NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary 37-58, Vol. III

Supporting Research and Advanced Development

For the Period June 1 to July 31, 1969

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

August 31, 1969

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Space Programs Summary 37-58, Vol. III

Supporting Research and Advanced Development

For the Period June 1 to July 31, 1969

JET PROPULSION LABORATORY
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA

August 31, 1969

SPACE PROGRAMS SUMMARY 37-58, VOL. III

Prepared Under Contract No. NAS 7-100
National Aeronautics and Space Administration

Preface

The Space Programs Summary is a multivolume, bimonthly publication that presents a review of technical information resulting from current engineering and scientific work performed, or managed, by the Jet Propulsion Laboratory for the National Aeronautics and Space Administration. The Space Programs Summary is currently composed of four volumes:

Vol. I. *Flight Projects* (Unclassified)

Vol. II. *The Deep Space Network* (Unclassified)

Vol. III. *Supporting Research and Advanced Development* (Unclassified)

Vol. IV. *Flight Projects and Supporting Research and Advanced Development* (Confidential)

Contents

SPACE SCIENCES DIVISION

TELECOMMUNICATIONS DIVISION

Contents (contd)

GUIDANCE AND CONTROL DIVISION

Contents (contd)

ENGINEERING MECHANICS DIVISION

Contents (contd)

Contents (contd)

MISSION ANALYSIS DIVISION

I. Lunar and Planetary Instruments

SPACE SCIENCES DIVISION

A. Microminiaturization of the Peak Detector–Analog-to-Pulse-Width Converter, *J. R. Locke*

1. Introduction

The peak detector–analog-to-pulse-width converter (PD–A/PWC) is a circuit originally designed and developed at JPL in discrete component form. The intent of the design was to provide a simple circuit that could be used to process the output signals of a mass spectrometer or gas chromatograph. Both instruments provide data in the form of pulses. A knowledge of when the pulses occur and what their amplitude is characterizes the primary information obtainable from these instruments. Data economy is achieved with this circuit because only peak amplitude and peak time of occurrence data are encoded, as opposed to the alternative approach of taking multiple data points, to describe the complete pulse, and abstracting information about the peak. Because the output of the PD–A/PWC is a pulse width proportional to the peak, with its leading edge occurring at the time of the peak, these data are readily converted to digital form.

This article will cover: (1) a functional description of the circuit, (2) component specification criteria, (3) circuit fabrication, and (4) test results taken on the prototype unit.

2. Circuit Functional Description

Figure 1 is a schematic of the PD–A/PWC. Peak detection is achieved by storing a voltage proportional to the peak input voltage on capacitor C1; as the input voltage drops below the peak, CR1 is reverse-biased, while CR2 continues to be driven by the input signal. This condition produces an unbalance across Q1AB that triggers Q5 from a normally high to a low state. The collector of Q5 going low produces a signal that fires a one-shot and enables a current source. The one-shot controls a series-shunt switch that disconnects the input from the signal source for a period of time adequate to allow discharge of the maximum input signal seen by capacitor C1. By discharging C1 through a constant current source Q6 to a level that returns Q5 to its high state, the resultant pulse width defined by the leading and trailing edges of the PD–A/PWC output is proportional to the peak input voltage.

The two parameters that most characterize the circuit's performance are the peak delay and the pulse width of the output signal.

Peak delay refers to the elapsed period occurring between the time of the peak and the PD–A/PWC output signal leading edge. This delay occurs because of the finite gain of the circuit and the voltage difference

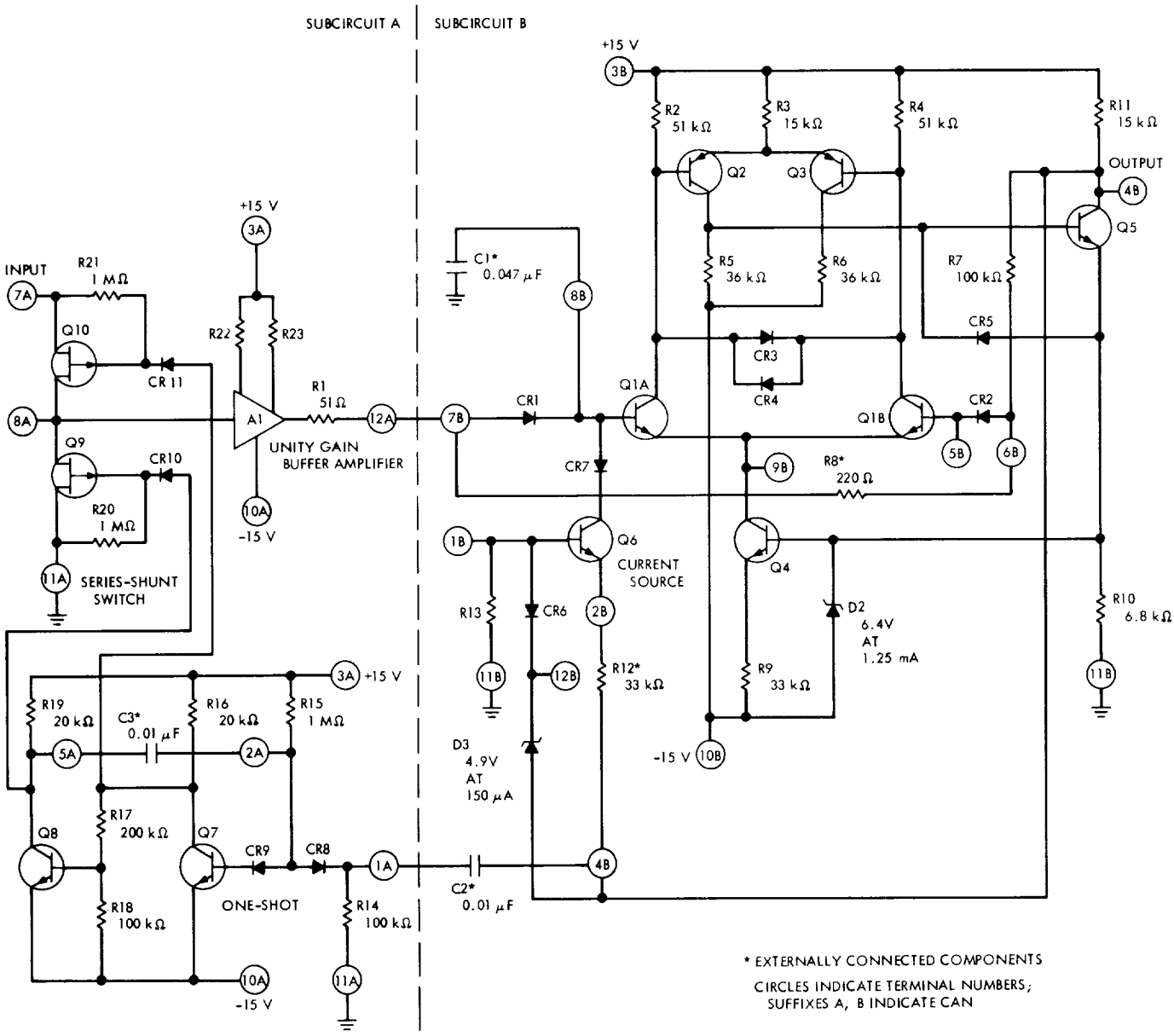


Fig. 1. PD–A/PWC schematic

existing across the bases of Q1A and Q1B at the time of the peak. For the leading edge trigger to occur, the input voltage must drop below the peak by an amount called the offset voltage. This offset voltage V_{os} is indicated in Fig. 2. The input waveform shown is a biased sinewave, as is used in testing the circuit. For such an input, it can be shown that the peak delay t_D is given approximately by the expression

$$t_D = \frac{1}{w} \arccos\left(\frac{V_p - 2V_{os}}{V_p}\right)$$

and

$$V_{os} \cong V_p - \frac{R_7 + R_8 + R_{11}}{R_7 + R_{11}}\left[V_p - V_{cc}\frac{R_8}{R_7 + R_8 + R_{11}} - \frac{KT}{q}\ln\left(\frac{I_{s_2}}{I_{s_1}} \cdot \frac{i_{b_A}(t_p) + C_1\dfrac{dV_{c_1}(t_p)}{dt}}{i_{b_B}(t_x)}\right)\right]$$

where

t_D = delay time in seconds

w = frequency in radians per second

V_p = peak input voltage in volts

V_{os} = offset voltage in volts

R_7 = 100 kΩ

R_8 = 220 Ω

R_{11} = 15 kΩ

V_{cc} = +15 V

K = Boltzmann's constant

T = temperature in °K

q = charge on an electron

I_{s_1}/I_{s_2} = ratio of the leakage currents in diodes CR1 and CR2

$i_{b_B}(t_x)$ = base current of Q1B at the time of the leading edge trigger

$i_{b_A}(t_p)$ = base current of Q1A at the time of input peak voltage

$dV_{c_1}(t_p)/dt$ = voltage rate of change across C1 at the time of input peak voltage

From the delay-time expression, it can be seen that the minimum detectable signal is comparable to the offset voltage, and that, for a constant pulse width, maximum delay occurs for the minimum detectable signal. In general, it can also be inferred that delay time increases inversely with the rate at which the input drops from the peak, and directly with pulse width. From the expression for offset voltage, it should be noted that a signal having a high rate of change with time before and near the peak will increase offset voltage, thus resulting in a higher minimum detectable input signal. It should further be noted that the offset voltage may be controlled to some extent by an appropriate choice of R_8. This choice is an advantage in applications where

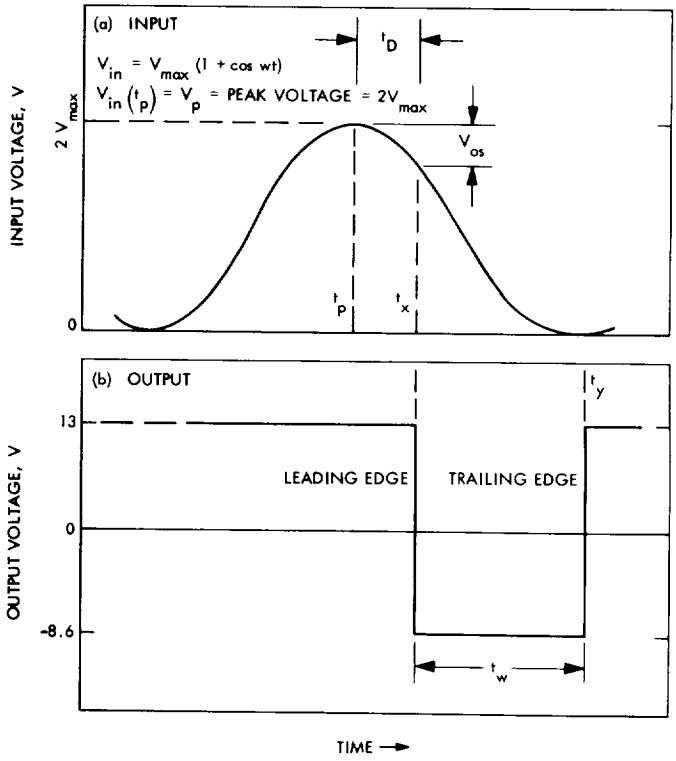


Fig. 2. PD–A/PWC waveforms

noise is a problem, since increasing the offset voltage improves the circuit's noise immunity.

The second parameter, pulse width, is the main subject of the next subsection. The accuracy of how well the PD–A/PWC can measure peak amplitude is largely determined by how repeatable the pulse width is for a fixed amplitude input signal. This repeatability is most significantly affected by the difference in voltage drops across CR1 and CR2, the stability of the current source used to discharge C1, and the constancy of the C1 lower threshold voltage that causes the output to return to its original state.

3. Component Specification Criteria

In preparation for the PD–A/PWC microminiaturization contract, a computer-aided analysis of the circuit was performed (Ref. 1). This analysis included a component sensitivity study of those components affecting the leading edge trigger, the trailing edge trigger, and the capacitor discharge circuit. From these sensitivities have come the component specification criteria. This subsection deals with how these sensitivities were used.

It was stated in the previous subsection that the two parameters that most characterize the PD–A/PWC are the peak delay and pulse width. For this reason, it was decided that the component sensitivity data would best be used by studying the stability of these two functional characteristics. As was suggested in the previous subsection, many of the factors affecting the delay time are associated with the input waveform and not readily related to the component specifications. Also, it was noted that delay time and pulse width were commonly affected by those component sensitivities associated with the leading edge trigger, whereas only the trailing edge trigger and the capacitor discharge circuit were significantly related to the pulse width. For these reasons, pulse-width stability alone was chosen as the criterion for the component specifications.

The discussion of the pulse width that follows excludes any contributing effects that the one-shot, series-shunt switch, or unity gain buffer amplifier might have. Component requirements for these subcircuits were based on conventional hand analysis techniques and will not be discussed in this article.

The first step in determining the component requirements in terms of the pulse-width stability was to write the expression for pulse width:

$$t_w = \frac{C_1[V_c(t_x) - V_c(t_y)]}{I_K + i_{b_{1A}}} \approx \frac{C_1[V_c(t_x) - V_c(t_y)]}{I_K}$$

where

$t_w = t_y - t_x$ = pulse width

t_x = time of the leading edge trigger

t_y = time of the trailing edge trigger

C_1 = capacitance of capacitor C1 = 0.047 μF

$V_c(t_x)$ = voltage across C1 at the time of the leading edge trigger

$V_c(t_y)$ = voltage across C1 at the time of the trailing edge trigger (≈ -0.30 V)

I_K = on-collector current of Q6 $\approx 200\ \mu$A

$i_{b_{1A}}$ = base current of Q1A ($i_{b_{1A}}$ (max) = 950 nA)

This expression was then differentiated with respect to p, a term chosen to represent any one particular component in the circuit, with the exception of C1, which was assumed to be a constant. The resultant expression is

$$\frac{\partial t_w}{\partial p} = \frac{C_1\left[\frac{\partial V_c(t_x)}{\partial p} - \frac{\partial V_c(t_y)}{\partial p}\right]}{I_K} - C_1 \frac{\partial I_K}{\partial p} \frac{[V_c(t_x) - V_c(t_y)]}{I_K^2}$$

where

$\frac{\partial t_w}{\partial p}$ = change in pulse width for a change in a particular parameter p when all others are held constant

$\frac{\partial V_c(t_x)}{\partial p}, \frac{\partial V_c(t_y)}{\partial p}$ = sensitivities of the leading and trailing edge trigger points as a function of a particular parameter p when all others are held constant

$\frac{\partial I_K}{\partial p}$ = change in the current discharging C1 as a function of a particular parameter p when all others are held constant.

Worst-case numerical values for

$$\frac{\partial V_c(t_x)}{\partial p}, \frac{\partial V_c(t_y)}{\partial p}, \text{ and } \frac{\partial I_K}{\partial p}$$

were obtained from the computer simulation of the circuit, from which Table 1 was prepared for the pulse-width component sensitivities. The total change in pulse width can then be expressed as the sum of the changes contributed by the resistors, junction voltages, current gains, and zener voltages:

$$dt_w = (dt_w)_R + (dt_w)_J + (dt_w)_\beta + (dt_w)_v$$

$= \text{total change in pulse width}$

where

$$(dt_w)_R = \sum_i \left(\frac{\partial t_w}{\partial p_R} dp_R\right)_i = \text{total change in pulse width contributed by resistance change}$$

$$(dt_w)_J = \sum_j \left(\frac{\partial t_w}{\partial p_J} dp_J\right)_j = \text{total change in pulse width contributed by transistor and diode junction voltages changes}$$

$$(dt_w)_\beta = \sum_k \left(\frac{\partial t_w}{\partial p_\beta} dp_\beta\right)_k = \text{total change in pulse width contributed by current gain changes}$$

$$(dt_w)_v = \sum_l \left(\frac{\partial t_w}{\partial p_v} dp_v\right)_l = \text{total change in pulse width contributed by zener voltage changes}$$

and

$\partial t_w/\partial p_R, \partial t_w/\partial p_J, \partial t_w/\partial p_\beta, \partial t_w/\partial p_v$ = sensitivities listed in Table 1.

dp_R = percentage change of resistance

dp_J = change in junction voltage expressed in millivolts

dp_β = change in current gain

dp_v = change in zener voltage expressed in millivolts

i, j, k, l = indices given to the resistors, junction voltages, current gains, and zener voltages, respectively.

Using the relationships just outlined and the sensitivities of Table 1 as guidelines, the minimum component requirements were set (Table 2). Imposing these requirements on the circuit, the worst-case change in pulse width was calculated for a 10-V peak input (full-scale) and a 50°C temperature change. The nominal pulse width for a 10-V peak input voltage is 2.350 ms and the calculated change was -13.8 μs. This change corresponds to -0.59%, which was considered reasonable and obtainable in terms of the known component performance that would be available in microminiature form.

4. Circuit Fabrication

Motorola's Government Electronics Division was awarded the contract to fabricate the PD–A/PWC in microminiature form. The hybrid circuit technique was chosen as the fabrication method to be used. This technique has several virtues with respect to JPL's needs, for the following reasons: (1) well-suited for small production quantities, (2) development of circuitry in discrete form generally applicable to hybrid version, and (3) quick startup in getting into production.

The specific hybrid technique used by Motorola is a multichip hybrid approach. This approach involves the attachment of both active and passive components in chip form upon a thick film conductor array supplemented with wire bonding.

The thick film system used for conductors and die pads consists of silk screening and firing a molybdenum–manganese paste onto a ceramic substrate. It is then successively tinned with nickel, gold, and a gold–germanium (AuGe) braze alloy. Component parts are then attached to the substrate by heating the AuGe to its melting temperature in an inert atmosphere and scrubbing the die into place. Once properly oriented, the die is trapped in place by allowing the AuGe braze alloy to freeze.

Wire bonding performed on the PD–A/PWC is accomplished by utilizing thermo-compression wire bonding methods. The principle of the bond is based on the theory that the metal members being bonded have large numbers of atoms with unsatisfied bonds. Heat and pressure produce a plastic deformation that brings the atoms

Table 1. Pulse-width component sensitivity

Resistor sensitivity	
Resistor	$\partial t_w/\partial p_R$, μs/ %
R1	+ 1.36
R2	+ 0.114
R3	− 0.00366
R4	− 0.109
R5	+ 0.00396
R6	+ 0.00000
R7	− 3.13
R8	+ 2.22
R9	− 0.532
R10	+ 0.00000
R11	− 0.820
R12	+ 2.305 $[V_c(t_x) - V_c(t_y)]$
R13	+ 0.022 $[V_c(t_x) - V_c(t_y)]$
Junction voltage sensitivity	
Junction[a]	$\partial t_w/\partial p_J$, μs/mV
Q1A(BE)	− 0.146
Q1B(BE)	+ 0.146
Q2 (BE)	− 0.00236
Q3 (BE)	+ 0.00226
Q4 (BE)	− 0.000124
Q5 (BE)	− 0.000069
Q6 (BE)	+ 0.0474 $[V_c(t_x) - V_c(t_y)]$
CR1	− 0.1562
CR2	+ 0.1462
CR3	− 0.0000039
CR4	+ 0.0000070
CR5	+ 0.0000000
CR6	− 0.0469 $[V_c(t_x) - V_c(t_y)]$
CR7	− 0.0000023 $[V_c(t_x) - V_c(t_y)]$
Current gain sensitivity	
Transistor[b]	$\partial t_w/\partial p_\beta$, μs/$\Delta\beta$
Q1A, h_{fe} = 200	+ 0.0346
Q1B, h_{fe} = 200	− 0.0325
Q2, h_{fe} = 150	+ 0.00915
Q3, h_{fe} = 150	− 0.00524
Q4, h_{fe} = 150	+ 0.0423
Q5, h_{fe} = 150	+ 0.000189
Q6, h_{fe} = 200	− 0.00570 $[V_c(t_x) - V_c(t_y)]$
Zener diode voltage sensitivity	
Zener diode	$\partial t_w/\partial p_v$, μs/mV
D2	+ 0.000194
D3	− 0.0464 $[V_c(t_x) - V_c(t_y)]$

[a]BE = base-emitter junction of transistor.
[b]h_{fe} = forward current gain of transistor.

Table 2. Minimum component requirements

Component	Requirement
Resistors	R12 temperature coefficient ≤ 25 ppm/°C Sum of R1 and R8 tracks R7 to within ±20 ppm/°C All other temperature coefficients ≤ 200 ppm/°C
Junctions	Q1A base-emitter junction voltage tracks Q1B voltage to within ±20 ppm/°C (at I_E = 100 μA) CR1 junction voltage tracks CR2 voltage to within ±20 ppm/°C (at I_E = 100 μA) Q6 base-emitter voltage (at I_E = 100 μA) tracks the CR6 junction voltage (at I_D = 100 μA) to within ±20 ppm/°C
Transistors	Room temperature current gain is as indicated in Table 1, stable to 1%/°C Q1A tracks Q1B to within 20% Q2 tracks Q3 to within 30%
Zener diodes	D2 voltage temperature coefficient ≤ 300 ppm/°C D3 voltage temperature coefficient ≤ 20 ppm/°C

in intimate contact, and a bond is formed. Aluminum wire is used to avoid the reliability problems associated with a gold–aluminum interface (so-called purple plaque problems).

The package design used for the PD–A/PWC consists of separating the circuit into two parts and packaging them in two 12-leaded TO-8 weldable hermetic cans. Two resistors (R8 and R12) and three capacitors (C1, C2, and C3) are connected externally. R12 and C1 are external to allow flexibility in scaling pulse width to peak input voltage. C3 is external to allow adjustment of the blanking period for different minimum intervals between pulses. R8 provides a variable measure of noise immunity, as discussed in *Subsection 2*. C2 was placed external because of the large amount of space required by it inside the can and the fact that it provides a convenient way of joining the two packages, while allowing additional circuit accessibility.

All components, with the exception of the capacitors, are in chip form. The diodes and transistors are in die form, and the resistors are thin film vacuum-deposited nichrome elements on a silicon base. It should be noted that the external resistors R8 and R12 are also of this form, but packaged in TO-18 cans.

Figure 3 shows the layouts of the two subcircuits, and Fig. 4 is a photograph comparing a discrete version of the PD–A/PWC with the prototype hybrid version mounted in a chassis used for testing.

