordinary diode

$= \tau/\tau_R$ photoconductor

$M_n = 1$ ordinary diode

$= 2\tau/\tau_R$ photoconductor

Note that the transit and recombination time factors affect the shot and generation-recombination noise but not the thermal noise.

The mean square signal and noise voltages appearing across the load resistors R_L in the equivalent circuits are written and compared to obtain the signal-to-noise ratio of the detector. The results are

$$S = \overline{v_s^2(\omega)} = \frac{1}{2} e^2 \eta^2 M^2 F_{so}^2 |S(\omega)|^2 |Z(\omega)|^2 \Delta f$$

$$N = \overline{v_n^2} = \left\{ 2e \left[I_i M_n + e \eta M M_n (F_B + F_{Lo} + F_{so}) \right] |S(\omega)|^2 + 4kT/R_p \right\} |Z(\omega)|^2 \Delta f$$

$Z(\omega) = R_p (1 + j\omega C R_p)^{-1}$ where $R_p = R R_L (R + R_L)$. R is the detector resistance; R_L is the load resistor and C is the sum of all shunt capacities in the input network. Δf is a small frequency interval centered around $f = \omega/2\pi$, the signal frequency. The shunt capacity C is the sum of detector capacity and circuit capacity. For diodes, their own capacitance, even for very fast detectors, will usually be at least several picofarads. For photoconductors, this capacitance is predominantly that of the input circuitry.

Normally in very high frequency, wide bandwidth operation R_L , the load resistance or equivalent amplifier input resistance, will be small — of the order of 50Ω . Except for very poor detectors with large internal currents I_i , the thermal noise term $4kT/R_p$ will be larger than the shot or g-r term. In this condition the optimum signal-to-noise ratio results when maximum signal power transfer to the load takes place.

The minimum detectable photon flux in this thermal-noise-limited condition is given for narrow bandwidths by

$$F_{so, \min} = \frac{(8kT/R_p)^{1/2}}{e \eta M} (\Delta f)^{1/2}$$

If the operating bandwidth is taken as the effective noise bandwidth, determined from the impedance function $Z(\omega) = R_p (1 + j\omega R_p C)^{-1}$ by the relation

$$B = R_p^{-2} \int_0^\infty |Z(\omega)|^2 df = (4R_p C)^{-1}$$

then the minimum detectable photon flux in this bandwidth is

$$F_{so, \min} = \frac{4(2kTC)^{1/2} B}{e \eta M}$$

and the detectivity for signals at optical frequency ν is

$$D = \frac{e \eta M}{h \nu} \frac{1}{4(2kTC)^{1/2} B}$$

DETECTION LIMITS OF SOLID STATE DETECTORS

Considering the other extreme of low frequency, narrow bandwidth operation, where $R_p = (4BC)^{-1}$ can be large, it is possible to reach a bandwidth region where the shot or g-r noise term dominates. For simple direct detection $F_{Lo} = 0$ and, in the case of a good detector for which the internal currents are negligible, the signal-to-noise ratio will be given by

$$\frac{S}{N} = \frac{\overline{v_s^2}}{\overline{v_n^2}} = \frac{\eta M F_{so}^2}{4 M_n (F_B + F_{so}) \Delta f}$$

and the minimum detectable photon flux is given by

$$F_{so, \min} = \left(\frac{2 M_n F_B B}{\eta M} \right)^{1/2} = \frac{1}{h \nu D}$$

which is seen to vary with the square root of the bandwidth B . The transition from the low pass-band shot or g-r noise limited condition to the thermal noise limited condition occurs when the shunt resistance R_p has value

$$R_p = \frac{2kT}{e^2 \eta M M_n (F_B + F_{so})}$$

The pass-band at which the change from shot to thermal noise limited performance takes place is

$$B = \frac{e^2 \eta M M_n F_B}{8kTC}$$

The goal of detector development is to depress the thermal noise portion of the curve so that the shot or g-r noise determined minimum detectable signal can be realized over any desired pass-band.

From the expressions for S and N given earlier, it is apparent that if the signal term and the photon-generated current noise term could be amplified, shot and g-r noise limited performance would be obtained for smaller values of R_p , i. e., over wider pass-bands, B . The techniques which have been suggested for accomplishing this in the case of incoherent detection include (1) high frequency parametric amplification in the photodiode; (2) building diodes in which internal gain is achieved by charge carrier multiplication caused by impact ionization; (3) making the photocurrent gain ratio τ/τ_R greater than 1 in photoconductors.

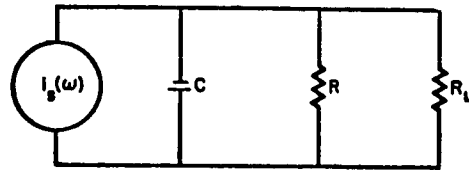
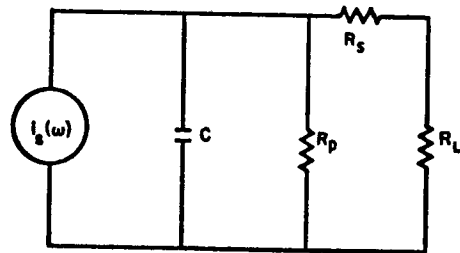


Figure A. Constant Current Equivalent Circuits of the Photodiode (Upper) and Photoconductor (Lower)

DETECTION LIMITS OF SOLID STATE DETECTORS

Heterodyne operation of a detector provides a very effective means of achieving shot and g-r noise limited performance and, when sufficient local oscillator power can be applied, may even render background contributions to the noise negligible so that a signal fluctuation limit is reached. The expression for the signal-to-noise ratio in this case becomes

$$\frac{S}{N} = \frac{2e^2 \eta^2 M^2 F_{so} F_{Lo} S^2}{\Delta f \left\{ 4e^2 \eta M M_n (F_B + F_{Lo}) S^2 + 4kT/R_p \right\}}$$

For F_{Lo} sufficiently large the only term of significance in the denominator is

$$4e^2 \eta M M_n F_{Lo} S^2,$$

and S/N is simply

$$\frac{S}{N} = \frac{\eta M F_{so}}{2M_n \Delta f}$$

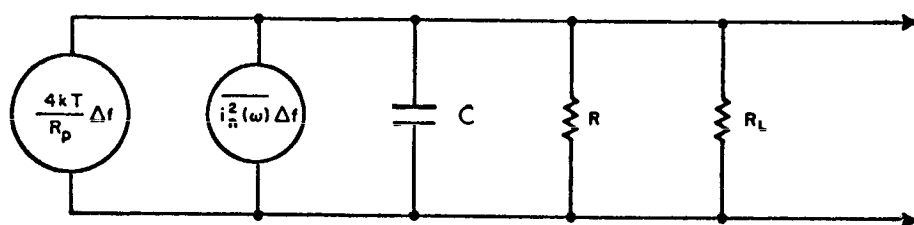


Figure B. Noise Equivalent Circuit for the Photodiode and Photoconductor

$\overline{v_R^2} \Delta f$ represents excess noise arising in the preamplifier and has been neglected in the text.

PERFORMANCE OF PHOTOEMISSIVE DETECTORS

Detail Performance Characteristics are given for available photoemissive detectors at the frequencies for ruby, helium-neon, neodymium, and neon lasers.

The following discussion will be devoted to presenting detail on available photomultipliers at several available laser wavelengths.

The following four wavelengths are particularly important in laser technology for the visible and near visible part of the spectrum: 6328 Å, 6943 Å, 1.06μ, and 1.13μ. These are the wavelengths of helium-neon, ruby, neodymium, and neon lasers, respectively. Table A contains quantum efficiencies and relative responsivities of the photosurfaces S-1, -10, -17, and -20 at these important frequencies.* Clearly only the S-1 surface is suitable for 1.06 and 1.13 radiation, and the S-17 surface is inferior to the others at the two shorter wavelengths.

A number of photomultiplier tubes are available with these surfaces. The responsivity and the noise equivalent input of the photomultipliers are usually given in microamperes per lumen** and in lumens. A test lamp of color temperature 2870°K is employed in place of a monochromatic source to determine responsivity and NEP. The values of NEP are given in Table B.

The absolute responsivity of a photomultiplier depends on the voltages applied to its dynodes. The manufacturers usually state the responsivity under specified conditions in microamperes per lumen, the source of irradiation being a black body at 2870°K temperature. The calculation of the signal-to-noise ratio is based on the NEP of Table B plus the shot noise plus other noise that may be entering the detector with the signal, including the noise due to signal fluctuations.

D. Gunter, et al., have reported signal enhancement in photomultipliers by factors as large as four by making use of multiple reflection in the glass face of the tube on which the photocathode is deposited. They added a glass quadrant to the tube and noted that when the angle of incidence increases above that characteristic of total internal reflection the response increased a factor of 4 in the red and 2 in the blue.

* See topic entitled "Theory of Photomultiplier Detectors" for optical frequency response of these detectors.

** 1 lumen = 1.496×10^{-3} watts, at max visibility, $\lambda = 5560 \text{ Å}$.

Table A. Quantum Efficiencies and Responsivities of Red-Sensitivity Photosurfaces

Photo-surface	6328 Å		6934 Å		1.06 μ		1.13 μ	
	Quantum Efficiency	Respon-sivity	Quantum Efficiency	Respon-sivity	Quantum Efficiency	Respon-sivity	Quantum Efficiency	Respon-sivity
S-1	0.0033	0.56	0.0040	0.80	0.0009	0.29	0.00016	0.03
S-10	0.0070	0.30	0.0028	0.10	0	0	0	0
S-17	0.0080	0.08	0.0004	0.017	0	0	0	0
S-20	0.045	0.41	0.028	0.22	0	0	0	0

Table B. NEP in One-Cycle Band at 6328 Å and 6943 Å for Several Photomultipliers

Tube	NEP, W-sec ^{-1/2}		Temperature, °C
	6328 Å	6943 Å	
7102	3.0×10^{-12}	2.1×10^{-12}	25
6217	2.6×10^{-13}	7.9×10^{-13}	25
7265	4.3×10^{-15}	8.0×10^{-15}	25
7265	5.6×10^{-16}	10.4×10^{-16}	-80
7326	1.07×10^{-14}	2.0×10^{-14}	25
7326	1.6×10^{-15}	3.2×10^{-15}	-80
<p>For $\lambda = 1.06 \mu$ the NEP of the 7102 tube is 5.9×10^{-12} W-sec^{-1/2}. For $\lambda = 1.13 \mu$ it is 5.7×10^{-11} W-sec^{-1/2}.</p>			

DETECTORS FOR 10.6 MICRONS

Five solid state detectors can be considered for detection of 10.6 μ energy. Of these copper doped germanium is preferred.

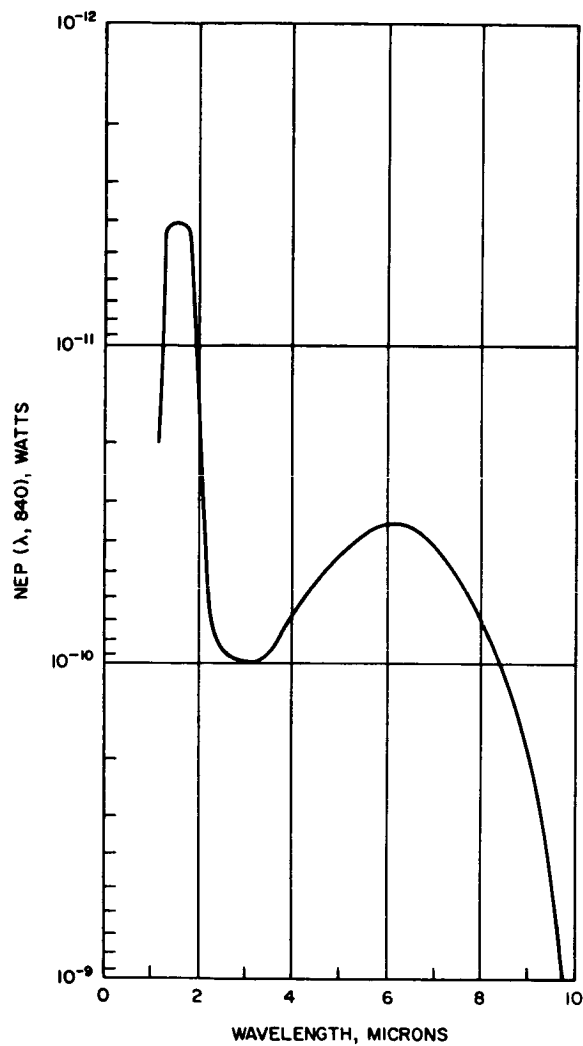
Five semiconductor materials can be considered for the detector in a 10.6 μ laser system. These materials, with their characteristics cutoff wavelength and maximum operating temperature, are listed in the Table. Spectral response curves and plots of the temperature variation of detectivity for several of these materials are shown in Figures A and B, respectively.

The first four semiconductor materials in the Table are extrinsic (impurity) photoconductors. The acceptor impurity concentrations for these materials are typically $2-3 \times 10^{14} \text{ cm}^{-3}$ for cadmium, $2-3 \times 10^{15} \text{ cm}^{-3}$ for mercury and gold, and $1-2 \times 10^{16} \text{ cm}^{-3}$ for copper. Above the critical temperature T_{max} , carriers are produced by thermal ionization of these impurity centers and carrier concentration increases exponentially with temperature. Below this critical temperature the concentration of carriers in the semiconductor is determined by the magnitude of the background radiation, and the detector is defined as operating in the background limited photoconductivity condition. The upper limit of detectivity (D^*) is set by the level of background radiation on the detector and the quantum efficiency. Above the critical temperature in the region of thermal ionization, the detectivity decreases exponentially with temperature.

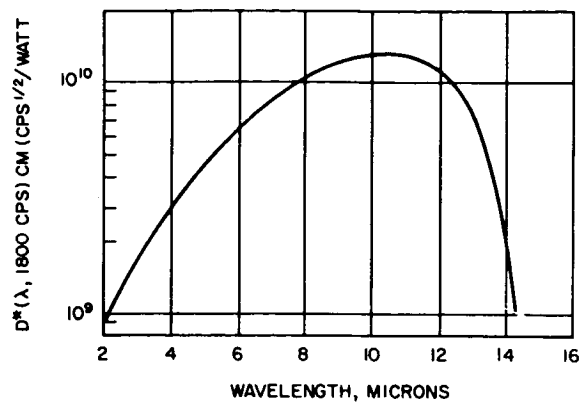
Of the three extrinsic materials listed in the Table, Ge:Cu is preferred for many applications for the following reasons:

1. Ge:Cu is the easiest to prepare since the copper is introduced by solid state diffusion. Copper has a very high diffusion coefficient in germanium. (Conversely Ge:Hg and Ge:Cd are prepared by horizontal zone leveling under a high vapor pressure of the constituent impurity.)
2. The higher impurity concentrations possible in Ge:Cu produce a high absorption coefficient and permit higher quantum efficiencies in a smaller sample thickness.
3. In general, Ge:Cu detector samples exhibit a lower resistance than either Ge:Hg or Ge:Cd. This factor is a distinct advantage in designing a detector for very high frequencies where a low RC time constant is mandatory.

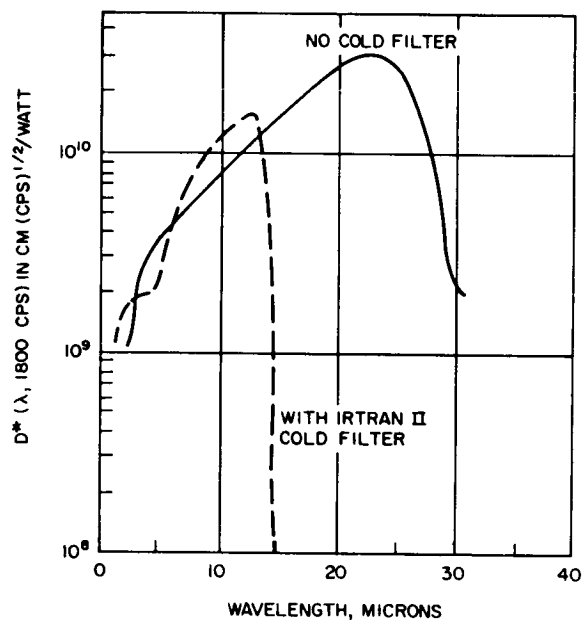
The chief disadvantage of the copper-doped germanium detector is the requirement for a very low operating temperature, i. e., liquid helium cooling. In closed cycle cooling a two-stage Stirling cycle cooler is necessary. The mercury-doped germanium detector has the major advantage of operation at temperatures up to about twice the absolute maximum temperature allowed for Ge:Cu. From the practical point of view, operation at liquid hydrogen and liquid neon temperature, as well as liquid helium, represents a distinct advantage. The use of single stage Stirling cycle refrigeration is an additional advantage when closed cycle cooling is desired. Ge:Au operates at the highest temperature of



1. Gold-doped germanium.



2. Mercury-doped germanium.



3. Copper-doped germanium.

Figure A. Monochromatic Detectivity as a Function of Wavelength for Several Extrinsic Photoconductors

DETECTORS FOR 10.6 MICRONS

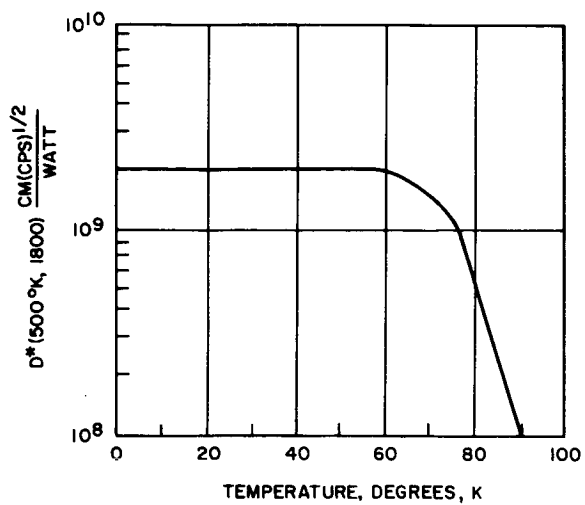
all the extrinsic photoconductors and can even be used at 77°K (liquid nitrogen). However, at this temperature its detectivity is a factor of two less than its peak value. Also, its detectivity at 10 μ is two orders of magnitude less than its peak value. Preliminary results of measurements of the frequency response of especially compensated mercury- and copper-doped germanium detectors¹ manufactured by the Santa Barbara Research Center indicate they exhibit a flat frequency response up to 300 MHz. From this it is concluded that the inherent lifetime of these materials is less than 5×10^{-10} seconds.

The last material listed in the Table ($\text{Hg}_{1-x}\text{Cd}_x\text{Te}$) is an intrinsic semiconductor. This material currently is just emerging from the status of a laboratory curiosity. The yield of good material with proper composition and purity has been very low, but research efforts at a number of laboratories are gradually solving some of these problems. As technology advances and the yield improves, this particular material offers great promise for operation to 12 μ at liquid nitrogen temperature (77°K).

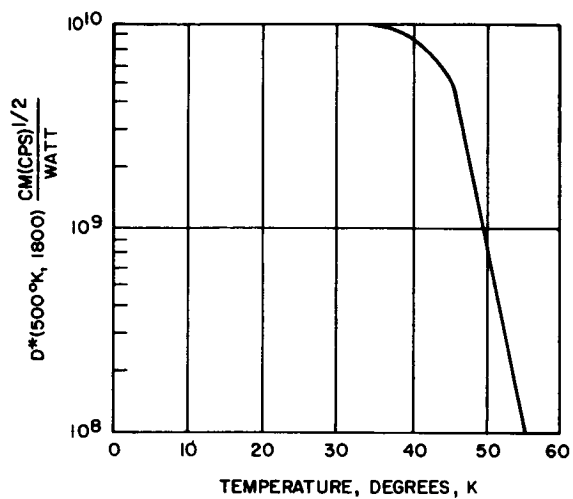
Characteristics of Semiconductor Materials for 10.6 μ Detector

Detector Material	$\lambda^{1/2}, \mu^a$	$T_{\text{max}}, ^\circ\text{K}^b$
Ge:Au	~9	70
Ge:Hg	14	40
Ge:Cd	22	25
Ge:Cu	28	18
$\text{Hg}_{1-x}\text{Cd}_x\text{Te}$	12	77
^a Wavelength at which detectivity decreases to 1/2 its peak value. ^b Temperature at which detectivity decreases to $1/\sqrt{2}$ of its maximum value.		

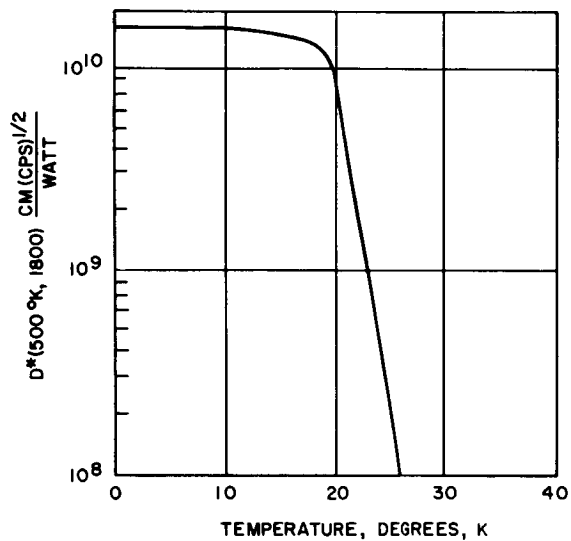
¹Picus, G. S., Interim Technical Report No. 2 Contract AF33(615)-3847, Hughes Research Laboratories, Malibu, California.



a. Gold-doped germanium.



b. Mercury-doped germanium.



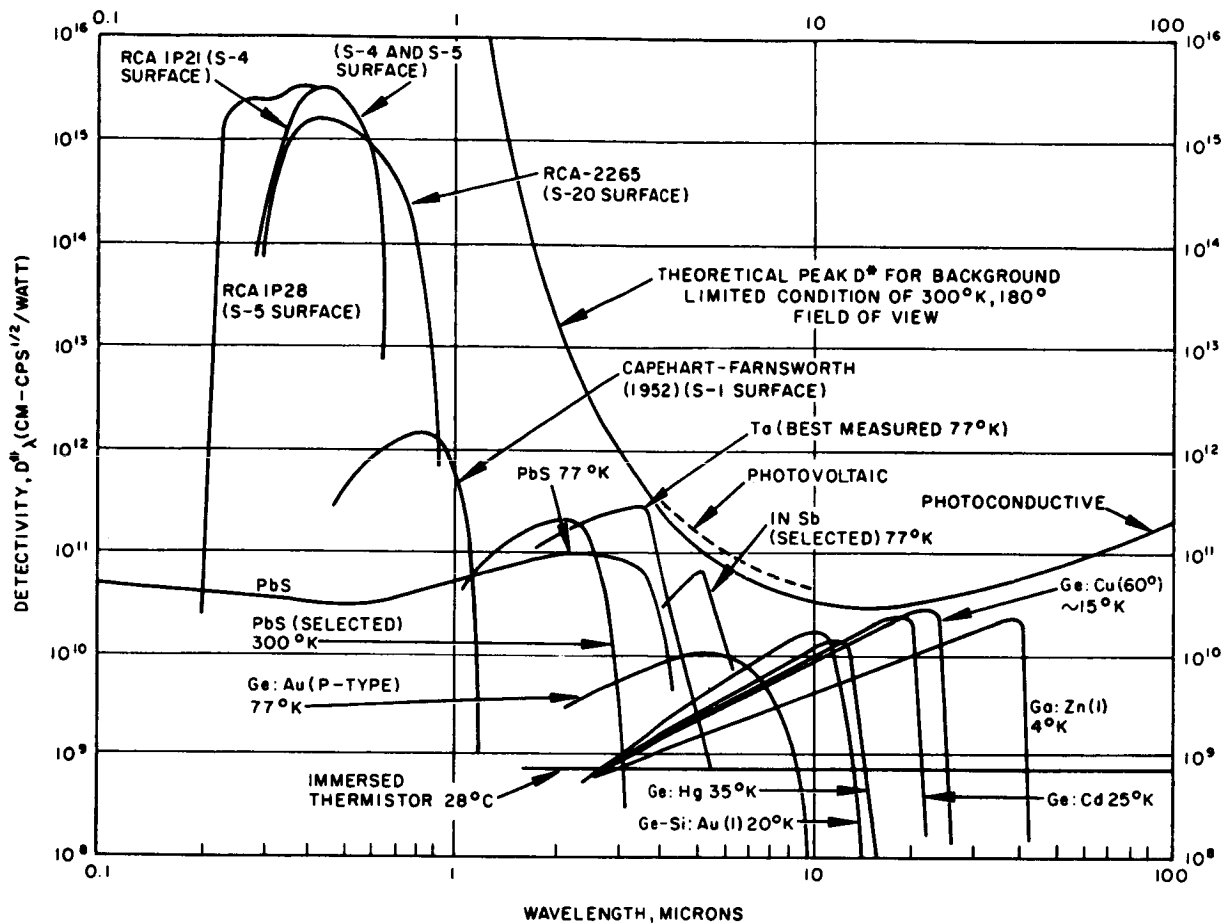
c. Copper-doped germanium.

Figure B. Blackbody Detectivity as a Function of Temperature for Several Extrinsic Photoconductors

DETECTOR PERFORMANCE SUMMARY

A summary of detector detectivity is given which spans wavelengths of 0.1 micron to 40 microns.

A summary of detector performance is given in the Figure. This material was first published in ASD-TDR-63-185, "Investigation of Optical Spectral Regions for Space Communications." May 1963, AF Avionics Laboratory, Wright-Patterson AFB, Ohio. (Unclassified)



Detectivity versus Wavelength for the Best Detectors

BURDENS FOR OPTICAL FREQUENCY DETECTORS

Burdens of weight, cost, and input power requirements are given for optical demodulations, which include the detector. The total demodulator also includes the circuitry required to produce a noise free (but not necessarily error free) bit stream.

A complete description of a detector must cover the following three areas in order to be useful to the systems designer.

1. The optical configuration—this means the physical size and shape of the detector and its housing and the extent to which these can be modified. These properties are important in the selection and design of both system optics and electronics.
2. The electrical configuration—the most important feature here is the characterization of the detector as an active circuit element. This information is required in order to calculate the electrical signal delivered to the electronics as a function of the radiation signal input and in order to investigate the noise performance of the detector and its associated receiver.
3. The support configuration—included in this area are such factors as the electrical bias required to make the detector operative, cooling requirements, and, in the case of heterodyne systems, the local oscillator requirement.

The electrical configuration has been described in some detail in the previous topics. It is the purpose of this topic to present general relationships between the bit rate (bits per second) detected by the detector and the weight, cost, and power required by the entire demodulator assembly.

These relationships are given in Figures A, B, and C. As expected the more complicated detection system, heterodyne detection, requires the most weight and power and is the highest cost system.

The curves of Figures A, B, and C provide the demodulator representation needed by the Optimization Methodology described in Volume II, Part 1. The values given in the Figures are for representative sample configurations, detailed designs will undoubtedly modify these values.

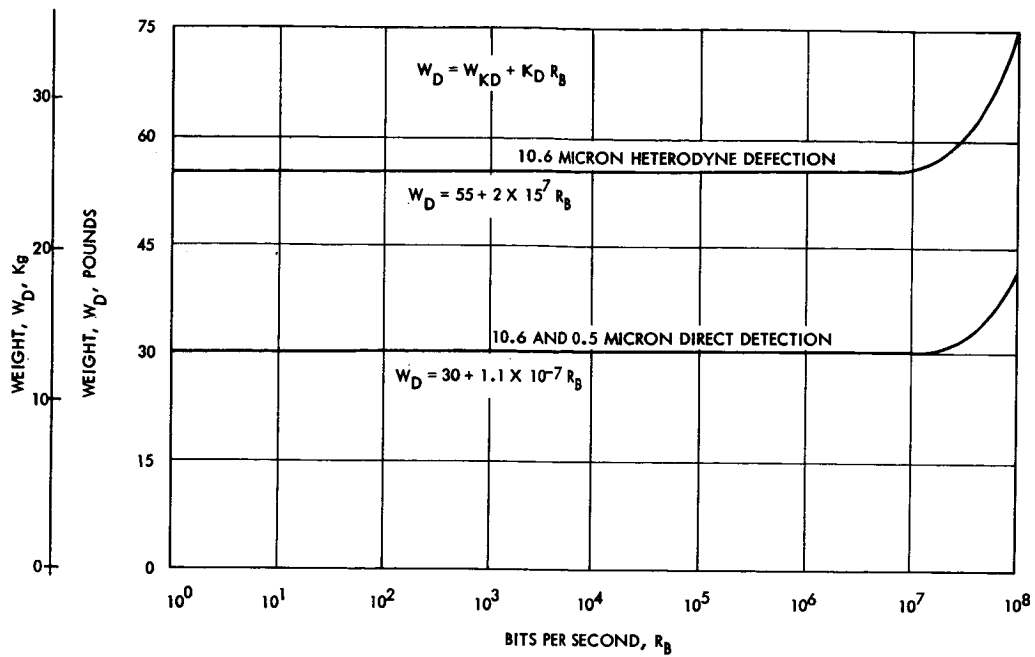


Figure A. Demodulator Weight

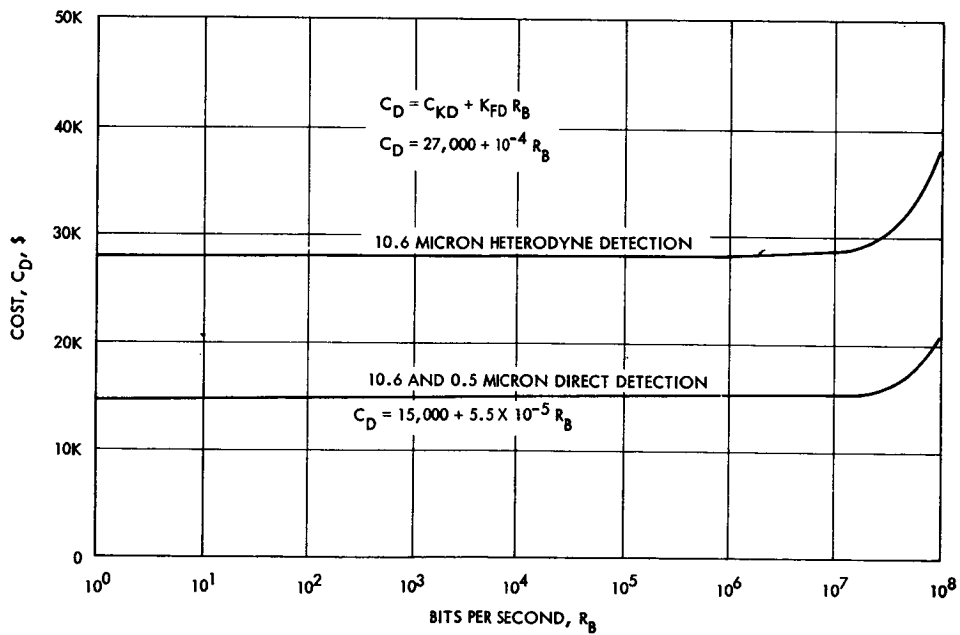


Figure B. Demodulator Cost

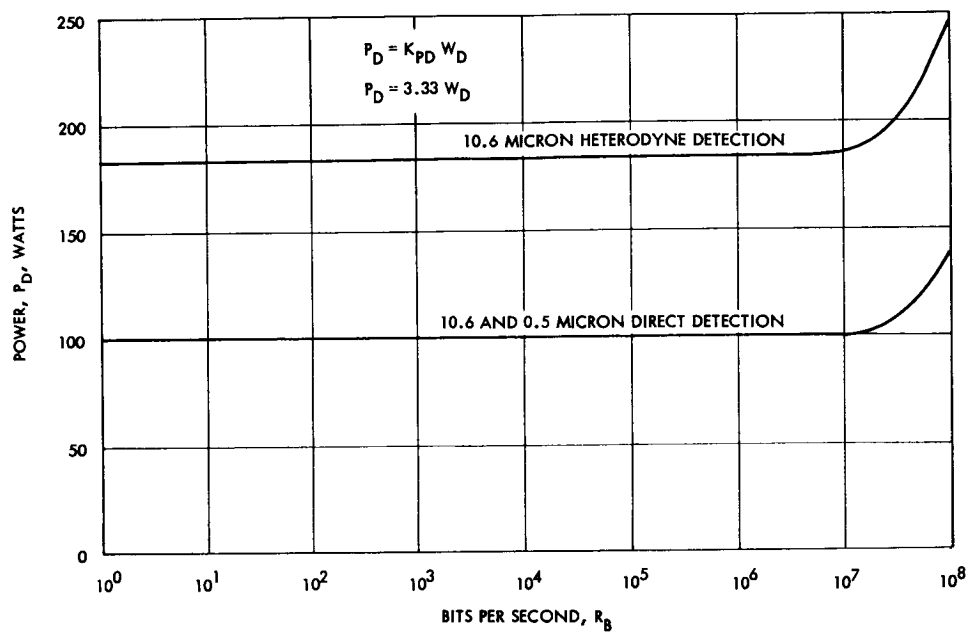


Figure C. Demodulator Power

PART 4 – TRANSMITTING AND RECEIVING APERTURES

Section	Page
Radio Frequency Antennas	234
Optical Frequency Apertures	282
Optical Apertures – Optical Configurations	296
Optical Frequency Apertures – Weight and Cost Relationships	318

INTRODUCTION

Antenna theory and state-of-the art are related to weight and cost of the antennas in the subsequent sections.

This section presents the theory and state of the art of radio frequency antennas potentially capable of fulfilling the characteristics dictated by space communication and tracking applications.

Antenna Theory for Large Apertures is discussed first. The fundamental parameters for characterizing antenna performance are defined. In addition, an analysis relating the degradation of performance to phase errors caused by both bending and random distortion of large antennas is given.

The Antennas for Space Communication Section describes antenna types potentially suitable for space communication and tracking applications. Also indicated is the performance and relative advantages of the various antenna types for different space applications. Curves showing degradation of antenna gain as a function of aperture distortion for various types of antennas are given.

The radio frequency antenna discussion is concluded by giving antenna burden relationships. Included are cost and weight relationships.

SUMMARY OF RF TRANSMITTING AND RECEIVING APERTURES

Phased array and parabolic dish transmitting and receiving antennas are the most useful antenna types for deep space missions.

High gain antennas may be designed as arrays of low or moderate gain elements, or as large area reflector surfaces illuminated by moderate gain feed elements. Array elements, for the operating frequencies of interest for space communication, are in general heavier than a reflector antenna of the same gain. This is attributable to need for relatively complex radiating element structures as contrasted to a simple reflector surface, and to the requirement for a complex feed system for the array antenna as contrasted to free space for the reflector antenna. Conversely array antennas have an inherent capability for forming multiple beams which may be switched rapidly from one beam to another with no moving parts. These characteristics overpower weight considerations and lead to their selection for specialized communication missions. The reflector antenna is, however, considered suitable for specific area coverage broadcast satellites and for wide bandwidth data links for planetary and deep space probes.

Two types of directive antennas are most commonly used for space communications. These are the parabolic dish and the planar array antenna.

Reflector-Type Antenna. This antenna has two basic components: a relatively large reflecting surface (most often paraboloidal) and a feed structure. When maximum antenna gain is required, as for space communication and tracking, the reflector size is chosen to be as large as practical and the feed is normally designed to illuminate the reflector with an intensity at the reflector edges that is approximately 10 db below that at the center.

The efficiencies of reflector-type antennas with a front-mounted feed are typically 55 percent, with 65 percent being the upper realizable bound. The efficiencies of cassegrainian type antennas are typically 60 percent with 70 percent being the upper realizable bound.

The figure illustrates the performance of several large ground based antennas as a function of wavelength, as documented by Ruze.¹

In addition to earth antennas, parabolic antennas have been used on the Pioneer spacecraft and the Mariner spacecraft.

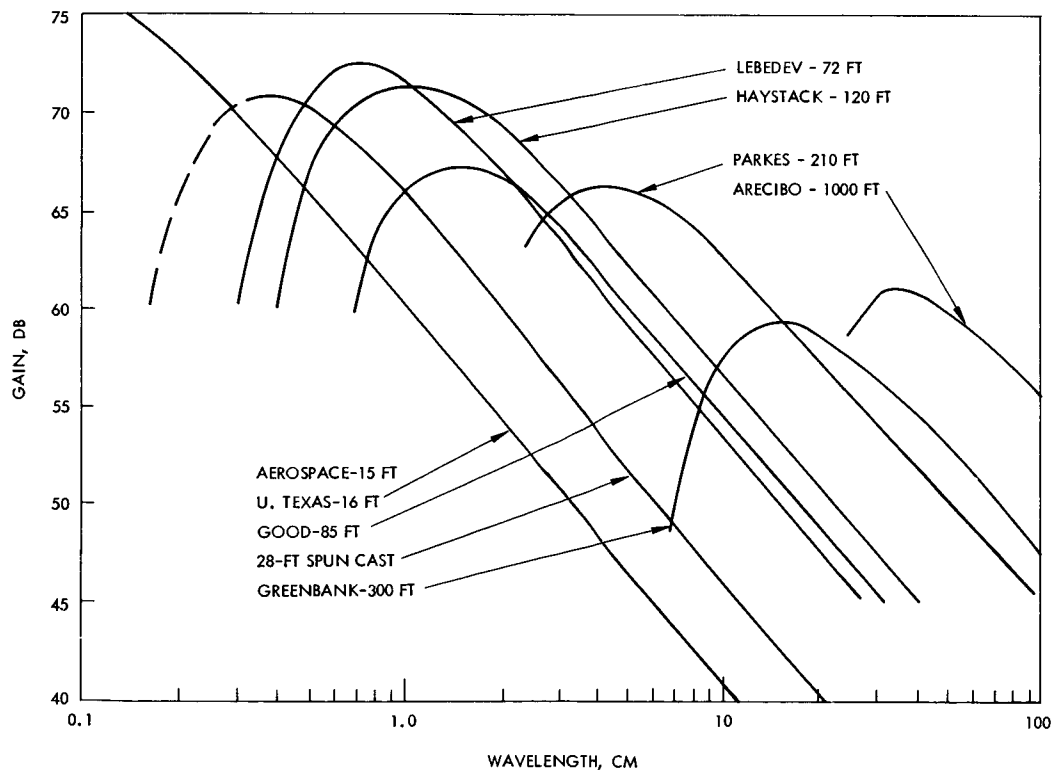
Phased Array Antenna. This type of antenna consists of an array of radiating elements with either fixed or variable relative phase differences. Those with fixed relative phase differences are referred to as planar

¹ Ruze, J., "Antenna Tolerance Theory - A Review, " Proc. IEEE, pp. 633-640, April 1966.

arrays and require mechanical pointing. Those with variable relative phase differences require external electronic controls to properly phase the elements to form a beam in a desired direction. When maximum antenna gain is required, as for space communication and tracking, all the elements of the array are excited equally and the relative phase between elements is adjusted for a beam normal to the plane of the array.

The weight, complexity, and cost of the variable phase shifters needed for those antennas with variable phase differences have deterred space applications of electronically scanned phased arrays. Planar arrays have been used on the ground and even in space when the type of space vehicle stabilization permitted mechanical beam steering.

A planar array antenna was used on the Surveyor spacecraft. This antenna measured 38 x 38 x 2 inches and had a gain of 27 db at 2300 MHz. Planar arrays have advantages of higher efficiency and lower volume than parabolic antennas of equivalent gain. Their chief disadvantage is higher cost.



Gain of Large Paraboloids

TRANSMITTING AND RECEIVING APERTURES
RADIO FREQUENCY ANTENNAS

Antenna Theory for Large Antennas Apertures

	Page
Antenna Gain and Aperture Relationships	240
Gain Degradation Due to Predictable Systematic (Non-Random) Phase Errors	244
Gain Degradation Due to Random Errors	248
Effects of Random Errors on Antenna Parameter Values	252

ANTENNA GAIN AND APERTURE RELATIONSHIPS

Relationships of antenna gain, aperture, and aperture energy distribution are given.

A microwave antenna is a transducer between electromagnetic waves contained by a transmission line and those radiated through space. The spatial distribution of the radiated electromagnetic waves in terms of intensity of energy flow is called the antenna pattern. More precisely, the antenna pattern in a specified plane is a plot of the power radiated per unit solid angle versus a space coordinate, which is usually an angle. Antennas in general are reciprocal and therefore the spatial antenna patterns are independent of whether the antenna is radiating into or receiving from space.

For a fixed frequency; beamwidth, gain, effective area, and sidelobe level are the parameters which are most important in characterizing the performance of antennas. The beamwidth is defined as the angular width at the half power or 3 db points of the main beam of the antenna pattern. The half power points on the main beam are those points where the power received or transmitted is one-half of the value at the beam peak. The portions of the antenna pattern, other than the main beam, are called sidelobes. The sidelobe level, usually quoted in decibels, is the ratio of the maximum power density in the largest sidelobe (usually the sidelobe adjacent to the main beam) to the power density in the main beam maximum. Sidelobes play an important part of determining antenna system noise temperatures, especially for high sensitivity earth antennas.

The antenna gain, G , expresses the ability of an antenna to concentrate the radiated power in a particular direction or, conversely, to absorb power incident from a particular direction more effectively than from other directions. G is defined as the ratio of power per unit solid angle in the direction of the peak of the main beam to the average radiated per unit solid angle. Let P_M be the power radiated per unit solid angle in the direction of the peak of the main beam and let P_t be the total power radiated when the antenna is perfectly matched.

Then

$$G = \frac{P_M}{P_t/4\pi}$$

Another measure of receiving antenna performance is effective area, or receiving cross section. Effective area is the area of a perfect antenna which absorbs the same amount of power from an incident plane wave as the actual antenna. Gain and effective area A_E are simply related as

$$\frac{G}{A_E} = \frac{4\pi}{\lambda^2}$$

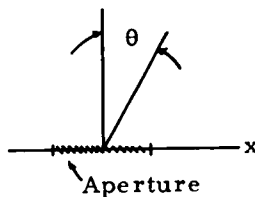
Let G_T and G_R be the respective antenna gains of a transmitting antenna and a receiving antenna separated by a distance R . If the total power transmitted is P_T , the power radiated in the direction of the main beam aimed at the receiver, per unit solid angle will be $P_T G_T / 4\pi$. The receiving antenna will present a receiving cross section $G_R \lambda^2 / 4\pi$ to the incident wave; it will, in effect, subtend a solid angle $G_R \lambda^2 / 4\pi R^2$ at the transmitter. The power absorbed at the receiver will thus be

$$P_R = P_T \frac{G_T G_R \lambda^2}{16\pi^2 R^2}$$

Here it is seen that the power received in space communication is directly related to the power transmitted as well as to the gain of both the receiving and transmitting antennas. Hence, for efficient transmission the gain of both the relatively small spacecraft antenna and the gain of the large ground station antenna have to be maximized within practical limitations. The most evident practical limitation on the gain of spacecraft antennas is the size allowed by the shroud of the launch vehicles. However, inflatable or unfurlable spacecraft antennas are exceptions to this size limitation. The gain associated with all antennas is dependent on the ability to maintain the desired amplitude and phase of the antenna's aperture illumination.

The spatial antenna pattern is $|g(y)|^2$, where $y = \sin \theta$, and is related to the aperture distribution $f(x)$ by:

$$g(u) = \int_a f(x) e^{jkyx} dx, \quad K = \frac{2\pi}{\lambda}$$



where a is the aperture. If the aperture consists of discrete sources, the above integration degenerates into a summation. Some typical, but common aperture distributions and their associated antenna patterns are given in Table 1¹. The following general observations may be made:

1. A uniform amplitude distribution provides highest gain (constant phase cases).

¹ Silver, S., Microwave Antenna Theory and Design, 12, Rad. Lab Series, McGraw Hill.

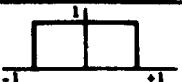



ANTENNA GAIN AND APERTURE RELATIONSHIPS

2. Tapering the amplitude from a high value near the center to a low value near the edges reduces the sidelobe level, increases the beamwidth, and decreases the gain from the uniformly illuminated case.
3. Tapering the amplitude from a low value near the center to a high value near the edges reduces the beamwidth, increases the sidelobe level, and decreases the gain from the uniformly illuminated case.

Relations among the various antenna parameters for many popular, specific cases are given in Figures A² and B².

²ITE Antenna Handbook.

Table 1. Secondary Pattern Characteristics Produced by Various Types of Aperture Distributions

 $f(x) = 1 \quad x \leq 1$ $= 0 \quad x > 1$ $g(u) = \frac{\sin u}{u}$				
	Gain factor \mathcal{A}_n ‡	Full width at half power Θ , radians	Angular position θ of first zero	Intensity of first side lobe: db below peak intensity
	1	$0.88 \frac{\lambda}{a}$	$\frac{\lambda}{a}$	13.2
 $f(x) = 1 - (1 - \Delta)x^2, x < 1$ $= 0 \quad x \geq 1$ $g(u) = a \left[\frac{\sin u}{u} + (1 - \Delta) \frac{d^2}{du^2} \left(\frac{\sin u}{u} \right) \right]$ $\mathcal{A}(\Delta) = \frac{(2 + \Delta)^2}{9 \left[1 - \frac{2}{3}(1 - \Delta) + \frac{1}{3}(1 - \Delta)^2 \right]}$				
$\Delta = 1.0$	1	$0.88 \frac{\lambda}{a}$	$\frac{\lambda}{a}$	13.2
0.8	0.994	$0.92 \frac{\lambda}{a}$	$1.06 \frac{\lambda}{a}$	15.8
0.5	0.970	$0.97 \frac{\lambda}{a}$	$1.14 \frac{\lambda}{a}$	17.1
0.0	0.833	$1.15 \frac{\lambda}{a}$	$1.43 \frac{\lambda}{a}$	20.6
 $f(x) = \cos^n \frac{\pi x}{2}, x < 1$ $= 0 \quad x \geq 1$ $g(u) = \frac{2a}{\pi} \frac{n! \cos u}{\sum_{k=0}^{\frac{n-1}{2}} \left[(2k+1)^2 - \frac{4u^2}{\pi^2} \right]}$, n , odd; $g(u) = a \frac{n!}{\sum_{k=1}^{\frac{n}{2}} \left[(2k)^2 - \frac{4u^2}{\pi^2} \right]}$, n , even $\mathcal{A}_n = \frac{4}{\pi^2} \left[\frac{2 \cdot 4 \cdot 6 \cdots (n-1)}{1 \cdot 3 \cdot 5 \cdots n} \right]^2$ $\left(\frac{2 \cdot 4 \cdot 6 \cdots 2n}{1 \cdot 3 \cdot 5 \cdots 2n-1} \right), n$, odd $\mathcal{A}_n = \left[\frac{1 \cdot 3 \cdot 5 \cdots (n-1)}{2 \cdot 4 \cdot 6 \cdots n} \right]$ $\left[\frac{(n+2)(n+4) \cdots 2n}{(n+1)(n+3) \cdots 2n-1} \right], n$, even				
$n = 0$	1	$0.88 \frac{\lambda}{a}$	$\frac{\lambda}{a}$	13.2
1	0.810	$1.2 \frac{\lambda}{a}$	$1.5 \frac{\lambda}{a}$	23
2	0.667	$1.45 \frac{\lambda}{a}$	$2 \frac{\lambda}{a}$	32
3	0.575	$1.66 \frac{\lambda}{a}$	$2.5 \frac{\lambda}{a}$	40
4	0.515	$1.93 \frac{\lambda}{a}$	$3 \frac{\lambda}{a}$	48
 $f(x) = 1 - x , x < 1$ $= 0 \quad x \geq 1$ $g(u) = 4a \left(\frac{\sin \frac{u}{2}}{\frac{u}{2}} \right)^2$				
	0.75	$1.28 \frac{\lambda}{a}$	$2 \frac{\lambda}{a}$	26.4
<p>* $u = \frac{\pi a}{\lambda} \sin \theta$</p> <p>‡ a = aperture length</p> <p>§ The efficiency of the aperture in concentrating the available energy into the peak intensity of the beam. Values are less than unity.</p>				

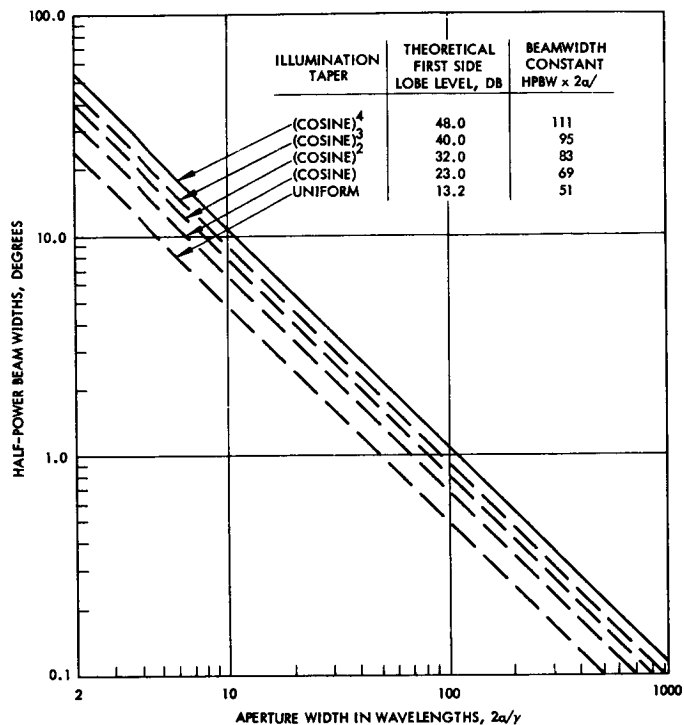


Figure A. Antenna Half-Power Beamwidths Versus Aperture Widths for Rectangular Apertures

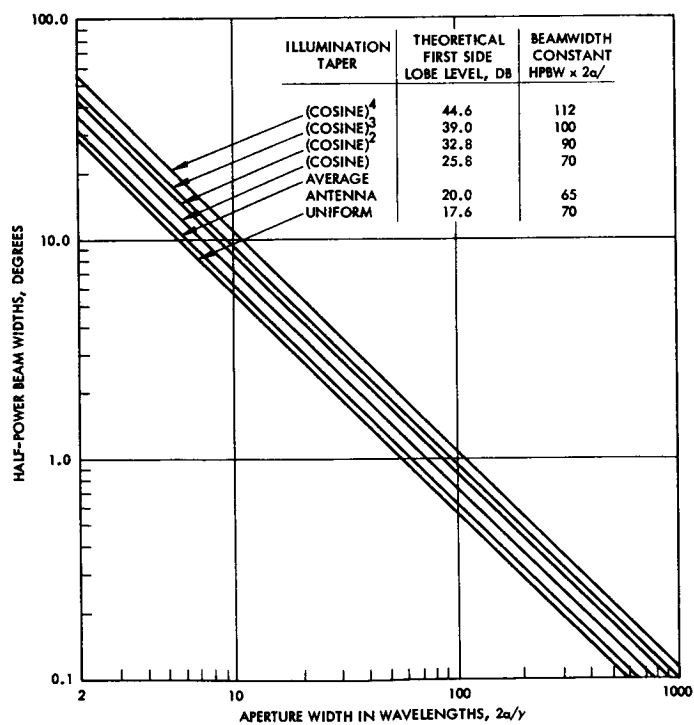


Figure B. Antenna Half-Power Beamwidths Versus Aperture Width for Circular or Elliptical Apertures

GAIN DEGRADATION DUE TO PREDICTABLE SYSTEMATIC (NON RANDOM) PHASE ERRORS

Maximum antenna gain degradation due to "worst case" distortion is calculated and plotted.

Bailin¹ has investigated the effect of array bending on the far field radiation pattern of a large linear array of closely spaced elements. The particular type of bending which was shown to produce the severest gain degradation is the bending of an array supported at one fourth the array length from each end as shown in Figure A. The far field radiation pattern of such a bent array was shown to be proportional to

$$g(W, B) \sim \frac{\sin W \pi / 2}{W \pi / 2} \left[1 - \frac{B^2}{(2^2 - W^2)} \left(1 - \frac{W^2}{2!} \right) + \frac{B^4}{(2^2 - W^2)(4^2 - W^2)} \left(1 - \frac{W^2}{2!} + \frac{W^2(2^2 - W^2)}{4!} \right) + \dots \right] - j \frac{2W}{\pi} \cos \frac{W \pi}{2} \left[\frac{B}{1 - W^2} - \frac{B^3}{(1 - W^2)(3^2 - W^2)} \left(1 - \frac{1 - W^2}{3!} \right) + \frac{B^5 + (1 - W^2)/3! + (1 - W^2)(3^2 - W^2)/5!}{(1 - W^2)(3^2 - W^2)(5^2 - W^2)} + \dots \right] \quad (1)$$

where

$$W = \frac{kc}{\pi} \sin \theta \quad (= 0 \text{ at } \theta = 0^\circ)$$

and

$$B = kB_0 \cos \theta = \frac{2\pi}{\lambda} \frac{\lambda}{N} \cos \theta \quad (= \frac{2\pi}{N} \text{ at } \theta = 0^\circ)$$

B_0 = the maximum amplitude of the phase distortion measured in radians as a fraction of a wavelength, λ/N

$$K = \frac{2\pi}{\lambda}$$

N is an integer

¹Bailin, L. L., "Fundamental Limitations of Long Arrays," Hughes Aircraft Company, Technical Memorandum TM330, October 1953.

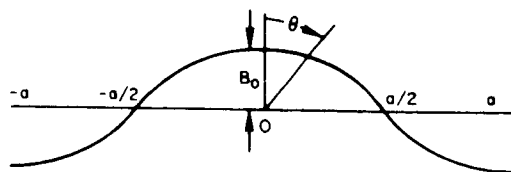


Figure A. Bending of an Array
Supported at One Fourth the
Array Length From Each
End

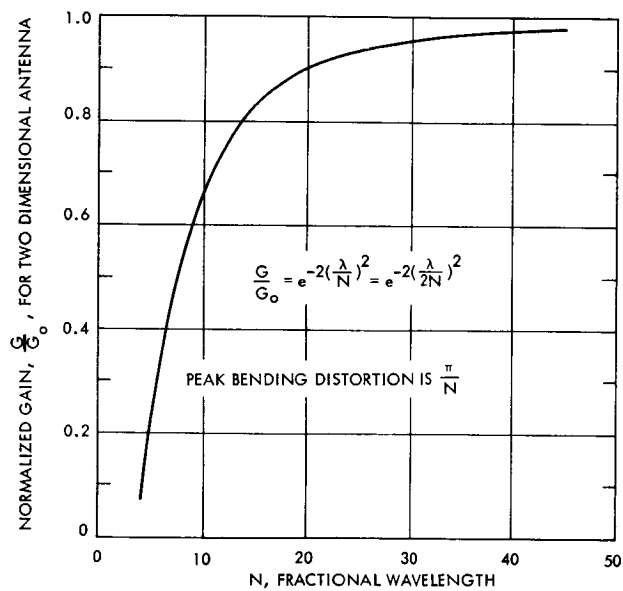


Figure B. Gain Loss in a Two Dimensional
Antenna Due to Non-Random Bending

GAIN DEGRADATION DUE TO PREDICTABLE SYSTEMATIC (NON RANDOM) PHASE ERRORS

In order to simplify the derivation, the gain degradation in this non random phase error portion of the analysis will be calculated as the reduction of P_M , the power radiated per solid angle in the direction of the peak of the main beam ($\theta = 0^\circ$). This equality is only approximate since P_t , the total power radiated, can also vary as a function of phase distortion. With this approximation the gain degradation of the linear array can readily be obtained by the square of the far field radiation pattern, Equation (1), evaluated at $\theta = 0^\circ$.

$$\frac{G}{G_{Lo}} = \left[1 - \frac{B^2}{4} \right]^2 \quad (2)$$

$$= \left[1 - \left(\frac{\pi}{N} \right)^2 \right]^2 \quad (3)$$

$$= e^{-2(\pi/N)^2} \quad (4)$$

for small values of (π/n)

where G_L is the gain of a one-dimensional antenna and G_{Lo} is the no-error gain of the same antenna.

Squaring Equation 4 results in the gain degradation of a two-dimensional antenna.

$$\frac{G}{G_o} = e^{-4(\pi/N)^2} \quad (5)$$

where G is the gain of a two-dimensional antenna, and G_o is the no-error gain of a two-dimensional antenna. This relationship is plotted in Figure B. Figure C is a plot of the db loss for two types of bending distortion analyzed by Bailin. Figure D relates this distortion to antenna gain.

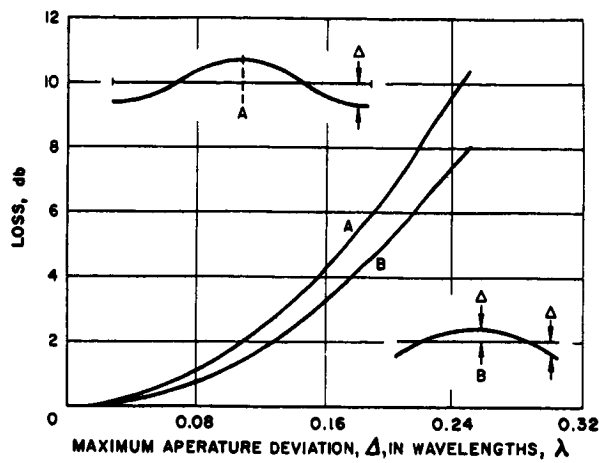


Figure C. Degradation in Gain Due to Sinusoidal Deflection of Paraboloid

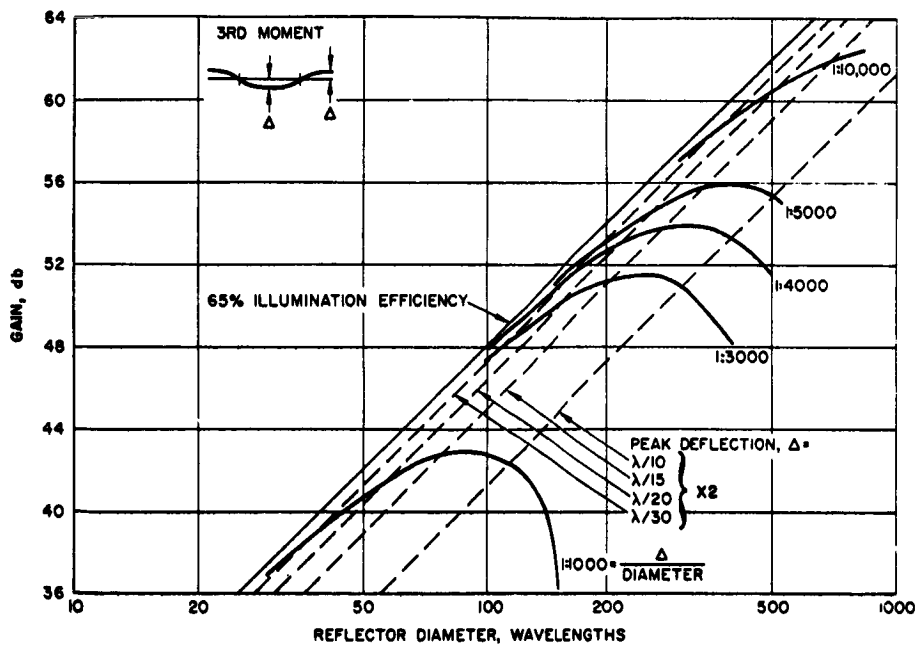


Figure D. Gain versus Reflector Distortion

GAIN DEGRADATION DUE TO RANDOM ERRORS

Random phase errors produce a degradation which are the square root of degradation produced by non-random errors.

Ruze¹ has statistically analyzed the degradation of two-dimensional antenna gain due to random phase errors. The phase errors of concern here are those random phase errors caused by loose machining tolerances and random distortion of the aperture surface. Ruze considered both discrete array and continuous aperture antennas. In general the same statistical considerations apply to the continuous aperture antenna as to the discrete array antenna. In the discrete array case, the error in one array element is independent of the errors in adjacent elements. However, this assumption is untenable in an aperture antenna since if the error is large at one point it will probably be large in its immediate neighborhood. Therefore a "correlation interval," C , is defined as that distance "on average" where the errors become essentially independent.

For small correlation intervals $C/\lambda \ll 1$

$$\frac{G}{G_0} \approx 1 - \frac{3}{4} \overline{\delta^2} \frac{C^2 \pi^2}{\lambda^2} \quad (1)$$

$$\approx \exp \left[-\frac{3}{4} \overline{\delta^2} \frac{C^2 \pi^2}{\lambda^2} \right] \text{ for small values of } \frac{3}{4} \overline{\delta^2} \frac{C^2 \pi^2}{\lambda^2} \quad (2)$$

where $\overline{\delta^2}$ is the mean square phase deviation in radians.

For large correlation intervals

$$\frac{G}{G_0} \approx 1 - \overline{\delta^2} \quad (3)$$

$$\approx e^{-\overline{\delta^2}} \text{ for small values of } \overline{\delta^2} \quad (4)$$

With equations describing the gain degradation of two-dimensional antennas due to random phase errors and another describing the gain degradation due to non random phase errors caused by array bending, * it would be of interest to compare the gain degradation predicted by the

¹Ruze, J., "The Effect of Aperture Errors on the Antenna Radiation Pattern," Supplemento AL Volume IX, Serie IX Del Nuovo Cimento No. 3, 1952.

*See previous topic.

two different analyses for identical rms phase errors. For convenience, the gain degradation is calculated for non random phase errors caused by bending the array supported at one fourth the array length from each end as shown in Figure A of the previous topic. This value of gain degradation is to be compared with that introduced by a random phase error of identical rms value, for the same two-dimensional antenna.

For a $2\pi/N$ value of maximum amplitude of phase distortion, the mean square phase degradation, $\overline{\delta^2}$, in radians can be determined as follows along the path " ℓ " for an aperture size " a " to be

$$\overline{\delta^2} = \frac{1}{a} \int_{-a/2}^{a/2} f^2(\ell) d\ell \quad (5)$$

$$\overline{\delta^2} = \frac{1}{a} \int_{-a/2}^{a/2} \left(\frac{2\pi}{N}\right)^2 \cos^2\left(\frac{2\pi}{a}\ell\right) d\ell \quad (6)$$

$$\overline{\delta^2} = \frac{1}{a} \left(\frac{2\pi}{N}\right)^2 \left[\left(\frac{\cos(2\pi\ell/a)}{4(2\pi/a)^2} \right) \left(2 \frac{2\pi}{a} \sin \frac{2\pi\ell}{a} \right) + \frac{2(2\pi/a)^2}{4(2\pi/a)^2} \ell \right] \Bigg|_{-\frac{a}{2}}^{\frac{a}{2}} \quad (7)$$

$$\overline{\delta^2} = 2 \left(\frac{\pi}{N}\right)^2 \quad (8)$$

Substituting Equation (8) into Equations (3) and (4) yields a gain degradation for random errors of:

$$\frac{G}{G_o} \approx 1 - 2 \left(\frac{\pi}{N}\right)^2 \quad (9)$$

or for small values of π/N

$$\frac{G}{G_o} \approx e^{-2(\pi/N)^2} \quad (10)$$

This is plotted in Figure A.

GAIN DEGRADATION DUE TO RANDOM ERRORS

Random phase error Equation (4) predicts a gain degradation which is the square root of the gain degradation predicted by non random phase error for the same rms phase errors. This result can be explained heuristically as follows. In the non random case, waves from relatively large portions of the aperture are in phase and of different phase than waves from other relatively large portions of the aperture. The two groups of waves destructively interfere with each other much more effectively than groups consisting of waves with random phase even though the mean square phase deviation is the same in both cases.

Antenna gain degradation is due to two distinct types of aperture distortion, random and non random. Curves depicting antenna gain as a function of aperture size in wavelengths for various amounts of random aperture phase distortion are presented in the next topic.

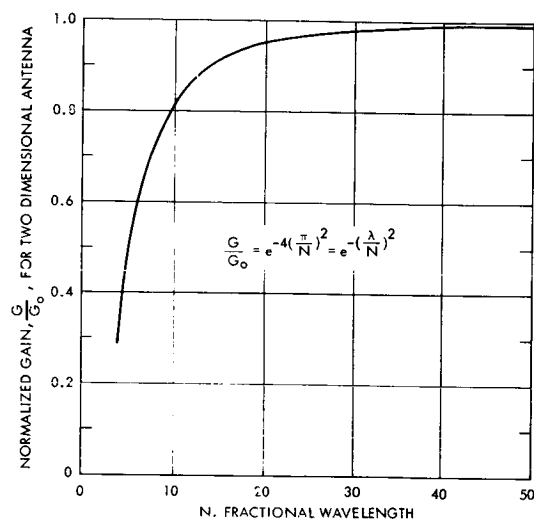


Figure A. Gain Loss in a Two Dimensional Antenna Due to Random Errors

EFFECTS OF RANDOM ERRORS ON ANTENNA PARAMETER VALUES

Curves are presented in this topic which depict antenna gain as a function of aperture size in wavelengths for various amounts of random aperture phase distortion and for various types of antennas.

The efficiency of an antenna is the ratio of its effective area A_E to its actual area A_A .

$$\text{Eff} = \frac{A_E}{A_A} \quad (1)$$

Recalling that

$$\frac{G}{A_E} = \frac{4\pi}{\lambda^2} \quad (2)$$

allows the gain to be written as

$$G = \frac{4\pi \text{Eff} A_A}{\lambda^2} \quad (3)$$

Efficiencies of reflector type antennas range from 55 to 70 percent and efficiencies of planar arrays range from 70 to 85 percent. Efficiencies of horn radiators range from 85 to 95 percent. For convenience it is assumed circular apertures of area $\pi D^2/4$, where D is the diameter.

For a circular aperture antenna with uniform phase and amplitude aperture illumination the gain may be expressed as

$$G = \frac{4\pi \text{Eff}}{\lambda^2} \left(\frac{\pi D^2}{4} \right) = \text{Eff} \left(\frac{\pi D}{\lambda} \right)^2 \quad (4)$$

Figures A, B, and C present the error-free antenna gain and gain with random phase errors as a function of aperture size in wavelengths. The root mean square phase error in each figure is a particular fraction of the electrical diameter in radians. Therefore, the rms phase error on one curve is greater for the large diameter circular planar arrays and parabolic reflectors. The random phase error curves are valid only for small rms values of phase error.

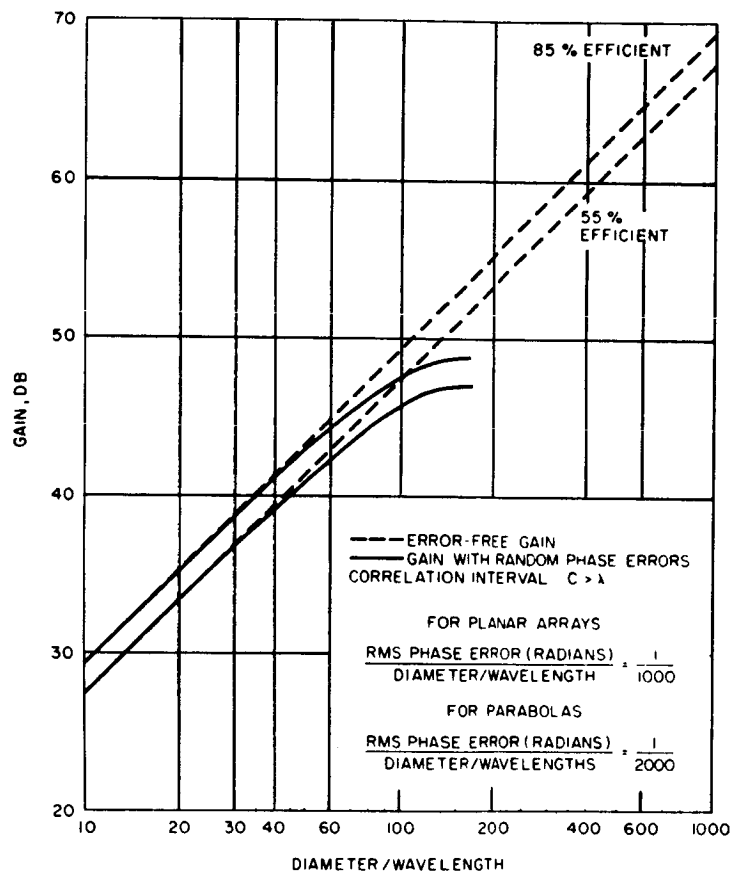


Figure A. Antenna Gain Degradation Due to Random Phase Errors in Aperture Illumination

EFFECTS OF RANDOM ERRORS ON ANTENNA PARAMETER VALUES

Figure D¹ presents the loss of gain due to reflector tolerance as a function of frequency.

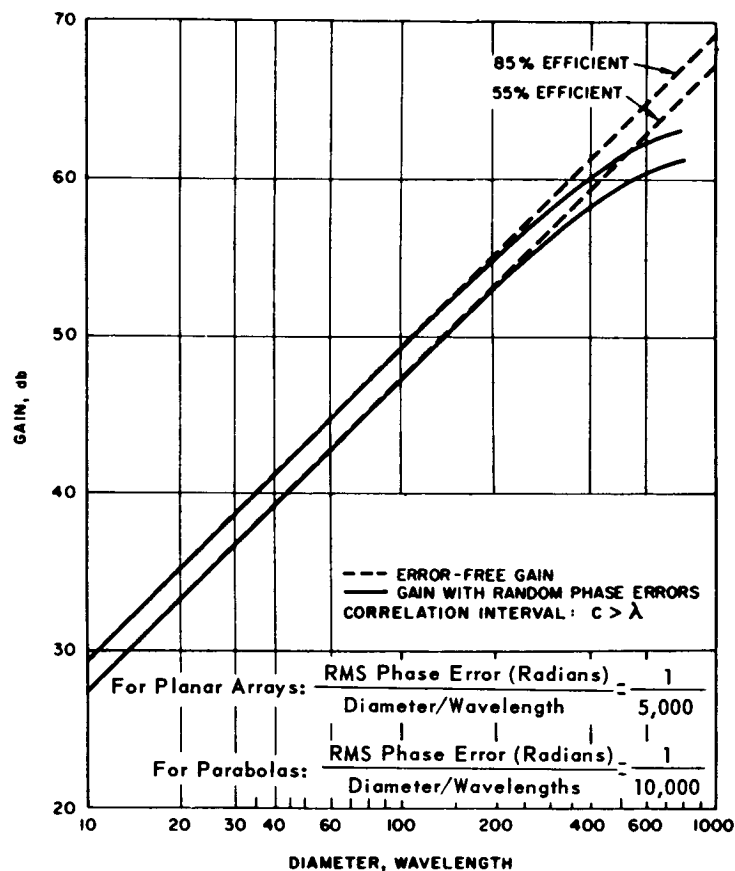


Figure B. Antenna Gain Degradation Due to Random Phase Errors in Aperture Illumination

¹ Ruze, J., "Antenna Tolerance Theory - A Review Proc.", IEEE, pp 633-640, April 1966.

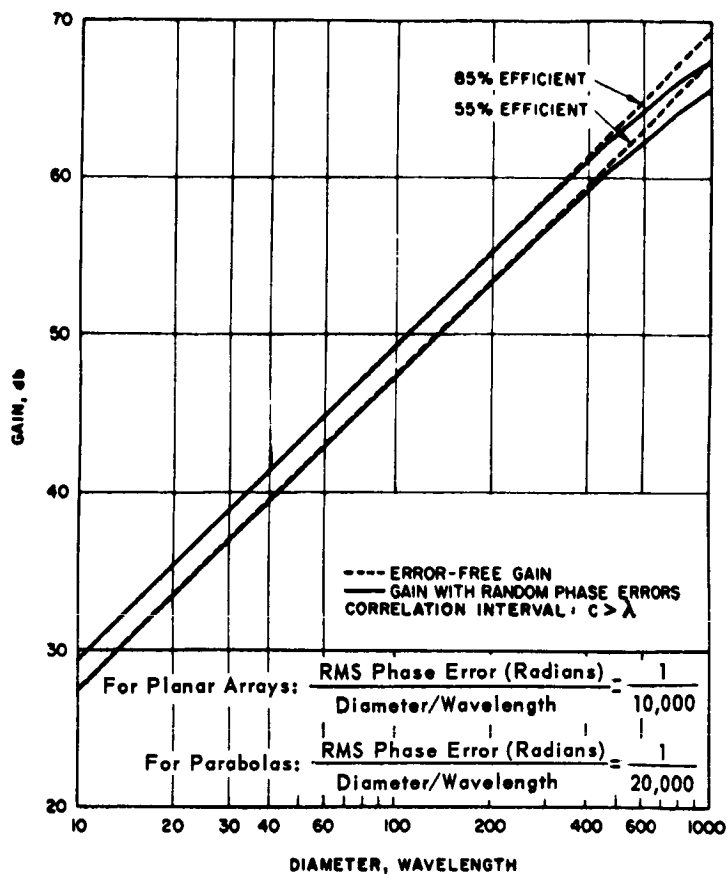
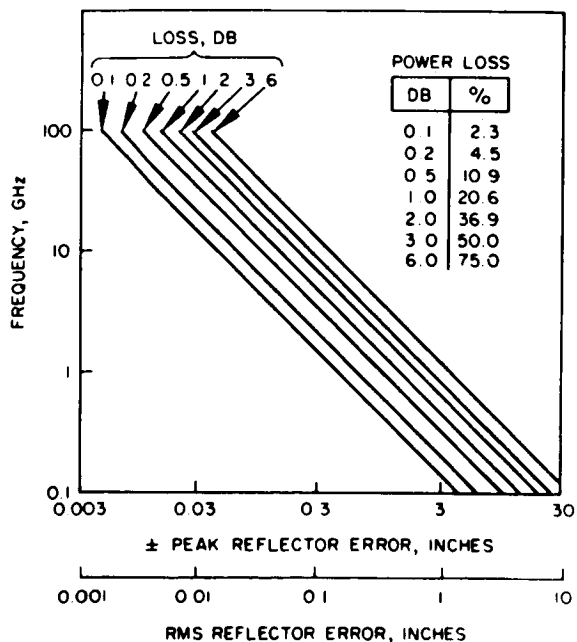


Figure C. Antenna Gain Degradation Due to Random Phase Errors in Aperture Illumination

Figure D. Gain Loss Due to Reflector Tolerance



TRANSMITTING AND RECEIVING APERTURES

Radio Frequency Antennas

	Page
Antennas for Space Communication — Design Considerations	258
Antennas for Space Communication — Antenna Types	260
Antennas for Space Communication — Deployable Paraboloids	262
Antennas for Space Communication — Weight Burdens for Paraboloids	264
Special Purpose Multibeam and Self Steering Antennas	266
High Gain, Self-Steering Antenna System for Satellite-Earth Communications	268
Antennas for Space Communication — Surface Station Cost Burdens	274
Antennas for Space Communication — Spacecraft Cost Burdens	276

ANTENNAS FOR SPACE COMMUNICATION — DESIGN CONSIDERATIONS

High gain antennas are considered for ranges greater than synchronous orbit.

Antenna design is limited by considerations of beamwidth, frequency, launch vehicle payload weight limits, and restrictions on payload package dimensions during boost. It seems reasonable to restrict consideration of high gain spacecraft antenna to satellites in synchronous earth orbits and to deep space and planetary probes. Low altitude satellites require wide angle coverage for effective coverage, and as the range is not too great, moderate transmitter levels are adequate for communication with the earth. Synchronous altitude satellites require an antenna with a beamwidth of 20 degrees to provide for earth coverage and to allow for spacecraft attitude errors. A lower limit of 0.7 degree is suggested from the standpoint of spacecraft attitude control.¹ If the antenna beam is nominally centered on the receiving station, a variation of spacecraft attitude by ± 0.1 degree will result in power level variations of approximately 1 db. Finer control may perhaps be accomplished but only at the expense of added complexity of the control system.

Booster payload weight and size limitations are important design limits for the spacecraft and its subsystems. Allowable weight bogies for spacecraft antenna subsystems cover many orders of magnitude for the currently available and projected launch vehicles. Satellite payloads approximately 9 feet in diameter can be accommodated by currently available launch vehicles without resorting to deployment mechanisms.

¹Patton, W. T. and Glenn, A. B., "Component Problems in a Microwave Deep Space Communication System," Conference Record, WINCON, Los Angeles, Calif., 1966.

ANTENNAS FOR SPACE COMMUNICATION — ANTENNA TYPES

Phased array antennas are compared with parabolic dish antennas. The latter is the subject of subsequent topics.

High gain antennas may be designed as arrays of low or moderate gain elements, or as large area reflector surfaces illuminated by moderate gain feed elements. Array elements, for the operating frequencies of interest for space communication, are in general heavier than a reflector antenna of the same gain. This is attributable to need for relatively complex radiating element structures as contrasted to a simple reflector surface, and to the requirement for a complex feed system for the array antenna as contrasted to free space for the reflector antenna. Array antennas have an inherent capability for forming multiple beams which may be switched rapidly from one beam to another with no moving parts. These characteristics overpower weight considerations and lead to their selection for specialized communication missions. The reflector antenna is, however, considered suitable for specific area coverage broadcast satellites and for wide bandwidth data links for planetary and deep space probes.

Examination of world geography shows that pencil beams are desirable to conserve radiated power by directing it toward the areas of interest. North America, Europe, Africa or the whole earth can be effectively covered by circular pencil beams. Paraboloid antennas can be designed with primary focus feed or with secondary reflector feeds. The primary reflector structural design approach affects weight to a larger degree than does the feed. It is, therefore, proposed to classify antenna designs by means of their structural design approaches.

Two types of directive antennas are most commonly used for space communications. These are the parabolic dish and the planar array antenna. Some performance characteristics for these antennas has been given in prior topics. Further data are given in this topic with brief comments on construction of these two types of antennas.