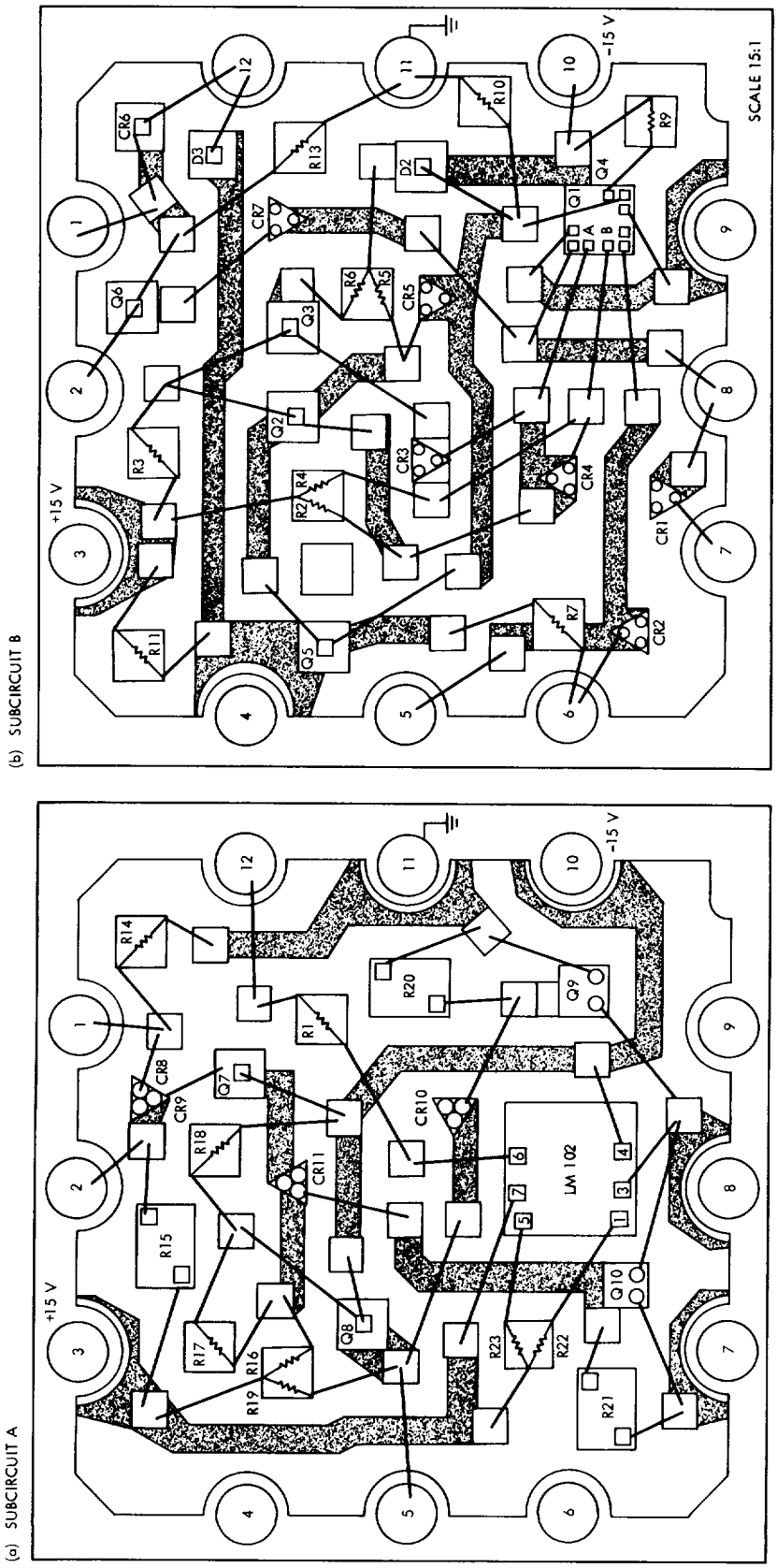


Fig. 3. PD–A/PWC subcircuit layouts (Fig. 1 shows subcircuit schematics)

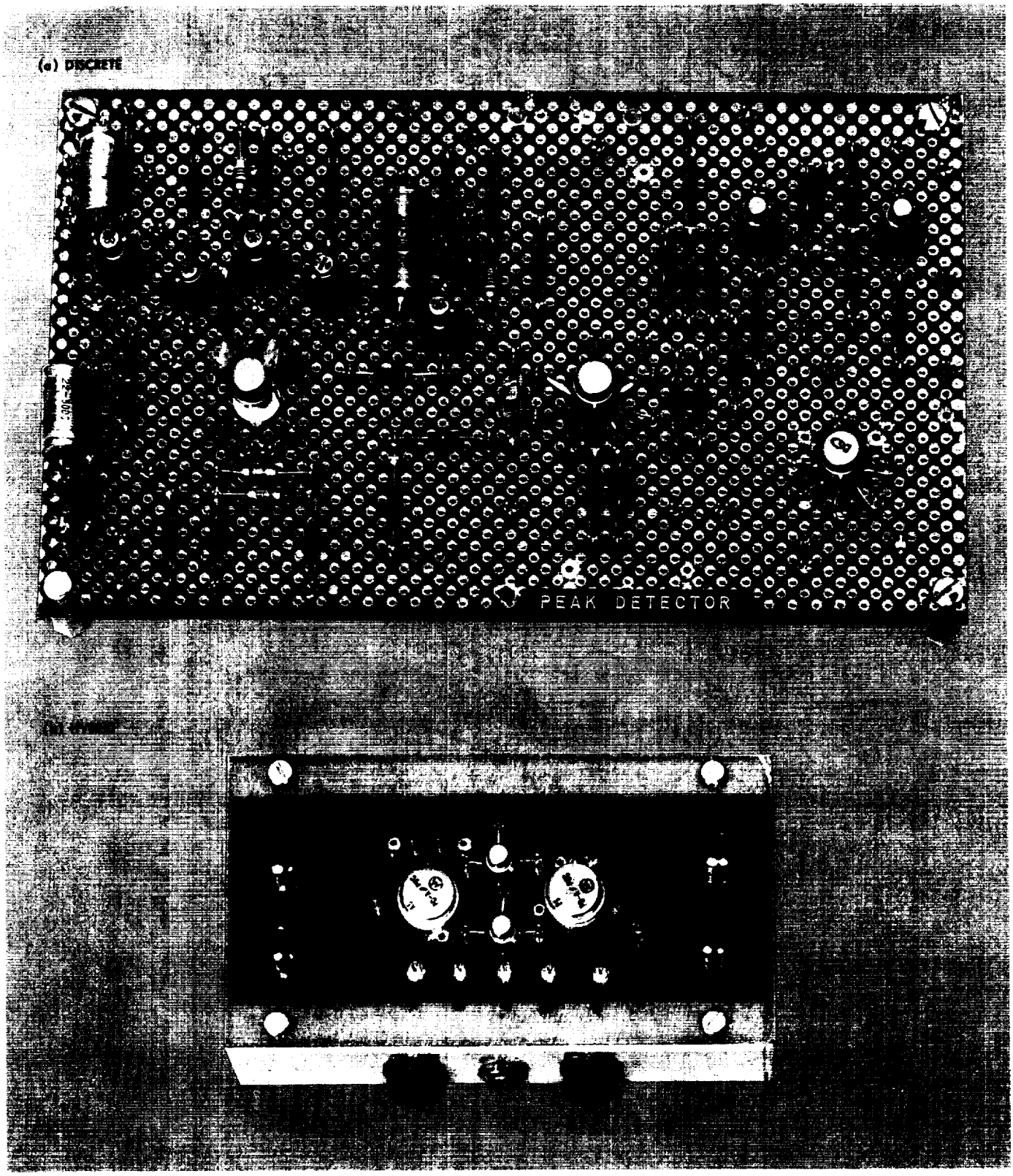


Fig. 4. PD–A/PWC versions

5. Test Results of the Prototype

Under the terms of the Motorola contract, all units are subjected to the following environmental conditions:

(1) *Heat soak.* The unit under test is raised to 135°C and maintained there for a 24-h period. At the end of this period, the unit is returned to room temperature for a minimum of 1 h. This sequence is repeated two more times for a total of three cycles.

(2) *Thermal stress.* The PD–A/PWC is cycled five times from −55 to 125°C. The unit is maintained at each of these temperatures for a period of 1 h. The transition time between temperatures is a maximum of 15 min.

(3) *Humidity.* The PD–A/PWC is raised to a temperature of 50°C and exposed to a relative humidity of 60 to 70%. The unit is then energized, and pulse-width and delay-time data are taken 30 min. after energizing the circuit.

(4) *Centrifuge.* The PD–A/PWC is subjected to 20,000-*g* centrifuge tests that will stress the unit in each of three orthogonal directions.

(5) *Vibration.* The PD–A/PWC is subjected to vibration along each of three orthogonal axes. The vibration program is similar to JPL SPEC TS 500437 B for class II instruments.

Pulse-width and delay-time data are taken before exposing the PD–A/PWC to the environmental condition and after exposure. Acceptance of the unit is based on no discernible change in performance.

The prototype unit shown in Fig. 4b was subjected to these tests and successfully met the test criteria. Table 3 consists of pulse-width and peak-delay data taken on the prototype unit at JPL. Pulse-width drift for a 50°C change was −0.44% for the 10-V peak amplitude. This change compares favorably with the −0.59% worst-case change predicted in *Subsection 4.* The full-scale pulse width used in the analysis and the full-scale pulse width shown in Table 3 differ because different values of discharge current I_K were used.

In addition to improving reliability, and decreasing the size and weight of the PD–A/PWC, it was anticipated that the improved thermal coupling of the components in the hybrid configuration would improve performance. This belief has apparently been borne out in comparing the pulse-width stability of the prototype with that of the discrete component version shown in Fig. 4a. Similar peak-delay and pulse-width data taken on the discrete component version show the drift to be 3.3 times greater than that of the prototype when normalized with respect to pulse width.

Reference

1. Overbey, J. L., and Locke, J. R., *Computer Aided Analysis of a Peak Detector and Analog-to-Pulse-Width Converter,* Technical Memorandum 33-401. Jet Propulsion Laboratory, Pasadena, Calif., Mar. 15, 1969.

Table 3. Prototype functional test data

Peak input voltage, V	0°C		25°C		50°C		t_w (25°C) − t_w (0°C), μs	t_w (50°C) − t_w (25°C), μs
	t_D, ms	t_w, μs	t_D, ms	t_w, μs	t_D, ms	t_w, μs		
10.000	0.68	3268.1	0.76	3261.1	0.80	3253.3	−7.0	−7.3
5.000	1.00	1630.8	1.08	1625.3	1.14	1620.2	−5.5	−5.1
3.000	1.25	977.1	1.36	972.3	1.40	968.7	−4.8	−3.6
1.000	1.90	327.1	2.10	323.6	2.16	321.4	−3.5	−2.2
0.500	2.55	166.7	2.75	164.1	2.84	162.5	−2.6	−1.6
0.300	3.08	103.7	3.36	101.6	3.40	100.4	2.1	−1.2
0.100	4.62	42.4	4.3	41.6	4.8	41.0	0.8	−0.6
0.050	6.6	27.5	6.4	27.2	6.5	26.9	0.3	−0.3
0.030	—	—	9.1 ±0.1	21.0	8.8 ±0.2	21.2	—	0.2
0.010	—	—	—	—	—	—	—	—
Threshold 0.036	9.1 ±0.2	21.8	—	—	—	—	—	—

II. Space Instruments

SPACE SCIENCES DIVISION

A. Breadboard Magnetic Core Memory, *L. L. Lewyn*

1. Introduction

A breadboard magnetic core memory for a pulse height analysis system has been developed and tested. The memory was designed specifically for operation in a lunar or planetary orbiting gamma-ray spectrometer instrument. Significant goals were simplicity of design, low power consumption, wide temperature operation range, and broad performance margins. The memory was procured under contract to Analog Technology Corporation where the major portion of the design effort was accomplished by Mr. T. Harrington and Dr. J. H. Marshall.

2. Theory of Operation

A core memory is used in a pulse height analyzer to accumulate a histogram of the spectrum analyzed by the system and to read out the spectrum when required. The memory accumulates 512 channels of pulse height analysis. Each channel corresponds to a quantization level in the histogram. The capacity of each channel is 65,535 counts, corresponding to a 16-bit binary word.

The core memory is organized around a 1024-word by 8-bit stack. Each channel, therefore, requires two words of storage. In normal data accumulation, an event is quantized and a count of one is added to the appropriate channel. The address corresponding to the 8 least significant bits of the channel is selected and a count of *one* is added. If an overflow occurs, the address associated with the most significant 8 bits is selected and a count of *one* is added.

Conventional pulse height analyzer memories are arranged so that the accumulation cycle is stopped when the memory is being read out. On planetary missions where the telemetry bandwidth is limited, the memory must be read out at a low bit rate. A readout contains approximately 8000 bits and must occur every 3 to 5 min. The time required for readout reduces the accumulation duty cycle.

This memory is arranged so that readout may occur during data accumulation. The readout system operates so that no data are lost during the period of time when data are being transferred from the stack to the readout

data register. The data may be read out destructively or non-destructively.

3. System Description

The memory system block diagram is shown in Fig. 1. The 9-bit address register contains the channel number of the pulse height analysis event to be recorded in the memory. Each channel number is associated with two words in the memory. These words contain the least and most significant 8 bits of data in each channel. The odd/even line from the programmer determines whether the most or least significant 8 bits are to be operated on. The odd/even line combined with the state of the address register controls the state of the X and Y switch drivers in the word selection circuitry.

The X and Y switch drivers operate the X and Y switches which control the routing of the memory read and write currents. The X and Y half currents are combined in the stack wiring to produce a full read or write current in one word at a time. The read currents set all of

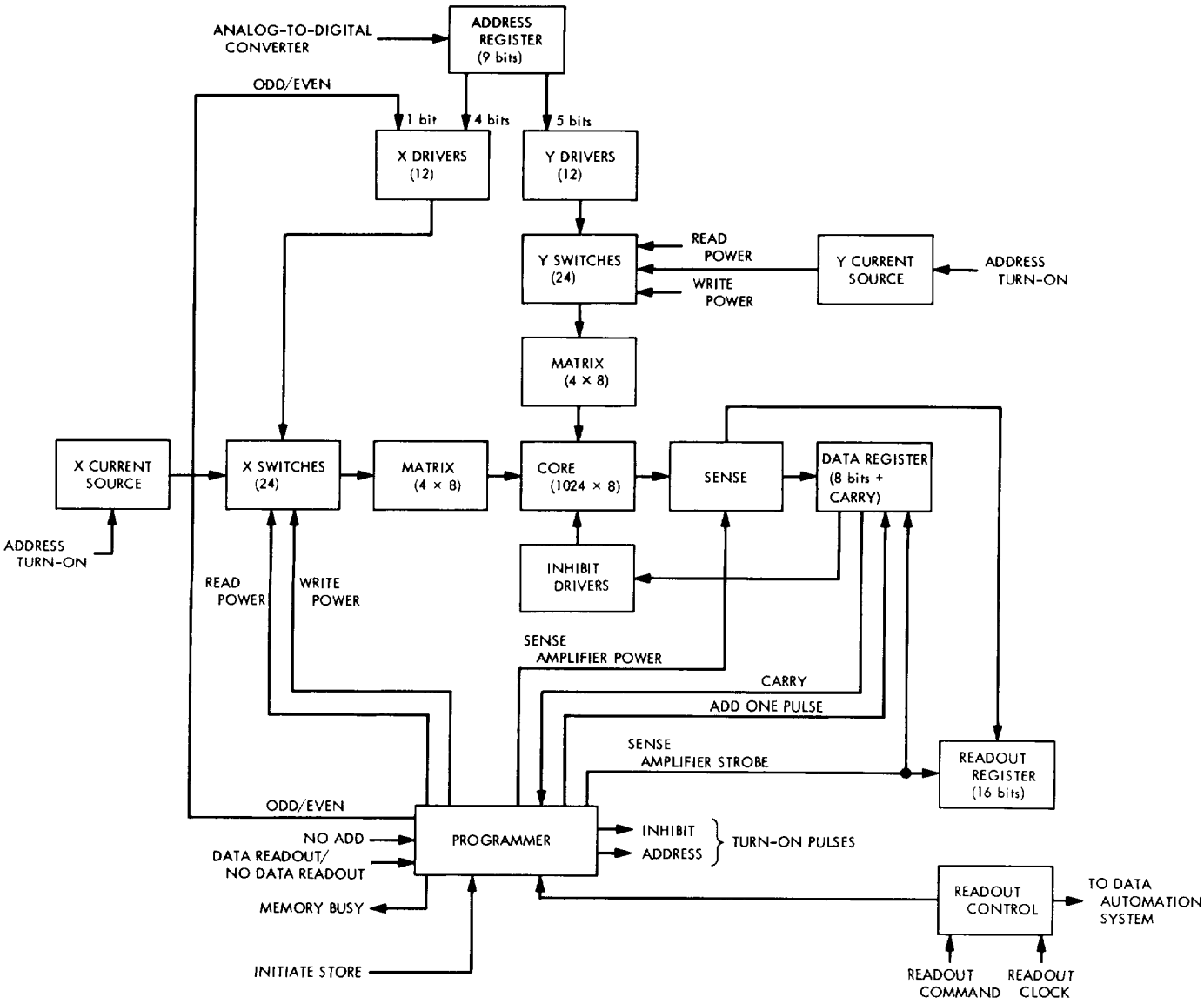


Fig. 1. Magnetic core memory system

the cores in the selected word to a *zero* state. The 8 sense amplifiers detect the flux transitions of the cores in the *one* state and set the data register flip-flops accordingly. During accumulation, an add *one* pulse is transmitted to the data register and its contents are written back into memory.

The data are written back into memory by allowing the half X and Y write currents combined in the stack to produce a full write current in all the cores of the selected word. Those bits which must be prevented from making a transition from *zero* to *one* are given a half current which opposes the write current. The state of the data register determines which inhibit drivers are to furnish opposing current.

If during the add *one* cycle on the least significant bits an overflow occurs, then the programmer changes the state of the odd/even line and causes the add *one* process to be repeated on the most significant bits.

The readout process is similar to the accumulation process except the add *one* cycle is not performed. During the readout process, both least and most significant bits of the channel to be read out are entered into the readout register in rapid succession. The memory then remains free to perform additional accumulation while the readout register shifts out the data at a rate determined by the readout control logic.

4. Test Results

The read, write, and inhibit currents specified for the stack as a function of temperature are summarized in Table 1. The performance of the address driver circuit is summarized in Table 2. The performance indicates that the address driver temperature tracking is in excellent conformance with the stack requirements. The nominal rise-time specification for the stack is 200 ns.

Worst-case sense winding differential output signals at 90, 25, and −40°C are shown in Fig. 2. The signal occurring 2 cm after the start of the trace is of interest. The large signal represents the sense voltage resulting from the transition of a single core. The small disturbance which

Table 1. Core stack current requirements

Temperature, °C	Current, mA		
	One-half read	One-half write	Inhibit
−30	331	331	321
25	315	315	305
80	298	298	288

Table 2. Address driver circuit performance

Temperature, °C	Measured current, mA	Deviation from desired current, %	Minimum supply voltage, V	Current 10–90% rise time, ns	Current delay, ns
80	300	0.7	3.57	215	110
25	318	1.0	3.65	210	130
−30	330	−0.3	3.92	200	140
−50	335	−1.3	4.15	190	145

Table 3. Memory operating margins

Temperature, °C	V_{sense}, V		Clock, MHz		Inhibit V_{ref}, V		Address V_{ref}, V		Inhibit voltage (+5 V) lower limit, V	Address voltage (+5 V) lower limit, V
	Upper limit	Lower limit	Upper limit	Lower limit	Upper limit	Lower limit	Upper limit	Lower limit		
−40	14	2.9	9.85	6.5	0.50	0.30	0.46	0.32	2.0	3.4
−30	8[a]	2.4	9.4	6.0	0.49	0.29	0.45	0.30	1.91	3.4
−10	8.2[a]	2 3	9.65	6.0	0.47	0.30	0.44	0.31	1.85	3.5
25	14	1.9	10.0	6.5	0.48	0.27	0.45	0.31	2.0	3.5
40	8.0[a]	1.9	9.45	6.0	0.46	0.30	0.43	0.32	1.75	3.5
80	8[a]	1.5	9.6	5.8	0.41	0.32	0.40	0.30	2.65	3.75
90	8[a]	1.5	8.9	6.7	0.40	0.31	0.38	0.30	2.70	3.75

[a]Readings were taken with sense amplifier–strobe line tied to V_{sense}.

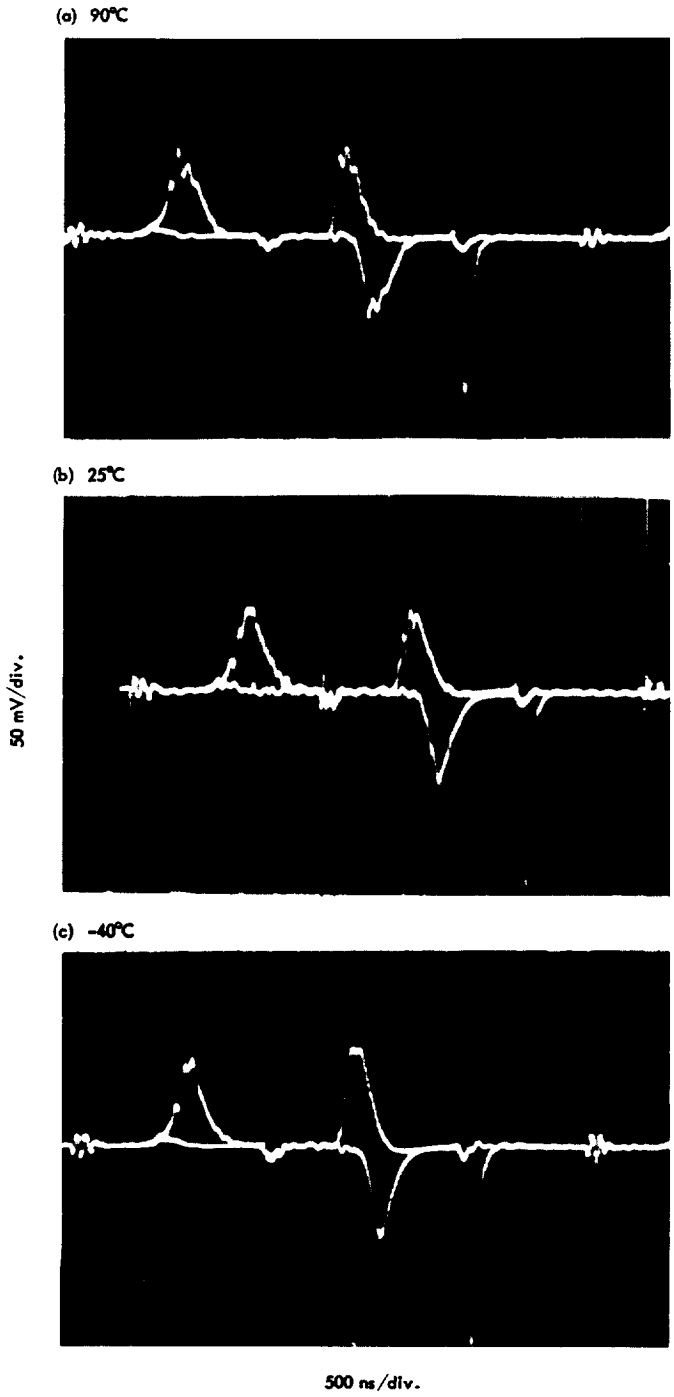


Fig. 2. Sense winding differential output signals (worst-case pattern)

rises to less than 1/10 the amplitude of the large signal represents the disturbance resulting from the selection of a worst-case pattern of cores with none making an actual transition. The large signal-to-noise margin and signal stability as a function of temperature indicate that sense amplifier thresholds may be easily established without the use of fine adjustment or temperature compensation.

An extensive series of temperature tests were performed to determine memory operating margins. The results of these tests are summarized in Table 3. The V_{sense} tabulations refer to the variation of the supply voltage to the sense threshold voltage divider and are, therefore, proportional to the threshold variations. Note that the thresholds may be varied widely at all temperatures without affecting the operation of the memory.

The inhibit and address reference voltages V_{ref} are in normal operation programmed to compensate for the core drive temperature variations. Note that the lack of overlap in limit voltages at the temperature extremes indicates that the memory will operate with constant reference voltages. Reference voltage temperature programming will be retained to keep broad operating margins.

Switch resistance and core inductance result in several voltage drops which are additive and must be exceeded by the driver voltage during the rise of the current pulse. The value of the drive-current pulse amplitude and width, together with the supply voltage, determines the power consumed by the memory during each cycle. The values of inhibit and address voltage lower limits indicate that the voltage margins are conservative at all operating temperatures.

The memory was constructed using Signetics 8400 and 8200 series logic elements. As a result, the standby power for the logic system was 2.35 W. If Texas Instruments series 54L logic elements were substituted for the Signetics logic, then the standby power for the entire memory system would be 258 mW. The memory power consumption with 54L logic elements is summarized in Table 4.

5. Conclusions

The test results indicate that the breadboard magnetic core memory can operate with broad performance margins over a temperature range in excess of 130°C. The standby power consumption of 258 mW utilizing 54L logic elements is extremely low for a memory of this size. Proprietary switching techniques developed by Analog Technology Corporation result in simplicity of circuit design and compact packaging.

The breadboard memory has exceeded the design objectives in operating temperature, power consumption, performance margins, and simplicity. In the near future,

Table 4. Power requirements of individual circuit elements

Element	Power, mW		
	Standby	500-Hz rate	5×10^4-Hz rate
Inhibit drivers	46	48	281
Address drivers	11	14	254
Address-matrix switches	0	0.04	4
Sense amplifiers and power switch	1	1.8	78
Logic, voltage divider, and buffers (estimate)	200	200	200
Double cycle	0	0	2
	258	264	819

the memory will be packaged for flight. It should prove to be a useful part of the gamma-ray pulse height analysis system or any other instrument system requiring a flight-packaged state-of-the-art core memory.

B. Low-Power Cold Gas Valve, *H. E. Geise*

1. Introduction

The plans for a balloon flight in the spring of 1966, to test the infrared multidetector spectrometer, started a search for a commercially available valve that could regulate the flow of gas at near liquid-nitrogen temperature. This gas was to be used for cooling the infrared detectors and maintaining their temperature at the desired level. The valve would be expected to satisfy the following requirements:

(1) Reliability.

(2) Operation in the balloon flight environment.

(3) Low power consumption.

Such a valve could not be purchased commercially and was, therefore, developed by the JPL Infrared Instruments Group. Its design and performance under test is described in this article.

2. Specific Requirements

a. Reliability. The requirement for reliability was of paramount importance. The costly investment in time, manpower, and money would largely be wasted in the event of a valve failure. Failure in the open position might result in a "run-away" of temperature. The infrared detectors used in the spectrometer required a temperature controlled at 193°K. The flow through the valve had to be metered accurately enough to hold this temperature within ±1 deg over a 6- to 8-h period of continuous operation.

b. Environment. The float altitude was expected to be around 125,000 ft (about 4 mbar). Ascending to this altitude, the gondola would be exposed to temperatures ranging from 210°K at approximately 53,000 ft to 240°K at 125,000 ft.

c. Power. While it was the least critical of the three requirements, power communications still had to be considered from two viewpoints:

(1) The only power available on the gondola was from 28-V batteries.

(2) Any heat generated by solenoids in a "power-on" condition would have a tendency to heat the valve body and that, in turn, would increase the temperature of the cooling gas and cause the LN_2 supply to be used at an increased rate. Since the available quantity of LN_2 determined the duration of the flight, it should not be spent needlessly.

3. Design

After considerable searching for a commercial unit, the decision was made to develop in-house a valve to meet these requirements. The original breadboard model was designed around hardware already on hand. A brass body was machined to accept two existing 28-V solenoids at opposite ends (Fig. 3). A through hole is provided for a shaft (Fig. 4) which connects the solenoid plungers and provides a shuttle action as one or the other solenoid is energized. The shaft is held in either of the two positions by means of grooves in the shaft and detent balls, springs, and pressure-adjusting set screws. This feature enables operation in either the "flow" or "no flow" configuration, while current is used only for transition from one configuration to the other.