Reflector-Type Antenna. This antenna has two basic components: a relatively large reflecting surface (most often paraboloidal) and a feed structure. When maximum antenna gain is required, as for space communication and tracking, the reflector size is chosen to be as large as practical and the feed is normally designed to illuminate the reflector with an intensity at the reflector edges that is approximately 10 db below that at the center.

The efficiencies of reflector-type antennas with a front-mounted feed are typically 55 percent, with 65 percent being the upper realizable bound. The efficiencies of cassegrainian type antennas are typically 60 percent with 70 percent being the upper realizable bound.

The figure illustrates the performance of several large ground based antennas as a function of wavelength, as documented by Ruze.¹

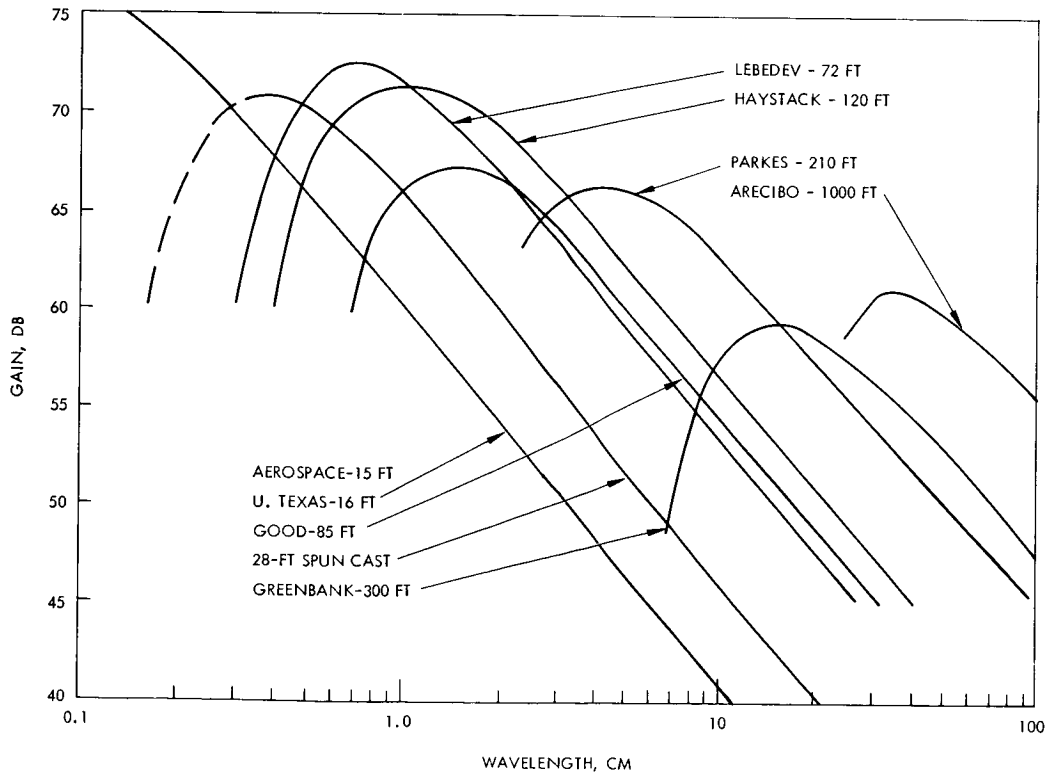
¹Ruze, J., "Antenna Tolerance Theory — A Review," Proc. IEEE, pp. 633-640, April 1966.

In addition to earth antennas, parabolic antennas have been used on the Pioneer spacecraft and the Mariner spacecraft.

Phased Array Antenna. This type of antenna consists of an array of radiating elements with either fixed or variable relative phase differences. Those with fixed relative phase differences are referred to as planar arrays and require mechanical pointing. Those with variable relative phase differences require external electronic controls to properly phase the elements to form a beam in a desired direction. When maximum antenna gain is required, as for space communication and tracking, all the elements of the array are excited equally and the relative phase between elements is adjusted for a beam normal to the plane of the array.

The weight, complexity, and cost of the variable phase shifters needed for those antennas with variable phase differences have deterred space applications of electronically scanned phased arrays. Planar arrays have been used on the ground and even in space when the type of space vehicle stabilization permitted mechanical beam steering.

A planar array antenna was used on the Surveyor spacecraft. This antenna measured 38 x 38 x 2 inches and had a gain of 27 db at 2300 MHz. Planar arrays have advantages of higher efficiency and lower volume than parabolic antennas of equivalent gain. Their chief disadvantage is higher cost.



Gain of Large Ground Based Paraboloids

ANTENNAS FOR SPACE COMMUNICATION — DEPLOYABLE PARABOLOIDS

Three types of extensible (deployed) spacecraft antennas are discussed: umbrella antenna, inflatable antenna, and petaline antenna.

Qualities of importance for a paraboloid reflector surface are r-f reflectivity, construction accuracy, and low weight. These qualities are closely associated with the type of reflector surfaces material used. Reflectivity can be obtained by thin metalized films, knitted metalized fabric meshes, wire meshes, or solid metallic panels. Aluminum mylar sandwich materials are representative of film materials. Knitted metalized nylon reflective meshes have been developed for radar targets. Similar materials using fused quartz appear ideal for long lifetime applications.

Design approaches may be classified according to the deployment technique used. A technique for deploying metalized film or mesh over a framework of radial ribs forms one class of antenna designs. This general class is illustrated by the design sketch of Figure A. For this design, normally straight ribs are deflected by alignment cables into a parabolic shape. The reflector material is supported by the deflected ribs. This general design approach is called the umbrella type design.

Inflatable structures form a second design classification. As the inflatable will escape in time, these structures must be designed to remain in the deployed configuration following loss of the inflatable. This may be accomplished by utilizing rigidizable materials or by deploying a rigid frame initially. Figures B and C show design sketches of this type of antenna structure. This design approach is generally termed the inflatable type design.

Mechanical deployment of rigid panels is a technique modeled after large ground antennas. Deployment can be accomplished by rotating the individual panels into their position in the paraboloid as illustrated by Figure D. This design approach is termed the petaline type design.

The foregoing design approaches are representative of antenna design approaches in terms of weight and accuracy. Deployment techniques and materials are key areas for advanced design and development. Performance and weight comparisons of the three types are presented in the next topic to develop design comparisons and to show general areas of applicability.

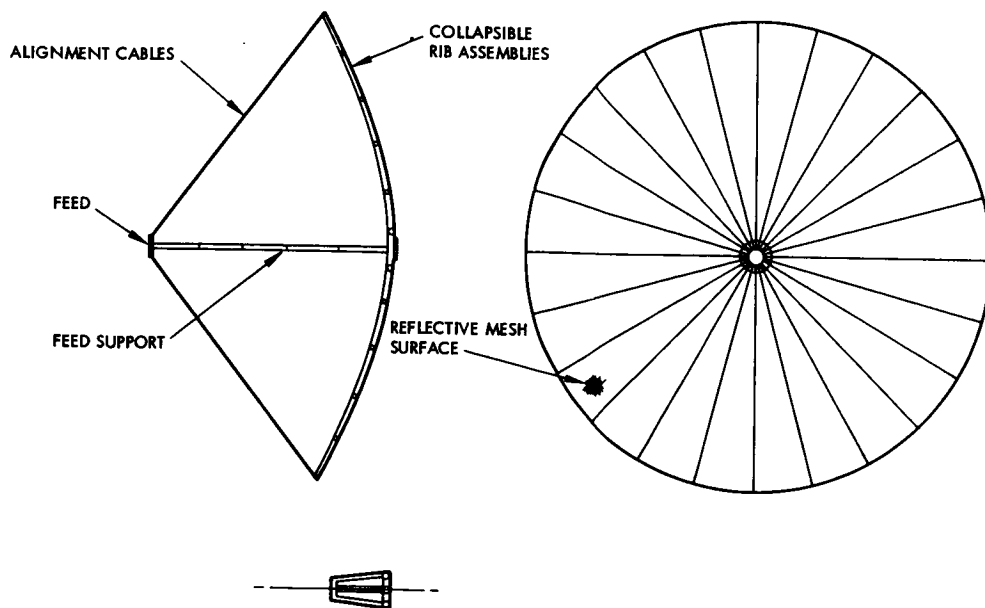


Figure A. Umbrella Antenna Design Concept

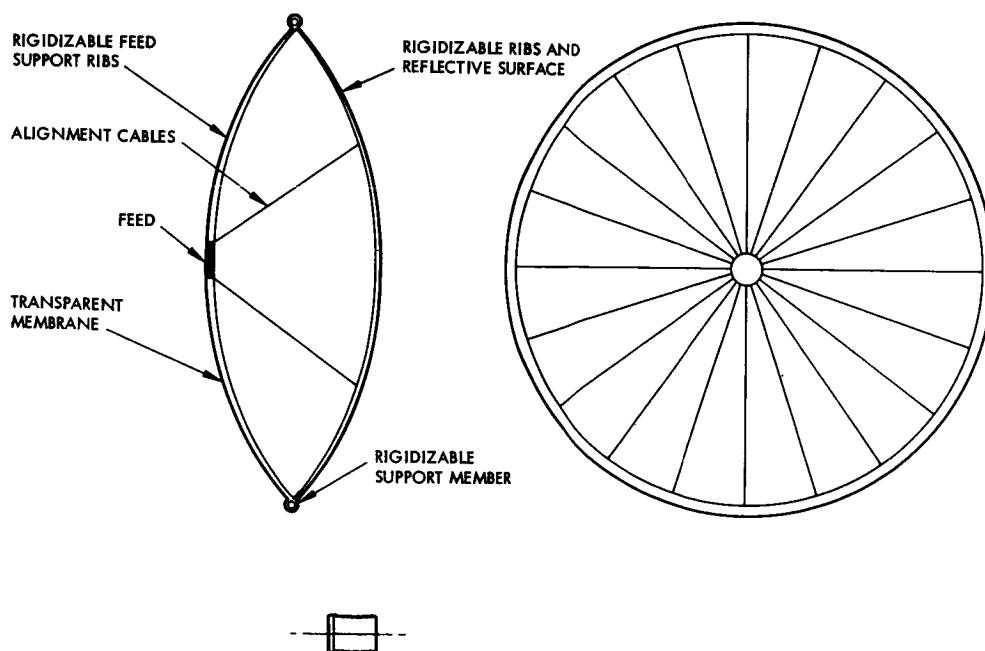


Figure B. Inflatable Antenna Design Concept Rigidizable Ribs and Reflector Surface

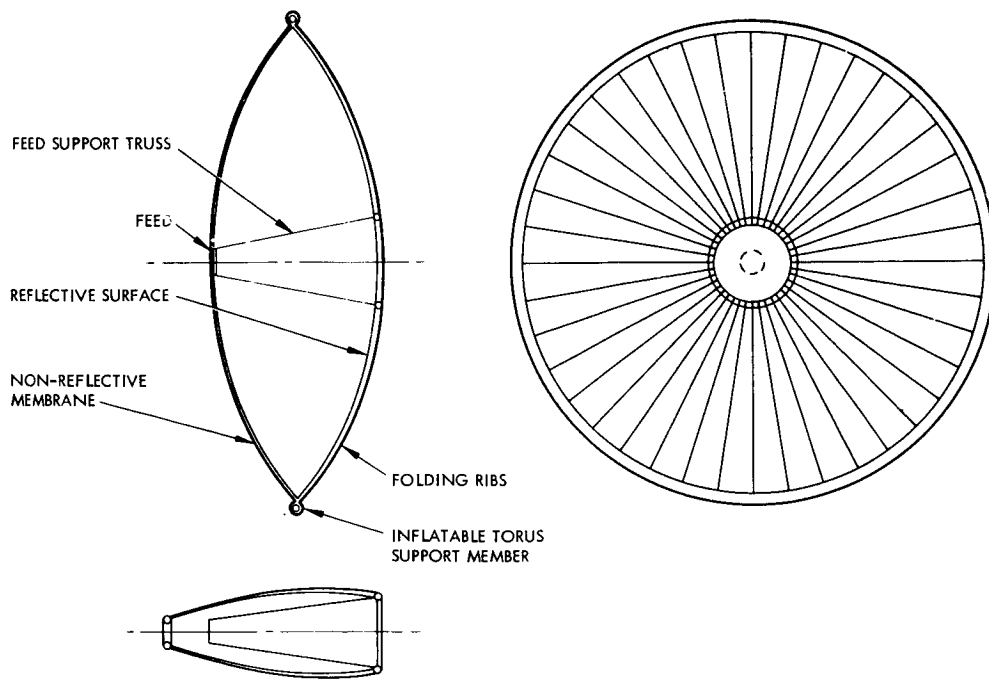


Figure C. Inflatable Antenna Design Concept Rigid Ribs,
Rigidizable Reflector Surface

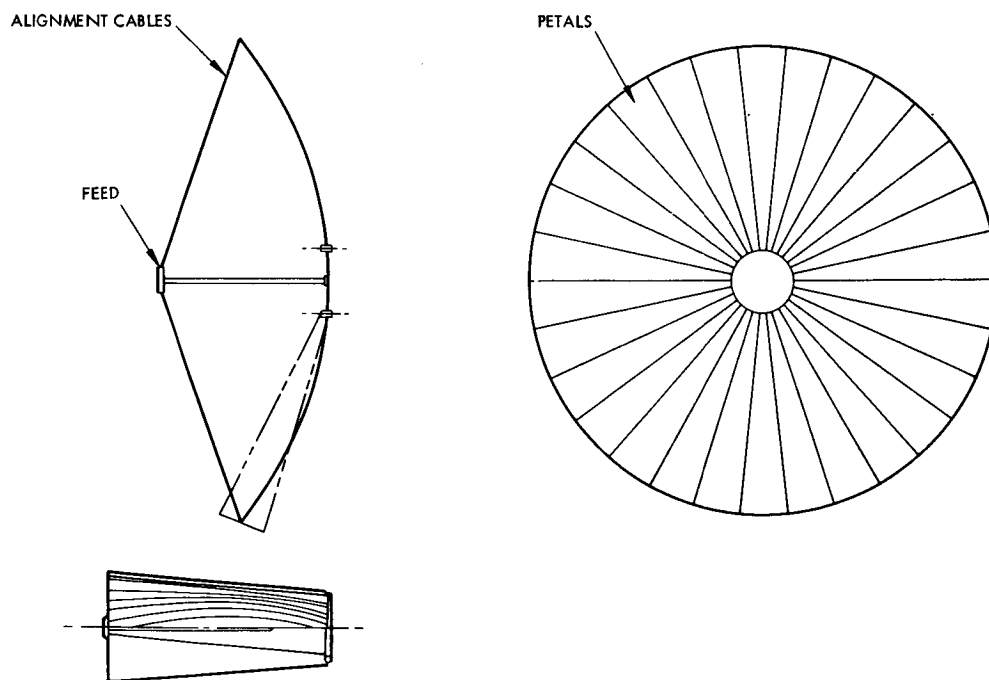


Figure D. Petaline Antenna Design Concept

ANTENNAS FOR SPACE COMMUNICATION – WEIGHT BURDENS FOR PARABOLOIDS

Antenna weight burdens are derived for extensible antennas and small (less than 10 feet equivalent diameter) antennas used as antennas on spacecraft.

The umbrella antenna design utilizes a lightweight mesh reflector spread over a framework of radial ribs. A metalized fabric or wire mesh forms a useful reflector for rf energy for an umbrella type of antenna if the mesh size is a small fraction of a wavelength. Mesh weight is minimized by selecting relatively small mesh size. This is limited by strength required for handling and deployment. Typical meshes of metalized nylon are available with a specific weight of 0.4 oz/yd². The number of ribs in an umbrella antenna influences the degree of conformity of the design to the desired paraboloidal shape.

An inflatable antenna design concept is presented for comparison which consists of a mylar aluminum sandwich material reflector similar to the material used in the ECHO II satellite. This material exhibits a degree of rigidity when stretched beyond its yield point. In order to restrain these loads in a lens shaped design, a torus support structure is required. A pressure stabilized toroidal support member is required for the initial inflation. Rigidizable materials could be used to retain strength following loss of the inflatent.¹

Petaline antenna design is the approach required for high precision and high gain applications. This is at the expense of much higher weights than are required for the umbrella or inflatable design approach. Honeycomb sandwich construction with aluminum or beryllium face sheets are one candidate construction method. Electro-deposited nickel structures are also adaptable to construction of antenna petals.

Weight estimates for antennas of different size is shown in the Figure for the three types of constructions. A 500 pound weight limit for the antenna results in a 46 foot diameter petaline antenna, a 68 foot diameter inflatable antenna, and a 120 foot diameter umbrella antenna.

In terms of the weight burdens of the overall communications system methodology the figure may be reinterpreted as given in Table I.

Since

$$W_{dT} = K_{dT} (d_T)^{n_T} + W_{KT}$$

where

W_{dT} = Antenna weight (measured in pounds)

d_T = Aperture diameter (measured in cm)

¹Keller, L. B. and Schwartz, S., "Rigidization Techniques for Integrally Woven Composite Constructions," Air Force Material Laboratory Report, ML-TDR-64-299, September 1964.

K_{dT} = Constant rotating antenna weight to apertures diameter

W_{KT} = Antenna weight independent of aperture diameter (measured in pounds)

n_T = A constant

and K_{dT} , W_{KT} , and n_T are the weight burdens given by the values of Table I for large antennas, 10 to 100 feet in diameter.

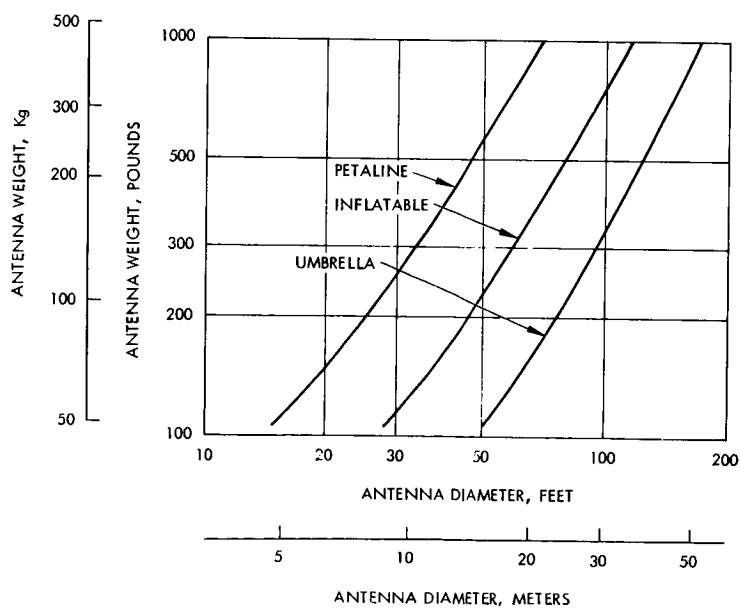
Table I. Weight Burdens

Antenna Type	K_{dT}	W_{KT}	n_T
Umbrella	0.012	42	2.2
Inflatable	0.03	105	2.2
Petaline	0.075	263	2.2

For smaller antennas, 0.5 to 10 feet in diameter, the values of Table II apply.

Table II. Weight Burdens for Small Antennas

	K_{dT}	W_{KT}	n_T
Non-extensible antennas	4.32×10^{-4}	0	2.0



Weight Comparison for Candidate
Spacecraft Antenna Designs

SPECIAL PURPOSE MULTIBEAM AND SELF STEERING ANTENNAS

Several multibeam and self steering spacecraft antennas have been developed as breadboard hardware. These antennas hold special potential for near earth relay satellites rather than for high efficiency deep space antennas.

The multibeam and self steering antennas discussed briefly in this section achieve antenna gain in the desired direction with electronic means as opposed to mechanical means. These antennas fall into two generic groups: those that require external controls to properly phase the elements and those that are self steering. The externally controlled systems, such as the conventional phased array, need (1) an external sensor (i-f, r-f, or ground station), (2) a computer, (3) a phasing network, and (4) an attitude sensing device to point the beam appropriately. In the self-steering system, however, attitude information is determined by the antenna system using a pilot beam from a ground station, and internal electronic circuitry which senses the phase of incoming pilot signals to position a beam in that direction.

Three general groups of self-steerable antennas are considered potentially capable of fulfilling the characteristics dictated by the space mission. These are briefly reviewed in the following paragraphs.

Switched Multiple-Beam Antennas. This group uses multiple-beam antennas with appropriate switching and control circuitry to select the proper beam with on command from the transmitter station or as indicated by a pilot signal from the receiver station. Several configurations are possible, including multiple feed lenses, reflectors, and beam-forming matrices.

The Transdirective Array, a special configuration of the multiple-beam array. This array utilizes a hybrid beam-forming matrix and array elements to form a high-gain antenna that receives incident signals from arbitrary directions, processes them to have arbitrary amplification and frequency before reradiating the signals from the same (or another) matrix toward arbitrary, desired directions. The system selects for reception, signals from stations which identify themselves with unique pilot tones and reradiates these signals toward other stations. The re-directing of the transmitter beam is accomplished either through the use of pilot tones transmitted from the desired station or by a command signal. The participating stations all utilize the high gain and directivity of the Transdirective Array as they simultaneously communicate with one another through the spacecraft.

Self-Phasing Arrays. This group of antennas can be composed of conformal arrays of elements. Each element has its own electronic circuitry that automatically phases the elements to produce a beam in the direction dictated by a pilot signal. (The pilot signal is sent from a station desiring to communicate with a transmitting station.)

In the simplest form of the self-phasing (retrodirective array) phase inversion is needed. This is derived from a mixing action. A c-w signal received by the nth element of an array is subtracted from (mixed with) a common local oscillator to obtain a transmit signal close to the received signal except that it has an inverted phase angle. The phase angle is similarly inverted at every element to create the condition by which

the transmit signal is formed into a beam essentially in the same direction as the received signal. A circulator or branching filter is used to separate the transmit and receive signals at the antenna elements. A slight frequency shift is used to improve isolation between transmitted and received signals and to make provision for suitable amplification.

In a more advanced form of this idea, the total incoming signal at the n th element consists of a narrow band pilot signal (ω_{p1}), offset from a broad information band (ω_m). The signals are subtracted from a suitable offset frequency (ω_1) and the difference frequencies are filtered, separately amplified, and mixed once more. The signal at the difference frequency ($\omega_m - \omega_{p1}$) is independent of the incoming interelement phase shift and is summed with all similar outputs from the other elements. At this point, the total array gain is realized.

A pilot (ω_{p2}) transmitted from the receiving ground station is also processed to form a beam toward that station. The information from the ground transmitting station, which was summed and amplified is now frequency translated and added to the pilot signal (ω_{p2}) to send the information toward the receiving station. Thus a complete communication channel is established electronically.

Adaptive Arrays. This class of arrays is closely related to self phasing arrays but uses phase-locked loops at each element to accomplish the appropriate phasing across the antenna aperture. The term "adaptive" comes from the adaptive properties of the phase locked loops which adjust the phase out of each element to a standard reference. This allows the signals to be added in phase and to realize the maximum gain of the array regardless of the direction of the incoming signal. Adaptive arrays can be made to transmit retrodirectively by appropriate additions to the circuitry at each element.

HIGH GAIN, SELF-STEERING ANTENNA SYSTEM FOR SATELLITE-EARTH COMMUNICATIONS

A description of a self-steering spacecraft antenna system is given. Antenna gain of 29.8 db may be pointed electronically at any point within a 30 degree cone.

An engineering model of a self-phasing antenna system for satellite-earth communications has been designed and fabricated at the Hughes Aircraft Company (NAS 5-10101). This system, shown in block diagram in Figure A, incorporates two channels, each with a 125-MHz RF bandwidth (for the sake of clarity, only one channel is illustrated completely in the figure). The system is designed to receive at 8 GHz and transmit at 7.3 GHz. The design is based on application to a gravity-gradient oriented and stabilized satellite in synchronous orbit with a conical coverage angle of ± 15 degrees. This coverage allows for uncertainties in the attitude of the spacecraft. The transmitting and receiving portions will steer appropriate beams along arbitrary directions within that cone. Two independent channels will be included which provides four independent steered beams. The beam designations and the frequency bands utilized are shown in Figure B.

This system is intended to serve as a communication link to relay information transmitted from one station to another station via high-gain beams. The positions of these beams are controlled by the phase information obtained from CW pilot signals which are generated by the ground stations which communicate to one another through the relay satellite.

For the channel shown in Figure A, a receiving pilot, a transmitting pilot, and a modulated signal are received by the receiving element, passed through a high-pass filter, down-converted to an intermediate frequency, and amplified by a wide-band IF preamplifier. After pre-amplification, the information signal and the pilots are separated by means of a triplexer filter. The pilots are then down-converted to a second, lower, IF to allow utilization of very narrow-band-pass filters to establish a good signal-to-noise ratio for the pilots. These band-pass filters comprise the quadruplexer which, in addition to limiting the noise bandwidth of the pilot channels, serve to separate the pilot signals. After passing through the quadruplexer, the pilot signals are up-converted to about 200 MHz to enable these pilots to be mixed with the wide-band modulated signals without overlap of the power spectra.

With reference to Figure A, the modulated signal, 450 to 575 MHz, passes from the triplexer to a wide-band mixer, and the receiving pilot signal, 206 MHz, also passes to this mixer. The modulated signal is denoted by $\cos \{[\omega_c + f(t)]t - \phi_i\}$ and the pilot signal by $\cos [\omega_p t - \phi_i + \beta]$, where ω_c is the carrier frequency, ω_p is the pilot frequency, $f(t)$ is a modulating signal, ϕ_i is the phase angle of the received signal relative to an arbitrary reference for the i th element, and β is the phase shift of the pilot signal relative to the modulated signal, common for all elements. If these two signals are mixed and the lower sideband retained, there results $\cos \{[\omega_c - \omega_p + f(t)]t - \beta\}$; therefore, it is seen that the phase of the resultant IF signal is independent of the relative phase angle of the signals at the elements. The signals from the output of these mixers, (one for each element) which are in phase, are summed. At the point of summation, the receiver array gain is realized for the information signals.

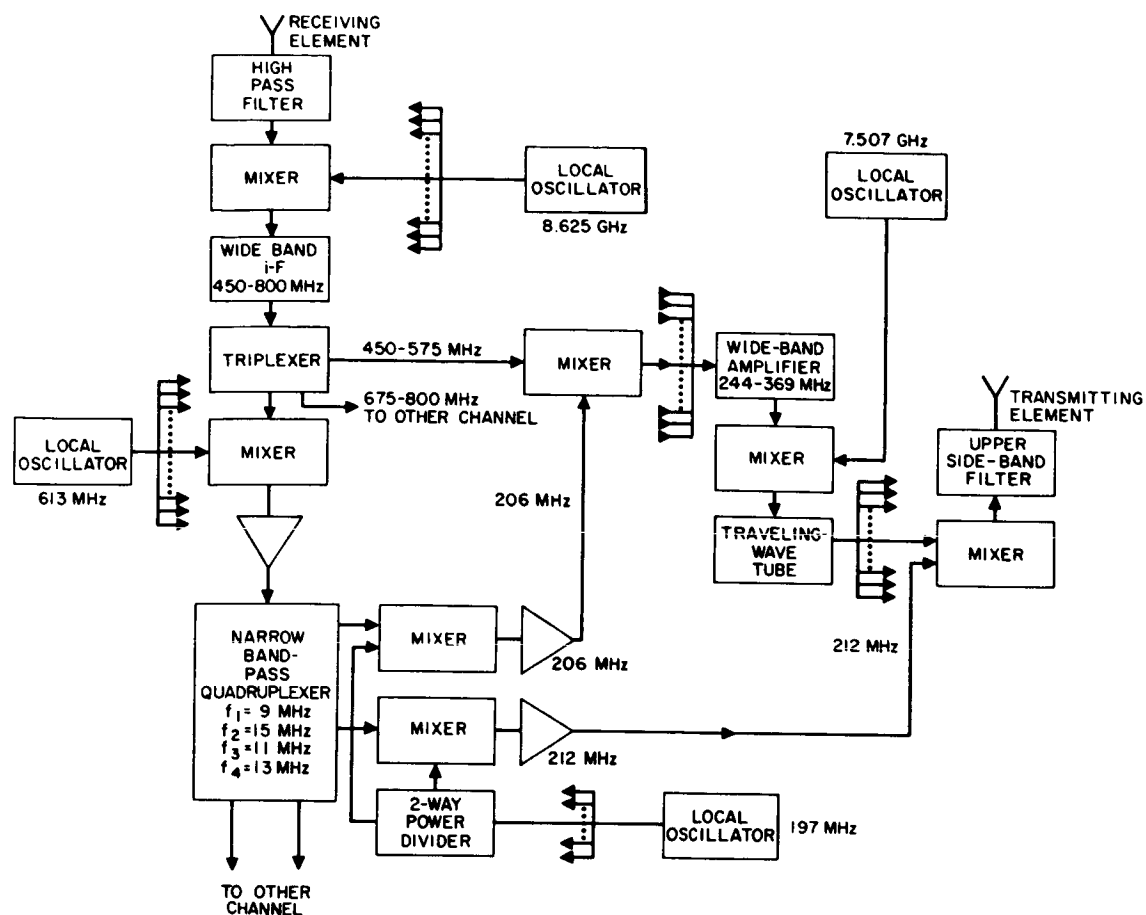


Figure A. High-Gain , Self-Steering Engineering Model Schematic

HIGH GAIN, SELF-STEERING ANTENNA SYSTEM FOR SATELLITE-EARTH COMMUNICATIONS

The signal is then amplified at IF, up-converted to RF, amplified at RF, and then distributed to the final transmitting mixers. At these mixers the transmitting pilot is mixed with the modulated signal and the upper sideband is selected by the band-pass filter that follows. A modulated signal is produced at a transmitting element; this signal has a phase angle which has the opposite sense from the phase angle of the transmitting pilot at the corresponding receiving element. The condition necessary to transmit the information from the antenna system in the direction of the transmitting pilot is that the recovery and transmitting arrays be scaled in wavelength.

The Table presents the electrical and physical characteristics of the engineering model, and the projected characteristics for a flight model of a similar system. Figure C shows the configuration for an engineering model. The flight model will be configured similarly.

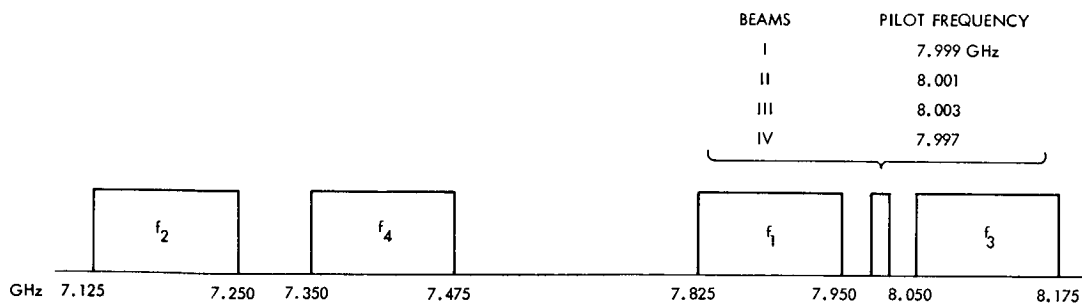
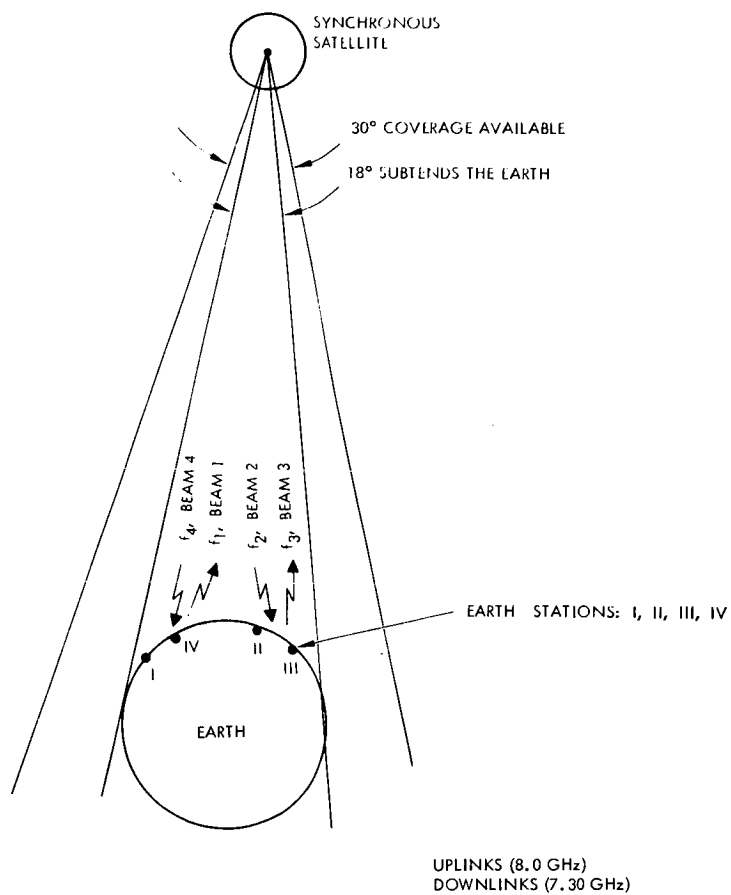


Figure B. Synchronous Altitude Gravity-Gradient 30-Degree Cone of Coverage

Transmitting and Receiving Apertures
Radio Frequency Antennas

HIGH GAIN, SELF-STEERING ANTENNA SYSTEM FOR SATELLITE-EARTH
COMMUNICATIONS

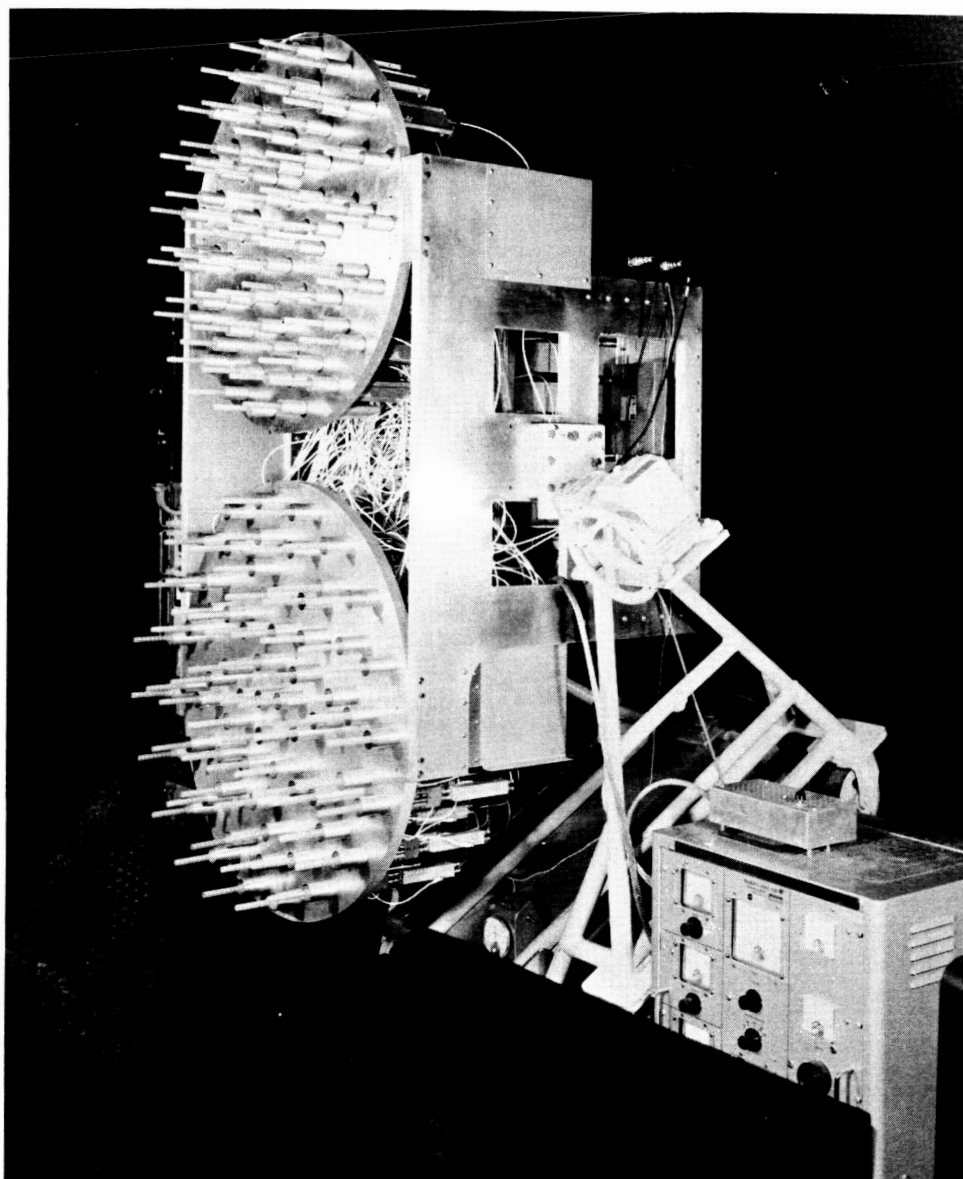


Figure C. Engineering Model of Self Steering Antenna

Characteristics of the Repeater

Parameter	Design Goals for Engineering Model	Measured Performance	Projected Characteristics of Flight Model
Number of elements, each for two arrays	64		64
Number of channels	2		2
R-f bandwidth, each channel	125 MHz		125 MHz
Guard band, between channels	100 MHz		100 MHz
Total cone angle of coverage	30 degrees	30 degrees	30 degrees
Element gain, minimum	11.6 db	12.4 db	11.6 db
Array gain, minimum	29.8 db	30.4 db	29.8 db
Polarization	Circular	Circular axial ratio = 0.8 db	Circular
Average mixed-filter pre- amplifier noise figure	15.2 db (max)	14.4 db	7 db
Effective radiated power	28.0 dbw	27.8 dbw	35.0 dbw
Ratio of pilot to modulated signal power when 125 MHz bandwidth is utilized	-10 db	*	-10 db
Power consumption: receiver	32.0 watts excluding L. O.		21.5 watts including L. O.
Power consumption: pilot processor	108.7 watts	182.6 watts	108.7 watts
Power consumption: attitude readout	0.9 watt		0.9 watt
Power consumption: total transmitter	201.1 watts excluding L. O.	87.9 watts** excluding L. O.	204.5 watts including L. O.
Prime power (exclusive of power supplies)	342.7 watts	270.5 watts	335.6 watts
Power consumption power supplies	—	—	73.3 watts
Power consumption: total prime power	—	—	408.9 watts
Total weight		180.0 Kg	79.0 Kg
<p>* Pilot signal ratios much lower than the design values were measured in the T-V tests (see discussion).</p> <p>** Power consumption applies to one channel of transmitter. For the channel operation less than twice this value would be needed.</p>			

ANTENNAS FOR SPACE COMMUNICATION – SURFACE STATION COST BURDENS

Surface station cost burdens are based on the cost of the DSIF antennas.

The cost to construct a surface station antenna has been evaluated as a function of antenna diameter. The cost evaluation has been made compatible with the communications system methodology developed under this contract and used to evaluate different communications systems.

The relationship used by the communication system methodology to relate surface station antenna cost to antenna diameter is as follows:

$$C_{\theta_R} = K_{\theta_R} (d_R)^{m_R} + C_{KR}$$

where

C_{θ_R} = the antenna fabrication cost

K_{θ_R} = a constant relating antenna diameter to cost

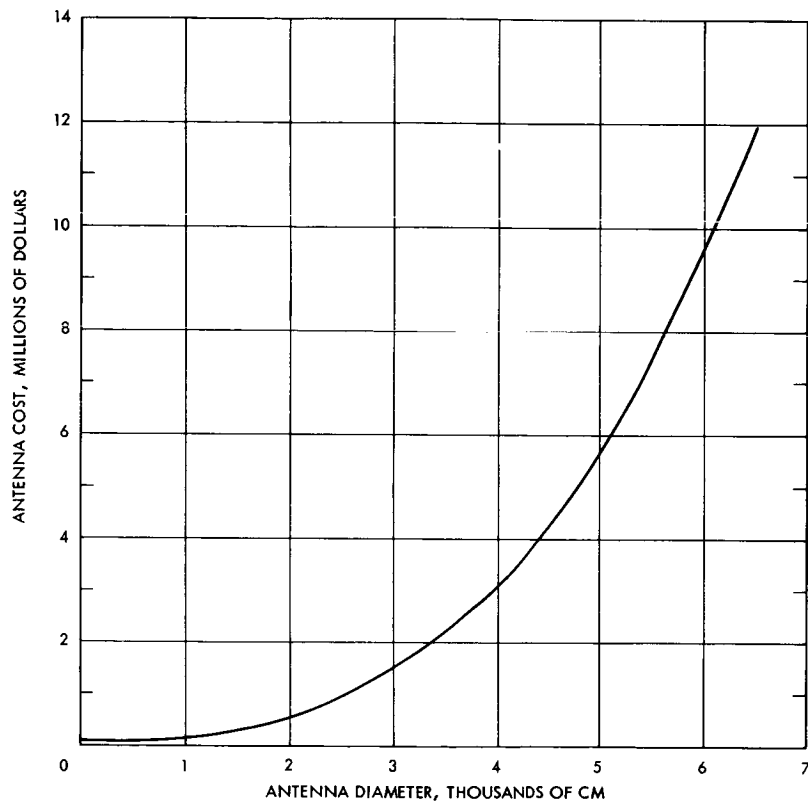
d_R = the receiver antenna diameter

m_R = a constant

C_{KR} = a fixed cost independent of antenna

The burden relationships are a function of frequency. Since 2.3 GHz is the assigned deep space communication frequency, the burdens, K_{θ_R} , m_R , and C_{KR} have been evaluated at this frequency. The data used in this evaluation is the cost data of the DSIF 85 foot and 210 foot antennas. The cost for these antennas are \$980,000 and \$12,000,000 respectively. These values are related by the diameter to the 2.8 power. These will be a certain amount of fixed cost for each station site (represented by C_{KR}), this is taken to be \$100,000 for a generalized surface station. Combining this value with the DSIF data gives values for the surface receiving aperture as indicated in the table. This relationship is plotted in the Figure.

Burden Constant	Surface Station Antenna Burdens	
	Antenna Diameter in cm	Antenna Diameter in feet
C_{KR}	100,000	100,000
m_R	2.8	2.8
K_{θ_R}	2.5×10^{-4}	3.5



Earth Station Antenna Costs

ANTENNAS FOR SPACE COMMUNICATION – SPACECRAFT COST BURDENS

Cost burdens for spacecraft antennas are documented which are suitable for use in the Communication System Methodology developed for this contract.

The cost to fabricate an antenna for spacecraft has been evaluated as a function of the antenna diameter. These relationships, with many similar ones for other parts of a deep space communications link, are used to determine the optimum values of spacecraft and surface station antenna size to provide a minimum cost system.

The relationship used in the antenna cost modeling is:

$$C_{\theta T} = K_{\theta T} (d_T)^{m_T} + C_{KT} \quad (1)$$

where

$C_{\theta T}$ = the transmitting antenna cost

$K_{\theta T}$ = a constant relating transmitting antenna fabrication cost to diameter

d_T = the transmitting antenna diameter

m_T = a constant

C_{KT} = the transmitting antenna fabrication cost independent of aperture diameter.

It is required then to provide values for the three burden constants used in equations 1: $K_{\theta T}$, m_T , and C_{KT} .

These constants are derived from two basic sources. The first source is the cost data that may be obtained from the construction of mechanical structures similar to antennas and the second source is the actual construction costs for spaceborne antennas.

The antenna costs used for the communications system methodology are the costs for a single antenna with engineering development costs prorated over several antennas. Table I contains representative cost data for large spaceborne antennas, including the development costs. The costs may be interpreted in terms of the burdens for spacecraft transmitting antennas as indicated in Table II for 10 antennas and 15 antennas. These costs are for large extensible antennas. Burden costs for small (less than 10 feet equivalent diameter) spaceborne antennas are given by the burden values of Table III for planar arrays and parabolic dishes, the Figure has plots of the values of cost for these several antennas.

Table I. Cost Size Relationships for Large Spaceborne Antennas

Item	Remarks	Cost Dollars	
		Diameter d, in Feet	Diameter, d, in cm
Engineering	Fixed Cost 20 man years	600,000	600,000
Tooling	Surface dependent	$500d^2$	$0.538d^2$
Fabrication	Surface dependent	$170d^2$	$0.183d^2$
Material	Surface dependent	$10d^2$	$0.0106d^2$
Quality Control	Surface dependent	$80d^2$	$0.086d^2$
Handling and	Volume dependent	$0.5d^3$	$1.76 \times 10^{-5}d^3$

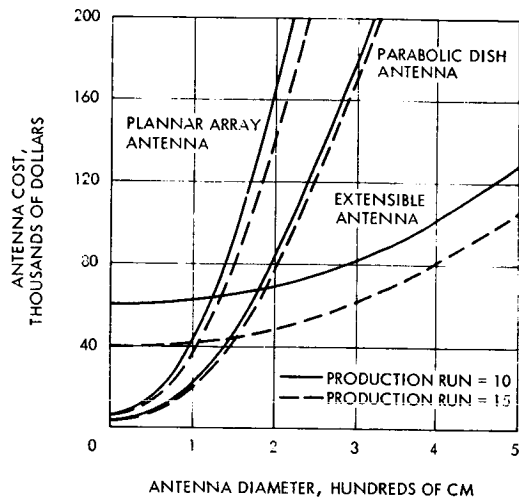
Table II. Antenna Burdens for Large Extensible
Spacecraft Antennas

Burden Constant	10 Antennas Constructed		15 Antennas Constructed	
	Diameter in cm	Diameter in Feet	Diameter in cm	Diameter in Feet
K_{θ_T}	0.084	150	0.08	143
C_{K_T}	60,000	60,000	40,000	40,000
m_T	2.19	2.19	2.19	2.19

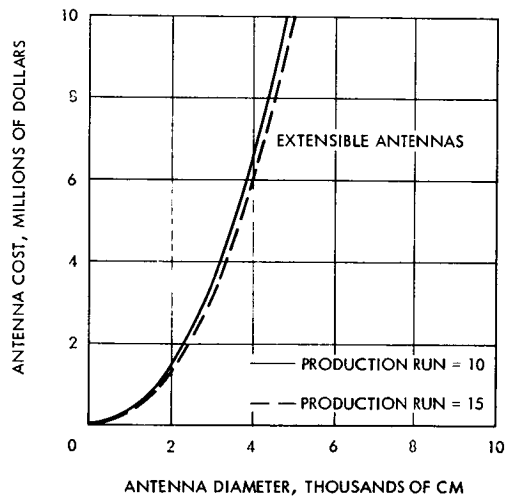
ANTENNAS FOR SPACE COMMUNICATION – SPACECRAFT COST BURDENS

Table III. Antenna Burdens for Small Spacecraft Antennas

Burden Constant	10 Antennas Constructed		15 Antennas Constructed	
	Diameter in cm	Diameter in Feet	Diameter in cm	Diameter in Feet
(a) Planar Array				
K_{θ_T}	4	3710	3.4	3160
C_{KT}	5000	5000	5000	5000
m_T	2	2	2	2
(b) Parabolic Dish				
K_{θ_T}	2	1850	1.92	1780
C_{KT}	2500	2500	2500	2500
m_T	2	2	2	2



(a)



(b)

Cost of Spacecraft Transmitting Antennas

TRANSMITTING AND RECEIVING APERTURES

Optical Frequency Apertures

	Page
Introduction	282
Optical Configurations	284
Optical Configurations — Optics with a Large Field of View	286
Optical Configurations — Use of a Tracking Mirror	290
Optical Configurations — Reflectance and Attenuation in Optical Systems	292

INTRODUCTION

Subsequent subtopics include: optical configurations, manufacturing techniques and tolerances, and material choices for optical beamwidths from 1 to 100 microradians.

The optics used in a laser communications system are a major design consideration. To obtain improved communication system performance using optical wavelengths requires each area of the technology to be examined such that potential implementation choices, basic limitations and interface requirements be understood.

The material of this section is organized in the following manner: The first subsection deals with optical configurations and the concomitant tolerances of such configurations. The next subsection deals with manufacturing techniques and thermal considerations while the third subsection deals with implementation factors such as material choices and their associated weight and cost.

The optical transmission aperture must provide a beam which is as narrow as possible to concentrate the transmitted energy at the receiver. This beamwidth produces one of the most difficult problems for optical communications and is therefore examined briefly below.

The beam spread of a perfect, unobstructed optical system can be depicted by a plot of the Airy Disk shape where the abscissa would be intensity of the energy and the ordinate would be the angular spread. The angular spread as measured by the diameter of the first ring of zero energy would be given by

$$\theta = \frac{2.44\lambda}{D} \quad (1)$$

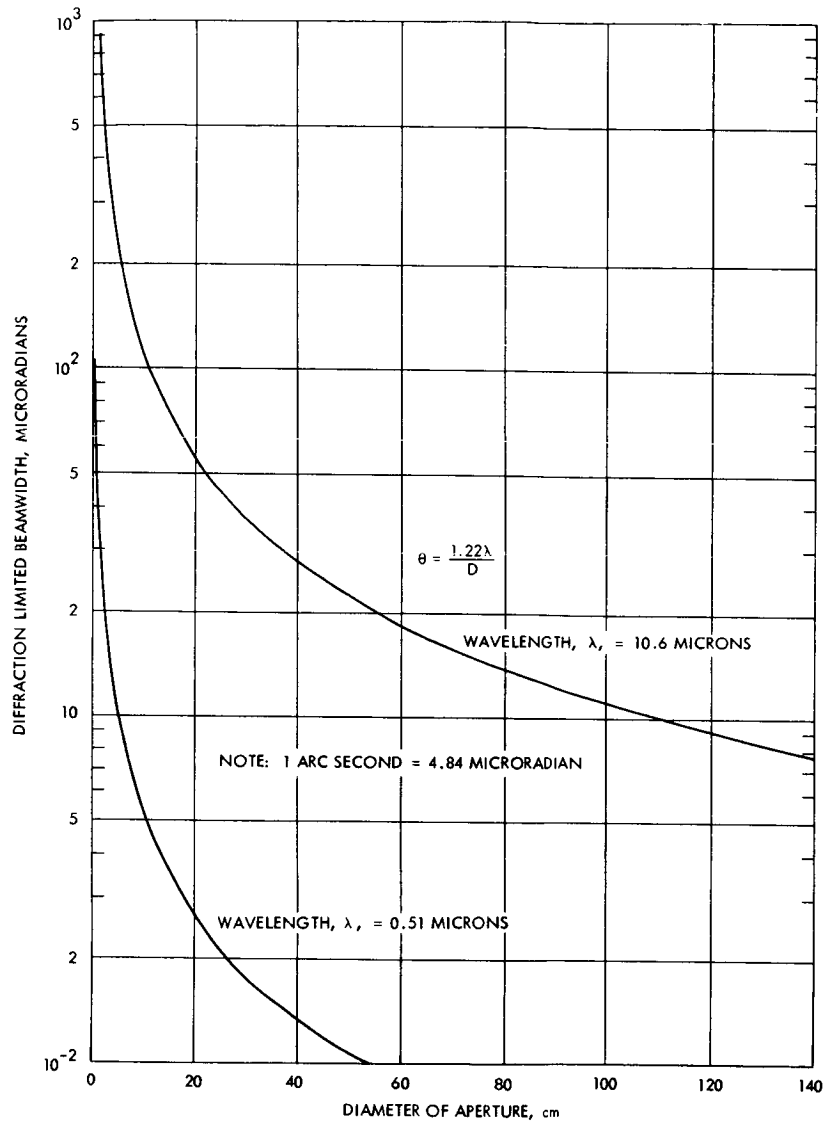
where

θ is the angular spread, radians

λ is the wave length

D is the aperture diameter

In practice however, the angular diameter that is useable is necessarily less than this and is taken at the half power points to be $\theta = 1.04\lambda/D$. (Another measure sometimes used for measuring beamwidths is the resolution of the aperture. Two adjacent point sources are considered to be resolvable when separated by $\theta = 1.22\lambda/D$.) The fact that the beam spread is inversely proportional to the aperture favors the use of a large diameter transmission unit. The narrow beamwidths of diffraction limited optics place severe pointing problems on the pointing and tracking system. As a measure of the pointing accuracy required, the diffraction limited beamwidth is plotted in the figure as a function of aperture for 0.51 and 10.6 microns. These results are possible only for excellent seeing when using the larger diameters. Such seeing may well be found only outside the Earth's atmosphere. As is seen from the figure, beamwidths as small as 1 microradian may be considered for visible light, beamwidths in the order of tens of microns are more appropriate for 10.6 micron transmission.



Diffraction Limited Beamwidths at
Optical Frequencies

Transmitting and Receiving Apertures Optical Frequency Apertures

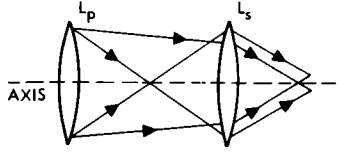
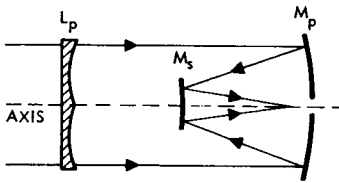
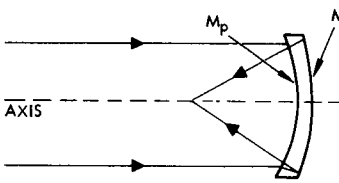
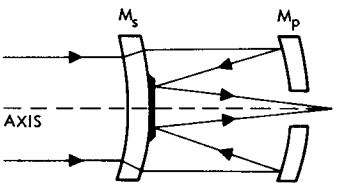
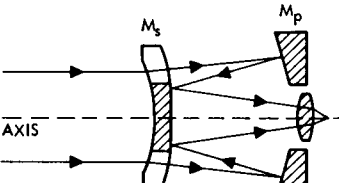
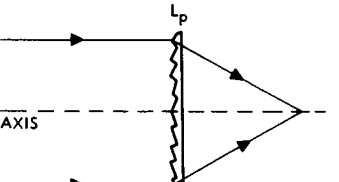
OPTICAL CONFIGURATIONS

Thirteen optical systems are shown in tabular form and pertinent design characteristics are given.

Optical Systems

TYPE	RAY DIAGRAM	OPTICAL ELEMENTS	PERTINENT DESIGN CHARACTERISTICS
PARABOLOID		REFLECTIVE M_p = PARABOLOIDAL MIRROR	<ol style="list-style-type: none"> 1. FREE FROM SPHERICAL ABERRATION. 2. SUFFERS FROM OFF-AXIS COMA. 3. AVAILABLE IN SMALL AND LARGE DIAMETERS AND f/NUMBERS. 4. LOW IR LOSS (REFLECTIVE). 5. DETECTOR MUST BE LOCATED IN FRONT OF OPTICS.
CASSEGRAIN		REFLECTIVE M_p = PARABOLOIDAL MIRROR M_s = HYPERBOLOIDAL MIRROR	<ol style="list-style-type: none"> 1. FREE FROM SPHERICAL ABERRATION. 2. SHORTER THAN GREGORIAN. 3. PERMITS LOCATION OF DETECTOR BEHIND OPTICAL SYSTEM. 4. QUITE EXTENSIVELY USED.
GREGORIAN		REFLECTIVE M_p = PARABOLOIDAL MIRROR M_s = ELLIPSOIDAL MIRROR	<ol style="list-style-type: none"> 1. FREE FROM SPHERICAL ABERRATION. 2. LONGER THAN CASSEGRAIN. 3. PERMITS LOCATION OF DETECTOR BEHIND OPTICAL SYSTEM. 4. GREGORIAN LESS COMMON THAN CASSEGRAIN.
NEWTONIAN		REFLECTIVE M_p = PARABOLOIDAL MIRROR M_s = REFLECTING PRISM OR PLANE MIRROR	<ol style="list-style-type: none"> 1. SUFFERS FROM OFF-AXIS COMA. 2. CENTRAL OBSTRUCTION BY PRISM OR MIRROR.
HERSCHELIAN		REFLECTIVE M_p = PARABOLOIDAL MIRROR INCLINED AXIS	<ol style="list-style-type: none"> 1. NOT WIDELY USED NOW. 2. NO CENTRAL OBSTRUCTION BY AUXILIARY LENS. 3. SIMPLE CONSTRUCTION. 4. SUFFERS FROM SOME COMA.
SCHMIDT		REFLECTIVE-REFRACTIVE M_p = SPHERICAL MIRROR M_s = REFRACTIVE CORRECTOR PLATE	<ol style="list-style-type: none"> 1. PRODUCES A CURVED FIELD. 2. FREE OF SPHERICAL ABERRATION AND COMA. 3. CENTRAL OBSTRUCTION BY ITS OWN FIELD SURFACE. 4. CAN OBTAIN LOW f/NUMBER. 5. SHARPER FOCUS OVER LARGER AREA THAN PARABOLOID. 6. MAY BE BUILT AS SOLID UNIT.
GALILEAN		REFRACTIVE L_p = BICONVEX LENS L_s = BICONCAVE LENS	<ol style="list-style-type: none"> 1. RADIATION GATHERING POWER LESS THAN REFLECTION SYSTEMS. 2. RELATIVELY SHORT LENGTH. 3. LIMITED FIELD OF VIEW. 4. SPECTRAL RESPONSE LIMITED BY LENS MATERIAL.

Optical Systems (continued)

KEPLERIAN		REFRACTIVE L_s = BICONVEX LENS L_p = BICONVEX LENS	<ol style="list-style-type: none"> 1. RADIATION GATHERING POWER LESS THAN REFLECTION SYSTEMS. 2. SPECTRAL RESPONSE LIMITED BY LENS MATERIAL. 3. NOT WIDELY USED NOW.
SCHMIDT-CASSEGRAIN OR BAKER		REFLECTIVE-REFRACTIVE M_p = ASPHERIC MIRROR M_s = ASPHERIC MIRROR L_p = REFRACTIVE CORRECTOR PLATE	<ol style="list-style-type: none"> 1. PRODUCES FLAT FIELD. 2. VERY SHORT IN LENGTH. 3. COVERS LARGE FIELD. 4. CORRECTOR PLATE HAS LARGER CURVATURE THAN SCHMIDT.
MANGIN MIRROR		REFRACTIVE-REFLECTIVE M_p = SPHERICAL REFRACTOR M_s = SPHERICAL REFLECTOR	<ol style="list-style-type: none"> 1. SUITABLE FOR IR SOURCE SYSTEMS. 2. FREE OF SPHERICAL ABERRATION AND COMA 3. MOST SUITABLE FOR SMALL APERTURES. 4. COVERS SMALL ANGULAR FIELD. 5. USES SPHERICAL SURFACES.
MAKSUTOV		REFRACTIVE-REFLECTIVE M_p = MENISCUS REFLECTOR M_s = MENISCUS REFRACTOR-REFLECTOR	<ol style="list-style-type: none"> 1. FREE OF SPHERICAL ABERRATION, COMA, AND CHROMATISM. 2. VERY COMPACT. 3. LARGE RELATIVE APERTURE. 4. MAY ALSO USE COMBINATIONS OF SPHERICAL AND ASPHERIC ELEMENTS.
GABOR		REFRACTIVE-REFLECTIVE M_p = SPHERICAL REFLECTOR M_s = SPHERICAL REFRACTOR-PLANO REFLECTOR	<ol style="list-style-type: none"> 1. HIGH APERTURE SYSTEM. 2. HAS MEAN CORRECTION OF SPHERICAL ABERRATION AND COMA. 3. SUITABLE FOR IR SOURCE SYSTEMS.
FRESNEL LENS		REFRACTIVE L_p SPECIAL FRESNEL LENS	<ol style="list-style-type: none"> 1. FREE OF SPHERICAL ABERRATION. 2. INHERENTLY LIGHTER WEIGHT. 3. SMALL AXIAL SPACE. 4. SMALL THICKNESS REDUCES INFRARED ABSORPTION. 5. DIFFICULT TO PRODUCE WITH PRESENT INFRARED TRANSMITTING MATERIALS.

OPTICAL CONFIGURATIONS – OPTICS WITH A LARGE FIELD OF VIEW

Optical systems having a field of view larger than beamwidth are examined for suitability to laser communications.

The previous topic noted the narrow beamwidths (and consequent accurate pointing requirements) of optical frequency apertures. This topic gives optical configuration choices which may be considered in meeting the requirements of optical frequency communication.

Because of the need to track with extremely small angular errors, it is quite desirable to design a system in which the fine pointing is accomplished by movement of small elements of the system, rather than movement of the entire system. This can be accomplished in one of three ways.

1. A larger field of view than is necessary can be provided and the transmitter source and receiver detector can be moved in the focal plane for tracking.
2. A reimaging optical system can be made with a small transfer lens that moves to provide fine pointing.
3. The optical system can incorporate a space of collimated light in which a mirror is inserted. Control of the mirror motion would control the image position in receiving, or the beam direction of a transmitter.

The first of these pointing methods is discussed below. The second is discussed extensively in the next Part of this Volume, Acquisition and Tracking. The third way is discussed in the next topic.

If the receiver and transmitting units are moved in the focal plane, there is a relatively wide choice of available systems. It is desirable, for thermal reasons, to keep the focal plane location away from the critical structure controlling alignment. This factor would perhaps favor a Cassegrain configuration, over a Newtonian configuration. In addition, the Cassegrain's shorter structure, lighter weight and lower moment of inertia favor the use of a Cassegrain over a Gregorian. The diameter of the diffraction limited field of Newtonian telescope – defined as that field where the length of the coma (see definition of coma below) is less than the Airy disc diameter – is about the same as a Cassegrain of equivalent focal ratios and is given by:

$$h = 30 \lambda f^3 \quad (10-2)$$

where:

h is field diameter

λ is wave length

f is focal ratio

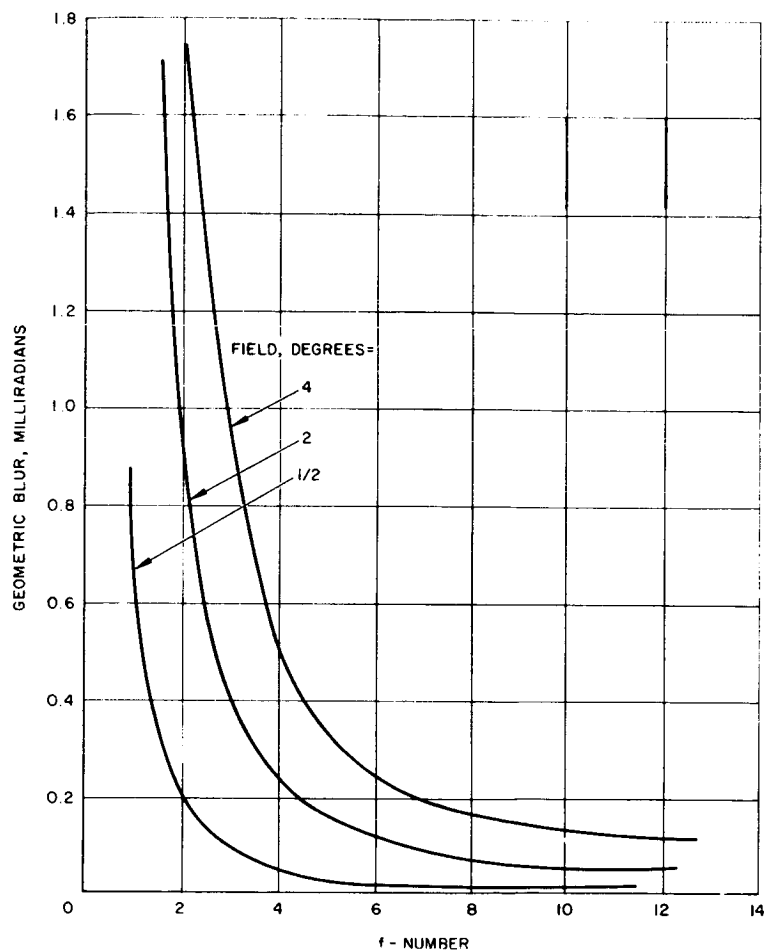


Figure A. Coma and Astigmatism Aberrations of a Parabolic Reflector Off Axis, Versus f Number

OPTICAL CONFIGURATIONS – OPTICS WITH A LARGE FIELD OF VIEW

Geometric blur is one of the penalties for off axis tracking. Figure A plots geometric blur due to coma and astigmatism versus f-number as a function of field of view for 0.51 microns. These plots strictly hold for only parabolic systems. Figure B compares the geometric blur circle of several types of telescopic systems at varying angular distances off-axis. Chosen for each category are typical examples found in use today. The f-number of these examples vary so that the comparison is not a strict one.

Parabolic types are represented by an f-10 instrument after K. Schwarzschild in "Telescopes & Accessories" by Baker.

The Ritchey-Chretien system is an f/15 telescope discussed in "Sky & Telescope" for April 1962. The Dall Kirkham system is also an f/15 telescope discussed in the same reference.

The parabolic telescope with a Ross Corrector is the 200" Hale telescope operating at about f/4, discussed in "Telescopes" by Kuiper.

The three catadioptric telescopes (the classical Schmidt, the Maksutov-Bouwers, and the compound catadioptric) are all f/2 systems discussed by Bouwers in "Achievements in Optics". The compound catadioptric is a corrected concentric arrangement employing both a meniscus lens and a Schmidt plate.

The Ritchey-Chretien is of the Cassegrain type, but uses an approximately hyperboloidal primary mirror (with greater asphericity than the paraboloidal mirror of the Cassegrain of the same speed) to achieve a larger field. The coma of a Cassegrain system is equal to the equivalent focal length Newtonian while the astigmatism would be that of the equivalent Newtonian multiplied by the magnification of the Cassegrain secondary. These data come primarily from K. Schwarzschild as reported in "Telescopes & Accessories" by Baker. The coma alone, in radians, equals:

$$\text{Coma} = \frac{0.1875\phi}{(efl/D)^2} = \frac{0.1875\phi}{f^2} \quad (10-3)$$

where ϕ is the half field angle in radians.

efl is the effective focal length

D is the aperture diameter

f is the f-number

This is plotted in Figure C.

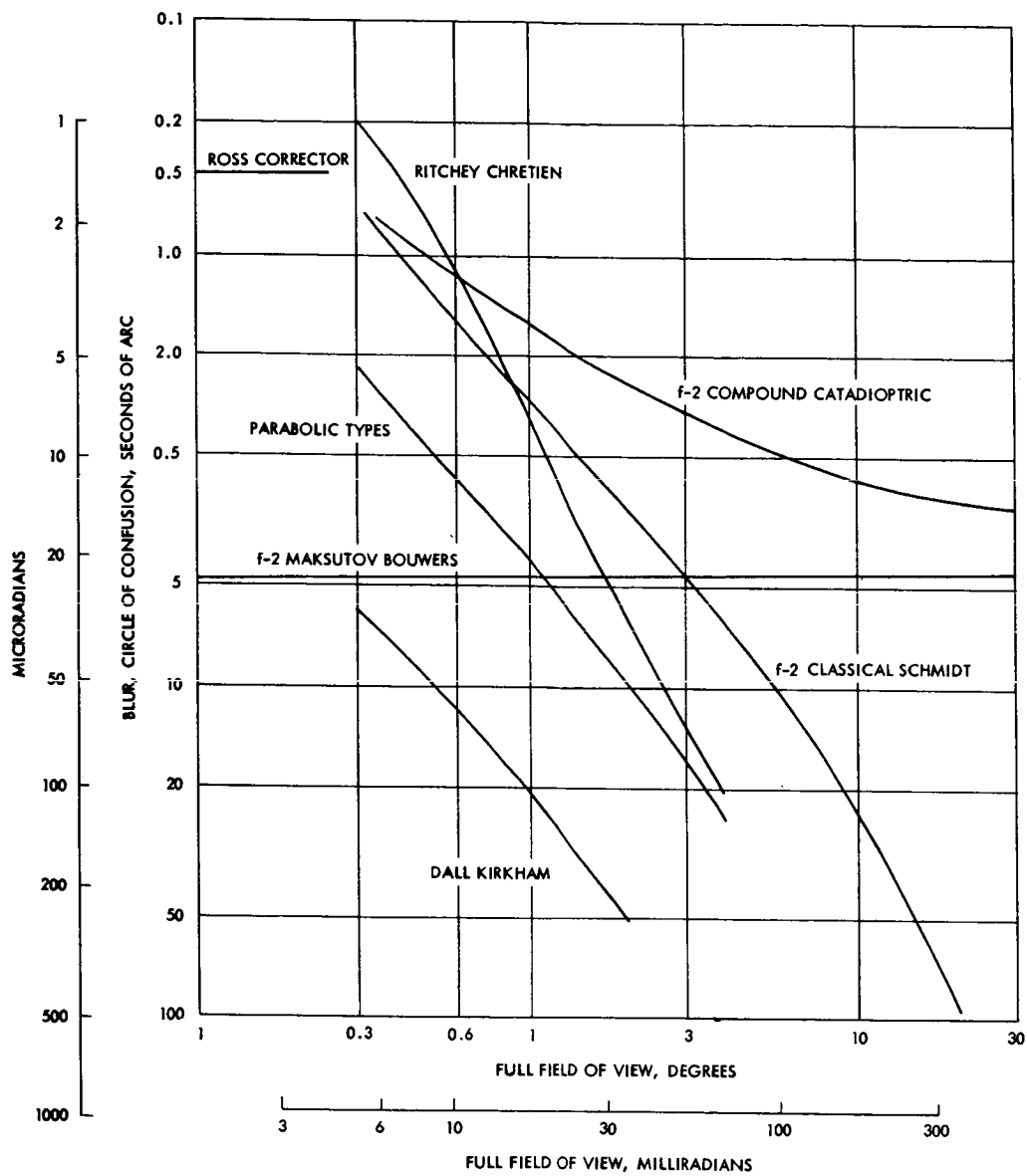


Figure B. Geometric Blur of Various Types
Versus Field of View

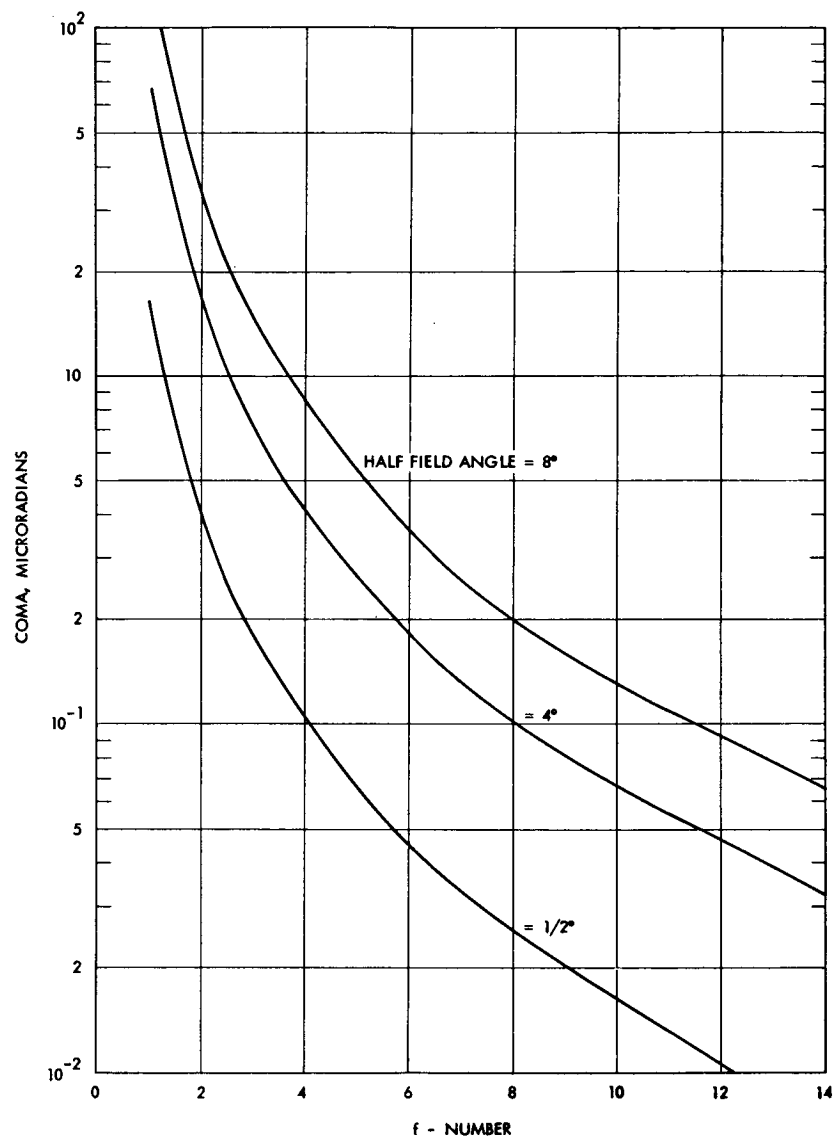


Figure C. Coma of a Cassegrain System

OPTICAL CONFIGURATIONS - USE OF A TRACKING MIRROR

Using a tracking mirror placed in collimated light a space of collimated light is very advantageous when using a mirror for fine pointing in an optical communication link.

If the spacecraft has vibration sources, tracking may be accomplished by using a space of collimated light into which a mirror is inserted. This has an added advantage in that the collimated space can also serve as a "soft link" and allow relative motion between the large optical system and the spacecraft. The situation is similar to that of a hand held telescope with an eyepiece providing collimated light. If the tracking mirror in the collimated space can remove image motion at the focal plane, large relative motions can be tolerated.

In a telescope using a mirror in a collimated space, there will be angular magnification proportional to the ratio of the aperture of the telescope to the aperture of the collimated space. Because of this, the guiding mirror must deviate the beam by an amount equal to the input error times the magnification. This reduces the sensitivity to the roughness or the error in mirror positioning and makes the sensing of the tracking signal probably the most critical item.

In summary, the three methods of providing fine guiding without pointing the entire telescope require a larger field than will actually be used. The only real field requirement of the system is that of the lead angle due to the relative velocities of the sending and receiving stations. The amplitude of the fine guidance system is bounded by the field of the telescope.

In a communication mode, an optical system would handle only monochromatic light. While there may be two different wave lengths for the transmitting and receiving functions, the optical system still need not be designed to function over a broad spectrum unless it is to serve other purposes in addition to communications. This gives the optical designer the freedom to design catadioptric systems for example without consideration of secondary color, which is the usual limiting factor in system performance.

OPTICAL CONFIGURATIONS – REFLECTANCE AND ATTENUATION IN OPTICAL SYSTEMS

Reflectance and attenuation values are given for visible and 10.6 micron radiation.

From the optical design point of view, the main effect of a change to 10.6 μ from visible optical wavelengths is that of the materials required for the refracting elements. Figure A shows the attenuation of a typical refracting or photographic type telescope due to its having glass components. A list of materials suitable for use at infrared wavelengths is given in the Table. It is immediately apparent that a system designed for 10.6 μ would be primarily a reflector.

Figure B shows the efficiency of a typical reflecting telescope due to the reflective coatings on its mirrors.

Figure C illustrates the difference in the reflectivity of both aluminum and silver coatings and their dependence on the orientation of the plane of polarization.

Optical Materials

Material	Transmission Range, Microns	Refractive Index at Midpoint of Transmission Range	Transmittance, Percent	Diameter, Inches	1 in. dia x 0.040 in. Unpol. Substrate, Typical Price, Dollars	Water Solubility gm/100 gm Water
Irtran 1	0.5-9.	1.31	90.	8.	37.00	Insoluble
Irtran 2	0.7-14.5	2.20	75.*	8.	35.00	Insoluble
Irtran 3	0.4-11.5	1.35	90.	6.	62.00	Practically insoluble
Irtran 4	0.5-22.	2.40	70.*	6.	84.50	Insoluble
Irtran 5	0.4-9.5	1.62	80.*	6.	75.50	Insoluble
Irtran 6	1.5-31.	2.71	65.*	3.	209.00	Insoluble
Calcium fluoride	0.13-12.	1.42	95.	7.	3.20	0.002
Calcium fluoride (rare earth free)	0.13-12.	1.42	95.	7.	9.00	0.002
Barium fluoride	0.15-15.	1.44	94.	6.	8.90	0.2
Fused quartz	0.2-4.5	1.46	94.	18.	1.75	Insoluble
Glass	0.3-2.7	1.51-1.90	82.*-92.	200.	1.00-25.00	Insoluble
Sapphire	0.2-6.5	1.70	88.*	5.	15.00	Insoluble
Arsenic trisulfide glass	1.0-12.	2.40	75.*	18.	22.50	Insoluble
Silicon	1.2-15.	3.44	54.*	12.	7.95	Insoluble
Germanium	1.8-22.	4.03	47.*	20.	7.95	Insoluble
Cesium iodide	0.25-70.	1.7	86.	6.	12.45	44.
Indium arsenide	3.8-7.	3.5	53.*	3.	71.00	Insoluble
KRS-5	0.5-40.	2.35	72.	4.	16.00	0.05
Sodium chloride	0.2-26.	1.52	92.	12.	0.75	35.7
Sodium fluoride	0.2-15.	1.30	96.	3.	8.90	4.22
Potassium bromide	0.2-40.	1.53	92.	12.	1.35	53.5
Potassium chloride	0.2-30.	1.5	93.	12.	3.55	34.7
Potassium iodide	0.4-42.	1.60	89.	8.	3.55	140.

* Can be antireflection coated so that transmittance is greater than 95 percent.

Sources: "State-of-the-Art Report Optical Materials for Infrared Instrumentation," University of Michigan Willow Willow Run Laboratories, January 1957.

Tech Bits, Eastman Kodak Publication for Scientists and Engineers, 1966, No. 2, pp. 11, 12, 13.

Herron Optical Company, private communication

Harshaw Chemical Company, private communication

MIL Handbook 141

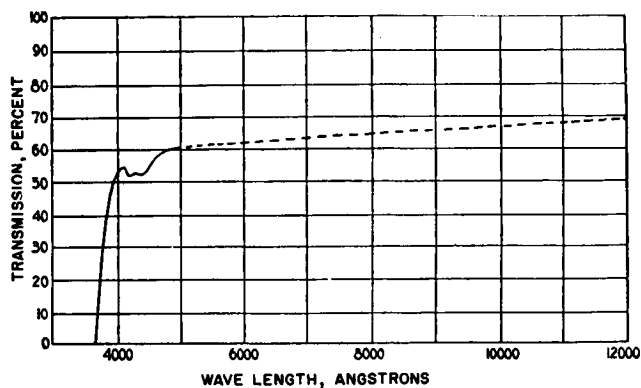


Figure A. Transmission Curve of the Lick 36-Inch Refractor and its Photographic Correcting Lens

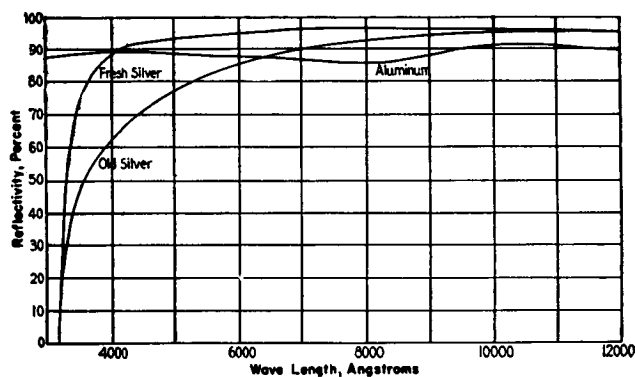


Figure B. Reflectivity of Evaporated Aluminum and Chemically Deposited Silver

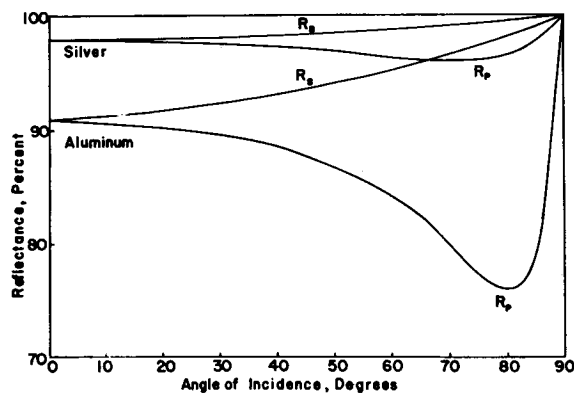


Figure C. The Reflectances of Aluminum and Silver for Different Angles of Incidence, When the Plane of Polarization is Parallel (R_p) and Perpendicular (R_s) to the Plane of Incidence

TRANSMITTING AND RECEIVING APERTURES

Optical Apertures — Optical Configurations

	Page
Effect of Surface Tolerances	296
Alignment Tolerances	298
Speed of Optical Configuration	310
Thermal Effects on Optical Configurations	312
Low Temperature Coefficient Material, CER-VIT	314

EFFECT OF SURFACE TOLERANCES

Surface tolerances must be reduced to a small fraction of a wavelength to maintain maximum energy density in the optical beam.

The accuracy of the optical system, i. e. , the optical surfaces and alignment, must be quite good to keep beam spread to a minimum. Using the perfect optical system as a reference, the loss of energy due to wave front errors can be calculated based on the work of Marechal and Francon*. The correlation between reduction in energy and wave front errors as calculated by this method is given in the Table. From these values, the need for keeping errors to a minimum can readily be seen.

Optical tolerances can be expressed in terms of the root-mean-square wave front error. The wavefront error is a combination of errors due to mirror surface errors, glass inhomogeneities, misalignment of elements and defocussing. (Discussion of alignment tolerances is given in the next topic.)

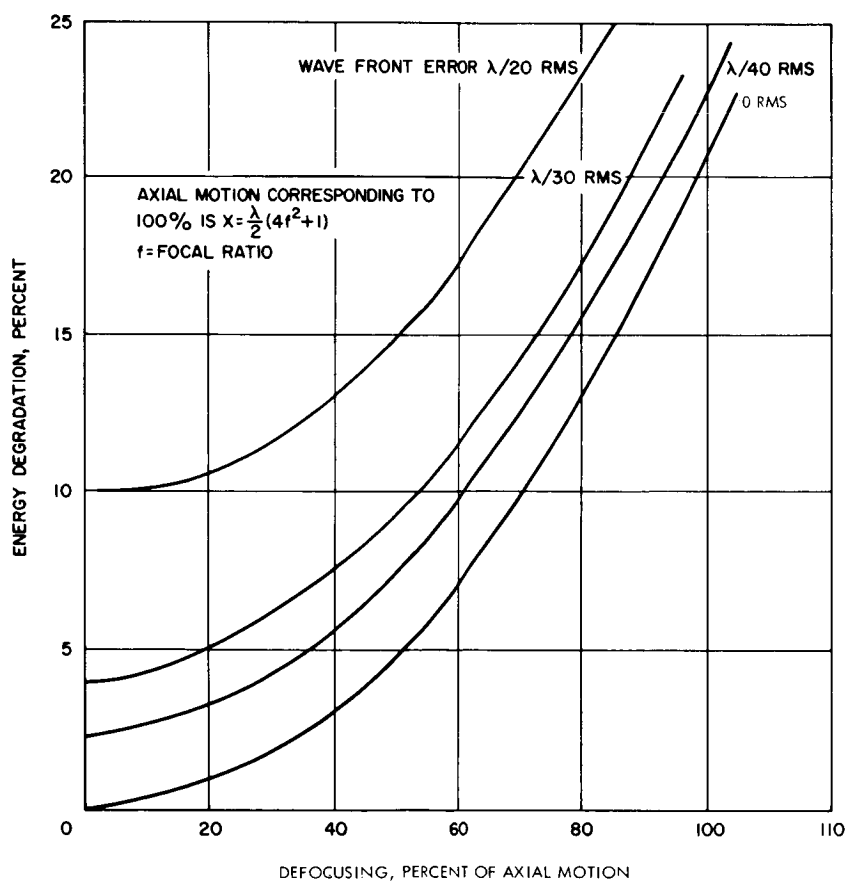
The addition of errors expressed in RMS form is not strictly correct although wave front errors along any path are additive (and can cancel each other). The figure** shows how focussing error combines with a random wavefront error to reduce contrast in a receiver or energy level in a transmitter.

* Marchal, A. and Francon, M. , Diffraction, Edit. Rv. d'Optique, Paris, 1960.

** Based on unpublished calculations by Dr. J. Kiewiet de Jonge, Allegheny Observatory and American Optical Company.

Effect of Wave Front Errors in Energy Density
at the Center of the Beam

Wave Front Errors RMS Values	Percent Reduction of Energy Density at the Center of the Beam
0	0%
$\lambda/28$	5%
$\lambda/20$	10%
$\lambda/16$	15%
$\lambda/14$	20%



Combined Defocussing and Wave Front Errors

ALIGNMENT TOLERANCES

Several alignment tolerances are given as a function of system magnification, wavelength and f-number. These tolerances are also plotted.

In general, the alignment tolerances on elements of a telescope are determined by their function in the system and the wave length of light. Thus, as the size of a telescope is increased, the tolerances do not scale and maintaining alignment becomes more difficult.

The order of magnitude of these tolerances can be seen from the following tolerance relationships for Cassegrain telescopes. All positions are relative to the primary mirror. The tolerances are:

1. Axial position of the secondary mirror

$$\Delta_{\max} = \frac{\lambda}{2} \left[4 \left(\frac{f}{m} \right)^2 + 1 \right] \quad (1)$$

based on the criterion – wavefront error not to exceed $\lambda/4$.
This equation is plotted in Figures A and B.

2. Decentration of secondary mirror

$$\delta_{\max} = \frac{13 f^3 \lambda}{m^2 - m} \quad (2)$$

based on a comatic image not to exceed Airy Disc Diameter.
This equation is plotted in Figures C and D.

3. Tilt of secondary mirror

$$\psi_{\max} = \frac{13 f^3 \lambda}{B (m^3 - m)} \quad (3)$$

based on a "comatic" image – not to exceed airy disc diameter.
This equation is plotted in Figures E and F as $(B) (\psi_{\max})$. The dimension B is defined by Figure G.

4. Position of focal plane

$$\Delta C_{\max} = \frac{\lambda}{2} (4 f^2 + 1) \quad (4)$$

based on the criterion that wavefront error is not to exceed $\lambda/4$. This equation is plotted in Figure H.

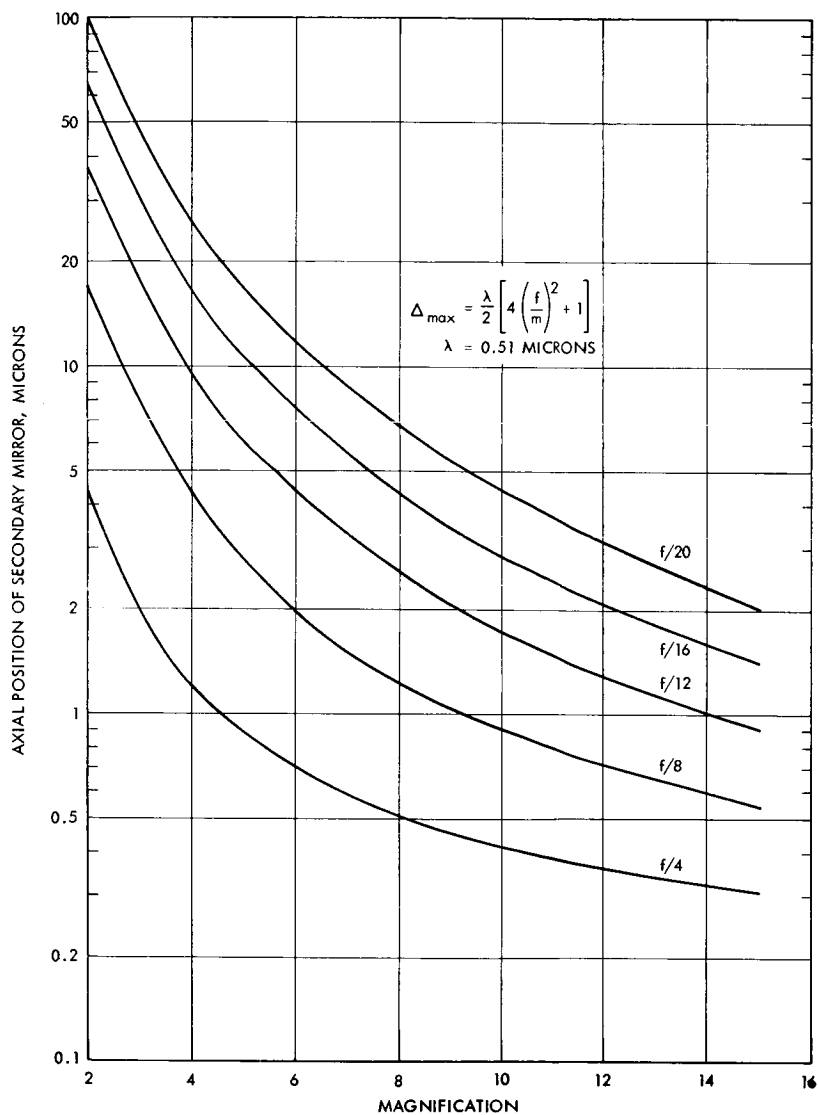


Figure A. Tolerance for Axial Position of Secondary Mirror for a Cassegrain Telescope

ALIGNMENT TOLERANCES

Symbols are defined as:

λ = wave length

f = focal ratio of Cassegrain system

m = magnification of secondary mirror of the Cassegrain system =
 C/B

A, B and C are shown in Figure G.

The tolerance budget of a system must be realistic to allow the designer to have maximum freedom possible. As an example, the equations just presented for a Cassegrain telescope tacitly assume that the focus is maintained by the control of dimensions A and C. If the focus were maintained by observation of the image in the focal plane (by an observer or automatic focusing device), then the half range of axial position of the secondary mirror, without introducing wave front errors exceeding one quarter wave, becomes

$$x_h = \frac{256 \lambda f^4}{m^5 - m^4 - m^3 + 1} \quad (5)$$

This provides considerable relaxation of this tolerance, and makes use of an optical focusing device very desirable. This equation is plotted in Figures I and J. These figures may be compared with Figures A and B to note the relaxation in tolerances.

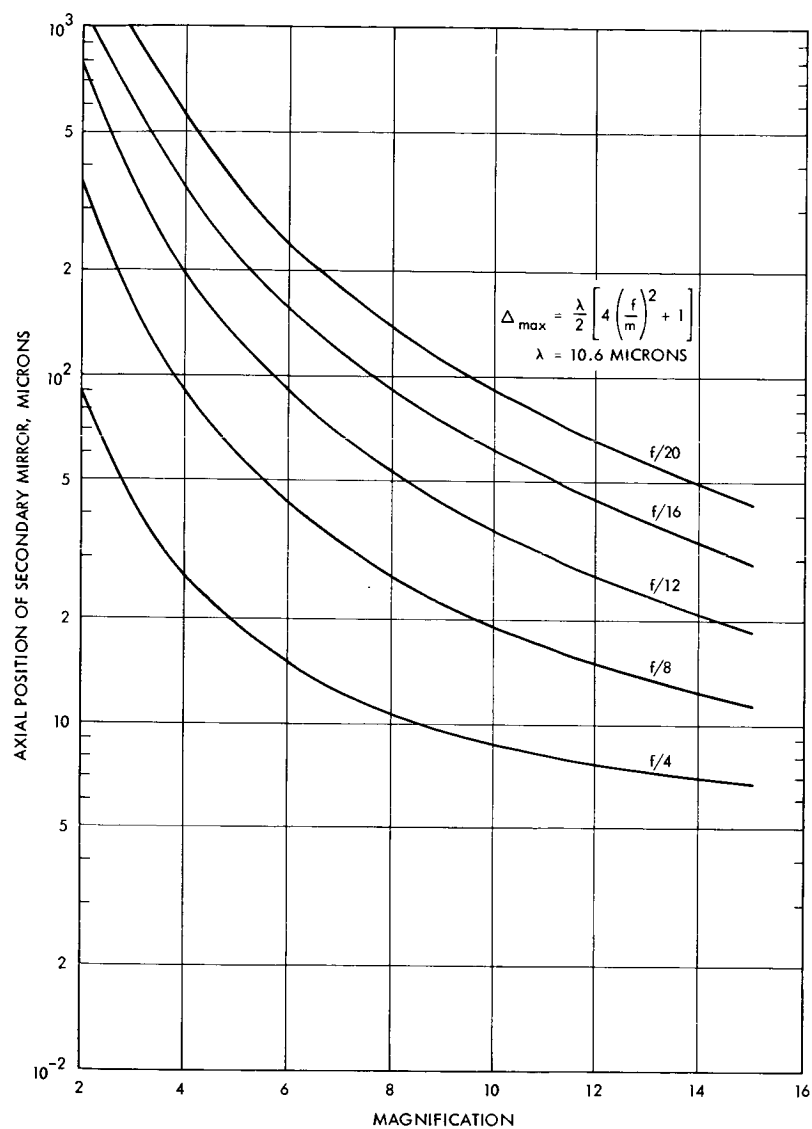


Figure B. Tolerance for Axial Position of Secondary Mirror for a Cassegrain Telescope

ALIGNMENT TOLERANCES

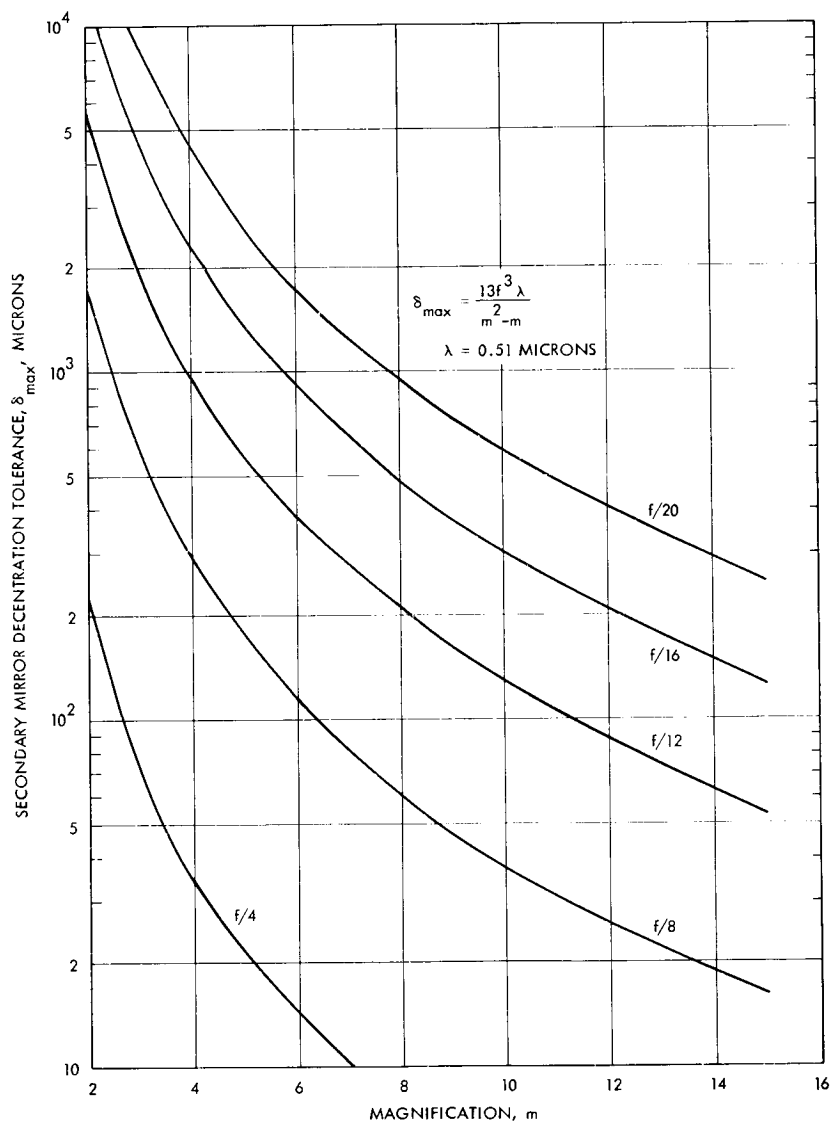


Figure C. Tolerance for Decentration of Secondary Mirror for a Cassegrain Telescope

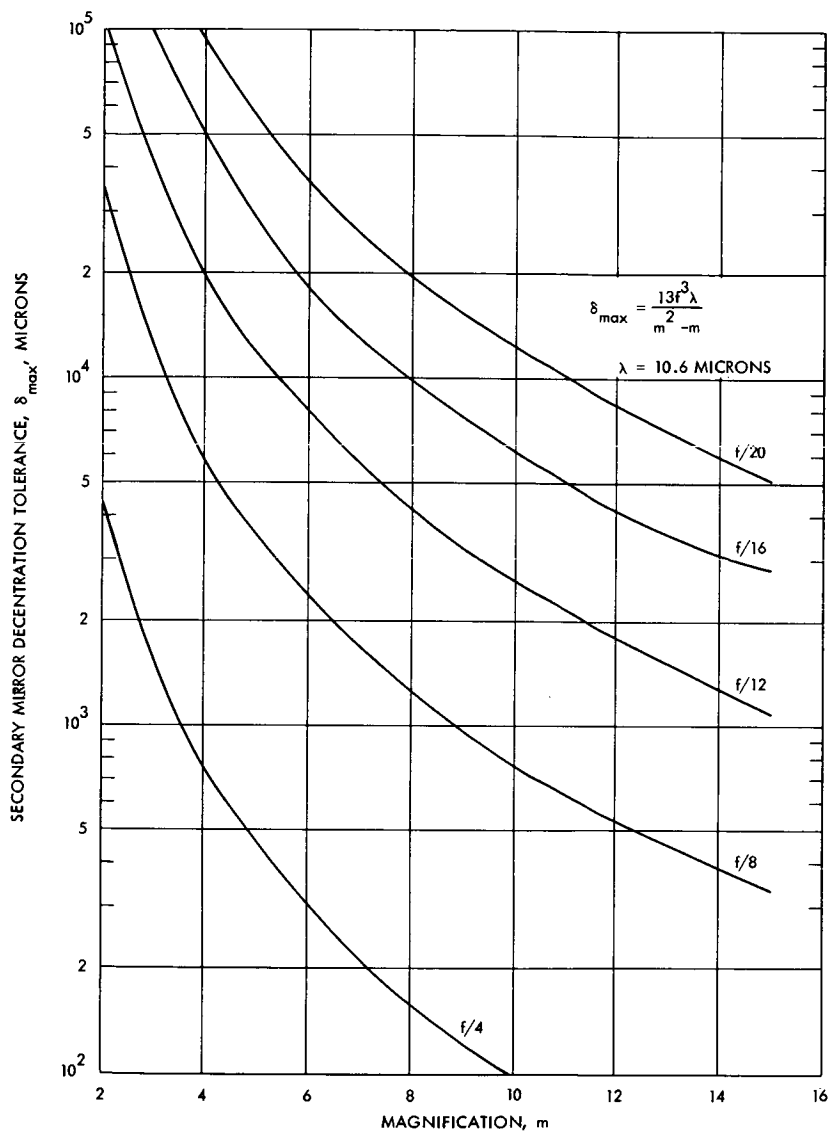


Figure D. Tolerance for Decentration of Secondary Mirror for a Cassegrain Telescope

ALIGNMENT TOLERANCES

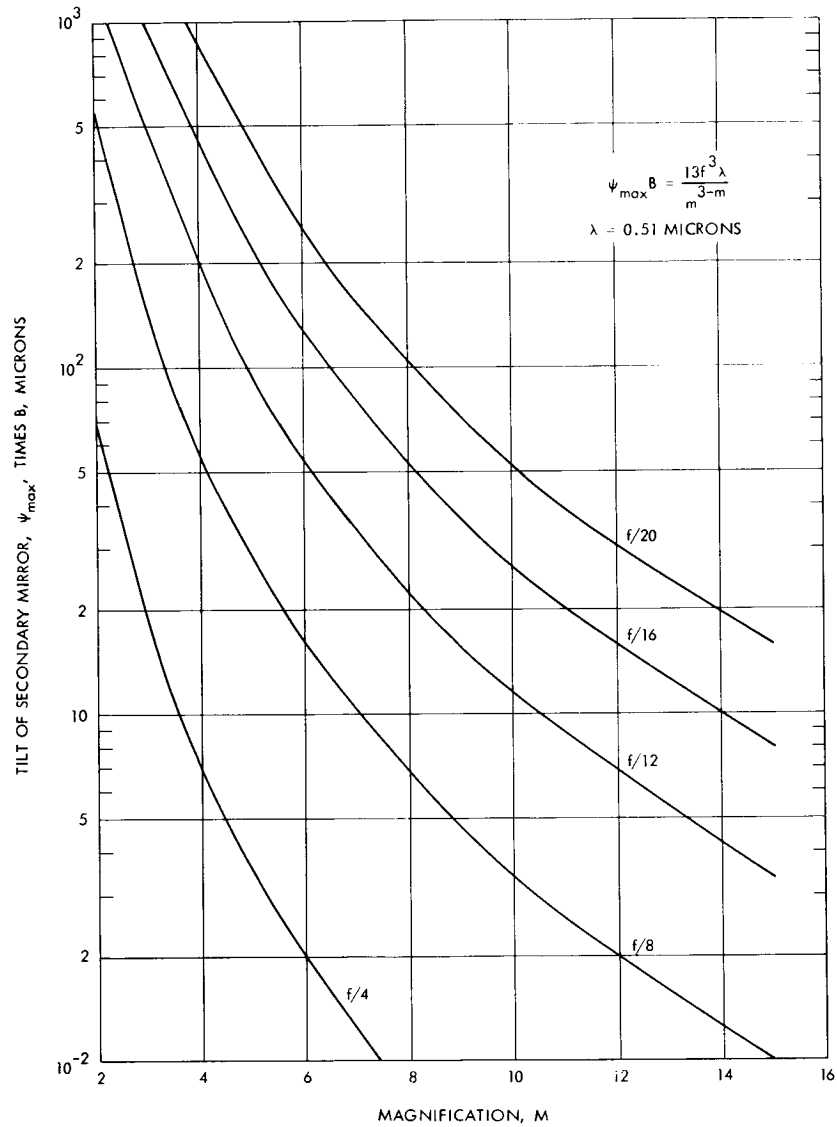


Figure E. Tolerance for Tilt of Secondary Mirror for a Cassegrain Telescope

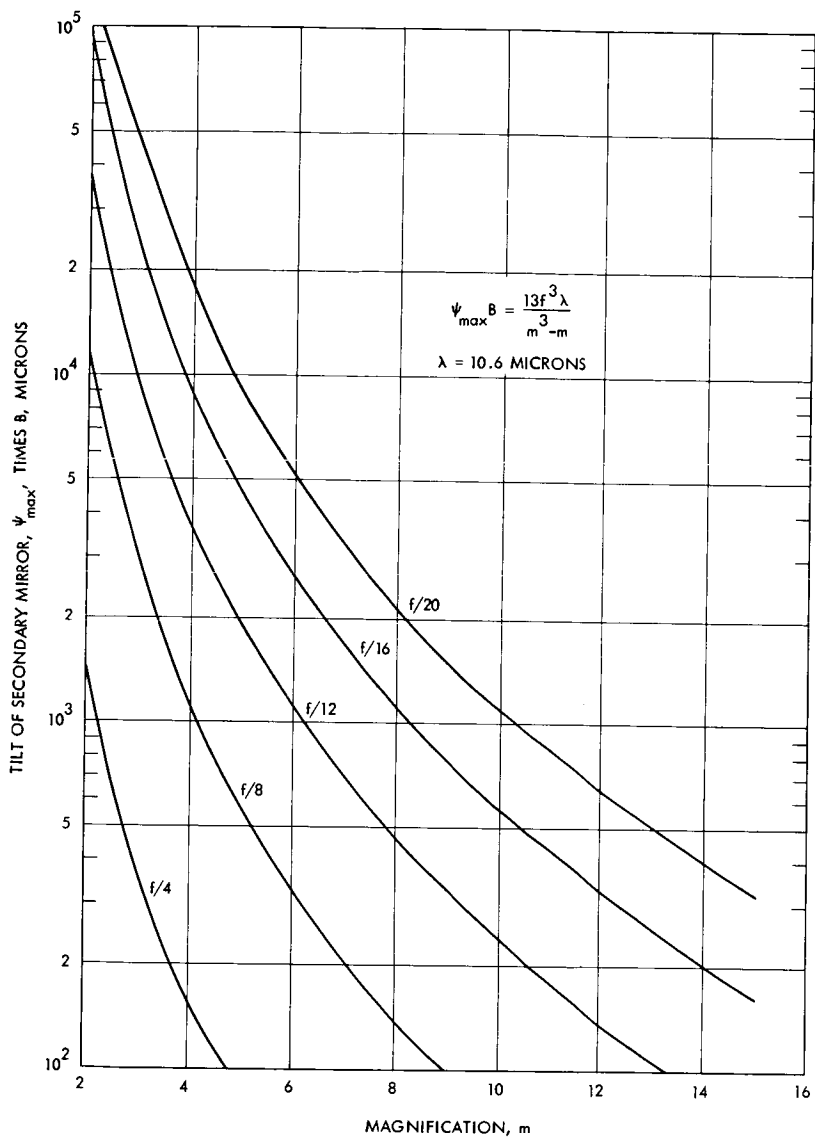


Figure F. Tolerance for Tilt of Secondary Mirror for a Cassegrain Telescope

ALIGNMENT TOLERANCES

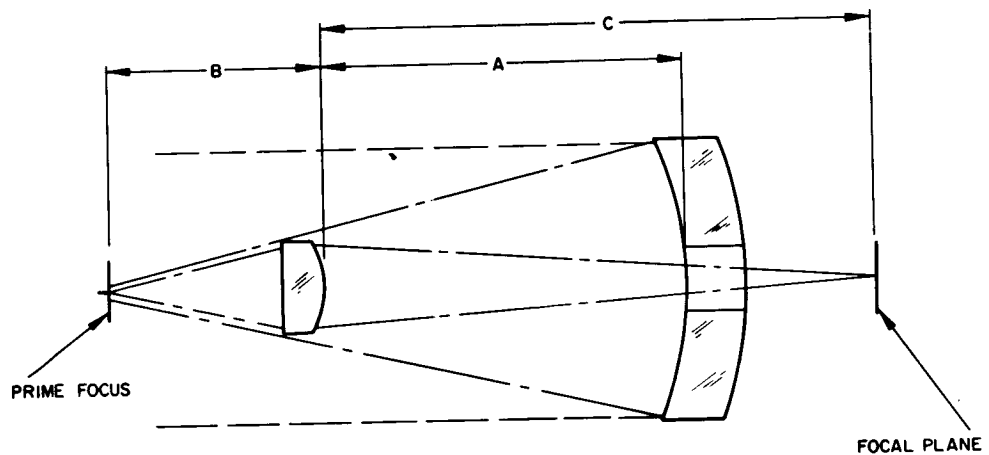


Figure G. Components of a Cassegrain System

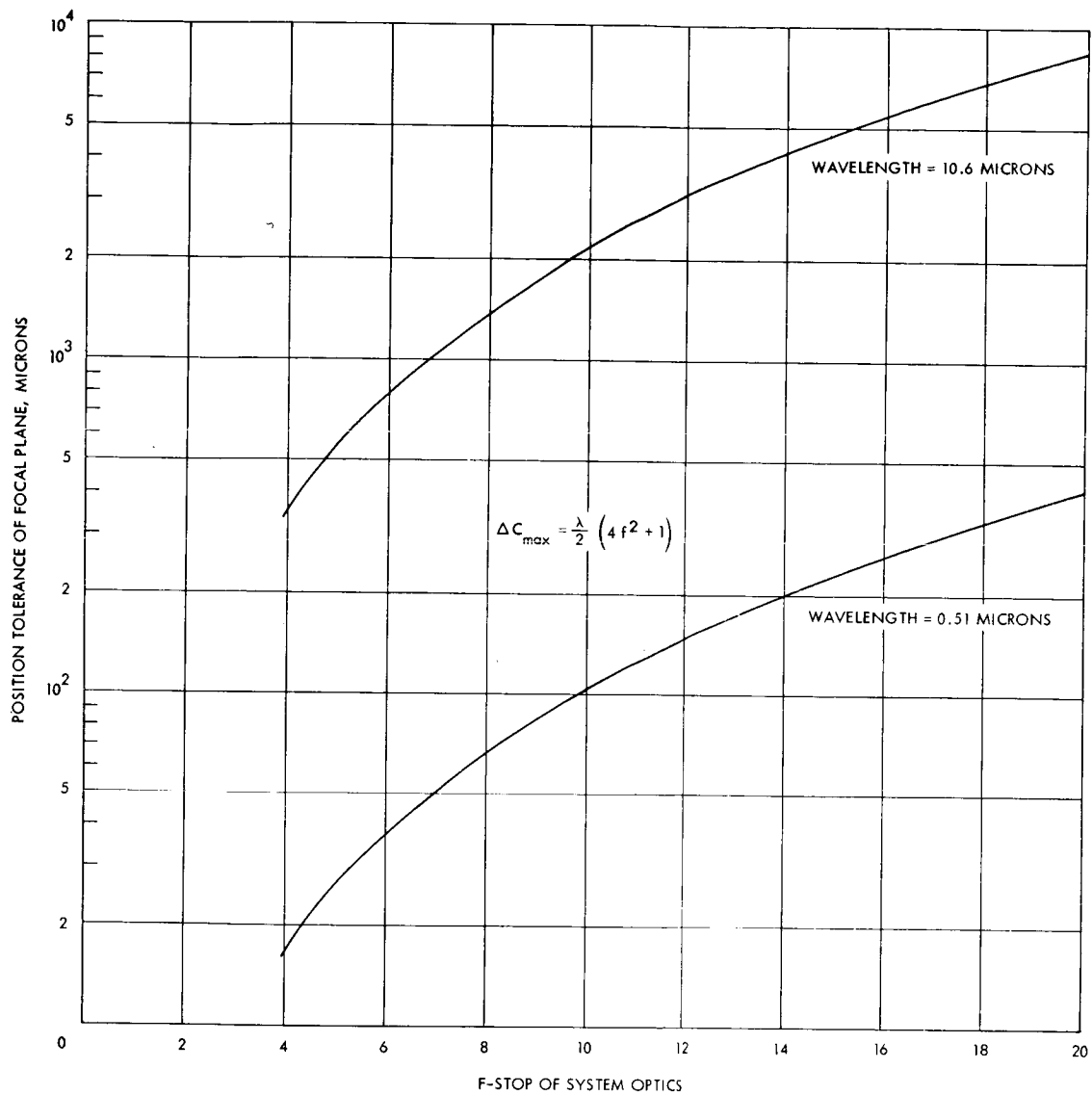


Figure H. Tolerance on Position of Focal Plane for a Cassegrain Telescope

ALIGNMENT TOLERANCES

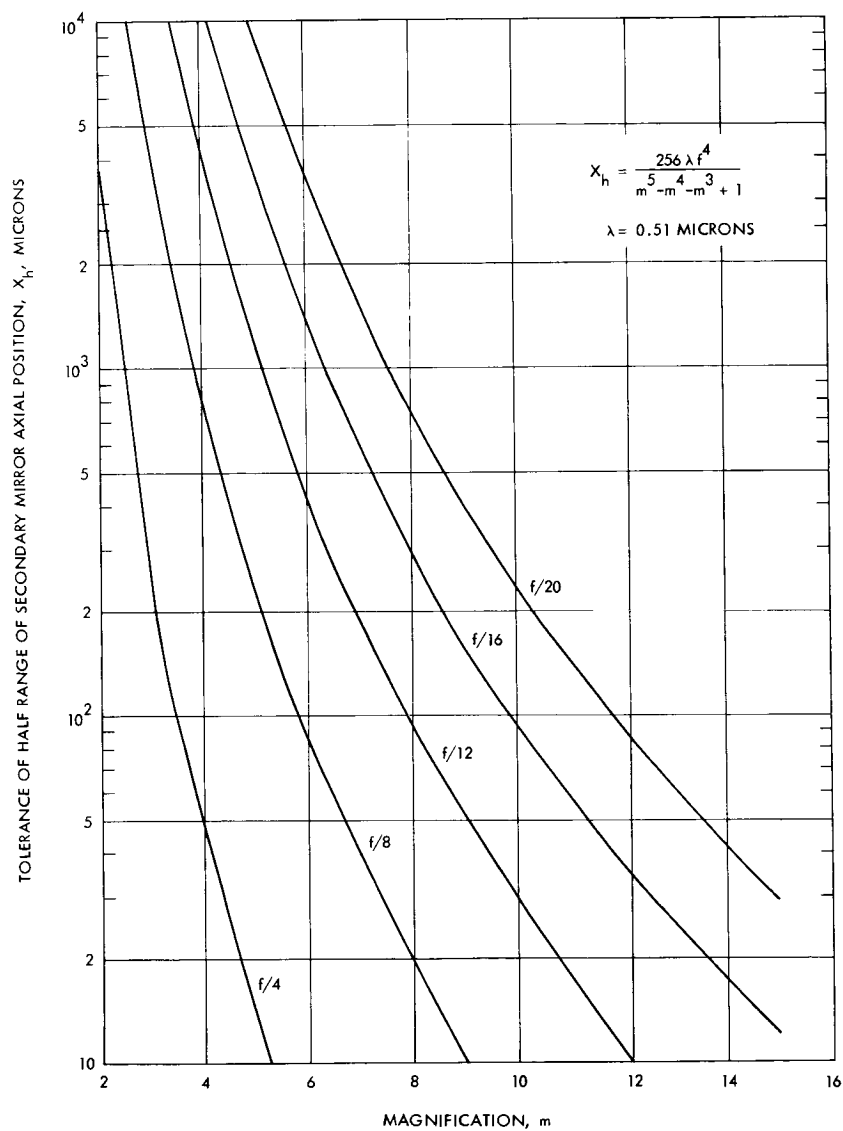


Figure I. Half Range of the Secondary Mirror Axial Position for a Cassegrain Telescope

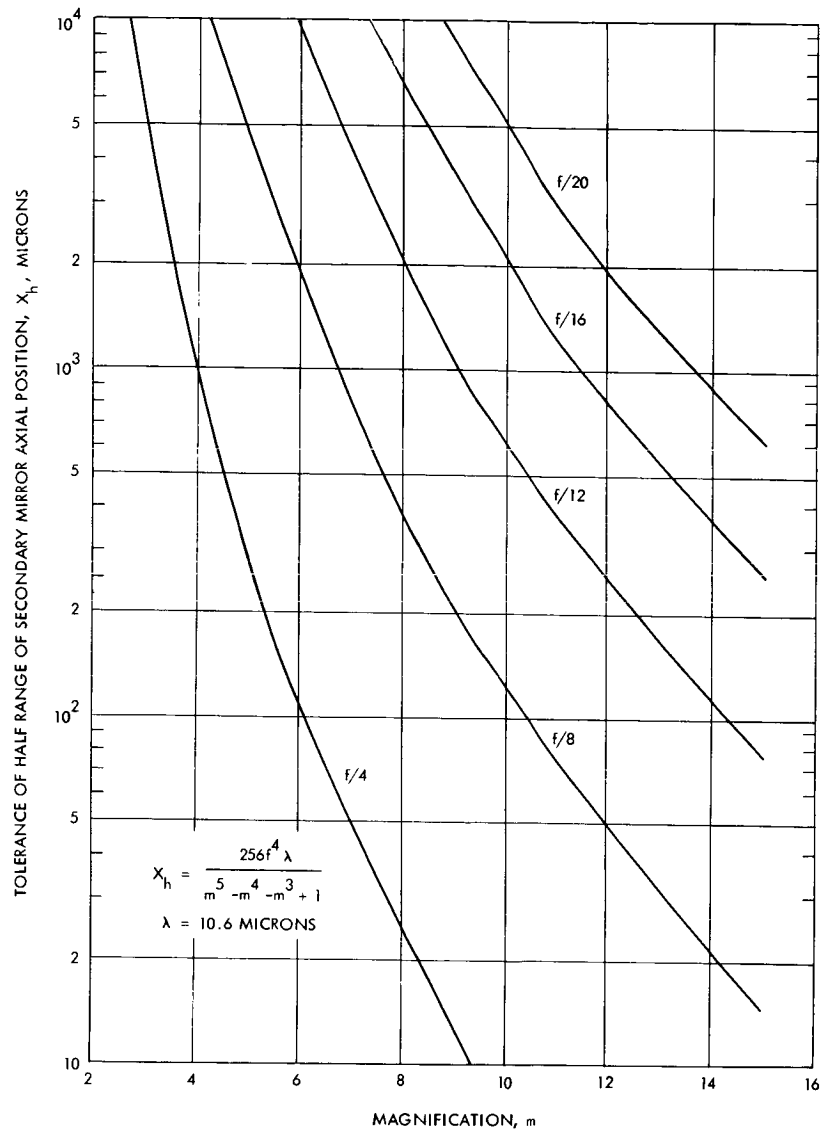


Figure J. Half Range of the Secondary Mirror Axial Position for a Cassegrain Telescope

SPEED OF OPTICAL CONFIGURATION

Higher speed optics (low f-number) are desirable for space configurations from the point of view of inertia, weight, rigidity and cost.

For a high performance spaceborn optical system, the addition of any weight to the system adds greatly to the cost. Thus, first consideration is given to lightweight optical mirror design which may have large apertures, and systems which can be built around such mirrors.

In a reflecting system, the primary mirror focal ratio of "speed" is one of the more difficult parameters to evaluate clearly since there are several trade-offs that must be made simultaneously. As far as optical manufacturability is concerned, it is easier to make a "slow" (large f-stop or ratio of focal length to diameter) paraboloid by the conventional optical processes. As the speed of a paraboloid is increased, the deviation from spherical becomes greater and optical manufacturing difficulties increase very rapidly. Testing also increases in difficulty with very "fast" (shorter focal length) mirrors. But the overall performance penalties for using a slow primary are severe. The system becomes much longer, has greater weight and inertia and is harder to hold rigid. There are also difficulties in obtaining large field coverage without excessive obscuration. Some of the same types of tradeoffs affect the design of conventional astronomical telescopes, since the use of a slow primary increases the size and weight of the telescope and dome. It is fairly well established that faster the primary mirror, the lower the overall cost of an astronomical telescope despite the increased time required to figure the primary mirror.

Under the pressure of these factors, there have been some recent improvements in the techniques of making and testing large aspherics and additional work is continuing. Probably the most significant technique makes use of an extremely accurate turntable to generate an aspheric curve with an absolute minimum of astigmatism. This allows the manufacture of faster, higher quality mirrors than were heretofore possible. Thus, large mirrors (approximately 1 meter) of $f/2$ or $f/1.5$ are now quite feasible.

THERMAL EFFECTS ON OPTICAL CONFIGURATIONS

Lateral and front to back thermal gradients are considered for the primary mirror of an optical system and a means for accommodating temperature effects is suggested.

A critical area in the design of a large optical systems for space use is that of thermal control. The effect of variations in temperature, with non-uniformity in temperature distribution, will be to distort the optical system. Both the structure and the elements themselves will be distorted. However, an automatic alignment system can be provided to compensate for the effects on the structure. Thus, it is the effect on the largest element (the primary mirror) that is of greatest concern.

The effects of a change in average temperature on a large lightweight mirror have not been determined. If the effect was a uniform expansion, there would be no change other than a change of the focal length. However, the actual mirrors are not perfectly uniform. For instance, beryllium sandwich mirrors are coated with nickel that has an expansion coefficient close to but not identical with that of beryllium. This produces a bi-metal effect if the nickel is not of the same thickness on front and back. Even fused silica sandwich mirrors have some impurity content and slight devitrification that would cause distortion with a change in ambient temperature.

The effects of non-uniformities in temperature can pose an extremely complex problem unless some simplifying assumptions are made. The assumption will be made that the temperature distribution in primary mirror can be approximated by linear gradients. The practical necessity of thermally isolating the mirror will result in any change taking place slowly and will result in the actual temperature distribution being reasonably close to linear.

A linear front-to-back gradient will cause a bending of the mirror such that the change in focal length, ΔF , will be given by:

$$\Delta F = \frac{2F^2 n \Delta T}{t}$$

where

F = focal length

n = coefficient of expansion

ΔT = temperature difference, front to back of mirror

t = thickness of the mirror

The first order effect can be removed by focussing, but the accompanying higher order effects will determine the actual limit of permissible gradient. This gradient is controlled both by the thermal isolation (with control of shielding temperatures) and by changing the effective

cross-sectional area of the core of the sandwich mirror to provide sufficient area for conduction. In this regard, if a large cross-sectional area is required, there is a heavy weight penalty.

Control of lateral gradients in the mirror will require careful design of thermal shields. A linear lateral temperature gradient results in wavefront distortion given by:

$$\delta = \frac{n \Delta T D^2}{16 F}$$

where

δ is the total wavefront distortion

n is the coefficient of expansion of the mirror material

D is the diameter of the mirror

F is the focal length of the mirror

ΔT is the difference in temperature, edge to edge along the meridian in the direction of the gradient

The curvature of the mirror varies with the position along the meridian in the direction of the gradient. The result is an elongated image in this direction. Applying this equation and a quarter wave distortion limit to large mirrors of reasonable focal lengths, results in a rather small temperature gradient tolerance; particularly for beryllium mirrors. However, beryllium, because of its high thermal conductivity, will handle a greater heat flow without exceeding the tolerance than will fused silica. Beryllium also has a peculiar property at cryogenic temperatures that is most valuable. Below 70°K the coefficient of expansion is essentially zero.

Two causes for thermal gradients in the mirror are the heat within the surrounding structure and the differences in radiation to the various parts of the front face from the exterior. Both can be controlled by shielding, although the latter source is less controllable.

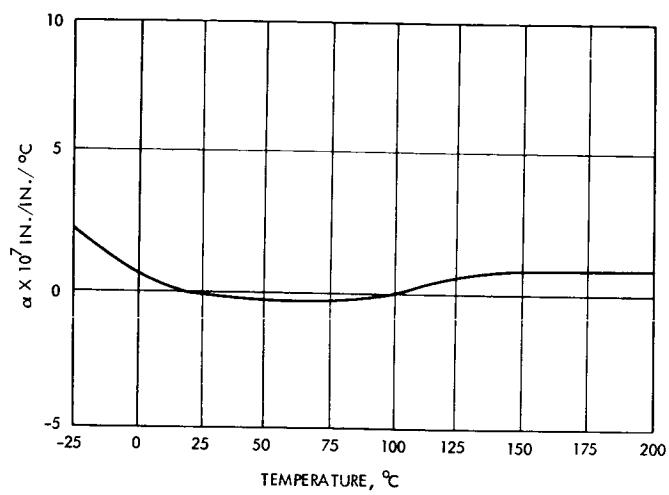
Many optical systems have been built to accommodate variations in ambient temperature. Some very ingenious mechanisms have been tried, often employing invar (low expansion) rods. However, a very simple solution is often possible. In this simple method a passive means of compensation is employed. If the optical system and supporting structure can all be made of materials having the same or nearly equal coefficients of expansion, then the optical system will only change in scale. That is, the image plane will always remain in focus but with a slightly different magnification or equivalent focal length. Beryllium optical components can often be supported by a beryllium structure to achieve this result. Care should be taken to design around beryllium's notch sensitivity. Glass components can be supported by titanium or 400 series stainless steels to achieve the same effect. Fused silica or Cer-Vit (see next topic) components would require an invar flotation system to accomplish the same result.

LOW TEMPERATURE COEFFICIENT MATERIAL, CER-VIT

Cer-Vit has a very low temperature coefficient which makes it a candidate material for spaceborn optical apertures.

Cer-Vit, similar to fused silica, has a peculiar coefficient of thermal expansion. It varies, but does not exceed $2 \times 10^{-7}/^{\circ}\text{C}$ over the range $0 - 300^{\circ}\text{C}$. The null points of the expansion coefficient curve can be shifted back and forth in the initial processing to a marked degree. A typical expansion curve is shown in the Figure.

Owens-Illinois have made mirror blanks of Cer-Vit up to 41 inches diameter to date. Segmented mirrors from 60 to 72 inches have been made. Cer-Vit is melted and formed as a glass, then by treatment is converted to a partially crystalline body. The most valuable property of Cer-Vit is its low coefficient of expansion - adjustable over a wide temperature region. It can be cored by casting or by subsequent grind-out. Present cost of a six-inch diameter blank one inch thick is \$65.00. Cer-Vit will also transmit to 4 microns for infrared uses, but exhibits some back scatter. It is both harder and stiffer than fused silica but has slightly less thermal shock tolerance and a lower maximum service temperature.



A Typical Expansivity Curve for Low Expansion Cer-Vit Mirror Material

TRANSMITTING AND RECEIVING APERTURES

Optical Frequency Apertures – Weight and Cost Relationships

	Page
Weight and Cost of Beryllium Mirrors	318
Weight and Cost for Fused Silica Mirrors	322
Weight and Cost Burden Relationships	324

WEIGHT AND COST OF BERYLLIUM MIRRORS

Weight and cost estimates are given for beryllium mirrors and for mirrors plus structure as a function of mirror aperture.

Mirrors Weight and Cost

Beryllium mirrors of high quality have been made in sizes up to 24 inches. Larger sizes have been made with lower quality. At present, there is no reason to believe that there is any basic size limitation for properly designed, high quality beryllium mirrors. The present beryllium pressing capacity limits sizes to about 7-feet in diameter without additional expenditures for facilities.

A cored center, sandwich construction, for beryllium mirrors is sometimes used to save weight and retain a very high bending stiffness. High bending stiffness helps resist bimetal bending caused by differential thermal expansion of the beryllium and the electroless nickel coating that is applied to these mirrors to provide a polishable surface. This effect can be minimized by coating both sides of the blank. The high bending stiffness also helps in resisting the vibration environment of the rocket launch.

An approximate curve relating beryllium mirror weight to aperture is included as Figure A, based on sandwich construction of reasonable proportions. A curve giving the approximate cost of beryllium mirrors is shown in Figure B.

Mounted Mirrors Weight and Cost

The primary mirror cell weight depends strongly on its orientation during launch relative to the steady acceleration and shock loads. Beryllium mirrors would require less weight for launch support than fused silica, making the total weight saving using a beryllium mirror considerably greater than the savings of the blank weight alone.

The weight range for the mechanical and optical structure of the telescope proper is shown in Figure C. These are for a telescope using beryllium optical components given as a function of aperture.

The cost range for the mechanical and optical structure of the telescope proper is shown in Figure D. These costs are for a telescope using beryllium optical components as a function of aperture.

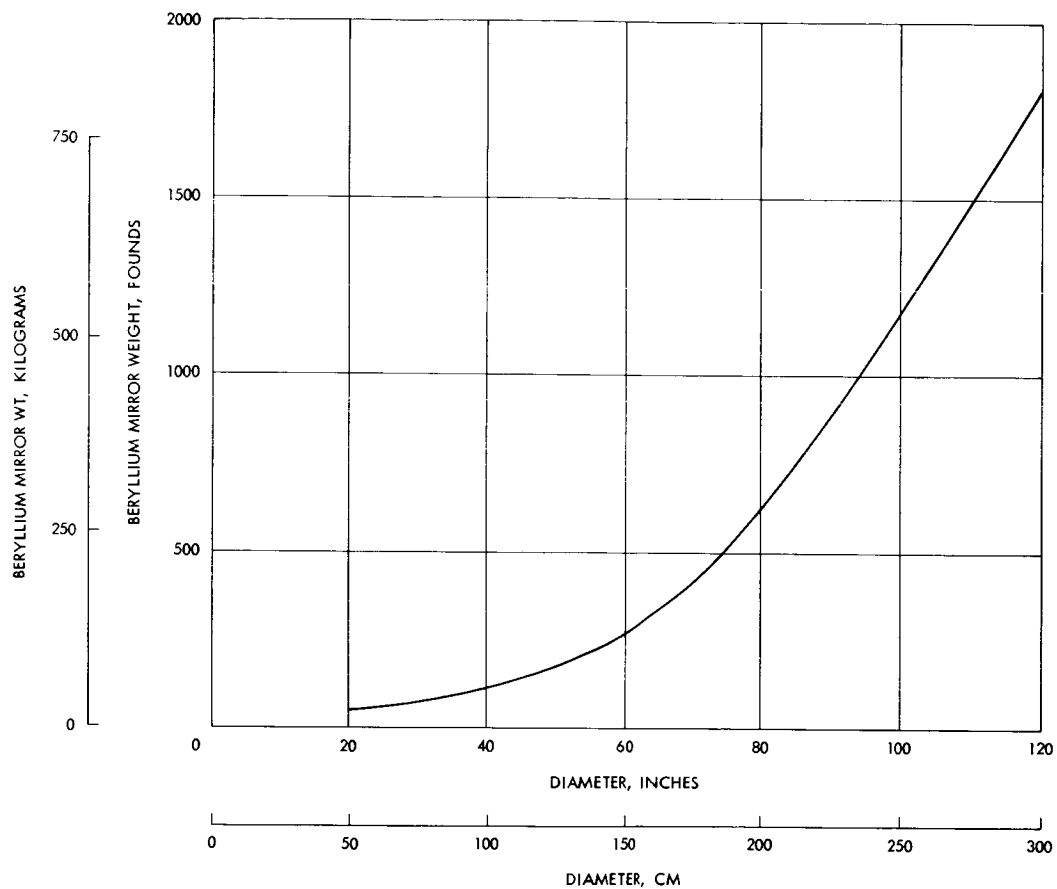


Figure A. Approximate Mirror Weight of Beryllium Sandwich Mirrors

WEIGHT AND COST OF BERYLLIUM MIRRORS

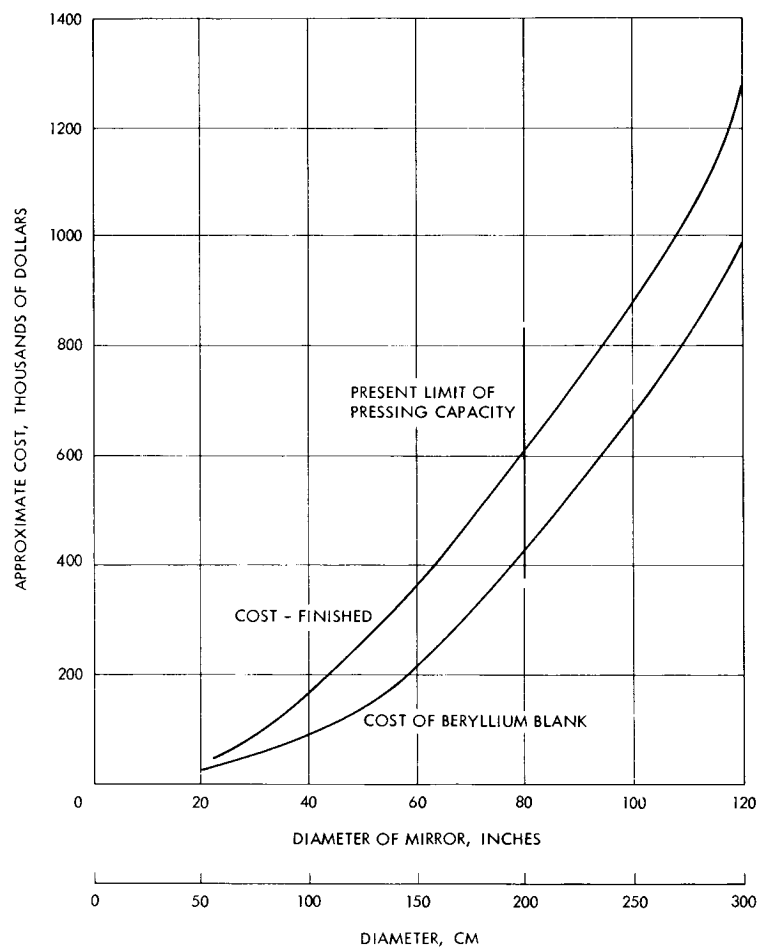


Figure B. Approximate Cost of Beryllium Sandwich Mirrors

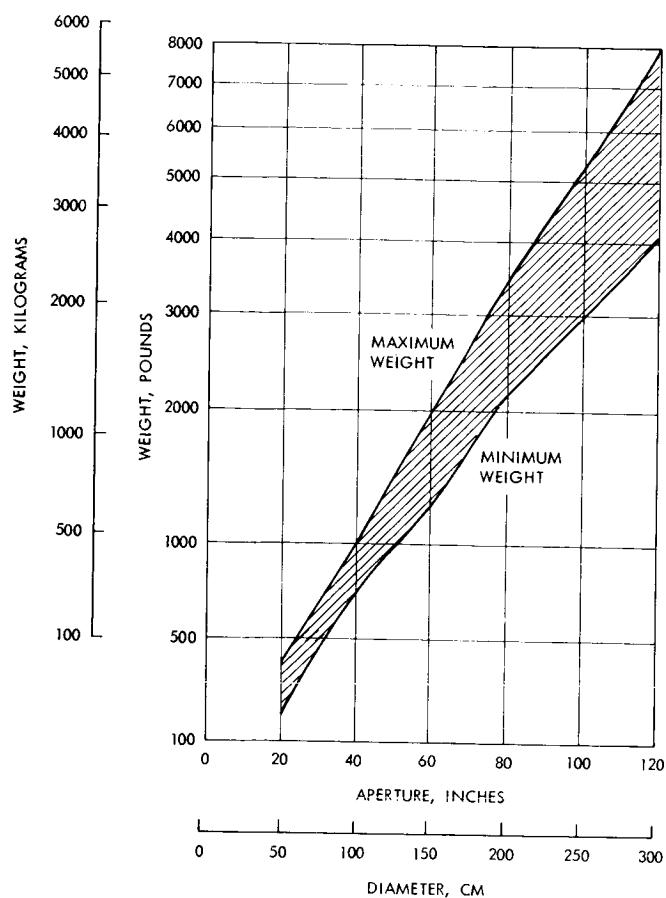
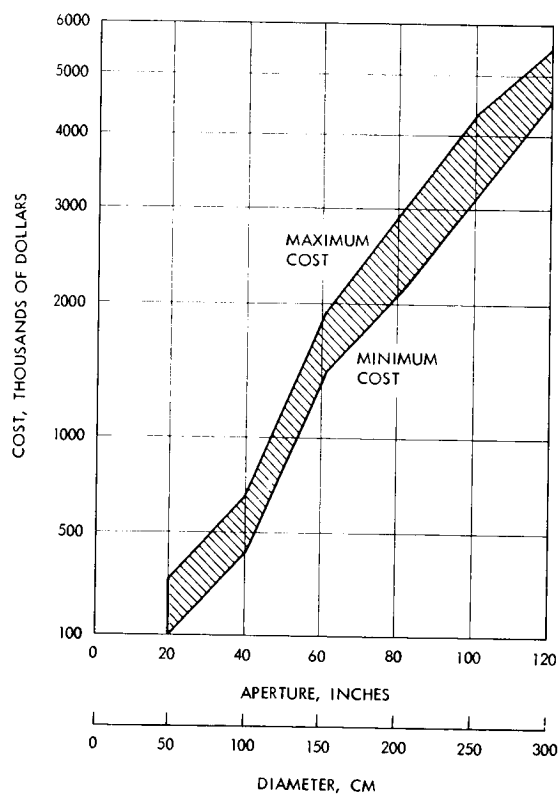


Figure C. Preliminary Weight Estimate for the Telescope Mechanical and Optical Structure Using Beryllium Optical Components

Figure D. Preliminary Cost Estimate for Mechanical and Optical Structure Using Beryllium Optical Components



WEIGHT AND COST FOR FUSED SILICA MIRRORS

Weight and cost estimates are given for lightweight, cored center silica mirrors.

Corning Glass Works have made lightweight, cored center, sandwich type, fused silica mirror blanks in sizes up to about 45 inches in diameter. Their present capacity to make bowls of fused silica limits them to about 80-inch sizes without making face plates by fusing smaller pieces together, a procedure that would require additional development work.

General Electric is developing a procedure to make lightweight blanks that would involve cutting the face plates and center core from solid blocks. These are made by fusing many hexagonal ingots together.

The approximate weights of fused silica mirror blanks are plotted against size in Figure A for normal proportions. A similar curve giving the approximate cost of fused silica mirrors is shown in Figure B.

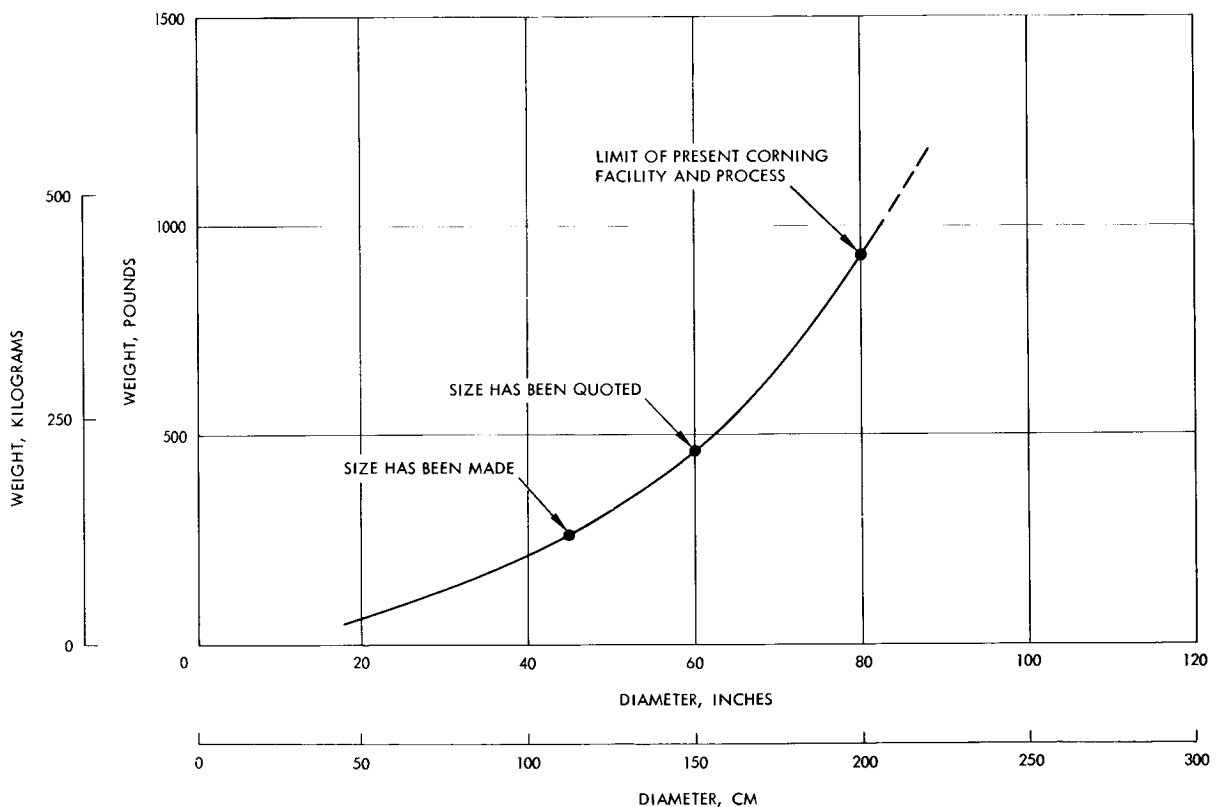


Figure A. Approximate Mirror Weight of Fused Silica Sandwich Mirrors

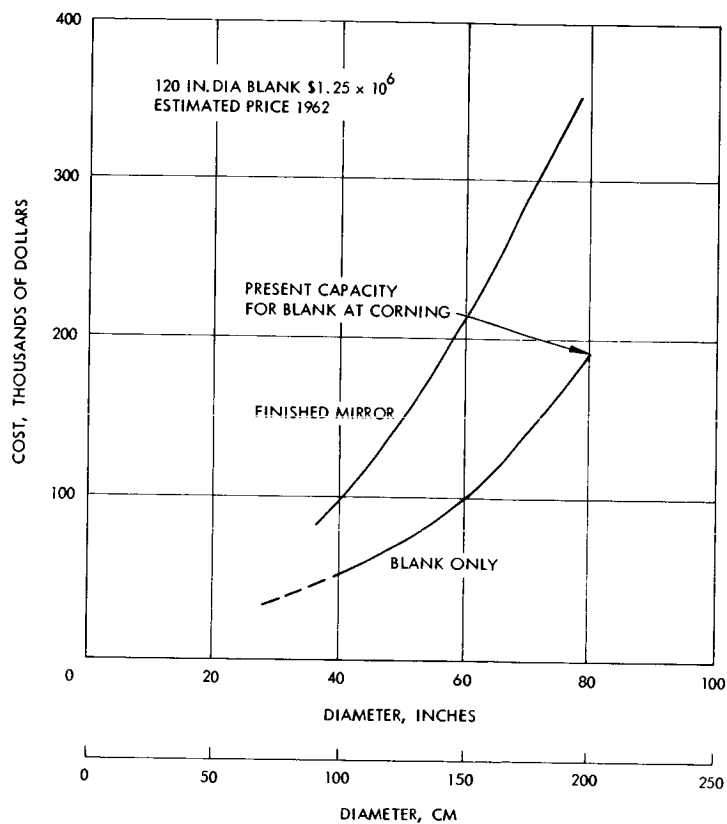


Figure B. Approximate Cost of Fused Silica Sandwich Mirrors

WEIGHT AND COST BURDEN RELATIONSHIPS

The relationships between weight and aperture diameter and cost and aperture diameter are modeled in a form suitable for the Communications System methodology.

The previous two topics gave estimate bounds for the cost and weight of optical assemblies as a function of aperture diameter. These may be modeled using the cost and weight models of the communications system methodology discussed in Volume II of this final report. The model used for weight and cost are as follows. The weight of the transmitting aperture is given by the relationship

$$W_{d_T} = K_{d_T} (d_T)^{n_T} + W_{KT} \quad (1)$$

where

d_T = the transmitting aperture diameter

K_{d_T} = a constant relating transmitting antenna weight to transmitting aperture diameter

W_{KT} = transmitting antenna weight independent of transmitter aperture diameter

n_T = a constant

and the aperture cost is given by

$$C_{\theta_T} = K_{\theta_T} (d_T)^{m_T} + C_{KT} \quad (2)$$

where

K_{θ_T} = constant relating transmitter antenna fabrication cost to transmitter aperture diameter

m_T = a constant

d_T = transmitting antenna diameter

C_{KT} = transmitter antenna fabrication cost independent of transmitter diameter

Equation 1 and 2 refer to the transmitting aperture cost and weight. These can be made to refer to the receiver cost and weight by replacing the "T" in the subscripts with "R".

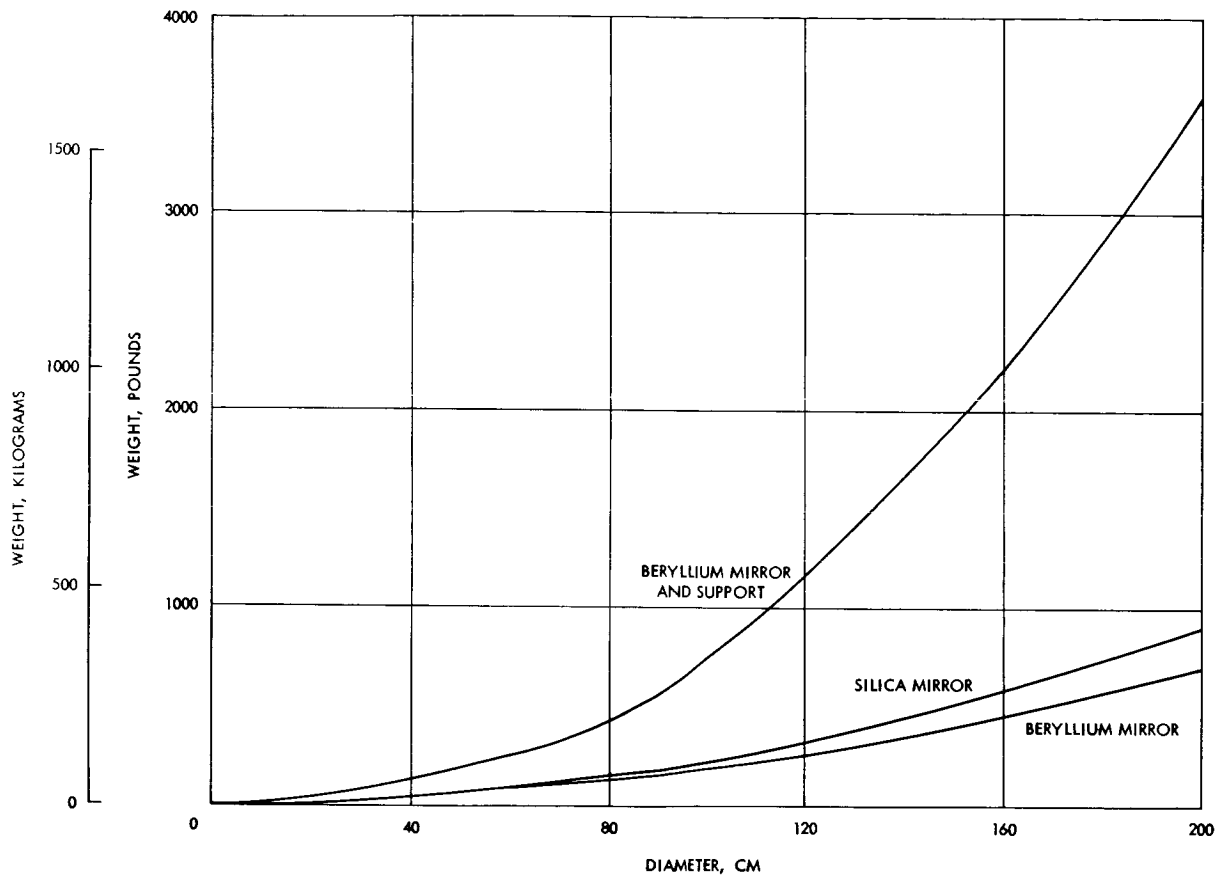


Figure A. Weight Burden Relationships for Optical Apertures

Transmitting and Receiving Apertures
Optical Frequency Apertures – Weight and Cost Relationships

WEIGHT AND COST BURDEN RELATIONSHIPS

The values for the burden constants, based on the previous two topics, are given in Tables A and B. Equations (1) and (2) are then plotted in Figures A and B.

Table A. Weight Burdens for Optical Apertures

Burden Constant	Mirror Only		Mirror and Support Structure
	Beryllium	Silica	Beryllium
W_{KT}, W_{KR}	5	10	25
K_{dT}, K_{dR}	0.0057	0.0074	0.0303
n_T, n_R	2.2	2.2	2.2

Table B. Cost Burdens for Optical Apertures

Burden Constant	Mirror Only		Mirror and Support Structure
	Beryllium	Silica	Beryllium
C_{KT}, C_{KR}	25,000	20,000	40,000
$K_{\theta T}, K_{\theta R}$	15	8.75	72
m_T, m_R	2	2	2

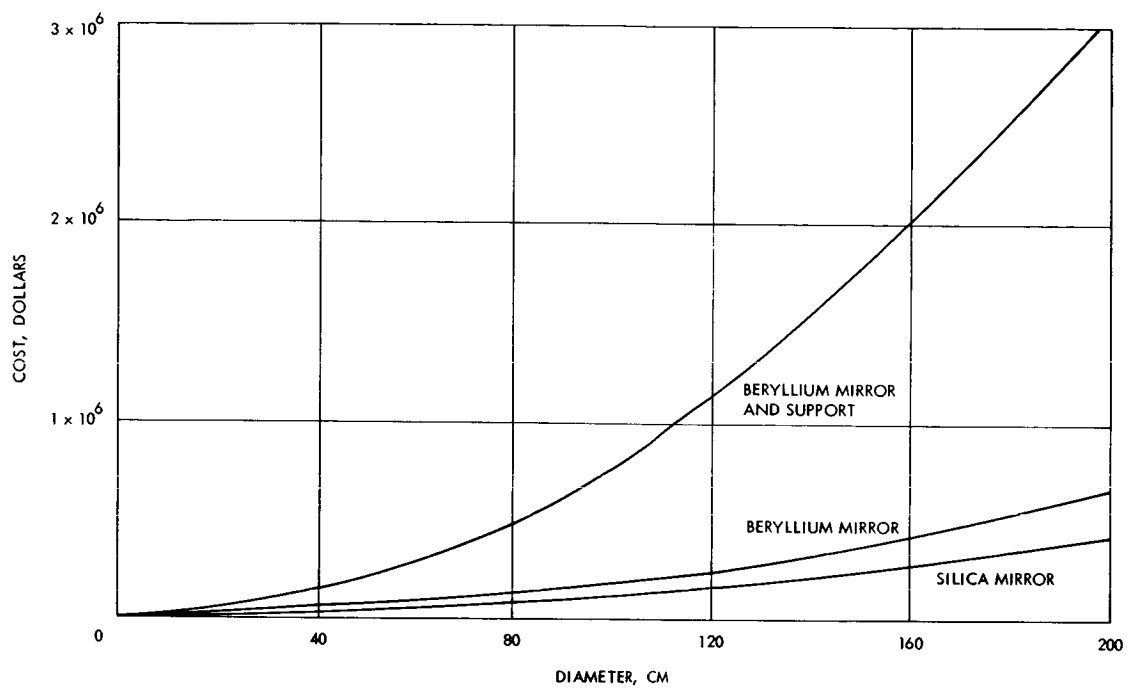


Figure B. Cost Burden Relationships for Optical Apertures

PART 5 – ACQUISITION AND TRACKING

Section	Page
GENERAL ACQUISITION AND TRACKING SYSTEM CONSIDERATIONS	334
Mission Associated Considerations	334
Receiver Location Considerations	346
ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS	353
The Tracking Subsystem – Introduction	358
Acquisition	380
Detection Theory	386
Angle Noise Error in Optical Tracking Systems	396
COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS	424
Attitude and Tracking Sensors	424
Attitude Control Techniques	440
Passive Attitude Control Techniques	444
Active Attitude Control Devices	452
Burden Relationships	466

INTRODUCTION AND SUMMARY

Acquisition and tracking will include subtopics of system requirements, performance analysis, tracker functions, tracking performance measured on a probability basis, and component burden relationships.

The advantage promised by laser communication is gained through the use of very narrow optical transmitter beamwidths allowing transmitter power requirements to be correspondingly small. This, in turn, requires very accurate pointing of the laser transmitter. Laser system pointing requirements are sufficiently more severe than microwave that this section deals mainly with optical acquisition and tracking.

The optimum acquisition and tracking system for a particular communication task depends on a host of mission parameters. For instance the transmission of data from a deep space vehicle (DSV) to an earth base receiver requires the spacecraft orientation with respect to a reference coordinate system to be determined. Then the spacecraft must be oriented so as to acquire a cooperative laser beacon at the receiver site. The ground beacon must be pointed to illumine the spacecraft taking proper account of atmospheric irregularities. An optional intermediate step is to have the ground based optical tracker acquire the spacecraft (by means of a broad beam on-board beacon) refining the knowledge of its position so that the ground beacon beamwidth may be narrowed. After the DSV transmitter has been pointed so that it irradiates the earth receiver, the tracking system must continue to point with sufficient accuracy that contact is maintained. The tracking system may be open or closed loop, depending on whether error information required to keep the transmitter beam properly oriented is generated at the receiver or at the transmitter. In either case, the acquisition and tracking system must take into account such factors as:

Relative motion between the tracker and the target.

Coordinate reference errors.

Signal propagation delays.

Aberration effects due to relative acceleration of the transmitter and receiver.

Perturbations of the spacecraft.

Atmospheric effects.

This study of the acquisition and tracking problem begins by considering the requirements imposed on the system by the peculiarities of the mission and the receiver location. Next, the acquisition and tracking system performance is analyzed in terms of these constraints and the system parameters which contribute to the overall pointing error. Then the various functions performed by the general (typical) optical tracker are delineated and a mathematical description of the performance of these functions in the presence of noise is presented. In particular, system performance measures such as probability of detection, probability of

acquisition, probability of false alarm, loss rate, tracking accuracy, etc., are established in terms of the system parameters such as beacon beam-width, power, receiver FOV, background noise, dark current noise, etc.

Various modes of tracking implementation are discussed. The conventional forms of position encoding which are treated are the following: pulsed beacon (monopulse) system utilizing a quadrant photomultiplier and a CW beacon using pulsed position modulation (PPM), amplitude modulation (AM) and frequency modulation (FM).

Finally, the state of the art and burden of components which significantly affect acquisition and tracking system performance, such as star sensors, sun sensors and attitude control and stabilization devices, is surveyed.

GENERAL ACQUISITION AND TRACKING SYSTEM CONSIDERATIONS

Mission Associated Considerations

Signal Propagation Delays	Page 334
Relative Motion Between Transmitter and Receiver	338
Coordinate Reference Frame Error	340
Manned Versus Unmanned Vehicles	342

SIGNAL PROPAGATION DELAYS

Signal propagation delays over deep space distances require that laser beams be "pointed ahead," typically 0 to 100 microradians.

The acquisition and tracking system configuration is determined by the mission constraints and by the receiver location. Mission related constraints include range, system life, angular motion rates, reference coordinate system errors, and torques produced by equipment or personnel on-board the vehicle which perturb it. Receiver location considerations include atmospheric turbulence and attenuation associated with a ground base, and environmental and operation difficulties posed by a lunar or satellite base.

In the case of earth communication to and from a deep space vehicle (DSV) where significant propagation times are involved, appropriate considerations must be taken of "point ahead" angles. That is, the earth and space transmitters must each be directed toward the point in space where the respective receivers will be when the beams arrive. Thus, a lead angle, α_T , is required for both the earth transmitter and the DSV transmitter. The line-of-sight (LOS) vectors are defined in the figure. These vectors will be called the earth-to-deep space vehicle (E-DSV) and the deep space vehicle-to-earth (DSV-E) vectors, respectively. The E-DSV line-of-sight vector is directed to the DSV from a point in space where the earth was when the received light left the earth on its way to the DSV. Similarly, the DSV-E line-of-sight vector is directed to the earth from a point in space where the DSV was when the light left the DSV. The earth's receiver and DSV's receiver must be directed along the DSV-E and the E-DSV lines-of-sight, respectively.