Since this valve was expected to operate at low temperatures, the decision was made to use Teflon for the sealing surface. A poppet valve was machined similar to that shown in Fig. 4, but entirely of Teflon. A sealing surface to match that of the poppet was machined in the brass body, and flow holes intersecting this surface for handling the in or out flow of gas were drilled. The correct angle was determined from the following considerations: If the included angle on the poppet and body was

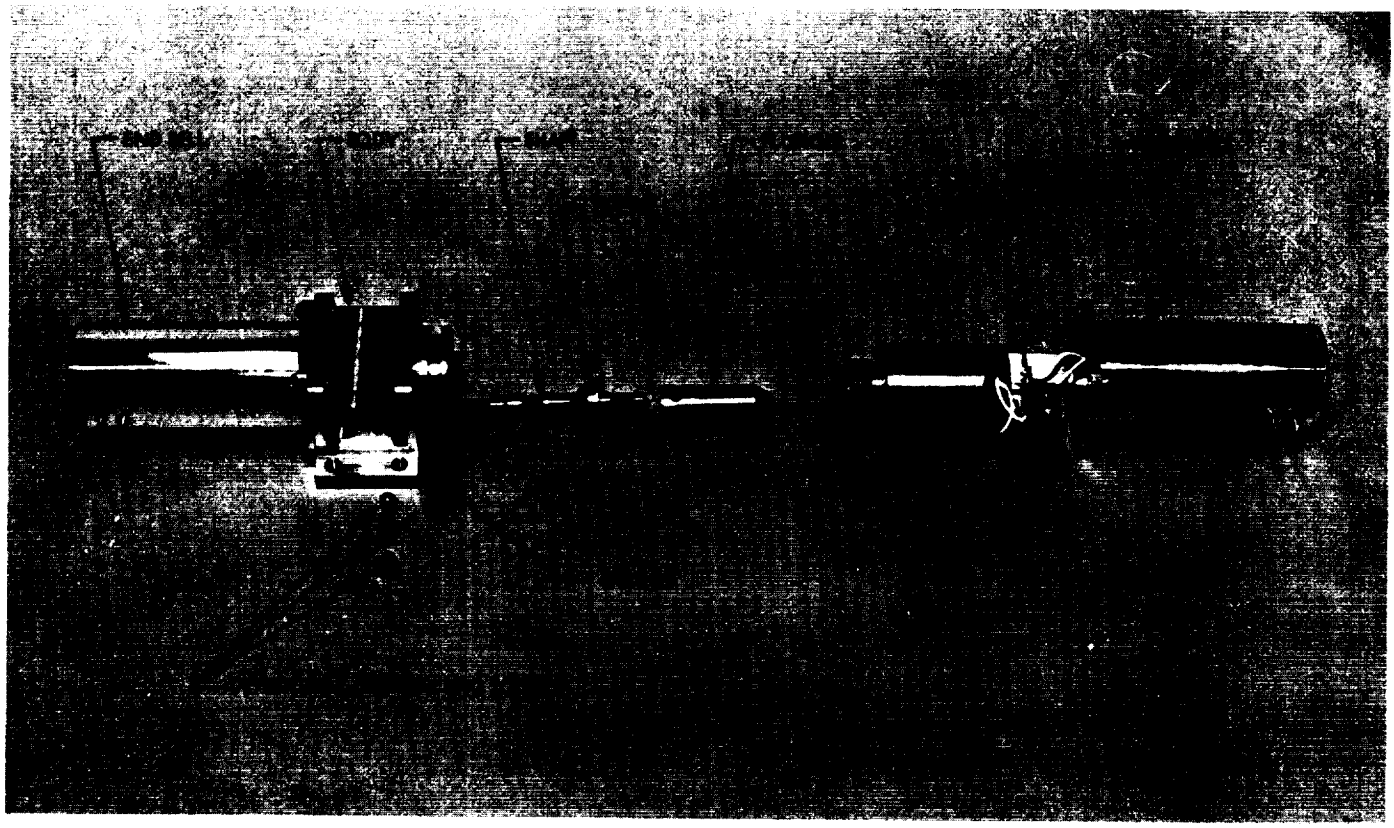


Fig. 3. Exploded view of cold gas valve

too shallow, the poppet would stick in the "no flow" position (similar to a tapered drill in a holder); if the included angle was too large, the gas pressure would lift it off the seat and the valve would leak. After several angle modifications and additional testing, the most promising included angle was found to be 75 deg.

Figures 3 and 4 show the final version of the valve as used during the balloon flights. The microswitch at the far end of each solenoid is used to switch power between solenoids. End bells with electrical feedthroughs protect the solenoids from the balloon environment.

4. Testing

Extensive environmental tests were conducted in a cold chamber with the valve in a "flow" condition for 5 s then switched to a "no flow" condition for 15 s. Tests were run continuously for 48 h at 50, 0, −50, and −100°C; operation of the valve was monitored. The output line from the valve body was connected to a flow meter, with photoelectric cells at top and bottom of the scale to record the position of the flow ball when the valve was in either a "flow" or "no flow" condition. A leaky valve would keep the flow ball off the seat, and this would be recorded on a chart recorder.

During this series of tests, the valve showed an occasional small leak in the "no flow" condition. This brought about two changes: The porting holes in the body were relocated so that the "in" pressure would help seal the poppet; and the Teflon poppet was replaced with a stainless steel poppet that incorporated an O-ring groove. A silastic O-ring was installed. The valve was reassembled and testing was resumed. This modification corrected the occasional leak, and the cold test was successfully concluded.

The valves were then installed on the dewars, as shown in Fig. 5, and complete systems were started. These tests have been extensive and include:

(1) Chrysler high-altitude chamber (1966).

(2) JPL 10-ft space simulator (1966).

(3) JPL vacuum chamber for instrument calibration (1966).

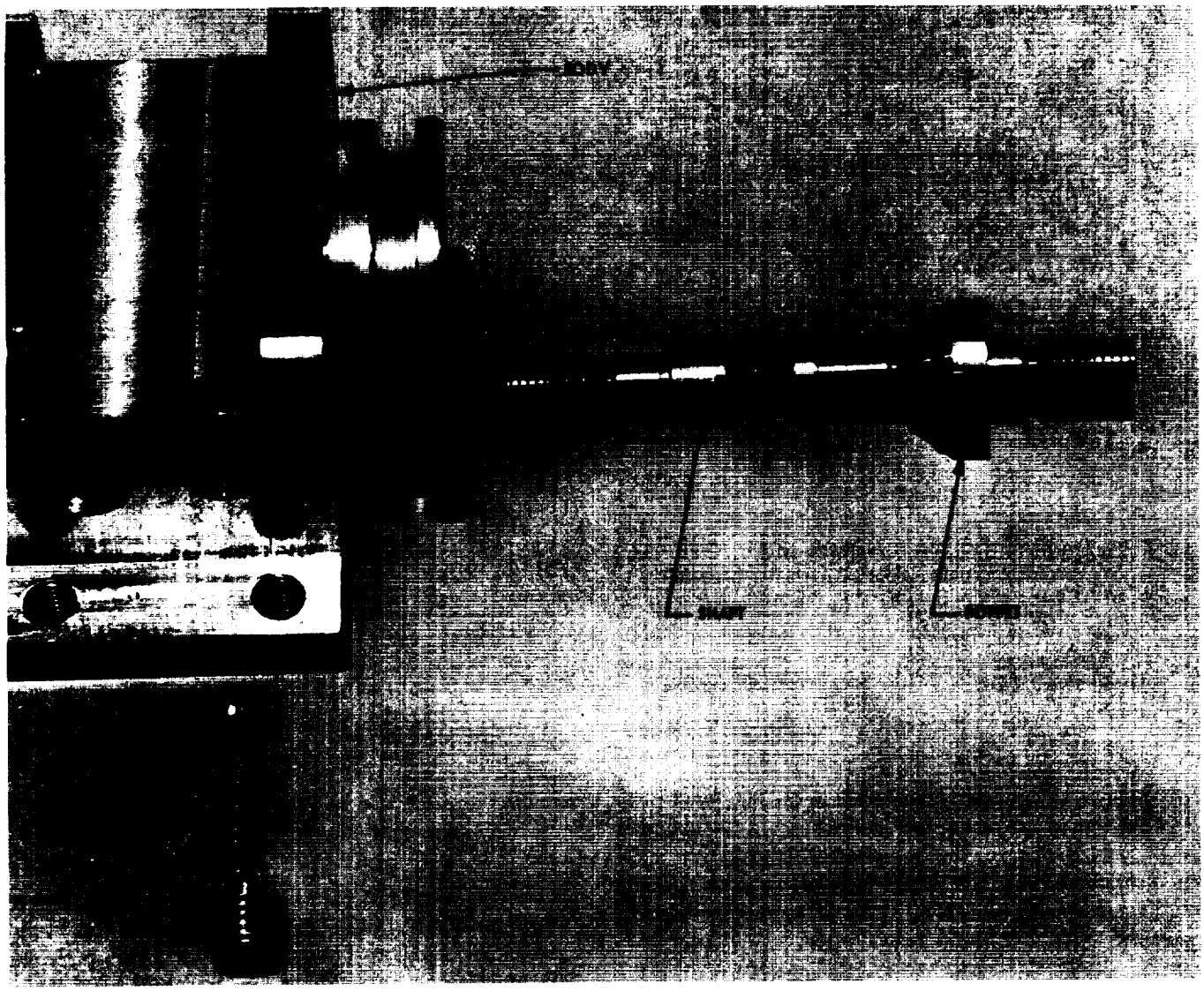


Fig. 4. Gas valve shaft and body

(4) JPL vacuum chamber (1967).

(5) Successful balloon flight from Page, Arizona (1967).

(6) JPL vacuum chamber (1968).

(7) Successful balloon flight from Palestine, Texas (1968).

5. Conclusion

The development of this valve proved to be a good investment. It fulfilled the requirements of reliability, operation in a balloon flight environment, and low power consumption. During the last 3 yr, it has performed to expectations in thousands of hours of tests and has supported two successful balloon flights.

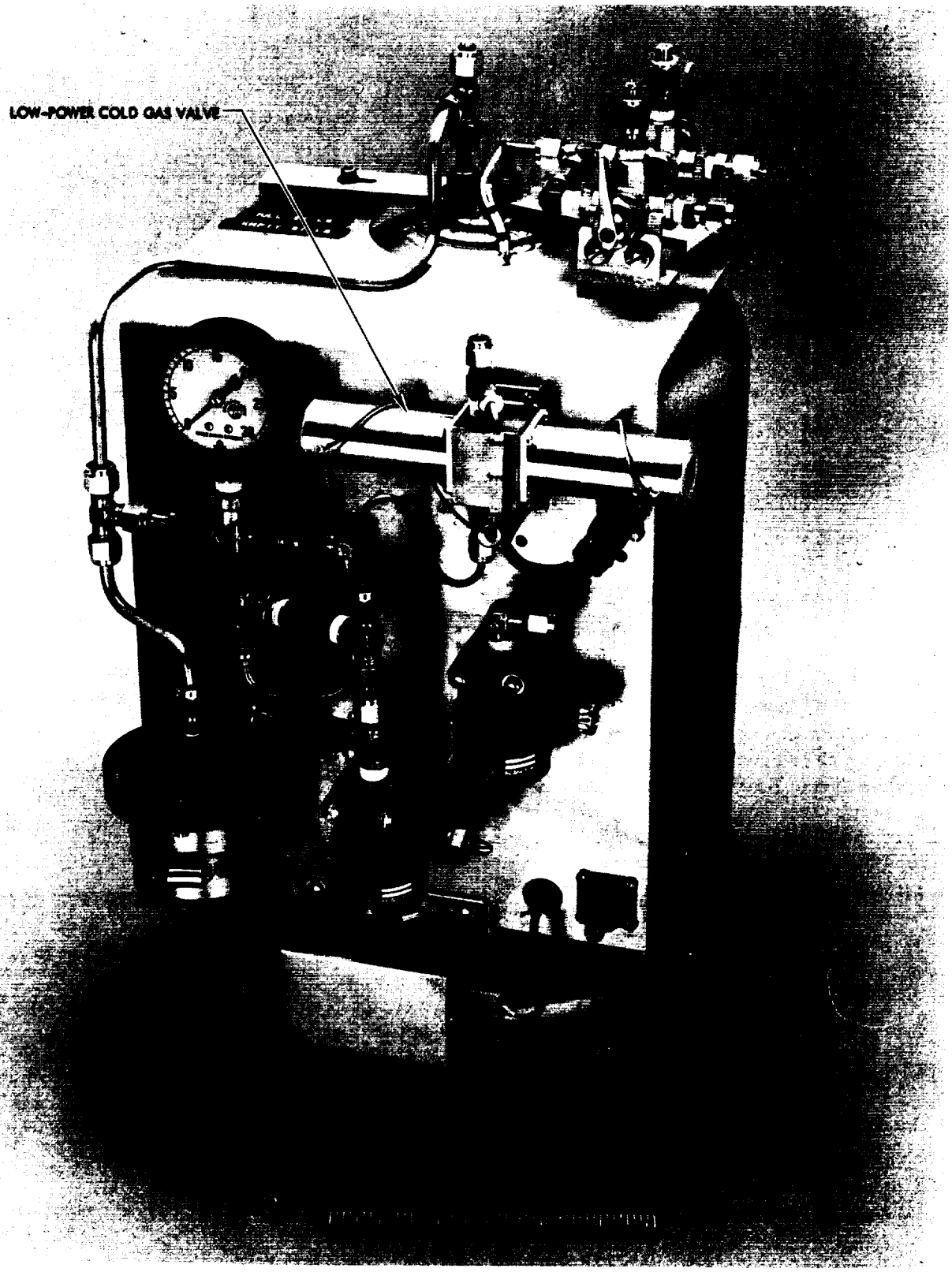


Fig. 5. Balloon-flight liquid-nitrogen dewar with cold gas valve installed

III. Science Data Systems

SPACE SCIENCES DIVISION

A. Image Recording Systems for the Electron Microscope, *J. J. Volkoff*

1. Introduction

The image in an electron microscope can be recorded electro-optically or photographically. In the selection of a recording system, its performance in the image domain must be determined. Resolution, noise, gray-scale reproduction, and dynamic range are some parameters which define the recording capabilities of a system. Where resolvability of an image is critical, a comparative analysis of recording systems may best be made with modulation transfer functions (MTF). The MTF must include the effect of object contrast.

The objective of this analysis is to determine the degree to which a recording system can resolve the object magnified on the electron microscope screen and to estimate the quality of the reproduced image.

2. Design Criteria

The electron microscope is to image unstained organic crystals to resolve atomic structure. These crystals are heat-sensitive, which limits the exposure to illumination. When these crystals are to be viewed through the electron microscope, the resulting image is not only dim but of low contrast. The brightness level is assumed to correspond to the level of image brightness on the electron microscope screen at which an image can just be resolved. It is estimated that this level results when an electron flux onto the electron microscope screen is 0.1 electron/μm^2-s. The object contrast is estimated to be 0.10. The maximum exposure period of 10 s is assumed for the recording system.

a. Resolution. The desired resolution of the object is 1 Å. For an object magnification of 500,000, the image resolution requirement is 20 cycles/mm or 40 picture elements/mm.

An electron microscope image size of 50×37.5mm (Ref. 1) at a spacial frequency of 20 cycles/mm corresponds to 1500 TV lines/frame. The required contrast ratio for the eye to just resolve a 1500-TV-line image displayed on a monitor screen must be greater than 0.30 at an image brightness of 10 ft-L (Ref. 1). By reducing the number of picture elements displayed, the required contrast ratio can be reduced. This is shown in Fig. 1 for three displayed image brightness levels. The resolution sensitivity shown by Fig. 1 was determined for average viewing conditions.

b. Luminance of the electron microscope faceplate. The threshold level of the electron flux upon the phosphor of the electron microscope screen is assumed as 0.1 electron/μm^2-s. Since there are 1.6×10^{-19} C/electron, this flux represents a current density of 1.6×10^{-12} A/cm^2.

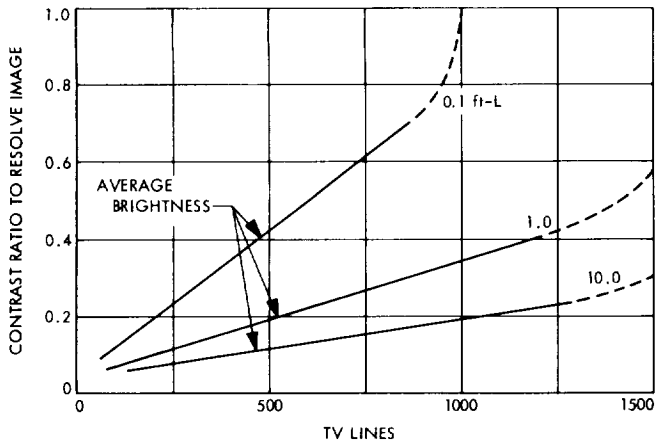


Fig. 1. Resolution sensitivity of human vision

When the electrons are accelerated through a potential of 100 kV, a luminous excitance M_v of 1.6×10^{-7} W/cm^2 is imparted to the phosphor.

The luminous efficacy K for a representative P-11 phosphor ranges from 25 to 50 lm/W (Ref. 1). The luminous efficacy for P-20 phosphor is greater than that for P-11 phosphor. Even though efficacy levels can be increased with larger phosphor grain size, the graininess of the image also is increased, which results in a degradation of resolution. The grain size of phosphor used on electron microscope screens usually vary from 5 to 10 μm. But for very high-resolution requirements, a grain size of about 3.5 μm is specified. Hence, for a high-resolution P-20 phosphor, a representative value for the luminous efficacy may be 25 lm/W. Assuming the light transmission efficiency η of the fiber optics of the electron microscope screen to be 0.8 (Ref. 3), then the threshold luminance is $M_v \cdot K \cdot \eta = 3.2 \times 10^{-6}$ cd/cm^2, or about 3×10^{-3} cd/ft^2. Thus, an illumination of 3×10^{-3} lm/ft^2 is the level at which the recording systems must operate.

3. Electro-Optical Systems

Four electro-optical camera systems shown in Fig. 2 are selected as possibly being capable of producing a resolvable image at the threshold illumination level available. These are an image orthicon (MgO), a single-stage image-intensifier coupled to a secondary electron conduction (SEC) vidicon, a three-stage image-intensifier coupled to a high-quality vidicon, and an intensifier-SEC vidicon. It is estimated that the selected maximum intensification is adequate to cause the respective camera tubes to operate at their saturation point, corresponding normally to the highest signal-to-noise ratio.

Since the transformed electron image passes through the fiber-optics screen, the effect which the screen has on the system MTF must be considered. The MTF for various optical devices are shown in Fig. 3a. These curves can be used with similar MTF curves for various electro-optical systems. The MTF curves shown in Fig. 3b are for single, dual, and three-stage image-intensifier tubes having fiber-optic faceplates. Similarly shown in Fig. 3c are the MTF curves for the four described camera tubes. The curves of Figs. 3a, b, and c are applicable for an image contrast of 1.0. By multiplying the respective modulation transfer factors of each component, the overall MTF for the total system can be obtained.

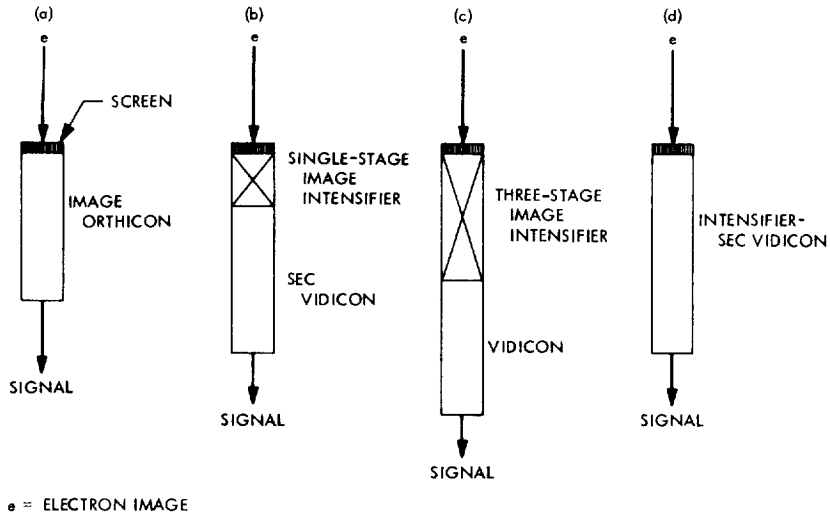


Fig. 2. Four electro-optical recording systems

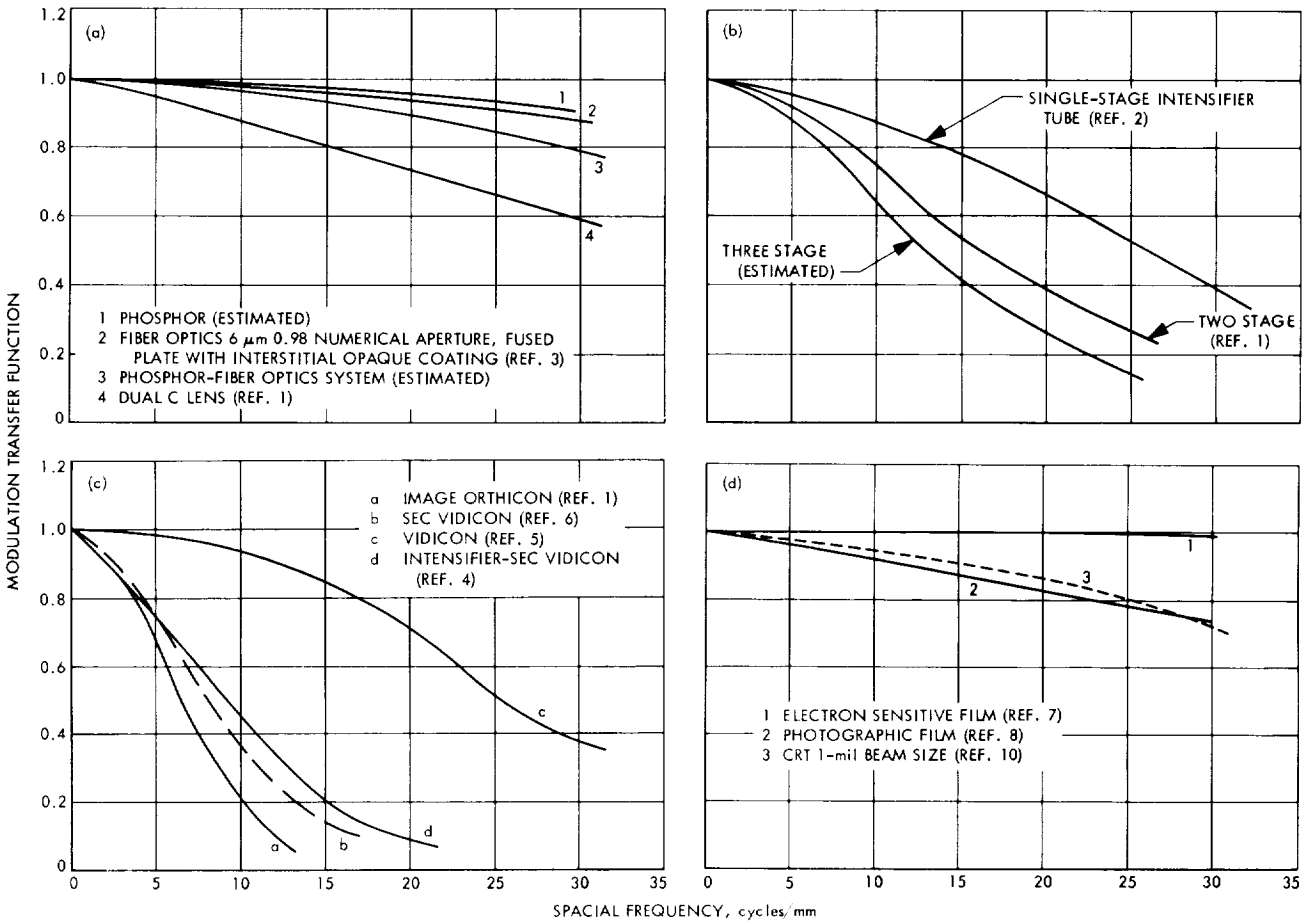


Fig. 3. Modulation transfer function for: (a) various optical devices, (b) various image intensifier tubes, (c) various camera tubes, (d) film and CRT display systems

4. Graphic Systems

The image can be reproduced unto film by exposing the film with electrons as shown in Fig. 4a, or by exposing the film with photons as shown in Figs. 4b and c. The electron microscope screen is not included in electron-beam recording of the film, whereas the film for the other systems (Figs. 4b and c), must accept the light emanating from the electron microscope screen.

a. Light exposure. The luminous excitance of light available from the electron microscope screen at the threshold illuminance of 3×10^{-3} lm/ft^2 is 10^{-8} W/cm^2. A representative value for the luminous efficacy of the phosphor to be used is assumed to be 300 lm/W for the emanating light. An exposure of 4×10^{-3} m-cd-s is required to achieve a density of 1 for a high-speed film emulsion. This exposure is equivalent to about 12×10^{-3} ft-cd of illumination for an exposure period of 1 s. Hence 4 s is required to adequately expose this film. However, if a slower emulsion is used to reduce graininess, an image-intensifier can be used between the electron microscope screen and the film as shown in Fig. 4b and c to increase the light level as required by the emulsion, but with a loss of resolution due to the intensifier.

b. Electron exposure. At threshold level, 0.10 electron/μm^2-s is the electron flux onto the phosphor. This flux is equivalent to 10^7 electrons/cm^2-s. For the desired resolution of 40 picture elements/mm, this flux is equal to 62.5 electrons/picture-elements second. Assuming that the maximum exposure period is 10 s, the total electron exposure density is 10^8 electrons/cm^2. At this exposure an optimum film density on a medium-grain emulsion can be obtained.

c. MTF curves. The MTF curves for the two graphic recording methods are shown in Fig. 3d. These curves

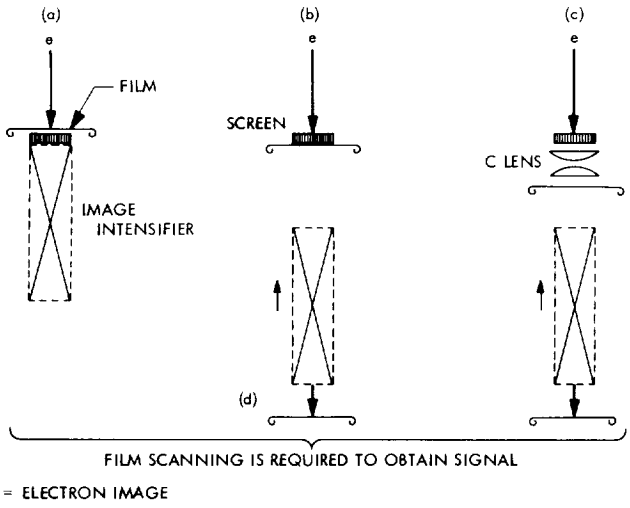


Fig. 4. Graphic recording systems

are applicable for an image contrast of 1.0. Included in Fig. 3d is the MTF for a cathode ray tube (CRT) display system.