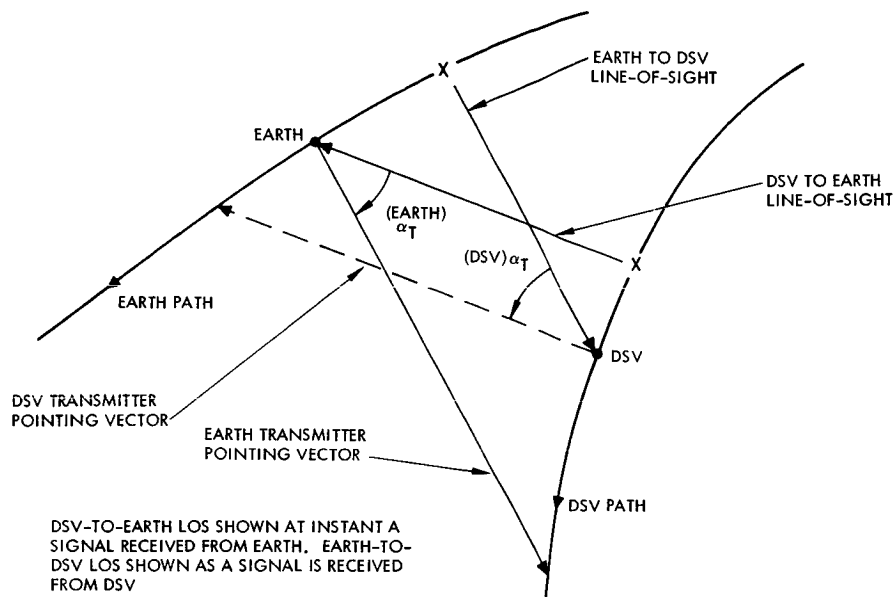
If atmospheric effects and curved propagation paths are neglected, the earth transmitter axis will be parallel to the E-DSV LOS. Likewise the DSV transmitter axis will be parallel to the DSV-E LOS. Atmospheric effects will cause the transmitted beams to be distorted and bent. This effect is only partially self compensating since the effect is different for a far field wave front than for a near field front. Thus $\alpha_T(\text{Earth}) \neq \alpha_T(\text{DSV})$ in general. This effect must be accounted for if it is desired to update the DSV lead angle from knowledge of the earth's lead angle. The earth receiver and the DSV receiver must be directed along the DSV-E and E-DSV lines-of-sight, respectively.

The position of the transmitter pointing vector relative to the receiver pointing vector may be determined if it is assumed that the lead angle, α_T , is small and that during the propagation times involved, the relative velocity angular direction (aspect angle) is constant. The length of the DSV-to-earth position vector will be approximately equal to that of the earth-to-DSV position vector. If this distance is R , and T is the round trip transit time,

then,

$$T = \frac{2R}{c} \quad (1)$$

where c = speed of light.



Earth and DSV Line-of-Sight Geometry

SIGNAL PROPAGATION DELAYS

The distance the earth station moves relative to the DSV during the round trip transit time is

$$d_E = V_T T \quad (2)$$

where V_T = the tangential velocity of the earth station relative to the DSV. For small angles, the lead angle may be written as

$$a_T = \frac{d_E}{R} \quad (3)$$

Combining equation 1 and 2 gives

$$R = \frac{cd_E}{2V_T} \quad (4)$$

Substituting equation (4) in equation (3)

$$a_T = \frac{2V_T}{c} \quad (5)$$

and it is seen that the lead angle is independent of the magnitude of the position vector, being only a function of the relative tangential velocity between the transmitter and respective receiver. Since $V_T \ll c$ for interplanetary missions in the foreseeable future, the small angle approximations made previously were valid. Lead angles will range from 0 to approximately 100 μ rad during the transfer orbit phase of the mission. For landers and orbiters, lead angle calculations must include the rotation of the planet or the orbital angular velocity. The relative tangential velocity, V_T , is not measured directly but is determined from the range rate and the angular orientation of the LOS. Typically these would be measured from earth using the optical carrier frequency or by using an auxiliary r-f system operating at DSIF frequencies. For instance, the two-way DSIF system presently achieves a range rate accuracy of ± 0.03 m/sec while doppler tracking accuracy in the 1970's will be on the order of 10^{-3} m/sec. Uncertainty in the orientation of the LOS introduces an additional error into V_T which is on the order of $V_T \Delta\theta$, where $\Delta\theta$ is the angular uncertainty. For $\Delta\theta = 0.001^\circ$ (17.5 μ rad) and $V_T = 50$ km/sec, ΔV_T due to this angular uncertainty is ≈ 0.88 m/sec. In order to achieve pointing accuracies of 0.001° it is necessary to use optical trackers rather than radio tracking.

RELATIVE MOTION BETWEEN TRANSMITTER AND RECEIVER

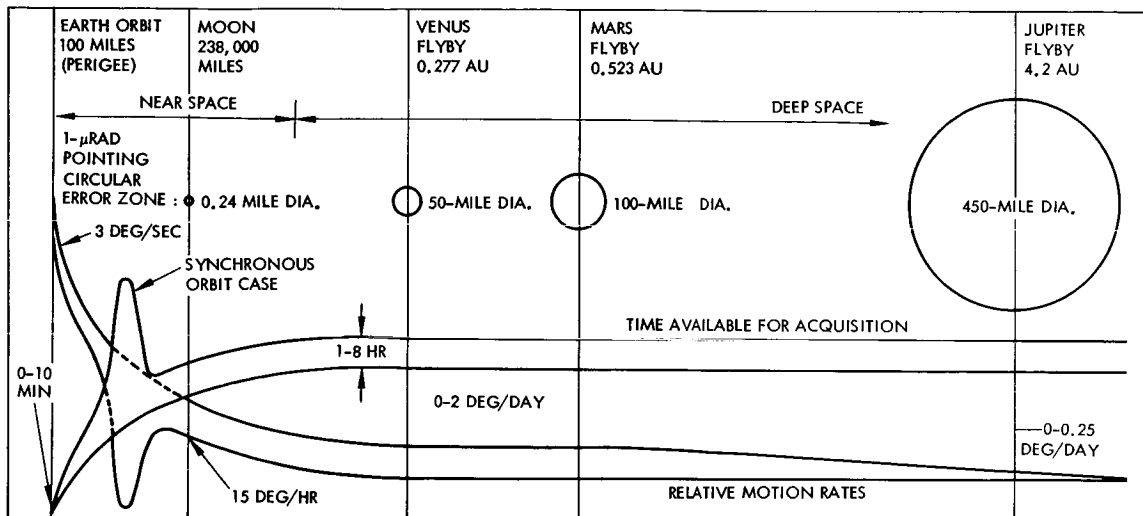
Angular rate of the LOS depends upon the range between the transmitter and receiver. Three range zones may be distinguished.

Angular line-of-sight (LOS) which must be accommodated by a space vehicle clearly are a function of the range to the receiving site from the spacecraft. Three general zones may be distinguished based upon acquisition and tracking criteria. These are near earth, near space, and deep space and are indicated in the figure.¹

High angular LOS rates such as encountered between a ground station and a near earth satellite are typically $3^\circ/\text{sec}$ for a 100 mile orbit. Such LOS rates, limit the viewing time (<10 minutes) and require more rapid acquisition. This in turn demands higher acquisition signal-to-noise ratios and wider transmitter beamwidths. At deep space ranges, LOS rates are much lower 0 to $2^\circ/\text{day}$ typically for a Mars fly-by, which is more compatible with acquisition ease and with narrow beamwidths. In addition, a slew rate requirements are appreciably higher for high LOS rates, ranging from $60^\circ/\text{sec}$ for near earth satellite to $10\text{-}20^\circ/\text{minute}$ for a deep space vehicle.

LOS angular rates for the DSV tracker are significantly different for a fly-by mission than for orbiters and landers since orbiters and landers add the rotation of the target planet to their angular rates. This means that the tracking system used on such missions must either be capable of tracking at higher rates or of biasing at the nominal rate, which is known approximately from astronomical data.

¹N.G. Lozins, "Pointing in Space," Space Aeronautics, August 1966, pp. 76-83.



Angular Rates and Available Acquisition Time

Note that there are three range zones which may be distinguished on the basis of range rate and time available for acquisition.

COORDINATE REFERENCE FRAME ERROR

Celestial coordinate reference frames have errors of the same or greater order of magnitude as the pointing accuracy of a laser system.

A fundamental limitation on the orientation of the DSV acquisition and tracking system is the accuracy with which a reference coordinate system can be specified. The available coordinate systems each have peculiarities which must be taken into consideration.

For instance, in the earth-centered equatorial reference system, one of the three orthogonal axes is defined by the line from the center of the earth to the first point of Aries which is that point where the path of the sun crosses the celestial equator from south to north on or about March 22. Currently there is no visually observable star at this location. This point moves westward along the ecliptic at the rate of 50.26 arc-sec per year as a result of the precessional motion of the earth's axis (other motions are also involved). The reference axes form an orthogonal set along this direction, the direction to the north pole, and a third direction normal to these two.

Similar problems arise in the heliocentric system reference system which uses the line from the center of the sun to the first point of Aries as an axis. Deep space acquisition and tracking systems using a celestial reference will have to compensate for these and other stellar motions noted below.

Motions to be considered include both the motions of binary stars and multiple star systems about a common center of gravity, and so-called "proper" motions, which are small changes in stars positions that increase steadily over the passage of years. (Each star has its own peculiar motion, which is called proper motion by astronomers.) For instance, Arcturus and Rigel Kent, two zero-magnitude stars have proper motions which exceed 2 and 3 arc-sec/year, respectively.

In addition to these real motions, there are apparent motions which include the parallactic displacement of nearby stars with respect to the distant stars in the background. This motion is due to the radial displacement of the earth from one side of its orbit to the other. The closer the star, the greater the parallax. The nearest star, Proxima Centuri, has a parallax of 0.763 arc-sec. Space ships venturing beyond 1 AU from the sun would observe proportionately larger values.

The effect of these coordinate errors upon an acquisition and tracking system clearly depends upon the relative size of the transmitted beam-width to these errors. In the situations where they are comparable and where long flight times are involved, it will be necessary to compensate for these errors by such means as programming pointing corrections or by searching over larger angular sectors during the acquisition.

MANNED VERSUS UNMANNED VEHICLES

Manned and unmanned missions have complementary advantages and disadvantages; in a manned vehicle, the man can correct errors in a tracker but his presence causes disturbances. The unmanned mission has fewer disturbing torques but may need redundancy to provide continuous tracking.

The advantages of a manned mission from an acquisition and tracking point of view are manifold. In conjunction with an on-board computer, man can make very accurate course and attitude correction and navigational sightings. Thus man is better capable of determining the DSV's course, attitude, and position than an earth based observer. Furthermore the astronaut can align an inertial platform reference device and make corrections for its drift. Two other important advantages of manned flight are the facts that continuous tracking and attitude control are not required as are for the unmanned mission. A manned mission has the added advantage of accurate lead angle alignment, eliminating the errors associated with the servo device necessary to the unmanned system. The lead angle may be determined by optical techniques and adjusted when desired. An additional advantage to a manned mission is the ability of the man to perform routine maintenance and repairs; and unmanned mission would require complete redundant systems for the equivalent reliability.

The primary disadvantage of a manned mission is the effect of man motion on the spacecraft, with resulting additional stringent restrictions on the tracking and pointing control system in order to avoid degradation of the tracking and pointing accuracy. For the manned mission this disadvantage may be circumvented and several advantages of the unmanned system acquired if a separate "satellite" vehicle were used for optical communication. If a failure should occur, such a vehicle would be maintained by the astronauts, the lead angle accurately reset, and its platform could be realigned. When communication is desired, the "satellite" vehicle would be separated from the mother vehicle, and maneuvered in space by remote control or tethered to the mother vehicle for convenient retrieval.

An unmanned mission will have the restraints of continuous tracking and continuous attitude control. An inertial platform reference device is not feasible for this system since drift correction would be very difficult. In place of a platform, star trackers and sun sensors could be used for attitude reference and control.

The general block diagrams for the "satellite" vehicle for the manned mission and for the unmanned mission are shown in Figures A and B, respectively. For both missions, the electromagnetic transit delay is involved twice in the DSV-E-DSV communication loop. Cross-coupling of the dynamic relationships is indicated by the dotted lines. An inertial platform and rate gyro stabilization is used on the satellite vehicle for the manned missions. Due to offset and drift problems, star trackers (or similar devices) are used for rate and attitude stabilization for the unmanned missions.

The computer shown is intended for the complex navigational and lead angle computations. However much of this burden may be removed by an earth based computer.

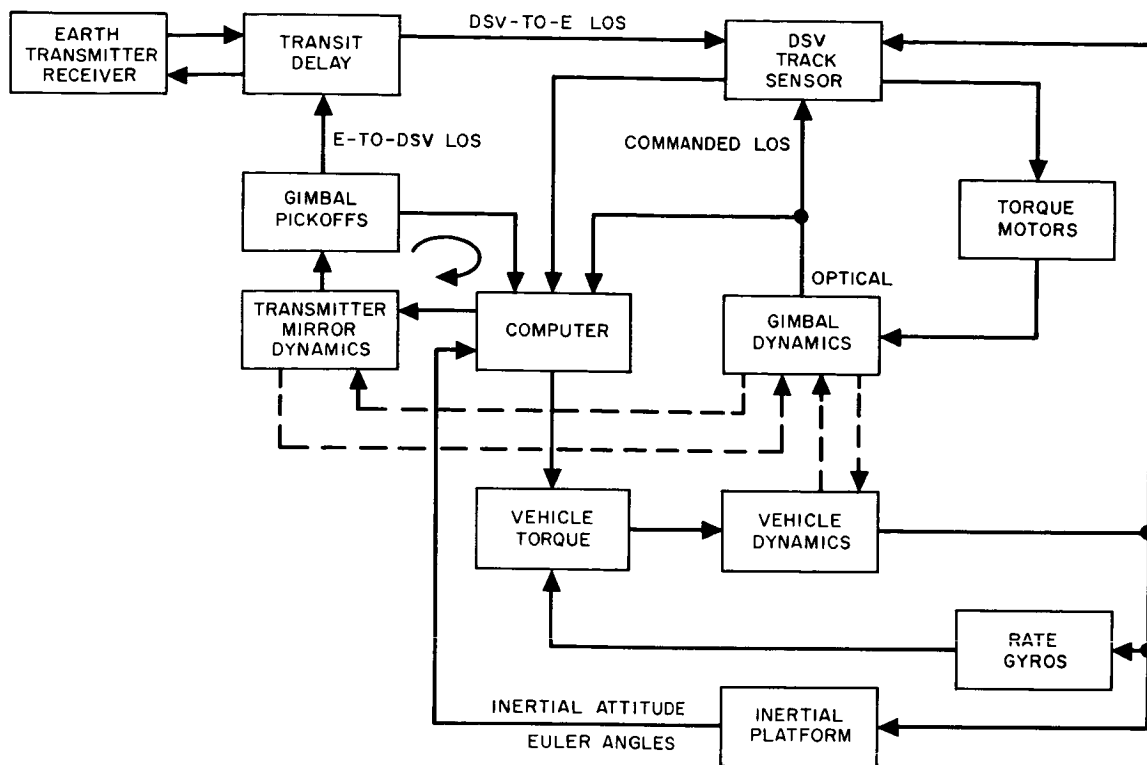


Figure A. Deep Space Vehicle (DSV) Communication Control System Block Diagram for a Manned Mission Communication "Satellite"

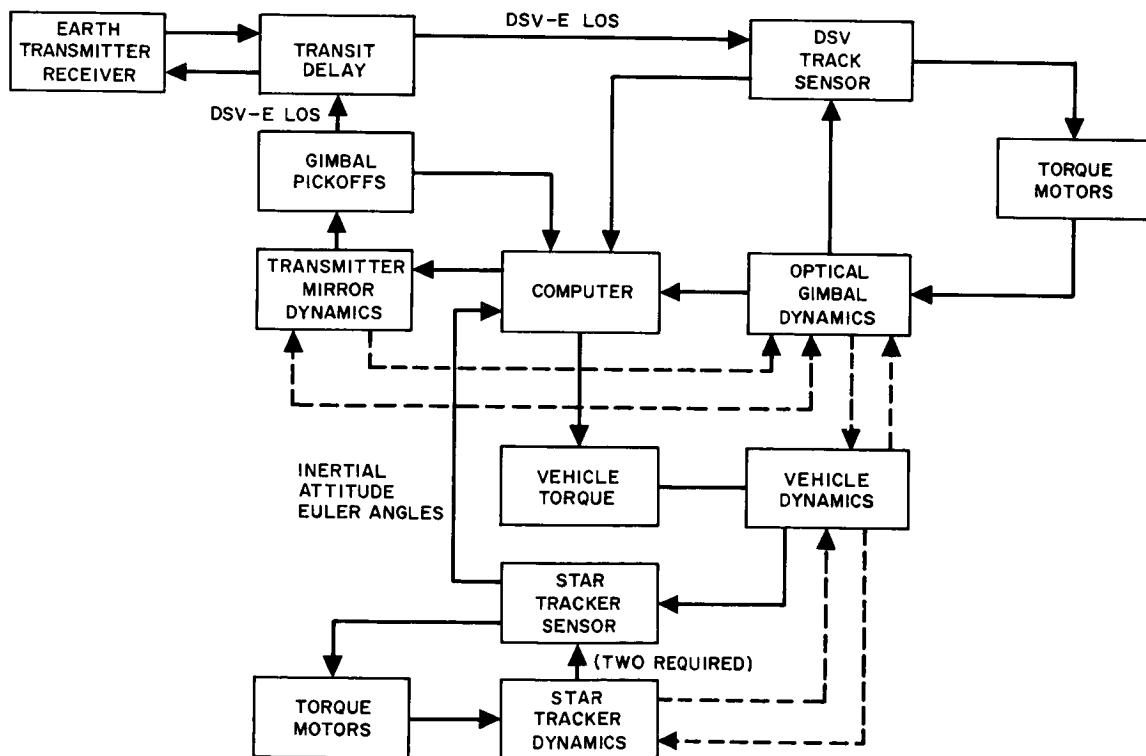


Figure B. Deep Space Vehicle (DSV) Communication Control System Block Diagram for an Unmanned Mission

GENERAL ACQUISITION AND TRACKING SYSTEM CONSIDERATIONS

Receiver Location Considerations

	Page
Earth Based Receiver Atmospheric Considerations	346
Earth-Satellite Based Receiver and Receiver Site Comparison Summary	350

EARTH BASED RECEIVER ATMOSPHERIC CONSIDERATIONS

An earth based receiver for an optical link must cope with atmospheric beam pointing effects including refraction and beam steering.

The deep space link receiver location has considerable bearing on the acquisition and tracking problem. Two possible locations, earth and earth-satellite, each have unique operational and economic advantages or disadvantages. The most significant of these will be considered briefly in this topic and the next topic.

The greatest disadvantage of an earth based receiver for deep space communications is presence of the earth's atmosphere. The first consideration is the attenuation of the optical or radio frequency carrier by absorption and scattering. Cloud coverage, fog, and/or smog may make optical communications impossible for a large percentage of the space mission time. Therefore earth locations must be found where the possibility of such interferences are minimized*.

In addition to attenuation, the atmosphere interferes with optical propagation by refraction, and by such turbulence induced effects as beam spreading, beam steering, and scintillation. Exhaustive measurements of steady state refraction have been made by astronomers. These are summarized in the table¹.

Qualitatively it should be noted that there is a fundamental difference between the fluctuations in signal level during transmission and those during reception. During reception, although diffraction at the spacecraft spreads the beam over a large portion of the earth, all the energy incident on the receiver aperture can be detected if a sufficiently large field stop (detector) is used. However, during transmission, only that portion of the beam that leaves the atmosphere in the direction of the spacecraft is used. There is only a slight spreading of the beam during passage through the atmosphere, but angular or phase disturbances are created because the plane wave front has been distorted. These disturbances may result in a large spreading of the beam after subsequent propagation. Angular divergence here, perhaps not yet affecting beam diameter because of the large initial diameter, will ultimately be the determining factor in beams spread. In antenna phraseology, the top of the atmosphere is still in the near-field region of an optical transmitter. The basic difference between transmission and reception can be summarized as follows: In transmission, cumulative phase fluctuations (which cause angular divergences) are important; however, in reception, only the cumulative amplitude fluctuations (produced by phase fluctuations near the top of the atmosphere) are significant.

Beam steering arises from time-dependent atmospheric inhomogeneities and introduces a random angular error in specifying the true direction of

*Suggestions are given in Part I of Volume IV of this report "Background Radiation and Atmospheric Propagation".

¹Perkin-Elmer Report No. 7846, Contract NAS8-11408, SPO 26471, November 1964.

Average Refraction Angle Versus Apparent Zenith Angle¹
for Visible Light

Apparent Zenith Angle (degrees)	Refraction Angle (minutes and seconds of arc)		Apparent Zenith Angle (degrees)	Refraction Angle (minutes and seconds of arc)	
0	0	0.0	70	2	35.7
5	0	5.0	75	3	30.0
10	0	10.1	80	5	13.1
15	0	15.3	81	5	46.0
20	0	20.8	82	6	26.0
25	0	26.7	83	7	15.0
30	0	33.0	84	8	19.0
35	0	35.7	85	9	40.0
40	0	47.9	86	11	31.0
45	0	57.1	87	14	7.0
50	1	8.0	88	17	55.0
55	1	21.4	89	23	53.0
60	1	38.7	90	33	51.0
65	2	1.9			

EARTH BASED RECEIVER ATMOSPHERIC CONSIDERATIONS

the earth-to-DSV line of sight vector which is superimposed on the predictable steady state refraction. This requires the earth-based transmitter to transmit over a correspondingly wider angle to insure that the DSV is illuminated.

The amount of beam deflection which can be expected depends directly on the strength of the turbulence. As the turbulence goes from weak to strong, beam steering angles typically vary from ± 1 to $\pm 15 \mu\text{rad}$ (rms). Very strong turbulence can produce deflections on the order of $\pm 50 \mu\text{rad}$ (rms). In addition, the apparent direction of the LOS will vary in time due to quivering at frequencies on the order of a hundred cps or less. These (relatively) rapid angular variations can amount to the same order of magnitude as in the slowly varying part mentioned above. The upper limits quoted refer to daytime conditions. At night conditions are markedly improved due to decreased thermal gradients.

Scintillation introduces random angle noise into the angle trackers with a power spectrum that varies as the $-2/3$ power of frequency at very low frequencies and has a high frequency cutoff determined by wind velocity and telescope aperture.

EARTH-SATELLITE BASED RECEIVER AND RECEIVER SITE COMPARISON SUMMARY

The advantages and disadvantages of an earth-satellite station are compared with an earth station.

A satellite base offers the possibility of continuously tracking a deep space probe with no outage due to eclipsing or due to atmospheric effects. A synchronous satellite, using a proper inclination angle with respect to the earth's equator, would have a small probability of being occluded by either the earth or the moon during a particular space mission for a time period in the order of months. In any case, not more than two such satellites should be required to provide uninterrupted link performance.

A synchronous satellite, rotating at the same speed as the earth, remains continuously visible from a given earth position.

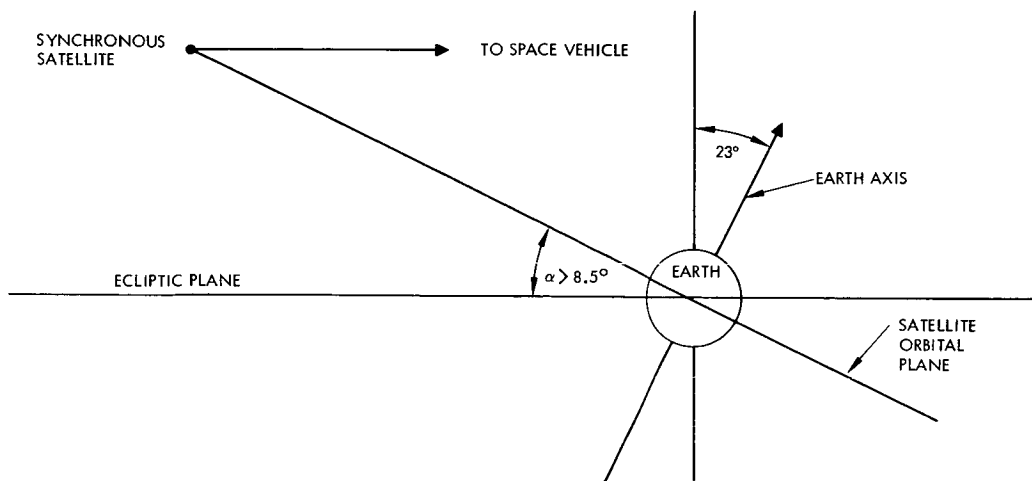
A synchronous satellite orbit, inclined with respect to the ecliptic plane by an angle somewhat greater than 8.5° ($\sin^{-1} 26,500/3900$) will be continuously visible to a deep space probe which is on a line formed by the intersection of the plane of the ecliptic and a plane perpendicular to the synchronous satellite orbit. This is indicated in the figure. As the deep space probe moves off this line there will be periods where the earth will occlude the LOS. Using an orbit with greater inclination will alleviate this, however over a period of time the earth will eventually be in a position to occlude the LOS unless the orbital plane of the synchronous satellite is precessed.

The absence of atmosphere at the earth-satellite greatly reduces acquisition and tracking problems since transmission losses, scintillation, image motion, and scattering effects will be reduced or eliminated. Furthermore continuous view eliminates switchover and reacquisition problems.

Aside from economic considerations, the principal disadvantage of the satellite based receiver is that it imposes on the receiver site acquisition and tracking system the design, reliability, and operating requirements associated with a space vehicle.

The table gives a comparison between the two links considered. Each site considered has definite advantages. A ground base for instance may utilize as much prime power and as large an antenna size as is considered necessary. A satellite base offers no atmospheric effects, a good probability of continuous coverage, and excellent pointing accuracies. The disadvantages are also well defined. A ground base is subjected to atmospheric effects, high background noise during the day, lack of continuous coverage from a single station, frequency reacquisition and long switchover time.

A satellite base is limited by antenna size and prime power, requires more complex and costlier equipment than a ground station, and in the event of occultations poses difficult reacquisition problems. Further, a satellite site requires the acquisition and tracking system to withstand large dynamic loads during boost.



Field of View of a Synchronous Satellite

General Acquisition and Tracking System Considerations
Receiver Location Considerations

EARTH-SATELLITE BASED RECEIVER AND RECEIVER SITE COMPARISON
SUMMARY

From analysis of this report it would appear that although the earth's atmosphere poses severe problems, they are not insurmountable. That is, a workable deep space system can be built which has its receiving station located on the earth's surface rather than a satellite receiver.

Comparison of Optical Receiver Sites

Base	Advantages	Disadvantages
Earth	<p>Power limited by laser state of art</p> <p>Antenna size limited by variable flexure of structure</p> <p>Logistics and maintenance simplified</p> <p>Sophisticated data processing and trajectory prediction equipment available</p>	<p>Pointing accuracy limited by image motion, beam spread</p> <p>Power reduced by absorption and scattering</p> <p>High background noise during daytime operation</p> <p>Possibility of operation depends on meteorological condition</p> <p>Several ground station required for continuous coverage</p> <p>Switchover and reacquisition problems difficult</p> <p>Long and frequent switchover time</p>
Satellite	<p>No atmospheric effects</p> <p>Low background noise</p> <p>Continuous coverage probable from single base</p> <p>Excellent pointing accuracy</p>	<p>Power and antenna size limited in near future by payload requirements</p> <p>Monitoring and control ground station required</p> <p>Complex equipment</p> <p>No maintenance</p> <p>Switchover and reacquisition difficult</p>

INTRODUCTION

In this section the primary emphasis is placed on one-way transmission from an unmanned deep space vehicle (DSV) to a receiving terminal located either on or near the earth.

The station at the DSV end of the communications link and the earth station are similar, in that both consist of input and output devices, a tracker, and signal processing electronics. Three important differences are: higher data rate assumed to be required for DSV-to-earth transmission, limited available space and power in the DSV, and environment of the DSV. As a result of the first two differences, a narrower beam is required for the down link and as a result of the absence of an atmosphere at the DSV, such a narrow transmit beam is possible. Accurately pointing and controlling such a narrow beam (to values as small as 1 microradian) are the major tasks of the acquisition and tracking control system.

In order to insure boresight integrity between the DSV transmitter and receiver, it would be desirable to use the same primary optical system. If the transmitter and receiver use different wavelengths (e. g. different laser modes) separation of the transmitted and received signals can be accomplished spectrally, and the full aperture can be used for each.

The subsections which follow include discussions of the following topics. 1) Acquisition subsystem operational consideration, 2) the tracking subsystem, 3) Signal-to-noise analysis of optical tracking systems, 4) acquisition, 5) detection theory, and 6) angle noise in optical tracking systems.

Subsequent Subsections

- Acquisition subsystem, operational considerations
- The tracking subsystem
- S/N analysis of optical tracking systems
- Acquisition
- Detection theory
- Angle noise error in optical tracking systems.

THE ACQUISITION SUBSYSTEM OPERATIONAL SEQUENCE

The acquisition sequence for the DSV consists of 3 phases: acquisition of inertial coordinates, acquisition of earth beacon, and tracking of earth beacon in narrow beam mode.

The sequence of events during the acquisition of the receiver site beacon by the deep space vehicle (DSV) is outlined in the following paragraphs.

Initially, a beacon at the receiver site capable of providing an adequate signal-to-noise ratio for DSV acquisition will be assumed. Thus the problem of acquisition reduces to orienting the DSV receiver field-of-view so the beacon falls within it.

The spacecraft is oriented in three phases. First it must be oriented such that the earth falls within a solid angle specified by the system gimbal limits. This is accomplished as follows: any large residual angular rates are first eliminated through operation of gas jets controlled by signals from a set of three rate gyros, one for each principal axis. The vehicle will then be oriented to point the telescope directly away from the sun by means of the gas jets and two-axis sun sensors. Roll rate about the telescope axis will be reduced to the limit cycle and held. Attitude signals for the pitch and yaw axes will be generated by the sun sensors, which are nulled when the telescope looks away from the sun. Next is the acquisition of the roll reference star (Canopus or a similar star) by the roll star tracker. This will be accomplished by a command roll rate about the sun line of sight (via gas jets) until the reference star enters the field of view. The generated star tracker error signal is then switched into the roll loop. Care must be taken to select the vehicle search rate small enough that the reference star may be acquired before the star passes through the field of view. Once acquisition of the star line-of-sight is complete an inertial reference has been established. Command signals are now given for a pre-programmed attitude maneuver and the spacecraft is rotated to point the telescope in the vicinity of the earth. (At near earth ranges, the horizon sensor can be used at this point.) When limit cycle operation has been achieved, the system is ready for the second phase of the acquisition.

The second phase of the acquisition consists of searching volume of space by scanning the DSV receiver using the telescope in a wide*angle field-of-view mode. The scan optics of the telescope start at the edge of the solid angle formed by the known uncertainty limits and search toward the telescope axis until the earth beacon is detected. When the beacon is detected, the error signals are switched into the gimbal drives of the telescope and the telescope axis is oriented along the DSV-E line of sight. The system is now in a coarse error detection mode.

*This phase may not be used when the DSV beam is 100 microradians or greater.

In the third phase of the acquisition the space vehicle's communication system is switched from the wide-angle search to the narrow angle track mode. (It may be necessary to do this in several steps if the wide angle tracking error is greater than the telescope's narrow angle field of view.) The system is now in the fine error detection or detection or tracking mode of operation.

The sun sensor can now be switched out of the vehicle pitch-yaw control channels, and be replaced by the tracking error signals, from the beacon. The star tracker must be retained however, since the two-axis control error commands. This marks the completion of the acquisition sequence and tracking control mode begins.

PHASES IN THE ACQUISITION SEQUENCE

- Acquisition of Inertial Coordinates
- Acquisition of Earth Beacon
- Tracking of Earth Beacon

ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS

The Tracking Subsystem

	Page
The Tracking Subsystem – Introduction	358
DSV Tracking Subsystem – Pointing Error	360
DSV Tracking Subsystem – Description	364
Stabilization Subsystems	368
Earth Station – Pointing Error Budget	372
Signal to Noise Ratios for Star Trackers	374

THE TRACKING SUBSYSTEM - INTRODUCTION

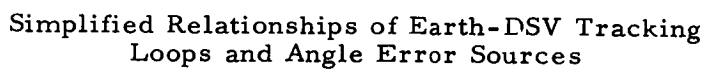
The Deep Space Vehicle (DSV) tracking loop, its errors and potential error reductions are noted.

The major contributions to pointing error in relation to stabilization and tracking characteristics are defined in this subsection for the DSV and for ground based tracking systems. The possible use of a near earth relay station to circumvent atmospheric effects on the downlink is discussed. The nature of pointing error budgets and interfaces and constraints with other subsystems are discussed for each case.

The relationship of system elements and the major angular error sources are shown in the figure for a tracking system.

The DSV-earth-DSV closed loop indicated by the dashed lines has the potential disadvantages of requiring several separated receivers at each earth-station for continuous coverage and of being able to correct errors which occur at frequencies corresponding to the 2-way transit time. However, this is only servo loop which can enclose the transmitter bore-sight and lead angle error sources. The transmitter-tracker relative alignment problem is made more difficult by the fact that the variable lead angle adjustment precludes mechanical locking following an in-flight alignment procedure. The major categories of pointing errors, typical causes and means of reducing their effects are summarized in the table.

Accurate pointing of massive structures such as a telescope is most efficiently done by having the entire spacecraft react to the torquing of a small inertia wheel control system. However, in certain applications, e.g., man motions or pointing two telescopes in different direction simultaneously, this is not possible. In such circumstances the telescope must be free to move relative to the spacecraft, and a means of controlling the telescope must be provided. It appears that the most difficult aspect of stabilization to the accuracies required for the DSV is associated with the generation of error signals of sufficient resolution and the alignment of the sensor sensitive axes with the control axes. The extremely precise control required necessitates vehicle configuration that will minimize the external disturbance torque effects as well as internal disturbance torques caused by inertial crosscoupling, equipment motion, temperature gradients, etc. Disturbance torques due to internal moving parts can be reduced by restricting activity during the fine tracking. However, disturbance torques due to inertial crosscoupling can be significant unless care is taken to balance the vehicle such that the inertias in all three axes are approximately equal. In addition, care must be taken to minimize the angular momentum stored in the vehicle.



**Pointing Error Causes and Means of Correction
for Optical Servo Systems**

Pointing Error Cause	Means of Correction
1. Mechanical disturbances of stabilized platform. (Bearing friction and misalignment, gimbal c.g. misalignment, spring torque from leads)	<ul style="list-style-type: none"> a. Attenuated by inertial stabilization b. Employ focal plane stabilization c. Separate from DSV d. Improve state-of-the-art
2. Stabilization system errors. Gyro drift (static and G-sensitive), resolver inaccuracies, accelerometer, tachometer	<ul style="list-style-type: none"> a. Feedback compensation b. On-gimbal star sensors c. Track loop design d. State-of-the-art improvement
3. Track errors. Sensor noise, error curve inaccuracies, resolver errors, focal plane tolerances	<ul style="list-style-type: none"> a. Lower tracking loop bandwidth b. Minimize tracking field-of-view c. Null tracking modes d. Increase beacon power
4. Mechanical alignment errors of optical axes. (Mechanical tolerance, lead angle errors)	<ul style="list-style-type: none"> a. Require only relative alignment b. Use closed loop where possible c. In-flight alignment/calibration
5. Atmospherics	<ul style="list-style-type: none"> a. Near earth relay b. Ground site selection c. Spatial averaging by distributed receivers

DSV TRACKING SUBSYSTEM - POINTING ERROR

The interrelationship of several errors associated with the DSV tracker is defined.

The correct pointing angle, in inertial coordinates, is given by the sum of the apparent line of sight (LOS) to the earth beacon, θ_{Bd} , and the lead angle, θ_{ld} (see the figure).

In general, pointing errors may be classified into three categories. These are listed below and illustrated in the figure.

1. Boresight and Lead Angle Errors (ϵ_B) - These errors contribute a static or nearly constant term to pointing error and are especially troublesome since they cannot be enclosed in a closed loop other than a DSV-earth closed loop. The penalty for excessive boresight errors is severe as the result is likely to be loss of transmission for an extended period of time. At ranges of 1 AU or more the round trip transit time exceeds 15 minutes and therefore correcting this error from the earth is slow.

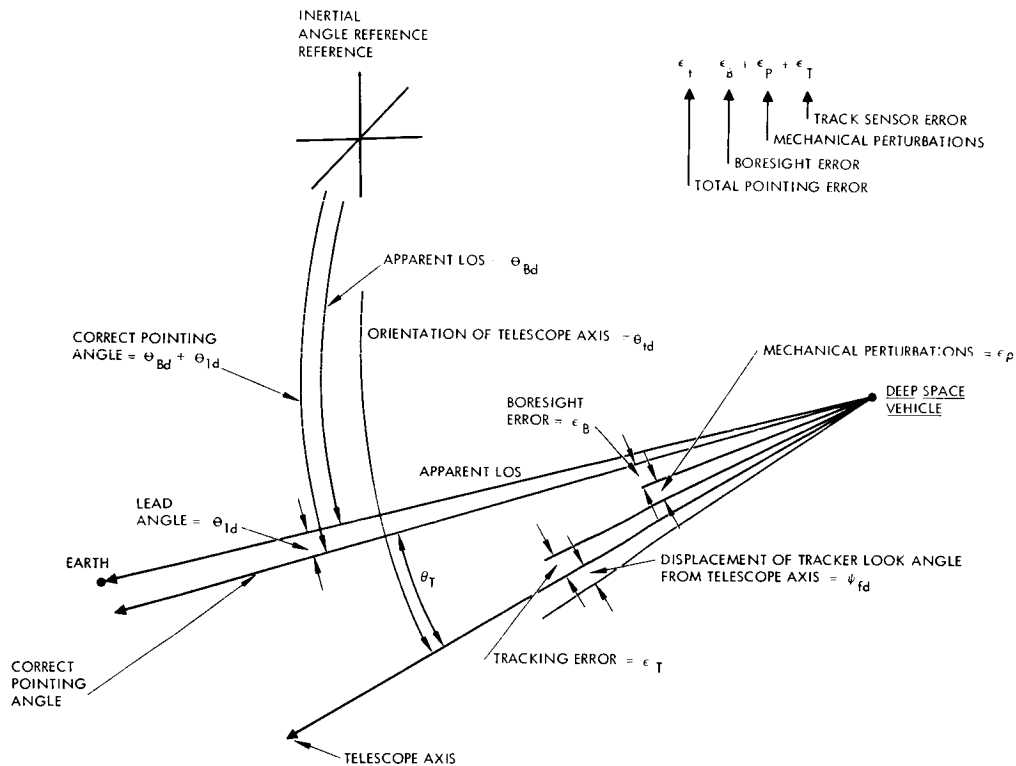
2. Mechanical Telescope Perturbations (ϵ_p) - The steady state response of the stabilization and tracking system may often be sufficient to suppress low frequency mechanical perturbations to a tolerable level. However, due to limited frequency response, high frequency components may cause transients in the inertial telescope pointing angle in excess of desired limits. Since these errors are sensed as apparent line of sight motions by the tracker they will be reduced by the combined action of track and stabilization loops. Under maximum spacecraft maneuvers the tracking system may not be able to contain the pointing error within prescribed limits so that an outage during maneuvers may accrue.

3. Tracking Errors (ϵ_T) - The true tracking error signal, θ_T , of the DSV tracker is given by:

$$\epsilon_T = \theta_{Bd} - \theta_{td} - \psi_{fd}$$

where θ_{td} is the orientation of the telescope axis in inertial coordinates under static conditions, i. e. $\theta_T = 0$, and, ψ_{fd} is the displacement of the tracker look angle (LOS) from the telescope axis (see the figure).

The angle noise in the tracker, originating from such sources as the beacon tracking sensor and the inertial reference sensor of the stabilization system, may be treated by standard Gaussian noise techniques. The angle noise power spectrum is modified by the closed loop response of the tracker and is included as a random term appearing in both θ_{td} and ψ_{fd} . This is considered in more detail in a subsequent topic.



Correct Pointing Angle and Errors Associated
with a Tracking System

DSV TRACKING SUBSYSTEM – POINTING ERROR

The total pointing error is then:

$$\epsilon_t = \epsilon_B + \epsilon_P + \epsilon_T$$

where the right-hand terms are the boresight, telescope perturbations, and tracking noise respectively. A typical system pointing accuracy specification therefore would include:

1. An absolute limit on boresight errors in the presence of thermal telescope environment, etc., and the accuracy of lead angle commands.
2. A specification on the tracking and stabilization system which reduces pointing errors as a function of deterministic mechanical disturbances of the telescope including expected rates and the torques due to the attitude control system and spacecraft-earth relative motion. In final form this type of specification may include the mechanical transfer characteristics of the Deep Space Vehicle-tracker-pointer mechanical interface.
3. A specification of the random tracking noise in the tracker to reduce loss rate in the presence of the aforementioned pointing error to be the prescribed level.

DSV TRACKING SUBSYSTEM - DESCRIPTION

The relationships of errors and implementation for a typical DSV tracking system are noted.

The servo loop of a typical vehicle track system using a mechanical inertial stabilization system is shown in Figure A. For simplicity only the outer gimbal loop is shown. The outer gimbal axis is denoted by the subscript d and is considered as part of a three axis system, attached to the inner gimbal of which is the track sensor platform. The three axis gimbal coordinate system is aligned with an i, j, k spacecraft coordinate system with the d axis coincident with the k axis when the inner and outer gimbal angles (θ_{td} and ψ_{fd} respectively) are zero. The angular track error signal, θ_T , detected by the track sensor is given by the difference between the LOS to the earth beacon, θ_{Bd} , (in inertial coordinates) and the course telescope angle, θ_{td} , plus the fine deflection pointing angle, ψ_{fd} .

$$\theta_T = \theta_{Bd} - (\theta_{td} + \psi_{fd}) \quad (1)$$

The detected error signal is given by the sum of θ_T and sensor errors, ϵ_t . The track angle error is modified by the fine deflection track closed loop function¹ to give:

$$\theta_T = \frac{s(\theta_{Bd} - \theta_{td})}{2 + K_1 K_2 g_1(s)} + \frac{\epsilon_t}{1 + s/K_1 K_2 g_1(s)} \quad (2)$$

The servo compensation function $g_1(s)$ generally has the character of a low-pass (or lag/lead) network. The effect of stabilization errors (spurious telescope motions, ϵ_p due to mechanical disturbances) may thus be reduced by the fine deflection tracking system up to the limiting response frequency of this loop. The frequency cut-off of the fine deflection track loop is limited primarily by the output information bandwidth of the track sensor which as a rule of thumb must be at least 6 times the bandwidth of the track loop to maintain phase margins. As can be seen from Equation (2) however, a high loop bandwidth makes the tracker more sensitive to sensor tracking noise (in proportion to the square root of the track loop bandwidth) so that even when the track sensor and deflection control mechanism are capable of arbitrarily fast reaction, the optimum track loop bandwidth should be limited.

In addition to correction of the track error by the fine deflection loop, the error signals from the track sensor are smoothed and used to correct the telescope pointing angle. The telescope must correct at a rate sufficiently fast to insure that the track angle θ_{fd} does not exceed the

¹ See Figure A for definition of K_1 , K_2 , $g_1(s)$

dynamic range of the fine deflection loop. This mechanism forms an outer or telescope tracking loop which encompasses the inner or fine deflection loop.

In a typical application the inner or fine deflection tracking loop contains the major contribution to the rms tracking noise. For the case where the angle noise power density, $W_{\eta}(f)$ of the track sensor is essentially white over the frequency response of the track loop, the rms tracker noise is from Equation (2).

$$\epsilon_N = \left[W_{\eta}(0) 2\pi \int_0^{\infty} df \frac{|g_1(f)|^2 K_1^2 K_2^2}{|K_1 K_2 g_1(f) + j2\pi f|^2} \right]^{1/2} \quad (3)$$

$$= \left[W_{\eta}(0) (\Delta f)_s \right]^{1/2}$$

where $(\Delta f)_s$ is the closed loop bandwidth.

The tracker noise power density at low frequency $W_{\eta}(0)$ is related to the solid FOV of the tracker, Ω_R ; the tracker voltage signal to noise ratio, SNR_V ; and the tracker information bandwidth $(\Delta f)_i$ by:

$$W_{\eta}(0) \leq \frac{\Omega_R}{(SNR_V)^2 (\Delta f)_i} \quad (4)$$

thus given the track field of view, track loop response time, $(\Delta f)_i$, and rms angle accuracy $j\epsilon_N$, the tracker sensitivity, SNR_V , is determined.

There are two major potential tradeoff quantities in the vehicle acquisition and tracking subsystem: (1) the bandwidth of the closed fine tracking system, f_t , and (2) the field of view of the tracker, Ω_R . The closed track loop bandwidth must be chosen large enough to reduce the residual stabilization error and small enough to avoid a large tracker noise contribution. Similarly the track field must be chosen sufficiently large to reduce tracker loss rate to a negligible level. However, an increased track field of view decreases the angular accuracy of the track sensor since the angle noise component due to sensor noise increases linearly with the size of the track FOV. In addition the errors due to tracker non-linearities and resolver readouts increase with the size of the FOV to be covered.

Sensor Boresight Errors. In a typical image sensor (vidicon, orthicon, image dissector) and non-linearities relating beam position to sweep voltage produce a tracking error which is proportional to the angular track field. In most cases this error is of the order of 1/2 to 1 percent.

DSV TRACKING SUBSYSTEM - DESCRIPTION

The quadrant photodetector depends for angle accuracy on the balance of energy in the blurred image of a point target among the four sections of the field of view. Even in the absence of noise, the null accuracy is limited by the degree to which the gains of the four channels can be balanced. Figure B shows the boresight error as a function of gain imbalance in a single channel. Thus in order to reduce boresight errors to less than 1 percent of the linear response portion of the field of view, gain imbalance must be held to less than 10 percent over the dynamic range of signal levels.

Resolver and Non-Orthogonality Errors. In the following, the method of analysis and parametric relations for tracking errors due to resolver inaccuracies and misalignment of gimbal axes is presented in some detail. Resolver error is the error that exists between the actual resolver shaft position and the indicated shaft position.

Base plate misalignment error is defined to be the angular error that arises when a gimbaled system is mounted on a reference base. The outer gimbal axis of rotation (see Figure C) is taken as the reference for determining these mounting errors and the mount may be misaligned with respect to a similar axis contained in the base. Since angles measured about three orthogonal axes completely specify the mounting misalignment, angles corresponding to roll ($\Delta\alpha$), pitch ($\Delta\beta$), and yaw ($\Delta\lambda$) are used for convenience.

In addition to base plate misalignment, the inner gimbal axis may not be orthogonal to the outer gimbal axis and cause pointing errors. This angle ($\Delta\gamma$) is measured in a plane orthogonal to the gimbal pointing direction.

The line-of-sight of each sensor may be misaligned with respect to the gimbal pointing direction for the mounts. These misalignment angles (Δx and Δy) are measured about an orthogonal set of axes (e and d) contained in a plane orthogonal to the gimbal pointing direction. Δx is measured about the d axis and Δy about the e axis.

Servo error is the error introduced by the tracking servo as a result of base motion inputs and servo noise. These sources of error (with the exception of servo dynamic error) are due to electrical and mechanical inaccuracies (measurement and fabrication) that can only be described in a statistical manner. It is assumed that the error parameters are statistically independent and that each parameter is normally distributed with zero mean.

Since the tracking servo base motion inputs and line-of-sight tracking rates are random with time the servo errors due to these inputs can also be considered to be random variables with zero mean.

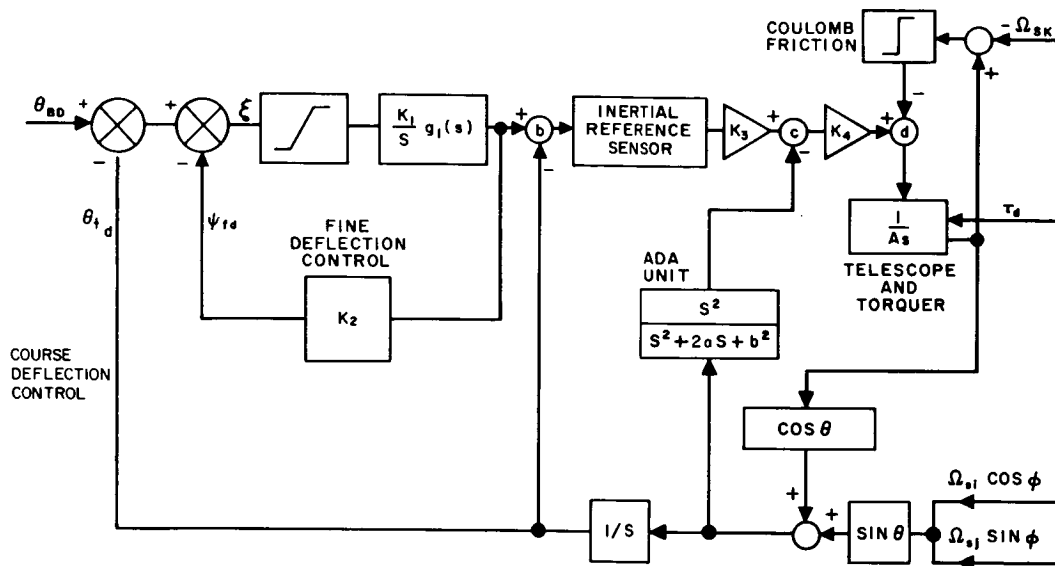


Figure A. Typical Outer Gimbal Servo Loop with Inertial Stabilization

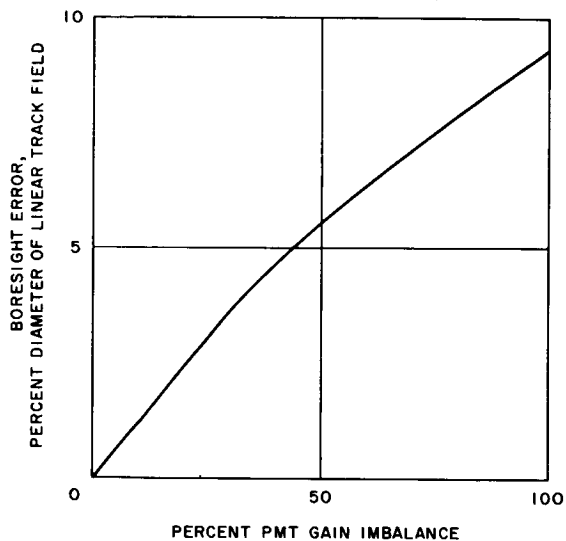


Figure B. Quadrant Detector Boresight Errors

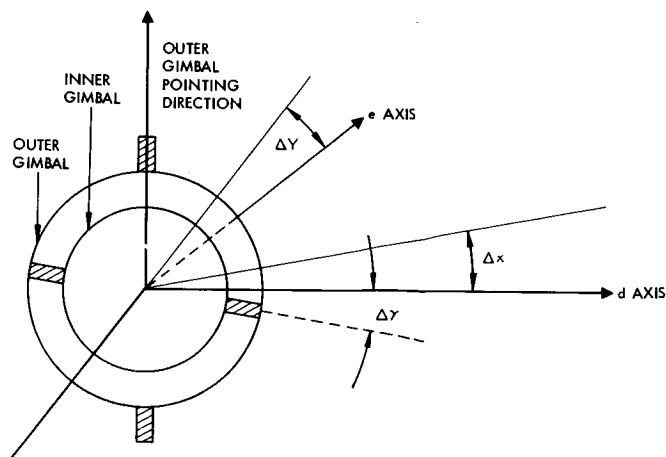


Figure C. Gimbal Alignment Geometry

STABILIZATION SUBSYSTEMS

Stabilization requirements and their effect on the tracking system are reviewed.

The stabilization loop shown in the figure illustrates typical problems in reducing the magnitude of mechanical perturbations to a tolerable level in the 0.1 - 1.0 arc-sec pointing accuracy regions. The disturbances may generally be classed as rate and torque disturbances. Rate disturbances are caused by vehicle maneuvers and the action of vehicle attitude control sensors. Torque disturbances may be caused by vehicle thrusting, meteorite impact or spring torques such as lead connections from the gimbals.

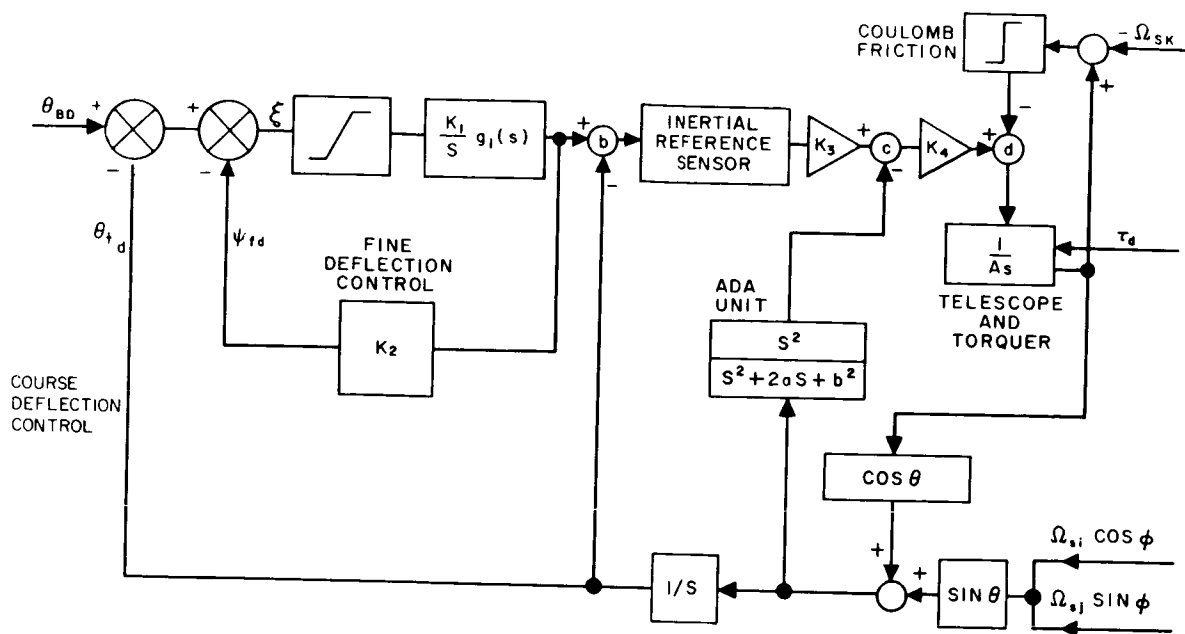
Angular rate disturbances denoted by $\Omega_{s;i,j,k}$ of the vehicle couple into the telescope dynamics primarily through bearing friction which is generally treated as ideal viscous (coulomb) friction and enters the servo loop as a torque at point (d). Disturbing torques may be minimized by placing the telescope mounting coaxial with the main thrust vector. However, a residual moment arm due to misalignment of the telescope center-of-gravity with the center of the gimbal axes remains. In addition, the vehicle structure will have an anisotropic transfer component so that torques directly about the gimbal axis will occur.

The stabilization shown reduces the input mechanical disturbances by (1) comparing the telescope position to an inertial reference sensor, point 6, and (2) by ADA (or accelerometer), point c, feedback loops around the telescope. In the typical stabilization scheme a rate integrating gyro mounted on the gimbal structure is used to sense motions of the telescope in inertial space. The output voltage of the gyro is then used to torque the gimbals so as to null the telescope inertial rate. When, as in the track mode, it is required to move the telescope in inertial space a torquing command is fed to the gyro. The major problems with current gyros are gyro drift and limited reliability. Performance of the best gyros to date is a few hundredths deg/hr (static) drift and greater than 20/million hours failure rate. The contribution of gyro drift to overall pointing and tracking accuracy is minimized by the low frequency error rejection of the closed track loop.

An alternate mechanization of the inertial reference sensor are star trackers mounted on the telescope platform. Since these must have a separate gimbal system to allow narrow field tracking of individual stars the inertial reference frame determined by the star trackers must be referenced to gimbal axis system through a coordinate transformation.

The steady state equivalent input angular rate error, $e_{da}(ss)$, reflected into the track loop at point (a) from input periodic torque disturbances which are lower in frequency than the natural frequency of the stabilization loop is:

$$e_{da}(ss) = \tau_d / K_3 K_4$$



Typical Outer Gimbal Servo Loop
with Inertial Stabilization

STABILIZATION SUBSYSTEMS

thus the limitation on the ability of the stabilization system is to minimize disturbances of this sort depends on the open loop gain (and frequency response) which can be supplied in the stabilization system.

Optical Design Constraints. In addition to sensitivity and resolution requirements, the design of systems in which the tracker shares the same aperture with the transmitter and receiver of the optical communication system imposes additional constraints on the optical design:

1. The field angle over which the optics are capable of collimating the transmitter beacon to the desired beamwidth must include in addition to the tracker field of view a range of angles sufficient to allow the transmitter lead angles to be generated. For a typical Mars transfer orbit this amounts to an extra 40 seconds of arc.
2. In systems employing aperture sharing rather than time sharing of earth beacon provision must be made to provide tracker-transmitter optical isolation of a very high degree.

Inertial Reference Unit and Guidance and Navigation Unit Interfaces. Data concerning relative range and range rate must be made available to compute the tracker lead angles. In addition, initial pointing angles for the acquisition process must be referenced to the gimbal coordinates through the gimbal angle readouts.

Vehicle Mechanical Interface. The reduction of high frequency mechanical disturbances transmitted over the optical platform — Vehicle interface is of prime importance to maintaining high pointing accuracy within acquisition and tracking subsystem. Where high pointing accuracy is required during periods of operation of reciprocating machinery, attitude control limit cycling, and thrusting; a solution to the problem is to modify the optical platform mount so as to reduce the high frequency perturbations. This may take the form of soft mounting or mechanically disconnecting the platform. Means must be provided however to allow referencing of spacecraft and gimbal axis coordinate systems.

Boresight Maintenance. The relative boresighting of the tracker and transmitter must be maintained to high accuracy in the thermal and mechanical environment of the spacecraft.

EARTH STATION - POINTING ERROR BUDGET

The error budget is described in terms of statistical and bias errors, the interrelationship is documented and sample normalized loss rate curve presented.

The major sources of pointing error in a ground tracking system are:

1. Atmospheric Scintillation (ϵ_S) - Angle scintillation due to random phase errors is introduced into the spacecraft beacon by the atmosphere. From the point of view of the angle tracker, scintillation introduces random angle noise with a power spectrum which varies as the $-2/3$ power of frequency near zero frequency and has a high frequency cut off determined by wind velocity which for large aperture telescopes will usually occur at a few cps.

The maximum blur due to atmospheric scintillation imposes a limitation on the tracker FOV. Typically the blur size can vary from 0.5 to 3 arc-sec at night and from 1 to 6 arc-sec for daytime observation.

2. Boresight and Lead Angle Errors (ϵ_B) - The pointing errors due to misalignment of the earth beacon and earth tracker and errors in introducing the beacon lead angle are of the form of bias errors.

3. Tracker Errors (ϵ_T) - Errors introduced in the tracker system such as tracker noise and track sensor boresight errors. The total pointing error is thus

$$\epsilon = \epsilon_S + \epsilon_B + \epsilon_T$$

which is the same form noted in a previous topic that described the Deep Space Vehicle pointing error.

A figure of merit for trackers is loss rate, Λ , defined as the inverse of the mean time between loss of track. The loss rate may be expressed as a function of the tracker error parameters, statistical and biased.

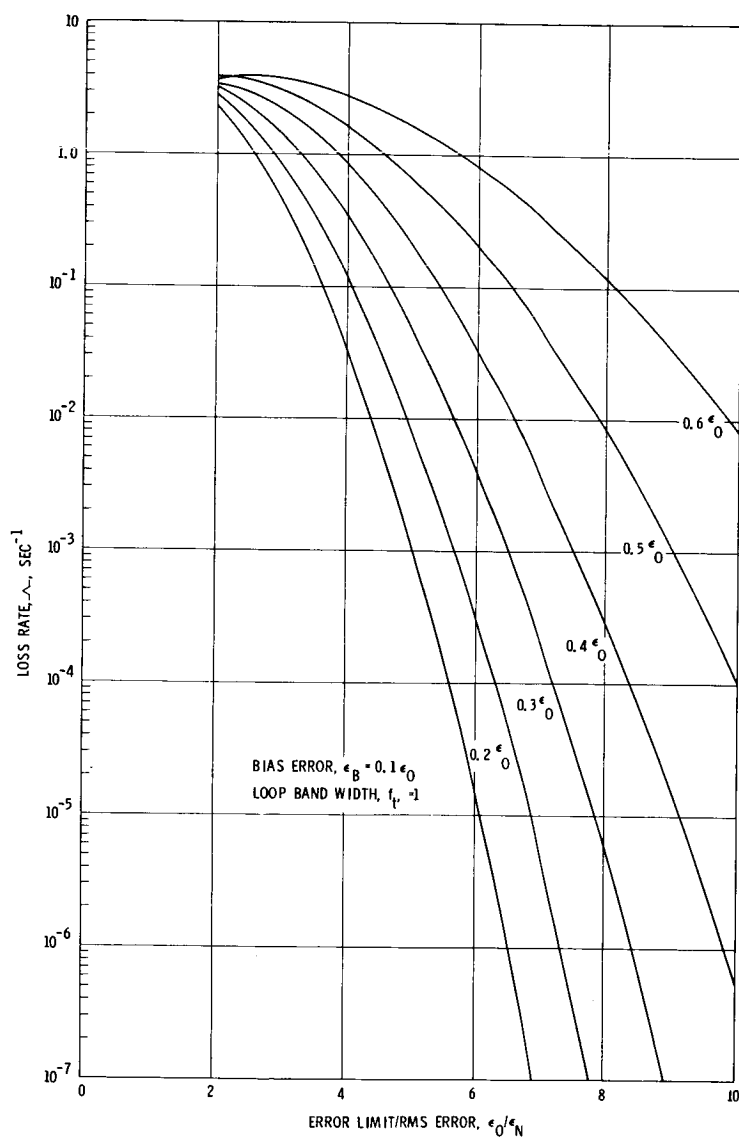
By denoting ϵ_N to be the rms sum of the statistical scintillation and tracker noise contributions to the tracking error, and ϵ_o as the angular radius limit of an earth beacon, the loss rate, Λ , is given approximately by:

$$\Lambda \approx 2\pi f_t \left(\frac{\epsilon_o - \epsilon_B}{\epsilon_N} \right) \exp \left\{ -1/2 \left[\left(\frac{\epsilon_o - \epsilon_B}{\epsilon_N} \right)^2 \right] \right\}$$

where f_t is the closed loop bandwidth of the tracker. (The multiplicative constant will vary somewhat depending upon the filtering used but this formulation is representative.)

The figure plots the loss rate as a function of ϵ_0/ϵ_N using ϵ_B as a parameter and normalized with $f_t = 1$.

The mean time between losses should obviously be large when compared to the largest time lag in the tracking system. This lag is the round trip time, which is in the order of 15 minutes for a Mars encounter. If the mean time to loss is specified as 20 times this, a loss is allowed once for every 18,000 seconds or a loss rate of 5.56×10^{-5} . As may be seen from the figure this requires the ratio of ϵ_0/ϵ_N to be between 6.2 and 9.2 as ϵ_B varies between 0.1 and 0.4.



Tracking Loss Rate

SIGNAL TO NOISE RATIOS FOR STAR TRACKERS

Signal to noise examples are given for a star reticle tracker as a function of star magnitude for 3 implementations.

Star trackers operate in the visible spectrum and therefore will use direct detection due to its high efficiency and its simplicity. Beacon trackers may use direct detection or heterodyne detection but due to the simplicity, direct detection operating in the visible light range is favored.

The equation describing the signal-to-noise ratio for direct detection has been developed in Volume Ii under "Detection Noise Analysis." It is repeated here for convenience.

$$\frac{S}{N} = \frac{\left(\frac{G\eta q}{hf} P_C \right)^2 R_L}{kTB_o + 2qB_o G^2 \left(\frac{\eta q}{hf} P_C + \frac{\eta q}{hf} P_B + I_D \right) R_L} \quad (1)$$

where:

- G = detector gain
- η = detector quantum efficiency
- q = electronic charge, 1.602×10^{-19} coulombs/electron
- h = Plank's constant 6.624×10^{-34} watt sec . sec
- f = light frequency, Hz
- P_C = received carrier power, watts
- R_L = load resistance, ohms
- k = Boltzmann's constant, 1.38×10^{-23} watts/Hz °K
- T = Amplifier noise temperature, °K
- B_o = Amplifier bandwidth, Hz
- P_B = Background received power, watts
- I_D = dark current, amps

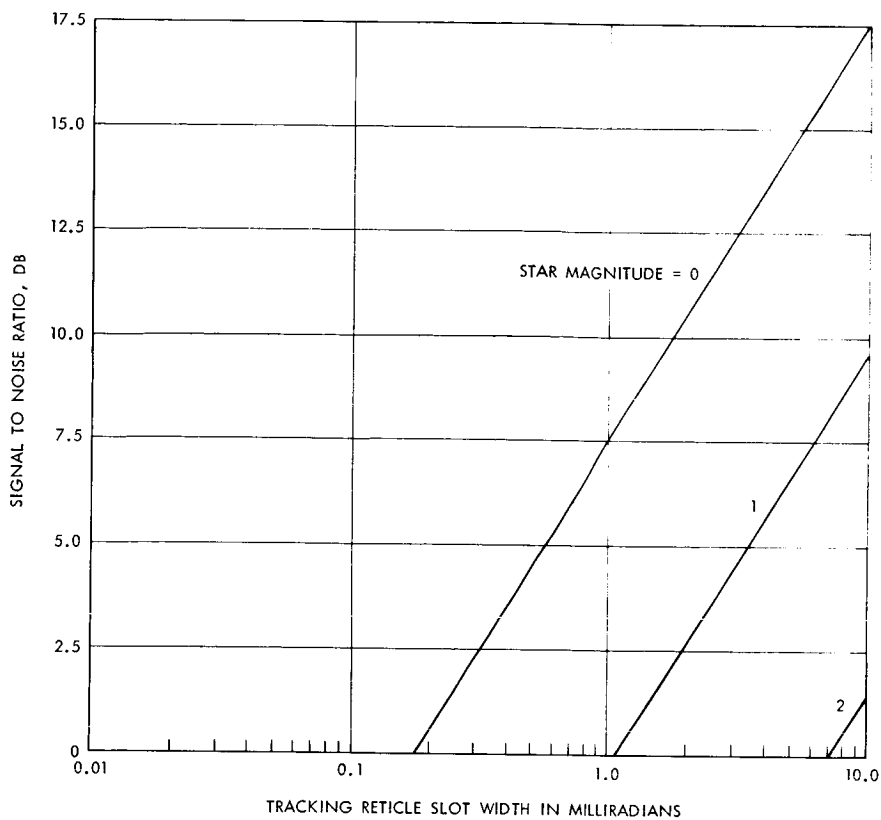


Figure A Sample Signal to Noise Ratio, Using a Photomultiplier Diode (Thermal Noise Limited)

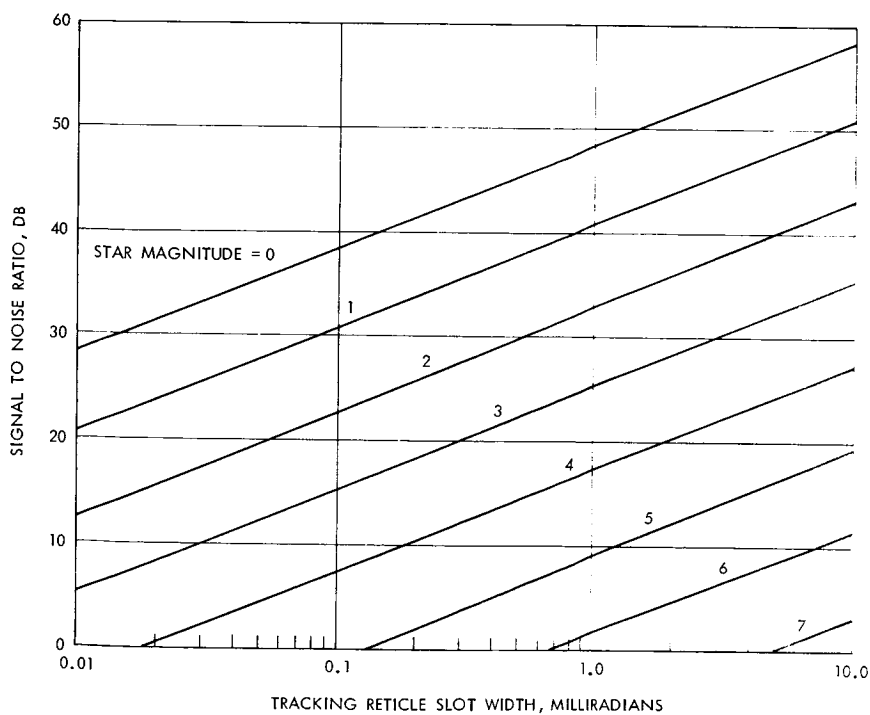


Figure B. Sample Signal to Noise Ratio Using a Photomultiplier Detector (Dark Current Limited)

SIGNAL TO NOISE RATIOS FOR STAR TRACKERS

Symbolically this is

$$\frac{S}{N} = \frac{\text{Signal Power}}{\text{Thermal noise power} + \text{shot noise power} + \text{background noise power} + \text{dark current noise power}}$$

The power received from the star, P_c , is

$$P_c = (H_s)(A_o)(B_1) \quad (2)$$

where

H_s = Star irradiance watts/cm²-micron

A_o = Star tracker effective receiving area

B_1 = Optical filter bandwidth

The video bandwidth of the tracker depends upon the type of tracker used. If a nutating reticle is used, the incoming position is encoded by the reticle rotation. This can be done by allowing the incoming signal to pass through a slit of width W . If the slit is offset by r milliradians, $2r$ corresponds to the nutation circle. If the nutation rate is f_s , the time the star is in the slit is t_o or:

$$t_o = \frac{\omega}{2\pi r f_s} \text{ seconds}$$

This time may be related to the required tracking bandwidth as

$$B_o \approx \frac{1}{t_o} = \frac{2\pi r f_s}{\omega} \quad (3)$$

If equation (2) and (3) are substituted into equation (1) and certain parameters values assumed, the signal to noise ratio may be calculated. Values for such a calculation are given in Figure A for a photo diode ($G = 1$) and for a photo multiplier tube ($G = 10^5$) in Figures B and C using star magnitude as a parameter. Other assumed parameter values are noted below.

$$\eta = 0.585$$

$$f = 3 \times 10^8 / .5 \times 10^{-6}$$

$$R_L = 300 \text{ ohms}$$

$$T = 350^\circ\text{K}$$

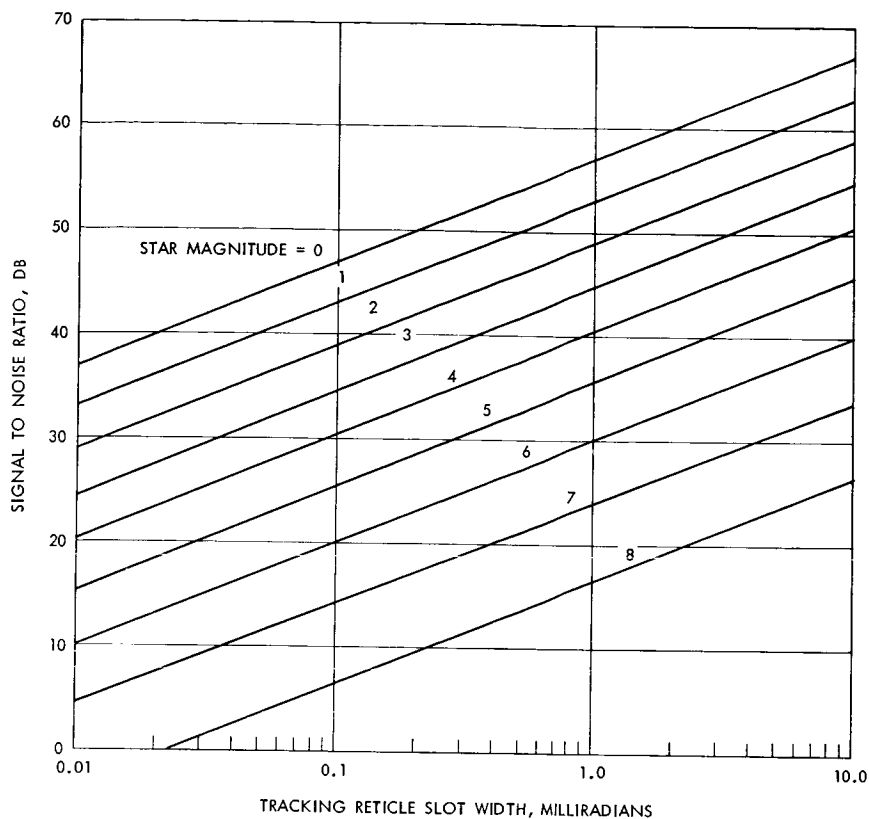


Figure C. Sample Signal to Noise Ratio Using a Photomultiplier Detector (Signal Limited to Background Limited)

$r = 5$ milliradians

$f_s = 32$ rev/sec

$P_B =$ equivalent to 6th magnitude star (1.1×10^{-14} watts/cm²-micron)

$I_D = 10^{-8}$ amps and zero amps (Figure C only)

$A_o = 10$ cm²

$B_l = 3$ microns

$H_s = 3 \times 10^{-12}$ watts/cm² - micron (0 magnitude - star)

It should be noted that the signal to noise ratio of Figure A is thermal noise limited while the signal-to-noise ratio of Figure B is dark current noise limited. Figure C which has the dark current set to zero, is signal noise limited for strong signals (near zero magnitude) and background noise limited for weaker signals (6th magnitude and greater). (Note that the background value assumed correspond to a 6th magnitude star.)

ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS

Acquisition

Mean Time to Acquisition	Page 380
Acquisition Probabilities	382

MEAN TIME TO ACQUISITION

The mean time to acquire is developed in terms of the statistical and nonstatistical detection parameters.

It is generally desirable to acquire the transmitter-beacon as rapidly as possible. The speed of acquisition is limited by: (1) the possibility of passing over the transmitted signal without detecting it, and (2) the possibility of acquiring false signals produced by noise if the threshold is reduced to increase the sensitivity of the receiver to the transmitted signals.

A quantity \bar{T}_a is defined as the mean time to acquire. The system is to be designed to minimize this number subject to various constraints imposed by other conditions. The average cost, in lost time, due to false acquisitions during the interval required to scan the acquisition field once, is

$$C_o = R_f T_\ell (T_\Sigma - T_\sigma)$$

where R_f is the average rate of false target acquisitions, T_ℓ is the lost time due to acquiring a false target, T_Σ is the minimum time to scan the search field of view, and T_σ is the time spent on target during a single scan. The average time needed to scan the acquisition field is then

$$\bar{T}_\Sigma = (T_\Sigma + C_o)$$

If no signal is sensed it is necessary to scan the complete acquisition field again before acquisition can be made. If the decision is "no target" on the second scan, the field scan is repeated, etc. If the target has equal probability, P_o , of being anywhere within the field, the mean time to acquire is

$$\begin{aligned} \bar{T}_a &= \frac{\bar{T}_\Sigma}{2} P_o + \frac{3}{2} \bar{T}_\Sigma P_o (1-P_o) + \frac{5}{2} \bar{T}_\Sigma P_o (1-P_o)^2 + \dots \\ &= \frac{\bar{T}_\Sigma}{2} \frac{(2-P_o)}{P_o} \end{aligned}$$

The minimum time in which the field can be scanned is

$$T_{\Sigma} = \frac{\Sigma}{R_s}$$

where Σ is the size of the search field of view and R_s the scanning rate. If the receiver field of view is σ , the time spent on target during a single scan, T_{σ} , is

$$T_{\sigma} = \frac{\sigma}{R_s}$$

In addition the number of scan elements, N_s ,

$$N_s = \frac{\Sigma}{\sigma}$$

Since the target may be in any one of these elements with equal probability, P_o , this probability is given by

$$P_o = \frac{1}{N_s} = \frac{\sigma}{\Sigma}$$

In terms of the fundamental parameters the mean time to acquire becomes

$$\bar{T}_a = \frac{\Sigma}{2R_s} \left[1 + R_f T_{\ell} \left(1 - \frac{\sigma}{\Sigma} \right) \right] \left[2 \frac{\Sigma}{\sigma} - 1 \right]$$

The quantity R_f is dependent on the noise which arrives at the thresholding device, the scan rate, R_s , and the receiver field of view, σ . Sources of noise are the photoelectric detector, background radiation, fluctuations in the signal and background due to atmospheric effects, and the random distribution in time of the photons which constitute the received power. These quantities, the false alarm rate and the probability of detection respectively, are discussed in the following for both the low incident flux level (Poisson) and high incident flux level (Gaussian) cases.

ACQUISITION PROBABILITIES

The acquisition probability is given as a function of the number of angle bins searched, the probability of signal plus noise exceeding the threshold, and the probability of noise alone exceeding the threshold.

The acquisition beamwidth, σ , must scan over the uncertainty solid angle, Σ . Using a simplifying assumption relative to beam overlap, defines K angle bins where $K = \Sigma/\sigma$. If the probability of the noise exceeding the detection threshold is taken as P_N and the probability of the signal plus noise exceeding the threshold is P_S , then probability of acquisition, P_{acq} , may be determined as follows:

$$P_{acq} = \sum_{n=1}^K P(m) (1 - P_N)^{n-1} P_S \quad (1)$$

where $P(m)$ is the probability that the signal is in the m^{th} bin. The acquisition implementation used is one where the beam is scanned until a target is detected. Thus the entire frame will not be scanned unless the target is not detected. If a false target is detected the false target angle coordinates are tracked until the nature of the false target is determined. If the simplifying assumption is made that $P(m) = 1/K$ equation (1) reduces to

$$P_{acq} = \frac{P_S}{K} \sum_{n=1}^K (1 - P_N)^{n-1} \quad (2)$$

this is a geometric series which has as the sum

$$P_{acq} = \frac{P_S}{K} \left[\frac{1 - (1 - P_N)^K}{1 - (1 - P_N)} \right]$$

or

$$P_{acq} = \frac{P_S}{KP_N} \left[1 - (1 - P_N)^K \right] \quad (3)$$

Equation 3 has been plotted for three values of K , the number of angle bins, in Figures A, B and C. In each figure the probability of acquisition, P_{acq} , is given as a function of P_S , the probability of signal and noise exceeding the threshold, using P_N , the probability of noise only exceeding the threshold, as a parameter. P_S and P_N may be a result of several types of statistics, Poisson, Gaussian, etc. Values for P_N and P_S in terms of such statistics are given in subsequent sections.

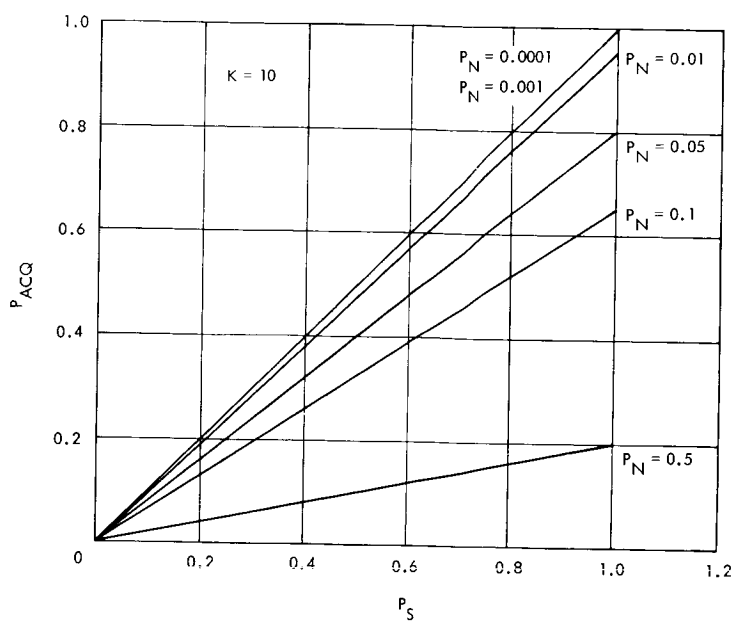


Figure A. Probability of Acquisition when 10 Angle Bins are Used

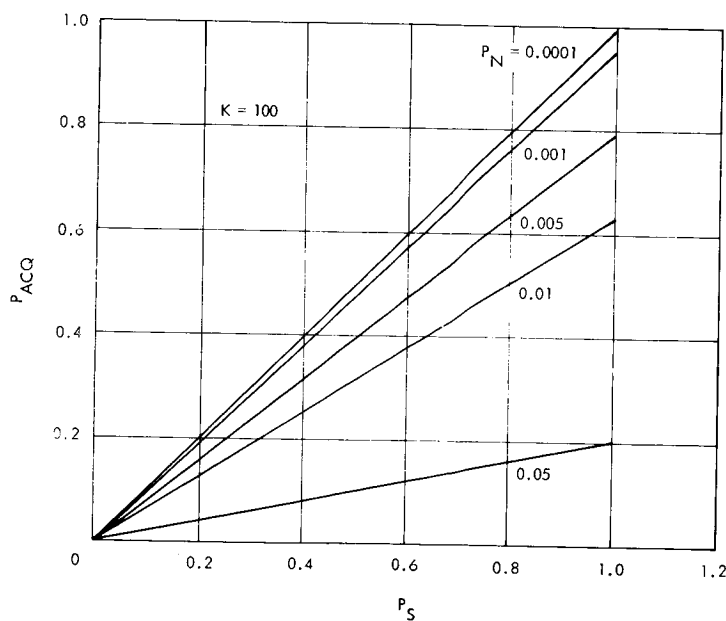


Figure B. Probability of Acquisition when 100 Angle Bins are Used

ACQUISITION PROBABILITIES

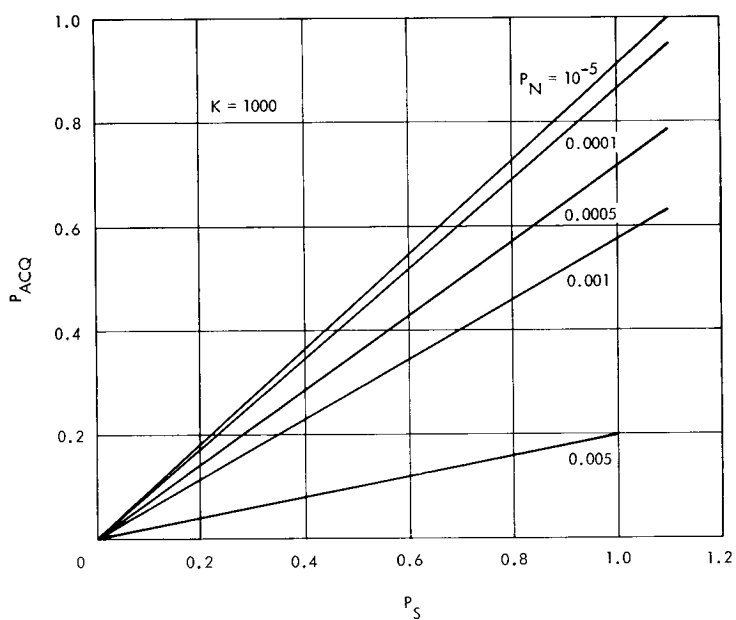


Figure C. Probability of Acquisition when
1000 Angle Bins Are Used

ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS

Detection Theory

The Probability of Detection and False Alarm (Gaussian Case)	Page 386
Probability of Detection and False Alarm (Poisson Case)	392

THE PROBABILITY OF DETECTION AND FALSE ALARM (GAUSSIAN CASE)

The probability of detection for an ideal matched filter detection is derived for Gaussian statistics.

The classical detection problem of a known signal in additive (colored) Gaussian noise may be described mathematically as follows:

Given a finite record of observed data,

$$v(t) = a s(t) + n(t), \quad 0 < t < T$$

determine if the known signal $s(t)$ is present or not (i.e., $a = 0$ or 1) given that $n(t)$ is not necessarily stationary Gaussian noise with vanishing mean and autocovariance function $K(u, t)$. The solution of this problem may be found in a number of textbooks.^{1, 2, 3}

The important results are the following:

The optimum detector is a filter with an impulse response

$$h(\tau) = \begin{cases} g(T-\tau) & \text{for } 0 < \tau < T \\ 0 & \text{for } \tau < 0 \text{ or } \tau > T \end{cases}$$

where $g(t)$ is the solution of the integral equation

$$\int_0^T K(u, t) g(u) du = s(t)$$

¹C. W. Helstrom, Statistical Theory of Signal Detection, Pergamon Press, New York, 1960.

²D. Middleton, An Introduction to Statistical Communication Theory, McGraw-Hill Book Company, Inc., New York, 1960.

³Y. W. Lee, Statistical Theory of Communications, J. Wiley and Sons, Inc., New York, 1960.

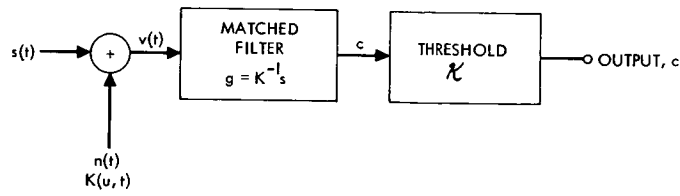


Figure A. Optimum Detector

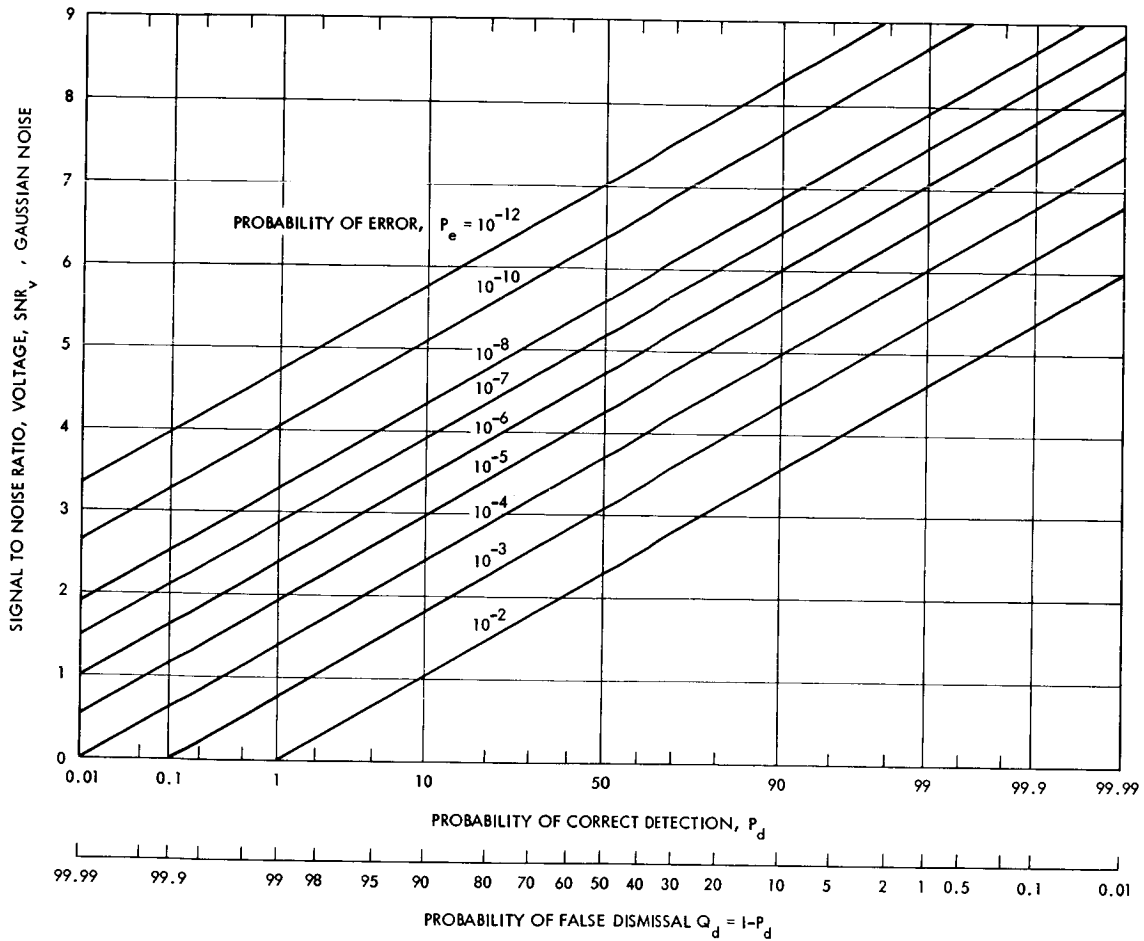


Figure B. General Detection Probability

THE PROBABILITY OF DETECTION AND FALSE ALARM (GAUSSIAN CASE)

followed by a threshold (bias level) given by

$$\chi = \ln \Lambda_0 + \frac{1}{2} \text{SNR}$$

where Λ_0 depends on the particular decision rule (Bayes, Minimax, Neyman-Pearson), and SNR means signal-to-noise ratio (power). Diagrammatically this is illustrated in Figure A.

If \mathcal{C} denotes the output of this filter, then $\mathcal{C} > \chi$ implies $s(t)$ is present while $\mathcal{C} < \chi$ indicates the $s(t)$ is not present. As expected, these results are not always true. Thus, sometimes it is $\mathcal{C} > \chi$ while the signal $s(t)$ is not present. This is a false detection. Similarly, it may happen that $\mathcal{C} < \chi$ while $s(t)$ is present. This is a false dismissal. The probability of false detection for a given threshold level, χ , depends on the SNR and is given by:

$$P_e = P_r [\mathcal{C} > \chi | a = 0] = \frac{1}{\sqrt{2\pi}} \int_{\frac{\chi}{\text{SNR}_v}}^{\infty} e^{-\frac{x^2}{2}} dx \quad \begin{array}{l} \text{Prob-} \\ \text{ability} \\ \text{of} \\ \text{Error} \end{array}$$

or

$$P_e = \text{erfc} \left[\frac{\chi}{\text{SNR}_v} \right] \text{ which is the (tabulated) error function.}$$

The probability of (correct) detection, P_d , is given by

$$P_d = P_r [\mathcal{C} > \chi | a = 1] = \frac{1}{\sqrt{2\pi}} \int_{\frac{\chi - (\text{SNR})_v^2}{(\text{SNR})_v}}^{\infty} e^{-\frac{x^2}{2}} dx$$

or

$$P_d = \text{erfc} \left[\frac{\chi - (\text{SNR})_v^2}{(\text{SNR})_v} \right]$$

THE PROBABILITY OF DETECTION AND FALSE ALARM (GAUSSIAN CASE)

Thus given an upper bound for the probability of false alarm, $P_e, \chi/\sqrt{\text{SNR}}$ can be determined and the corresponding value of the probability of detection, P_d , can be obtained for various (SNR_v) . The results are given in Figure B.

Indication of Proof. Observe that \mathcal{C} is a Gaussian random variable, for it is the result of a linear operation (namely, integration) performed on the Gaussian random variable $v(t)$. Its mean value under hypothesis $a = 0$ is:

$$E [\mathcal{C} \mid a = 0] = 0$$

while under the hypothesis $a = 1$ is:

$$E [\mathcal{C} \mid a = 1] = \mu$$

The variance of \mathcal{C} under both hypotheses is the same and is found to be:

$$E [(\mathcal{C} - \mu)^2 \mid a = 1] = E [(\mathcal{C} - 0)^2 \mid a = 0]$$

It can be shown for the case under consideration that $\mu = (\text{SNR}_v)^2$, i.e., mean value of \mathcal{C} under hypothesis $a = 1$ is equal to the variance of \mathcal{C} . Because \mathcal{C} is Gaussian its p.d.f.'s* are given by:

$$p(\mathcal{C} \mid a = 0) = (2\pi (\text{SNR}_v)^2)^{-1/2} \exp - \frac{\mathcal{C}^2}{2(\text{SNR}_v)^2}$$

and

$$p(\mathcal{C} \mid a = 1) = (2\pi (\text{SNR}_v)^2)^{-1/2} \exp - \frac{(\mathcal{C} - \text{SNR}_v^2)^2}{2(\text{SNR}_v)^2}$$

The false detection probability P_e is:

$$P_e = \Pr [\mathcal{C} > \chi \mid a = 0] = \Pr \left[\frac{\mathcal{C}}{(\text{SNR}_v)} > \frac{\chi}{(\text{SNR}_v)} \mid a = 0 \right]$$

*Probability distribution function

The random variable $x = \mathcal{C} / (\text{SNR}_v)$ has unit variance.

Hence:

$$P_e = \Pr \left[x > \frac{\mathcal{K}}{(\text{SNR})_v} \mid a = 0 \right] = \frac{1}{\sqrt{2\pi}} \int_{\frac{\mathcal{K}}{(\text{SNR})_v}}^{\infty} \exp -\frac{x^2}{2} dx$$

or

$$P_e = \text{erfc} \frac{\mathcal{K}}{(\text{SNR})_v}$$

In a similar way the correct detection probability P_d is:

$$P_d = \Pr [\mathcal{C} > \mathcal{K} \mid a = 1] = P \left[\frac{\mathcal{C} - (\text{SNR}_v)^2}{(\text{SNR})_v} > \frac{\mathcal{K} - (\text{SNR}_v)^2}{(\text{SNR})_v} \mid a = 1 \right]$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{\mathcal{K} - (\text{SNR}_v)^2}{(\text{SNR})_v}}^{\infty} \exp \left(-\frac{x^2}{2} \right) dx$$

or

$$P_d = \text{erfc} \frac{\mathcal{K} - \text{SNR}_v^2}{(\text{SNR})_v} = \text{erfc} \left(\frac{\mathcal{K}}{\text{SNR}_v} - \text{SNR}_v \right)$$

The error function, $\text{erfc } x$, is extensively tabulated in many publications under various forms. Here the form

$$\text{erfc } x = \int_x^{\infty} e^{-t^2/2} dt$$

is used.

PROBABILITY OF DETECTION AND FALSE ALARM (POISSON CASE)

A means for determining the probability of detection and probability of false alarm using Poisson statistics is given.

For this case, the approach of Woodbury¹ is followed. Since both the signal and noise photons are governed by Poisson statistics, the average number of received signal photons N_R is \bar{N}_R and the average number of received noise photons N_N is \bar{N}_N . The probability of detection, P_d , is defined as the probability that the number of signal plus noise events, $N = N_R + N_N$, be equal to, or greater than, a certain threshold, M , when the laser signal is present.

$$P_d = \sum_{X=M}^{\infty} \exp \left(-\bar{N} \frac{(\bar{N})^X}{X!} \right)$$

where the distribution of N is taken to be Poisson. Figure A plots M versus N with P_d as a parameter.

The probability of false alarm, P_e , per cell is defined as the probability that the number of noise events N_N is equal to or greater than M :

$$P_e = \sum_{X=M}^{\infty} \exp \left(-\bar{N}_N \frac{(\bar{N}_N)^X}{X!} \right)$$

This relationship is given in Figure B, where the M (threshold) is given as a function of \bar{N}_N , with P_e as a parameter.

¹E. J. Woodbury, Annals of the New York Academy of Sciences, 122, 661 (1965).

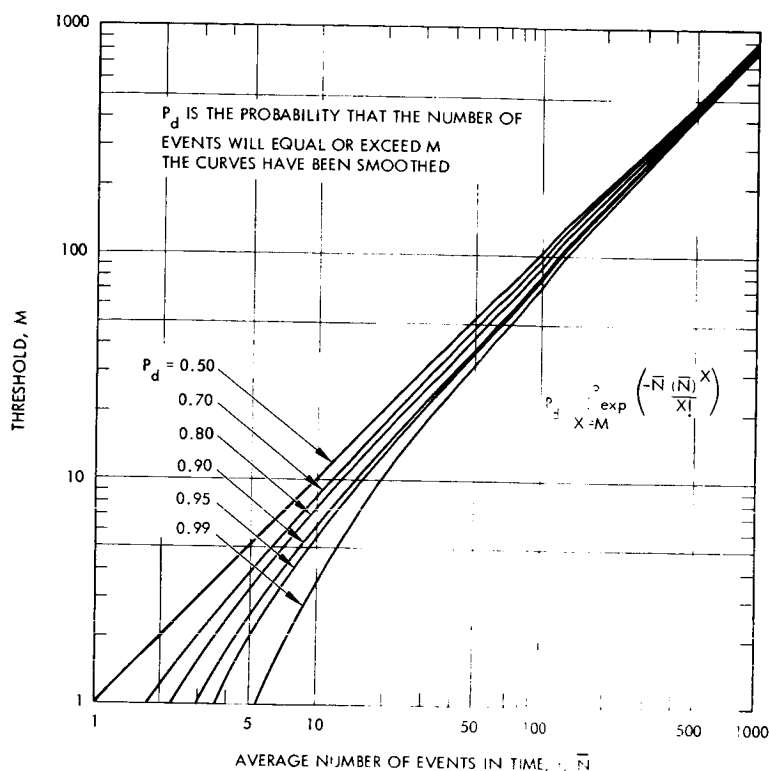


Figure A. Threshold as a Function of Average Number of Events, $\bar{N} = \bar{N}_R + \bar{N}_N$ in Time t , Dwell Time

Note: When it is desired to detect with a given probability, P_d , an event which emits on the average \bar{N}_R photons, in the presence of \bar{N}_N noise photons in t seconds, the threshold should be set at a number, M . Under these conditions there is always a probability of false detection, P_e , given in Figure B.

PROBABILITY OF DETECTION AND FALSE ALARM (POISSON CASE)

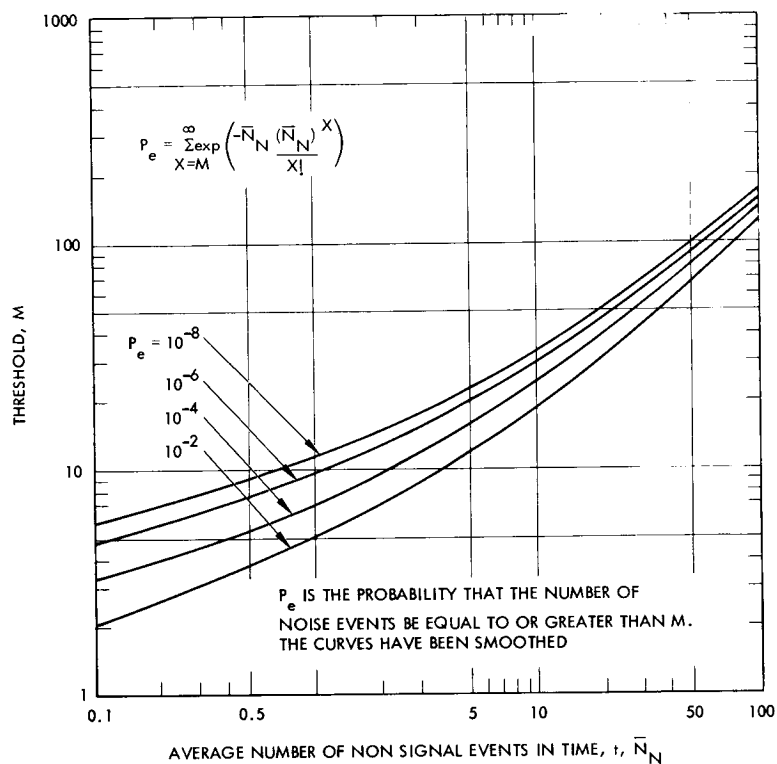


Figure B. Threshold as a Function of Average Number of Noise Events in Time t

Note: This curve may be used to estimate the probability of false detection, P_e , when the threshold of the decision mechanism has been set at M (see Figure A). The knowledge of the average number of noise photons, \bar{N}_N , determine P_e .

ACQUISITION AND TRACKING SYSTEM PERFORMANCE ANALYSIS

Angle Noise Error in Optical Tracking Systems

	Page
Introduction	396
Monopulse Quadrant Tracking System	400
Angle Noise Analysis of Monopulse Quadrant Tracking System	406
Beam Lobing PPM Tracking System	412
AM Reticle Tracking System	416
FM Reticle Tracking System	420

INTRODUCTION

The rms angle tracking error has a form which is relatively independent of the specific tracking implementation.

Angle tracking systems may be classified into the following four general categories:

Quadrant Angle Tracking Systems

Frame Scanning Angle Tracking Systems

Beam Lobing Angle Tracking Systems

Reticle Angle Tracking Systems

In a quadrant angle tracking system, the pulsed or continuous wave (CW) carrier is tracked by defocusing the received beam onto a four quadrant sensor. The relative strengths of the four sensor signals give the off-axis angle deviation.

A frame scanning system consists of a moving sensor describing a raster or spiral scan, or an array of point sensors such as vidicon elements which are sequentially examined. The coordinates of the detected image determine the tracking angle. Because of the time sampling nature of the sensor, the system is usually limited to use with CW carriers.

In a beam lobing system the received beam is focused to a spot which is mechanically rotated about its axis to illuminate four "cross-hair" slit sensors. The relative time position of detections from the sensors determines the tracking angle. Beam lobing systems are usually limited to operation with CW carriers.

The reticle system intensity modulates a received CW or pulsed beam by a rotating "pin wheel" type of transparency. The relative position of the beam on the reticle produces an AM or FM signal, depending on the reticle code. The modulation is detected using the scanning frequency and phase to yield the tracking angle and off axis magnitude.

The following topics present a description of the various angle tracking systems and an error analysis of their performance. One reservation must be made; the relations given are derived under the assumption of Gaussian statistics. This condition is achieved when the number of photons utilized in the decision process is sufficiently large to assume the law of large numbers. For low photon levels the relations will not be a function of the signal-to-noise ratio, but some function of the signal and noise power.

Angle Tracking Systems

- Quadrant
- Frame Scanning
- Beam Lobing
- Reticle

General Form of Angle Error Equation

$$\epsilon_{\theta} = \frac{k_{\theta}}{(\text{SNR})_v} \left(\frac{\Delta f_s}{f} \right)^{1/2}$$

INTRODUCTION

The random angular position error of an optical tracker is measured by the RMS tracking angle noise error, ϵ_{θ} , given by a function of the following form:

$$\epsilon_{\theta} = \frac{k_{\theta}}{(\text{SNR})_v} \left(\frac{\Delta f_s}{f} \right)^{1/2}$$

where

$(\text{SNR})_v$ = voltage signal-to-noise ratio

Δf_s = servo noise bandwidth

f = pulse repetition rate (or modulation frequency)

and

k_{θ} = modulation (resolution coefficient)

k_{θ} depends upon the nature of the specific tracker and is in a sense a measure of the ultimate geometrical accuracy limitation which the particular position encoder places upon the tracker, such as angular diameter of the Airydisk, angular width of the reticle slit, etc.

A description of the monopulse system used for the basic analysis is given in the following. In addition, CW system utilizing pulse position modulation (PPM), amplitude modulation (AM) and frequency modulation (FM) are discussed and the corresponding forms for these cases are given.

MONOPULSE QUADRANT TRACKING SYSTEM

The bias error due to unbalance of photomultiplier detectors is desired.

The monopulse system consists of a pulsed (laser) beacon and a quadrant photomultiplier (PM) tracker. The receiving optics form a diffraction limited spot at the apex of a four-sided prism which reflects portions of the blur circle onto separate photomultipliers, as shown in Figure A.

The error signal is derived by comparing the amount of energy in each of the four PM channels. In order to determine the error in the X and Y coordinates the following differences are found:

$$\epsilon \text{ Coordinate Error: } (S_1 + S_2) - (S_3 + S_4) = \Delta \epsilon$$

$$\theta \text{ Coordinate Error: } (S_1 + S_4) - (S_2 + S_3) = \Delta \theta$$

where S_i = signal from i th quadrant.

When the center of the "blur circle" falls at the apex of the beam splitting prism, the error signals are zero. Clearly both the absolute sensitivity of each of the PMs (considered in this topic) and the noise (considered in the next topic) will limit the ultimate accuracy of the system so that the center of the blur circle will perform excursions about this mean null position.

Bias Error Caused by Photodetector Gain Unbalance. The voltages generated at each of the photodetectors are given by

$$V_1 = \frac{H G_{D1}}{(\pi \delta^2 / 4)} [\pi \delta^2 / 16 + f(\epsilon, \theta)] \quad (1)$$

$$V_2 = \frac{H G_{D2}}{(\pi \delta^2 / 4)} [\pi \delta^2 / 16 + f(+\epsilon, -\theta)] \quad (2)$$

$$V_3 = \frac{H G_{D3}}{(\pi \delta^2 / 4)} [\pi \delta^2 / 16 + f(-\epsilon, -\theta)] \quad (3)$$

$$V_4 = \frac{H G_{D4}}{(\pi \delta^2 / 4)} [\pi \delta^2 / 16 + f(-\epsilon, +\theta)] \quad (4)$$

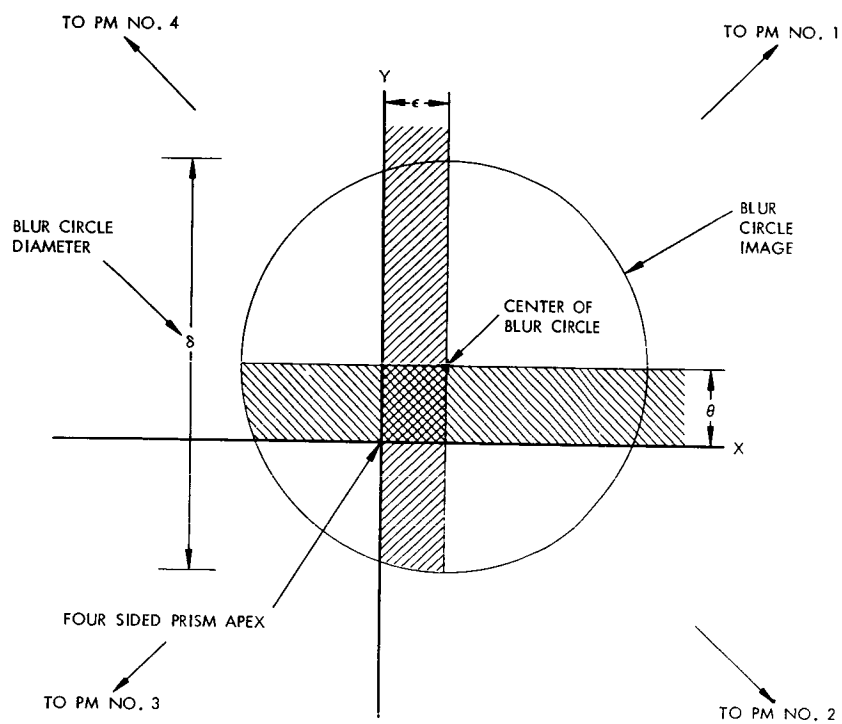


Figure A. Blur Circle and Quadrant Photomultiplier (PM) Geometry

MONOPULSE QUADRANT TRACKING SYSTEM

where H is the total radiant energy incident on the system, G_{Dn} is the gain of the n th photodetector, and δ is the diameter of the blur circle. The function, $f(\epsilon, \theta)$ is the incremental area of the blur circle in a quadrant as a function of ϵ, θ ; at boresight, $f(0, 0) = 0$. The bias error is determined by forming the steering signals and setting them equal to zero, thus,

$$\Delta \text{ azimuth} = (V_1 + V_2) - (V_3 + V_4) = 0 \quad (5)$$

$$\Delta \text{ elevation} = (V_1 + V_4) - (V_2 + V_3) = 0 \quad (6)$$

Substituting Equations (1) to (4) into Equations (5) and (6) yields after some algebraic manipulation

$$0 = 1/4 [G_{D1} + G_{D2} - G_{D3} - G_{D4}] + \frac{4}{\pi \delta^2} [G_{D1} f(\epsilon, \theta) + G_{D2} f(+\epsilon, -\theta) - G_{D3} f(-\epsilon, -\theta) - G_{D4} f(-\epsilon, +\theta)] \quad (7)$$

and

$$0 = 1/4 [G_{D1} - G_{D2} - G_{D3} + G_{D4}] + \frac{4}{\pi \delta^2} [G_{D1} f(\epsilon, \theta) - G_{D2} f(+\epsilon, -\theta) - G_{D3} f(-\epsilon, -\theta) + G_{D4} f(-\epsilon, +\theta)] \quad (8)$$

The exact function $f(\epsilon, \theta)$ is rather complex. However for small angles a good approximation is

$$f(\epsilon, \theta) \approx \frac{\epsilon \delta}{2} + \frac{\theta \delta}{2} \quad (9)$$

Substituting Equation (9) into Equations (7) and (8) yields

$$0 = [G_{D1} + G_{D2} - G_{D3} - G_{D4}] + b_o [G_{D1} (\epsilon + \theta) + G_{D2} (+\epsilon - \theta) - G_{D3} (-\epsilon - \theta) - G_{D4} (-\epsilon + \theta)] \quad (10)$$

and

$$0 = [G_{D1} - G_{D2} - G_{D3} + G_{D4}] + b_o [G_{D1} (\epsilon + \theta) - G_{D2} (+\epsilon - \theta) - G_{D3} (-\epsilon - \theta) + G_{D4} (-\epsilon + \theta)] \quad (11)$$

where b_o is the modulation index given by

$$b_o = \frac{8}{\pi \delta} \quad (12)$$

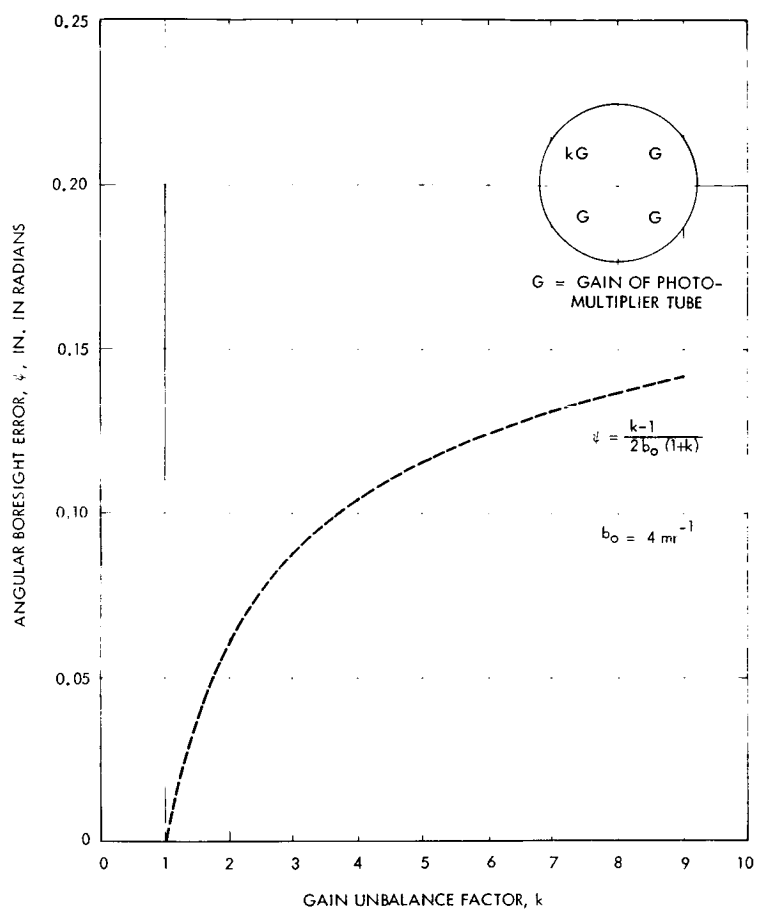


Figure B. Total Error Bias, ψ , Produced by PMT Gain Unbalance

MONOPULSE QUADRANT TRACKING SYSTEM

Equations (10) and (11) are a set of linear simultaneous equations in ϵ and θ which may be easily evaluated for any photodetector gains. For example let

$$G_{D1} = G_{D2} = G_{D3} = G$$

and

$$G_{D4} = kG \quad (13)$$

Solving for the bias gives

$$\epsilon = \frac{k - 1}{2b_o (1 + k)}$$

and

(14)

$$\theta = - \left[\frac{k - 1}{2b_o (1 + k)} \right]$$

The total bias angle is

$$\psi = \sqrt{\epsilon^2 + \theta^2} = \frac{k-1}{\sqrt{2} b_o (1 + k)} \quad (15)$$

where

ψ = total error bias produced by the gain change

k = gain unbalance factor

b_o = the modulation coefficient before the gain change occurred.

This equation takes into account the elevation and azimuth component biases and the effect of the gain factor on the modulation coefficient, b_o , of the system. The result is plotted in Figure 3 for $b_o = 4 \text{ mr}^{-1}$. The gain balancing techniques presently used in the PMTs will hold the gain factor to below 1.7 and hence the bias error to less than 0.05 mr for this example.

ANGLE NOISE ANALYSIS OF MONOPULSE QUADRANT TRACKING SYSTEM

Signal and noise relationships for a monopulse quadrant tracking system are derived.

A block diagram of a typical monopulse quadrant tracking system is shown in the figure. It consists of entrance optics which form a diffraction limited spot at the apex of a four sided prism which reflects portions of the blur circle onto four separate photomultipliers. The photomultipliers convert the optical signal into electrical signals which are then processed by a series of amplifiers and filters. The four quadrant signals are used to derive the azimuth and elevation tracking error signals by combining their sums and differences. These error signals in turn control the entrance optics in such a manner as to seek their cancellation. The object of this section is to consider the angular accuracy of which such a monopulse trackup system is capable. This will be done in the following steps. First the process by which the beacon signal is converted to electrical tracking error signals is described. Second, the effect of noise being introduced in this process is observed. Third, the loop error signal is derived and fourth, the angular noise error spectrum is evaluated.

Signal Process

The optical signal power captured by the entrance optics is

$$S_T = H(t)A_O \quad (1)$$

where $H(t)$ is the beacon irradiance and A is the optics aperture. Assuming small position errors the signal captured by any one of the four photomultipliers is given by

$$S_i = \frac{H(t)A_O}{\frac{\pi \delta^2}{4}} \left[\frac{\pi \delta^2}{16} + \left(\epsilon_i(t) + \theta_i(t) \right) \frac{\delta}{2} \right] \quad (2)$$

where δ is the diameter of the blur circle and $\epsilon_i(t)$ and $\theta_i(t)$ are the time dependent quadrant position errors of the blur circle defined in the figure. To simplify the analysis and with no loss in generality it will be assumed that only an azimuth error, ϵ_i , exists. Thus for $\theta_i = 0$ the quadrant signals are

$$S_1 = S_2 = H(t)A_O \left[\frac{1}{4} + \frac{2}{\pi \delta} \epsilon(t) \right] \quad (3)$$

$$S_3 = S_4 = H(t)A_O \left[\frac{1}{4} - \frac{2}{\pi \delta} \epsilon(t) \right]$$

The photomultiplier converts the optical signal power into electrical voltages which are then amplified by the AGC amplifiers. The signal is then filtered. If the optical beacon were cw, a simple low pass filter with sufficient bandwidth to follow the signal fluctuations induced by the tracker would be used. If the beacon is pulsed the filter would consist of a sample and hold circuit consisting of a switch which is turned on and off at the beacon pulse rate, and a low pass filter. At the filter output the LaPlace transform of the signal voltages are

$$E_i(s) = G_D B(s) S_i(s) \quad (4)$$

where G_D is the gain of the photomultiplier assumed to be independent of frequency and $S_i(s)$ and $B(s)$ are the respective Laplace transforms of the optical signal in the i^{th} quadrant and the filter.

The Laplace transform of the voltage signals thus depends on the beacon modulation and the type of filter employed. If the beacon is a cw signal of irradiance H , the Laplace transform is

$$E_i(s) = G_D A B(s) H A_o \left[\frac{1}{4s} \pm \frac{2}{\pi \delta} \epsilon(s) \right] \quad (5)$$

where $\epsilon(s)$ is the Laplace transform of the two dependent azimuth position error. If the beacon is a pulsed signal and the detector filter consists of a sample and hold type network, the signal voltage is

$$E_i(s) = \left\{ P \right\} G_D A A_o \left[\frac{1}{4s} \pm \frac{2}{\pi \delta} \epsilon(s) \right] \quad (6)$$

where $\left\{ P \right\}$ is the value at which the holding network peaks and is given by

$$\left\{ P \right\} = \left\{ H(t) \oplus B(t) \right\}_{PK} \quad (7)$$

In both the cw and pulsed beacon cases, the signal voltage can be expressed as

$$E_i(s) = K \left[\frac{1}{4s} \pm \frac{2}{\pi \delta} \epsilon(s) \right] \quad (8)$$

where K is a constant derived from either equations (5) or (6).

After filtering, the signal voltages are combined in the sum and difference networks. The sum signal is given by

$$\sum_i E_i(s) = \frac{K}{s} \quad (9)$$

Finally the difference signal is given by

$$\Delta E_i(s) = (E_1 + E_2) - (E_3 + E_c) = \frac{8}{\pi \delta} K \epsilon(s) \quad (10)$$

ANGLE NOISE ANALYSIS OF MONOPULSE QUADRANT TRACKING SYSTEM

Noise Analysis

Together with the signal, the photomultiplier generates an irreducible noise voltage which like the signal voltage, is amplified and filtered. The noise power at the output of the photomultiplier will be assumed uniformly distributed across the frequency spectrum of the signal. Its spectral density is thus independent of frequency and of N_D volt² per cycle. At the filter output, the Laplace transform of the mean square noise voltage is given by

$$\overline{V_n(s)}^2 = N_D B(s) B(-s) \quad (11)$$

If the filter has a noise equivalent bandwidth W_N , the mean square noise voltage is

$$\overline{V_n(t)}^2 = N_D W_N \quad (12)$$

The power signal to noise ratio of the total beacon signal received is obtained from (10) and (12)

$$(SNR) = \frac{(\sum E_i(t))^2}{V_n(\epsilon)^2} = \frac{K^2}{N_D W_N} \quad (13)$$

Loop Error Signal

Since all of the devices following the photomultiplier are linear, the signal and noise voltages are added to one another. The total difference network azimuth error signal is

$$\Delta E_T = \Delta E_i(s) + \sqrt{4V_n(s)^2} \quad (14)$$

This signal is transformed to the azimuth pointing error angle $\epsilon(s)$ by the servo. The angle is then given by

$$\epsilon(s) = -\Delta E_T G(s) \quad (15)$$

where $G(s)$ is the transfer function of the feedback loop. Substituting (10) and (14) into (15) yields

$$\epsilon(s) = -\left[K \frac{8}{\pi \delta} \epsilon(s) + 2 \sqrt{V_n(s)^2} \right] G(s) \quad (16)$$

Solving for the azimuth pointing angle gives

$$\epsilon(s) = \frac{-2(G(s)\sqrt{V_n(s)^2}}{1 + KG(s)\frac{8}{\pi\delta}} \quad (17)$$

RMS Angular Noise

The RMS angular noise may be obtained from

$$\left| \overline{\epsilon(j\omega)} \right|^2 = \frac{1}{W_o} \int_0^{W_o} \left| \epsilon(j\omega) \right|^2 d\omega \quad (18)$$

where W_o is the equivalent noise bandwidth of the feedback loop transfer function and

$$\epsilon(j\omega)^2 = \frac{4G(s)^2 \overline{V_n(s)^2}}{\left[1 - KG(s)\frac{8}{\pi\delta} \right]^2} \bigg|_{s=j\omega} \quad (19)$$

It is interesting to note that if the servo loop is designed to make

$$KG(s)\frac{8}{\pi\delta} \gg 1$$

equation (19) becomes

$$\left| \epsilon(j\omega) \right|^2 = \frac{1}{\frac{8}{\pi\delta}^2 \frac{K^2}{\overline{V_n(s)^2}}} \quad (20)$$

which in turn, after substitution of (13) becomes

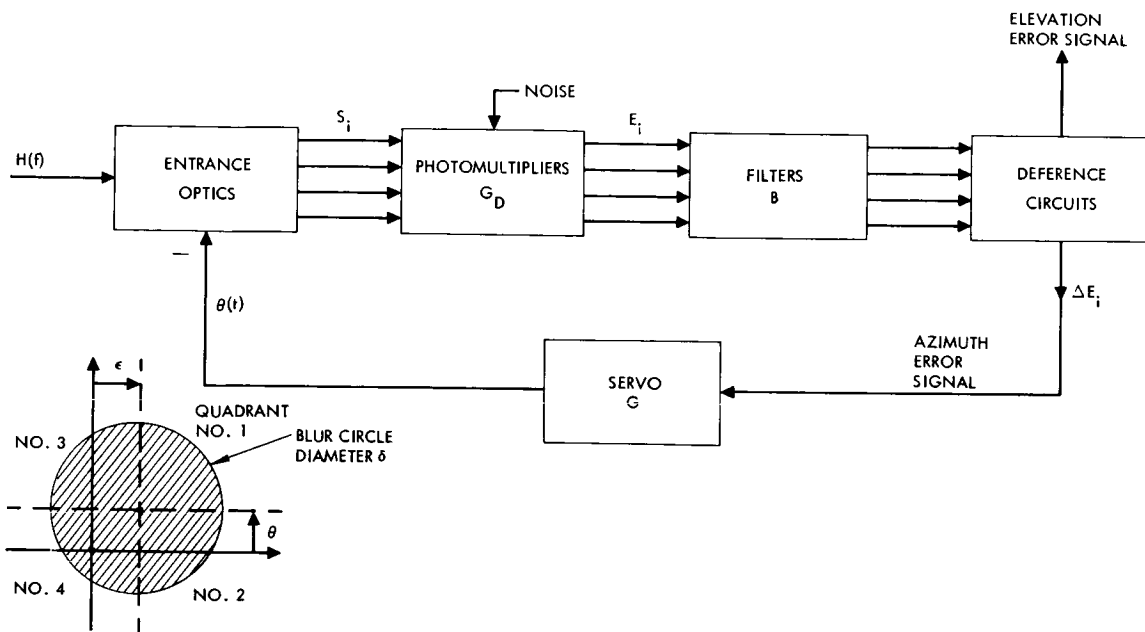
$$\left| \epsilon(j\omega) \right|^2 = \frac{1}{\left(\frac{8}{\pi\delta} \right)^2 (\text{SNR})} \quad (21)$$

ANGLE NOISE ANALYSIS OF MONOPULSE QUADRANT TRACKING SYSTEM

Substitution of (21) into (18) yields the familiar result

$$\left| \bar{\epsilon} \right|^2 = \frac{1}{b_o^2 (\text{SNR})}$$

where b_o is modulation index defined as $8/\pi\delta$.



Monopulse Quadrant Tracking System Block Diagram

BEAM LOBING PPM TRACKING SYSTEM

The total tracking error is derived for a PPM tracker.

The operation of the beam lobing pulse position modulation (PPM) tracking system is illustrated in Figures A, B, and C. The image of a point source (or target) at boresight is deflected by a rotating prism (Figure A) so that it describes a periphery just touching the outer ends of the sensor cells (dotted circle in Figure B). The image of a target off boresight, but within the field of view, is described by a periphery with the same radius as before, but centered at the point where the image would be without the prism. When the target image passes over a cell (slit) a pulse, called a "target pulse," is realized. The duration, τ , of the target pulse is the time the nutating image requires to cross the cell. The position of the target pulse thus generated is varied with respect to four fixed and equally-spaced reference pulses. From Figures B and C, it is clear that at least two target pulses are needed for locating the target position with respect to boresight, giving the yaw (x) and pitch (y) angles. There are four separate detection channels — one corresponding to each cell.

The pulse position modulation, measured between the reference pulses and the signal pulses (see Figure C), is used to obtain the pitch and yaw signals. The error in tracking is due to the error in measuring the time at which the blur circle passes the sensor.

From Figure C it is seen that the error in the x axis, x , is:

$$x = R \sin \omega t \quad (1)$$

and

$$\Delta x = -\Delta t \left[R \omega \cos \omega t \right] \quad (2)$$

when the tracking error is small, i.e., ωt is close to zero,

$$\cos \omega t \cong 1$$

and

$$\Delta x = -R \omega \Delta t = -\frac{2\pi R}{T} \Delta t \quad (3)$$

where

T is the period of nutation

Δt is an error in determining the time when the light spot crosses the detector due to noise

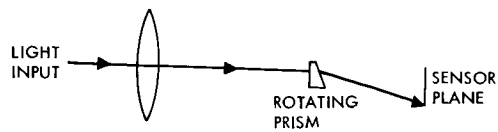


Figure A. Optical Schematic of Beam Lobing PPM System

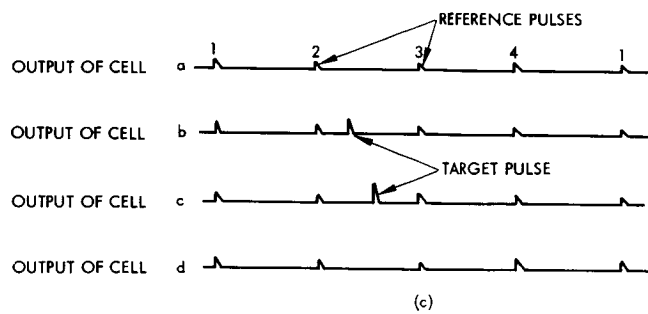


Figure B. Image Plane Geometry of PPM Beam Lobing System

BEAM LOBING PPM TRACKING SYSTEM

The noise uncertainty, Δt , can be reduced to a minimum value (for a given signal to noise ratio) by using a matched filter. To find the value of Δt , consider first the time required for the blur circle to cross the detector, τ .

$$\tau = \frac{\Delta l + f}{\omega R} \quad (4)$$

for an optimum filter Δt is then

$$\Delta t = \sqrt{\frac{N}{S_P}} \sqrt{\frac{1}{\frac{3}{2\tau}}} = \sqrt{\frac{N}{S_P}} \sqrt{\frac{2}{3}} \sqrt{\frac{\Delta l + f}{\omega R}} \quad (5)$$

where N is the noise power spectral density at the output of the photo-detector and S_P is the peak signal measured at the same place.

The position uncertainty in the x direction Δx , is then

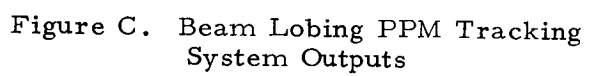
$$\Delta x = -\frac{2}{3} \frac{\sqrt{\omega R(\Delta l + \delta)}}{S_P / \sqrt{N}} \quad (6)$$

and the uncertainty error in the y direction will be the same. The total error, $\epsilon_P = \sqrt{2} \Delta x$ since $\tau = (\omega/2\pi)$

$$\epsilon_P = \frac{4}{3} \sqrt{\pi(\Delta l + \delta)R} (1/T) (S_P / \sqrt{N}) \quad (7)$$

Equation (7) holds for white thermal noise. In the background limited case, noise is proportional to the square root of the signal and the total error is

$$\epsilon_P = \frac{4}{3} \sqrt{\pi(\Delta l + \delta)R} (1/\sqrt{S_P T}) \quad (8)$$



AM RETICLE TRACKING SYSTEM

The implementation and performance of an AM reticle tracking system is described.

Figure A illustrates a block diagram of an AM Reticle tracking system. The received beam passes through a rotating reticle such as the one shown in Figure B and is focused on the photodetector. The reticle intensity modulates the beam in such a manner that the angular position of the beam from the reticle axis, which is boresighted to the optical reference system, may be determined by the electrical detection system following the photodetector. With the reticle shown in Figure B, if the beam is not passing through the center of the rotating reticle, the radiant power on the detector surface will be of the general form of Figure C. The frequency spectrum of this signal consists of a fundamental modulation frequency dependent upon the angular slit spacing and angular velocity of the reticle, plus harmonic sidebands. The narrow band filter passes only the fundamental and the first pair of sidebands. Next, the demodulator shifts the spectrum to zero frequency. The filtered output of the demodulator ideally is a sine wave whose amplitude is a function of the radial displacement of the beam, and whose phase is proportional to the beam displacement along orthogonal axes. The reference for the phase detectors is the frequency corresponding the reticle rotation rate. The beam displacement voltages are used to control servo motors which reposition the optical sensor to the beam center.

The performance analysis of the AM tracking is complicated by the fact that the signal modulation is not a monotonic function of the radial error. As the beam moves from the center of the rotating reticle to the outer edge, initially the target spot will be partially covered, then partially uncovered by slits resulting in only partial modulation. Furthermore, at certain radial positions the rate of covering and uncovering of the spot will be nearly equal and the percentage of modulation will drop. These modulation perturbations, however, are reasonably small and may be ignored in a first order analysis.

Even assuming Gaussian noise statistics at the optical detector output, the analysis of the angle error fluctuations is difficult because the electrical receiver contains nonlinear elements. It has been shown¹ that the tracking error due to detector noise is

$$\text{Large Signal Input } \epsilon_t = \frac{K\pi a_i \left(\frac{f_c}{f_m} \right)}{2 \left(\frac{S}{N} \right)_v}$$

¹P. E. Mengers, "Tracking Accuracy of Infrared Trackers," General Electric Report No. R59ELC100 Defense Electronics Division, December 1959.

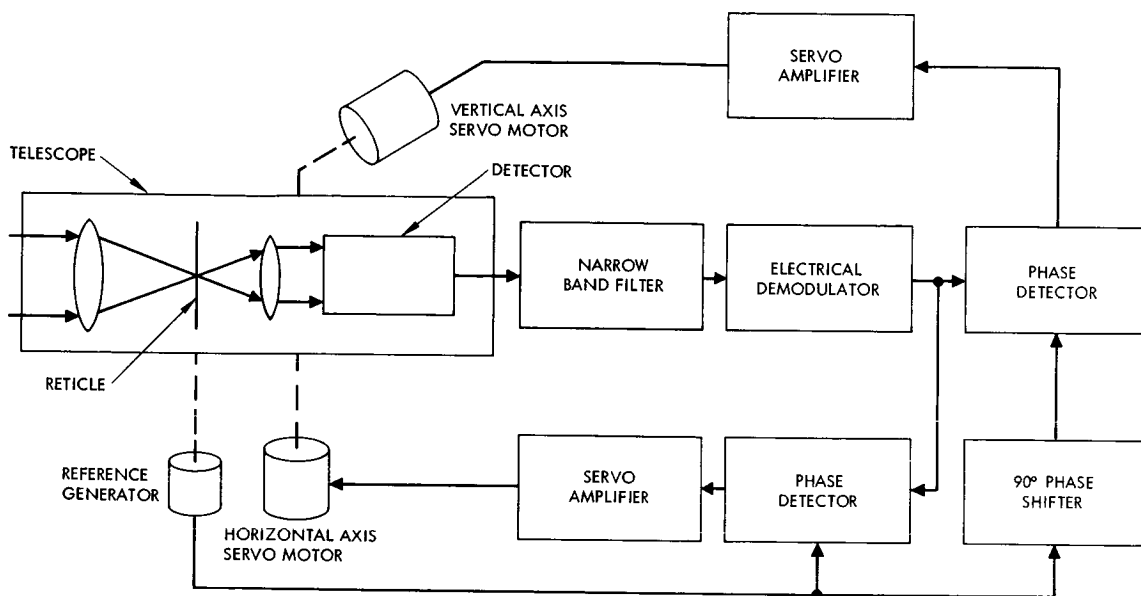


Figure A. AM Tracking System, Block Diagram

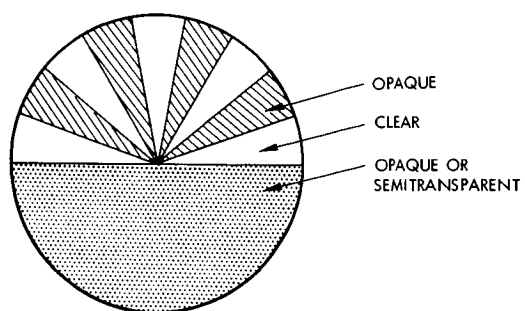


Figure B. Simple Optical
AM Reticle

AM RETICLE TRACKING SYSTEM

$$\text{Small Signal Input } \epsilon_t = \frac{K \pi a_i \left(\frac{f_c}{f_m} \right)}{2 \left(\frac{S}{N} \right)_v^{1/2} \left(2^{1/8} \right)}$$

where

a_i = angular size of image

f_c = reticle rotational frequency

$K = 0.64$

f_m = modulation frequency

$\left(\frac{S}{N} \right)_v$ = receiver output voltage signal-to-noise ratio

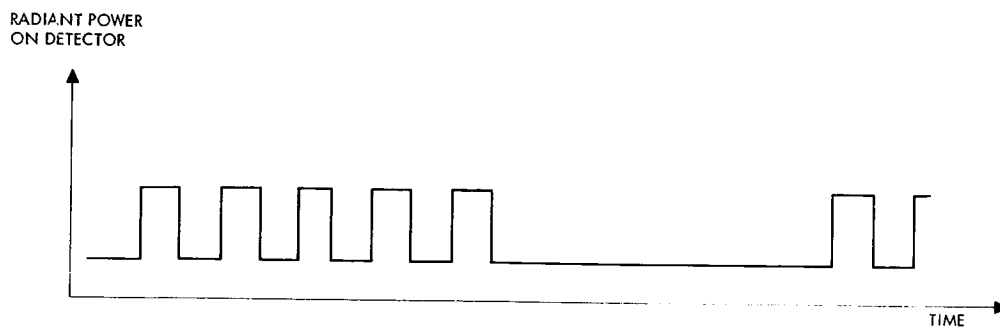


Figure C. Radiation Signal After Modulation by AM Reticle

FM RETICLE TRACKING SYSTEM

Typical FM reticle implementation is shown and the noise error signals are documented.

An FM reticle tracker operates in a similar manner to the AM reticle tracker. One type of FM reticle consists of slits which are shifted slightly as a function of the radial distance from the center of the reticle. The result is a frequency modulation about the mean chopping frequency. Another type of FM reticle consists of a reticle with evenly spaced slits rotated off-axis to the received beam as shown in Figure A. Figure B illustrates the relationship of the instantaneous frequency to the angular position. The tracking error fluctuations have been shown¹ to be:

$$\text{Small Signal Case } \epsilon_t = \frac{2 a_m \left(\frac{\omega_d}{\omega_F} \right)}{\left(\frac{S}{N} \right)}$$

$$\text{Large Signal Case } \epsilon_t = \frac{2 a_m}{\left(\frac{S}{N} \right)}$$

where

a_m = maximum angular displacement of beam

$\Delta F_i = \frac{\omega_F}{2\pi}$ = IF input filter bandwidth (halfwidth)

$\Delta F_d = \frac{\omega_d}{2\pi}$ = predetection filter bandwidth (halfwidth)

$\left(\frac{S}{V} \right)$ = receiver output voltage signal to noise ratio.

¹P. E. Mengers, "Tracking Accuracy of Infrared Trackers," General Electric Report No. R59ELG100 Defense Electronics Division, December 1959.

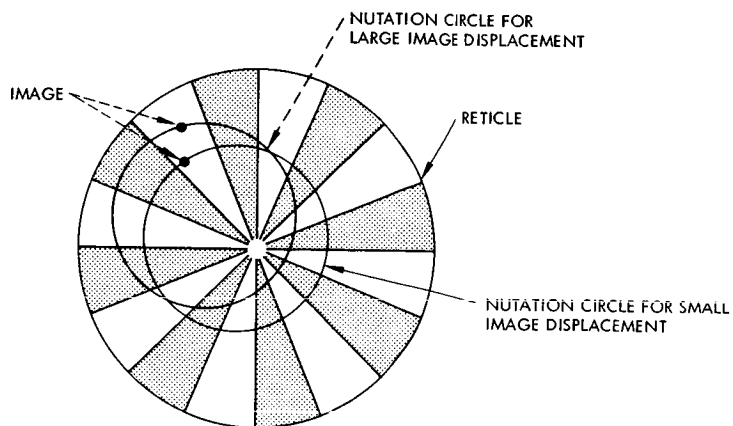


Figure A. Nutation Circle and Reticle
Relative Geometry

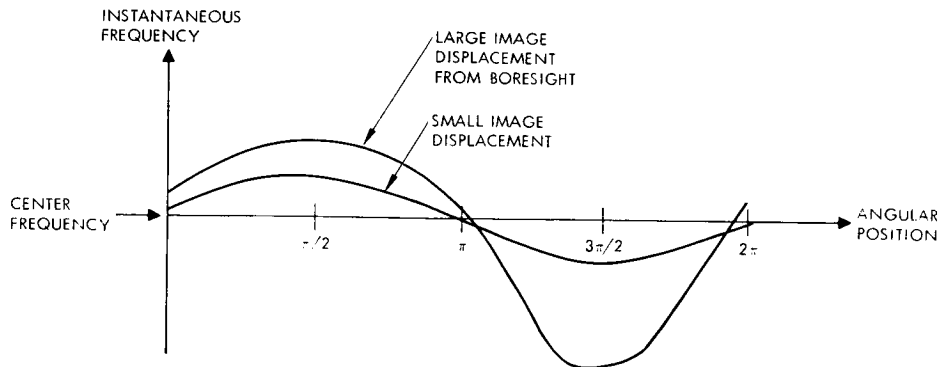


Figure B. Instantaneous Frequency as Function
of Rotation of Reticle Axis About Boresight
for a FM Precessional Reticle

COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS

Attitude and Tracking Sensors

	Page
Sun Sensors	424
Conversion Chart for Angular Measure	428
Star Sensors	430
Star Tracker Detectors	434
Planet Sensors	436

SUN SENSORS

Attitude sensors are introduced and the first of these, sun sensors, are described.

Active sensors are required for stabilization since gyroscope devices are unsuitable for extended missions due to their short life (under 1000 hours typically) and unpredictable drift errors. Similarly ambient field sensors e. g., magnetic field, are unsuitable because field lines are not predictable with sufficient accuracy and are too weak to be of use in interplanetary space.

Three basic types of attitude and tracking sensors are of interest with respect to the acquisition and tracking systems. They are:

1. Sun Sensors
2. Star Sensors
3. Planet Sensors

Sun sensors are described in this topic while star and planet sensors are described in subsequent topics.

The sun is the most common attitude reference for non-earth oriented vehicles. The principal advantage of the sun as a reference is the relative ease with which it may be acquired as a consequence of its high intensity. One significant disadvantage is that solar activity may shift the center of radiance by as much as 0.75 arc second.** Among the various types of sun sensors which have been used to date are the following.

Shadow Masked Sun Sensors. These consist of shadow masked arrays of photovoltaic or photoconductive cells as indicated in Figure A. The cells are connected as differential pairs such that the output electrical signal changes sign at a center null point. These sensors are small and simple and are capable of null accuracies of the order of 0.1 degree. The chief sources of error are stray light and unequal aging and thermal drift between cells.

Lens Type Sun Sensors. The important design parameters of lens type sun sensors are the focal length and the distance between the reticle and the detector, indicated in Figure B. The use of a lens allows an "angular gain" over that obtainable with the shadow mask. Two units such as are indicated in Figure B are used to provide null positioning.

Null accuracies from 0.01 to 0.1 degree can be obtained using these sensors.

*Much of the material of this section is found in "Optical Attitude Sensors for Space Vehicle Applications: Descriptive Survey of Recent Literature and Error Studies" M.S. Thesis by James Harold Spotts, 1965, UCLA.

**Observed from 1AU.

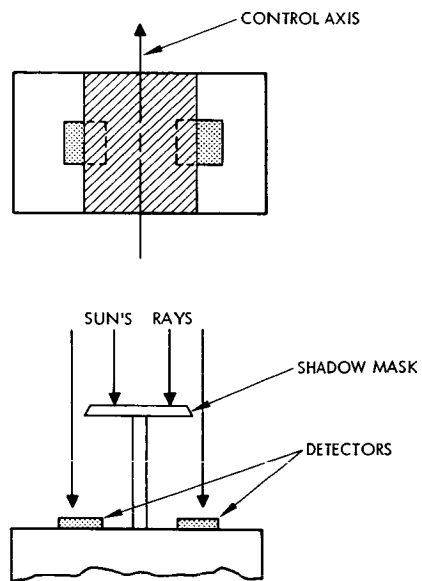


Figure A. Simple Shadow Mask Sun Sensor

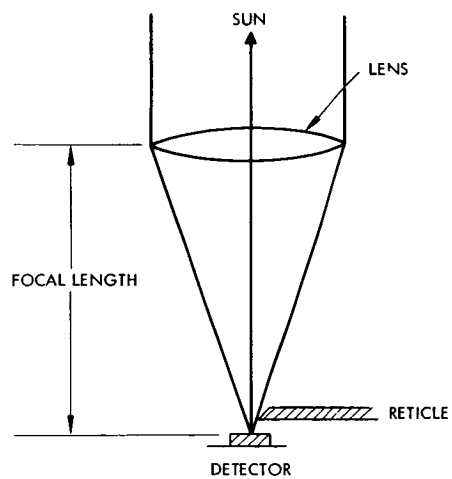


Figure B. Simple Lens Type Sun Sensor

Component Performance and Burden Relationships Attitude and Tracking Sensors

SUN SENSORS

"Critical Angle Prism" Sun Sensor.¹ Generally for higher accuracies, more complex systems are needed. However, null accuracies from 0.001 to 0.01 degree can easily be obtained using the critical angle prism sensor, a simple device composed of a glass prism and two photovoltaic cells, shown in Figure C. This sensor makes use of the fact that when the sun is in the null plane, almost all of the light will be reflected in the prism and will not reach the photo detector. Due to the critical angle, deviations from the null create a sharp error signal in the differentially connected photo cells. Null accuracies of the order of 30 arc seconds have been obtained. Sun sensors of this type have been proposed for use in the AOSO and claim accuracies as 1.3 arc seconds or 0.00036 degrees.

Digital Sun Sensors. Many stabilized vehicles do not require a sun sensor for attitude stabilization but use the sun as a reference for initial attitude determination. The digital sun sensor, shown in Figure 7, actually encodes the sun angle for digital communication. If the field of view of the digital sensor is 0 degrees and if n cells are used, then the resolution is $1/2^n$ degrees. The only limiting factor is the angle subtended by the sun, approximately 0.5 degree at 1AU.

Two digital sun sensors are used on the Saturn Meteoroid Satellite and on the gravity gradient stabilized version of the ATS. They have field of view of $128^\circ \times 128^\circ$ and 1 degree resolution.

¹Seward, Harold H. "A Sunfinder for an Interplanetary Vehicle." Massachusetts Institute of Technology Instrumentation Laboratory Report, E-965 (Revision A), Dec. 1960.

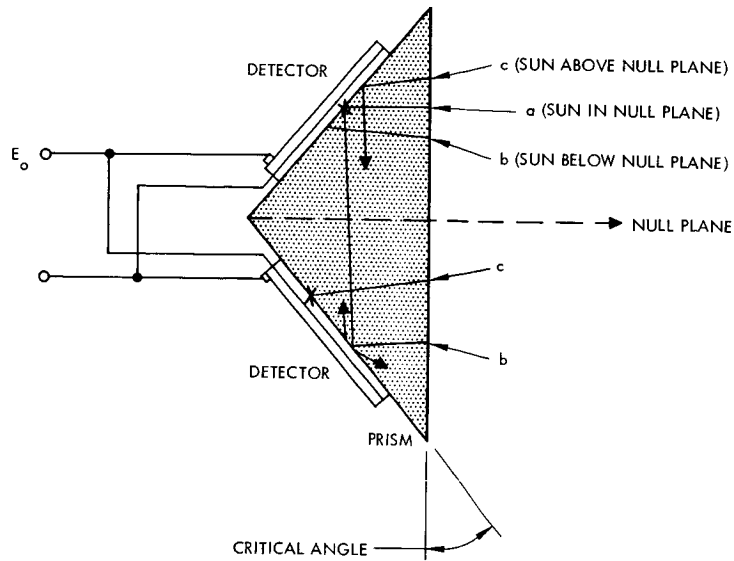


Figure C. Critical Angle Prism Sensor Design

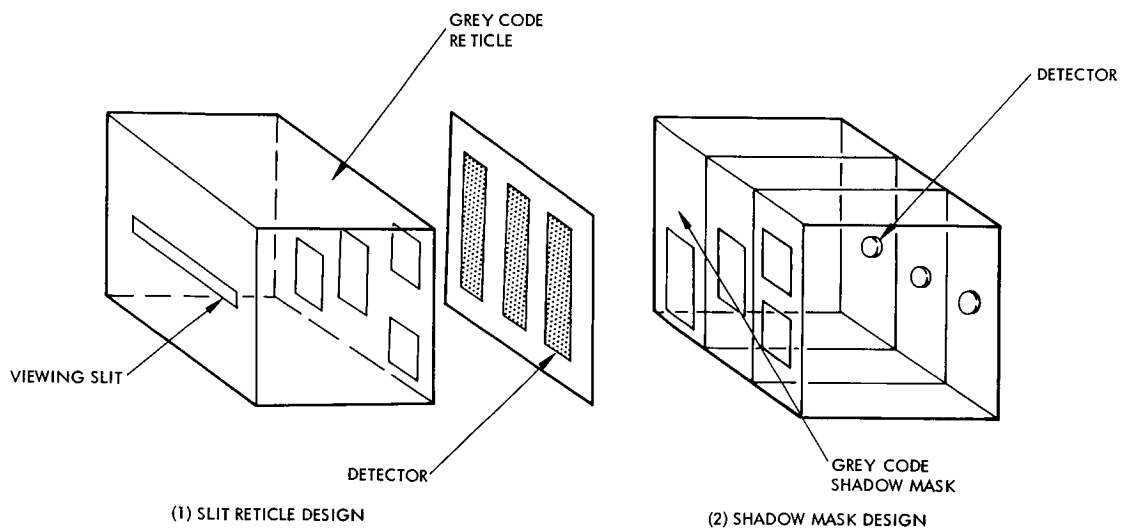
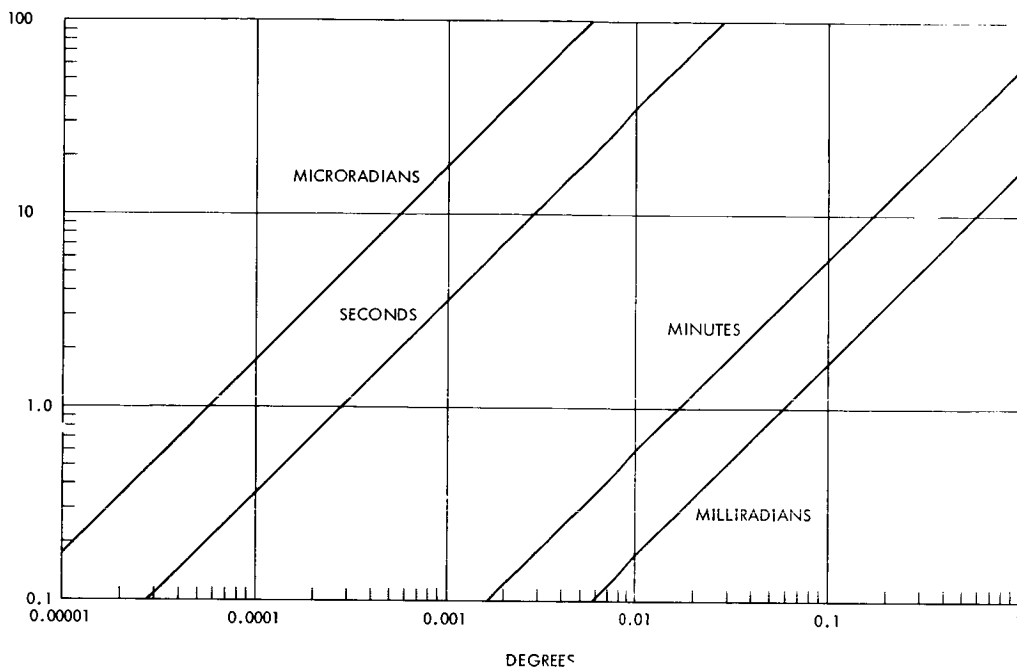


Figure D. Digital-Type Sun Sensors

CONVERSION CHART FOR ANGULAR MEASURE

A conversion chart for degrees, minutes, seconds, milliradians and microradians is given.

In this section various forms of angular measure are used. While each usage is appropriate, relative sizes are not always appreciated. For this reason a single conversion chart has been constructed which relates degrees to minutes, seconds, milliradians and microradians.



Conversion Chart for Angular Measure

STAR SENSORS

Scanning techniques used with conventional photomultiplier are described and accuracy vs weight is given for certain space designs.

There are three principal classes of star sensors, according to the type detector used: (1) the conventional photomultiplier, (2) the image detector, and (3) the quadrant photomultiplier. Photomultiplier tube tracking are discussed in this topic. Further description of photomultipliers, image dissectors and quadrant photomultipliers are discussed in the next topic.

A star appears essentially as a point source. (The apparent angular size of useful stars ranges from 0.0068 arc second for Sirius to 0.0410 arc second for Antares.) Hence the limitations of star tracker accuracy are mainly due to the background noise and internal instrument errors. By using larger optics, the signal to noise ratio (S/N) of the sensor output can be increased and higher angular discrimination achieved. Additional discrimination can be obtained by measuring differences in intensities and spectral densities. Also several sensors may be used to discriminate by recognizing a certain pattern of stars.

Initial acquisition is difficult due to the small angular size of the stars and the small angular beam of the star tracker. Scanning during acquisition is required and may be provided internally in the tracker or by maneuvering the vehicle. Mirror and vibrating reed scanning techniques are depicted schematically in Figures A and B. The table lists detailed specifications of some typical instruments. Values from this table and other data is plotted in Figure C.