The spread function of electron-exposed images at 100 kV has not been published. However, it is believed to be less than that for light-exposed images. Even though the MTF is reduced at increased accelerating voltages in electron-beam recording, the reduction is not significant at resolutions less than 25 cycles/mm.

d. Image focusing. Visual focusing of the image for the film recording systems is probably best performed with a three-stage image intensifier coupled to the electron microscope screen as schematically shown in Fig. 4. The response for this intensification can be obtained from Fig. 3b. At a resolution of 20 cycles/mm the MTF is 0.26; but for an image contrast of 0.10, the resultant contrast is 0.026. When the electron microscope image is intensified, an image brightness of about 370 ft-cd results on the phosphor faceplate of the intensifier. The required image contrast for the unaided eye to just resolve 20 cycles/mm when viewed at this brightness level and at a distance of 14 in. (Ref. 9) is 1.0. By magnifying the image to 20X, the image brightness is reduced to about 0.5 ft-cd, but the image contrast required is also reduced to a level estimated to be 0.006. Hence the image can be resolved; therefore, focusing the image by means of an image-intensifier system appears to be feasible.

5. Results

A comparison of the response capabilities of the electro-optical and graphic recording systems can be made on a monitor-displayed image basis. To display the film containing the reproduced image, a CRT-film scanner-display system may be used.

The MTF curves for the various recording systems are shown in Fig. 5 for an object contrast of 1.0. The response curves for the graphic recording systems include the MTF of the CRT display system. The response of the graphic recording systems as shown in Fig. 5 exceeds the response of electro-optical systems by a factor greater than three at a spacial frequency of 20 cycles/mm.

The response for each recording system at a resolution of 20 cycles/mm and object contrast of 0.10 is shown in Table 1. The three-stage image-intensifier/vidicon system has the highest response and reproduced image quality of all of the four candidate camera tube systems. Assuming that the displayed image from this camera tube system is at 20 ft-L, the required threshold contrast (Ref. 9) to just resolve the image at this luminescence and at a viewing distance of 10 in. is about 0.23. It is very unlikely that any of the electro-optical recording systems are capable of reproducing a resolvable displayed image of the entire object field as shown on the electron

CURVE[1]	DESCRIPTION	SYSTEM TYPE
2a	IMAGE ORTHICON (MgO)	ELECTRO-OPTICAL
2b	SINGLE-STAGE IMAGE-INTENSIFIER/ SEC VIDICON	
2c	THREE-STAGE IMAGE-INTENSIFIER/ VIDICON	
2d	INTENSIFIER-SEC VIDICON	
4a	ELECTRON BEAM RECORDING ON FILM	GRAPHIC
4b	ELECTRON-MICROSCOPE SCREEN SURFACE-CONTACT PHOTOGRAPHING	
4c	LENS COUPLING PHOTOGRAPHIC FILM	
4d	IMAGE-INTENSIFIER/PHOTOGRAPHIC FILM	

[1]CURVE NUMBERS CORRESPOND TO CONFIGURATION NUMBERS OF FIGS. 2 AND 4

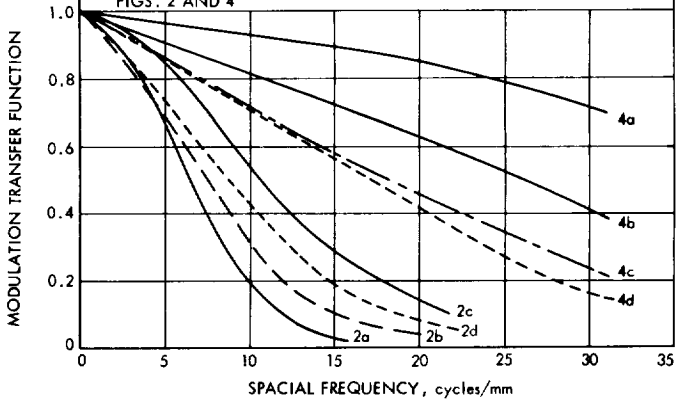


Fig. 5. Modulation transfer function for the various recording systems

Table 1. Performance of the various image-recording systems

Type	Recording system	Configuration	SNR[a] at high light	Response at resolution of 20 cycles/mm object contrast = 0.10
Electro optical	Image orthicon (MgO)	Fig. 2a	20[a]	0.002
	Single-stage image-intensifier/SEC vidicon	Fig. 2b	40[a]	0.004
	Three-stage image-intensifier/ vidicon	Fig. 2c	80[a]	0.014
	Intensifier-SEC vidicon	Fig. 2d	43[a]	0.008
Graphic	Electron-beam film recording	Fig. 4a	150[b]	0.085
	Surface contact photographing	Fig. 4b	160[a,b]	0.063
	Lens coupling photographic film	Fig. 4c	160[a,b]	0.046
	Image-intensifier —photographic film	Fig. 4d	140[a,b]	0.042

[a]Signal-to-noise ratio does not include phosphor noise.
[b]Granularity and quantum efficiency dependent.

microscope screen. However, by recording the image and displaying about a fifteenth of the total picture elements on a monitor, the resulting displayed image can just be resolved with a three-stage image-intensifier/vidicon system.

The responses of the photographic systems shown in Table 1 are also lower than the threshold contrast of 0.23 required to resolve the displayed image. By displaying about a quarter of the total picture elements, a just resolvable displayed image can be obtained with the image-intensifier photographic recording system, configuration Fig. 5 (curve 4d). The most responsive recording system is electron-beam recording. One-third of the total picture elements displayed from film containing the electron-beam image can be resolved on a monitor.

Estimates of image quality for the various recording systems are shown in Table 1 by the overall signal-to-noise ratio. The image quality of the electron-beam film recording system is higher than any of the other systems. It is the only system which is not affected by phosphor noise of the electron microscope screen. The electron beam recorded image can be reliably digitized to 7 binary bits, and all the photographic images can be digitized most probably to 6 binary bits. However, the highest digitizing level which the three-stage image-intensifier/vidicon system is capable of is most probably 5 binary bits.

6. Conclusion

Recording the image of an electron microscope which is to resolve 1 Å of an object is most feasible with electron-beam recording on film. Focusing the image for this system can be performed with a three-stage image-intensifier coupled to the electron microscope screen.

Of the electro-optical recording systems considered, the highest responsive system is the image-intensifier coupled to a vidicon camera tube. A displayed portion of the total image having about 500 TV lines can be just resolved on a monitor screen with this camera tube configuration. The quality of the displayed image from this system is not as high as that which may be obtained by recording the image on film. This system can be used for focusing the image in conjunction with electron-beam recording.

References

1. Sadashige, K., "Image Intensifier TV Camera System, Its Performance and Applications," *Appl. Opt.*, Vol. 6, No. 12, December, 1967.
2. *40–25mm Image Intensifier*, Westinghouse Electric Corp., WX-30677, 1968.
3. Kapany, N. S., *Fiber Optics, Principles and Applications*, Academic Press, New York, 1967.
4. *Intensifier SEC Camera Tube, WL-32000*, Westinghouse Electric Corp., Bulletin TD 86-827, 1968.
5. *Vidicons*, Data Sheet ETR-3837A, General Electric Corp., 1964.
6. *SEC Camera Tube, WL-30691*, Bulletin TD-86-817, Westinghouse Electric Corp., 1968.
7. Soule, H. V., *Electro-Optical Photography at Low Illumination Levels*, John Wiley & Sons, Inc., New York, 1968.
8. Kingslake, R., *Applied Optics and Optical Engineering, VII*, Academic Press, New York, 1965.
9. DePalma, J. J., and Lowry, E. M., "Sine-Wave Response of the Visual System II. Sine-Wave and Square-Wave Contrast Sensitivity," *J. Opt. Soc. Amer.*, Vol. 52, No. 3, March, 1962.
10. Potter, R. J., "An Optical Character Scanner," *SPIE J.*, Vol. 2, February–March 1969.

IV. Physics

SPACE SCIENCES DIVISION

A. The Magnetic Neutral Point and Flux Reconnection, *A. Bratenahl and C. M. Yeates*

For more than twenty years the peculiar dynamics that occur at X-type magnetic neutral points have provided the most persistent and suggestive theoretical basis for the solar flare. More recently this dynamical process of severing and reconnecting field lines has been introduced to explain many other phenomena, e.g., the astrophysical dynamo, the open magnetosphere, and plasma injection. Developing the adequate theory for the neutral point process, however, has proved exasperatingly difficult. The difficulty arises because there are a large number of possible effects to consider, and their relative importance depends on how they enter the "mix." A particular mix depends on how the process is carried out; that is, besides plasma conditions, it depends on initial conditions and boundary conditions. For some time, it has appeared essential to examine at least one pathway in detail, experimentally,[1] and thereby furnish a guide for future theoretical development. The double inverse pinch experiment has proved to be an informative method of approach. We summarize our results here; a more complete account will be presented elsewhere[2].

The double inverse pinch device sets up a double source plane hydromagnetic flow, generally in argon plasma at 100 μm pressure. There is a stagnation point where the flows impinge at the center of the device and the stagnation point coincides with a magnetic neutral point. Since the neutral point is of the X-type, it lies on and forms a part of the separatrix of the field line system. The separatrix is of the form of a figure eight and thus divides the flux into three cells within which the field lines encircle one or the other or both of the source current rods. Near the X-point the separatrix divides the flow into pairs of inflow and outflow sectors.

The schlieren photographs (Fig. 1), when combined with other data, show that the double inverse pinch in preionized argon consists of a pair of wave heads of

[1]SPS 37-34, Vol. IV, pp. 215–218; SPS 37-37, Vol. IV, pp. 185–186; SPS 37-52, Vol. III, pp. 176–182.

[2]Bratenahl, A., and Yeates, C. M., "Experimental Study of Magnetic Flux Transfer at the Hyperbolic Neutral Point," submitted to *The Physics of Fluids*.

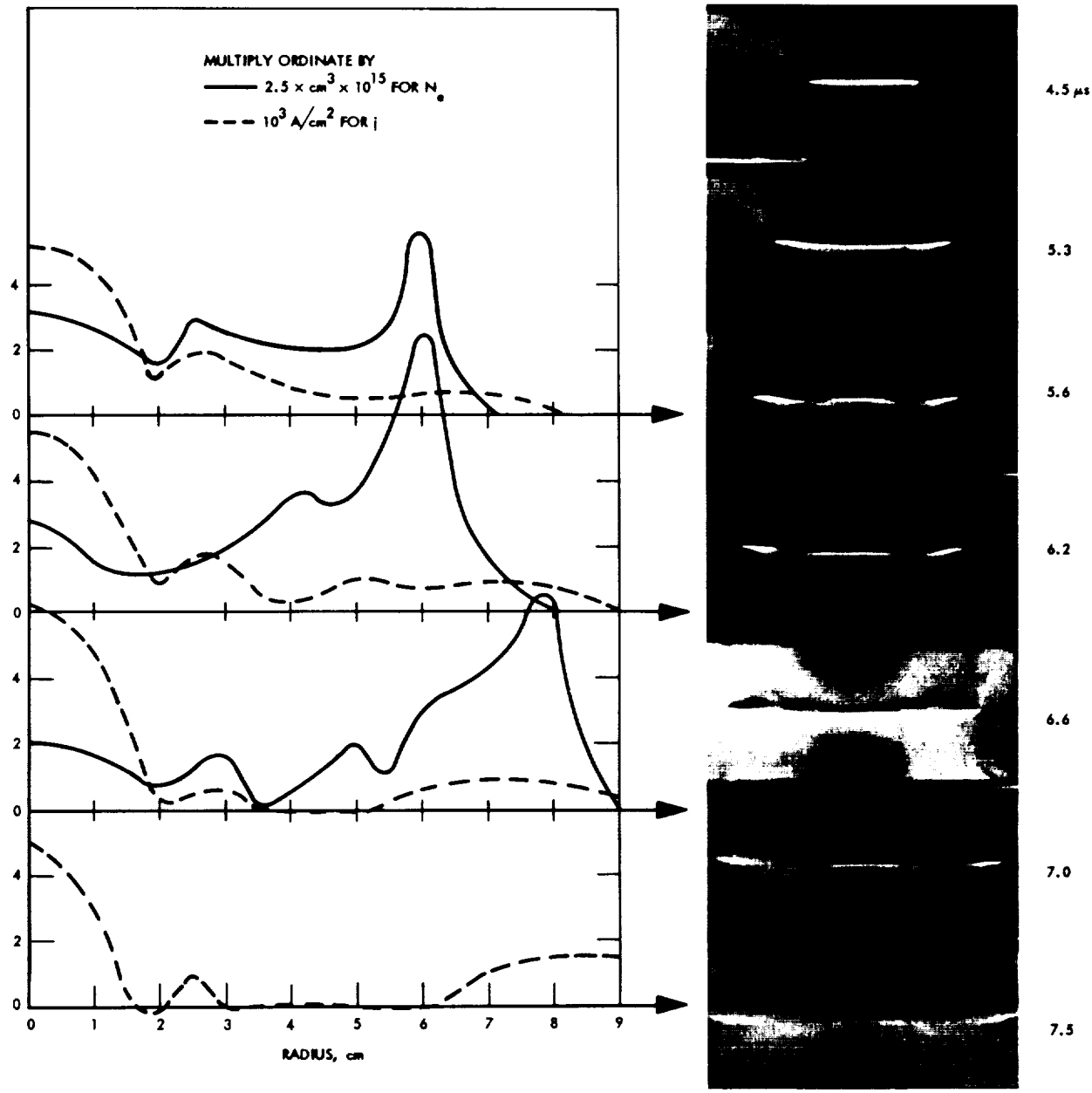


Fig. 1. Sequence of schlieren photographs compared with electron density profiles along the y-axis

enhanced density 2 cm thick, followed by tails of diminishing density extending back to the current rods. Following collision, the wave systems interact to form a high-density collision layer which is then vigorously ejected along its length by the Lorentz force. Clearly this double-source flow is supersonic but sub-Alfvénic. Consequently, a shock-like transition is required where the flows converge on the X-point at which the magnetic field B and the velocity u vanish. It is also required to redirect the flows converging toward the y-axis into flows diverging away from the x-axis. As the post-collision time increases, it becomes clear that the collision layer is chiefly wave-head plasma, the tail plasma being turned through shocks located just downstream of the separatrix. These shocks may be seen branching out to the sides of the collision layer. Later their density gradients drop below our detection threshold (at about 7.0 μs). During their period of visibility, the flow downstream is an expansion fan, forming an extension of the same Lorentz force ejection process already mentioned. Figure 2 shows,

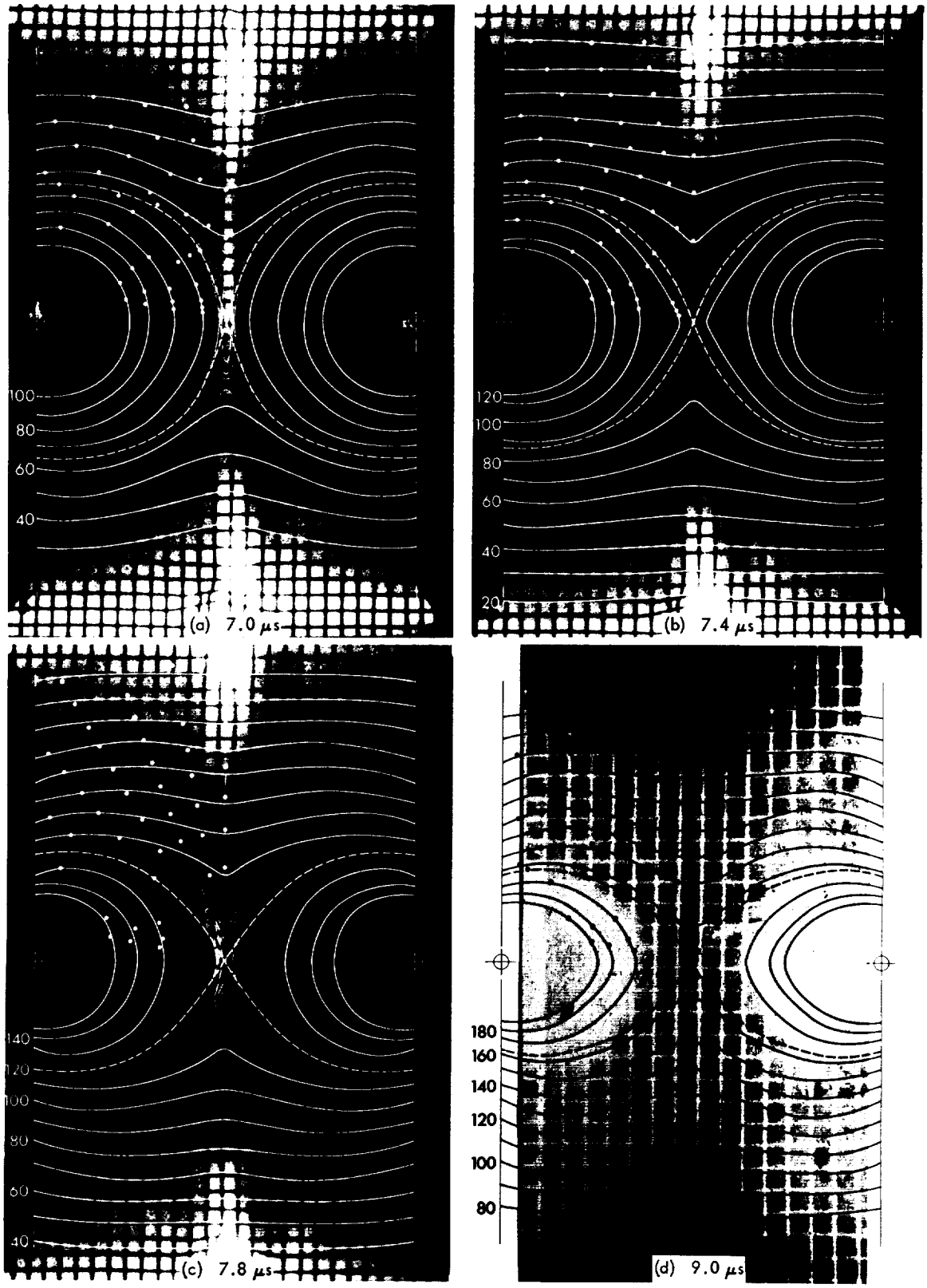


Fig. 2. Superposition of field line maps and contours of constant current density on Kerr-cell photographs, at the indicated times; (a) is reversed to bring out the coincidence of faint luminosity with the separatrix (dashed curve)

in a superposition of Kerr-cell photographs, field line maps and contours of constant current density, that the current and luminosity are "switched on" close to the separatrix. Current contours are 0.5, 1, 2, 3, and $4 \times A \times 10^3/cm^2$; field lines carry value of vector potential in $Wb \times 10^{-4}/m$. Also, when the shocks lose schlieren visibility, the current becomes progressively concentrated along lines which detailed calculations prove to be shock loci. These stationary shocks are of the slow-mode type, close to the switch-off limit. (We have been cautious in announcing their appearance here, as it is our understanding that they have never been seen before, outside of theory.) At 7.4 μs the configuration bears a striking resemblance to Petschek's proposal (Ref. 1).

Accurate measurement of the current density at the X-point was achieved by means of a miniature 200-turn toroidal coil probe, 0.140 and 0.017 in. major and minor diameters, respectively. Used in conjunction with the flux probe, it became immediately clear that during the 4 μs characterized by Petschek's reconnection mode with its switch-off shocks, the X-point current and electric field are strictly proportional, satisfying the normal ohmic relation. Therefore, the flux transfer rate at this time is determined by resistive diffusion in a small (hyperbolic pinch) region whose dimensions are controlled by the positions taken up by the slow-mode shocks.

However, at 7.8 μs this ohmic relation fails. In a time less than 0.3 μs, the 7-kA current density is cut off, and the electric field increases by a factor of three. In that short time, therefore, the resistivity increases more than an order of magnitude; the electric field, enabling an impulsive flux-transfer process to take place, initiates a fast-mode large-amplitude wave disturbance system which effects the redistribution of flux between the cells. Detailed study of conditions leading up to this anomalous behavior of the resistivity η (Fig. 3) provides unmistakable evidence that as the current density j_x increases, the plasma density N_e decreases with the result that the electron drift velocity V_d increases, approaching the thermal agitation speed $(kT/m_e)^{1/2}$. That this is a critical condition leading to instability is well known.

It turns out that all the factors producing this critical condition are readily determined in the experimental arrangement. We have available, therefore, a facility which not only presents the phenomenon of anomalous resistivity in an exceptionally "pure" form, but also is particularly convenient for detailed diagnostic studies.

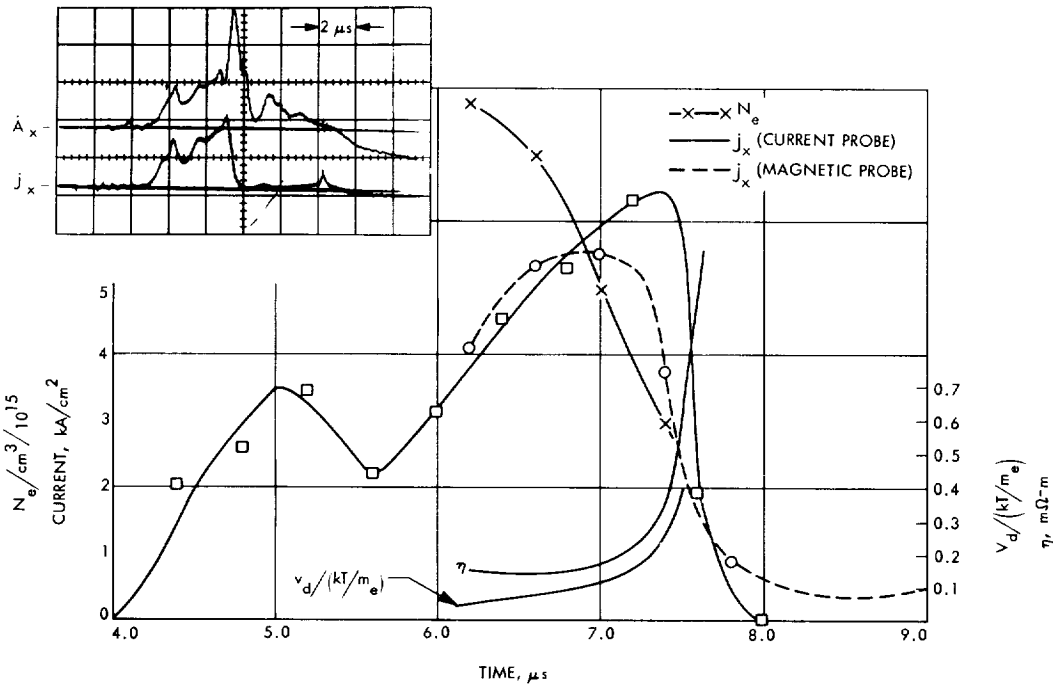


Fig. 3. Conditions at X-point, current density, resistivity, electron density, and ratio of electron drift velocity to thermal agitation speed. Inset shows typical behavior (upper trace is rate of change of vector potential A_r), and the lower trace is the current density j_r at the neutral point

Such studies could be of great value, as there are many controversies among the various versions of present-day theory, and there is an urgent and widespread interest in this basic problem of plasma physics.