Comparison Matrix of Star Sensors

Candidate Instrument	Performance (at Synchronous Altitude)	Interface Considerations	Reliability	Weight, Power and Size
Star (Polaris) sensors (Polaris is a +2.1 magnitude star)				
1. IIT Federal Labs. OAO Bore-sighted Star Tracker	<ol style="list-style-type: none"> 1. Star magnitude sensitivity +6 2. Field of view 10 arc minutes 3. Accuracy (rms) ± 1.5 arc seconds 4. Gimbaled (electronically) ± 1.5 degrees 5. Two axis capability 15 arc second steps 	<ol style="list-style-type: none"> 1. Small field-of-view acquisition may be difficult. Also may be difficult to constrain vehicle motions within field of view in normal operation 2. Precision and gimbaling capability not required. 3. High weight of system. 	<ol style="list-style-type: none"> 1. To be used on OAO. 2. Reliability number not available. 3. All digital circuitry. 	<ol style="list-style-type: none"> 1. Weight 25 pounds 2. Power ± 28 volts dc 3. Size (diameter) 3 x 15 inches 5 x 11 x 12 inches electronics
2. IIT - Dual Mode Star Tracker	<ol style="list-style-type: none"> 1. Sensitivity 2. Field of view 8 x 8 degree acquisition, 32 x 32 minute track 3. Accuracy (rms) ± 5 arc seconds 4. Not gimbaled 5. Two axis capability 	<ol style="list-style-type: none"> 1. Adequate field of view for acquisition. 2. More precise possibly than required. 	<ol style="list-style-type: none"> 1. No moving parts - all electronic gimbaling 2. Reliability number not available. 	<ol style="list-style-type: none"> 1. Weight 9.5 pounds 2. Power ± 26 volts dc, 8.0 watts 3. Size 5 x 10.5 x 5 inches
3. IIT - Electro-Optical Housing Head for Star Tracking	<ol style="list-style-type: none"> 1. Sensitivity 2. Field of view 3. Accuracy (rms) ± 1.0 x 1.0 degree 4. Not gimbaled 5. Two axis capability 	<ol style="list-style-type: none"> 1. Small field of view relative to 2 above - acquisition more difficult. 		
4. Kollsman Instrument Company. Gimbaled Seeker	<ol style="list-style-type: none"> 1. Sensitivity 2. Field of view 3. Accuracy (rms) ± 5 arc seconds 4. Gimbaled (mechanically) 5. Two axis capability 	<ol style="list-style-type: none"> 1. Excessive weight - high precision and gimbaling not required for Polaris one or two axis sensing. 	<ol style="list-style-type: none"> 1. OAO flight instrument 2. Reliability number not available. 	<ol style="list-style-type: none"> 1. Weight 42.5 pounds 2. Power 12.9 watts 3. Size Not available
5. Bendix Corporation Star Sensor (Proposed for SERT-III Mission)	<ol style="list-style-type: none"> 1. Sensitivity 2. Field of view 10 x 10 degrees 3. Accuracy <0.1 degree 4. Not gimbaled 5. Single axis design 	<ol style="list-style-type: none"> 1. Proposed sensor with off-the-shelf components - sensor development program necessary to a certain degree. 	<ol style="list-style-type: none"> 1. Proposed design utilizing off-the-shelf components - hence reliability not available 2. Reliability development program required. 	<ol style="list-style-type: none"> 1. Weight 5.0 pounds 2. Power 8.0 watts 3. Size 4 x 7 x 10 inches
6. IIT Sensor Design (Proposed for SERT-III)	<ol style="list-style-type: none"> 1. Sensitivity 2. Field of view 10 x 10 degrees 3. Accuracy 0.012 degree 4. Single axis design 5. Information not available. Sensitivity capability possibly questionable. 	<ol style="list-style-type: none"> 1. Similar to (2) above except single axis; however, some development required. 	<ol style="list-style-type: none"> 1. Proposed design utilizing off-the-shelf components, reliability not available. 	<ol style="list-style-type: none"> 1. Weight 8.0 pounds 2. Power Not available 3. Size 2-1/4 diameter x 8.5 inches

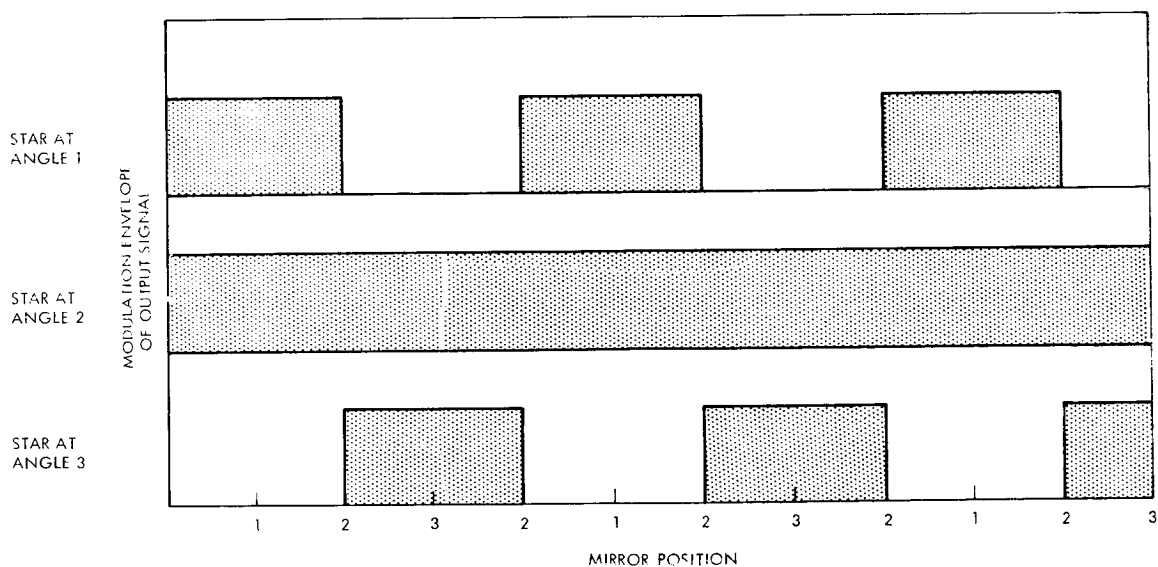
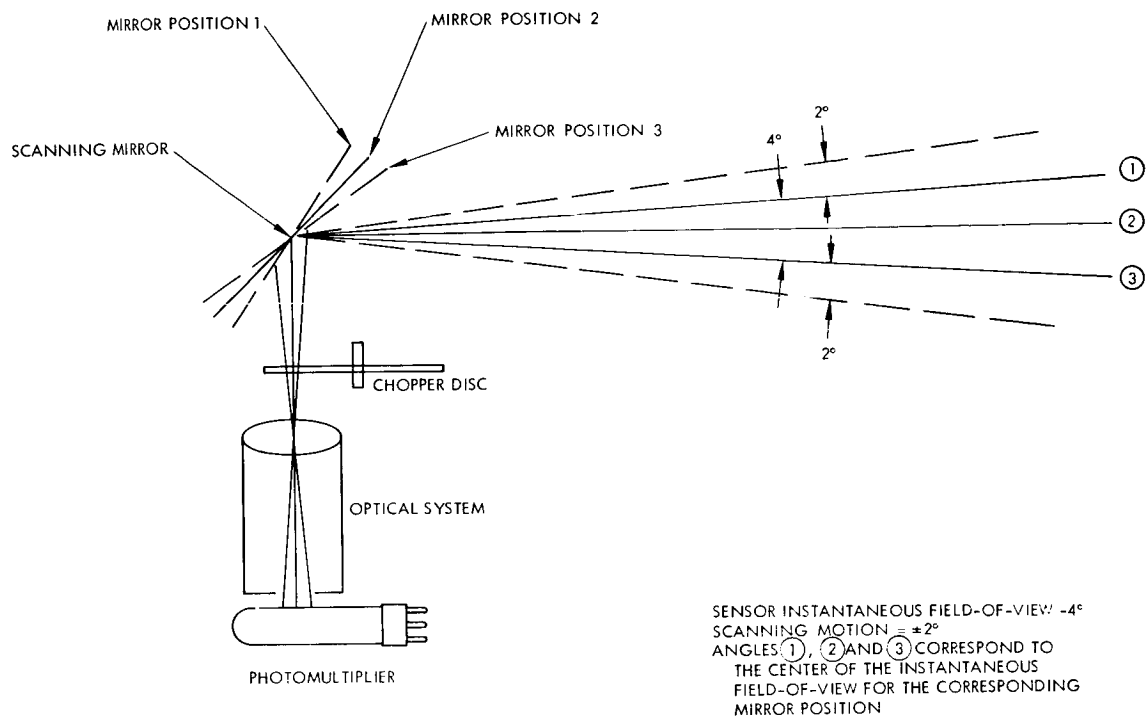


Figure A. Scanning Mirror Star Sensors

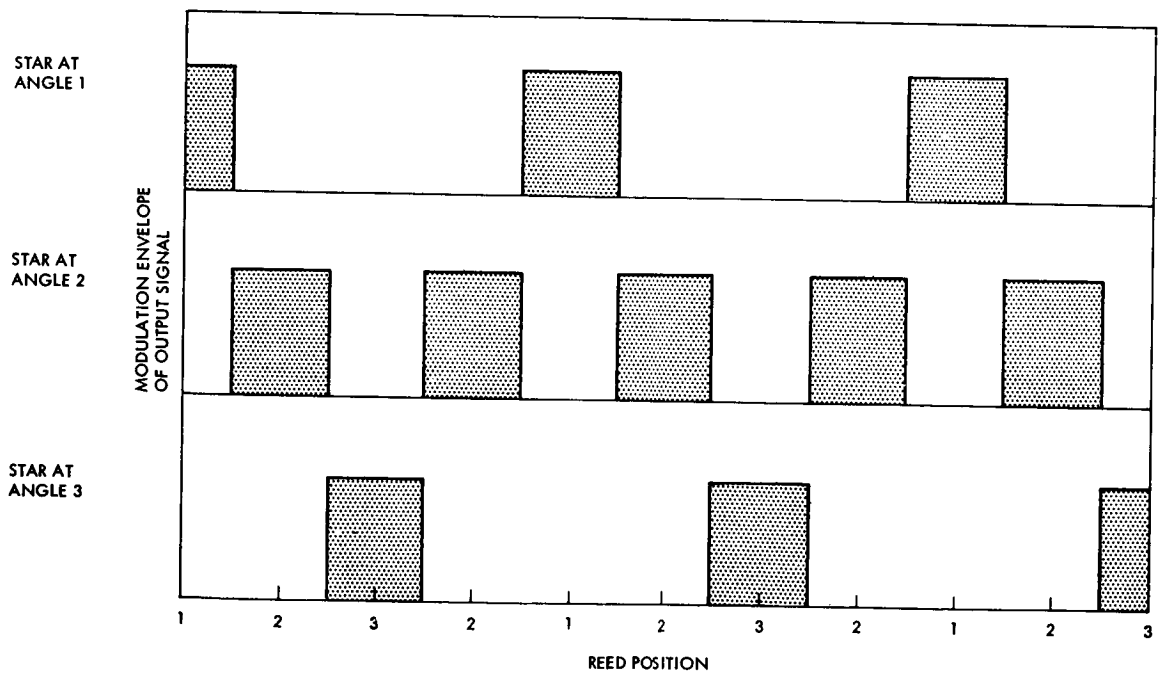
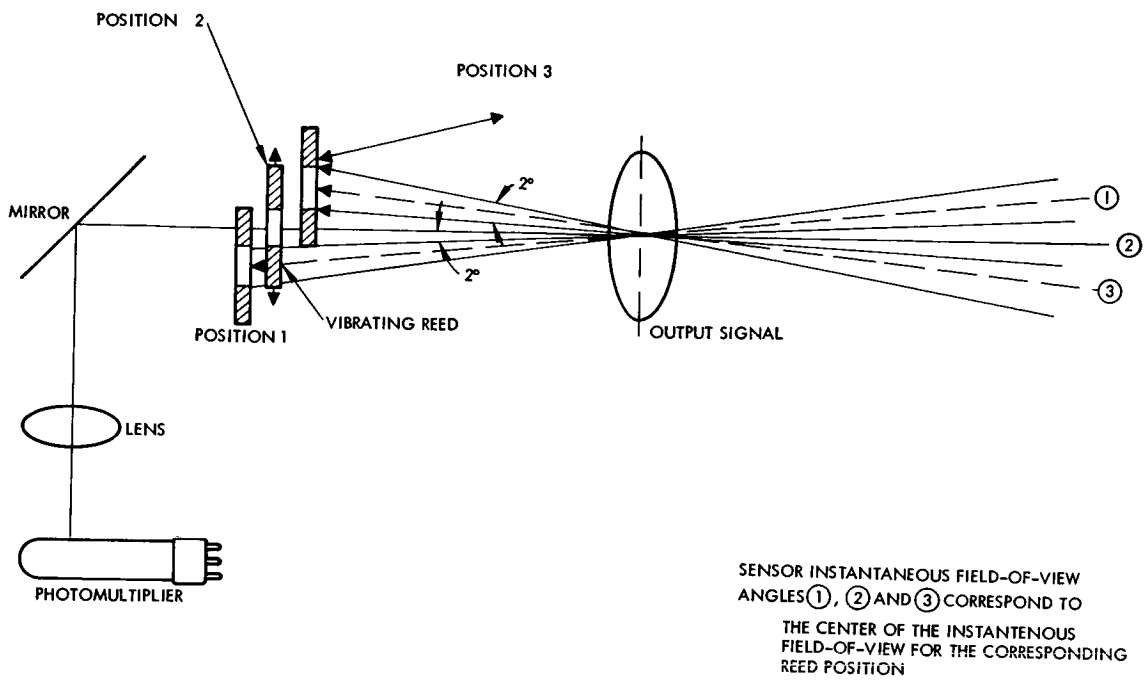


Figure B. Vibrating Reed Star Sensor

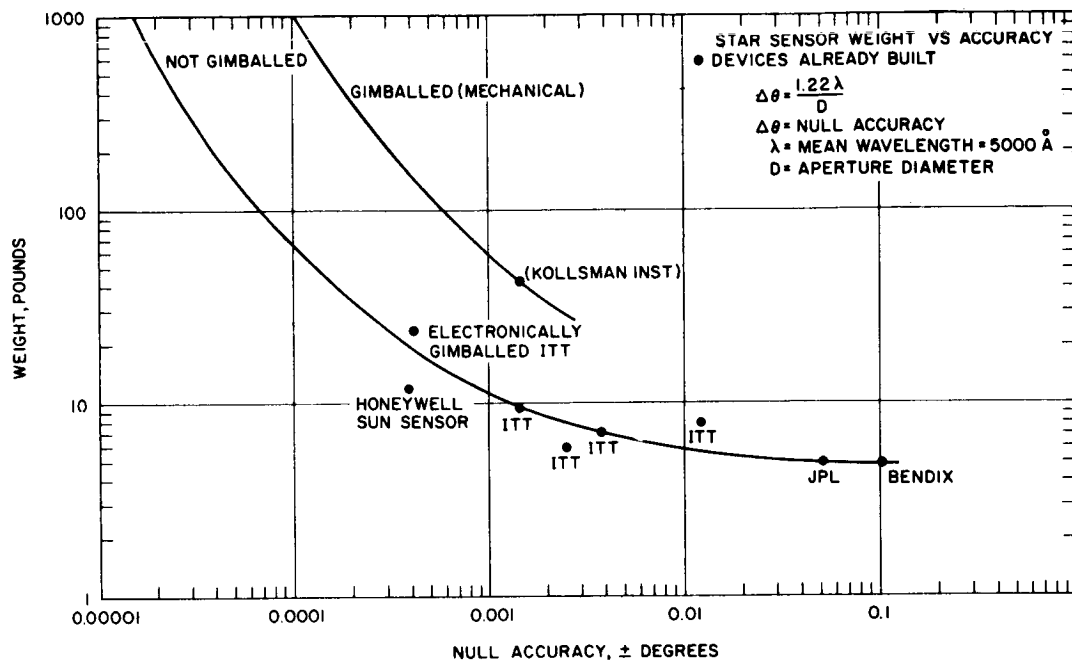


Figure C. Star Sensor Weight versus Accuracy

STAR TRACKER DETECTORS

Examples of star tracker implementation using photomultiplier, image dissector, and quadrant photomultiplier are given

Photomultiplier Tube. The conventional photomultiplier tube detector has been widely used in star sensors. Since only the magnitude of the energy incident on the photocathode is sensed, scanning is required to give position information. To avoid dc drift, modulation is provided either by a chopper or by the mechanical scanning.

A one-axis Canopus sensor of this type is used on Surveyor. Error signals are generated by comparing the modulation envelope of the sensor output with a reference signal square wave. The Surveyor sensor was designed for a null accuracy of approximately 0.1 degree with a 4 x 5 degree instantaneous field of view (FOV).

On the orbiting astronomical observators (OAO), six photomultiplier trackers are used, each having 1 degree FOV. Only three of the six trackers are required to lock on their guide stars for vehicle acquisition. This system is capable of 30 arc second accuracy for second magnitude or brighter stars. High accuracy (± 0.1 arc second) star sensing with a 2 arc minute FOV is provided by the 80 cm Cassegrain telescope used for astronomical observations.

Image Dissector. The image dissector tube detector (Figure A) allows mechanical scanning and modulating to be performed electronically. It consists of a scanning section and of a photomultiplier. A one axis Canopus sensor of this type having a null accuracy of 0.1 degree was used on Mariner II. It had a total field of view of 4 degrees in roll and 32 degrees in pitch and an instantaneous field of view of 0.86 degree in roll and 10.6 degrees in pitch.

The boresighted star tracker used on the OAO incorporated two image dissector tube and had a null accuracy of 2 arc seconds. It had a total field of view of 3 degrees and an instantaneous field of view of 10 arc minutes.

Quadrant Photomultiplier. The quadrant photomultiplier (Figure B) uses four photocathode segments. The segments are sequentially sampled and the currents are then compared to obtain the attitude signal. The Canopus sensor of this type used on the Advanced Orbiting Sun Observatory, AOSO, had a null accuracy of ± 0.5 arc minute.

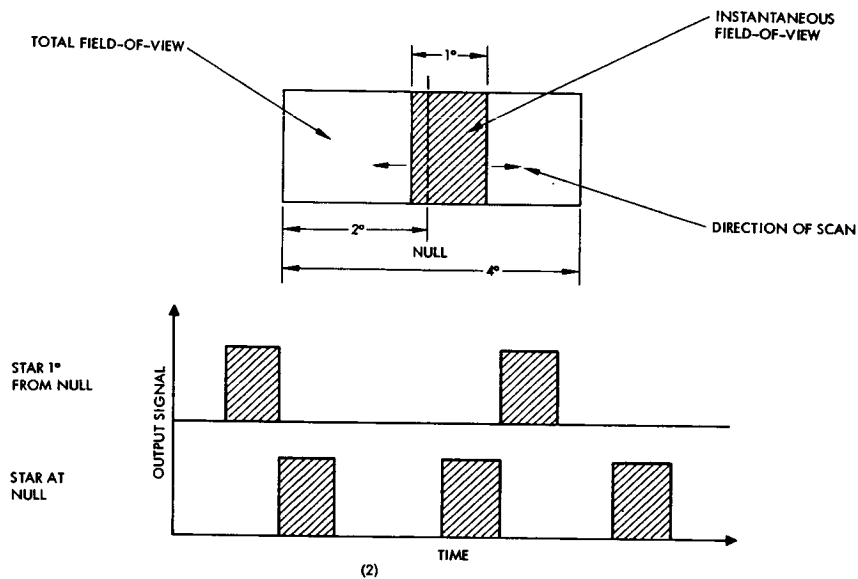
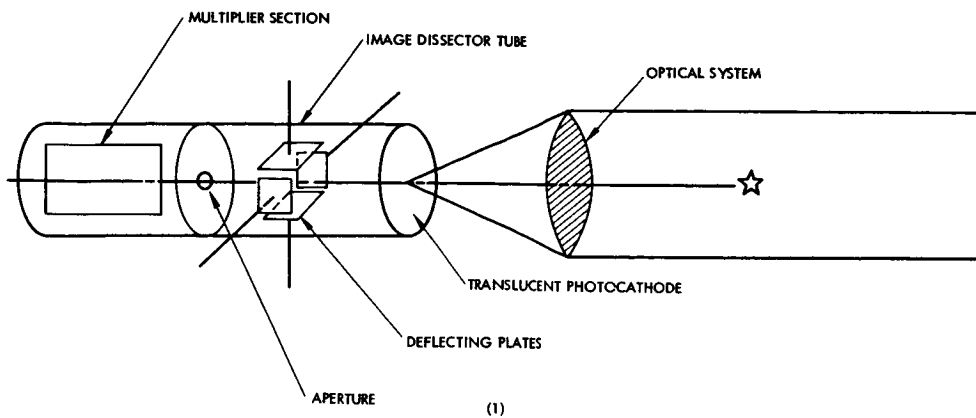


Figure A. Image Dissector Star Sensor

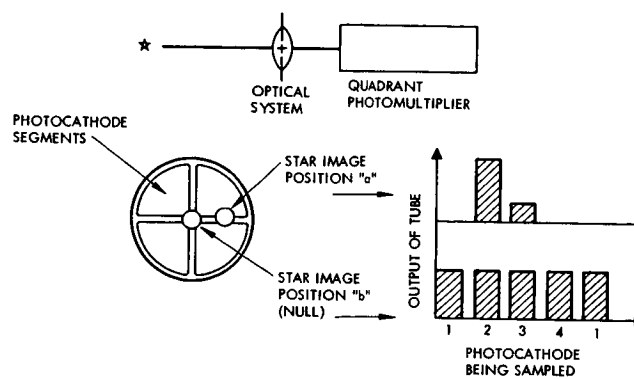


Figure B. Schematic Diagram of a Quadrant Photomultiplier Type of Star Sensor

Component Performance and Burden Relationships Attitude and Tracking Sensors

PLANET SENSORS

The principle of operation for Planet Sensors is described and several implementations are schematically illustrated.

Planet sensors are applicable to bodies subtending angles greater than 0.25 degree. As a result of the larger angular size of the target object, target irregularities are the dominant error source. Planet sensors may be conveniently grouped according to their spectral response as visual or infrared.

Visual Sensors. The Ranger vehicles used shadow masked photomultiplier tubes to sense the earth at ranges of 90,000 miles. These sensors (Figure A) were designed to have a linear range of 2.5 degrees and a null accuracy of ± 0.1 degree. Similar performance was reported for the Mars sensors used on Mariner II.

Infrared Sensors. Horizon scanners (Figure B) utilize a scanning sensor with a small field of view. Pulses are generated when a body enters or leaves the field of view. The Mercury capsule used two such sensors, each having an instantaneous field of view of 2×3 degrees and providing an accuracy of $\pm 1/2$ degree. Nimbus used a similar system to provide accuracy of $\pm 1/2$ to $\pm 1-1/2$ degrees. Horizon scanners of various types are compared in the table.

Edge trackers (Figures A and C) lock onto the space-target boundary and oscillate the sensor field of view about this edge. Attitude information thus obtained is based on the average mirror orientation. Three sensors of this type provide 2-axis information for the OGO, with accuracies of better than 1 degree at altitudes from 100 miles to 100,000 miles. Contemporary infrared planet trackers are not capable of accuracies much better than ± 0.5 degree.

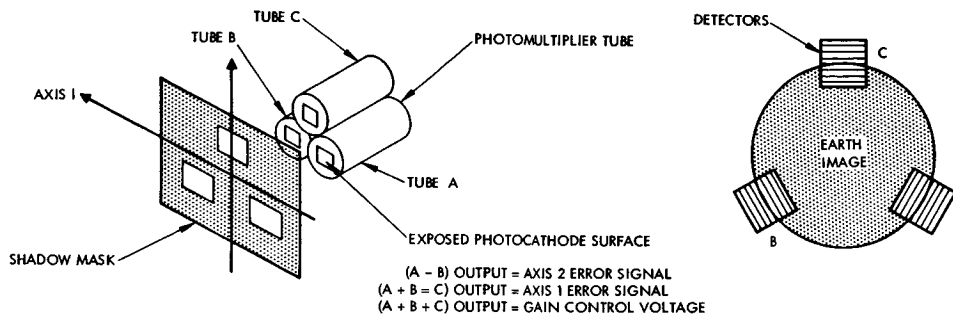


Figure A. Ranger Earth Sensor Schematic

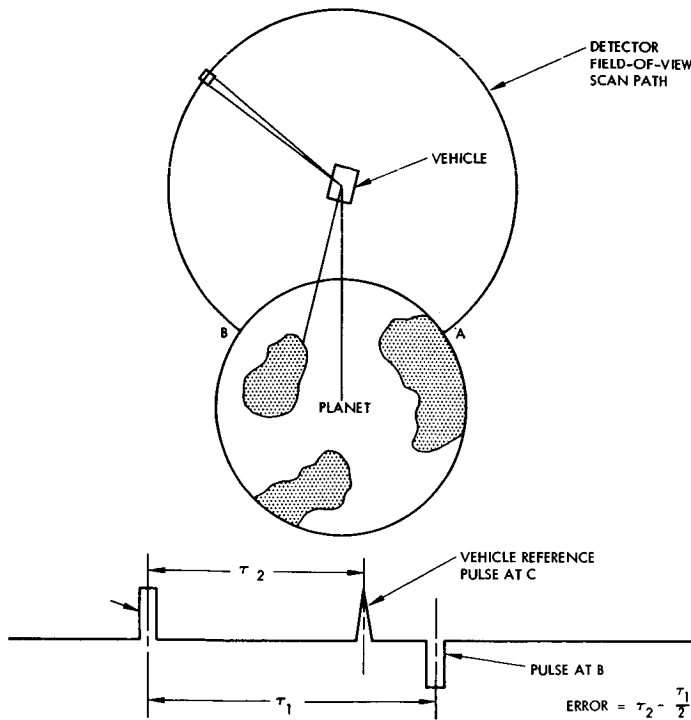


Figure B. Operating Principle of Horizon Scanners

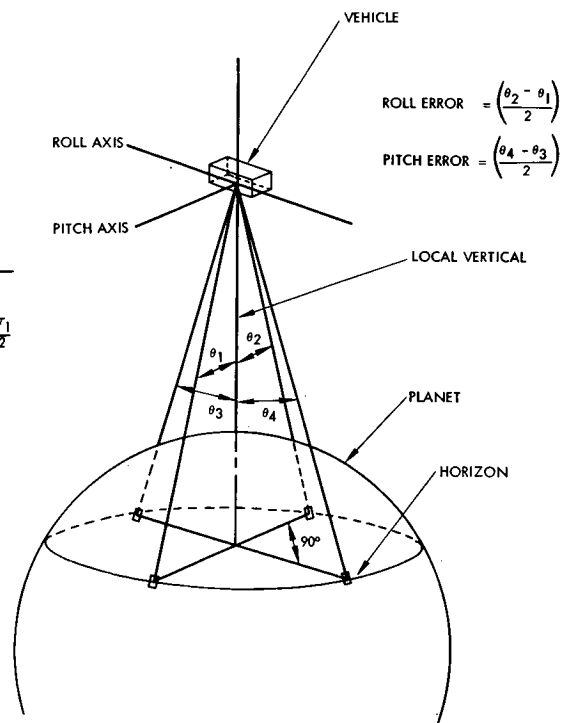


Figure C. Operating Principle of Edge Tracker Devices

Comparison Matrix of Horizon Scanners

Candidate Instrument	Performance (at Synchronous Altitude)	Reliability	Weight, Power and Size
Earth Horizon Scanners 1. Advanced Technology Labs (ATL) Gemini Horizon Scanner	<ol style="list-style-type: none"> Altitude range 50 to 2000 n.mi. Accuracy 0.1 degree Scanning Azimuth, edge tracking Field of view 92 to 102 degrees above vertical - 1.4 x 1.4 degrees instantaneous 8 to 22μ bandpass Two axis capability ±20 degrees allowable at angle Accuracy versus tilt 850 n.mi. - accuracy degrades with tilt angle 	<ol style="list-style-type: none"> MTBF = 25, 000 hours estimated. Currently flying on Gemini and classified projects. 	<ol style="list-style-type: none"> Weight 4.5 pound sensor 4.2 pound electronics 10 watts at 26 volts 4-18 x 5-3/4 x 5-1/8 inches Power Size
2. ATL OGO Horizon Scanner	<ol style="list-style-type: none"> Altitude range 100 n.mi. to 80, 000 n.mi. Accuracy 0.1 - 0.2 degree Scanning Edge tracking - two heads per axis Field of view 1.4 x 1.4 degrees instantaneous 90 degrees above vertical acquisition 8.5 to 20 microns bandpass Two axis capability Accuracy versus tilt Error increases above 5.0 degrees tilt angle maximum tilt is 6.0 degrees 	<ol style="list-style-type: none"> 0.874 reliability for 1 year In production, flight experience on OGO program. 	<ol style="list-style-type: none"> Weight 13.2 pounds Power 9.8 watts Size 7 x 6 x 2.5 inches each dual head 7 x 5 x 3 inches electronics
3. ATL Advanced OGO Horizon Scanner	<p>Performance identical to (2) above except that:</p> <ol style="list-style-type: none"> Accuracy <0.05 degree Two axis capability 14 to 16 microns bandpass 	<ol style="list-style-type: none"> 0.9934 reliability for 1 year. Scheduled to fly on later OGO's - in qualified testing. 	<ol style="list-style-type: none"> Weight 10.0 pounds Power 6.0 watts Size 7 x 6 x 2.5 inches each dual head 7 x 5 x 2 inches electronics
4. Barnes All Altitude Horizon Sensing System - Model 13-160	<ol style="list-style-type: none"> Altitude range 100 to 22, 000 n.mi. Accuracy 0.1 degree instrument <0.1 - 0.2 degree system error Scanning Three edging tracking heads, one per sensor Field of view 70 degree acquisition scan, 0.5 x 3.0 degrees instantaneous 1.8 to 20 microns bandpass Two axis capability Accuracy versus tilt angle Similar to 2.6 above 	<ol style="list-style-type: none"> MTBF not available. 	<ol style="list-style-type: none"> Weight 4 pounds sensor head each 7 pounds Power 4.0 watts average 12.0 watts maximum 5 x 6.5 x 5 inches each Size 13 x 8.5 x 4 inches electronics

COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS

Passive Attitude Control Techniques

	Page
Solar Radiation Pressure	444
Gravity Gradient Forces	446
Magnetic Forces	448

Active Attitude Control Devices

Reaction Wheels	452
Momentum Spheres and Fluid Flywheels	454
Momentum Wheels	456
Reaction Jets	458

INTRODUCTION

Types of control techniques are introduced and accuracy ranges are given.

Attitude control techniques are classified as active or passive, depending on whether they consume energy or not. Passive control techniques produce restoring torques by use of such natural forces as:

Solar radiation pressure

Gravity gradients forces

Magnetic forces

Aerodynamic forces

Active attitude control devices use:

Spin stabilization

Reaction wheels

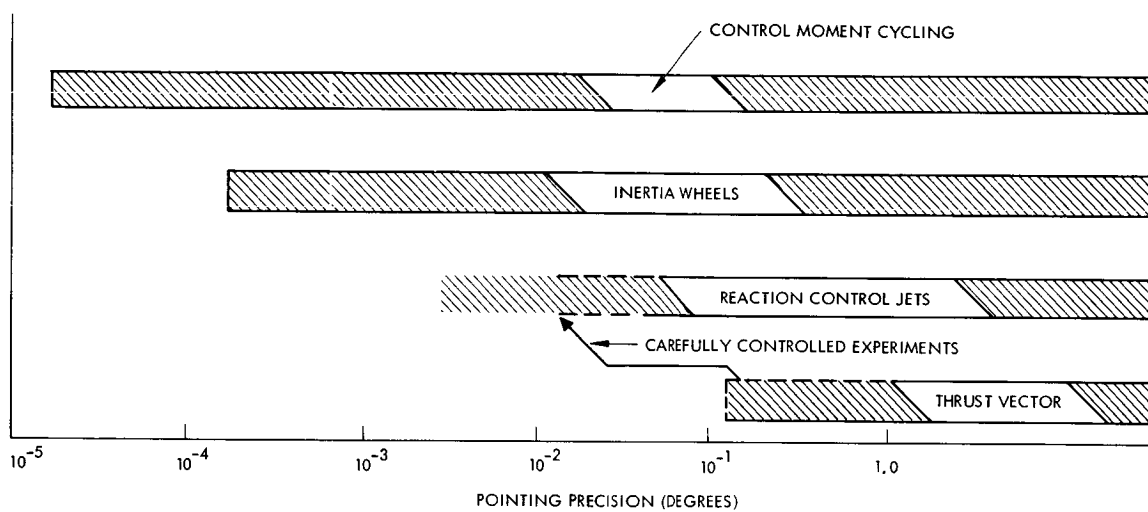
Momentum wheels

Reaction jets

Passive attitude control techniques other than solar pressure are limited in use to regions near planets. Restoring torques produced by natural forces are relatively quite small compared to those of active devices so that the equilibrium time, to correct significant vehicle displacements, may be too long for a primary attitude control system and an optical communication system. Attitude control techniques are summarized in the table. Accuracy limitations of the above attitude control techniques and others are summarized in the figure¹.

Spin Stabilization has an accuracy range to 0.2 degree and passive control techniques produce angular accuracies of 2 to 4 degrees.

¹"High Accuracy Attitude Control for Space Astronomy," D.C. Fosth and W.H. Zimmerman, JACC 1967, pp. 753-761.



Accuracy Comparison of Space Type Actuators

Component Performance and Burden Relationships Attitude Control Techniques

INTRODUCTION

Comparison Matrix of Attitude Control Techniques

Candidate Systems	Performance Considerations	Interface Considerations	Reliability Considerations	Weight Considerations
I. On-Off reaction jet systems	<ol style="list-style-type: none"> Limit cycle mode of operation causes average attitude errors to be one-half the deadband plus sensor errors. System fuel consumption susceptible to sensor noise and disturbance torque magnitude. System provides maneuver capability and acquisition capability for small fuel weight penalty. Duty cycle inversely proportional to attitude deadband (for small disturbance torques), hence fuel consumption in steady state possibly large. 	<ol style="list-style-type: none"> Sensor and system physical alignment errors add to system attitude errors. System requires either low thrust levels and small thrust misalignments for orbital Δv maneuver jets (or auxiliary control mode during Δv maneuvers). Force levels may be selected to perform some transfer orbit control; however, if supplemental apogee propulsion is required, spin stabilization will be required during apogee thrusting. 	<ol style="list-style-type: none"> Relatively simple, reliable system. Extensive space flight experience exists. Cold gas system requires no ignition-only solenoid valves and pressure regulation are active elements. Liquid propellant requires ignition, and zero-g feed system. 	<ol style="list-style-type: none"> Fuel weight may be excessive, especially if cold gas selected as propellant.
II. Magnetic control systems	<ol style="list-style-type: none"> Missing control axis will result in poor pointing accuracy at certain times in orbit. Magnetic control introduces cross-coupling between axes in most cases. Weak earth field requires large coils with reasonable control torque magnitudes. Somewhat complex signal processing is required on board spacecraft. Lifetime is not limited by fuel supply provided power remains available throughout mission. 	<ol style="list-style-type: none"> System requires extremely low thrust from orbital Δv maneuver thrusters due to rather weak control forces. Acquisition performance is relatively poor with primary magnetic control. Sensing of earth's magnetic field is required in addition to other 3 axis attitude sensing. Large power supply required to produce sufficient control torque. Mounting of coils may present alignment and configurational problems. 	<ol style="list-style-type: none"> No moving parts, but electronic circuitry is more complex than reaction jets. Limited space flight experience exists. Thermal heating of coils may introduce lifetime degradation. 	<ol style="list-style-type: none"> Weight of large coils may be excessive. Weight of additional power supply may be excessive.
III. Reaction wheels with reaction jet momentum removal	<ol style="list-style-type: none"> Cyclic disturbance torques and maneuvers can be performed for no additional fuel penalty. Precise accuracy (limited to sensor accuracy) and attitude stability can be achieved. Tracking accuracy can be better than for other candidates by proper design. Reaction jet system provides acquisition capability and back-up attitude control capability (until fuel is depleted). Some crosscoupling is introduced via the angular momentum vector of the wheels. 	<ol style="list-style-type: none"> Low Δv maneuver thrust levels and misalignments are required. Sensing for three control axes must be provided, however, system is somewhat less sensitive to sensor noise than reaction jet system. Separate transfer orbit control and control during apogee thrusting (if required) may be required. Any number of maneuvers for experiments may be performed with no weight penalty. 	<ol style="list-style-type: none"> Wheels and motors in each axis constitute some moving parts; electronic circuitry is also more complex than for reaction jet system. High speed bearing in motor-flywheel sets may degrade lifetime. Some space flight experience will or does exist (Nimbus, OAO, OGO). 	<ol style="list-style-type: none"> Weight tradeoff must be performed versus all reaction jet systems for 2-year lifetime.

Comparison Matrix of Attitude Control Techniques (Continued)

Candidate Systems	Performance Considerations	Interface Considerations	Reliability Considerations	Weight Considerations
IV. Reaction wheels with magnetic momentum dumping.	<ol style="list-style-type: none"> Same as III-1, 2, 3, 5 above. Back up attitude control capability does not exist essentially due to II-1, 2, 3, 4; signal processing is not set up for primary attitude pointing. Momentum dumping only reduces the crosscoupling and missing axis difficulties somewhat. 	<ol style="list-style-type: none"> Same as III-1, 3, 4 above. Same as II-3, 5 above. 	<ol style="list-style-type: none"> Same as III-1, 2, 3 above for wheel portion. Momentum dumping circuitry more complex than for reaction jet momentum dumping. 	<ol style="list-style-type: none"> Same as II-1, 2 above. Total system weight including wheels larger than for III since momentum unloaded expected to require only small fuel weight.
V. Momentum bias system with reaction jet momentum removal	<ol style="list-style-type: none"> Momentum vector stiffness allows rather gross Δv maneuver thrust levels and misalignments. Reaction jet system provides acquisition capability; wheel and jet may be used from possible transition from spin stabilized transfer orbit mode to non-spinning operational mode. Most cyclic angular momentum due to disturbances can be stored without fuel penalty. Passive dumping is sufficient in roll-yaw axes; active only (electronically) in pitch axis. 	<ol style="list-style-type: none"> Only two axis earth sensing may be sufficient due to inertial stability and fact that roll error (relative to earth) becomes yaw error 90 degrees later in orbit. Large Δv thrust levels and misalignments are tolerable. Experiment maneuvers will be limited in N-S direction. 	<ol style="list-style-type: none"> One wheel instead of three; however, larger wheel may have more bearing lifetime problems. Control circuitry is least complex of all candidates. No space flight experience. 	<ol style="list-style-type: none"> Wheel weight and structural support weight may be large. Fuel weight will be excessive if frequent reorientation maneuvers are required.
VI. Gyro torquers with reaction jet momentum removal	<ol style="list-style-type: none"> Same as III-1, 2, 3, 4. Performance essentially identical with reaction flywheels. 	<ol style="list-style-type: none"> Same as III-1, 2, 3, 4. 	<ol style="list-style-type: none"> Mechanical motion of constant speed gyros replaces speed variations of flywheels. Six spinning masses rather than three (unless two axis gimbal used with one set of 2 gyros). No appreciable space flight experience (scheduled for one classified military project but program cancelled). 	<ol style="list-style-type: none"> Same as III-1.

SOLAR RADIATION PRESSURE

Solar pressure values and usefulness to spacecraft control are described.

Solar Radiation pressure may be used in attitude control at distances greater than 300 miles from the earth. The radiation pressure at a solar distance of 1 AU is approximately 10^{-7} lb/ft² and varies inversely with the square of the solar distance. The solar radiation torque on a vehicle is given by

$$\begin{aligned} L_s &= C_r A_s \left| \vec{b} \times \vec{P}_s \right| \\ &= C_r P_s A_s b \sin \psi_s \end{aligned}$$

where

L_s = solar-radiation torque

A_s = area of solar sail

b = distance from center of mass to center of pressure of vehicle and solar sail

ψ_s = angle of incidence of solar radiation on sail surface

P_s = solar-radiation pressure

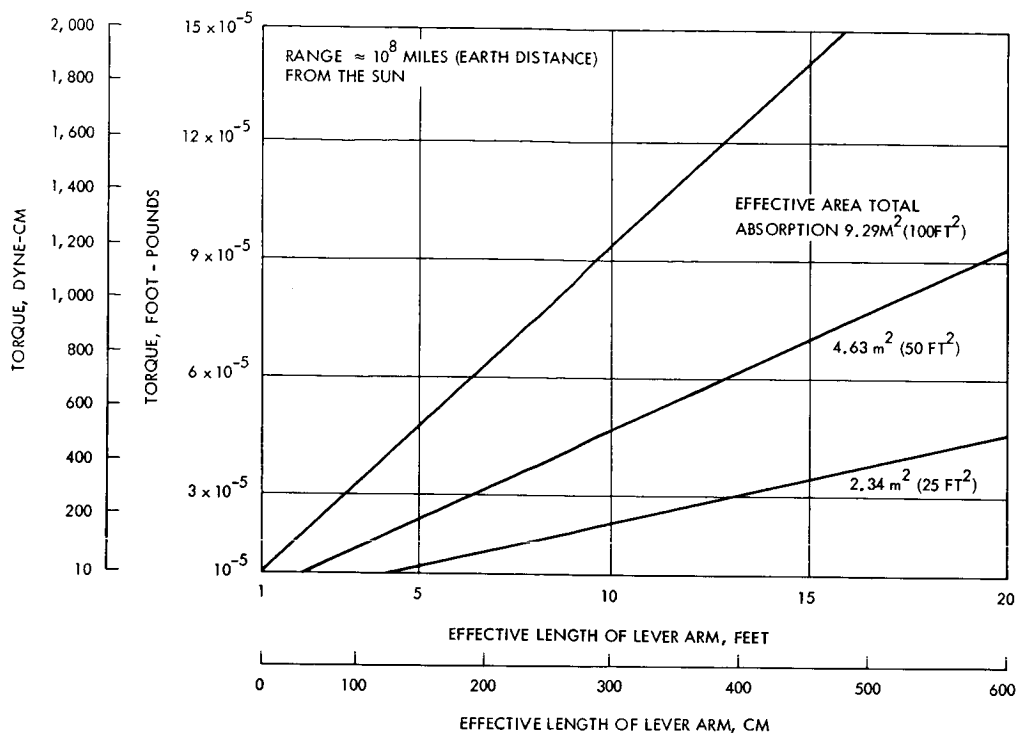
C_r = coefficients of radiation force

According to Thompson,¹ the angle ψ_s , is kept small, typically in the range of 10 degrees. Chin² provides a table of coefficients, C , which are a function of both vehicle geometry and surface reflectivity (see the table.)

The figure shows solar radiation torque as a function of effective length of lever arm with effective area as a parameter. Studies have indicated that for a typical 1000-pound vehicle at 1 AU having a 50 ft² stabilizer, a 5 degree disturbance would be reduced to less than 1 degree error in about 12 minutes. Attitude control by radiation pressure has the advantage of extreme reliability. The principle disadvantage is that restoring torques are so weak that equilibrium times are long and accuracy is limited to 1 degree. Another disadvantage is that large deployed areas of sail are required.

¹W. T. Thompson, "Passive Attitude Control of Satellite Vehicles," Guidance and Control of Aerospace Vehicles, edited by C. T. Leondes, McGraw-Hill, 1964.

²T. H. Chin, "Spacecraft Stabilization and Altitude Control," Space/Aeronautics, June 1963.



Solar Radiation Torque as a Function of Lever Arm Effective Length

(Note: Pressure varies as the inverse square of the distance to the sun.)

Coefficient of Solar Radiation, C_r

Vehicle Geometry	Coefficient
Plane surface	$0 < C_r < 2$
Sphere	$0.75 \leq C_r \leq 1.25$
Cylinder	$0.862 \leq C_r \leq 1.471$
Cone or paraboloid	$0 \leq C_r < 2$

One step beyond passive stabilization is the use of movable solar vanes to trim the vehicle in response to sensed attitude information. This system was used on Mariner C. The same disadvantages remain, however, as for the case of the completely passive system.

GRAVITY GRADIENT FORCES

Torque and natural period equations are given for gravity gradient attitude control.

Unless the gravitational force acts along a line passing through the center of mass, a torque tending to rotate the satellite will result. The conventional gravity gradient configuration is shown in the figure. The gravity gradient torque, L_g , is:

$$L_g = \frac{3}{2} \omega_o^2 (I_x - I_z) \sin 2\theta_\ell \quad (\text{dyne-cm})$$

where

I_x = moment of inertia about the x axis

I_z = moment of inertia about the z axis

$$\omega_o = \text{orbit angular rate} = \sqrt{\frac{GM_e}{r^3}}$$

G = gravitational constant

M_e = mass of earth

r = distance from center of earth

θ_ℓ = angle of libration, angle between the symmetric axis of the satellite and the local vertical

The natural period of oscillation (libration period) of the gravity-stabilized satellite is given by

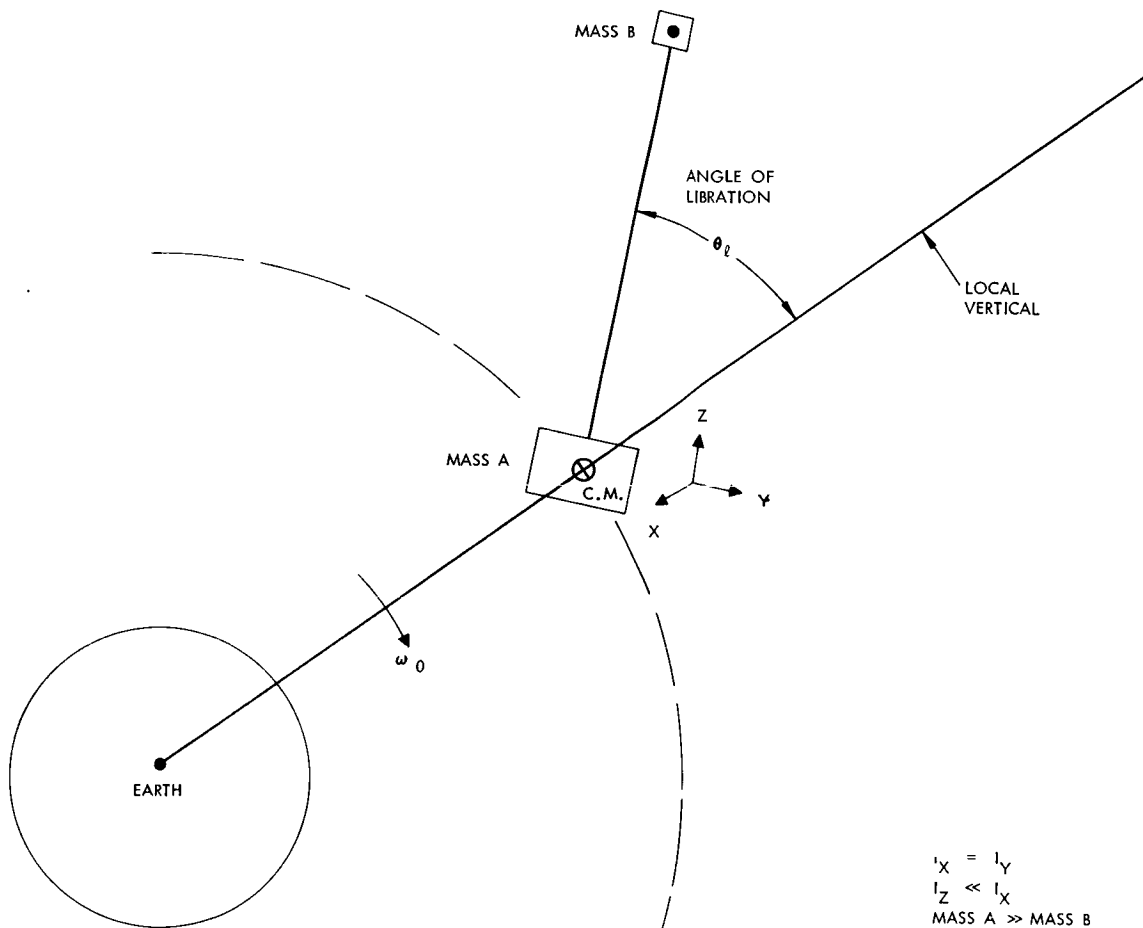
$$T_{||} = \frac{2\pi}{\sqrt{\frac{3GM_e}{r^3} (1 - I_z/I_x)}} \quad \text{seconds}$$

in the plane of the orbit and by

$$T_{\perp} = \frac{\pi}{\sqrt{\frac{GM_e}{r^3} (1 - I_z/I_x)}} \quad \text{seconds}$$

normal to the plane of the orbit.

Typical values for an orbit period of 100 minutes are $T_{||} = 57.8$ minutes and $T_{\perp} = 50.0$ minutes.



Gravity Gradient Schematic

The principle advantage of gravity gradient stabilization is its extreme simplicity. Its disadvantages are that restoring torques are small so that even with adequate damping, accuracy is limited to the order of ± 1 degree.¹

¹E. I. Ergin, "Current Status of Progress in Attitude Control, AIAA Guidance and Control Conference, Cambridge, Massachusetts, August 12-14, 1963.

MAGNETIC FORCES

Magnetic forces are described and altitude variations documented.

Torques are produced by the interaction of the earth's magnetic field with magnetic elements on the vehicle. The earth's magnetic field is that of a magnetic dipole with its axis precessing about the earth's spin axis. The total earth magnetic field intensity as a function of altitude is plotted in the figure.¹ The axial and normal components of the dipole field are given by

$$H_{\text{axial}} = 0.38 \frac{(1 - 3 \cos^2 \delta_m)}{(r/r_e)^3}$$

$$H_{\text{normal}} = 0.461 \frac{\sin^2 \delta_m}{(r/r_e)^3}$$

H_{axial} = the axial component of field intensity (oersteds)

H_{normal} = the normal component of field intensity (oersteds)

δ_m = the angle between the earth's magnetic dipole axis and the radius vector to the satellite

r = the distance to the satellite from the center of the earth (centimeters)

r_e = the radius of the earth (6.371×10^8 cm)

The torque tending to align the magnetic dipole of the on-board magnetic elements with the local magnetic field is given by

$$L_m = MH \sin \phi$$

where

M = magnetic dipole moment of the vehicle

H = local magnetic field intensity

ϕ = angle between the local magnetic field and the dipole moment of the vehicle

¹E. I. Ergin, "Current Status of Progress in Attitude Control, AIAA Guidance and Control Conference, Cambridge, Massachusetts, August 12-14, 1963.

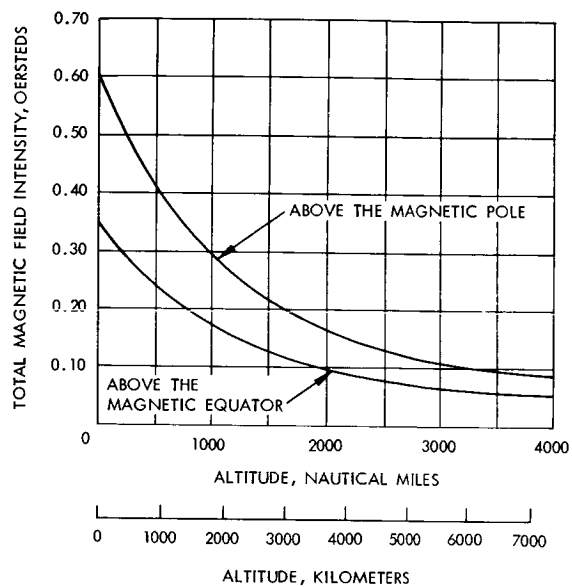


Figure A. Field Intensity Versus
Altitude for Earth's Magnetic
Field

MAGNETIC FORCES

In the case of the TRAAC satellite in a 500 mile orbit, $H = 0.3$ oersteds and $M = 10^3$ resulting in a maximum torque of $L_m = 300$ dyne cm. Since the earth's field precesses with its spin rate, the torque will, in general vary from one orbit to the next as well as being a function of position during an orbit, and may in fact require the expenditure of considerable on-board power. Thus, this is not strictly a passive control method.

Variable torques can be produced by current carrying coils. For a coil centered about the Z body-axis of a satellite the magnetic torque is given by

$$\vec{L}_m = \frac{\pi r_c^2 i}{10} n (B_y \hat{u}_x + B_x \hat{u}_y)$$

\hat{u}_x and \hat{u}_y = unit vectors

r_c = radius of coils (cm)

i = current (amps)

B = flux density (gauss)

n = number of turns

The maximum torque as a function of altitude over the magnetic equator is plotted in Figure B. The Tiros II and III used current loops to precess the satellite spin axis during the orbit. Significant control torques can be achieved by this technique at altitudes up to 10,000 miles with accuracies as great as 0.1 degree. Such accuracies however presuppose corresponding accuracy in knowledge of the magnetic field direction. Lack of such knowledge along with the variations of the magnetic field as the spacecraft traverses its orbit limit the usefulness of magnetic stabilization where high accuracy is required.

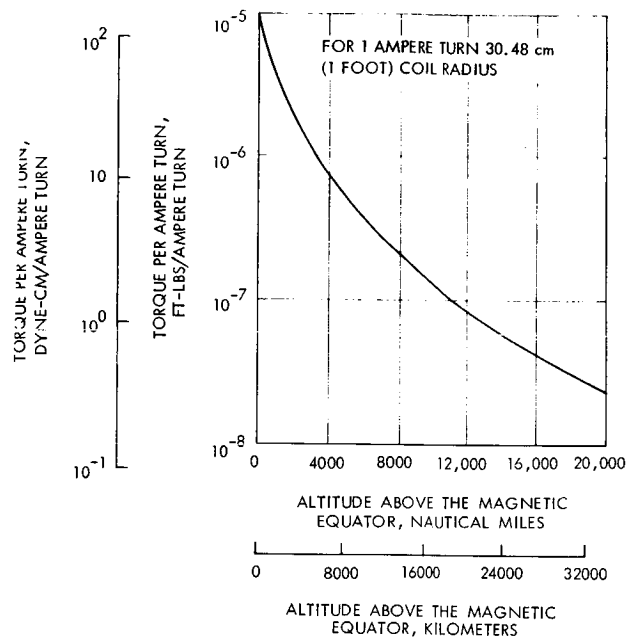


Figure B. Maximum Torque Versus
Altitude (at Magnetic Equator)

REACTION WHEELS

A reaction wheel control system can produce very good control at the cost of system complexity.

Reaction wheels produce a torque on the vehicle in response to an equal but opposite accelerating torque applied to the wheel itself. In doing so, the reaction wheel also acquires and stores angular momentum from the vehicle. Since there is a practical limitation on the moment of angular momentum which may be stored in the wheel, excess momentum must be periodically "dumped" by decelerating the wheel while applying a compensating torque by a coarse attitude control device, typically a reaction jet. Since the reaction wheel provides reaction torque only about its axis of rotation, a separate wheel is generally required for each of the three control axes. This requirement may be circumvented by utilizing gyroscopic crosscoupling between two wheels to provide torques about a third axis. Inadvertent crosscoupling is, in fact, a complicating design consideration for three wheel systems. Reaction wheel size is determined by the angular momentum storage capacity required and driving motor considerations. Figure A depicts wheel weight as a function of angular momentum for wheels of various radii. Figures B and C show wheel weight as a function of maximum speed for wheels of several radii. Figures A through C are based on a constant input torque of 2.71×10^2 dyne-cm (2×10^{-5} ft-lb). Figure D shows momentum capacity versus buildup time for several constant input torques. In a typical system application 705 Kg (1550-pound) communication satellite vehicle was stabilized to within $\pm 0.25^\circ$ /axis with 15.2 cm (6-inch) diameter 2.26 Kg (5-pound) reaction wheels. Momentum dumping with 4000 rpm wheels was required every 48 hours. Fuel weight required for momentum dumping for a 3-year period weighed only 1.6 Kg (3.54 pounds).

Reaction wheels are accurate and reliable. Tracking errors less than ± 0.5 arc second and accuracies of ± 0.1 arc second are predicted. Bendix reports 10,000 hour lifetimes for their present wheels and predicts future lifetimes of several years.

The disadvantages of a reaction wheel control system are its relative complexity with motors and wheels continuously operating, the electrical power requirements, and potentially the system weight (a reaction jet system is also required). Crosscoupling may also be a problem as the wheels possess a net angular momentum vector which introduces additional cross-coupling between the control axes.

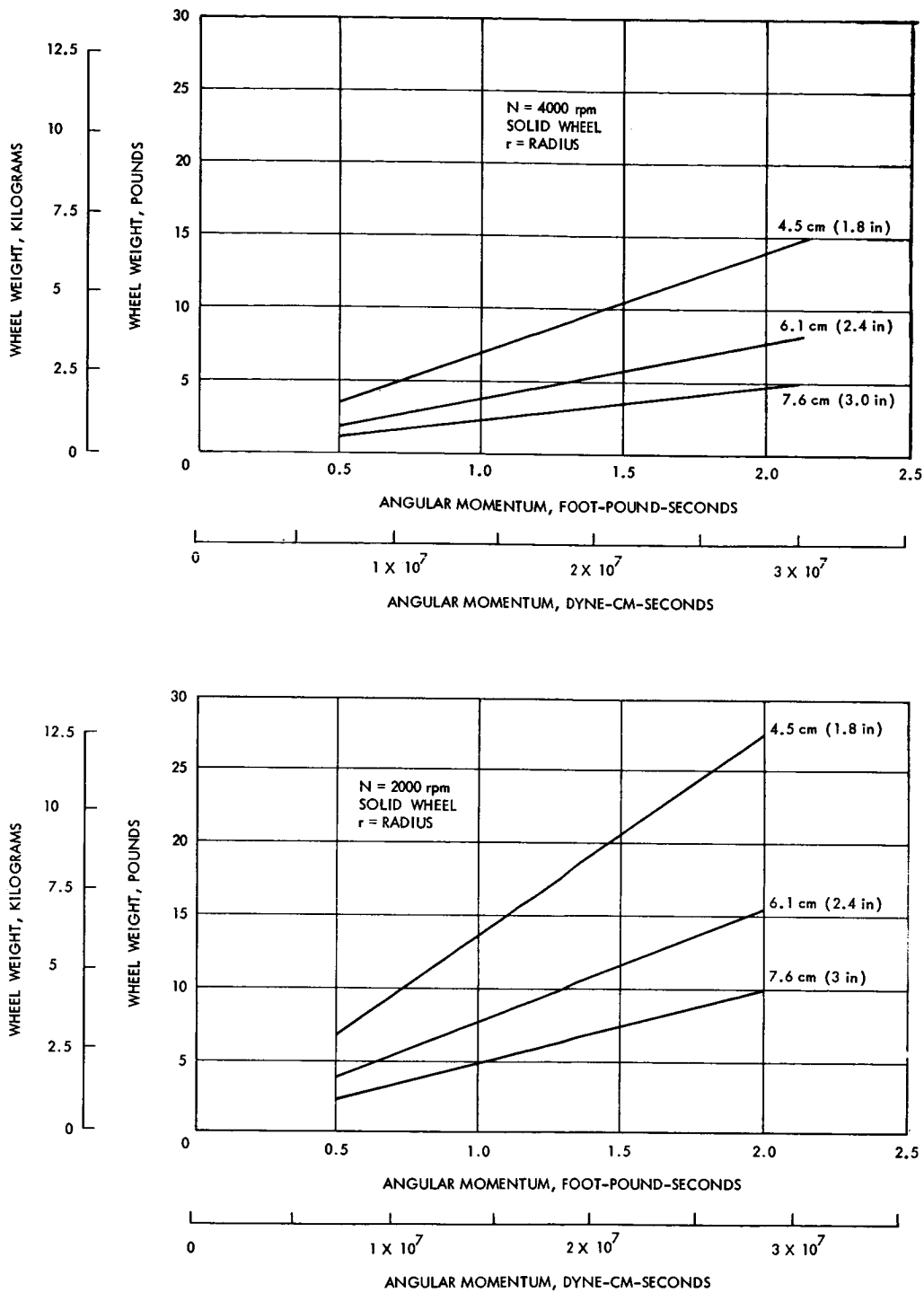


Figure A. Weight Versus Maximum Allowable Momentum Storage (max) in Reaction Wheel

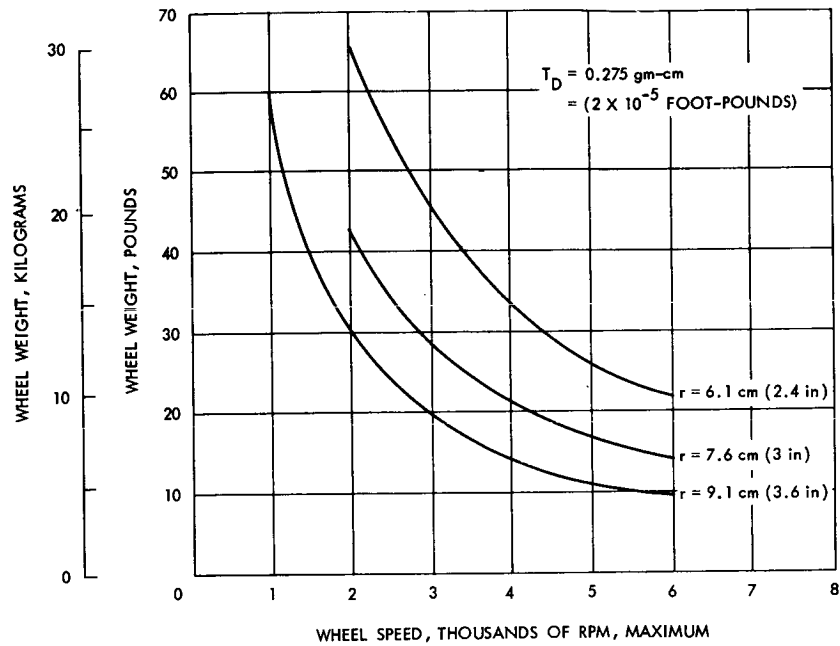


Figure B. Weight as a Function of Speed for 5-Day Reaction Wheel Dumping

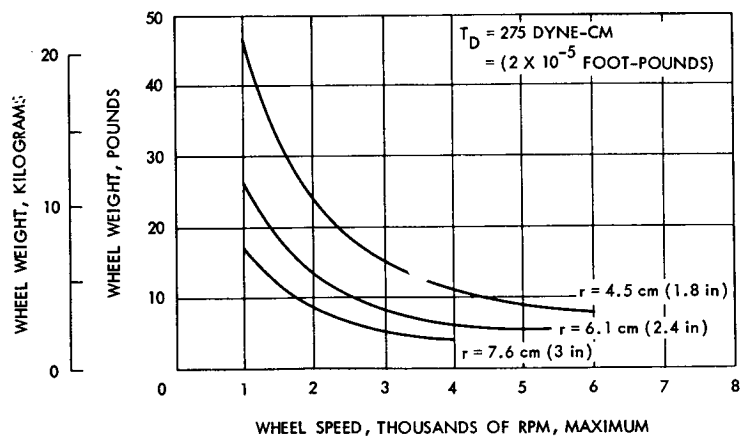


Figure C. Weight as a Function of Speed for 1-Day Reaction Wheel Dumping

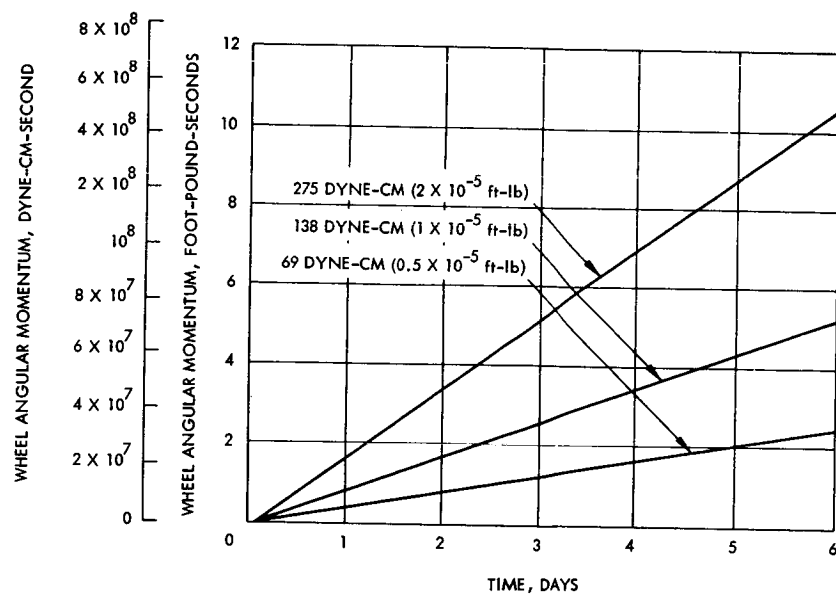


Figure D. Momentum Capacity as a Function of Time for a Constant Torque

MOMENTUM SPHERES AND FLUID FLYWHEELS

Momentum spheres and fluid flywheels are described and actual design are compound with reaction wheels.

Variations on the reaction wheel principle include the reaction sphere and the fluid flywheel. In place of individual flywheels for each vehicle control axis, a momentum sphere provides, in a single device, control torques for all three axes. As suggested by its name, the momentum sphere consists of a spherical rotor suspended by magnetic or electrostatic forces so that it is free to rotate about any axis. Driving motor elements are provided at mutually orthogonal locations and are excited in accordance with the control torques required for their respective axis. The sphere then rotates at a velocity and about an axis which are the result of the past history of vehicle torque requirements. As with other momentum transfer devices, the momentum sphere will saturate and requires resetting. A characteristics advantage of the momentum sphere is the lack of gyroscopic coupling from one axis to the other regardless of the component of momentum about any axis that may be stored in the sphere.

The fluid flywheel has been developed by GE for the Advent communication satellite. It is functionally equivalent to a conventional reaction wheel, however, the reaction torque is provided by circulating a liquid through a closed circular tube. In certain applications, this can be quite advantageous. The fluid within the tube is essentially the rim of a flywheel, yet the tube itself is a static element. As a matter of fact, it is not necessary that the tube follow a circular path; it can follow the overall outline of the vehicle. Therefore, the inertia of the fluid, the required momentum is reached at a low fluid velocity, which minimizes flow losses. Ideally, a low density fluid with low viscosity is desirable. The low density increases the inside diameter of the tube for a given mass per foot, which further reduces the flow losses for a given velocity.

With conducting fluids, the circulating pump can be of the Faraday type in which a driving dc current flows across a diameter of a section of the tube which is in the gap of a strong magnet. (ac pumps are also possible.) dc pumps are characterized by a high pumping head at low velocities or in equivalent terms a high stall torque. Since it has no wearable moving parts, it is, in principle, capable of very high reliability. A typical reaction wheel, reaction sphere, and fluid flywheel are compared in the table.

Device	Size and Burden	Reliability	Development Status	Advantages	Disadvantages
Wheels (3)	15.5 cm (6-inch) diameter 13.6 kg (30-pound) weight 15 watts	Improved materials the reliability fairly good. Still subject to mechanical failure (mainly bearing).	Used on OGO and OAO. Developed for several immediate applications.	Highly accurate, proven.	Crosscoupling effects.
Spheres	15.5 cm (6-inch) diameter 9.06 kg (20-pound) weight 10 watts	Questionable reliability. Major problem is the sphere suspension system.	None developed for space use. A few experimental models now in existence.	No cross-coupling. Lighter weight.	Subject to suspension system failure. Development status.
Fluid	15.5 cm (6-inch) diameter 16.3 kg (36-pound) weight 7 watts	Highly reliable (no moving parts).	Developed by GE for Advent Communication Satellite.	Low power consumption. High reliability. Simple control system. Fast response.	Crosscoupling effects. Development status.

Comparison of Momentum Storage Devices

MOMENTUM WHEELS

Momentum wheels are similar to reaction wheels except they are larger. Typical values of weight and momentum are given.

Momentum wheels provide restoring torques about the axis of rotation in a manner identical to the reaction wheel. However, they differ from the latter in having much greater momentum storage capacity so that they also provide gyroscopic stability about the remaining two axes. Orientation of the vehicle about the spin axes is controlled by applying torque to the momentum wheel just as if it were a reaction wheel. Although the other two axes are gyroscopically stabilized, another torquing device such as a reaction jet or reaction wheel must be used to process them to the proper attitude as well as to correct for drift and disturbing torques.

The advantage of the momentum wheel system is the much greater attitude stiffness compared to the reaction wheel system. Disadvantages are the greater weight and volume of the wheel and its required supporting structure and the potentially high power requirements for its driving motor. Some typical momentum wheel control system weights are shown in Figure A and B.

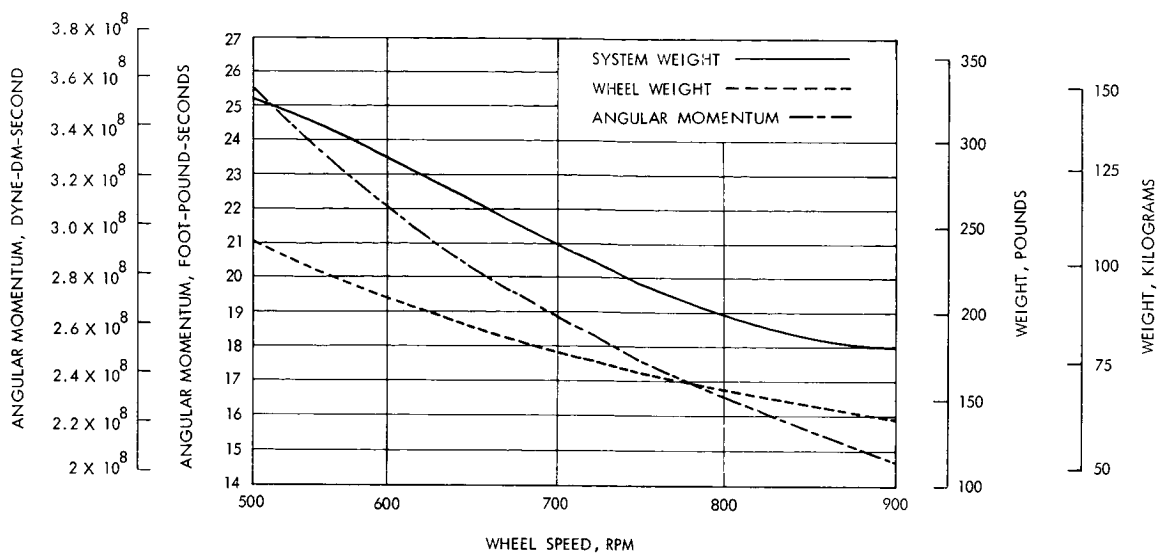


Figure A. Momentum Wheel System Weight and Angular Momentum Versus Wheel Speed for Constant Wheel Diameter, 152 cm, (5 feet)

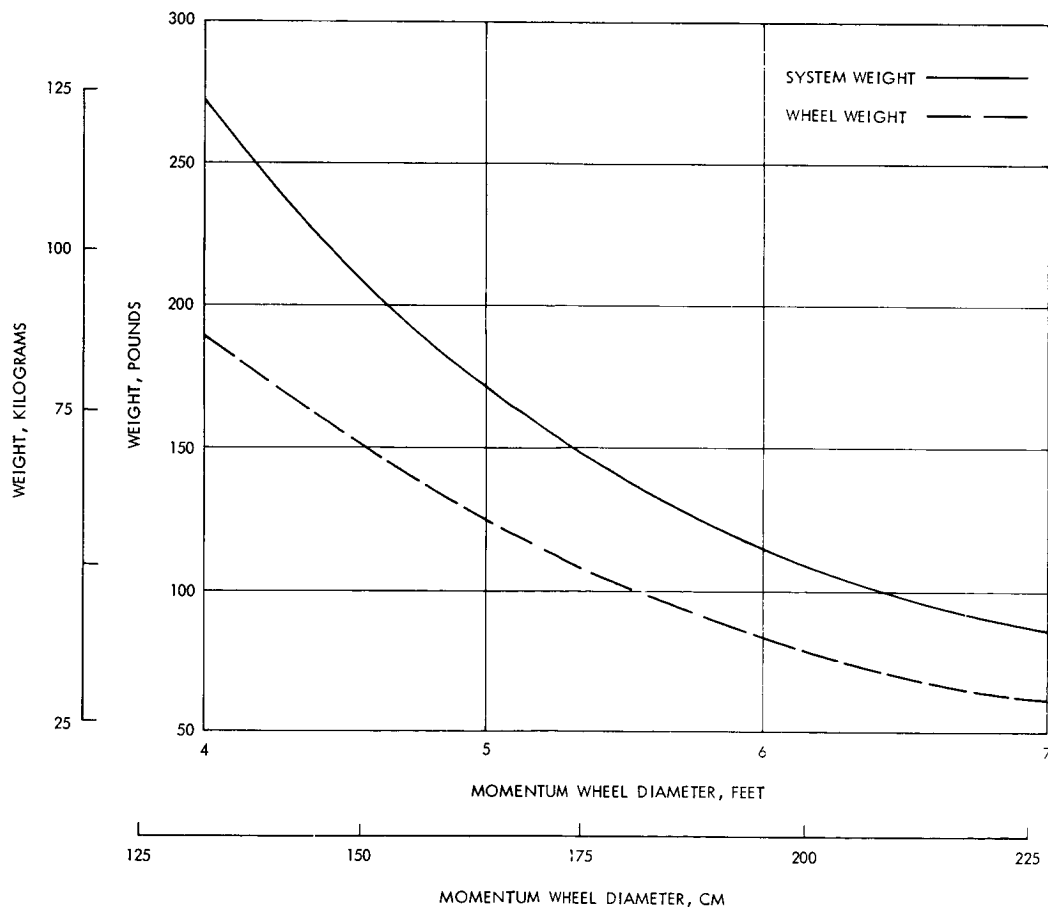


Figure B. Momentum Wheel System Weight Versus Wheel Diameter for Constant Wheel Speed (500 rpm) and Angular Momentum, 1.82×10^5 gram-cm-sec, (13.24 Foot-Pound-Seconds)

REACTION JETS

Reaction jets are widely used in attitude control. Typical system weights are given for cold gas systems, monopropellant systems and bipropellant systems.

Reaction jet systems, using gas or liquid propellant, constitute a simple, reliable, and space proven means of attitude stabilization. Reaction jet systems are customarily operated in an on-off mode characterized by limit cycle operation in which the jets are energized when the error in vehicle orientation exceeds a pre-determined limit.

For stability, velocity information is added to the attitude sensor signal. In the usual case of the two sided or "hard" limit cycle the vehicle enters a "deadband" angle between the limit positions with a velocity sufficient to coast across the deadband and emerge from the other side, whereupon the torquers are energized and the vehicle again crosses the deadband, etc. In a properly designed system the velocity with which the deadband is entered each time is less than that at which it left the deadband so that deadband rates decrease until a minimum value or limit cycle rate is reached. Fuel consumption during limit cycle operation varies inversely as the width of the deadband and directly as the square of the deadband rates, if the effects of the external disturbance torques are low compared to the deadband rates. For a two sided limit cycle the fuel weight required W , is given by

$$W = \frac{I \dot{\theta}_o^2 t}{\ell \theta_d I_{sp}}$$

where

θ_d = the half-deadband angle

$\dot{\theta}_o$ = characteristic limit cycle rate of the system

ℓ = moment arm

t = total mission time

I_{sp} = fuel specific impulse

In this mode, the fuel consumption is directly dependent upon system parameters, and the fuel weight is inversely proportional to θ_d . In general, θ_d must be set smaller than the desired pointing accuracy to allow for potential sensors errors; hence fuel weight can be expected to be high for applications where $\theta_d < 0.1$ degree.

When vehicle rates are very low, external disturbance torques may reverse the vehicle's motion before it crosses the deadband. In this case it emerges from the same side of the deadband as it entered.

This mode of operation is called a one sided or "soft" limit cycle. It is an attractive situation because no fuel is consumed to remove vehicle momentum which was imparted by the previous expenditure of fuel. The integral of the attitude control torques is equal to the integral of the disturbance torques, resulting in the minimum possible fuel consumption, namely

$$W = \frac{T_d t}{I_{sp} \ell}$$

where

t = time

T_d = the average disturbing torque

This one-sided mode of operation is not normally achieved by the use of position information and directly-sensed rate information. Derived rate techniques have been developed, however, which permit very low vehicle angular rates to be achieved. As a result one-sided limit cycle operation is feasible where the disturbance torques are relatively constant and can be accurately predicted.

Reaction jets can be used singly, or in pairs. The latter configuration has the advantage that the moment applied to the vehicle is constant regardless of the location of the nozzles relative to the c. g. This may be of importance where the c. g. shift is large (such as when a lander is separated from a parent vehicle). The use of couples is also advantageous in that it permits the nozzles to be located at any convenient part of the vehicle.

Reaction jet systems are customarily classified according to the fuel used as

Cold Gas

Hot Gas monopropellant

Hot Gas, bipropellant.

Cold gas systems operate by expelling compressed cold gas through expanding nozzles. The common fuels - nitrogen, helium, and the freons - are non-toxic, inexpensive, readily available, and pose no thermal or contamination problems to components or surfaces on which they may impinge. Specific impulse is modest but specific volume is relatively low. A wide range of thrust levels may be attained and thrust

REACTION JETS

response time is short. In contrast to the time and expense of developing and qualifying a new hot gas thrust chamber, cold gas nozzles require only proper sizing and brief testing to assure that they will have the intended performance. Cold gas nozzles were the first actuators used for attitude control of space vehicles and have been widely used since.

Along with these virtues there are some distinct disadvantages to cold gas actuators. Where large total impulse is required, the low specific impulse gives a definite weight penalty. For missions of long duration, especially where accurate orientation is required, the control valve may be operated many thousands of times. This places a severe reliability requirement on the components. Leakage poses a severe threat as a consequence of the limited total impulse carried by the vehicle.

As seen in Figure A,¹ there is a fairly well defined breakeven point where weight considerations favor the use of hot gas over cold gas. Other factors to be considered are the thrust level desired and the minimum impulse bit required. Hot gas engines typically have higher minimum thrust, but this is partially offset by a faster valve response than for cold gas. This is particularly true of bi-propellant systems. Figures B through D¹ show typical performance characteristics for explosive systems.

Mono-propellant hot gas systems rank intermediate between the bi-propellant and the cold gas. They have the advantage of simplicity in that a single valve is required, and the pressurization system does not have to guard against the possibility of accidental mixing of fuel and oxidizer due to leakage or permeation of diaphragms or bladders. Mono-propellants also have more modest specific impulse and a thrust buildup time constant that is both longer and more uncertain than with the bi-propellants. This is particularly true for short pulses.

Other factors to be considered in the use of hot gas systems are the gas temperature and the combustion products, both of which can be detrimental to surfaces they impinge upon, and also to optical devices whose line-of-sight intersects the plume. Radiation cooled thrusters pose an additional thermal problem to the vehicle and equipment near them. As a result of these considerations the tradeoff between hot gas and cold gas systems frequently favors the cold gas even though hot gas can demonstrate a definite weight saving.

The primary advantages of the reaction jet stabilization system are simplicity, flight proven capability, and the ability to provide acquisition, reacquisition, and maneuvers on ground command (limited only by sensor field-of-view capability). The primary disadvantages are the weight penalties associated with fuel and tankage weight for long missions and the accuracy limitations (the limit cycle mode implies an average pointing error approximately equal to one-half the attitude deadband width in addition to sensor contributed errors).

¹Woestemeyer, F. B., "General Considerations in the Selection of Attitude Control Systems," Conf. Proc. SAE/NASA Aerospace Vehicle Flight Control Conference, Los Angeles, California, July 13-15, 1965.

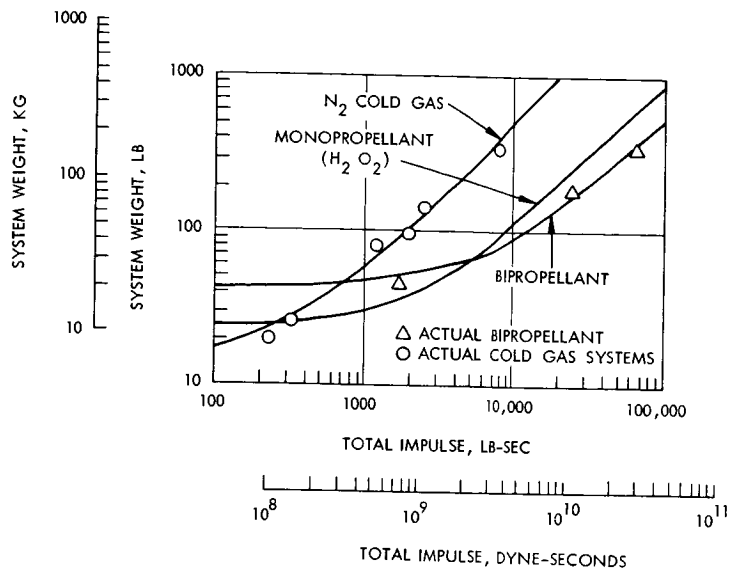


Figure A. Attitude Control System Weight Versus Total Impulse

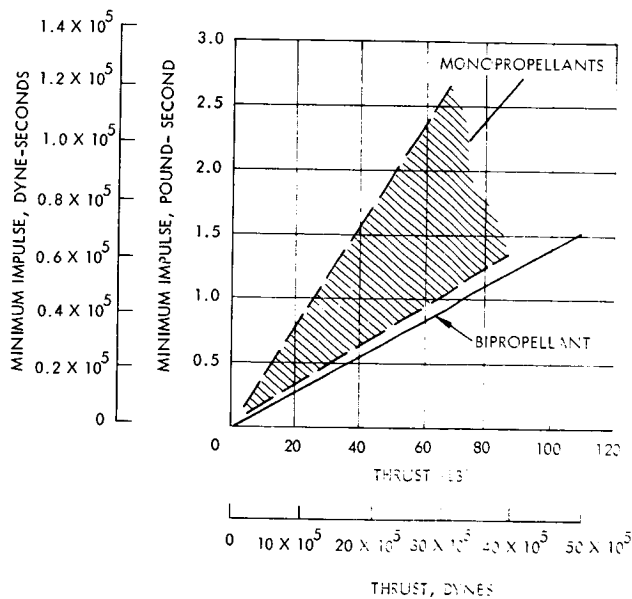


Figure B. Minimum Impulse Bit Versus Thrust

REACTION JETS

A further consideration is that for reaction jet systems energy consumption is a function of the instantaneous control torques required by the vehicle. Momentum removed from the vehicle leaves the system with the expelled gas and cannot be reclaimed. Momentum transfer devices, on the other hand, merely transfer momentum from the vehicle structure to a rotating mass such as a flywheel or a gyroscope located inside the vehicle. As a result, momentum that is transferred to a fly-wheel or gyro in resisting a clockwise torque on the vehicle can then be transferred back to the vehicle in resisting a subsequent counterclockwise torque. This difference is significant in the case of orbiting vehicles which frequently experience cyclic torques as a function of position in orbit. Large components of these will integrate to zero over the orbital period so that the net momentum required over the period is zero.

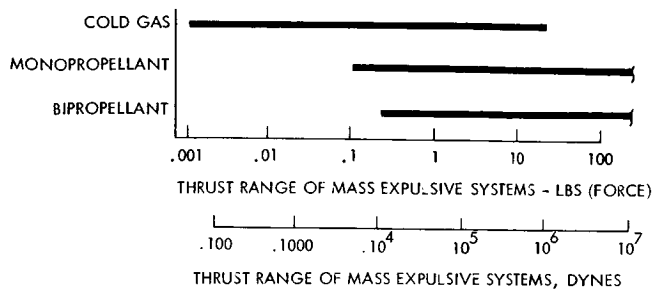


Figure C. Thrust Range of Mass Expulsive Systems

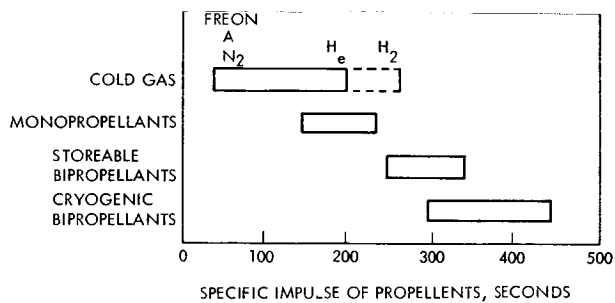


Figure D. Specific Impulse of Propellants

COMPONENT PERFORMANCE AND BURDEN RELATIONSHIPS

Burden Relationships

Weight Burdens	Page 466
Cost Burdens	468
Power Burdens	472

Component Performance and Burden Relationships

Burden Relationships

WEIGHT BURDENS

The acquisition and tracking weight needed for spaceborne apertures are modeled in a form compatible with the methodology described in Volume II of this final report.