It is noteworthy that Friedman and Hamberger (Ref. 2) predicted that Petschek's steady-state reconnection mode would inevitably culminate in such a resistivity instability, and our experiment now confirms both of these notions in a clear-cut fashion. The implications to flare theory should be obvious. The enhanced flux-transfer rate in our experiment is four times the steady-state rate, and is flux-transported-limited rather than resistive-diffusion-limited. This is because flux is transported at the fast-mode wave speed, and in our experiment, this speed is found to be 16 cm/μs at its slowest point (bottleneck), whereas the enhanced resistive diffusion through the hyperbolic pinch at the X-point is at least 40 cm/μs, offering no impediment by comparison. Prior to the resistance anomaly, resistive diffusion in the (Petschek) steady-state mode is found to be only 4 cm/μs. We assert the enhanced rate of flux-transfer that we have identified is the ultra-fast rate: its direct dependence on the fast-mode wave speed in conducting plasma can only be exceeded by the speed of light in a nonconducting medium.

Energy calculations have been made. The enhanced flux-transfer rate corresponds to a sudden increase in load inductance for the capacitor bank energy source. At the moment of onset, the source is supplying 150 MW. The event induces an additional surge of 15 MW, lasting for about 0.5 μs.

We believe that the results constitute a significant advance toward the solution of a nonlinear nonlocal whole-flow problem that hitherto has been unresolved theoretically. The experiment has demonstrated nature's solution to a very basic mechanism in plasma dynamics which has widespread application in both space research and plasma technology.

The results provide a basis for further related studies, notably in the areas of anomalous resistivity, microwave emission, X-rays, and high-energy particle acceleration, all of which have direct bearing on both the flare application and plasma technology in general. At the same time, the experiment has provided new insight into the role played by the structural morphology of magnetic fields in defining possible synoptic sequences leading to solar flares and other manifestations of solar activity.

References

1. Petschek, H. E., *AAS–NASA Symposium on the Physics of Solar Flares,* NASA SP-50, p. 425, National Aeronautics and Space Administration, Washington, D.C., 1964.
2. Friedman, M., and Hamberger, S. M., *Astrophys. J.*, Vol. 152, p. 667, 1968.

V. Communications Systems Research

TELECOMMUNICATIONS DIVISION

A. Combinatorial Communication: Quadratic Forms Over Finite Fields and Second-Order Reed–Muller Codes, *R. J. McEliece*

1. Introduction

Let $V_m(q)$ be the m-dimensional vector space of m-lists from $GF(q)$. If F_r is the set of all polynomials of degree r or less in m variables over $GF(q)$, the rth-order generalized Reed–Muller (GRM) code over $GF(q)$ is defined to be the code of block length q^m whose codewords are the truth tables of the polynomials in F_r. That is, if

$$v_0, v_1, \cdots, v_{q^m-1}$$

is an ordering of $V_m(q)$, a typical codeword in the rth-order GRM code is

$$\left[f(v_0), f(v_1), \cdots, f\left(v_{q^m-1}\right)\right], \qquad f \in F_r$$

It is the object of this article to investigate in close detail the second-order GRM code, and in particular to compute its complete weight spectrum. This has already been done by Berlekamp and Sloane (unpublished) for the case $q = 2$. And it turns out that much of the work has already been done by Dickson (Ref. 1) in relation to his investigation of quadratic forms over finite fields. Another object of this article is to give as far as possible a simultaneous treatment of the cases q odd and q even; it is the author's belief that this distinction has been overemphasized in the past.

The major results to be derived are these: First of all, if $f(x_1, \cdots, x_m)$ is a polynomial of degree ≤ 2 over $GF(q)$, the number of solutions to $f(x_1, \cdots, x_m) = a$ is of the form $q^{m-1} + \nu q^{m-j-1}$ where $\nu = 0$, ± 1, or $\pm(q-1)$ and $0 \leq j \leq (m/2)$. We shall derive a set of canonial quadratic forms over $GF(q)$, and obtain formulas for the number of quadratic forms of each rank and type. Finally, we shall present complete weight spectra for the second-order GRM code and a code closely related to it.

2. Canonical Quadratic Forms Over Finite Fields

By a *quadratic form* (QF) *in m variables* over a field F, we mean a polynomial homogeneous of degree two over F;

$$Q(x_1, \cdots, x_m) = \sum_{i,j=1}^{m} a_{ij} x_i x_j, \qquad a_{ij} \in F$$

Two forms Q and Q' are said to be *equivalent* if there is a nonsingular change of variables

$$x_i = \sum_k \beta_{ik} x'_k$$

which transforms Q into Q'. Q is *nonsingular* if it is not equivalent to a form in fewer than m variables, and more generally the *rank* of a QF is the least number of variables in which it can be expressed by a nonsingular change of variables. Finally, a QF is said to *represent zero* if there is an m-list $(\xi_1, \xi_2, \cdots, \xi_m) \neq (0, 0, \cdots, 0)$ such that $Q(\xi_1, \xi_2, \cdots, \xi_m) = 0$. The following theorem is valid over any field, finite or not.

Theorem 1. If a nonsingular QF $Q(x_1, x_2, \cdots, x_m)$ represents zero, then Q is equivalent to a form of the shape

$$x_1 x_2 + Q'(x_3, x_4, \cdots, x_m)$$

Proof. If $Q(\xi_1, \xi_2, \cdots, \xi_m) = 0$, and not all of the ξ_i are zero, it is possible to find a nonsingular transformation of the form

$$x_i = \xi_i x'_1 + \cdots \qquad \text{for all } i$$

which forces the new coefficient of x_1^2 to be zero. Then not all of the (new) coefficients a_{1j} can be zero since Q is nonsingular, and with a suitable renumbering we may assume $a_{12} \neq 0$. Then the transformation

$$x'_2 = \Sigma\, a_{1j} x_j, \qquad x'_i = x_i, i \neq 2$$

puts Q into the promised form. QED

Corollary. Every nonsingular QF is equivalent to one of the shape

$$x_1 x_2 + x_3 x_4 + \cdots + x_{2s-1} x_{2s} + Q'(x_{2s+1}, \cdots, x_m)$$

where Q' does not represent zero.

Thus, Theorem 1 reduces the question of finding a set of canonical quadratic forms to the same question for forms which do not represent zero. While for some fields (e.g., the rationals) this is not much help, for finite fields it is very useful, because of Theorem 2, first proved by Dickson, and later greatly generalized by Chevalley and others.

Theorem 2. Every nonsingular QF in $\geqq 3$ variables over a finite field represents zero.

Proof. If $Q(x_1, \cdots, x_m)$ has the unique solution $(0, 0, \cdots, 0)$, then the polynomial $1 - Q(x_1, \cdots, x_m)^{q-1}$ must be the same as the polynomial

$$(1 - x_1^{q-1}) \cdots (1 - x_m^{q-1})$$

since they have the same value for each choice of the x's. But the degree of $1 - Q^{q-1}$ is $\leq 2(q-1)$, while the degree of $(1 - x_1^{q-1}) \cdots (1 - x_m^{q-1})$ is $m(q-1)$, a conflict if $m \geqq 3$. QED

Thus, every QF over a finite field which does not represent zero is equivalent to a QF in zero, one, or two variables with the same property. The only QF in one variable is $ax^2, a \in GF(q)$. If q is even, then every element in $GF(q)$ is a square and ax^2 is equivalent to x^2. If q is odd and λ is a particular nonsquare in $GF(q)$, then every ax^2 is equivalent to either x^2 or λx^2, depending on whether a is a square or not. For QF's in two variables, we need to go into more detail.

Theorem 3. Every nonsingular QF in two variables which does not represent zero is equivalent to one of:

$$x^2 - \lambda y^2 \text{ if } q \text{ is odd}, \qquad \lambda \text{ is not a square}$$

$$\lambda x^2 + xy + \lambda y^2 \text{ if } q \text{ is even}, \qquad \text{trace}(\lambda) = 1$$

Proof. Let $Q(x, y) = Ax^2 + Bxy + Cy^2$. For odd q, if $A = 0$, Q represents zero. If $A \neq 0$, the transformation

$$x = x' - (B/2A)\, y', \qquad y = y'$$

diagonalizes Q into $\alpha x^2 + \beta y^2$ with $\alpha = A$, $\beta = C - B^2/4A$. If α and $-\beta$ both have the same quadratic character, Q represents zero since it can be written $\gamma(x^2 - y^2)$. Thus, we may assume that α is a square and $-\beta$ is a nonsquare. Then $\alpha x^2 + \beta y^2$ is equivalent to $x^2 - \lambda y^2$, as asserted.

If q is even, we may assume $ABC \neq 0$ since otherwise Q obviously represents zero. Then the substitution

$$x = \left(\frac{C}{AB^2}\right)^{1/4} x', \qquad y = \left(\frac{A}{CB^2}\right)^{1/4} y'$$

puts Q in the form $\alpha x^2 + xy + \alpha y^2$. If trace $(\alpha) = 0$, then there is a solution z to $z^2 + z = \alpha$ and $Q(z/\alpha, 1) = 0$. If trace $(\lambda) = 1$ as well, and if $u^2 + u + \alpha^2 = \lambda^2$, then the substitution

$$x = \left(\frac{\beta}{a}\right)^{1/2} x' + \frac{u}{(a\beta)^{1/2}} y', \qquad y = \left(\frac{\alpha}{\beta}\right)^{1/2} y'$$

changes $\alpha x^2 + xy + \alpha y^2$ into $\lambda x^2 + xy + \lambda y^2$. QED

We are, therefore, led to Table 1, which gives the canonical quadratic forms over $GF(q)$.

From now on, the parameter λ will be reserved to identify the *type* of a QF, as per Table 1. Furthermore, for even m we will frequently associate a sign $\epsilon = \pm 1$ with the two types: $\epsilon = +1$ for QF's with $\lambda = 1$ (odd q) or $\lambda = 0$ (even q), $\epsilon = -1$ for QF's with $\lambda =$ nonsquare (odd q) or trace $(\lambda) = 1$ (even q).

3. The Finite Orthogonal Group and the Number of QF's of Each Rank

There are $q^{m+\binom{m}{2}} = q^{m(m+1)/2}$ quadratic forms in m variables over $GF(q)$. The goal of this subsection is to compute F_k^λ, the number of forms of rank k and type λ, for each $0 \leq k \leq m$. From elementary group theory, we know that the number of QF's equivalent to a given QF Q is equal to the number of nonsingular linear transformations (LT's) divided by the number of LT's which leave Q invariant. Now it is easy to see that the total number of nonsingular LT's in m variables is

$$(q^m - 1)(q^m - q) \cdots (q^m - q^{m-1}) = q^{m(m-1)/2} \prod_{i=1}^{m} (q^i - 1)$$

The calculation of the number of LT's which fix a given canonical QF is a much more difficult problem, but fortunately almost all of the work has already been done for us: Let $Q(x_1, \cdots, x_m)$ be a given canonical form of rank k, i.e., Q involves only the first k variables, and suppose the nonsingular LT

$$\begin{array}{l|l}
\begin{array}{l} x_1 = a_{11} x'_1 + \cdots + a_{1k} x'_k \\ \cdots \\ x_k = a_{k1} x'_1 + \cdots + a_{kk} x'_k \end{array} & \begin{array}{l} + a_{1,k+1} x'_{k+1} + \cdots + a_{1m} x'_m \\ \cdots \\ + a_{k,k+1} x'_{k+1} + \cdots + a_{km} x'_m \end{array} \\
\hline
\begin{array}{l} x_{k+1} = a_{k+1,1} x'_1 + \cdots + a_{k+1,k} x'_k \\ \cdots \\ x_m = a_{m1} x'_1 + \cdots + a_{m,k} x'_k \end{array} & \begin{array}{l} + a_{k+1,k+1} x'_{k+1} + \cdots + a_{k+1,m} x'_m \\ \cdots \\ + a_{m,k+1} x'_{k+1} + \cdots + a_{m,m} x'_m \end{array}
\end{array}$$

fixes Q. Then clearly:

(1) The upper right quadrant must be all zeros.

(2) The upper left quadrant must be nonsingular and must fix Q.

(3) The lower right quadrant must be nonsingular.

(4) The lower left quadrant may be arbitrary.

Now the group of nonsingular LT's which fix a nonsingular QF in k variables is called the *orthogonal group* $O_k^\lambda(q)$ where λ is the type of the form, and the orders of the various orthogonal group were first calculated by Dickson: his results are given in Table 2. Combining Table 2 with restrictions (1)–(4), we arrive at Table 3.

Table 1. Canonical nonsingular quadratic forms over GF (q)

m odd	$x_1x_2 + x_3x_4 + \cdots + x_{m-2}x_{m-1} + \lambda x_m^2$ where $\lambda = 1$ or a special nonsquare (absent for even q)
m even, q odd	$x_1x_2 + x_3x_4 + \cdots + x_{m-1}^2 - \lambda x_m^2$ where $\lambda = 1$ or a special nonsquare
m even, q even	$x_1x_2 + x_3x_4 + \cdots + x_{m-1}x_m + \lambda(x_{m-1}^2 + x_m^2)$ where $\lambda = 0$ or a special value with trace $(\lambda) = 1$

Table 2. Orders of the finite orthogonal group $O_k^\lambda(q)$

$$|O_{2j+1}^\lambda(q)| = Aq^{j^2}\prod_{i=1}^{j}(q^{2i}-1) \text{ where } A = \begin{cases}1 \text{ if } q \text{ is even}\\ 2 \text{ if } q \text{ is odd}\end{cases} \text{ (notice no dependence on } \lambda)$$

$$|O_{2j}^\lambda(q)| = 2(q^j-\epsilon)\,q^{j^2-j}\prod_{i=1}^{j-1}(q^{2i}-1) \text{ where } \epsilon = \begin{cases}+1\\ -1\end{cases} \text{ as described in Subsection 2}$$

Table 3. F_k^λ, the number of QF's of type λ and rank k in m variables over $GF(q)$

$$F_{2j+1}^\lambda = Bq^{j^2+j}\frac{\prod_{i=m-2j}^{m}(q^i-1)}{\prod_{i=1}^{i}(q^{2i}-1)} \text{ where } B = \begin{cases}1 \text{ if } q \text{ is even}\\ \tfrac{1}{2} \text{ if } q \text{ is odd}\end{cases}$$

$$F_{2j}^\lambda = \tfrac{1}{2}q^{j^2}(q^j+\epsilon)\frac{\prod_{i=m-2j+1}^{m}(q^i-1)}{\prod_{i=1}^{j}(q^{2i}-1)} \text{ where } \epsilon \text{ is defined as before}$$

As a sidelight, we notice that if T_k represents the total number of forms of rank k, we obtain

$$T_k = q^{j(j+1)}\frac{\prod_{i=m-k+1}^{m}(q^i-1)}{\prod_{i=1}^{j}(q^i-1)}$$

where $j = [k/2]$. This shows $\Sigma T_k = q^{m(m+1)/2}$. A direct proof of this would probably be very difficult.

4. Counting Solutions

We shall present in this subsection a variety of results about the number of solutions $(\xi_1, \xi_2, \cdots, \xi_m)$ of $Q(x_1, x_2, \cdots, x_m) = a$, where Q is a QF or closely related to one. In view of the canonical QF's given in Subsection 2, the following result is fundamental.

Theorem 4. The number of solutions $(\xi_1, \xi_2, \cdots, \xi_m)$ to

$$x_1x_2 + x_3x_4 + \cdots + x_{2t-1}x_{2t} = a$$

is given by

$$q^{2t-1} + (q-1)q^{t-1} \qquad \text{if } a = 0$$

$$q^{2t-1} - q^{t-1} \qquad \text{if } a \neq 0$$

Proof. We induct on t, the case $t = 1$ being easy to verify directly. Let $N_t(a)$ represent the number of solutions, and consider the equation

$$x_{2t-1}x_{2t} = a - (x_1x_2 + \cdots + x_{2t-3}x_{2t-2})$$

If $a = 0$, the right side is zero in $N_{t-1}(0)$ ways corresponding to $(2q-1)N_{t-1}(0)$ solutions, and the right side is nonzero in $(q-1)N_{t-1}(*)$ ways [where $*$ is an arbitrary nonzero element in $GF(q)$], corresponding to $(q-1)^2N_{t-1}(*)$ solutions. Hence,

$$N_t(0) = (2q-1)N_{t-1}(0) + (q-1)^2N_{t-1}(*)$$

Similarly for $a \neq 0$, we obtain

$$N_t(*) = (q-1)N_{t-1}(0) + (q^2-q+1)N_{t-1}(*)$$

In matrix notation,

$$\begin{pmatrix}N_t(0)\\ N_t(*)\end{pmatrix} = \begin{pmatrix}2q-1 & (q-1)^2\\ q-1 & q^2-q+1\end{pmatrix}\begin{pmatrix}N_{t-1}(0)\\ N_{t-1}(*)\end{pmatrix}$$

so that by induction

$$\begin{pmatrix}N_t(0)\\ N_t(*)\end{pmatrix} = \begin{pmatrix}2q-1 & (q-1)^2\\ q-1 & q^2-q+1\end{pmatrix}^{t-1}\begin{pmatrix}2q-1\\ q-1\end{pmatrix} = M^{t-1}\begin{pmatrix}2q-1\\ q-1\end{pmatrix}$$

To raise M to high powers, it is convenient to diagonalize it:

$$M = \begin{pmatrix}1 & -1\\ 1 & q-1\end{pmatrix}^{-1}\begin{pmatrix}q & 0\\ 0 & q^2\end{pmatrix}\begin{pmatrix}1 & -1\\ 1 & q-1\end{pmatrix}$$

so that

$$M^{t-1} = \begin{pmatrix}1 & -1\\ 1 & q-1\end{pmatrix}^{-1}\begin{pmatrix}q^{t-1} & 0\\ 0 & q^{2t-2}\end{pmatrix}\begin{pmatrix}1 & -1\\ 1 & q-1\end{pmatrix}$$

from which the theorem follows at once. QED

Table 4. $N(Q, a)$, the number of solutions to $Q(x_1, \cdots, x_m) = a$ (rank $Q = k$)

$$\text{even } k\text{: } N(Q,a) = \begin{cases} q^{m-1} + \epsilon(q-1)^{m-(k/2)-1} & \text{if } a = 0 \\ q^{m-1} - \epsilon q^{m-(k/2)-1} & \text{if } a \neq 0 \end{cases}$$

$$\text{odd } k\text{: } N(Q,a) = \begin{cases} q^{m-1} & \text{if } a \neq 0 \text{ and } q \text{ is even} \\ q^{m-1} \pm q^{m-(k/2)-1/2} & \text{if } a = 0 \text{ or if } q \text{ is odd} \end{cases}$$

[+ if λa is a square, − if λa is a nonsquare in $GF(q)$]

Theorem 4 leads to Table 4 without much additional work.

We shall now present some results about the number of solutions to equations $P(x_1, x_2, \cdots, x_m) = a$ where P is now a polynomial of degree two, and $P(0, 0, \cdots, 0) = 0$. We write $P = Q + L$ where Q is a QF and L is a linear form $L = a_1x_1 + a_2x_2 + \cdots + a_mx_m$. We restrict attention to the case where Q is in canonical form. The identities

$$x_1x_2 + x_3x_4 + \cdots + x_{2t-1}x_{2t} + a_1x_1 + \cdots + a_{2t}x_{2t} =$$
$$(x_1 + a_1)(x_2 + a_2) + \cdots + (x_{2t-1} + a_{2t-1})(x_{2t} + a_{2t})$$
$$- (a_1a_2 + \cdots + a_{2t-1}a_{2t})$$

allow us to reduce the equations $P(x_1, \cdots, x_m) = a$ almost to the point where Table 4 applies, and the rest of the details are easy to supply.

5. The Weight Spectrum for the Second-Order Reed–Muller Code and Some of its Relatives

We now have collected enough information to calculate the complete weight spectrum for many codes whose codewords are the truth tables of a subgroup of the additive group of all polynomials whose degrees are ≤ 2. We shall consider only two of these codes, the quadratic forms and the group of all such polynomials. The first we shall call the second-order homogeneous Reed–Muller code (HRM), and the second is the usual GRM code.

We begin with the second-order HRM code, and using Table 4, we can make Table 5. Here the F_k's are as calculated in *Subsection 3*.

Since the weight of a codeword is the number of *nonzero* coordinates, by subtracting the number of zeros from q_m, we arrive at Table 6; since $Q(0, 0, \cdots, 0) = 0$ for any QF, we omit the zero vector from the truth table and obtain a code of block length $q^m - 1$.

Table 5. Number of QF's over GF (q) with a fixed number of zeros

Number of zeros	Number of QF's
q^{m-1}	$\sum_{k \text{ odd}} F_k$
$q^{m-1} + \epsilon(q-1)q^{m-j-1}$	F^{ϵ}_{2j} $\quad j = 0, 1, \cdots, \left[\frac{m}{2}\right]$

Table 6. Weight spectrum for the $\left(q^m - 1, \frac{m(m+1)}{2}\right)$ second-order HRM code over GF (q)

Weight	Words of this weight
$(q-1)(q^{m-1} \pm q^{m-j-1})$	$\frac{1}{2}q^{j^2}(q^j \pm 1)\frac{\prod_{i=m-2j+1}^{m}(q^i-1)}{\prod_{i=1}^{j}(q^{2i}-1)}$
$(q-1)q^{m-1}$	$\sum_{k \text{ odd}} F_k$

Before proceeding, let us observe that in the case $q = 2$, $x^2 = x$, so that the HRM code actually consists of the truth tables of all polynomials of degree ≤ 2 without constant term. Thus, Table 6 gives for $q = 2$ the weight spectrum of the (shortened) second-order Reed–Muller code over $GF(q)$. This result was recently obtained by Berlekamp and Sloane.

Let us now proceed to the second-order GRM code over $GF(q)$ for $q > 2$. There are $q^{m^2/2+3m/2+1}$ polynomials of degree ≤ 2. From Table 7, we see that if P is a polynomial of degree ≤ 2, the number of solutions to

Table 7. Number of solutions to $Q(x_1, x_2, \cdots, x_m) + L(x_1, \cdots, x_m) = a$ where Q is a canonical QF of rank k

If some $a_i, i > k$ is nonzero, the number of solutions is q^{m-1}.

Hence, we assume $\boxed{a_{k+1} = \cdots = a_m = 0}$ and:

Conditions		Number of solutions
odd k, odd q:	$\lambda\left(a + a_1a_2 + \cdots + a_{k-2}a_{k-1} + \frac{a_k^2}{4\lambda}\right) =$ square	$q^{m-1} + q^{m-(k/2)-1/2}$
	nonsquare	$q^{m-1} - q^{m-(k/2)-1/2}$
	0	q^{m-1}
odd k, even q:	$a_k = 0$	q^{m-1}
	$a_k \neq 0$, trace $\left(\frac{a + a_1a_2 + \cdots + a_{k-2}a_{k-1}}{a_k^2}\right) =$ 0	$q^{m-1} + q^{m-(k/2)-1/2}$
	1	$q^{m-1} - q^{m-(k/2)-1/2}$
even k, odd q:	$a + a_1a_2 + \cdots + a_{k-3}a_{k-2} + \frac{a_{k-1}^2}{4} - \frac{a_k^2}{4\lambda} =$ 0	$q^{m-1} + \epsilon(q-1)q^{m-(k/2)-1}$
	*	$q^{m-1} - \epsilon q^{m-(k/2)-1}$
even k, even q:	$a + a_1a_2 + \cdots + a_{k-1}a_k + \lambda(a_{k-1}^2 + a_k^2) =$ 0	$q^{m-1} + \epsilon(q-1)q^{m-(k/2)-1}$
	*	$q^{m-1} - \epsilon q^{m-(k/2)-1}$

$P(x_1, x_2, \cdots, x_m) = 0$ is of the form $q^{m-1} + \nu q^{m-j-1}$ where $\nu = 0, \pm 1$, or $\pm(q-1)$, and $0 \leq j \leq (m/2)$, and j is determined by the rank of the QF part of p which we call rank (P). Thus, from Table 7 P has $q^{m-1} + \nu q^{m-j-1}$ solutions for some $\nu \neq 0$ only if rank $(P) = 2j$ or $2j+1$. If we start with a P such that $P(0, 0, \cdots, 0) = 0$ and consider the set of polynomials $P + a, a \in GF(q)$, with the aid of Table 7 it is easy to arrive at Table 8.

Table 8. Number of solutions and a's to $P + a = 0$

If $P(x_1, \cdots, x_m) = Q(x_1, \cdots, x_m) + a_1x_1 + \cdots + a_mx_m$ with Q canonical, rank $Q = k$, and $a_{k+1} = \cdots = a_m = 0$ (otherwise always q^{m-1} solutions, the number of solutions to $P + a = 0$ is:

Number of solutions	Number of a's with this number of solutions
k odd, q odd: $q^{m-1} \pm q^{m-(k/2)-1/2}$	$\frac{1}{2}(q-1)$
q^{m-1}	1
k odd, q even: $q^{m-1} \pm q^{m-(k/2)-1/2}$	$\frac{1}{2}q$
k even: $q^{m-1} + \epsilon(q-1)q^{m-(k/2)-1}$	1
$q^{m-1} + \epsilon(q-1)q^{m-(k/2)-1}$	$q-1$

Table 8 leads immediately to Table 9. Finally, we get the weight spectrum for the second-order GRM, $q \neq 2$ (Table 10).

Reference

1. Dickson, L. E., *Linear Groups*, Dover Publications, Inc., New York, N.Y., 1958.

Table 9. Number of polynomials of degree ≤ 2 and rank k with a fixed number of zeros, $q \neq 2$

Number of zeros	Number of polynomials
k odd: $q^{m-1} \pm q^{m-(k/2)-1/2}$	$F_k \cdot \frac{1}{2}(q-1)q^k$
k even: $q^{m-1} \pm (q-1)q^{m-(k/2)-1}$	$F_k^+ \cdot q^k$
k even: $q^{m-1} \pm q^{m-(k/2)-1}$	$F_k^-(q-1)q^k$

Table 10. Weight spectrum for the $(q^m, m^2/2 + 3m/2 + 1)$ second-order GRM code, $q \neq 2$

Weight	Number
$(q-1)q^{m-1} \pm q^{m-j-1}$	$\frac{1}{2}(q-1)q^{2j+1}F_{2j+1} + (q-1)q^{2j}F_{2j}^+$
$(q-1)(q^{m-1} \pm q^{m-j-1})$	$q^{2j}F_{2j}^-$

B. Combinatorial Communication: Digital Quasi-Exponential Function Generator,

T. O. Anderson and W. J. Hurd

1. Introduction

In the implementation of the slope threshold data compression method of Kleinrock (SPS 37-49, Vol. III, pp. 325–328), it is necessary to generate an exponentially decaying function of time, with a programmable initial value and time constant. The value of the decaying exponential is compared to the value in a digital register, and it is necessary to be able to program the initial value and time constant of the exponential in order to select the desired data compression ratio.

An exponentially decaying time function can easily be obtained in an analog system by use of a simple resistance–capacitance circuit or with an integrate and reset circuit. Control of the time constant and initial value would be more difficult, however, and would involve some sort of analog switching. In a digital system, the straightforward implementation would be by means of a digital low-pass filter, which requires use of an adder and, in general, a multiplier. Another approach would be to use a hybrid system, with an integrate–reset circuit and an analog-to-digital converter. Both the straightforward digital approach and the hybrid approach would be fairly expensive, and the hybrid approach would suffer the usual disadvantages of analog circuitries.

This article describes a simple and inexpensive method for digital approximation of the exponential function to any desired degree of accuracy. The implementation uses only simple counters and gating logic.

2. Method

The basic procedure in the implementation of the quasi-exponential slope generator is to vary the clock rate of a countdown counter in such a manner that the value in the counter is decremented approximately exponentially with time. The initial value in the counter is preset to any desired value, and the time constant is chosen by selection of the basic clock frequency.

The first observation to make concerning the exponential function is that the slope at any point in time is proportional to the amplitude at that time. In particular, the slope when the counter is at full scale (FS) is twice the slope at one-half scale, which is in turn twice the slope at one-fourth scale, etc. Furthermore, the shape of the curve from full scale to one-half scale is exactly the same as from one-half scale to one-fourth scale, except for a factor of two in slope. This means that a counter which decays exponentially from full scale to half scale will also decay exponentially from half scale to one-fourth scale if the clock rate is halved.

In this implementation, the range of the counter is divided into a number of fields according to the value

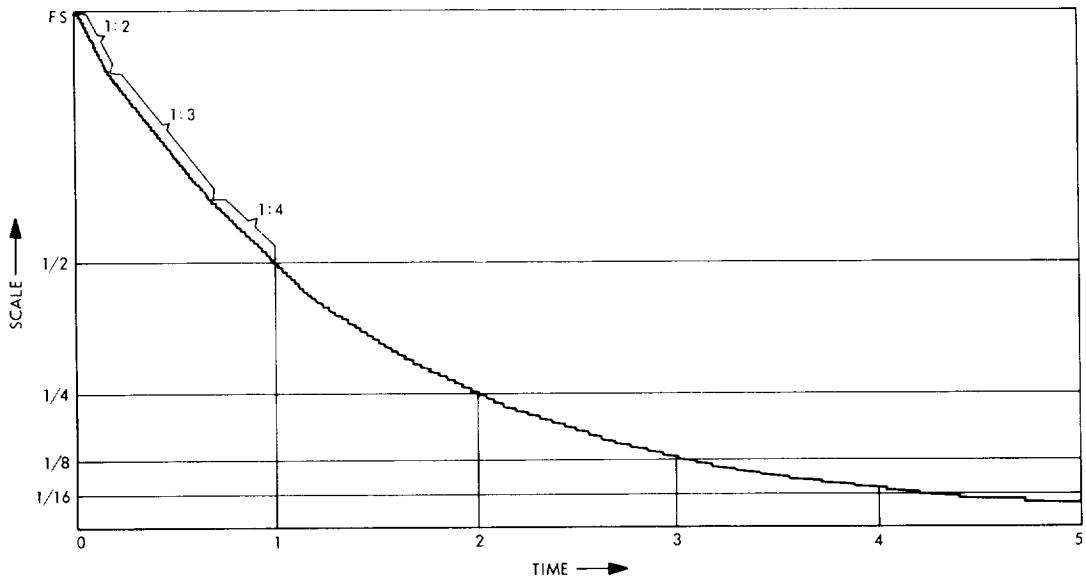


Fig. 1. Quasi-exponential waveform generator digital-to-analog converter output

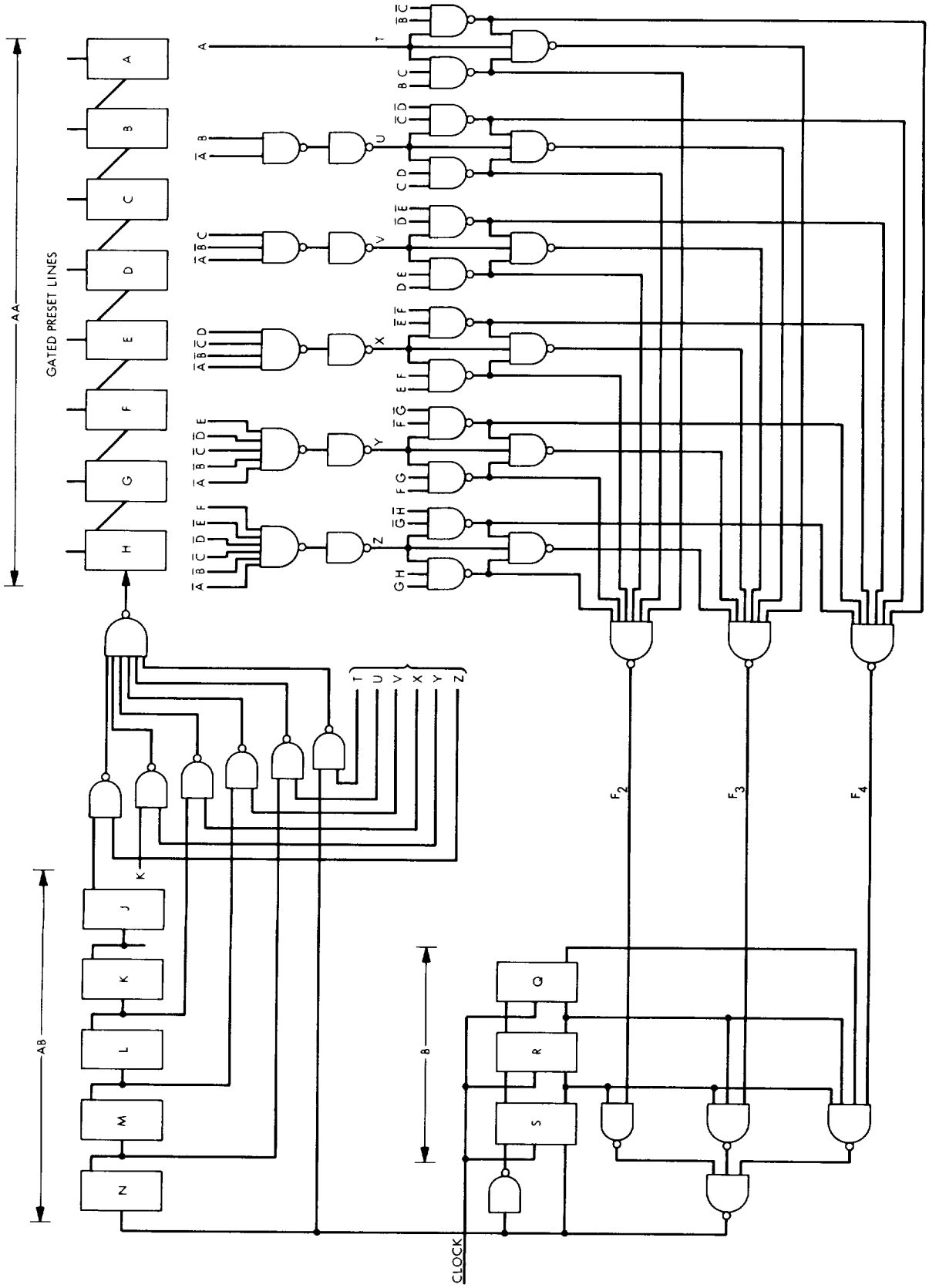


Fig. 2. Logic diagram of quasi-exponential waveform generator

in the counter, and the clock rate for each field is one-half of that for the preceding field, as follows:

full to ½ scale—full clock rate applied

½ to ¼ scale—½ clock rate applied

¼ to ⅛ scale—¼ clock rate applied

⅛ to 1/16 scale—⅛ clock rate applied

. .
. .
. .

It is now necessary only to approximate the exponential function within each field. This is accomplished by further dividing each field into several subfields, and by further dividing the basic clock by a different number in each subfield, thus determining the slope for that subfield. The slope in each subfield is proportional to the inverse of the number by which the basic clock is divided. In the application to the slope threshold data compressor, sufficient accuracy is obtained by dividing each field into three subfields. The ranges of the three subfields are the upper ¼, the center ½, and the lower ¼ of the major field, and the relative slopes in these three subfields are ½, ⅓, and ¼, respectively.

The approximation to the exponential obtained using these three subfields is shown in Fig. 1. The accuracy can be arbitrarily increased by increasing the number of subfields and properly selecting the slopes and ranges of these subfields. It is obvious that there is a clear-cut trade-off between accuracy and complexity of the control logic.

3. Implementation

The implementation of the quasi-exponential slope generator is shown in Fig. 2. The major components are three cascaded counters: a fixed length binary counter, AA; a variable length binary counter, AB; and a variable length ring counter, B. The value of the function is in counter AA, which can be preset to any value. The gating logic controls the number of stages in counter AB according to which field the value of counter AA is in, and controls the length of the ring counter according to the subfield. For the major field logic, functions T, U, V, · · · are 1 for the highest, next highest, third highest, · · · , major field, so that AB counts to $2^0, 2^1, 2^2, \cdots$. For control of the ring counter, functions F_2, F_3, and F_4 are 1 in the upper, center, and lower subfields so that counter B counts to 2, 3, or 4, respectively.

C. Decoding and Synchronization Research: Description and Operation of a Sequential Decoder Simulation Program, *J. A. Heller*

1. Introduction

A complete program has been written for the SDS 930 computer which simulates the operation of a sequential decoder, decoding real-time data. Any rate 1/3 convolutional code having a constraint length of at most 25 may be used. The channel simulated assumes binary phase-shift-keyed signaling, perfect phase coherence, and 3-bit receiver uniform output quantization. The details of the channel and quantizer thresholds are given in Ref. 1. All other channel and coder–decoder parameters are programmable over a wide range. For instance, the input data are broken up into blocks of programmable length. Between blocks the coder is resynchronized by inputting a constraint length of *zeros* into it. The energy-per-bit to noise ratio E_b/N_0 is also variable. The simulated decoder may have any size memory B, and any speed factor μ, over the data rate. In addition, the decoder threshold spacing is adjustable. Any number of blocks may be decoded in a given run before statistics are compiled and printed.

2. Operation of Decoder

Figure 3 shows a block diagram of the sequential decoder program. The control functions and all input–output are performed by a main program written in Fortran. The actual sequential decoding, data generation, and most of the compilation of statistics are performed by subroutines, written in symbol, which are called by the main program.

The magnetic tape containing the compiled program is loaded as follows:

(1) Place the tape on a tape transport at the load point.

(2) Set the transport for MT1, at a density of 800 bits/in.

(3) Input the Monarch commands:

ΔASSIGN S DF1Y BI MT1

ΔFORTLOAD BI

The loaded program will wait for data card input at the card reader.

a. Data card. This card is used to supply the program with all necessary simulation parameters for one complete

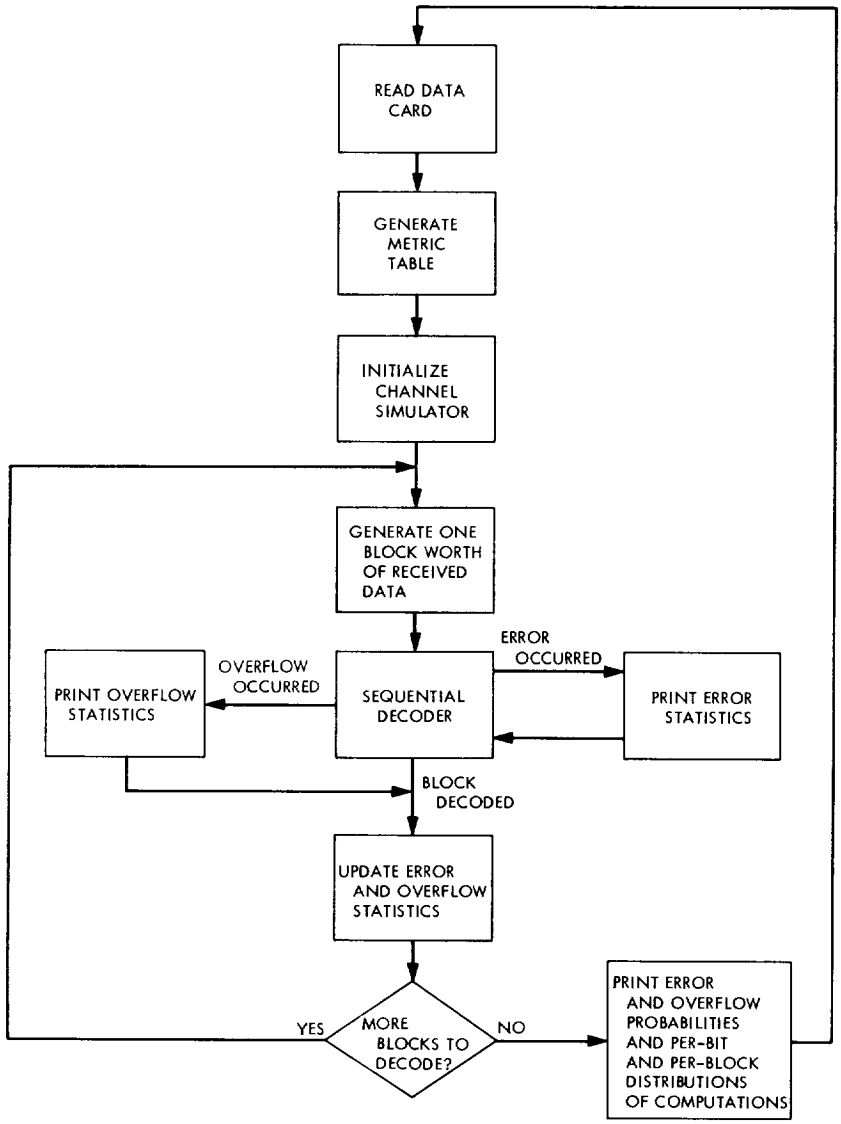


Fig. 3. Block diagram of the sequential decoder

run (with respect to Fig. 3, a run is defined as the sequence of events from the reading of a data card to the reading of the next data card). Figure 4 shows a typical data card. The items on this card are described in the following paragraphs.

Number of blocks. The integer in columns 1 through 6 (right justified to column 6) is the number of blocks to be decoded in this run. The box in Fig. 3 containing the question "more blocks to decode?" is answered by this number. In Fig. 4, the number of blocks is 7000.

Code constraint length. Columns 7 and 8 contain the integer code constraint length which may be any number not greater than 25. This number is the length of the resynchronization sequence between code blocks. It is also used in compiling error statistics. The constraint length is 23 in the example in Fig. 4.

Code. The convolutional code is determined by the octal contents of columns 10–17, 19–26, and 63–70. Each of the octal digits specifies the connections between a given coder shift register stage and the three mod-2 adders. It is assumed that the first coder stage is connected to *all* of the adders; thus, an octal 7 is implied to precede the 24-digit composite octal number formed by combining the three fields on the card. For example, the 24 octal digits from Fig. 4, preceded by a 7 for the

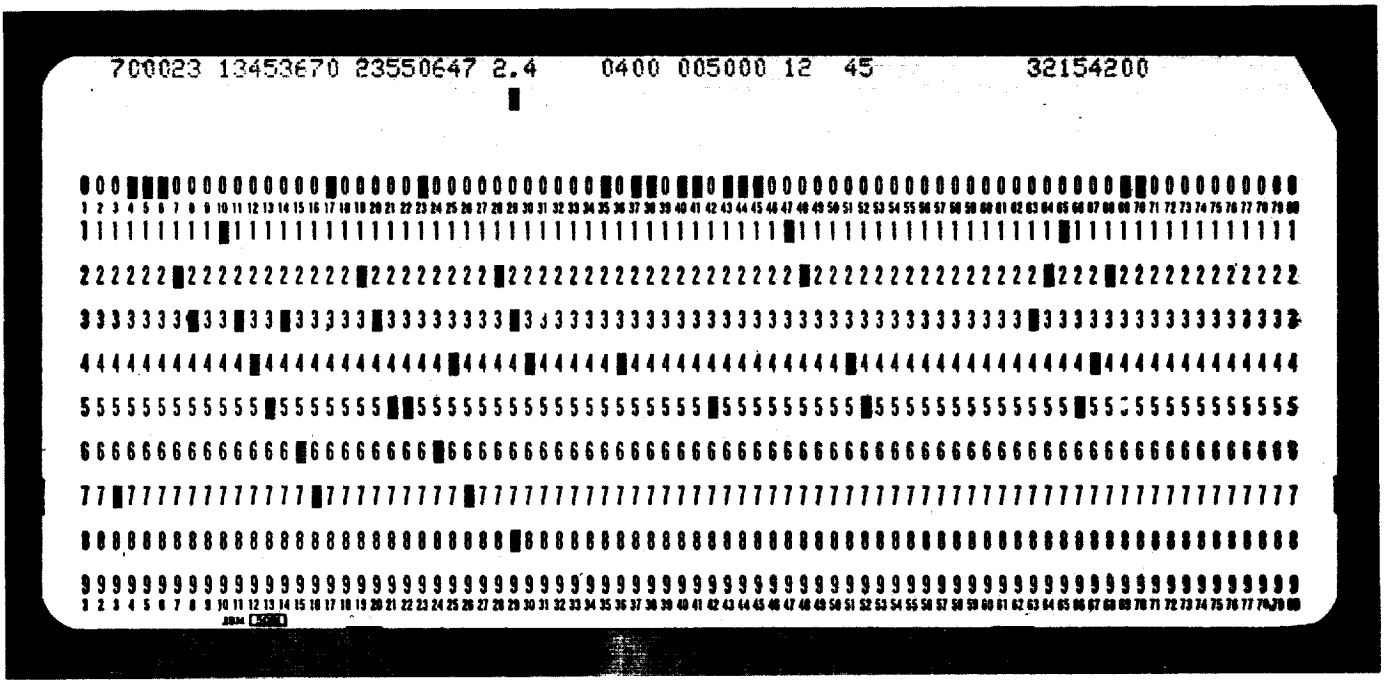


Fig. 4. Sample data card

first coder stage, are shown at the top in Fig. 5. Under each digit is its equivalent binary representation in a vertical format. Each of the three rows formed by these numbers is the code generator for one of the 3 mod-2 adders. A "1" indicates a connection and a "0," no connection. For instance, from g_1 in Fig. 5 we see that stages 1,4,5,7,8,12, · · · ,22 are connected to the first adder. It is seen from Fig. 5 that the code constraint length (longest non-*zero* generator span) is indeed 23, consistent with the number in columns 7 and 8. The composite octal code generator is always left justified in the three fields on the card. In other words, the octal digit in column 10 always specifies the connections from the second coder stage, column 11, the third coder stage, etc.

Signal-to-noise ratio. The F format floating point number in columns 28–33 is the per bit signal-to-noise ratio E_b/N_0 in dB. In Fig. 4, $E_b/N_0 = 2.4$ dB. For reference, an $E_b/N_0 \cong 2.2$ dB corresponds to operation at R_{comp} for a rate 1/3 code on the simulated channel; that is, when

	7	1	3	4	5	3	6	7	0	2	3	5	5	0	6	4	7	3	2	1	5	4	2	0	0
g_1 =	1	0	0	1	1	0	1	1	0	0	0	1	1	0	1	1	1	0	0	0	1	1	0	0	0
g_2 =	1	0	1	0	0	1	1	1	0	1	1	0	0	0	1	0	1	1	1	0	0	0	1	0	0
g_3 =	1	1	1	0	1	1	0	1	0	0	1	1	1	0	0	0	1	1	0	1	1	0	0	0	0

Fig. 5. The code generators

$E_b/N_0 = 2.2$ dB, $R_{comp} \approx 1/3$ bit/symbol (Ref. 1). As usual, when $E_b/N_0 > 2.2$ dB, $R_{comp} > 1/3$ and the decoder is operating *below* R_{comp}.

Code block length. The integer in columns 35–38 specifies the number of information bits encoded between decoder resynchronizations. The relative magnitude of this number compared with the constraint length determines the rate loss of the code. For instance, in Fig. 4 the block length is 400 and the constraint length is 23; hence, a fraction 23/423 of the data sent is a known sequence (in this case all *zeros*). The block length strongly influences the number of bits lost in a decoder buffer overflow. On the average, more than half of the bits in a block in which an overflow occurs are lost. When an overflow occurs in the simulated decoder, the strategy is to simply skip forward and start decoding the next block.

Buffer size. The integer in columns 40–45 specifies the number of branches of received data that the decoder can store in its buffer. In the example in Fig. 4, 5,000 branches of data can be stored. Since each information bit (branch) generates three code symbols, and each of these symbols is 3-bit quantized at the receiver, 9 bits must be stored for each branch. Thus, the 5,000 branch buffer represents a 45,000 bit memory. At the beginning of each run, the buffer is started as if an overflow had just occurred on the first bit of the previous block. That is, the buffer starts within a block of being full.

Speed factor. In columns 47–48 is placed an integer which is the speed factor of the simulated decoder. This is the ratio of the number of computations the decoder can perform per second, divided by the data rate in bits/s. The number used here should be greater than the average number of computations per bit or the decoder will not even be able to keep up with the average effort required to decode. A "computation" is defined as a look forward or backward from a tree node (Ref. 2, p. 473).

Threshold spacing. The threshold spacing used in the simulated Fano algorithm sequential decoder is determined by the integer in columns 51–52. If these columns are left blank, the threshold spacing is set at the *maximum* upward branch metric (about 27–34, depending on signal-to-noise ratio). If a larger spacing than 34 is desired, it can be entered on the card. If a threshold spacing less than 34 is inputted, the decoder may not work. The value 45 shown in Fig. 4 was chosen because it approximately minimizes the average computation per bit when $E_b/N_0 = 2.4$ dB.

b. Decoding metric. During decoding, the metric is obtained using a table look-up procedure. The table is generated using the standard Fano algorithm metric (Ref. 2, p. 463).

c. Channel simulator. For the purposes of the simulation, it is assumed that the all *zeros* data sequence is produced by the source. Thus, the output of the coder is also all *zeros*. This entails no loss in generality because decoder performance is invariant on the particular source sequence sent. The all *zeros* sequence is convenient because no coder is necessary to generate the codewords.

The effects of channel noise are obtained by using a multiplicative congruential noise generator (Ref. 3). Binary fractions uniformly distributed between *zero* and *one* are obtained using the recurrence relation

$$x_{i+1} = (cx_i) \bmod 2^{47}$$

The unit interval is subdivided into eight disjoint intervals. Each interval length corresponds to a channel transition probability. The channel output is determined by the interval in which the pseudo-randomly-generated number falls. The particular subdivision of the unit interval is, of course, a function of E_b/N_0. The parameters c and x_0 were chosen to make the x_i's appear independent, and to maximize the period of the recurrence relation. Using this generator, the Pareto exponents of the experimental distribution of computations per bit match closely the theoretically predicted values (Ref. 1). This is a particularly severe test of statistical independence.

At the beginning of each run (new data card), the noise generator is reset. That is, the recurrence relation begins again with $x_i = x_0$. Thus, for a fixed E_b/N_0, the same channel outputs are always generated.

3. Output Statistics

Figure 6 shows a sample printer output from a decoding run. It contains data card information, information on blocks that had errors or caused overflows, as well as compiled statistics.

Some items are self explanatory, others are explained below. In the second printed line, L is the code block length, B is the buffer size, and MU is the speed factor. In the fourth printed line, NBKTR is the threshold spacing; other items are irrelevant. The next line has E_b/N_0 and the corresponding probability that a hard decision on a symbol would be correct.

At this point, the program prints out data on errors and overflows as they occur.

a. Errors. Errors have occurred when the decoder has not followed the all *zero* information bit path in decoding a block. An error event begins at the first "1" in the decoded information stream and ends at the first succeeding "1" which is followed by a constraint length of information zeros. Thus, an error event begins with the first bit error which takes the decoder off the correct path and ends with the last bit error before the decoder remerges with the correct path. Several error events can occur in one decoded block and each error event has some bit errors and some correct bits contained within it.