A typical transmitter or receiving antenna pointing system consists of a gimbaled support unit, which holds the antenna, and an associated control system, which points the antenna. The weight of the antenna pointing system is dependent upon the weight it must support, the antenna weight, whose weight in turn is dependent upon the antenna size. The antenna pointing system weight is usually not dependent upon the pointing accuracy.

The weight of the acquisition and tracking system may then be modeled in terms of the diameter of the aperture being used.

Such modeling has been done below in a form suitable for the methodology described in Volume II of this final report.

$$W_{QT} = W_{BT} + K_{W_{AT}} K_{d_T} d_T^{n_T} \quad (1)$$

where

W_{QT} = total transmitter acquisition and tracking weight.

W_{BT} = transmitter acquisition and tracking weight independent of aperture size.

$K_{W_{AT}}$ = constant relating transmitter acquisition and tracking weight to transmitter antenna weight.

K_{d_T} = constant relating transmitter antenna weight to transmitter antenna diameter.

n_T = exponent relating transmitter antenna weight to transmitter antenna diameter.

The constants of equation 1 have been evaluated using the material of this section references quoted in this section and engineering judgement. The result of this determination is given in Figures A and B which plot the acquisition and tracking weight required for spaceborne optical and radio apertures respectively.

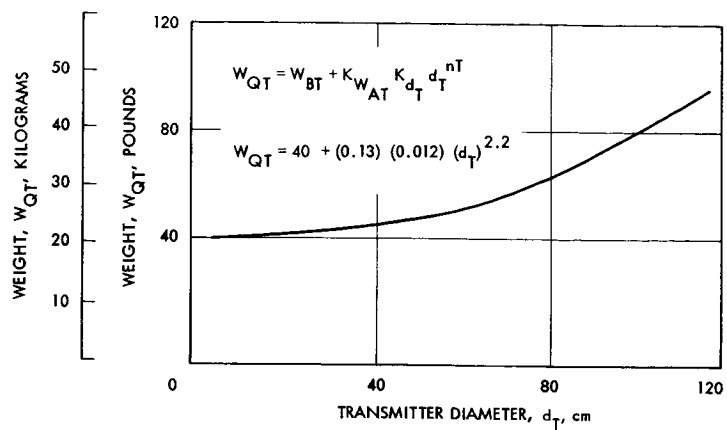


Figure A. Optical Transmission Acquisition and Track Weight

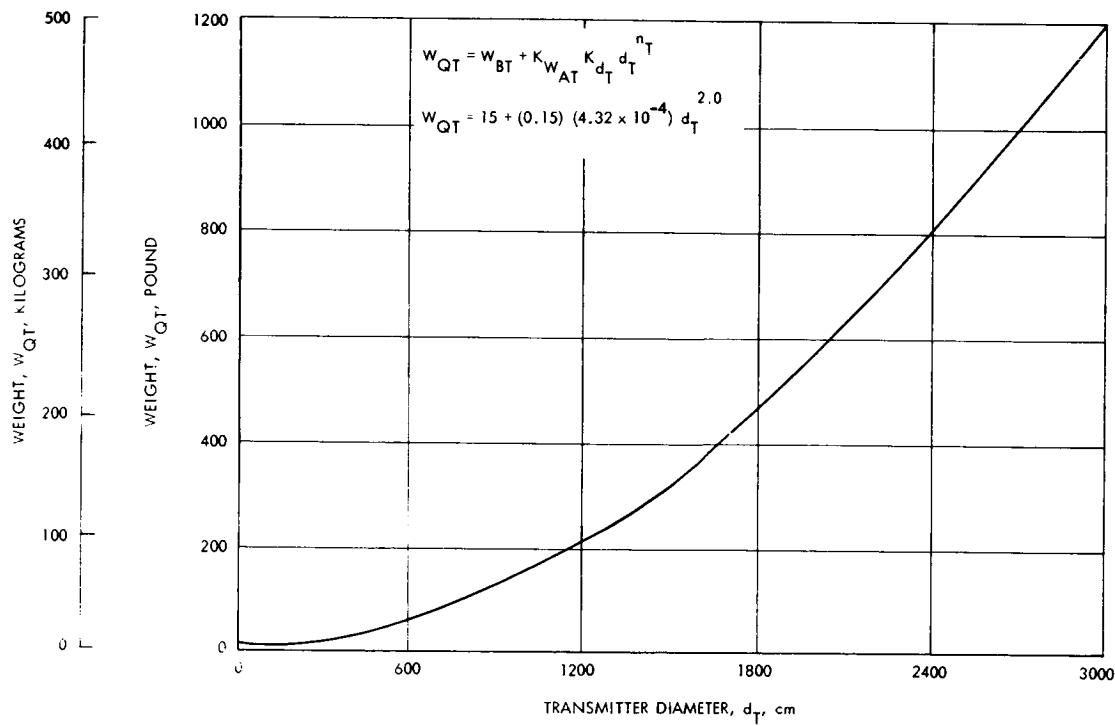


Figure B. Radio Transmitter Acquisition and Track Weight

Component Performance and Burden Relationships Burden Relationships

COST BURDENS

The acquisition and tracking cost for spaceborne and earth apertures is modeled in a form compatible with the methodology described in Volume II of this final report.

The fabrication cost of the transmitter antenna pointing equipment is inversely proportional to the pointing accuracy. Pointing accuracy is generally specified as a fixed percentage of the transmitter beamwidth. Since the transmitter antenna is usually diffraction limited, the fabrication cost of the transmitter antenna pointing equipment is dependent upon the transmitter antenna diameter or gain. A modeling dependent upon these considerations and compatible with the methodology described in Volume II of this Final Report is shown below.

$$C_{QT} = C_{AT} + K_{AT} \left(\frac{\lambda}{d_T} \right)^{-q_T} \quad (1)$$

where

C_{QT} = total transmitter acquisition and tracking fabrication cost.

C_{AT} = transmitter acquisition and tracking fabrication cost independent of transmitting beamwidth.

K_{AT} = constant relating transmitter acquisition and tracking cost to transmitter beamwidth.

λ = transmitted wavelength

d_T = transmitter aperture diameter

q_T = Exponent relating transmitting acquisition and tracking fabrication cost to transmitted beamwidth.

The fabrication cost of the receiver antenna pointing equipment is inversely proportional to the pointing accuracy. For a diffraction limited receiver antenna the fabrication cost of the pointing equipment is dependent upon the receiver antenna diameter or gain, and for a non-diffraction limited receiver antenna the fabrication cost is proportional to the receiver field of view. A modeling dependent upon these considerations and compatible with the methodology described in Volume II of this final report is shown below.

$$C_{QR} = C_{AR} + K_{AR} (\theta_R)^{-q_R} \quad (2)$$

where

C_{QR} = Total receiver acquisition and tracking fabrication cost.

C_{AR} = Receiver acquisition and tracking equipment fabrication cost independent of receiver beamwidth.

K_{AR} = Constant relating receiver acquisition and tracking equipment fabrication cost to receiver beamwidth.

θ_R = Receiving beamwidth, field of view

q_R = Exponent relating receiver acquisition and tracking equipment fabrication cost to receiver beamwidth.

The constants of equations 1 and 2 have been evaluated using the material of this section, references quoted in this section and engineering judgment. The results of this determination is given in Figures A, B, C and D.

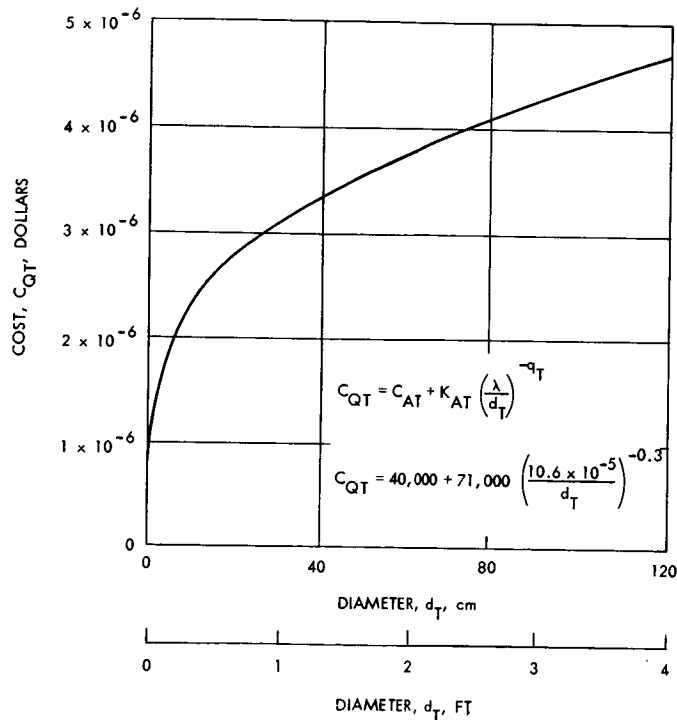


Figure A. Cost of Optical Transmitter Acquisition and Tracking

Component Performance and Burden Relationships Burden Relationships

COST BURDENS

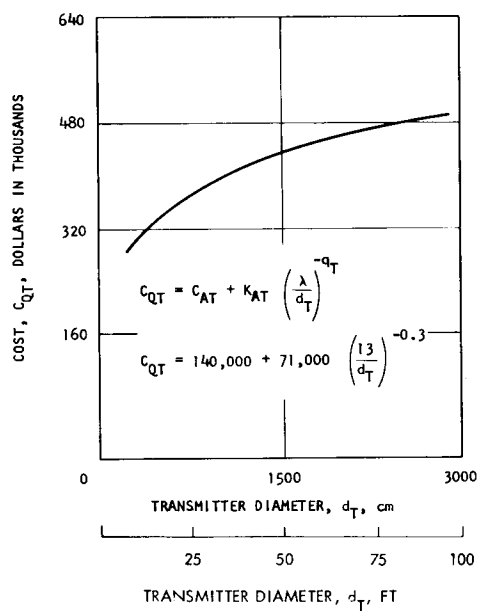


Figure B. Cost of Transmitter Acquisition and Pointing

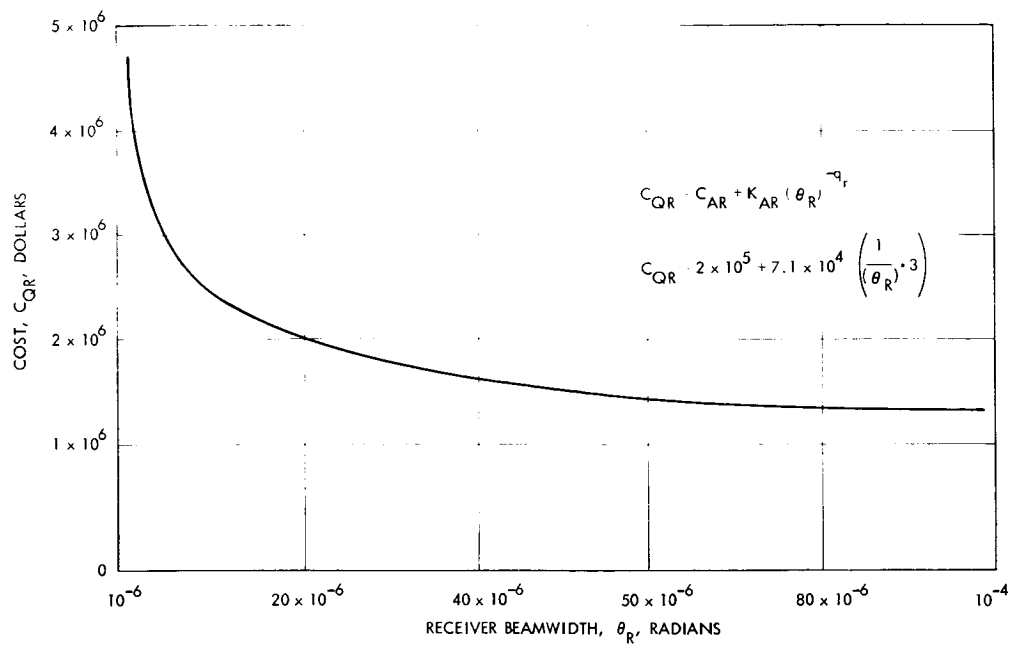


Figure C. Optical Receiver Acquisition and Pointing Cost

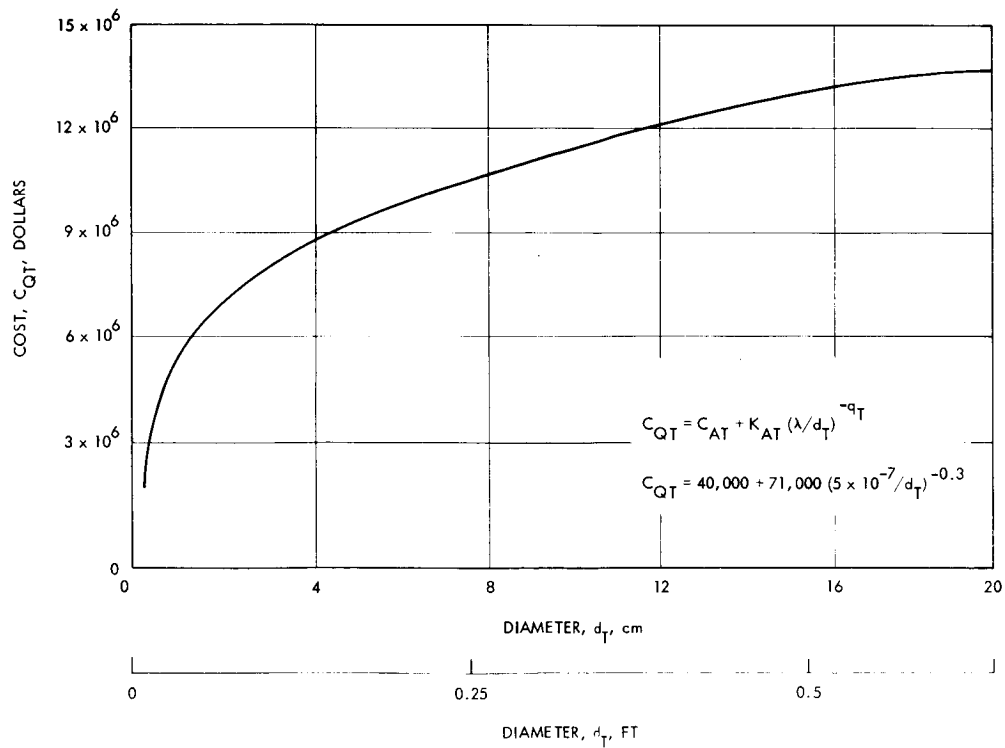


Figure D. Radio Transmitter Acquisition and Pointing Cost

POWER BURDENS

The acquisition and tracking power requirement for spaceborne apertures are modeled in a form compatible with the methodology described in Volume II of this final report.

The electrical power requirement for the transmitter or receiver antenna pointing equipment is primarily dependent upon the weight of the antenna that must be positioned by the gimbal motors. Hence, the power requirement is also proportional to the antenna diameter or gain. A modeling dependent upon this consideration and compatible with the methodology given in Volume II of this final report is shown below.

$$P_{QT} = K_{P_{QT}} W_{QT} \quad (1)$$

where

P_{QT} = The power required by the transmitter acquisition and tracking subsystem.

$K_{P_{QT}}$ = Constant relating transmitter acquisition and tracking equipment power to acquisition and tracking weight.

W_{QT} = Total transmitter acquisition and tracking weight.

The total transmitter acquisition and tracking weight, W_{QT} , may be expressed in terms of the transmitting aperture diameter. This was done in a previous topic but is repeated here for completeness.

$$W_{QT} = W_{BT} + K_{W_{AT}} K_{d_T} d_T^{n_T} \quad (2)$$

where

W_{QT} = Total transmitter acquisition and tracking weight

W_{BT} = Transmitter acquisition and tracking weight independent of aperture size

$K_{W_{AT}}$ = Constant relating transmitter acquisition and tracking weight to transmitter antenna weight

K_{d_T} = Constant relating transmitter antenna weight to transmitter antenna diameter

n_T = Exponent relating transmitter antenna weight

d_T = Transmitter Aperture Diameter

The total expression for the required power is then

$$P_{QT} = K_{P_{QT}} \left[W_{P_{QT}} + K_{W_{AT}} K_{d_T} d_T^{n_T} \right] \quad (3)$$

This is plotted in Figures A and B for a radio and optical system, respectively.

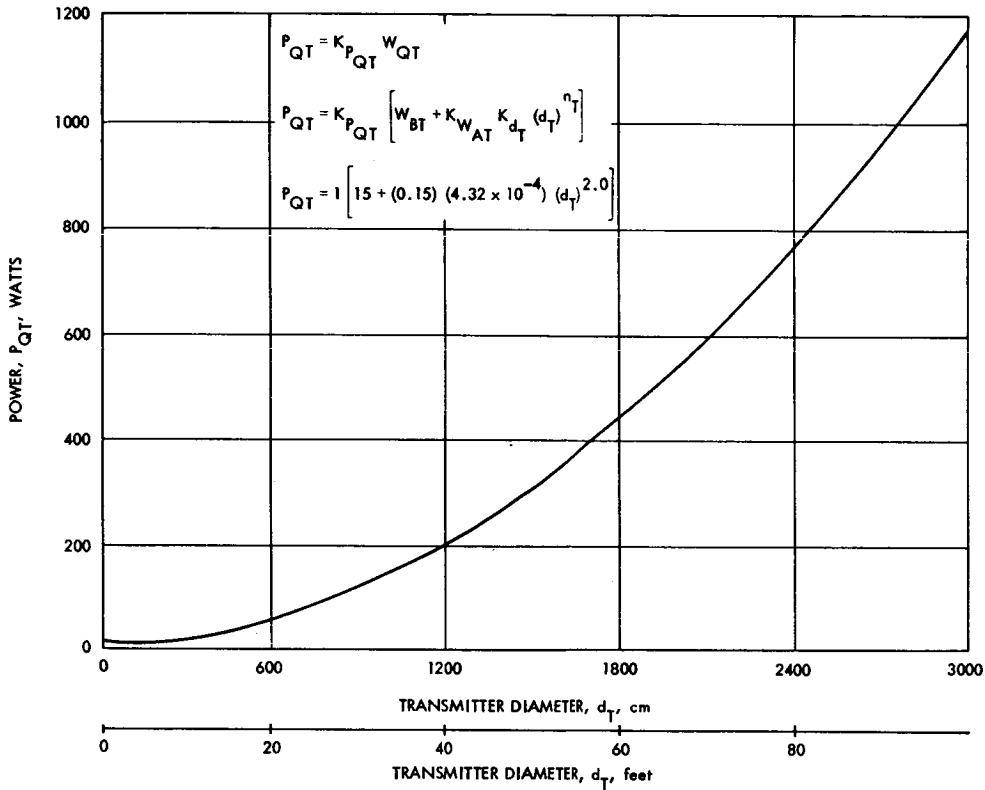


Figure A. Power for Radio Acquisition and Tracking (Transmitting or Receiving)

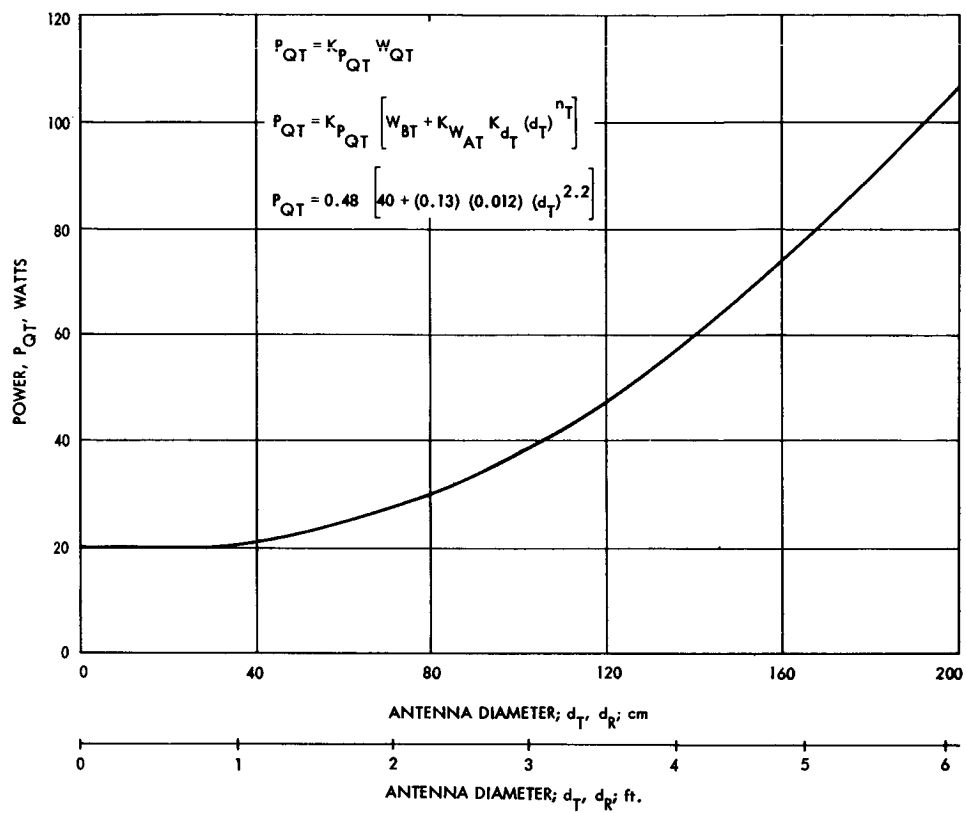


Figure B. Optical Transmitter (Receiver) Acquisition and Tracking, Power

PART 6 – PRIME POWER SYSTEMS

Section	Page
Solar Power Systems	482
Nuclear Power Systems	490
Chemical Power Systems	508
Power Summary	516

INTRODUCTION

Three types of prime power systems are discussed: Solar power systems, nuclear power systems, and chemical power systems.

Mission constraints of duration, power requirement, environment, and goals play an important role in the selection of the prime power source or sources. The three types of power sources used in space missions to date are documented in this part. Some implementations of these types of sources have been flown but not all, where applicable this is noted.

Solar Power Systems

The most useful means of converting solar power into electric power is by means of solar cells. Other means include solar thermoelectric systems, solar thermionic systems, and solar dynamic systems.

Nuclear Power Systems

The heat of nuclear fission is used to generate electric power. This may be done passively as in a radio isotope source or actively as in reactor dynamic power systems. These and different methods of converting the heat energy into electric energy are discussed.

Chemical Power Systems

Chemical power systems are in the form of batteries and fuel cells. The lifetime, cycling capability and capacity of these sources are given.

SUMMARY

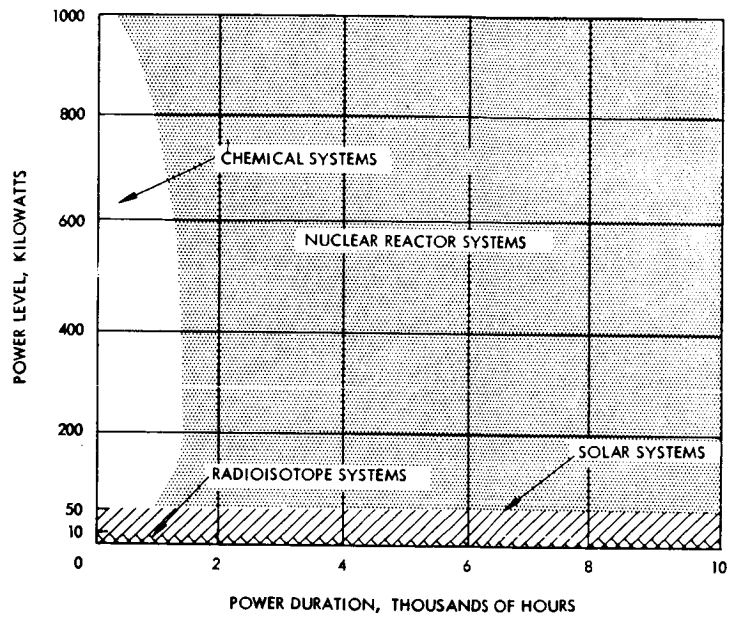
The type of power source used depends upon the mission, its duration and the power requirements.

The selection of a spacecraft power system for a specific mission depends on the power requirements of the mission, the mission duration, and the environment in which the system must operate. The effect of these constraints on the selection of power system of a few watts to kilowatts capacity is discussed in this Part.

Spacecraft power systems may be classified into three general categories according to the initial energy source as solar, nuclear, or chemical. The weight of solar and nuclear systems is generally not a strong function of mission duration, whereas the weight of the chemical system is decidedly so. Typical power system selections based on a solar distance of 1AU* are shown in the figure¹ as a function of power level and mission duration. Nearly every power system will include some provision for energy storage to provide for peak power demands and, in the case of solar systems, to provide continued power during periods of solar eclipse. The extent and type of energy storage required depends critically on the exact mission power history and, in the case of solar systems, on the solar illumination history.

*1AU (astronomical unit) $\approx 149.6 \times 10^6$ km.

¹1967 NASA Authorization, Part 4, United States Government Printing Office, Washington, D.C., 1966.



Power System Range of Application at Near-Earth
(1 AU) Solar Distance

PRIME POWER SYSTEMS

Solar Power Systems

	Page
Solar Voltaic Systems	482
Solar Cell Degradation in a Space Environment	484
Solar Thermal Systems	486

SOLAR VOLTAIC SYSTEMS

Solar cells are very useful as power sources in space. Constants are given which relate weight, power, cost and area of solar cells.

Basically there are two types of solar systems: photovoltaic and solar thermal systems. The first group will be discussed in this topic and the second group in a later topic. The latter group includes solar thermoelectric and solar thermionic systems as well as solar dynamic systems of various types.

Solar photovoltaic (solar cell) systems are presently by far the most appealing solar power system. At present they alone have proven reliability. They are considerably more efficient (hence lighter and more compact) than solar thermoelectric systems. By comparison with solar thermionic and solar dynamic systems, they are relatively insensitive to angular orientation with respect to the incident solar illumination. Photovoltaic array power output is directly proportional to surface area (hence weight) to powers of many kilowatts.

Over the range of solar illumination for which photovoltaic arrays are useful, cell efficiency is a function only of array temperature, i. e., power output is directly proportional to illumination intensity. Hence, for constant array temperature, the specific weight increases as the square of the solar distance. Conversely, the specific power (watts/kg) decreases as the square of the solar distance. Array specific power at constant illumination decreases as cell temperature increases. This is because solar cell efficiency is an inverse function of temperature as seen in Figure A. As a result, photovoltaic array power passes through a maximum as solar distance is reduced to about 0.5 AU and then is sharply reduced (Figures B and C).¹ Solar array weight and area as a function of unregulated output power are shown in the table for various solar distances based on expected capability in the near future. Solar photovoltaic power system cost constants are also given in the table. These estimates include the solar cells, supporting structure, mechanisms, and wiring, but not power conversion and distribution equipment.

At intermediate power levels, a choice must be made between an oriented or a non-oriented photovoltaic array and between various geometric array configurations. The choice depends on 1) the trade-off between decreased array weight, size, and cost and increased system complexity and development cost with orientation, 2) limitations on array moment of inertia imposed by the vehicle attitude control system, and 3) packaging limitations.

¹Gross, Sidney, "Discussion of Power Systems for Solar Probes - Solar Photovoltaic Concepts, " Proceedings of the Intersociety Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

Table A. Solar Cell Power, Weight, and Cost Parameters

Position of Solar Cells	Power Output Watts/cm ²	Weight Kilograms/watt	Cost Dollars/watt
Mercury (.39AU)	1.87×10^{-2}	0.0186	43
Venus (.72AU)	2.10×10^{-2}	0.01775	38.3
Earth (1.0AU)	1.52×10^{-2}	0.0227	53
Mars (1.52AU)	0.72×10^{-2}	0.0481	112
Jupiter (5.2AU)	0.06×10^{-2}	0.575	1350

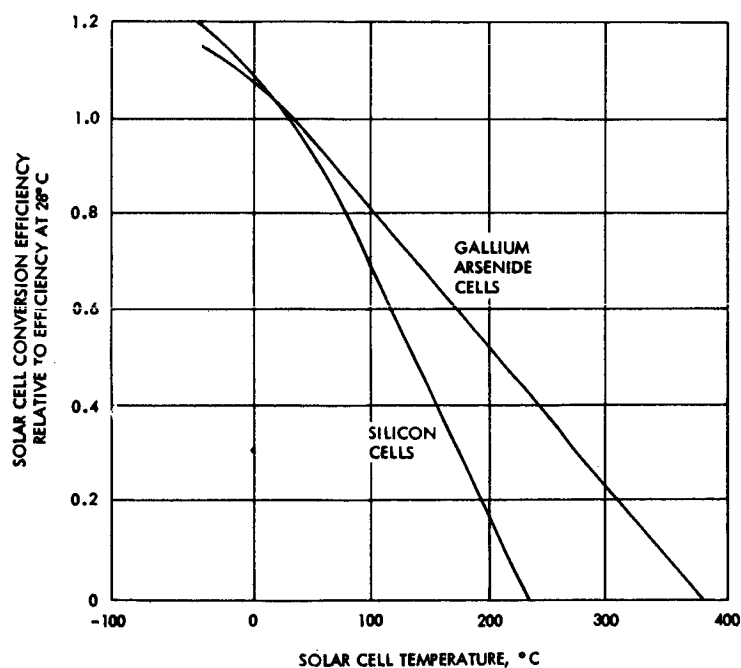


Figure A. Temperature Dependency of Solar Cell Efficiency

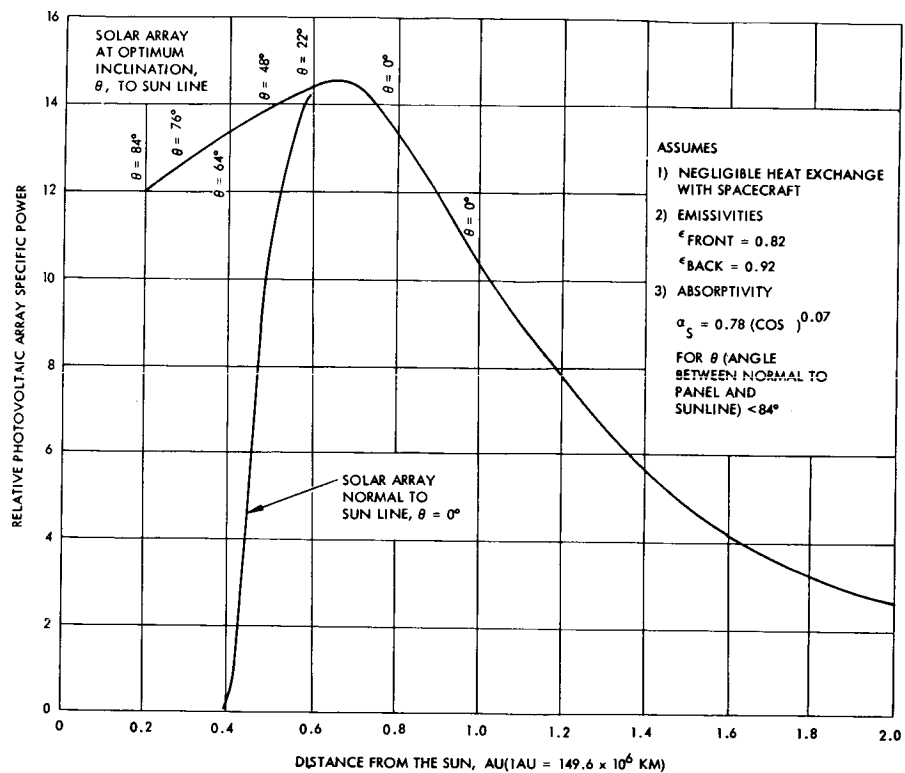


Figure B. Photovoltaic Array Specific Power Versus Solar Distance - Silicon Cells

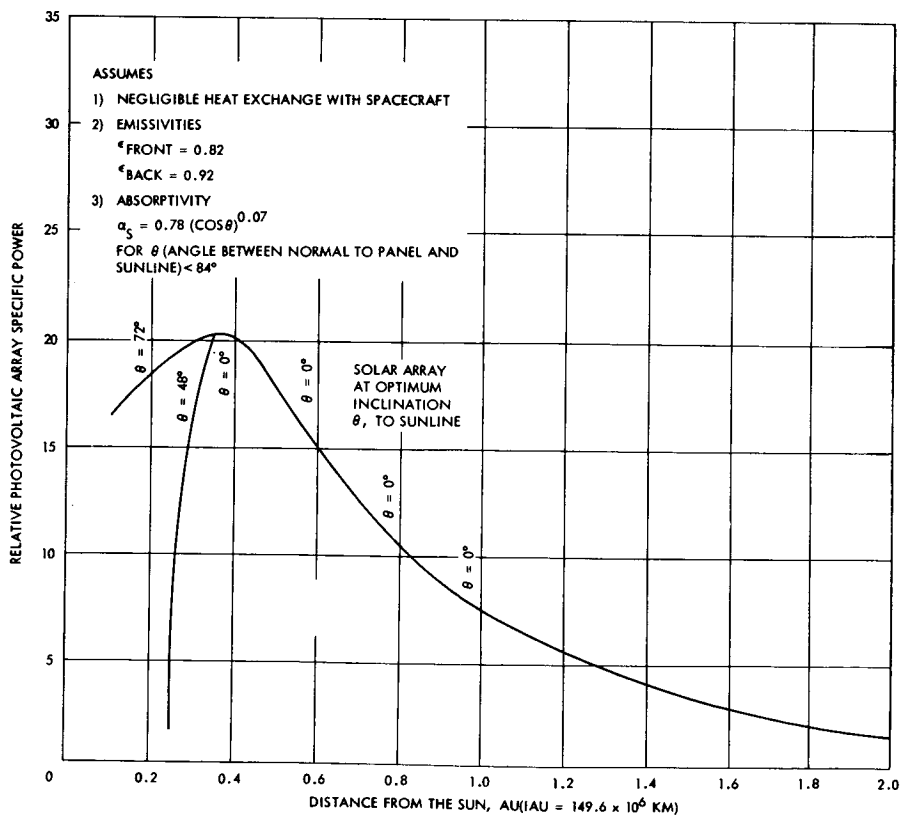


Figure C. Photovoltaic Array Specific Power Versus Solar Distance - Gallium Arsenide Cells

SOLAR CELL DEGRADATION IN A SPACE ENVIRONMENT

The amount of solar cell degradation is estimated, based upon a March 1969 launch to Mars.

Solar cell performance is degraded by the particle radiation environment encountered in space. The conventional technique for protection of solar cells against the degrading effects of particle radiation is the use of transparent cell covers as shielding. The degree of protection from radiation damage afforded by such methods depends of course on the thickness of the covers as well as upon the protective properties of the material used. A supplementary technique consists of overdesigning the array to an extent that it will maintain an acceptable power output for the duration of the mission after being degraded to the degree predicted on the basis of the known or anticipated radiation environment. For a given radiation environment and power requirement, there is an optimum compromise for minimum weight between increasing cell cover thickness to reduce degradation and increasing the size array so that it will continue to deliver the required power after being degraded. A study of solar array degradation as a function of time during Earth-Mars transit due to solar flare activity has been made for 30- and 45-mil quartz covers on the basis of an assumed March 1969 launch date. The principal cause of array degradation by solar flares is proton radiation. Thirty- and 45-mil quartz covers are impervious to protons having energies less than 10 mev and 15 mev, respectively. Tables A and B list approximate percent of initial solar array power capability remaining in successive months following the March 1969 launch for 30- and 45-mil quartz covers.

Table A. Solar Array Degradation in Earth-Mars Transit by
Solar Flare Activity Using n on p Silicon Cells with
30-Mil Quartz Covers

Month	Total Proton Flux (Protons/Cm ² having E > 10 mev)		Percent of Original Power Capability Remaining	
	Maximum	Minimum	Maximum	Minimum
1969				
March (Launch)	0	0	100	100
April	5×10^9	2.8×10^9	95	93
May	1.4×10^{10}	5×10^9	93	90
June	1.8×10^{10}	8×10^9	91.3	88
July	3×10^{10}	1.4×10^{10}	90	86.6
August	4×10^{10}	1.6×10^{10}	89	85.2
September	5×10^{10}	1.8×10^{10}	88	84.2
October	5.5×10^{10}	2.5×10^{10}	87.2	83.2
November	6.0×10^{10}	3×10^{10}	86.8	82.6
December	6.3×10^{10}	3.4×10^{10}	86.0	82.0
1970				
January	7×10^{10}	4×10^{10}	85.7	81.0
February	7.5×10^{10}	4.4×10^{10}	85.0	80.6
March (Intercept)	8×10^{10}	4.8×10^{10}	84.6	80.0

Table B. Solar Array Degradation in Earth-Mars Transit by
Solar Flare Activity Using n on p Silicon Cells with
45-Mil Quartz Covers

Month	Total Proton Flux (Protons/Cm ² having E > 10 mev)		Percent of Original Power Capability Remaining	
	Maximum	Minimum	Maximum	Minimum
1969				
March (Launch)	0	0	100	100
April	2.5×10^9	1.4×10^9	96	95
May	7×10^9	2.5×10^9	95	92
June	9×10^9	4×10^9	94	91
July	1.5×10^{10}	7×10^9	92	89
August	2×10^{10}	8×10^9	91.5	87.5
September	2.5×10^{10}	9×10^9	91.0	86.5
October	2.75×10^{10}	1.25×10^{10}	90.0	86
November	3×10^{10}	1.5×10^{10}	89.0	85.8
December	3.15×10^{10}	1.7×10^{10}	88.5	85.6
1970				
January	3.5×10^{10}	2×10^{10}	87.5	85.4
February	3.75×10^{10}	2.2×10^{10}	87.3	85.3
March (Intercept)	4×10^{10}	2.4×10^{10}	87.0	85.1

SOLAR THERMAL SYSTEMS

Solar thermoelectric, thermionic, and dynamic systems are described. They may find greatest use in a high solar flux near the sun.

Solar Thermoelectric Systems. Solar thermoelectric systems consist of thermoelectric elements heated by solar illumination either directly or using concentrators. In either case, the low efficiency of the thermoelectric conversion process (5 percent) leads to specific weights upward of 91 kg/kw¹ (200 lb/kw) or 10.1 watts/kg (5 watts/lb) at 1 AU solar distance. They may, however, be competitive with photovoltaic systems for solar probes where solar cell efficiency is severely degraded by high temperature. Estimated solar thermoelectric power system weight and area interdependancies upon power are given in the Table for 1 AU and 0.3 AU solar distances.²

Solar Thermionic Systems. Solar thermionic systems use concentrating mirrors to focus solar energy on thermionic converters. Although they are not competitive with photovoltaic systems on a weight basis, considerable developmental effort has been expended in the hope that they will be superior in high radiation environments or high operating temperatures (such as might be encountered on solar probe missions). Orientation accuracies required for the large collectors (5 minutes of arc) and the problems associated with their deployment and possible degradation by the space environment pose severe problems. Solar thermionic systems are still in too early a stage of development for accurate evaluation. Estimated solar thermionic power system weight and area interdependancies on power are also noted in the Table for solar distance regimes of 0.1 to 0.3 AU and 0.35 to 0.7 AU.²

Solar Dynamic Systems. Solar dynamic systems are characterized by the use of a heat engine (typically a turbine) to drive an electrical generator. Solar Brayton cycle systems are regarded as promising for high-temperature and high-radiation environments, but share with solar thermionic systems the problems associated with deployment and orientation of large collecting areas in space. They are not sufficiently developed to permit accurate evaluation.

¹Rappaport, Paul, "Space Power: The Next Step," Space/Aeronautics, 45, Number 4, 1. 76, September 1965.

²Brosens, P. J., "Discussion of Solar Power Systems for Solar Probes - Thermoelectris and Thermionics," Proceedings of the Intersociety Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

Power, Weight and Area Interdependencies for Thermoelectric
and Thermionic Power Systems

Distance from the Sun, AU	Thermoelectric Power		Thermionic Power	
	Watts/cm ²	Kg/cm ²	Watts/cm ²	Kg/cm ²
1.0	1.66×10^{-2}	0.09		
0.3	2.54×10^{-2}	0.051		
0.7 to 0.35			2.06×10^{-3}	0.0373
0.3 to 0.1			6.36×10^{-2}	0.0256

PRIME POWER SYSTEMS

Nuclear Power Systems

	Page
Introduction to Nuclear Power Systems	490
Thermoelectric Reactor Power Systems	492
Thermionic Reactor Systems	496
Reactor Dynamic Power Systems — Brayton Cycle	498
Reactor Dynamic Power Systems — Rankine Cycle	500
Radioisotope Thermoelectric Systems	502
Radioisotope Dynamic Systems	506

INTRODUCTION TO NUCLEAR POWER SYSTEMS

The advantages of nuclear power systems include their long life and independence of solar radiation.

Nuclear power systems convert the thermal energy generated by nuclear reactors or isotope decay to electrical energy by the same conversion cycles employed with solar thermal power systems, i.e., turboalternators, thermionics, or thermoelectrics. The primary advantages of nuclear systems over solar systems are the independence from solar illumination and their potentially lower specific weight at high power levels. Their main disadvantage is the nuclear radiation produced and shielding required to protect personnel or radiation-sensitive equipment from it.

Reactor power sources are generally applicable to high power levels (about 10kw), and may be competitive at much lower levels for deep space missions. This results from the relatively high specific weight for low power designs due to the heavy reactor and shielding. Problems with reactor system reliability and life have been experienced partly because of the necessity to integrate sophisticated nucleonic control systems and high temperature fluid systems.

Reactor problems may be avoided by using radioisotopes as energy sources because their thermal energy output is continuous and predictable. Radioisotope thermoelectric systems can be designed for very low power levels and like reactors can operate for very long times. Although their specific weight is one of the lowest for power levels from 1 to 10kw, radioisotope systems are seldom used where alternative systems can provide competitive performance. One reason for this is the higher cost of radioisotope systems resulting from the limited quantities of isotopes available; another is the complications introduced by radiation produced.

THERMOELECTRIC REACTOR POWER SYSTEMS

The theory of reactor power systems operation is given with data on performance.

Reactors can serve as a heat source for thermoelectric converters. Thermoelectric converters transform heat energy to electrical energy by means of the Seebeck effect in a thermoelectric couple. The configuration of a thermoelectric converter element is shown schematically in Figure A. Heat is transferred to the thermoelectric elements P and N (two dissimilar conductors or semiconductors) through the copper block on top and rejected through the two copper blocks at the bottom. As a result of the Seebeck effect, a voltage is developed between the bottom ends of the two thermoelectric elements. In practical converters, large numbers of such elements are combined in series as in Figure B and then in parallel to produce usable voltages and currents.

A typical thermoelectric power system is illustrated schematically in Figure C. An actual system configuration is shown in Figure D. Because of the relatively low efficiency of thermoelectric conversion (5 to 8 percent) extensive radiator area is required. A 570-watt reactor thermoelectric power system has been successfully tested in space for 43 days and has operated up to 10,000 hours in a simulated space environment. System parameters, including weight, cost, and volume parameters, for reactor thermoelectric systems of the SNAP 10A type developed by Atomics International are shown in the Table¹ for various power levels.

¹Glyfe, J.D., and Wimmer, R.E., "Reactor Thermoelectric Power Systems for Unmanned Satellite Applications," Proceedings of the Inter-society Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

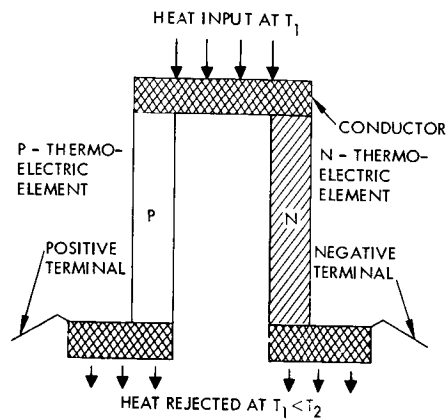


Figure A. Basic Thermoelectric Couple, Schematic Arrangement

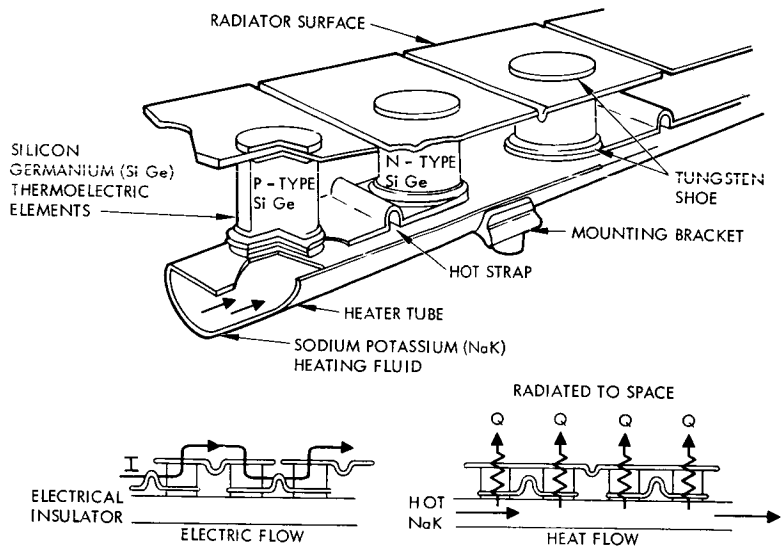


Figure B. Direct Radiating Thermoelectric Module

THERMOELECTRIC REACTOR POWER SYSTEMS

Reactor Thermoelectric System Performance (Atoms International)

Net Power, Electrical (kw)	0.5	1.0	2.0	5.0	10.0	15.0	20.0
Reactor Power, Thermal (kw)	28.4	47.1	84.8	212	424	635	850
Reactor Outlet Temperature ($^{\circ}\text{F}$)	1300	1300	1300	1300	1300	1300	1300
Design Life, Rated Power (yr)	3	3	3	3	3	3	3
Gross Radiator Area (ft^2)	24	48	98	245	525	865	1150
(cm^2)	2.23×10^4	4.46×10^4	9.1×10^4	22.8×10^4	48.8×10^4	80.4×10^4	107×10^4
Base Diameter (ft)	2.84	3.67	4.84	7.3	10.0	12.7	15.2
(cm)	86.7	111.5	147.5	222.5	305	388	463
Overall Height (ft)	7.67	10.17	13.58	21	31.5	41.5	50.5
(cm)	234	310	414	641	960	1265	1540
Unshielded System Weight (lb)	524	633	852	1740	2885	4215	5850
(kg)	238	288	387	791	1310	1915	2660
Reactor - Payload Separation Distance (ft)	45	52	65	80	100	100	100
(cm)	13.7×10^2	15.9×10^2	19.9×10^2	2.48×10^2	30.5×10^2	30.5×10^2	30.5×10^2
Total Shielded System Weight (lb)	685	829	1076	2385	3670	5060	6830
(kg)	311	377	489	1085	1670	2300	3100
Specific Power (10^3 lb/kw)	1.37	0.83	0.54	0.48	0.37	0.33	0.32
(kg/kw)	623	377	245	218	168	150	145
Specific Cost* (10^3 \$/kw)	1600			600		470	450
	1250			400		310	300

*Upper figure is current specific cost; lower figure is potential specific cost with quantity production.

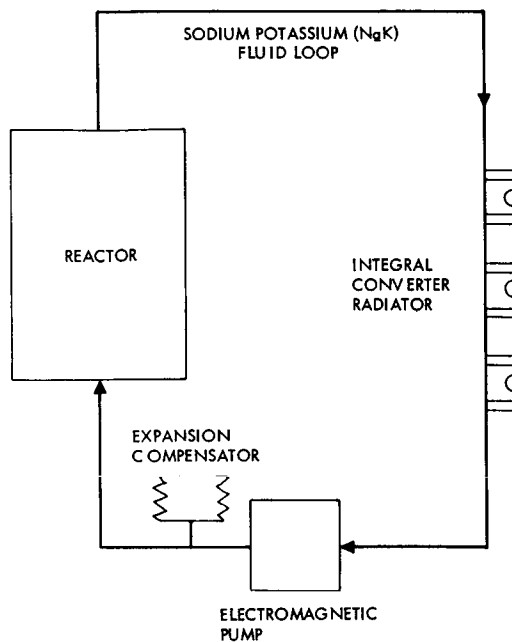


Figure C. Reactor-Thermoelectric System Schematic

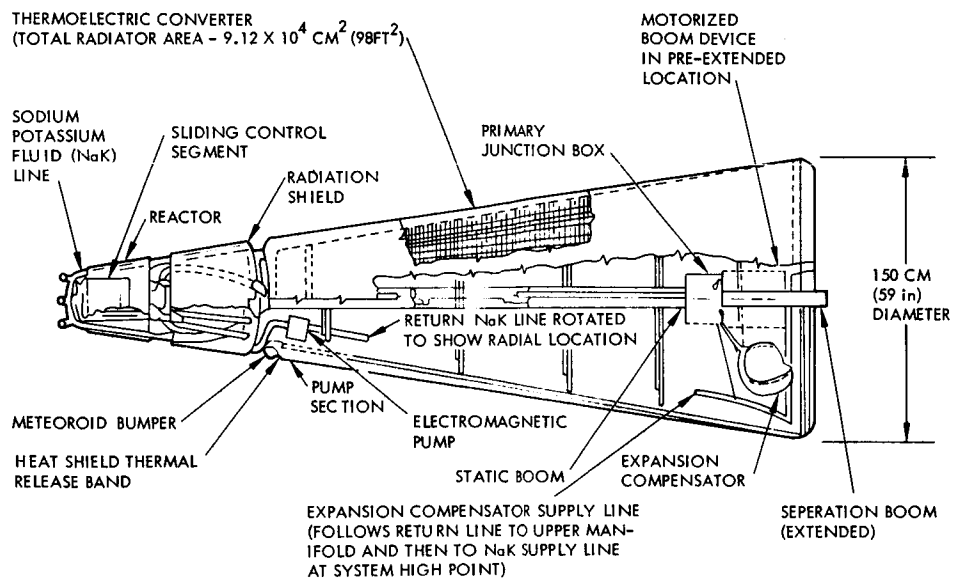


Figure D. Two Electrical Kilowatt Reactor Thermoelectric Power System (Atoms International)

THERMIONIC REACTOR SYSTEMS

Thermionic systems use thermionic diodes heated by the reactor to produce electricity. Temperatures in the order of 3000°C are used.

A thermionic diode converter can be used to convert heat into electricity. The device is illustrated schematically in Figure A. A hot cathode emits electrons which travel across a narrow gap to a relatively cool anode. If the cathode and anode are connected externally, the electrons collected on the anode return to the cathode through the external circuit. Thus an electric current i is established. If a load is inserted in the external circuit, a potential difference is developed across it with the signs as indicated. Insofar as the external circuit is concerned, the cathode is the positive terminal and the anode is the negative terminal of the thermionic generator.

Heat is continually supplied to the cathode to compensate for the energy taken by the emitted electrons and loss of heat from the cathode due to radiation, convection, and conduction. As the external surface of the cathode is entirely used for transfer of heat from the heat source to the cathode, this loss of heat from the cathode is mostly transferred to the anode. They may be rejected in order to keep the anode temperature from rising too high.

The thermionic diode is a heat engine: thermal energy is supplied to the cathode at a high temperature T_2 and a portion of it rejected from the anode at a lower temperature T_1 . The conversion efficiency is proportional to $(T_2 - T_1)/T_1$; the greater the temperature difference $T_2 - T_1$, the more efficient the diode is. Thus a reactor is theoretically well-suited as a thermionic heat source since it can provide heat at high emitter temperatures with the resultant increased efficiency and decreased specific weight. The major problems associated with thermionic converters concern the very high emitter temperatures that are required to obtain lightweight systems and the resultant reliability problems. Various reactor thermionic system configurations are illustrated in Figure B. Of these, the converter integral with the reactor offers potentially the lowest specific weight and volume, but poses the greatest reliability problems because it operates with the highest emitter temperature. The thermionic converter has higher efficiency (25 to 30 percent at 2000°C emitter temperature) than thermoelectric converters (5 to 8 percent) because of its high operating temperatures. Its higher operating temperatures also permit higher heat rejection temperatures which reduce the radiator area requirements. Reactor thermionic systems have not yet reached the flight hardware phase and are not likely to be adequately developed for flight use before the mid 1970's. It is not possible at this time to estimate accurately costs or volumes of space qualified reactor thermionic systems. Estimated weights relationships of reactor thermionic systems for various emitter operating temperatures are shown in the table.¹

¹ 1967 Authorization, Part 4, United States Government Printing Office, Washington, D. C., 1966.

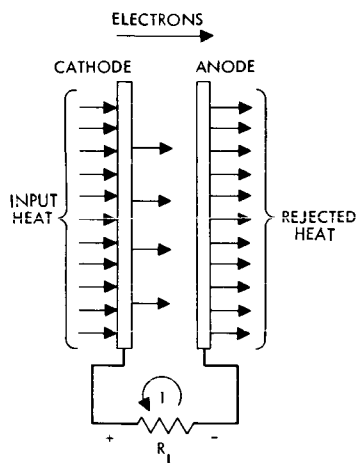


Figure A. A Thermionic Diode Converter of Heat to Electrical Energy

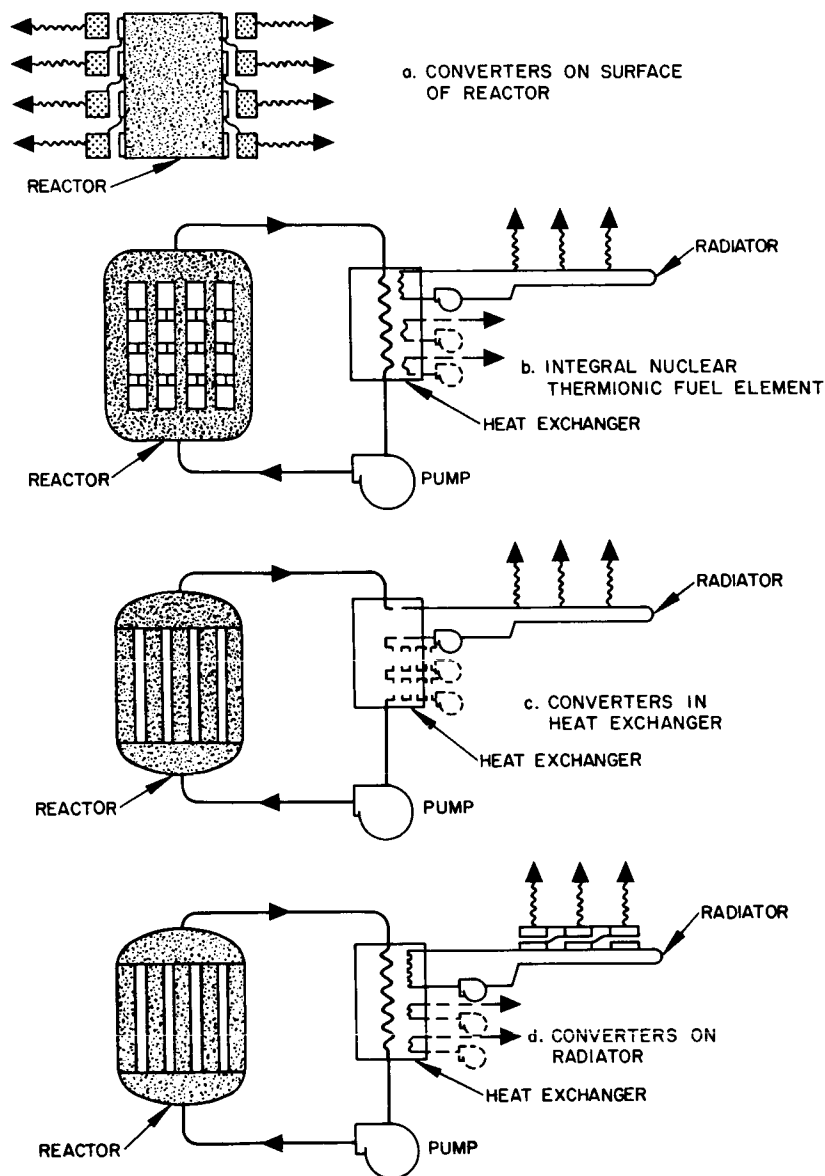


Figure B. Reactor Thermionic Power Systems

Nuclear Thermionic Power/Weight Relationship

Emitter Temperature, °C	Kilograms / Kilowatt	Pounds / Kilowatt
2700	38.6	85
2800	29.5	65
3000	18.2	40

REACTOR DYNAMIC POWER SYSTEMS – BRAYTON CYCLE

A Brayton cycle dynamic reactor using a turbine is described which produces powers in the order of 3 kw.

Dynamic power systems are characterized by the use of a heat engine (reciprocating engine or turbine) to drive an electrical generator. Reactor dynamic power systems appear attractive for high power requirements. Although no complete power system is presently being developed, component development continues on several types. The two heat engine types which are presently being studied are the Brayton cycle and the Rankine cycle. The Brayton cycle is described below and the Rankine cycle is described in the next topic.

The working fluid for the Brayton cycle is an inert gas such as argon or neon. Heat input is at constant pressure from a suitable heat source. The hot gas is expanded through a turbine and the waste heat is rejected in a radiator at a continuously decreasing temperature. The gas is compressed and the cycle repeated.

The principal advantages of the Brayton cycle are

1. The inherent simplicity of a single loop and a single phase working fluid enhances the system reliability.
2. The corrosion free atmosphere provided by the inert gas will allow use of uncoated high temperature refractory alloys without fear of corrosion or oxidation.
3. Similarly, the inert gas system should also be erosion free because there will be no solid or unburned particles in the working fluid.

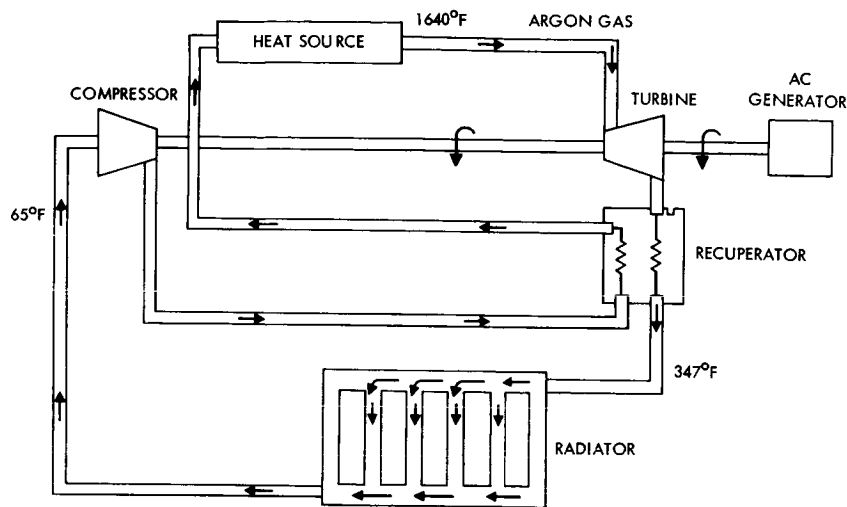
Its major disadvantages are

1. Since most of the heat of the cycle is not added at the highest temperature or rejected at the lowest temperature of the cycle, the efficiency of the simple Brayton cycle is low. However, this can be largely offset by the proper selection of cycle temperatures and the use of a regenerator.
2. Considerable pumping power is required for the compression process in the Brayton cycle as compared to pumping a liquid in the Rankine cycle.
3. The continuously decreasing temperature in the radiator and the low radiator outlet temperature increase the heat rejection problem. Since the only means of heat rejection is by radiation, a large radiating area is required.

A module of a SNAP 8 reactor, powered 10-kw Brayton-cycle system, is shown schematically in the figure.¹

¹ 1967 Authorization, Part 4, United States Government Printing Office, Washington, D.C., 1966.

The module has a design weight of approximately 330 kg (725 pounds) or 33 kg/kw (72.5 pounds/kw) and requires a radiator area of approximately $5.57 \times 10^5 \text{ cm}^2$ (600 ft²). The measured efficiency of a 3-kw demonstration Brayton-cycle system was 18 percent (electrical output/heat input). No flight tests of a complete system are presently scheduled, and an operational system is not expected until the 1970's.



Brayton Cycle Dynamic Power System

REACTOR DYNAMIC POWER SYSTEMS — RANKINE CYCLE

Rankine cycle power systems have a higher efficiency than Brayton cycle systems. Prototype units have produced 10 kw with an unshielded weight of 2000 pounds.

As mentioned in the previous topic reactor dynamic power systems use a reciprocating engine or a turbine operating from heat produced by a reactor. Heat input to the Rankine cycle is used to vaporize and, if required, superheat the working fluid until the desired conditions are achieved at the turbine inlet. The waste heat at the turbine exhaust is dissipated by radiation to space until the fluid is completely condensed to a liquid. The liquid is then pumped to the boiler where the cycle is repeated. Superheating is generally required to prevent the possibility of any vapor condensing during the expansion process which would cause erosion and a reduction of prime mover efficiency. A wide variety of working fluids can be used, but for space applications, liquid metals and organic fluids are receiving the most attention.

The principal advantages of the Rankine cycle are

1. The efficiency approaches that of the Carnot cycle since most of the heat is added isothermally and most of the waste heat is rejected isothermally.
2. Isothermal rejection of waste heat is desirable from the standpoint of minimizing the radiator area. In addition, the heat rejection temperature can be considerably higher than for the Brayton or Stirling cycles.
3. The Rankine cycle has received the greatest development effort.

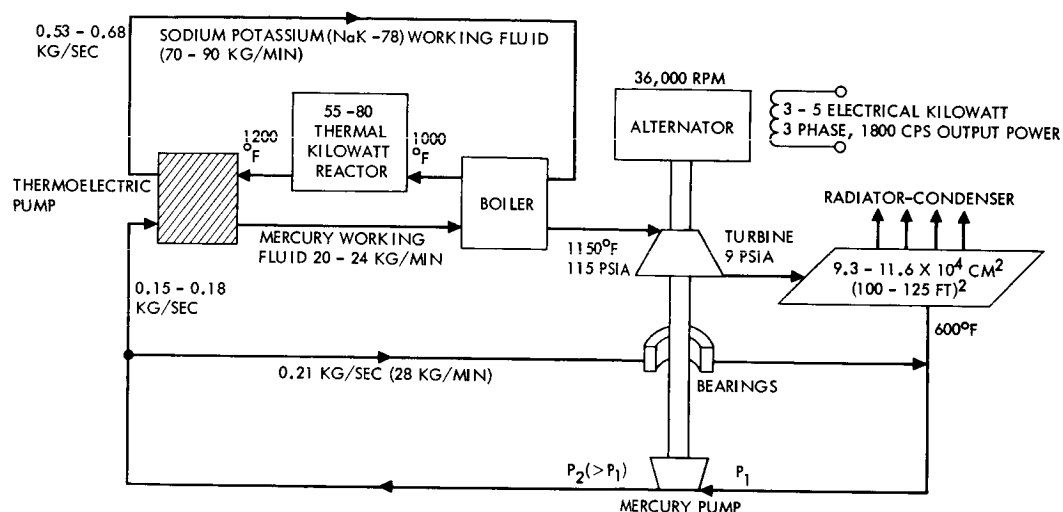
The principal disadvantages of the Rankine cycle are

1. The cycle is mechanically complex, two or more loops may be required, particularly in the case of nuclear heat sources.
2. The corrosion and erosion problems associated with metal vapors may adversely affect the system life.

A Mercury-Rankine component development program has been underway for a number of years. The objective has been to develop and qualify components for a basic 3 to 5 kw module. A typical Reactor Mercury Rankine power system based on present components is shown in the figure.¹ The reactor is used to heat sodium-potassium (NaK) fluid in a sealed recirculating loop. The hot sodium-potassium is circulated through a boiler to evaporate and superheat the mercury working fluid. The mercury vapor is then expanded within a turbine which drives a

¹Wallerstedt, R. L., and Owens, J. J., "SNAP Mercury Rankine Program," Intersociety Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

permanent magnet alternator and a mercury boiler feed pump. These three components are all mounted on a single rotating shaft and are supported by mercury lubricated bearings, the shaft rotating at 36,000 rpm. This assembly is called the combined rotating unit. The exhaust mercury is then condensed and subcooled by a radiator-condenser and is pumped back to the boiler by means of the high-pressure feed pump. The alternator generates alternating current electrical power at 1800 Hz. A 10-kw unshielded Mercury-Rakine system of this type is expected to weigh about 910 kg (2000 pounds) or 91 kg/kw (200 pounds/kw). It is not expected that space qualified Mercury-Rankine systems will be available before 1970. No specific mission applications have yet been defined.



Reactor-Rankine Cycle Dynamic Power System

RADIOISOTOPE THERMOELECTRIC SYSTEMS

Radioactive isotopes produce heat as a result of the radioactive decay process. In principle, they may be used as a heat source for thermionic, thermoelectric, and dynamic power systems. Thermoelectric systems are described in this topic.

Radioisotope thermoelectric generator (RTG) systems are applicable to long duration power requirements of less than 1 kw where independence from solar illumination or resistance to radiation degradation is a constraint. Specific weights of existing low power (e. g., the 50 watt SNAP 27) systems are in the range of 0.32 kg/watt (0.7 lb/watt). Higher power systems are expected to have specific weights of 0.23 kg/watt (0.5 lb/watt) or less.¹ Specific costs are strongly dependent upon the cost of the isotope used. The choice depends on the extent of shielding permitted and the operating lifetime. For extended missions an isotope having a long half-life is required in order to minimize the variation in heat output over the variation in heat output over the mission. Characteristics of typical radioisotope fuels are tabulated in Table A.

Since thermoelectric conversion is relatively inefficient (5 to 8 percent) an adequate heat rejection system must be included. The simplest such system consists of fins having a high emissivity coating. Since the heat output of the radioisotope source decays exponentially with time, more heat is produced at the beginning than at the end of the RTG's design life, requiring additional thermal control in order to maintain the hot junction temperature thermoelectric elements at a constant optimum level. The most common technique is the use of a high temperature radiating surface to bypass surplus heat away from the thermocouples at the beginning of the mission. The area of this radiator and hence the amount of heat bypassed is regulated by a thermostatically controlled shutter, the position of which is a function of radioisotope half-life and activity.

A further environmental consideration for RTG power systems is the effect of nuclear radiation produced by the decaying radioisotope fuel on the spacecraft and its payload. (To comply with limitations on the Surveyor Spacecraft the SNAP-11 incorporated shielding weighing 0.09 kg/watt (0.2 lb/watt).)

Characteristics of a number of existing RTG power supplies are tabulated in Table B.²

¹Rappaport, Paul, "Space Power: The Next Step," Space/Aeronautics, 45, Number 4, P. 76, September, 1965.

²Barney, R., "Radioisotope Thermoelectric Generators," Research Report No. 14, Hughes Aircraft Company, Space Systems Division, Power Systems Department, September, 1966.

Table A. Typical Radioisotope Fuels

Isotope	Sr-90	Ce-144	Pm-147	Po-210	Pu-238
Half Life (years)	28	0.78	2.7	0.38	89
Power Density (w/cc)	1.1	24.5	1.8	1210	3.9
Source		Fission Products		Neutron Irradiation	
Potential Availability (kwh/yr)	66	800	12	140	4
Lead Time (years)	2-5	1-5	2-5	1-2	2
Estimated Cost (\$/thermal watt)	20	1	100	10 to 20	500 to 1000
Shielding Required in uranium Typical Manned System*	(4 inches) 10.15 cm	(6-1/2 inches) 16.5 cm	(1 inch) 2.54 cm	(1 inch) 2.54 cm	(24 inches LiH) 61 cm
Typical Unmanned System**	(0.2 inches) 0.518 cm	(1.5 inches) 3.81 cm	0	0	0
*1 mr/hr at 7.6 cm (3 feet) per kwthermal **100 r/hr at 7.6 cm (3 feet) per kwthermal					

Table B. Characteristics of Existing Radioisotope Thermoelectric Generators

	SNAP 11	SNAP 17A	SNAP 17B	SNAP 19	SNAP 27	SNAP 29
Fuel	Curium-242	Strontium-90	Strontium-90	Plutonium-238	Plutonium-238	Polonium-210
Vendor	Martin-Marietta	Martin-Marietta	General Electric	Martin-Marietta	General Electric	Martin-Marietta
Voltage	28 \pm 0 percent	28 \pm 0 percent	28 \pm 0 percent	24 \pm 2 percent	29 \pm 1 percent (voltage regulator)	Not available
Initial power output, watts	25	27.8	26	~50	57 to 56 (end of mission)	400
Weight, kilograms (pounds)	13.9 (30.5)	11.8 (26)	11.7 (25.72)	13.6 (30)	17.5 (38.47) total (includes fuel capsule cask)	~400
Watts/kg (watts/pound)	1.76 (0.8)	2.2 (1.0)	2.2 (1.0)	3.74 (1.7)	2.64 (1.2)	Not available
Kg/watt (pounds/watt)	0.57 (1.25)	2.2 (1.0)	2.2 (1.0)	0.27 (0.6)	0.32 (0.7)	Not available
Efficiency, percent	4.65	5.13, end of life	5.94, end of life	5.1 end of life	4 end of life	Not available
Mission design life	90 days	5 years	5 years	5 years	1-year lunar, preceded by 2-year earth storage	90 days
Hot junction temperature, °F (beginning of life)	1050	1500	1142	842	1100	Not available
Cold junction temperature, °F (beginning of life)	402 (day) 350 (night)	450	320	276	525	Not available
Radiator fin temperature, °F	367	435	310	265	510	Not available
Number of fins	2	6, equally spaced	6, equally spaced	2 (180 degree spacing)	8, equally spaced	Not available
Dimensions, cm (inches)	50.8 (20) diameter x 30.4 (12) long	44.5 (17.5) diameter x 31.8 (12.5) long	4.5 (17.75) diameter x 3.65 (14.38) long; barrel diameter 14.4 (5.67)	56 (22) diameter x 28 (11) long	41 (16.14) diameter x 39 (15.28) long; barrel diameter 13.1 (5.14)	Not available
Watts/ft ³	11.5	16.0	12.5	20.6	31.4	
Remarks		Design and development plan completed. Study terminated.	Design and development plan completed. Study terminated.	Launch late 1967 on Nimbus B	Used for the ALSEP. Fuel capsule inserted after lunar landing.	Initial design phase

RADIOISOTOPE THERMOELECTRIC SYSTEMS

Typical RTG cost as a function of unregulated output power is \$5000/watt for limited production and \$2200/watt for quantity production.³

A 277-watt Strontium 90 powered RTG proposed by General Electric is illustrated in the Figure.⁴ Its significant operating characteristics are shown in Table C. The unit weighs 71 kg (156 pounds), distributed as follows:

Heat Source	53.2 kg (117 pounds)
Heat Rejection and Structure	10.4 kg (23 pounds)
Thermoelectric Elements	<u>7.28 kg (16 pounds)</u>
Total	70.9 kg (156 pounds)

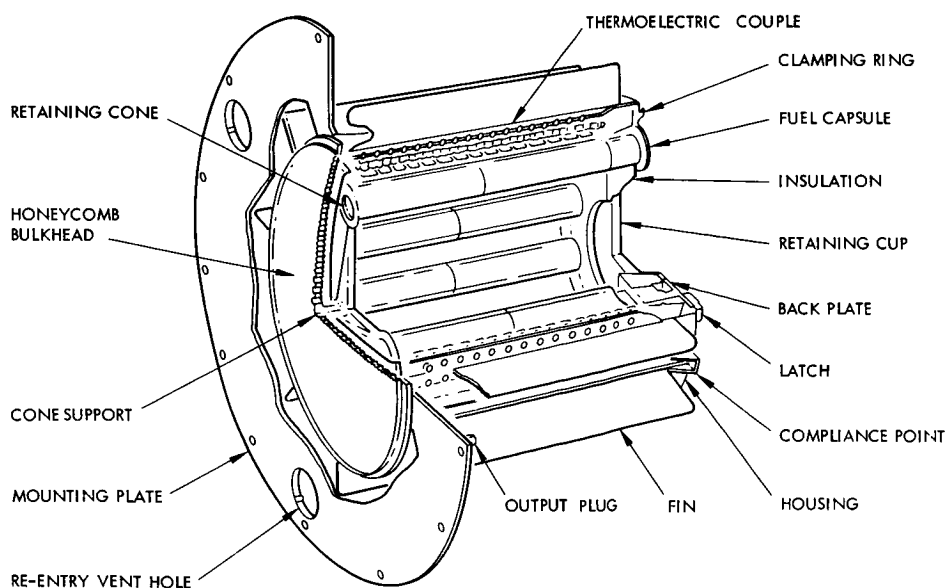
A reliability of 0.99 for one year and 0.95 to 0.75 for 5 years is claimed for this unit.

³Harris, E. D., and Dreyfuss, D. J., "Manned Spacecraft Electrical Power Systems: Requirements, Weight Correlation and Cost Implications," Proceedings of the Intersociety Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

⁴250 Watt Radioisotope Thermoelectric Power System, Presentation by the General Electric Corporation, Missile and Space Division.

Table C. 277-Watt Strontium 90 RTG Performance Summary

Operating lifetime	5 years	
Number of thermoelements	320 in 20 modules	
Number of isotope capsules	10	
Reliability		
1 year	0.99	
5 years	0.95 to 0.75	
Electrical-thermal history		
	<u>Beginning of Life</u>	<u>End of Life</u>
Power output (w)	364.0	277.0
Output voltage	28.5	28.5
Hot junction temperature (°F)	1820.0	1700.0
Cold junction temperature (°F)	775.0	740.0
Heat input (w)	7130.0	6317.0
Thermopile efficiency	6.16	5.32
Generator efficiency	5.1	4.38
Average capsule temperature (°F)	2000.0	1890.0
Maximum capsule temperature (°F)	2070.0	1960.0



277-Watt Strontium 90 Fueled RTG

RADIOISOTOPE DYNAMIC SYSTEMS

Radioisotope dynamic systems hold a potential of producing up to 10 kw using a Brayton cycle.

For radioisotope systems developing powers higher than about 1 kw there is a need for higher power conversion efficiency than can be obtained from thermoelectric converters. Because of this need the Brayton gas turbine discussed previously with potential efficiencies as high as 25 percent is the object of great interest. With the Brayton cycle power conversion system, it may be possible to obtain about 10 kw of radioisotope electric power, which is probably an upper limit considering radioisotope cost and availability. The radioisotope Brayton system is of particular interest in future manned missions such as orbital laboratories in which these higher powers are likely to be needed. A projected 11 kw isotope Brayton cycle system weighs 2260 kg (4967 lb) or 204 kg/kw (450 lb/kw) and has an overall efficiency of 21.6 percent.¹ This system is still in the preliminary design stage.

¹ Kirkland, Vern D., and McKhann, George G., "Preliminary Design and Vehicle, Integration of a Pu 238 Radioisotope Brayton Cycle Power System for MORL," Proceedings of the Intersociety Energy Engineering Conference, Los Angeles, California, September 26-68, 1966.

PRIME POWER SYSTEMS

Chemical Power Systems

	Page
Fuel Cells	508
Batteries	510

FUEL CELLS

Fuel cells are efficient power sources for relatively short duration missions.

A fuel cell is an electrochemical device in which the chemical energy of a conventional fuel is converted directly and efficiently into low voltage direct current electrical energy. One of the principal advantages of the fuel cell depends upon the conversion that can (at least in theory) be carried out isothermally, so that the Carnot limit on efficiency of heat engines does not apply. A fuel cell may be visualized as a primary battery in which the fuel and oxidizer are stored externally. The processes are illustrated schematically in Figure A. An actual fuel cell battery of the type developed by GE for the Gemini spacecraft is illustrated in Figure B.¹ This type uses a semi-permeable membrane electrolyte and hydrogen and oxygen as reactants. The hydrogen-oxygen fuel cell has received major emphasis in the manned spacecraft program because of its high efficiency and because it produces potable water as a by-product. Other types of hydrogen-oxygen fuel cells are the Bacon type and the capillary type. Although the differences between these types are basically confined to the cell itself, the operating pressures and temperatures are different, which in turn affects the reactant tank and radiator designs. In the Table power/weight relationships are indicated as a function of mission duration. The volume of fuel cells is $0.84 \times 10^4 \text{ cm}^3/\text{KW day}$ ($2.81 \text{ ft}^3/\text{KW day}$) for fuel and tankage volume.

As an example of fuel cell cost, a 2 KW system operating for 30 days costs 200/watt.²

Fuel Cell Weight per Watt

Days in Operation	Kilograms/watt	Pounds/watt
1	51	113
2	52.8	117
4	86.8	191
5	99.2	218
7	123	271
14	207	556
21	287	632
30	445	980
Fixed Weight	38.6	85

¹"Fuel Cells—Electrical Power Generation for Space Vehicles," General Electric Corporation, Lynn, Massachusetts, 1963.

²Allis Chalmers Manufacturing Company, private communication, May 29, 1967.