When an error event occurs, the program prints ERR IN BLOCK followed by the number of the block in which the error occurred. This is followed by the length of the error, in information bits, from the first to the last bit error in the event. This is followed by NUMB OF ONES, or the number of bit errors in the event. Finally, the total number of computations it took to decode the offending block is recorded. The first error event in Fig. 6 occurred in block 3, had bit length 6, of which 4 were in error (and hence 2 correct), and took 5534 computations. The last (negative) figure in an error line indicates the position of the first bit of the error event in the block. This can vary between minus the block length (for the beginning of the block) to *zero* (for the end).

```
200 BLOCKS   CONSTRAINT LENGTH= 8
L= 200  B=  700  MU=10
                              CODE=13453670 00000000 00000000

IJK= 0 JKL= 0 KLM= 0 NBKTR= 45 NXTRA=  0
   2.400DB-BITSNR GIVES PO=.85912
ERR IN BLOCK    3    LENGTH=  6    NUMB OF ONES=  4    COMP/BLOCK=  5534     .53
ERR IN BLOCK   13    LENGTH=  3    NUMB OF ONES=  3    COMP/BLOCK=  2244     .78
ERR IN BLOCK   25    LENGTH= 19    NUMB OF ONES= 10    COMP/BLOCK=  3882     .136
ERR IN BLOCK   36    LENGTH=  4    NUMB OF ONES=  3    COMP/BLOCK=  5652     .175
ERR IN BLOCK   64    LENGTH=  1    NUMB OF ONES=  1    COMP/BLOCK=  5253     .95
ERR IN BLOCK   81    LENGTH=  5    NUMB OF ONES=  4    COMP/BLOCK=  1198     .189
ERR IN BLOCK  111    LENGTH=  3    NUMB OF ONES=  3    COMP/BLOCK=   660     .92
          OVERFLOW IN BLOCK   125   COMP/BLK=    7373
ERR IN BLOCK  152    LENGTH=  5    NUMB OF ONES=  4    COMP/BLOCK=  1349     .85
ERR IN BLOCK  168    LENGTH= 16    NUMB OF ONES=  8    COMP/BLOCK=  2330     .86
ERR IN BLOCK  174    LENGTH=  9    NUMB OF ONES=  5    COMP/BLOCK=  3885     .128
          OVERFLOW IN BLOCK   177   COMP/BLK=    7278

AVERAGE COMP/BIT=   5.861
BIT ERROR PROB=.112E-02
UNDETECTED ERRORS PER BIT=.250E-03
PROB THAT A BIT IS WITHIN AN UNDET ERR=.177E-02
MEAN ERROR LENGTH=   7.100
ERROR LENGTH VARIANCE=      31.490
FREQ OF OCCURANCE OF 1 IN ERROR EVERT=.634

PROB OF OVERFLOW=.1000E-01
PN SEQ AT ENDRUN= 00000000 00000000
P[BITCOMP GE      1]=  .103E 01
P[BITCOMP GE      5]=  .107E 00
P[BITCOMP GE     10]=  .670E-01
P[BITCOMP GE     20]=  .402E-01
P[BITCOMP GE     50]=  .173E-01
P[BITCOMP GE    100]=  .835E-02
P[BITCOMP GE    200]=  .425E-02
P[BITCOMP GE    500]=  .137E-02
P[BITCOMP GE   1000]=  .350E-03
P[BITCOMP GE   2000]=  .100E-03
P[BITCOMP GE   5000]=  .000E 00
P[BITCOMP GE  10000]=  .000E 00
P[BITCOMP GE  20000]=  .000E 00
P[BITCOMP GE  50000]=  .000E 00
P[BITCOMP GE 100000]=  .000E 00
P[BITCOMP GE 200000]=  .000E 00
P[BITCOMP GE 500000]=  .000E 00
P[COMP.GE.     500]=.800E 00
P[COMP.GE.    1000]=.345E 00
P[COMP.GE.    2000]=.105E 00
P[COMP.GE.    5000]=.250E-01
P[COMP.GE.   10000]=.000E 00
P[COMP.GE.   20000]=.000E 00
P[COMP.GE.   50000]=.000E 00
P[COMP.GE.  100000]=.000E 00
P[COMP.GE.  200000]=.000E 00
P[COMP.GE.  500000]=.000E 00
```

Fig. 6. Sample printer output

b. Overflows. When an overflow occurs, OVERFLOW IN BLOCK, followed by the block number, is printed. Next is the number of computations performed in that block up until the time that the buffer overflowed.

After all blocks are either decoded or deleted due to overflow, the run statistics are printed. The AVERAGE COMP/BIT is the total number of computations in the run divided by the number of bits decoded. UNDETECTED ERRORS PER BIT is the number of error events divided by the number of bits decoded. The PROB OF OVERFLOW is the number of *blocks* in overflow divided by the number of *blocks* decoded. This is equal to the bit deletion rate only if *all* bits in an overflowed block are considered lost. In practice, on the average, perhaps 1/3 to 1/2 of the bits in such a block are decoded reliably. Thus, the deletion rate is somewhat lower than the printed overflow probability. The line following the probability of overflow is irrelevant.

The next 17 lines give the per bit distribution of computations. Each line is the frequency that the number of computations per bit was greater than or equal to 1,5,10,20, etc. Notice that in the first line the "probability" is 1.03, when it should be 1.00 (since each bit requires one computation with certainty). This anomaly is due to counting computations due to information bits and resynchronization bits, while only normalizing by the number of information bits alone. The number of computations performed on a bit i in a block is defined as the number

of computations done between the time the decoder first reaches any node at depth i in the tree to the time it first reaches any node at depth $i + 1$. If no errors occur, the distribution of the number of computations per bit will be Pareto.

The next 10 lines give the distribution of computations per block.

4. Timing

For the purposes of estimating the computer time for a simulation run, the following formula is convenient. On the average, the time (in milliseconds) necessary to decode one bit is given by

$$T_b \approx 0.6 + 0.1\,\bar{c}$$

where $\bar{c}$ is the average number of computations per bit expected in the run. $\bar{c}$ varies between 1+ at high E_b/N_0 and about 10 near R_{comp} to higher values at lower E_b/N_0 (decoding above R_{comp}). It is also a function of how many errors and overflows occur. In Fig. 6, 200 blocks of 200 bits each have been decoded. $\bar{c}$ is about 5.8. Using the formula, $T_b \cong 1.2$ ms. Therefore, the run of 40,000 bits took about 48 s.

5. An Application

The program was used to study the overflow behavior of a sequential decoder operating just below R_{comp}, as a function of buffer size. The results will be presented here as an example of an application of the simulator.

A constraint length 25 code given by the octal generator 7134536702355064732154207 was used in the simulation runs. The first eight digits in the generator were chosen because they generate an excellent $K = 8$ code (SPS 37-54, Vol. III, pp. 171–177). The rest of the digits were chosen at random. This constraint length was long enough that no errors occurred in the decoding of over 10^8 information bits. The purpose of the experiments was to compare actual frequencies of overflows with the probabilities predicted by a theoretical method.

Figure 7 is a plot of overflow probability versus buffer size for two values of E_b/N_0, and two different speed factors. The seven experimental points are the result of seven simulation runs with the number of blocks decoded varying from 7,000 for the highest probability point to 260,000 for the lowest point on the $E_b/N_0 = 2.4$ dB curve. The 260,000 block run took just less than two days of computer time, and generated 41 overflow events!

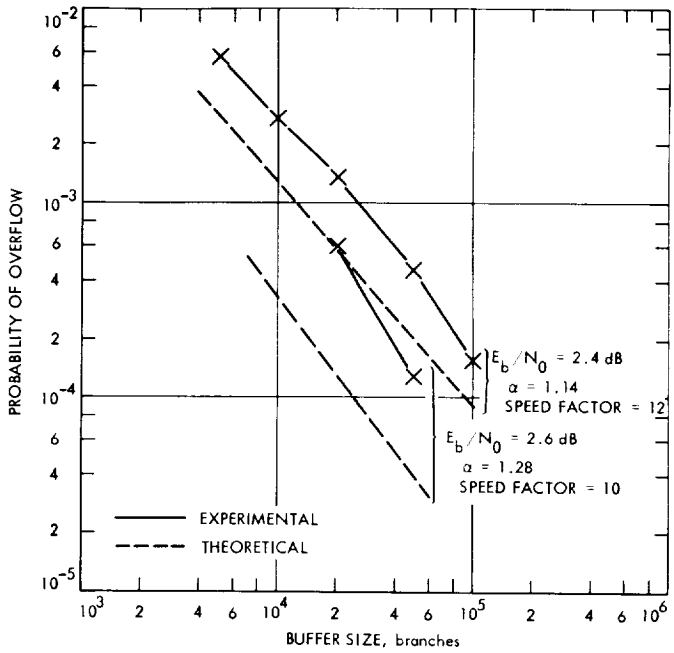


Fig. 7. Experimental and theoretical overflow probabilities vs buffer size for a code with constraint length 25 and block length 400 (no errors occurred)

It is well known both theoretically and experimentally that the per bit distribution of computations is of the Pareto form (Ref. 1), that is

$$\Pr(c > L) \approx kL^{-\alpha}, \qquad L >> 1$$

It has been argued that if overflows are caused mainly by long single searches rather than a concatenation of short searches, the probability of an overflow per bit is just the probability that an empty buffer fills before the decoder penetrates one branch further into the tree (SPS 37-56, Vol. III, pp. 78–83). But this is just

$$\Pr(c > \mu B) \approx k(\mu B)^{-\alpha}$$

where μ is the speed factor and B is the buffer size. An overflow can occur while decoding any of the L bits in a block, so using the union bound (and assuming all bits in an overflow block are lost)

$$P_o \approx L \Pr(c > \mu B) \approx kL(\mu B)^{-\alpha}$$

From the per bit distributions of computation obtained from these runs, it was determined that $k = 1.9$. Also, it was found that for $E_b/N_0 = 2.4$ dB, $\alpha = 1.14$ and for $E_b/N_0 = 2.6$ dB, $\alpha = 1.28$. These Pareto exponents agree closely with theory (Ref. 1). Using these values, along

with $L = 400$ and $\mu = 12$ at 2.4 dB and 10 at 2.6 dB, we get

$$P_o(2.6\,\text{dB}) \approx 40\,(B^{-1.28})$$

and

$$P_o(2.6\,\text{dB}) = 45\,(B^{-1.14})$$

These are plotted as the dashed lines in Fig. 7. (They are straight lines on log–log paper.)

The fact that the experimental curves are a factor of 2 or 3 times worse in overflow probability than the theoretical curves is due to the fact that overflows were not always isolated events due to one long search starting with an empty buffer. In fact, the output data clearly show that an overflow event which required about μB computations was often soon followed by more block overflows caused by much less than μB computations. In other words, overflows often occurred in bursts. The first overflow event leaves the buffer almost full; while the buffer is emptying, the probability of an additional overflow is much greater than usual—hence the bursts. This condition was anticipated in a more general theoretical treatment on overflow probability (SPS 37-56, Vol. III, pp. 78–83).

Much more simulation must be done before the parameters necessary to achieve low overflow probabilities with a sequential decoder are known. It is clear, however, that the original theory is too optimistic and may be even more unrealistic at lower erasure probabilities.

References

1. Jacobs, I. M., "Sequential Decoding for Efficient Communication from Deep Space," *IEEE Trans. on Commun. Technol.*, Vol. COM-15, No. 4, Aug. 1967.
2. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, John Wiley & Sons, New York, N.Y., 1965.
3. Chambers, R. P., "Random-Number Generation," *IEEE Spectrum*, Feb. 1967.

D. Decoding and Synchronization Research: On Doppler Jitter and Digital Error Rate in Coherent Systems, *C. L. Weber[1] and W. J. Hurd*

1. Introduction

In digital communication systems, there is often the need to maintain continuous doppler frequency information for range-rate estimation as well as the need to maintain the error rate below a specified value. The allocation of the total available transmitted power to the various information-bearing subcarrier signals then needs to be carried out based on knowledge of the effect that this choice will have on doppler tracking capability as well as error rates. This article presents a new analysis of doppler measurement performance for a general class of coherent digital communication systems. Quantitative tradeoffs are given between bit error probability and doppler tracking performance as a function of power allocation.

The type of coherent digital system to be considered is one which transmits the data signals by phase modulating the RF carrier with biphase-modulated sinewave data subcarriers. The frequencies of the several subcarriers are assumed to have been judiciously chosen so that the spectra of the modulated data are non-overlapping.

The demodulation of the subcarriers is assumed to be carried out by employing squaring loops, Costas loops, etc., so that all necessary subcarrier phase and synchronization reference information is obtained directly from the data signal, thereby eliminating the need for separate sync channels and eliminating the need for placing any power in a subcarrier reference signal.

2. System Model

A basic simplified diagram of the general type of system to be considered is depicted in Fig. 8. In the transmitter portion of the system, the data signals

$$\{s_k(t), k = 1, \cdots, K\}$$

biphase-modulate the frequency-multiplexed subcarrier waveforms $\{(2p_k)^{1/2} \cos \omega_k t, k = 1, \cdots, K\}$. The input to the carrier phase modulator $\theta(t)$ is therefore given by

$$\theta(t) = \sum_{k=1}^{K} (2p_k)^{1/2}\, s_k(t) \cos \omega_k t \qquad (1)$$

where $p_k, k = 1, \cdots, K$ is the average power in the kth biphase-modulated subcarrier waveform before carrier phase modulation.

The trend in modern digital communication systems is to employ subcarriers with 100% modulation, that is, there is no residual power at the subcarrier frequency for tracking purposes. The reason for this is, with the advent of squaring loops, Costas loops, delay-locked

[1]Consultant, Electrical Engineering Department, University of Southern California, Los Angeles, California.

loops, etc., coherent phase reference and bit synchronization information can be obtained directly from the 100%-modulated data signal. The $\{s_k(t), k = 1, \cdots, K\}$ are thus assumed to consist of a sequence of ± 1's with bit times $\{T_{b_k}, k = 1, \cdots, K\}$, respectively.

With this modulation scheme assumed, the output of the phase modulator is given by

$$s(t) = (2P)^{1/2} \sin(\omega_c t + \theta(t) + \theta_0) \tag{2}$$

where P is the overall average transmitted power, and θ_0 is some unknown constant reference angle.

The received waveform is then given by

$$y(t) = (2P)^{1/2} \sin\left[\omega_c t + \int^t \omega_d(\tau)\, d\tau + \theta(t) + \theta_0\right] + n(t) \tag{3}$$

where $n(t)$ is assumed to be white gaussian noise with one-sided spectral density N_0 watts/hertz, and $\omega_d(\tau)$ represents the doppler frequency shift due to the relative range-rate between transmitter and receiver.

Neglecting frequency shifters, frequency synthesizers, bandpass limiters, etc., the coherent carrier tracking loop generates the reference signal

$$r(t) = 2^{1/2} \cos\left[\omega_c t + \int^t \omega_d(\tau)\, d\tau + \hat{\theta}_0(t)\right] \tag{4}$$

The data bearing waveforms which comprise $\theta(t)$ are assumed to be at frequencies outside the bandwidth of the carrier phase-locked loop (PLL). The doppler frequency is assumed to be varying slowly enough to be within this bandwidth, however, so that the carrier PLL is able to track this signal. Therefore, the output data bearing signal of the carrier tracking loop which goes into the various subcarrier demodulators is given by

$$y_0(t) = s_0(t) + n_0(t) \tag{5}$$

where

$$s_0(t) = P^{1/2} \sin[\theta(t) + \phi_r(t)]$$

The additive noise $n_0(t)$ has the same statistics (Ref. 1) as $n(t)$, and $\phi_r(t) = \theta_0 - \hat{\theta}_0(t)$ is the carrier loop phase error.

With the doppler frequency slowly varying, any cycle slipping can be assumed to be due only to the additive noise. No detuning will be assumed to exist between the received carrier and the voltage-controlled oscillator rest frequency. Then the approximate steady-state mod 2π probability density function of ϕ_r is given by (Ref. 1)

$$p(\phi_r) \simeq \frac{\exp(\alpha_r \cos \phi_r)}{2\pi I_0(\alpha_r)}, \qquad -\pi \leq \phi_r \leq \pi \tag{6}$$

where

$$\alpha_r \triangleq \frac{P_c}{N_0 B_{L_r}} \tag{7}$$

is the signal-to-noise ratio (SNR) of the carrier tracking loop. In Eq. (7), B_{L_r} is the one-sided noise bandwidth of the carrier PLL based on the linear theory and P_c is the average power of the received signal at the carrier frequency. The overall SNR of the received signal is defined to be

$$\beta_r \triangleq \frac{P}{N_0 B_{L_r}} \tag{8}$$

In order to ultimately specify the trade-off between doppler tracking capability and digital demodulation capability, the distribution of power between the carrier tracking loop and the subcarrier demodulation loops must be specified. To do this, $s(t)$ is represented by the series [see, e.g., Lindsey (Ref. 2)]

$$\begin{aligned} s(t) &= (2P)^{1/2} \operatorname{Im}\{\exp[j(\omega_c t + \theta(t) + \theta_0)]\} \\ &= (2P)^{1/2} \operatorname{Im}\left(\exp[j(\omega_c t + \theta_0)]\right. \\ &\quad \left. \times \prod_{k=1}^{K} \sum_{m_k=-\infty}^{\infty} (j)^{m_k} J_{m_k}(2p_k)^{1/2} \exp\{j m_k [\omega_k t + s_k(t)]\}\right) \end{aligned} \tag{9}$$

where $J_{m_k}(\cdot)$ is the Bessel function of order m_k. The signal which enters the carrier tracking loop is

$$(2P_c)^{1/2} \sin(\omega_c t + \theta_0) + n(t)$$

where the average power in the tracking signal, P_c, is given by the average power of the component of $s(t)$ in Eq. (9) at ω_c. This is obtained by setting

$$m_k = 0, k = 1, \cdots, K$$

from which we obtain the fraction of the total power that enters the carrier tracking loop, namely,

$$\frac{P_c}{P} = \prod_{k=1}^{K} J_0^2(2p_k)^{1/2} \tag{10}$$

To obtain the power in the subcarriers, $s_0(t)$ is similarly expanded. The average power in the kth subcarrier data signal at the output of the kth extraction filter (Fig. 8) is obtained by assuming $\phi_r(t)$ is essentially constant over the bit time T_{b_k} of the kth data signal, and setting $m_k = \pm 1$ and $m_{k'} = 0$ for all $k' \neq k$. This average power P_{s_k} is then given by

$$\frac{P_{s_k}}{P} = 2J_1^2(2p_k)^{1/2} \prod_{\substack{k'=1 \\ k' \neq k}}^{K} J_0^2(2p_{k'})^{1/2} [\mathbf{E}(\cos\phi_r)]^2 \tag{11}$$

where $\mathbf{E}$ denotes expectation. From Ref. 1,

$$\mathbf{E}(\cos\phi_r) = \frac{I_1(\alpha_r)}{I_0(\alpha_r)}$$

For a given ϕ_r, the input signal to the kth subcarrier demodulator is then given by

$$y_k(t) = (2P_{s_k}(\phi_r))^{1/2} \cos\left[\omega_k t + \frac{\pi}{2} s_k(t)\right] + n_k(t) \tag{12}$$

where

$$[P_{s_k}(\phi_r)]^{1/2} \triangleq (P)^{1/2} D_{s_k} \cos\phi_r \tag{13}$$

and where $n_k(t)$ has the same statistics as $n(t)$. This follows from the fact that the extraction filters

$$H_k(j\omega), k = 1, \cdots, K$$

in Fig. 8 are normally broadband with respect to the bandwidths of the synchronization tracking loop, the phase tracking loop, and the matched filter in the subcarrier demodulator. In Eq. (13),

$$D_{s_k} \triangleq \left[2J_1^2(2p_k)^{1/2} \prod_{\substack{k'=1 \\ k' \neq k}}^{K} J_0^2(2p_{k'})^{1/2}\right]^{1/2}$$

In the important special case where all subcarrier channels are afforded the same performance level, we have that $p_k = p, k = 1, \cdots, K$, from which Eqs. (10)

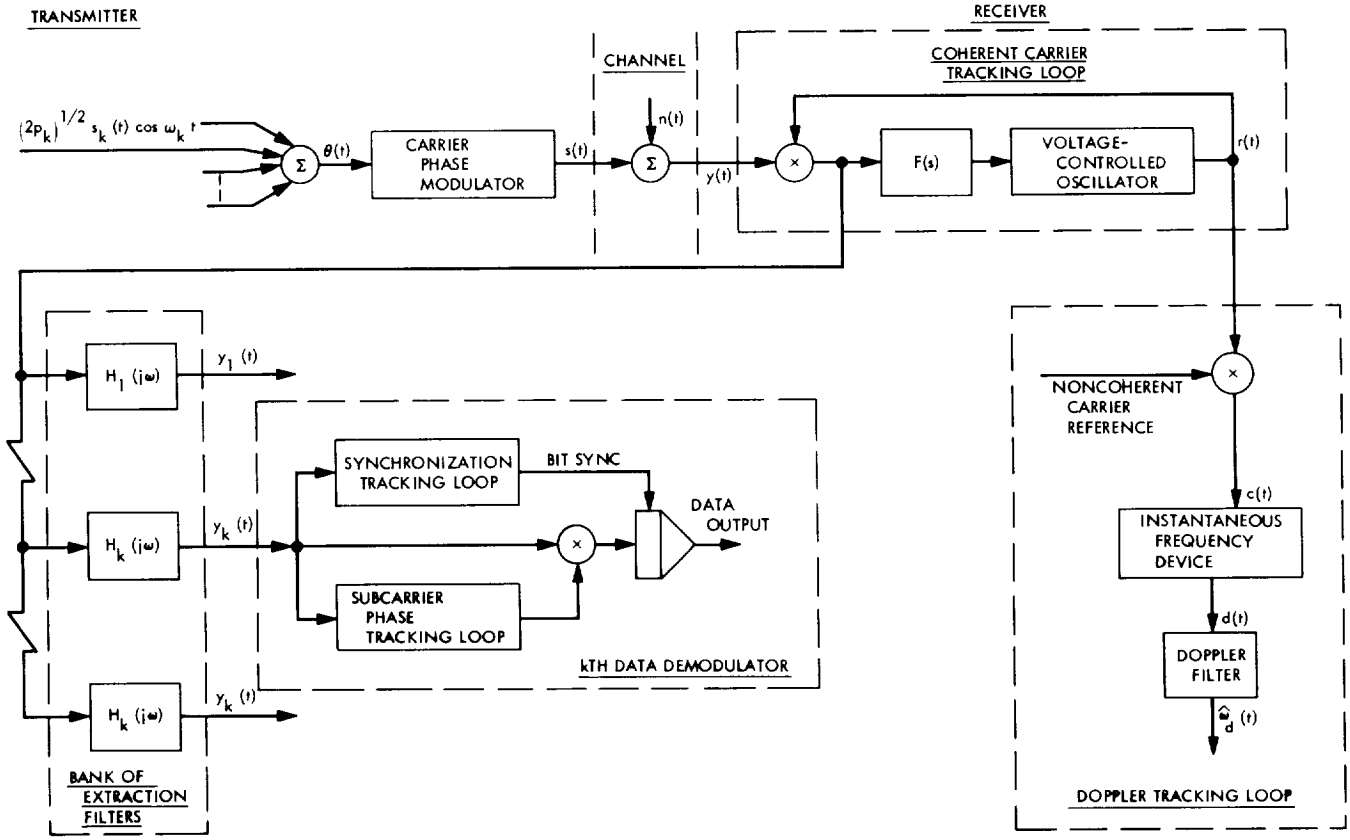


Fig. 8. General system configuration

and (11) reduce to

$$\frac{P_c}{P} = [J_0(2p)^{1/2}]^{2K} \tag{14}$$

and

$$\frac{P_{s_k}}{P} = 2J_1^2(2p)^{1/2}[J_0(2p)^{1/2}]^{2K-2}[\mathbf{E}(\cos\phi_r)]^2 \tag{15}$$

respectively. Note that p affects $\mathbf{E}(\cos\phi_r)$ as well as the other factors in Eqs. (14) and (15). Since P_{s_k}/P is dependent on $\mathbf{E}(\cos\phi_r)$, we see that the performance of the digital demodulation of the subcarrier signals is not directly dependent upon the occurrence of cycle slipping events, but only indirectly dependent to the extent that cycle slipping broadens the steady-state mod 2π probability density function of the carrier phase error ϕ_r. Cycle slipping, as would be expected, is a fundamental concern in determining doppler tracking capability.