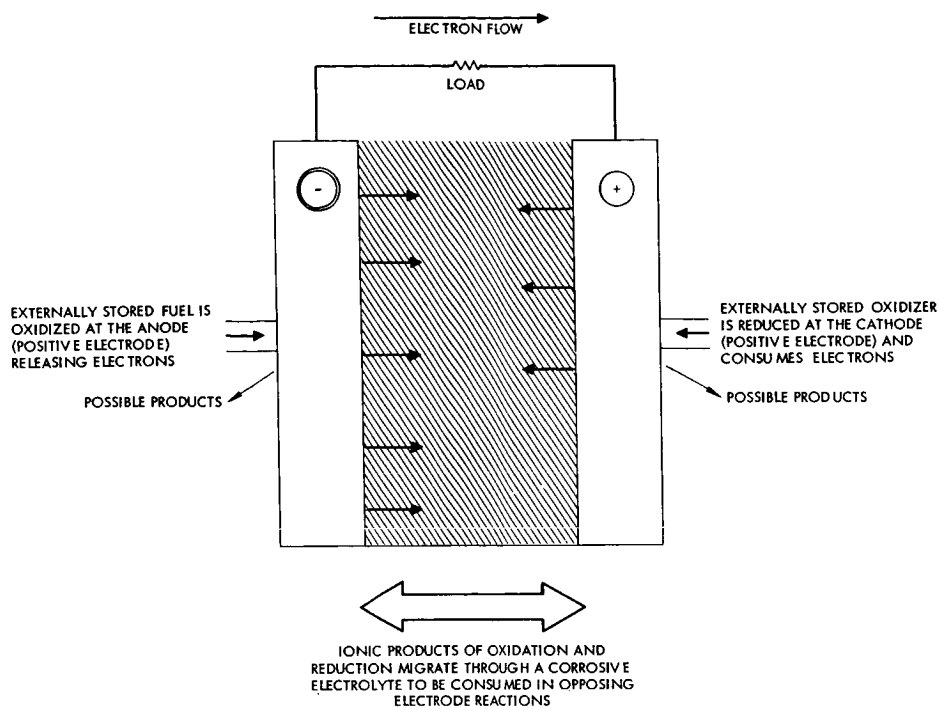


Figure A. Fuel Cell Chemical Processes

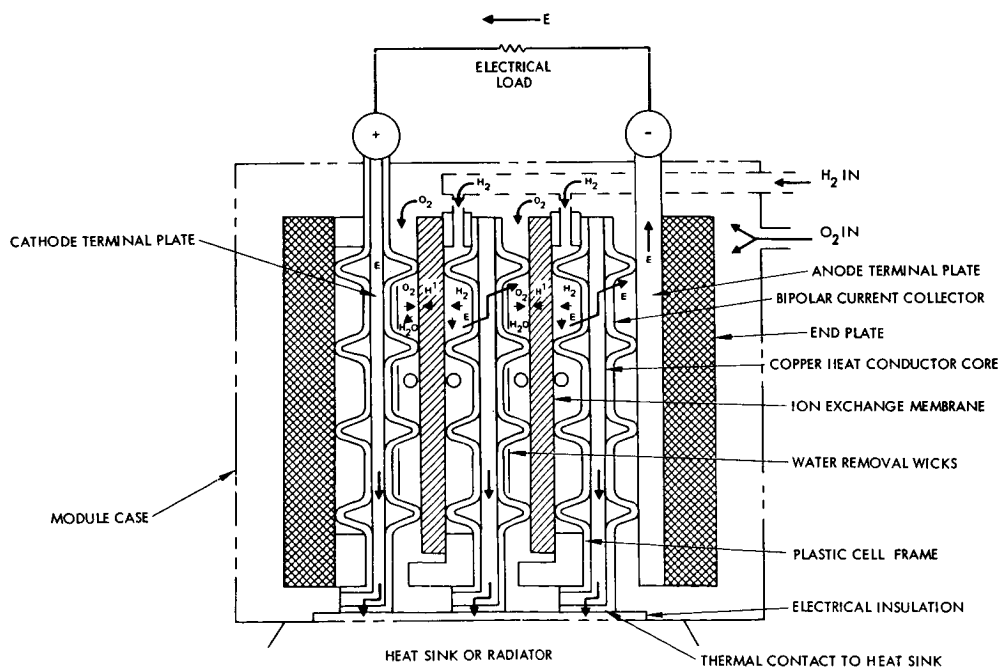


Figure B. Hydrogen-Oxygen Fuel Cell of the Ion-Exchange Membrane Type (General Electric)

BATTERIES

Primary (rechargeable for a few cycles) and secondary (rechargeable for many cycles) batteries are described.

Primary batteries (rechargeable for only a few cycles) are used for spacecraft power when the power levels are low and the mission duration is short. The silver-zinc battery is the most commonly used type, primarily because of its higher energy density (up to 100 watt-hour/lb or 220 watt-hours/kg). However, its loss of charge in storage is severe (40 percent in six months at 25°C). The pertinent characteristics of silver-zinc, silver-cadmium, and other less commonly used space primary batteries are summarized in the Table.¹

Secondary batteries (rechargeable for many cycles) are required as a part of nearly all space power systems to meet peak power demands and in solar power systems to provide power during periods of solar eclipse. The most useful secondary batteries are the nickel cadmium, the silver cadmium, and the silver zinc types. The nickel cadmium battery has the greatest cycle life but the lowest specific energy. The silver zinc type has the highest specific energy but much lower cycle life. The silver cadmium combines some of the advantages and disadvantages of both these types. The relative performance of these three systems with respect to energy storage density and cycle life are shown in Figures A and B.²

¹Szego, George C., "Space Power Systems," State of the Art, Institute for Defense Analysis, Washington, D. C., 1963.

²Mandel, Hymann, J., Recent Developments in Secondary Batteries, "Proceedings of the Intersociety Energy Conversion Engineering Conference, Los Angeles, California, September 26-28, 1966.

Primary Space Batteries

Anode Cathode Electrolyte	Zn AgO KOH	Zn Ag ₂ O ₃	Cd AgO KOH	Cd Ag ₂ O ₃	Mg AgCl KSCN- Ammonia	Mg or Ca Ca CrO ₄ Fused Salt (Thermal cell)
Separator(s)	Semi-permeable, Nylon + cellophane, Synpor Cellulosic	Semi-permeable or	Special	Special	Special	Special
Seal or Vent	Automatically activated, with pressure vents, or sealed	Automatically activated and with pressure vents	Automatically activated	Automatically activated	Automatically activated	Automatically activated
Case	Plastic or Nylon	Plastic or Nylon	Special	Special	Special	Drawn Stainless Steel
Theoretical performance at 25°C: voltage/cell watt-hr/kg (watt-hr/lb) watt-hr/cm ³ (watt-hr/in ³) Actual performance at 25°C: voltage/cell watt-hr/kg (watt-hr/lb) ¹ watt-hr/cm ³ (watt-hr/in ³) ¹	1.82 505 (230) 3.72 (61) 1.50-1.70 110-242 (50-110) 0.3-0.6 (5-10)	1.59 272 (124) 1.89 (31) 1.2-1.6 66-198 (30-90) 0.15-0.49 (2.5-8)	1.38 314 (143) 2.44 (40) 1.08-1.30 55-110 (25-50) 0.17-0.37 (2.8-6.0)	1.16 174 (79) 1.34 (22) 1.05 44-77 (20-35) 0.14-0.26 (2.3-4.2)	2.35 4.07 (185) 2.38 (39) 2.15 17.6 (8) (-65°F) 0.05 (0.9) (-65°F)	
Wet-stand life, percent loss of charge: ² at 0°C, 1 month 6 months 12 months at 25°C, 1 month 6 months 12 months at 50°C, 1 week 1 month 6 months	5 20 50 10 40 -- 15 60 --	0-5 20 50 10-15 40 -- 15-20 50-60 --	5 20 50 20 50 60 20 40 --	5 10 20 5 20 35 15 30 --	Automatically activated	Automatically activated
Cost, \$/watt-hr (at 25°C): vented sealed Available amp-hour sizes	\$0.35-\$1.00 \$0.35-\$1.50 1 to 400 A.H.	\$0.60-\$2.00 \$0.75-\$2.50 0.5 to 300 A.H.			Experimental models	\$500 to \$1000 1 x 10 ⁻³ to 0.5
1. Dependent upon cell size, discharge rate and number of cycles required. (See Figures 11-34 and 11-35.) 2. Based on percent loss of full and not rated capacity; i.e., calculated as capacity after stand divided by full capacity prior to stand.						

BATTERIES

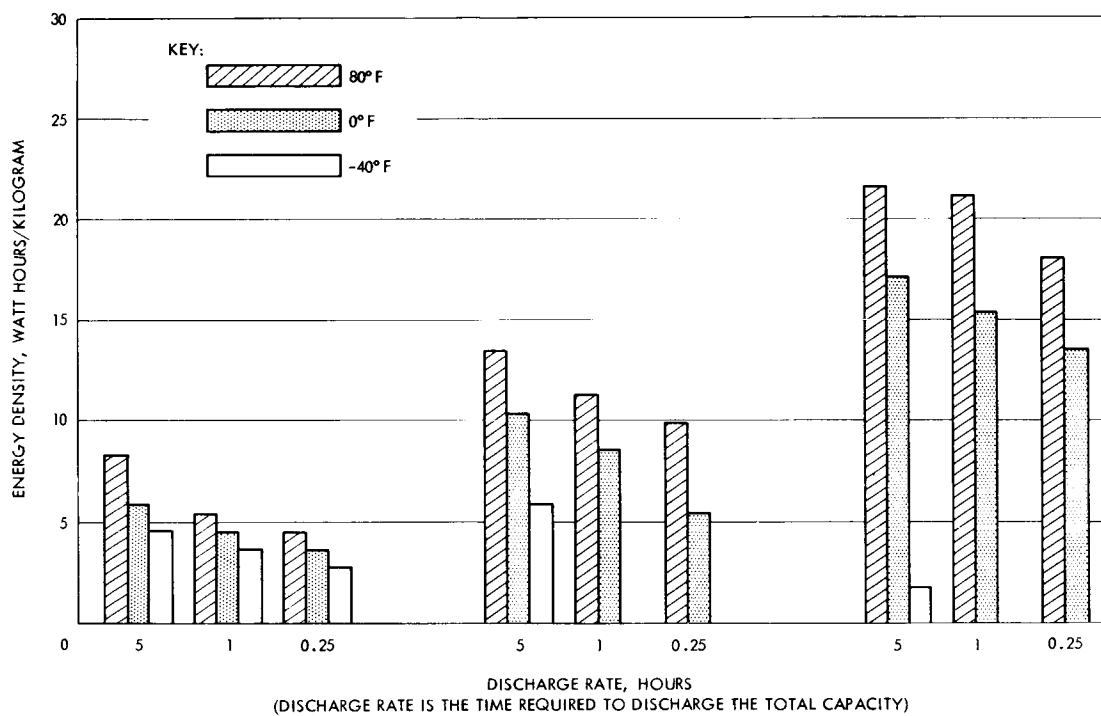


Figure A. Secondary Battery Energy Density versus Discharge Rate

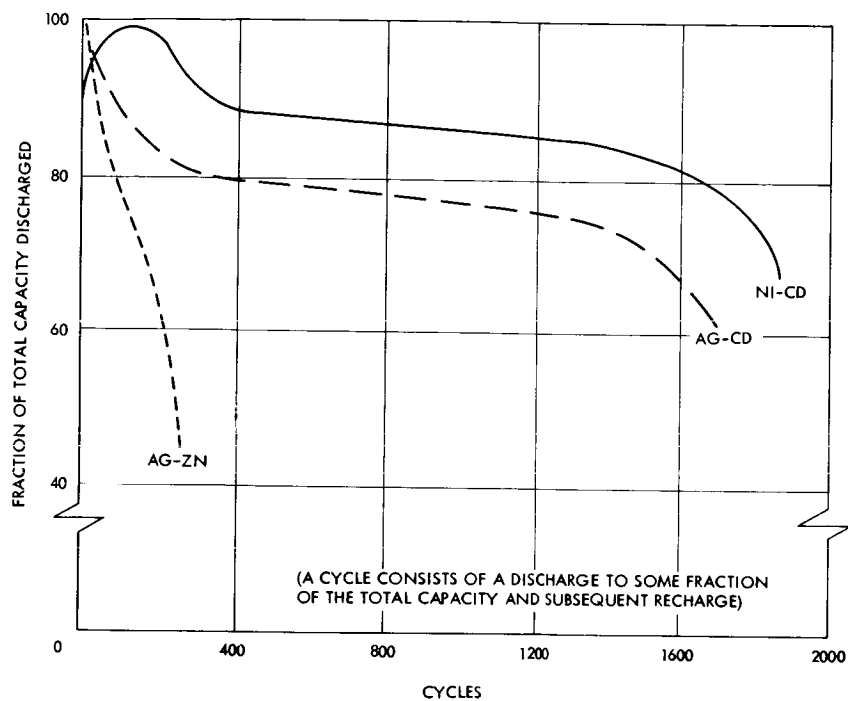


Figure B. Secondary Battery Capacity versus Cycle Life

PRIME POWER SYSTEMS

Power Summary

	Page
Cost, Volume, and Weight	516
Prime Power Burdens	522

Prime Power Systems
Power Summary

COST, VOLUME, AND WEIGHT

Cost, volume and weight relations are given as a function of prime power type, range and power level.

Power summary tables are given for different power levels requirements and different ranges (planets). There are three sets of tables. The first set, consisting of Tables A and B, gives the weight required by various prime power systems measured in kilograms (Table A) and pounds (Table B).

The second set of tables, Tables C and D, gives the volume (or area) of various prime power systems. Table C is the metric system, and Table D presents the same data in English units.

Table E is an estimate of cost for the different system types, ranges and power levels.

Power System Weight, Kilograms

Probe Near: Distance from Sun, km Expected Communication Distance from Earth, km Mission Duration, days	Mercury 57.7 x 10 ⁶ 115 to 190 x 10 ⁶ (1968) 82 to 152				Venus 108 x 10 ⁶ 40 to 190 x 10 ⁶ (1970) 72 to 210				Earth 149 x 10 ⁶ 10 ⁴ 30				Mars 227 x 10 ⁶ 90 to 310 x 10 ⁶ (1973) 118 to 260				Jupiter 775 x 10 ⁶ 600 to 890 x 10 ⁶ (1973) 450 to 1200+			
	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Fuel Cell	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric			
Power System Type	Total Power System Weight, kilograms																			
Output Power, watts																				
10	0.181	0.363		2.27	0.177	0.272		2.25	0.227	4.44		2.25	0.5		2.25	5.8		2.25		
25	0.454	0.953		5.67	0.434	0.643		5.67	0.567	11.1		5.67	1.22		5.67	14.4		5.67		
50	0.909	1.86		11.3	0.84	1.28		11.3	1.13	22.2		11.3	2.4		11.3	28.9		11.3		
100	1.81	3.72		22.7	1.77	2.56		22.7	2.27	44.4		22.7	4.81		22.7	57.8		22.7		
250	4.54	9.30		56.7	4.43	6.43		56.7	5.67	111		56.7	12.0		56.7	144		56.7		
500	9.09	186.0	311	117.0	8.4	12.8	310	117.0	11.7	222	310	117.0	24.0	310	117.0	289	310	117.0		
1000	18.1	37.2	375	227	17.7	25.6	376	227	22.7	444	376	227	48.1	375	227	577	375	227		
2000	36.3	74.4	488		35.4	51.3	488		45.4	890	487		96.2	488		1155	488			
5000	90.9	186	1080		84	127	1075		113.2	2220	1080		240	1085		2880	1085			
7500	136.0	279	1625		133	193	1625		170	3330	1625		356	1625		4330	1625			
10000	181.0	372	1665		177	256	1665		227	4440	1665		481	1665		5790	1665			

Notes:
1) Assumes no batteries.
2) Power conditioning losses and weights not included.

Notes: 1) Assumes no batteries.
2) Power conditioning losses and weights not included.

Prime Power Systems
Power Summary

COST, VOLUME, AND WEIGHT

Table B. Power System Weight, Pounds

Probe Near: Distance from Sun, km Expected Communication Distance from Earth, km Mission Duration, days	Mercury 57.7 x 10 ⁶ 115 to 190 x 10 ⁶ (1968) 82 to 152				Venus 108 x 10 ⁶ 40 to 190 x 10 ⁶ (1970) 72 to 210				Earth 149 x 10 ⁶ 104 30				Mars 227 x 10 ⁶ 90 to 310 x 10 ⁶ (1973) 118 to 260				Jupiter 775 x 10 ⁶ 600 to 890 x 10 ⁶ (1973) 450 to 1200+			
Power System Type	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Fuel Cell	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric		
	Total Power System Weight, pounds																			
Output Power, watts	0.4	0.8	5	0.39	0.6	5	0.5	9.8	5	1.1	5	12.5	2.7	12.5	5	12.8	31.9	12.5		
10	0.4	0.8	5	0.39	0.6	5	0.5	9.8	5	1.1	5	12.5	2.7	12.5	5	12.8	31.9	12.5		
25	1.0	2.1	12.5	0.98	1.42	12.5	1.25	24.5	12.5	2.7	12.5	31.9	5.3	63.8	25	63.8	25	25		
50	2.0	4.1	25	1.85	2.83	25	2.5	49	25	5.3	25	63.8	10.6	127.5	50	127.5	50	50		
100	4.0	8.2	50	3.90	5.65	50	5.0	98	50	10.6	50	127.5	21.2	255.0	100	255.0	100	100		
250	10	20.5	125	9.75	14.2	125	12.5	245	125	26.5	125	319	53	685	250	685	250	250		
500	20	41	250	18.5	28.3	250	25	490	250	53	250	637.5	106	829	500	829	500	500		
1000	40	82	500	39	56.5	500	50	980	500	106	500	1076	212	1076	1000	1076	1000	1000		
2000	80	164	1076	78	113	1076	100	1960	1076	212	1076	2385	425	3578	2000	3578	2000	2000		
5000	200	410	2385	185	283	2385	250	4900	2385	530	2385	6375	795	3578	5000	6375	5000	5000		
7500	300	615	3578	293	425	3578	375	7350	3578	795	3578	9563	1060	3670	7500	9563	7500	7500		
10000	400	820	3670	390	565	3670	500	9800	3670	1060	3670	12750	1060	3670	10000	12750	10000	10000		

Notes: 1) Assumes no batteries.
2) Power conditioning losses and weights not included.

Notes: 1) Assumes no batteries.
2) Power conditioning losses and weights not included.

Table C. Power System Volume, (or Area) cm^3 , (cm^2)

Probe Near: Distance from Sun, km Expected Communication Distance from Earth, km Mission Duration, days	Mercury 57.7 x 10 ⁶ 115 to 190 x 10 ⁶ (1968) 82 to 152			Venus 108 x 10 ⁶ 40 to 190 x 10 ⁶ (1976) 72 to 210			Earth 149 x 10 ⁶ 10 ⁴ 30			Mars 227 x 10 ⁶ 90 to 310 x 10 ⁶ (1973) 118 to 260			Jupiter 775 x 10 ⁶ 600 to 890 x 10 ⁶ (1973) 450 to 1200+					
Power System Type	Solar Photovoltaic (Area, cm ²)	Solar Thermionic (Area, cm ²)	Reactor Thermoelectric (Volume, cm ³)	Solar Photovoltaic (Area, cm ²)	Solar Thermionic (Area, cm ²)	Reactor Thermoelectric (Volume, cm ³)	Radioisotope Thermoelectric (Volume, cm ³)	Solar Photovoltaic (Area, cm ²)	Reactor Thermoelectric (Volume, cm ³)	Radioisotope Thermoelectric (Volume, cm ³)	Solar Photovoltaic (Area, cm ²)	Reactor Thermoelectric (Volume, cm ³)	Radioisotope Thermoelectric (Volume, cm ³)	Solar Photovoltaic (Area, cm ²)	Reactor Thermoelectric (Volume, cm ³)	Radioisotope Thermoelectric (Volume, cm ³)		
Total Power System Volume (Cm ³) or Area (Cm ²)																		
Output Power, watts																		
10	559			0.858 x 10 ⁴	0.465 x 10 ³		0.858 x 10 ⁴	0.65 x 10 ³		0.858 x 10 ⁴	1.39 x 10 ³		0.858 x 10 ⁴	16.85 x 10 ⁴		0.858 x 10 ⁴		
25	1400			2.28 x 10 ⁴	1.21 x 10 ³		2.28 x 10 ⁴	1.675 x 10 ³		2.28 x 10 ⁴	3.53 x 10 ³		2.28 x 10 ⁴	42.1 x 10 ³		2.28 x 10 ⁴		
50	2700	837		4.56 x 10 ⁴	2.42 x 10 ³	2.42 x 10 ³	4.56 x 10 ⁴	3.35 x 10 ³		4.56 x 10 ⁴	6.98 x 10 ³	4.56 x 10 ⁴	4.56 x 10 ⁴	84.1 x 10 ³	9.56 x 10 ⁴	9.56 x 10 ⁴		
100	5400	1580		9.13 x 10 ⁴	4.75 x 10 ³	4.82 x 10 ³	9.13 x 10 ⁴	6.63 x 10 ³		9.13 x 10 ⁴	13.95 x 10 ³	9.13 x 10 ⁴	9.13 x 10 ⁴	168.5 x 10 ³	9.13 x 10 ⁴	9.13 x 10 ⁴		
250	13.4 x 10 ³	4000		22.7 x 10 ⁴	12 x 10 ³	12.1 x 10 ³	22.7 x 10 ⁴	16.55 x 10 ³		22.7 x 10 ⁴	34.8 x 10 ³	22.7 x 10 ⁴	22.7 x 10 ⁴	421 x 10 ³	22.7 x 10 ⁴	22.7 x 10 ⁴		
500	26.8 x 10 ³	7.9 x 10 ³	45.8 x 10 ⁴	45.2 x 10 ⁴	23.9 x 10 ³	24.2 x 10 ³	45.2 x 10 ⁴	33 x 10 ³	45.8 x 10 ⁴	45.2 x 10 ⁴	69.7 x 10 ³	45.8 x 10 ⁴	45.2 x 10 ⁴	841 x 10 ³	45.2 x 10 ⁴	45.2 x 10 ⁴		
1000	53.5 x 10 ³	15.7 x 10 ³	92.5 x 10 ⁴	90.4 x 10 ⁴	47.6 x 10 ³	48.2 x 10 ³	90.4 x 10 ⁴	66 x 10 ³	92.5 x 10 ⁴	90.4 x 10 ⁴	139.5 x 10 ³	92.5 x 10 ⁴	90.4 x 10 ⁴	1685 x 10 ³	90.4 x 10 ⁴	90.4 x 10 ⁴		
2000	107.1 x 10 ³	31.4 x 10 ³	236 x 10 ⁴	236 x 10 ⁴	95.6 x 10 ³	96.8 x 10 ³	236 x 10 ⁴	132 x 10 ³	236 x 10 ⁴	236 x 10 ⁴	279 x 10 ³	236 x 10 ⁴	236 x 10 ⁴	3370 x 10 ³	236 x 10 ⁴	236 x 10 ⁴		
5000	268 x 10 ³	78.7 x 10 ³	790 x 10 ⁴	790 x 10 ⁴	23.9 x 10 ³	242 x 10 ³	790 x 10 ⁴	330 x 10 ³	790 x 10 ⁴	790 x 10 ⁴	699 x 10 ³	790 x 10 ⁴	790 x 10 ⁴	8420 x 10 ³	790 x 10 ⁴	790 x 10 ⁴		
7500	401 x 10 ³	118 x 10 ³	1450 x 10 ⁴	1450 x 10 ⁴	45.2 x 10 ³	363 x 10 ³	1450 x 10 ⁴	496 x 10 ³	1450 x 10 ⁴	1450 x 10 ⁴	1045 x 10 ³	1450 x 10 ⁴	1450 x 10 ⁴	12600 x 10 ³	1450 x 10 ⁴	1450 x 10 ⁴		
10000	535 x 10 ³	157 x 10 ³	2340 x 10 ⁴	2340 x 10 ⁴	477 x 10 ³	482 x 10 ³	2340 x 10 ⁴	660 x 10 ³	2340 x 10 ⁴	2340 x 10 ⁴	1395 x 10 ³	2340 x 10 ⁴	2340 x 10 ⁴	16850 x 10 ³	2340 x 10 ⁴	2340 x 10 ⁴		

Notes: 1) Assumes no batteries.
2) Power conditioning losses and volumes not included.

Notes: 1) Assumes no batteries.

2) Power conditioning losses and volumes not included.

Prime Power Systems
Power Summary

COST, VOLUME, AND WEIGHT

Table D. Power System Volume (or Area), feet³ (or feet²)

Probe Near: Distance from Sun, km Expected Communication Distance from Earth, km Mission Duration, days	Mercury 57.7 x 10 ⁶ 115 to 190 x 10 ⁶ (1968) 82 to 152				Venus 108 x 10 ⁶ 40 to 190 x 10 ⁶ (1970) 72 to 210				Earth 149 x 10 ⁶ 10 ⁴ 30				Mars 227 x 10 ⁶ 90 to 310 x 10 ⁶ (1971) 118 to 260				Jupiter 775 x 10 ⁶ 600 to 890 x 10 ⁶ (1973) 450 to 1200+			
	Solar Photovoltaic (Area, ft ²)	Solar Thermionic (Area, ft ²)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)	Solar Photovoltaic (Area, ft ²)	Solar Thermionic (Area, ft ²)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)	Fuel Cell (Volume, ft ³)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)	Solar Photovoltaic (Area, ft ²)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)	Solar Photovoltaic (Area, ft ²)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)	Solar Photovoltaic (Area, ft ²)	Reactor (Volume, ft ³)	Radioisotope (Volume, ft ³)
Total Power System Volume (ft ³) or Area (ft ²)																				
Output Power, watts	0.6	0.6	0.3	0.3	0.5	0.5	0.3	0.3	0.7	0.3	0.3	1.5	0.3	0.3	18.1	0.3	0.3	18.1	0.3	0.3
10	1.5	1.5	0.8	0.8	1.3	1.3	0.8	0.8	1.8	0.8	0.8	3.8	0.8	0.8	45.3	0.8	0.8	45.3	0.8	0.8
25	2.9	2.9	1.6	1.6	2.6	2.6	1.6	1.6	3.6	1.6	1.6	7.5	1.6	1.6	90.5	1.6	1.6	90.5	1.6	1.6
50	5.8	5.8	3.2	3.2	5.1	5.1	3.2	3.2	7.1	3.2	3.2	15.0	3.2	3.2	181	3.2	3.2	181	3.2	3.2
100	14.4	14.4	8.0	8.0	12.9	13	8.0	8.0	17.8	8.0	8.0	37.5	8.0	8.0	453	8.0	8.0	453	8.0	8.0
250	28.8	28.8	16.1	15.9	25.7	26	16.1	15.9	35.5	16.1	15.9	75	16.1	15.9	905	16.1	15.9	905	16.1	15.9
500	57.6	57.6	32.6	31.8	51.3	52	32.6	31.8	71	32.6	31.8	150	32.6	31.8	1810	32.6	31.8	1810	32.6	31.8
1000	115.2	115.2	83.3	83.3	102.6	104	83.3	83.3	142	83.3	83.3	300	83.3	83.3	3620	83.3	83.3	3620	83.3	83.3
2000	288	288	278	278	257	260	278	278	355	278	278	750	278	278	9050	278	278	9050	278	278
5000	432	432	510	510	486	390	510	510	533	510	510	1125	510	510	13575	510	510	13575	510	510
7500	576	576	824	824	513	520	824	824	710	824	824	1500	824	824	18100	824	824	18100	824	824
10000																				

Notes: 1) Assumes no batteries.
2) Power conditioning losses and volumes not included.

Table E. Power System Cost, Dollars

Probe Near: Distance from Sun, km Expected Communication Distance from Earth, km Mission Duration, days	Mercury 57.7 x 10 ⁶ 115 to 190 x 10 ⁶ (1968) 82 to 152				Venus 108 x 10 ⁶ 40 to 190 x 10 ⁶ (1970) 72 to 210				Earth 149 x 10 ⁶ 10 ⁴ 30				Mars 227 x 10 ⁶ 90 to 310 x 10 ⁶ (1973) 118 to 260				Jupiter 775 x 10 ⁶ 600 to 890 x 10 ⁶ (1973) 450 to 1200+			
Power System Type	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Solar Thermionic	Reactor Thermoelectric	Radioisotope Thermoelectric	Fuel Cell	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric	Solar Photovoltaic	Reactor Thermoelectric	Radioisotope Thermoelectric		
Output Power, watts	Total Power System Cost, dollars																			
10	4.30 x 10 ²	COSTS NOT AVAILABLE				3.00 x 10 ⁴	3.83 x 10 ²			3.00 x 10 ⁴	5.30 x 10 ²			1.12 x 10 ³		3.00 x 10 ⁴	1.35 x 10 ⁴	3.0 x 10 ⁴		
25	1.08 x 10 ³	COSTS NOT AVAILABLE				7.50 x 10 ⁴	9.58 x 10 ²			7.50 x 10 ⁴	1.33 x 10 ³			2.80 x 10 ³		7.50 x 10 ⁴	3.4 x 10 ⁴	7.5 x 10 ⁴		
50	2.15 x 10 ³	COSTS NOT AVAILABLE				1.50 x 10 ⁵	1.92 x 10 ³			1.50 x 10 ⁵	2.65 x 10 ³			5.60 x 10 ³		1.50 x 10 ⁵	6.75 x 10 ⁴	1.5 x 10 ⁵		
100	4.30 x 10 ³	COSTS NOT AVAILABLE				3.00 x 10 ⁵	3.83 x 10 ³			3.00 x 10 ⁵	5.30 x 10 ³			1.12 x 10 ⁴		3.00 x 10 ⁵	1.35 x 10 ⁵	3.0 x 10 ⁵		
250	1.08 x 10 ⁴	COSTS NOT AVAILABLE				7.50 x 10 ⁵	9.58 x 10 ³			7.50 x 10 ⁵	1.33 x 10 ⁴			2.80 x 10 ⁴		7.50 x 10 ⁵	3.4 x 10 ⁵	7.5 x 10 ⁵		
500	2.15 x 10 ⁴	COSTS NOT AVAILABLE				1.50 x 10 ⁶	1.92 x 10 ⁴			1.50 x 10 ⁶	2.65 x 10 ⁴			5.60 x 10 ⁴		1.50 x 10 ⁶	6.75 x 10 ⁵	1.5 x 10 ⁶		
1000	4.30 x 10 ⁴	COSTS NOT AVAILABLE				3.00 x 10 ⁶	3.83 x 10 ⁴			3.00 x 10 ⁶	5.30 x 10 ⁴			1.12 x 10 ⁵		3.00 x 10 ⁶	1.35 x 10 ⁶	3.0 x 10 ⁶		
2000	8.6 x 10 ⁴	COSTS NOT AVAILABLE				1.45 x 10 ⁶	7.66 x 10 ⁴			1.45 x 10 ⁶	1.06 x 10 ⁵			2.24 x 10 ⁵		3.00 x 10 ⁶	2.70 x 10 ⁶	3.0 x 10 ⁶		
5000	2.15 x 10 ⁴	COSTS NOT AVAILABLE				2.00 x 10 ⁶	1.92 x 10 ⁵			2.00 x 10 ⁶	2.65 x 10 ⁵			5.60 x 10 ⁵		3.00 x 10 ⁶	6.75 x 10 ⁶	3.0 x 10 ⁶		
7500	3.23 x 10 ⁵	COSTS NOT AVAILABLE				2.55 x 10 ⁶	2.78 x 10 ⁵			2.55 x 10 ⁶	3.97 x 10 ⁵			8.40 x 10 ⁵		3.00 x 10 ⁶	1.01 x 10 ⁷	3.0 x 10 ⁶		
10000	4.30 x 10 ⁵	COSTS NOT AVAILABLE				3.20 x 10 ⁶	3.83 x 10 ⁵			3.20 x 10 ⁶	5.30 x 10 ⁵			1.12 x 10 ⁶		3.00 x 10 ⁶	6.75 x 10 ⁷	3.20 x 10 ⁶		

Notes: 1) Assumes no batteries.
2) Power conditioning losses and costs not included.

Notes: 1) Assumes no batteries.
2) Power conditioning losses and costs not included.

PRIME POWER BURDENS

Power Burdens used for the communications system Methodology are tabulated from the data previously given.

Prime power burdens relate the prime power to weight or cost. They are used in the communications system methodology to determine the lightest or least expensive system. In the communication system modeling, the power supply weight is described by:

$$W_{ST} = K_{W_{ST}} P_{ST} + W_{KE}$$

where:

$K_{W_{ST}}$ = constant relating transmitter power supply weight to power requirement

P_{ST} = transmitter power supply power requirement

W_{KE} = transmitter power supply weight independent of transmitter power requirement

and the fabrication cost is given by:

$$C_{FT} = K_{ST} P_{ST} + C_{KE}$$

where

K_{ST} = constant relating transmitter power supply fabrication cost to power requirement.

P_{ST} = transmitter power supply requirement

C_{KE} = transmitter power supply fabrication cost independent of transmitter power requirement

These constants are summarized in the Table for the power system types discussed in "Prime Power Systems".

Power System Weight and Cost Burden Constants

System	$K_{W_{S_T}}$ kg/watt (lb/watt)	W_{K_E} kg (lb)	K_{S_T} \$/watt	C_{K_E} \$
Solar				
Photo-voltaic	0.0454 to 0.0238 (0.1 to 0.05) (1 AU, 28°C) ¹	Negligible	\$53 ²	Negligible
Thermo-electric	0.0906 (0.2) (1 AU) 0.0498 (0.11) (0.3 AU)	Negligible	*	Negligible
Thermi- onic	0.0454 (0.1) (1 AU) 0.0272 (0.06) (0.3 AU)	*	*	*
Dynamic	*	*	*	*
Reactor				
Thermo-electric	0.0145 (0.32) (20 KW) 0.621 (1.37) (0.5 KW)	181.5 (400)	\$400 (at 5 KW)	1.2×10^6
Thermi- onic	0.0454 to 0.0227 (0.1 to 0.05)	*	*	*
Dynamic	0.0771 to 0.0113 (0.16 to 0.25)	*	*	*
Radioisotope				
Thermo-electric	0.225 (0.5)	Negligible	\$3,000 (at 1 KW)	Negligible
Thermi- onic	0.0454 (0.1) (above 1 KW)	*	*	*
Dynamic	0.15 (0.33) (Brayton Cycle)	*	*	*
Fuel Cell	0.00386 (0.085) (except fuel)		\$200	
<p>*DEVELOPMENT STAGE: costs or weights not accurately known</p> <p>1. 1 AU = 1.496×10^8 km</p> <p>2. Assumes 7 cm x 10 cm, 12 percent efficient cells available at \$10/cell — Helioteks estimate of capability in five years.</p>				

PART 7 – HEAT EJECTION SYSTEMS

Section	Page
Heat Ejection Elements	532
Heat Ejection Elements – Radiators	540
Weight and Cost Burdens	550

GENERAL HEAT EJECTION CONSIDERATIONS

Heat ejection system parameters are based largely on the transmitter power, efficiency, operating temperature and spacecraft thermal environment.

Thermal control system requirements imposed by the communication system are presented in this section. The communication system characteristics which determine the thermal control requirements are largely the output power, efficiency, and operating temperature of the transmitting source. In addition, the thermal control system burden is also influenced by the mission thermal environment and the spacecraft configuration.

Since operation of the transmitter will not be continuous throughout the mission, the steady state thermal control of the spacecraft is dependent upon the rate of heat or power ejected. Thus, essentially all the heat produced by intermittent operation of the transmitter must be ejected from the spacecraft. The power to be ejected, W , is then

$$W = P_T \frac{1 - k_e}{k_e}$$

where

P_T = transmitter output power

k_e = transmitter efficiency

This approximation of ejected heat burden is totally conservative since it is assumed that none of the heat ejected by the transmitter is put to effective use in thermal control.

The system parameters of most significance in determining the radiator burden for a given transmitted power are the operating temperature of the transmitting source and its efficiency. Operating temperatures range from less than 40° C for some laser sources to 200 to 250° C for TWT microwave sources. The efficiencies vary even more widely. If the transmitter power and efficiency have been specified and if its operating temperature is known, the associated radiator weight, area, and cost are essentially determined.

TRANSMITTER SOURCE CHARACTERISTICS

Microwave and laser transmitting sources vary markedly in their sensitivity to operating temperature, their efficiency and their allowable operating temperature.

To gain an appreciation for the range of parameters into which a thermal head radiator must match, the power amplifiers for microwave links and laser links are considered below.

Microwave Sources. A microwave source of prime interest for long space missions is the traveling-wave amplifier tube (TWT). In a TWT the greatest heat is generated at the collector surface. These parts may reach temperatures as high as 200 to 250°C in present long life tubes. For lower power levels (less than 100 to 200 watts output) it is customary to conductively cool the collector by thermally connecting it to a heat sink. The heat sink in turn conducts the heat to an external radiating surface. Higher power tubes are customarily cooled by flowing a coolant fluid through integral passages in the collector and other critical parts. The upper limit in outlet fluid temperature is imposed by the collector temperature limitations although it is typically somewhat lower as a result of temperature drops in other parts of the internal tube coolant circuit.

For power levels beyond 1 kw, TWTs are generally built in a different configuration from that used at lower powers. The high power configuration uses a cavity resonator which requires a solenoid to provide the necessary magnetic field. The solenoid must be cooled as well in this case. With modern high temperature insulating materials the solenoid operating temperatures may be comparable with the collector temperature.

Traveling wave tubes operate at efficiencies as high as 30 percent including power supply losses. For high power tubes, the solenoid cooling requirement is reflected in the efficiency.

Optical Transmitting Sources. The two optical sources of primary interest are the CO₂ laser (10.6 μ wavelength) and the Argon laser (0.5145 μ wavelength). They differ drastically in operating efficiency and operating temperature requirements.

The CO₂ laser operates at efficiencies as high as 15 percent. To achieve this high efficiency, the gas mixture must be kept at temperatures of 20°C or less. Efficiency drops off rapidly at higher temperatures. For a typical low power device, output was reduced from 1.5 watts at 20°C to 0.7 watt at 60°C and 0.25 watt at 100°C. To maintain the required temperatures, most laboratory versions use water as a coolant, flowing it between the walls of the discharge tube and an outer concentric jacket. For the higher power levels envisioned for space transmitting sources, an active fluid cooling system is virtually a necessity.

The Argon laser is characterized by efficiencies of the order of 0.1 percent or less. The very large fraction of input power which must be rejected as a result of this inefficiency demands liquid cooling for all power levels under consideration. Efficiency is not a critical function of the operating temperature as is the case with the CO₂ laser.

The upper limit in operating temperature is imposed by the limits for safe operation of the solenoid which surrounds the discharge tube and provides the pumping magnetic field. The pumping solenoid generates a large amount of heat and both it and the discharge tube must be cooled. The most effective way to achieve this is to flow the coolant fluid through the annular passage between them. Maximum operating temperatures imposed by solenoid temperature as limited by modern high temperature insulating materials may be as high as 100 to 150°C.

Heat Ejection Systems

SUMMARY

Heat ejection design is based upon ejecting heat into space.

Spacecraft heat ejection systems are able to eject heat into space by radiation. The amount of heat that is ejected depends upon several factors. These include the temperature of the surface ejecting heat, T ; the surface emissivity, ϵ ; a constant, which is the Stefan-Boltzmann's constant, $\sigma = 5.7 \times 10^{-12}$ watts/cm² °K; the temperature of the sink, T_s , (0°K for free space, somewhat higher when near a planet.); and heat loading, which is predominantly from the sun. This loading depends upon the solar illumination, H ; the absorptivity of the surface, α_s ; and the angle at which the sun's rays strike the surface, θ . Expressed in a single equation, the heat ejected in watts/cm² is:

$$Q = \epsilon \sigma (T^4 - T_s^4) - \alpha_s H \cos \theta$$

Clearly there are many variables in this equation. Ranges for these are given with sample calculations in the material which follows.

HEAT EJECTION SYSTEMS

Heat Ejection Elements

	Page
Types of Heat Ejection Systems	532
Heat Pipe	534
Useful Heat Pipe Properties	536

TYPES OF HEAT EJECTION SYSTEMS

An active heat exchanger is one in which the heat is conveyed from the heat source to the radiating element by a moving coolant or moving mechanical parts, while a passive heat exchanger has no moving parts.

Heat ejection systems may be classified as active or passive (see the figure). In the most general sense, an active system is one which embodies moving parts (e. g., a coolant fluid or a thermal switch) while a passive system does not. In typical active systems heat is conveyed to the radiating surface by first transferring it to a fluid medium which is then physically transported to the radiator where its heat is ejected. In a passive system heat is conveyed to a radiating surface and dissipated from it by purely static processes.

Passive Heat Ejection

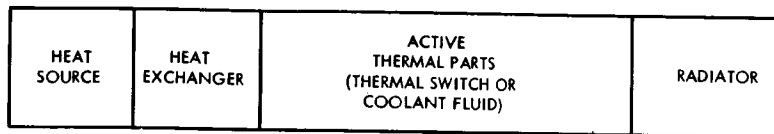
Passive heat ejection systems are preferable when they can meet the requirements because of their extreme simplicity and concomitant lighter weight, lower cost, and higher reliability. They consist merely of a conducting path between the heat source and an external radiating surface, often a part of the spacecraft structure, having a highly emissive surface coating with low solar absorptivity. The limitation on their utility is almost always excessive temperature gradients in the conducting path as a result of thermal resistance.

Passive and active heat ejection systems differ in the method of conducting the heat from one point to another but the same considerations apply to the actual radiators of heat. The heat radiators are considered in some detail in later topics of this section.

Active Heat Ejection

Active heat ejection systems generally consist of a heat exchanger to transfer heat from the transmitting source to the cooling fluid, the necessary plumbing to convey the fluid to the radiator, and the radiator itself. Of these, the radiator proper is the major contributor to the thermal control system cost, weight, and area burdens. The heat exchanger at the transmitting source is an integral part of the source and is characteristic of it. The burdens associated with transferring heat from the source to the cooling system are thus included in the transmitting source burdens and cannot meaningfully be divorced from them. The remaining system components — plumbing, pumps, controls and the coolant itself — are of less significance than the radiator with respect to cost, weight, and volume. They are, in any event, so peculiar to a specific vehicle and communication system configuration as to preclude meaningful treatment here.

Both condensing and non-condensing active heat ejection systems will be discussed in subsequent topics. Condensing (two phase) systems are most applicable to dynamic power systems and so are included as a matter of general interest. Non-condensing (single phase) systems appear more applicable to cooling transmitting sources since boiling of the coolant fluid in condensing systems introduces vapor pockets and would lead to local hot spots in critical areas.



(a) ACTIVE SYSTEM



(b) PASSIVE SYSTEM

Heat Ejection System Constituant Parts

HEAT PIPE¹

The unique feature of a heat pump is the use of capillary action to "pump" a fluid.

The heat pipe is essentially a closed, evacuated chamber whose inside walls are lined with a capillary structure, or wick, that is saturated with a volatile fluid (see the figure). The operation of a heat pipe combines two familiar principles of physics: vapor heat transfer and capillary action. Vapor heat transfer is responsible for transporting the heat energy from the evaporator section at one end of the pipe to the condenser section at the other end. What distinguishes the heat pipe is that the heat pipe capillary action is responsible for returning the condensed working fluid back to the evaporator section to complete the cycle.

The function of the working fluid within the heat pipe is to absorb the heat energy received at the evaporator section (by the latent heat at evaporation), transport it through the pipe and release this energy at the condenser end.

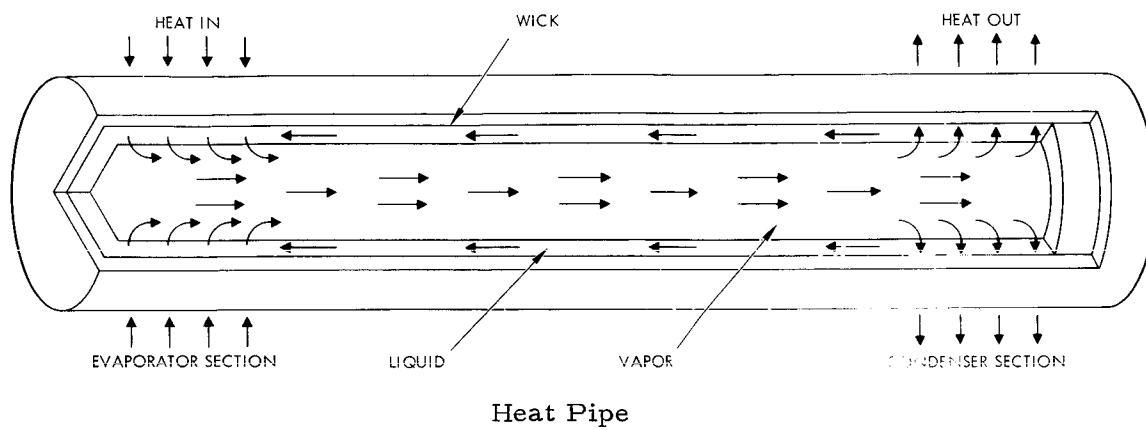
As the working fluid vaporizes, the pressure at the evaporator end of the pipe increases. The vapor pressure sets up a pressure difference between the ends of the pipe, and this pressure difference causes the vapor, and thus the heat energy, to move toward the condenser section. When the vapor arrives at the condenser section, it encounters a temperature lower than that of the evaporator. As a consequence the vapor turns back to a liquid and thereby releases the thermal energy stored in its heat of vaporization.

As the fluid condenses, the vapor pressure created by the molecules decreases, so that the necessary pressure difference for continual vapor heat flow is maintained.

It is important to note that the vaporized fluid stores heat energy at the temperature at which the vapor was created, and that it will retain the energy at that temperature until it meets a colder surface. The result is that the temperature along the entire length of the heat pipe tends to remain constant. It is this tendency to resist any difference in temperature within the heat pipe that is responsible for the device's high thermal conductance.

One of the requirements of any self-contained vapor heat-transfer system is a means of returning the condensed liquid to the evaporator to replenish the supply. In older systems this return was accomplished either by gravity or by a pump. The heat pipe, on the other hand, can operate against gravity and without a second external energy source. This is accomplished by capillary action within the wick that connects the condenser to the evaporator. The capillary action "pumps" the cooled fluid back to the heat source.

¹G. Y. Eastman "The Heat Pipe", Scientific American, May 1968, pp 38-46.



USEFUL HEAT PIPE PROPERTIES

Five useful properties of heat pumps, as an example of an active heat radiator element, are described.

Heat Pipe Properties

There are five properties of the heat pipe that deserve special mention because they serve to define the areas in which practical applications are to be found for the device. These are listed briefly below.

First, active devices that operate on the principle of active vapor heat transfer can have several thousand times the heat-transfer capacity of the best passive metallic conductors, such as silver and copper.

A second property of the heat pipe is called "temperature flattening." There are many heat-transfer applications in which a uniform temperature over a large surface area is required. Without the heat pipe special care must be taken to ensure a uniform temperature of the heat source. A heat pipe, however, can be coupled to a nonuniform heat-source to produce a uniform temperature at the output, regardless of the point-to-point variations of the heat source.

Third, the evaporation and condensation functions of a heat pipe are essentially independent operations connected only by the streams of vapor and liquid in the pipe. The patterns and area of evaporation and condensation are independent. Thus the process occurring at one end of the pipe can take place uniformly or nonuniformly, over a large or a small surface area, without significantly influencing what is going on at the other end.

A fourth property of the heat pipe is that it makes it possible to separate the heat source from the heat sink.

Fifth, the heat pipe can also be operated so that the thermal power and/or the temperature at which the power is delivered to the intended heat sink can be held constant in spite of large variations in the power input to the heat pipe.

Heat Pipe Implementation

Heat pipes have been made to operate at various temperatures spanning the range from below freezing to over 3,600 degrees F. The power transferred ranges from a few watts to more than 17,000 watts. Working fluids have included methanol, acetone, water, fluoridated hydrocarbons, mercury, indium, cesium, potassium, sodium, lithium, lead, bismuth and a range of inorganic salts. The containment vessels have been made of glass, ceramic, copper, stainless steel, nickel, tungsten, molybdenum, tantalum and various alloys. The wicks or capillary structures have included sintered porous matrixes, woven mesh, fiber glass, longitudinal slots and combinations of these structures in various geometries. In physical size heat pipes have ranged from a quarter of an inch to more than six inches in diameter and up to several feet in length. Moreover, heat pipes can be designed in almost any configuration.

An operating life in excess of 10,000 hours without failure or detectable degradation has been achieved with a range of fluid-container systems. The longest of these tests has currently passed 16,000 hours at 1,100 degrees F., using potassium as the working fluid in a nickel containment vessel.

HEAT EJECTION SYSTEMS

Heat Ejection Elements – Radiators

	Page
Radiant Heat Transfer from a Flat Surface	540
Radiator Fin Effectiveness	542
Radiator Area Requirements – Condensing Systems	544
Radiator Area Requirements – Non-Condensing Systems	546

RADIANT HEAT TRANSFER FROM A FLAT SURFACE

The equations describing radiant heat transfer from a flat surface are given and pertinent curves are drawn.

For typical spacecraft fin and tube radiators the controlling thermal resistance is conduction and radiation in the radiator fin. Therefore the fin design is a chief concern for the preliminary designer, while heat transfer from the working fluid to the fin is a second-order problem.

Radiant heat transfer from a flat surface at temperature, T , to a sink at absolute zero is described by the Stefan-Boltzmann equation:

$$Q = \epsilon \sigma T^4 \quad (1)$$

where

Q = radiative power (watts/cm²)

ϵ = surface emissivity

σ = Stefan-Boltzmann constant = 5.7×10^{-12} watts/cm² °K

T = radiating surface temperature (°K)

For a non-zero sink temperature, this expression becomes

$$Q = \sigma \epsilon (T^4 - T_s^4) \quad (2)$$

where

T_s = sink temperature (°K)

If solar illumination is incident on the radiator, it must also be ejected, reducing the effective radiative heat flux (i. e., dissipation of heat produced by an on-board source) to

$$Q = \epsilon \sigma (T^4 - T_s^4) - \alpha_s H \cos \theta \quad (3)$$

where

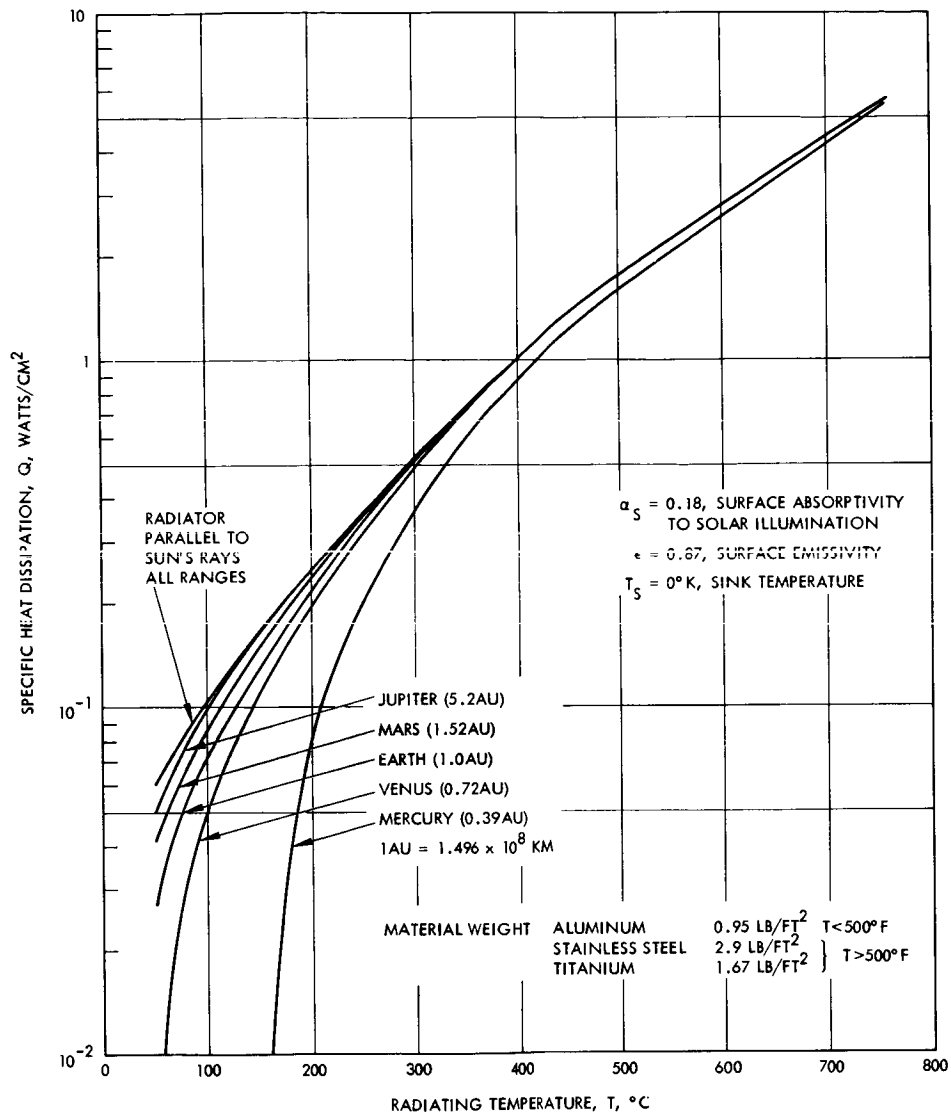
α_s = surface absorptivity to solar illumination

H = solar illumination intensity (0.138 watts/cm² at 1 A. U.)

θ = angle between the incident solar illumination and the normal to the radiator surface

ϵ = surface emissivity

Equation (3) is plotted in the figure indicating the dependance of the heat radiation upon solar radiation. As may be seen at ranges greater than 1 to 1.5 AU from the sun, the solar effect does not cause a major variation in the radiator design.



Specific Heat Dissipation Capacity Versus
 Temperature of Radiators Formed to
 Direct Sunlight at Various
 Solar Distances

RADIATOR FIN EFFECTIVENESS

The fin effectiveness compares the actual radiative effectiveness of a radiator with the maximum effectiveness.

A quantity termed the fin effectiveness is introduced to assist in the evaluation of the performance of a finned radiator. It is defined as the ratio of the heat ejected by the fin to that which would be ejected if the entire fin were maintained at the base temperature. The fin effectiveness, η , may be included in the equation defining the specific heat radiation Q , as follows.

$$Q = \epsilon \sigma \eta (T^4 - T_s^4) - \alpha_a H \cos \theta$$

where

- α_s = surface absorptivity to solar illumination
- H = solar illumination intensity (0.138 watts/cm² at 1 AU)
- θ = angle between the incident solar illumination and the normal to the radiator surface
- ϵ = surface emissivity
- σ = Stefan-Boltzmann constant = 5.7×10^{-12} watts/cm² °K
- T = radiating surface temperature (°K)
- T_s = sink temperature (°K)

Expressed mathematically:

$$\eta = \frac{\int_0^{\frac{B}{2}} T_x^4 dx}{\frac{B}{2} T^4} \quad (4)$$

where

- η = fin effectiveness
- B = tube spacing
- T_x = temperature at a point on the fin
- x = distance along fin
- T = fin base temperature

This equation was derived by Coombs¹ et al., and was solved numerically on an IBM-704 computer. The results are given in the figure as a function of the dimensionless radiation modulus M_r defined as:

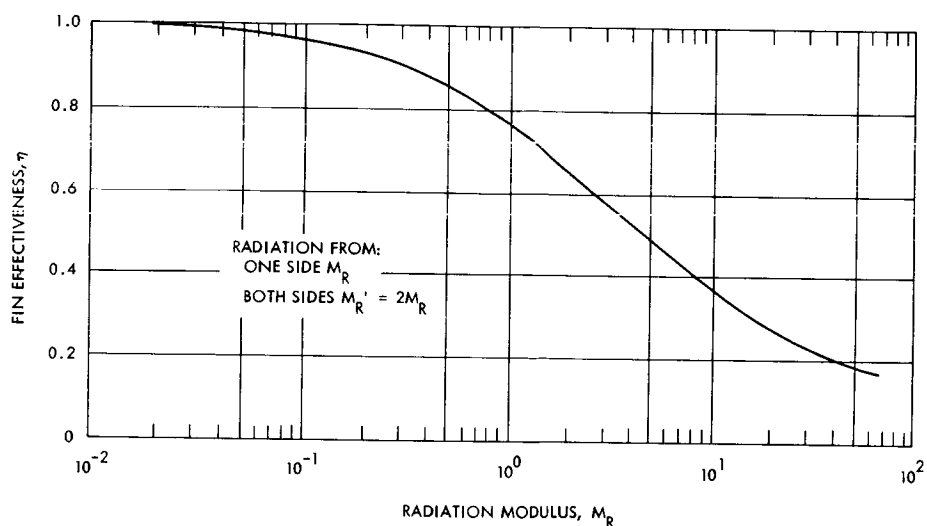
$$M_r = \frac{B^2 \epsilon \sigma T^3}{kt} \quad (5)$$

where

k = conductivity of fin material

t = fin thickness

By using the curve in the figure the fin effectiveness may be evaluated for a given material, geometry, and base temperature, T .



Fin Effectiveness versus Radiation Modulus

¹M. G. Coombs, R. A. Stone, and T. Kapus, "The SNAP 2 Radiative Condenser Analysis," NAA-SR-5317, July 1960.

RADIATOR AREA REQUIREMENTS – CONDENSING SYSTEMS

The specific heat radiated from a surface depends critically on the radiator orientation and emissivity especially at lower radiating temperatures.

For condensing radiators, the tube temperature remains constant until the fluid is completely condensed, as long as the static pressure drop is kept small. This follows since the condensate and condensing vapor are always in thermal equilibrium. If this condition is met, the area requirements for the condensing portion of the radiator can be obtained from the total power which must be dissipated and the radiative power per unit area, Q :

$$Q = \epsilon \sigma \eta (T^4 - T_s^4) - \alpha_s H \cos \theta \quad (1)$$

where

Q = radiative power (watts/cm²)

ϵ = surface emissivity

σ = Stefan-Boltzmann constant = 5.7×10^{-12} watts/cm² °K

T = radiating surface temperature (°K)

T_s = sink temperature (°K)

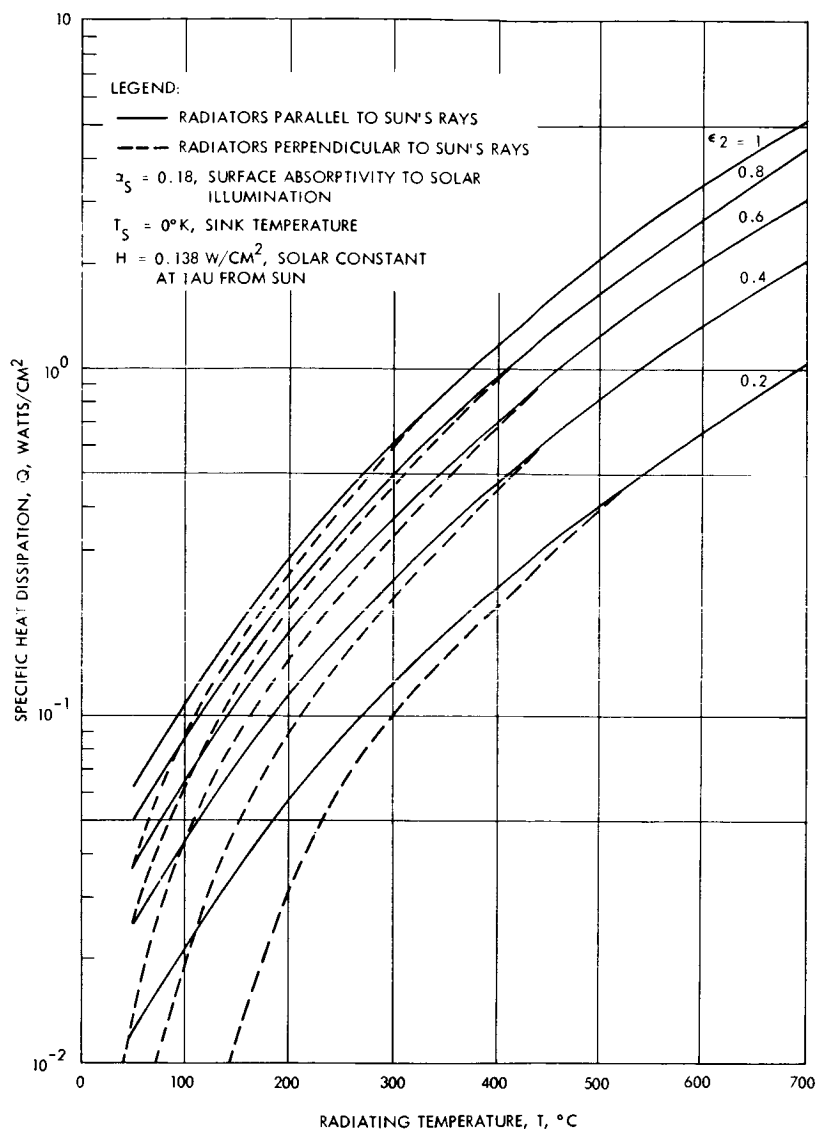
η = fin effectiveness

α_s = surface absorptivity to solar illumination

H = solar illumination intensity (0.138 watts/cm² at 1 AU)

θ = angle between the incident solar illumination and the normal to the radiator surface

Equation 1 has been plotted in the figure as a function of radiating temperature with the product of surface emissivity, ϵ , and fin effectiveness, η , as a parameter. Two sets of curves are shown in the figure. One set corresponds to the specific heat dissipation when the radiating surface is parallel to the sun's rays (best case) and the second set corresponds to the sun's rays being normal to the radiating surface (worst case). As may be seen from the figure the orientation of the radiating area becomes critical at lower radiating temperatures and lower effective emissivities.



Specific Heat Dissipation Capacity Versus Radiator Temperature

RADIATOR AREA REQUIREMENTS - NON CONDENSING SYSTEMS

A non-condensing radiator has a temperature gradient along the radiator length. This causes this type of radiator to be less efficient than a condensing radiator.

In non-condensing systems the radiant heat rejection is accompanied by a sensible heat loss of the fluid. The temperature decrease of the fluid results in temperature gradients both perpendicular and parallel to the direction of fluid flow. This complicates the analysis, but by combining the model of the condensing (constant temperature) fin with that of a radiator which experiences a coolant temperature drop, an expression can be derived to give the area requirements for the tube-fin configuration.¹ The result is given by:

$$Q = \eta \sigma \epsilon \left[\frac{3 T_{in}^3 T_{out}^3}{T_{in}^2 + T_{in} T_{out} + T_{out}^2} \right] \quad (1)$$

where

Q = radiative power (watts/cm²)

η = fin effectiveness

σ = Stefan-Boltzmann constant = 5.7×10^{-12} watts/cm²°K

ϵ = surface emissivity

T_{in} = fluid temperature into radiator (°K)

T_{out} = fluid temperature out of radiator (°K)

The radiative power, Q , is plotted in the figure as a function of temperature drop of the non-condensing fluid across the radiator length. The input temperature to the radiator is used as a parameter. The effectiveness is shown in the figure to decrease with increasing temperature drop along the radiator length.

Equation 1 has been written not considering the heat loading of the sun or the sink temperature. If these factors are included the new equation is as follows:

$$Q = \eta \sigma \epsilon (T_{eff}^4 - T_s^4) - \alpha s H \cos \theta$$

where

$$T_{eff} = \left[\frac{3 T_{in}^3 T_{out}^3}{T_{in}^2 + T_{in} T_{out} + T_{out}^2} \right]^{1/4}$$

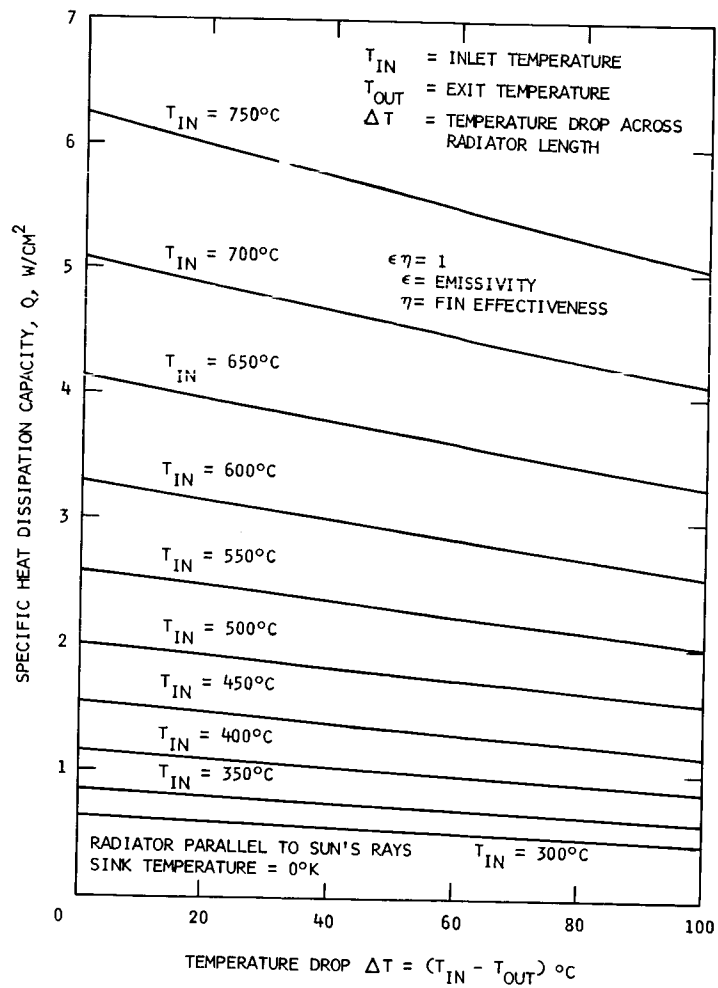
¹Energy Conversion Systems Reference Handbook, Volume X, Reactor System Design, Electro-Optical Systems, September 1960.

T_s = sink temperature

α_s = surface absorptivity to solar radiation

H = solar constant (0.138 w/cm^2 at IAU)

θ = angle between the incident solar illumination and the normal to the radiator surface



Area Requirements for Non-Condensing Radiator

HEAT EJECTION SYSTEMS

Weight and Cost Burdens

Radiator Weight and Cost Variations	Page 550
Weight and Cost Burden Constants	552

RADIATOR WEIGHT AND COST VARIATIONS

Radiator weights and cost vary over wide ranges depending upon the specific heat dissipation.

According to AiResearch Corporation,¹ low temperature radiator specific weight, assuming aluminum construction and structural rigidity as required for radiator areas greater than 50 ft², is approximately 0.95 lb/ft². For smaller radiator areas, depending on the amount of structural rigidity required, the specific weight may be as low as 0.045 lb/ft². Radiator heat dissipation capability versus weight is plotted in Figure A based on 0.95 lb/ft². Typical costs as quoted by the same source indicate development costs of \$50,000, exclusive of environmental testing, for one 10 to 30 ft² space qualified radiator. For production of a number of identical radiators with the above development cost amortized over five units, an approximate functional relationship between radiator cost, C_H , and area, A , of

$$C_H = \$13,750 + 75 A \quad (1)$$

can be inferred. For large production runs, with the development cost amortized over one hundred units, the radiator cost is reduced to

$$C_H = \$2,750 + 25 A \quad (2)$$

Radiator heat dissipation capability versus cost for a five unit production run is plotted in Figure B.

In addition to the heat dissipation capacity of the radiator, three significant factors strongly affect the required cost and weight of the radiator. These factors are: 1) the radiating temperature of the heat radiator (note that this can vary with the type of power amplifier on the spacecraft.); 2) the sink temperature, T_s , in to which the radiator radiates; and 3) the aspect of the radiator to the heat energy from the sun. These are, of course, all factors in determining the specific heat dissipation, Q . Figures A and B indicate typical variations in cost and weight with typical values of emissivity, fin effectiveness, and surface absorptivity.

1. Private Communication, AiResearch Corporation

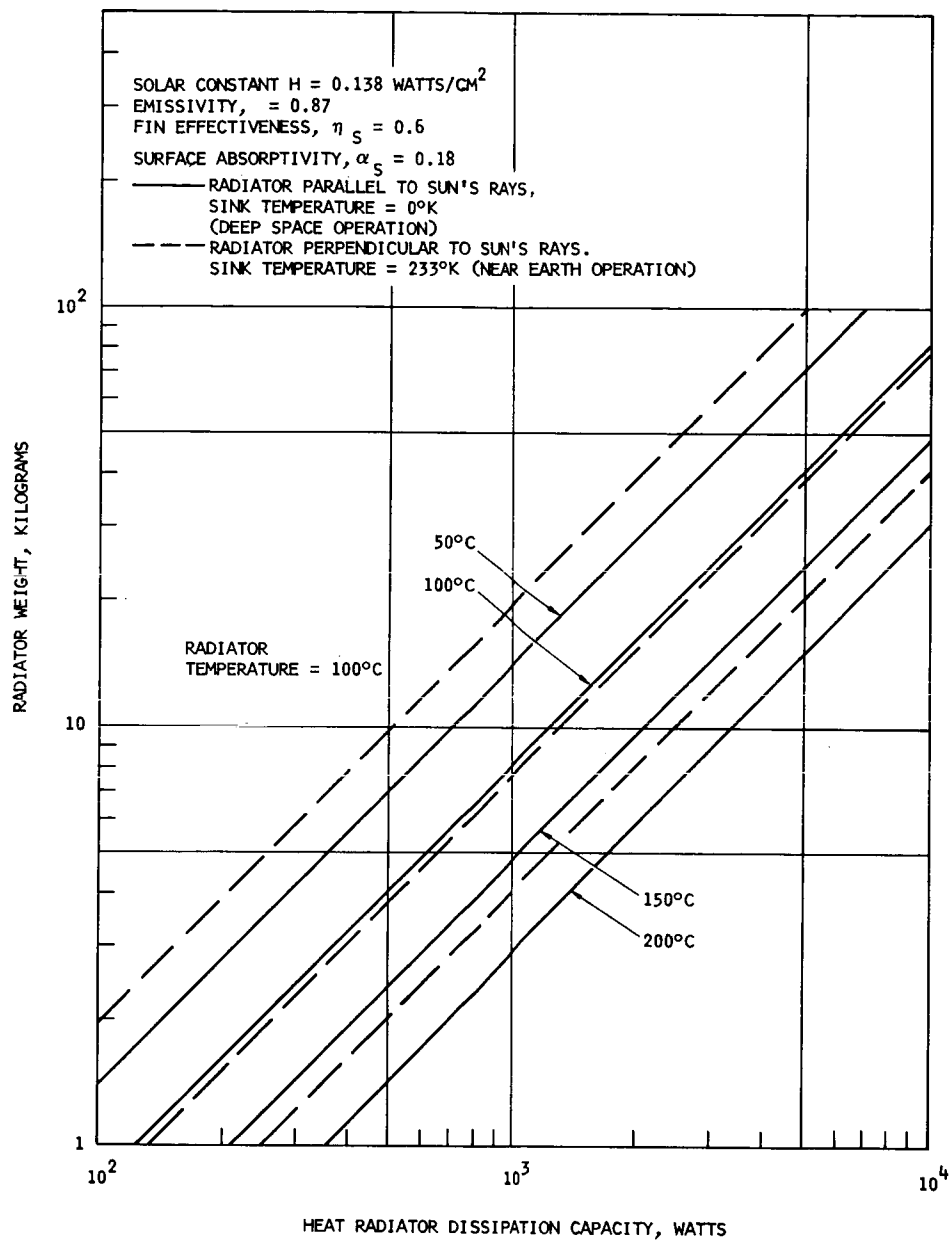


Figure A. Fin and Tube Radiator Weight, W_H (Kilograms),
 Versus Heat Dissipation Capacity at Various
 Radiator Temperatures

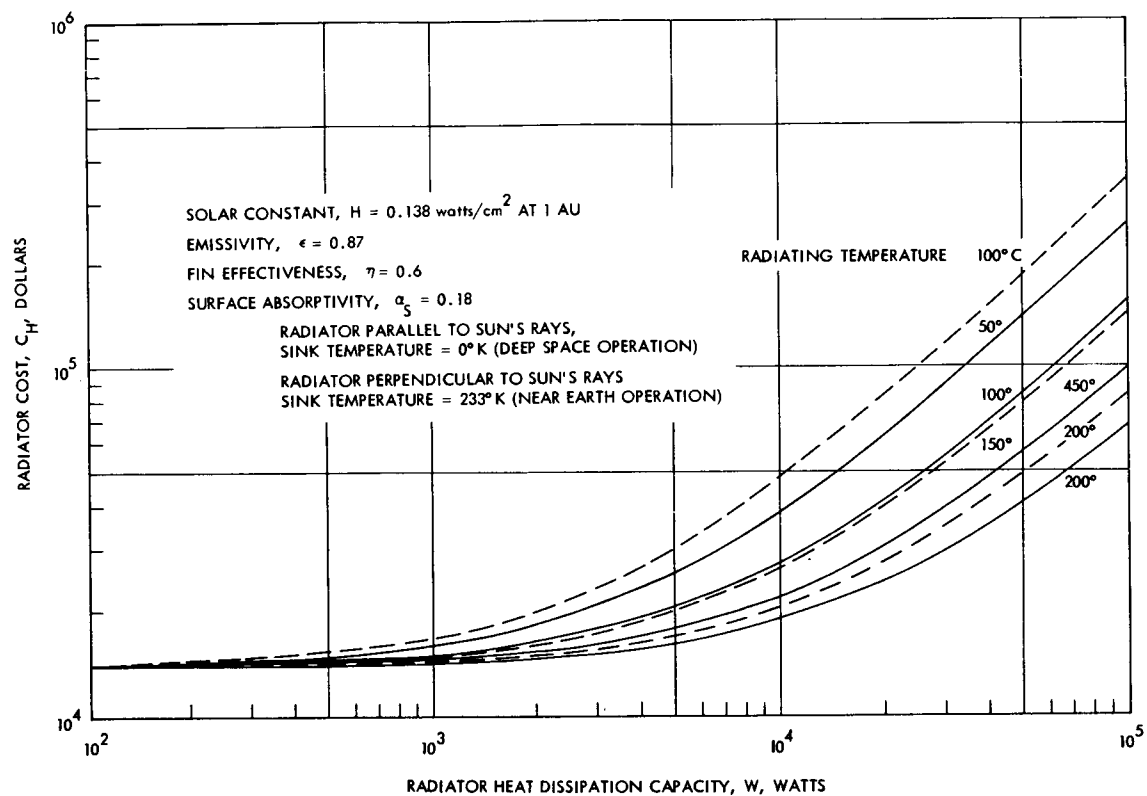


Figure B. Fin and Tube Radiator Cost, C_H , Versus Heat Dissipation Capacity at Various Radiator Temperatures

WEIGHT AND COST BURDEN CONSTANTS

Weight and cost burden for the communications system methodology are derived from the previous topic.

An overall goal of this final report is to provide a means for comparing various communication systems on the basis of cost and weight. To accomplish this it is necessary to express each component in terms of its pertained parameter(s) and weight and cost. This topic lists these relationships for the heat exchanger/radiator.

The fabrication cost and weight of a heat exchanger are proportional to the power which is to be radiated. The heat exchanger weight, W_H , given in the nomenclature of the communications system methodology is:

$$W_H = K_X \left(\frac{1 - k_e}{k_e} \right) P_T + W_{KH}$$

and the heat exchanger fabrication cost, C_H , is

$$C_H = K_H \left(\frac{1 - k_e}{k_e} \right) P_T + C_{KH}$$

where

- W_{KH} = transmitter heat exchanger weight independent of the radiated heat
- C_{KH} = heat exchanger fabrication cost independent of power dissipation
- K_X = constant relating heat exchanger weight to power dissipation
- K_H = constant relating heat exchanger fabrication cost to power dissipation
- k_e = power efficiency, from the prime power source to the output power
- P_T = transmitted power

It is the purpose of this topic to tabulate the values of K_X , W_{KH} , K_H , and C_{KH} . These may be derived from the previous topic and are given in Tables A and B. As may be noted from these tables the burden values are strongly dependent upon the radiating temperature and upon the sun's heat load on the radiator. Thus different values of these burdens are associated with different temperature maintainance requirements (e. g., CO₂ laser requires relatively low operating temperature as compared to a TWT which can operate at a much higher temperature.)

Table A. Values of the Heat Exchanger Weight Burdens¹

Radiating Temperature, °C	W _{KH}		K _X	
	Pounds	Kilograms	Pounds/Watt	Kilogram/Watt
Radiator parallel to sun's rays, sink temperature = 0°K (Deep space operation)				
50	0	0	0.031	0.014
100	0	0	0.0175	0.008
150	0	0	0.011	0.0049
200	0	0	0.0068	0.00311
Radiator perpendicular to sun's rays, sink temperature = -40°C (Near earth operation)				
50	0	0	*	*
100	0	0	0.042	0.019
150	0	0	0.0165	0.0075
200	0	0	0.0088	0.004
* Input heat exceeds radiated heat.				

Table B. Values of the Heat Exchanger Cost Burdens¹

Radiating Temperature, °C	C _{KH}	K _H
	Dollars	Dollars/Watt
Radiator parallel to the sun's rays, sink temperature = 0°K (Deep space operation)		
50	13,750	2.5
100	13,750	1.4
150	13,750	0.84
200	13,750	0.54
Radiator perpendicular to the sun's rays, sink temperature = -40°K (Near earth operation)		
50	13,750	*
100	13,750	3.36
150	13,750	1.31
200	13,750	0.70
* Input heat exceeds radiated heat.		

* Heat radiator parameters are emissivity, $\epsilon = 0.87$; surface absorptivity, $\alpha_s = 0.18$; fin effectivity, $\eta = 0.6$; solar constant $H = 0.138 \text{ watt/cm}^2$.