The distributions of the power into the subcarriers as functions of p, as given by Eq. (15), are shown in Fig. 9 for various values of β_r and K. It is important to note that for a fixed number of subcarriers, the value of p which maximizes the subcarrier power varies significantly with signal-to-noise ratio β_r. For example, for $K = 1$, the

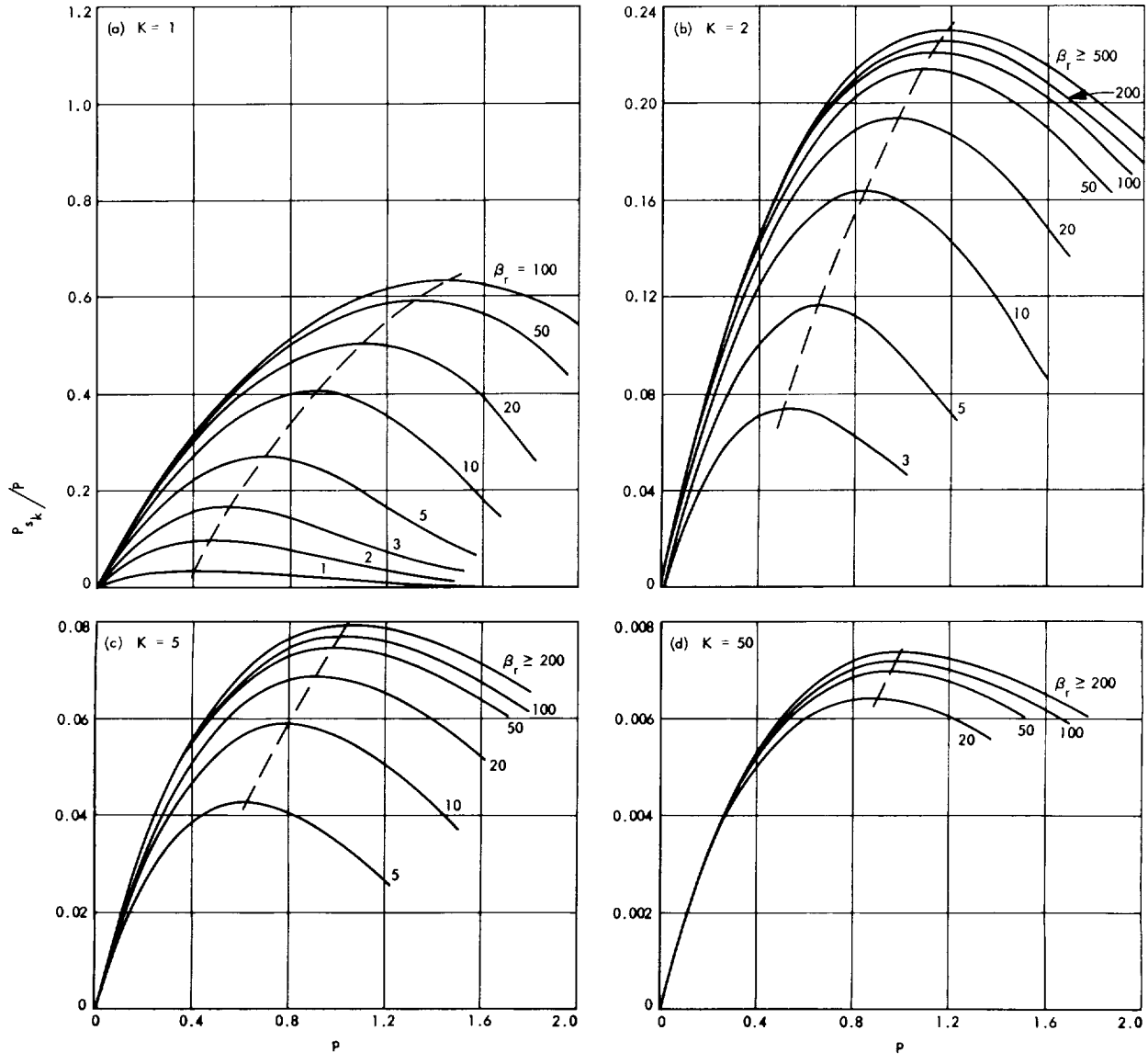


Fig. 9. P_{s_k}/P versus p for various values of β_r and K

maximum value of P_{s_k}/P occurs for values of p ranging from 0.4 at $\beta_r = 1$ to 1.45 at $\beta_r = 100$. This demonstrates that when designing a biphase digital communication system (as well as when designing a PLL), a design point must be picked; that is, an input signal-to-noise ratio β_r must be chosen at which the system will be designed to operate optimally.

The dependence of the optimal p on β_r decreases as the number of subcarriers is increased; this is displayed in Fig. 10, where the optimal choice of p, $\hat{p}$, is plotted against β_r for various K. We are defining $\hat{p}$ to be that value of p which maximizes the total power into the subcarriers.

It should also be noted that the optimal p often occurs at values which over-modulate the carrier. When $K = 1$, for example, $p = \pi^2/8 \simeq 1.23$ corresponds to a peak phase modulation of $\pi/2$ radians, so the carrier is 100% modulated. From Fig. 10, it is seen that $\hat{p}$ is greater than 1.23 for all $\beta_r > 35$ when $K = 1$, so that there is over-modulation. For $K \geq 2$, there is over-modulation for all β_r which produce reasonable error rates. Therefore, in the design of digital phase-modulated systems, over-modulation is not only good, it is optimum in most cases. The exact amount of over-modulation which is desirable can be obtained from Fig. 10.

The fraction of the total power that is lost in phase modulation due to noise and distortion can also be easily displayed. This fraction decreases as β_r increases. In Fig. 11, curve 1 is the maximum percent of the total received power that is transferred into the various subcarriers, namely, KP_s/P, which is plotted versus K at $\hat{p}$ and at high β_r. When $K = 1$, it is possible to transfer as much as 68% of the total power into the one subcarrier, while when $K = 50$, only 37% of the total power can be transferred into all 50 subcarrier channels. Curve 2 is a plot of the percent of total power which is needed by the carrier tracking loop at $\hat{p}$ and at high β_r. Equivalently stated, KP_s/P is maximized when P_c/P is equal to the value given by curve 2 in Fig. 11. Curve 3 in Fig. 11 is the sum of curves 1 and 2. The remainder of the power is lost. For one subcarrier, at least 22% of the total power is always lost, while for $K = 50$, at least 36% of the power is always lost.

The performance of the subcarrier demodulators and the doppler measuring subsystem is now determined so that their tradeoff can be established.

3. Bit Error Performance

The output signal of the kth extraction filter, which is the input signal to the kth subcarrier demodulator, is

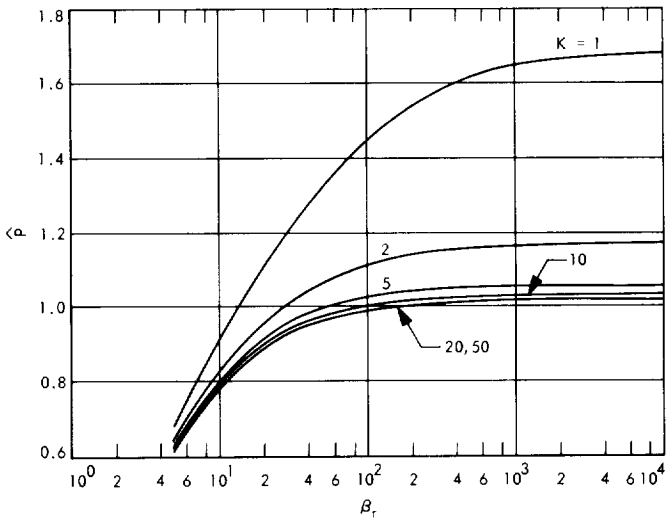


Fig. 10. $\hat{p}$ versus β_r for various values of K

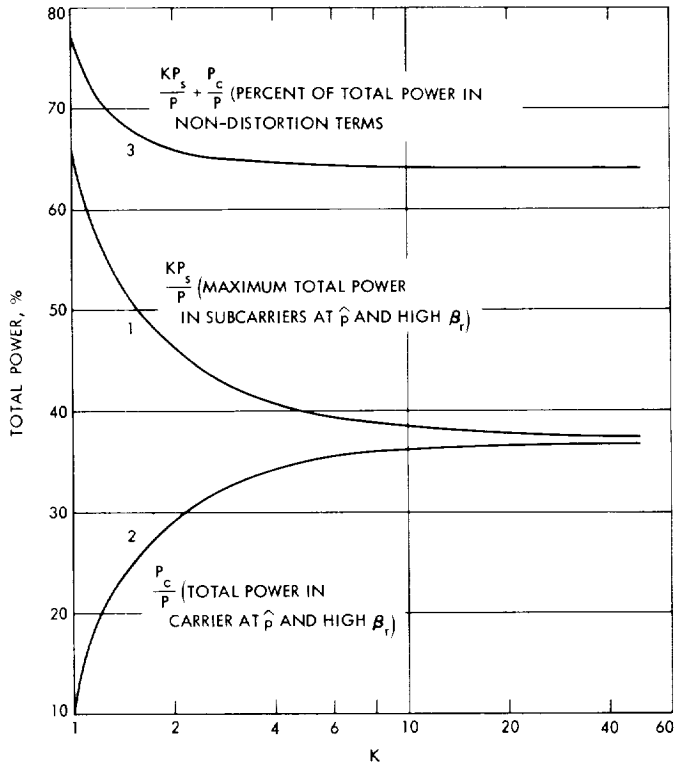


Fig. 11. Power allocations which maximize total subcarrier power

given by

$$y_k(t) = [2P_s(\phi_r)]^{1/2} \cos\left[\omega_k t + \frac{\pi}{2} s_k(t)\right] + n_k(t) \tag{16}$$

where $n_k(t)$ has the same statistics as $n(t)$, and where the assumption is maintained that the total available subcarrier power is shared equally among the K data channels. Each data signal $s_k(t)$ fully biphase-modulates its subcarrier reference signal $\cos \omega_k t$.

The assumption that the squaring loop, Costas loop, or other method which is employed to obtain coherent phase information is functioning with negligible jitter with respect to that in the carrier loop is now employed. This is realistic, since the noise bandwidth of the subcarrier tracking loop can often be made 10^{-3} to 10^{-6} times narrower than that of the carrier tracking loop. Therefore, the subcarrier telemetry demodulation will be assumed to perform like a perfectly coherent system which has perfect synchronization information. The probability of a bit error is then given by (Ref. 1)

$$P_e = \int_{-\pi}^{\pi} \operatorname{erfc}\left(\frac{2P_s(\phi_r)\,T_b}{N_0}\right)^{1/2} \frac{\exp(\alpha_r \cos\phi_r)}{2\pi I_0(\alpha_r)}\, d\phi_r \tag{17}$$

where $\operatorname{erfc}(\cdot)$ is the complementary error function. The average SNR per bit R_s is defined as the ratio of signal energy per bit to noise spectral density. Therefore,

$$R_s \triangleq \frac{P_s T_b}{N_0} = \frac{\{\mathbf{E}[P_s(\phi_r)]^{1/2}\}^2 T_b}{N_0} \tag{18}$$

where T_b is bit time, which has been assumed the same for all digital subcarrier channels. In terms of the parameters which describe the carrier tracking loop, R_s can be expressed as

$$R_s = \beta_r \frac{P_s}{P} \frac{1}{\delta_r} \tag{19}$$

where

$$\delta_r \triangleq \frac{1}{T_b B_{L_r}}$$

is defined as the reciprocal of the product of the bit time and the noise bandwidth of the carrier loop.

When α_r, the carrier loop SNR, is high, $\mathbf{E}\{\cos\phi_r\}$ is close to 1 and P_E is closely approximated by $\operatorname{erfc}(2R_s)^{1/2}$. For low α_r, however, P_E is often significantly higher than $\operatorname{erfc}(2R_s)^{1/2}$. Furthermore, the minimum value of P_E (minimized with respect to p) does not necessarily occur exactly at the maximum value of R_s. For preliminary design purposes, however, the P_E at $R_{s\max}$ is usually close enough to $P_{E\min}$ so that it may be adequate to design for $R_{s\max}$ rather than for $P_{E\min}$, for the sake of convenience. However, except at high loop SNR's, one should never approximate $P_E(R_{s\max})$ by $\operatorname{erfc}(2R_{s\max})^{1/2}$, but should perform the appropriate integration given in Eq. (17).

4. Doppler Measurement

The signal from which doppler frequency information is to be obtained is the carrier voltage-controlled-oscillator output $r(t)$ which has the representation

$$r(t) = 2^{1/2} \cos\left[\omega_c t + \int^t \omega_d(\tau)\, d\tau + \hat{\theta}(t)\right]$$

The carrier frequency is removed by mixing with a noncoherent carrier reference $2^{1/2}\sin(\omega_c t)$ so that the output of the doppler mixer, neglecting the double-frequency terms, is

$$c(t) = \cos\Delta(t) \tag{20}$$

where

$$\Delta(t) = \int^t \omega_d(\tau)\, d\tau + \phi_r(t) + \theta_1 \tag{21}$$

The signal $c(t)$ is the input waveform to the doppler measurement device. Whatever scheme is employed, we shall assume that this device ideally obtains the instantaneous frequency of $c(t)$, namely $\dot{\Delta}(t)$,

$$\dot{\Delta}(t) = \omega_d(t) + \dot{\phi}_r(t) \tag{22}$$

The measurement disturbances in Eq. (22) are seen to be $\dot{\phi}_r(t)$. In this initial approach to provide a tradeoff between doppler and error rate, we shall assume $\omega_d(t)$ is sufficiently slowly varying so that it can be assumed constant. In general, $\dot{\Delta}(t)$ would be filtered to provide the best estimate $\hat{\omega}_d(t)$ of $\omega_d(t)$ from $\dot{\Delta}(t)$.

The fundamental task in acquiring knowledge of doppler measurement capability is in obtaining necessary statistical information about $\dot{\phi}_r(t)$, namely, the first and second moments. The difficulty centers around the fact that, in order to obtain a tractable mathematical model of a PLL, the assumption is generally made that the additive noise is white. In most choices of loop filters, this leads to the conclusion that $\dot{\phi}_r(t)$ is also white. This problem can be partially overcome with the following

approach. Let us model the total phase error $\phi_r(t)$ as

$$\phi_r(t) = \int^t 2\pi N(\tau)\, d\tau + \phi_m(t) \tag{23}$$

where $N(\tau)$ is a stochastic process consisting of a sequence of pulses which are each of unit area and of short time duration. A pulse is positive whenever the loop slips a cycle to the right and negative whenever the loop slips to the left. The process $\phi_m(t)$ is the mod 2π phase process (Ref. 1). When $\phi_m(t)$ is modeled as in Eq. (23), the cycle slipping part and the phase jitter part between cycle slips are additive. At low loop SNR's the predominant contribution to overall phase error will be due to cycle slips, while at large SNR's, cycle slips will occur very rarely, and the predominant variation in $\phi_r(t)$ will be due to the mod 2π phase jitter $\phi_m(t)$.

Empirical data taken on phase-locked loops (Ref. 3) lead one to believe that the cycle slipping events in disjoint intervals are statistically independent. This, plus the fact that it is a jump process, is sufficient to conclude (Ref. 4) that $N(t)$ is a generalized Poisson process.

The cumulative number of cycle slips $N(t)$ can be expressed as

$$N(t) = N_+(t) - N_-(t) \tag{24}$$

where $N_+(t)$ consists of the positive pulses in $N(t)$ and thus represents cycle slips in the positive $\phi_r(t)$ direction, and, similarly $N_-(t)$ represents cycle slips in the negative $\phi_r(t)$ direction. If there is no detuning in the carrier voltage-controlled oscillator, then the expected number of cycle slips to the right will equal that to the left. Hence,

$$\mathbf{E}(N(t)) = 0$$

The individual processes $N_+(t)$ and $N_-(t)$ are Poisson processes, and will be assumed to be statistically independent. Therefore, since the steps are always taken to be of unit size, and because the variance of the Poisson distribution equals the mean,

$$\sigma^2_{N_+} = \sigma^2_{N_-} = \mathbf{E}(N_+) = \mathbf{E}(N_-)$$

which corresponds to the expected number of steps to the right or left, or equivalently to the expected number of cycles slipped to the right or left, respectively. Combining,

$$\sigma^2_N = \sigma^2_{N_+} + \sigma^2_{N_-} = \mathbf{E}(N_+) + \mathbf{E}(N_-) \tag{25}$$

The expected number of cycle slips to the right or left per unit time has been shown to be (Ref. 1)

$$\mathbf{E}(N_+) = \mathbf{E}(N_-) = \frac{B_{L_r}}{\pi^2 \alpha_r I_0^2(\alpha_r)} \tag{26}$$

This is the exact result for first-order PLL's and an approximation for higher-order loops. Since

$$\dot{\phi}_r(t) = 2\pi N(t) + \dot{\phi}_m(t) \tag{27}$$

we can write

$$\sigma^2_{\dot{\phi}_r} = 4\pi^2 \sigma^2_N + \sigma^2_{\dot{\phi}_m} = \frac{8B_{L_r}}{\alpha_r I_0^2(\alpha_r)} + \sigma^2_{\dot{\phi}_m} \tag{28}$$

where the additional assumption has been made that $N(t)$ is statistically independent of $\dot{\phi}_m(t)$.

The remaining task is to determine $\sigma^2_{\dot{\phi}_m}$. As previously indicated, models of PLL's have exclusively assumed the additive disturbance to be white and gaussian, with the conclusion that $\dot{\phi}$ is also white. Since doppler measurement is inherently concerned with cycle slipping, all of the various linear theories of PLL's break down when attempting to determine doppler measurement capability. One way, however, in which realistic statistical information can be determined about $\dot{\phi}_m(t)$ is to assume that

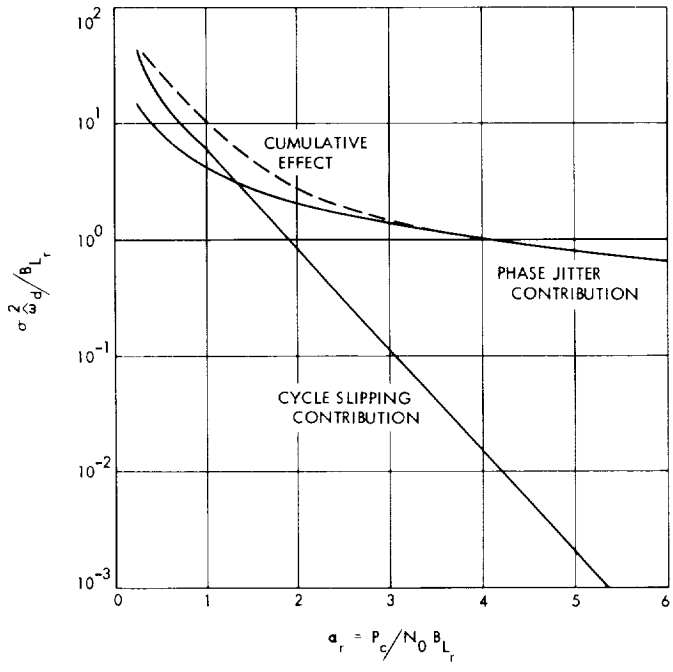


Fig. 12. Doppler error versus carrier loop signal-to-noise ratio

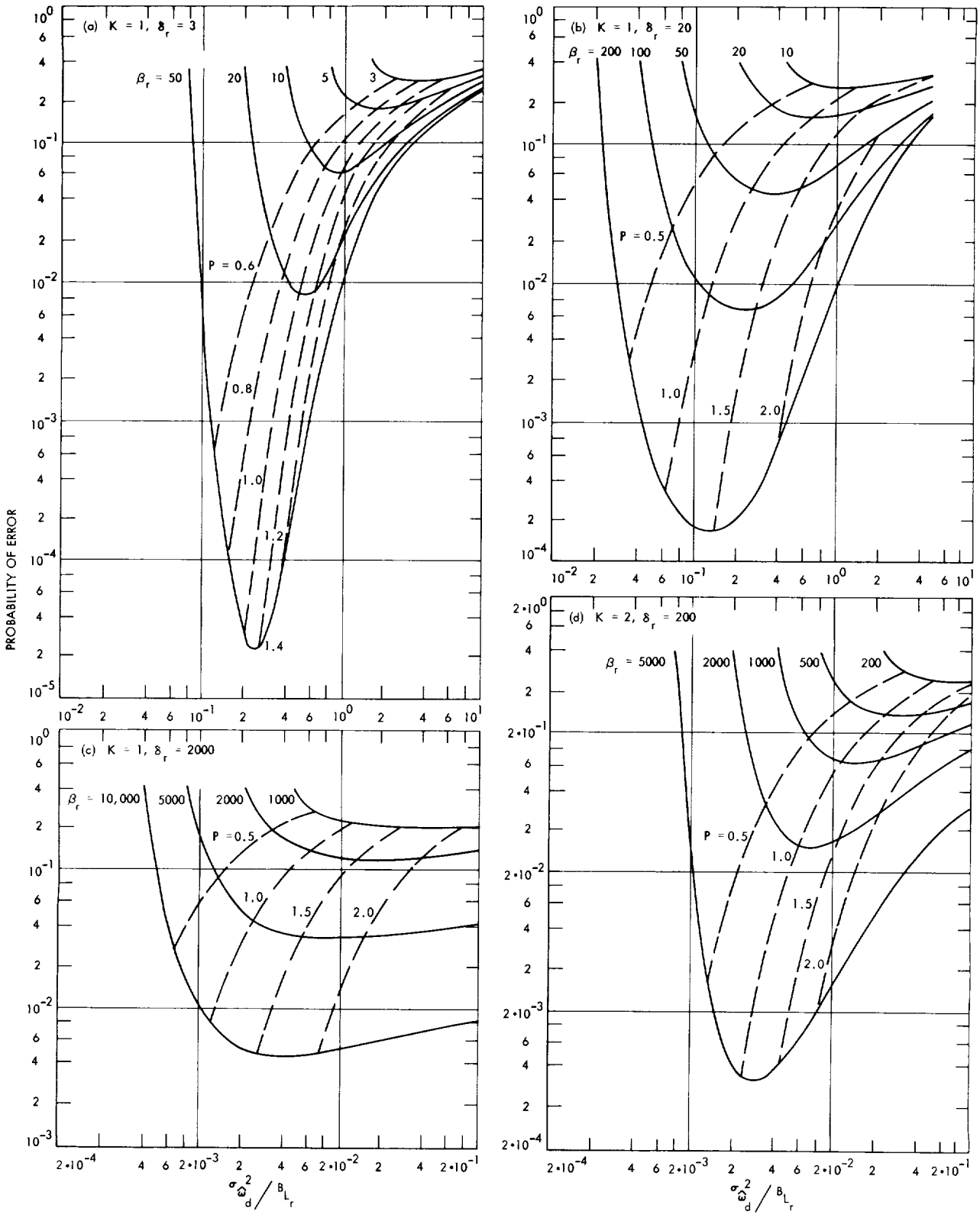


Fig. 13. Error rate versus doppler jitter for various values of K, δ_r, and β_r

the PLL is a second-order loop and that the loop filter is of the form

$$F(s) = \frac{1}{\tau s + 1}$$

With this filter, the Fokker–Planck equation for the probability density function of the Markov vector process $(\phi(t), \dot{\phi}(t))$ can be solved in the steady state (Ref. 5), from which the variance of $\dot{\phi}$, $\sigma^2_{\dot{\phi}}$, is given by

$$\sigma^2_{\dot{\phi}} = \frac{4B_{L_r}}{\alpha_r} \tag{29}$$

At high SNR, the principal disturbance in $\dot{\phi}_r$ is due to $\dot{\phi}_m$, since cycle slips occur rarely at high SNR. Therefore, at high SNR

$$\sigma^2_{\dot{\phi}_r} \simeq \sigma^2_{\dot{\phi}_m} = \frac{4B_{L_r}}{\alpha_r} \tag{30}$$

This representation has an additional assumption imbedded; namely, this result is approximately independent of the structure of the loop filter. These results, namely, the first term in Eq. (28) for the cycle slipping and Eq. (30) for phase jitter are plotted in Fig. 12. Cycle slipping decreases exponentially, while phase jitter between cycle slips decreases only as $(\alpha_r)^{-1}$. Thus, we can qualitatively determine when $\sigma^2_{\dot{\phi}_m}$ becomes the predominant disturbance in doppler measurement. The cumulative effect of cycle slipping and phase jitter is also shown in Fig. I2 as the dotted line.

5. Discussion of Doppler Jitter Versus Error Rate

Using the results of the previous subsections, the tradeoff between doppler measurement capability and error rate can be established. Figures 13a, b, c display the variation of probability of error and variance of doppler measurement normalized by the noise bandwidth of the carrier PLL, B_{L_r}, as a function of the modulation index parameter p for several β_r. These curves are plotted for rates corresponding to $\delta_r = 3$, 20, and 2000. At low data rates, it is noted that the minimums are sharper than at the higher data rates. This says the choice of p is more sensitive at lower data rates than at higher rates.

These curves also indicate the price in performance which is paid by varying p. For example, in Fig. 13a, in which $K = 1$, $\delta_r = 3$, when $\beta_r = 20$, the error rate is minimized when $p \simeq 0.9$. Increasing p beyond 0.9 deteriorates both error rate and doppler measurement, since much more power is going into distortion terms. If p is lowered from 0.9 to 0.6, for example, the error rate increases to a value slightly larger than 10^{-2} while the normalized doppler variance decreases from 0.55 to 0.38.

In Fig. 13d, the tradeoff is similarly presented for two subcarrier channels and $\delta_r = 200$.

It should be noted that, at low data rates and low β_r, the results are inexact as far as error probability is concerned. This is because the assumption that the carrier phase error ϕ_r is constant over the bit time gradually breaks down in this region.

References

1. Viterbi, A. J., *Principles of Coherent Communication*. McGraw-Hill Book Co., Inc., New York, N.Y., 1966.
2. Lindsey, W. C., "Phase-Shift-Keyed Signal Detection with Noisy Reference Signals," *IEEE Trans. Aerospace and Electronics Systems*, Vol. AES-2, pp. 393–401, Jul. 1966.
3. Charles, F. J., and Lindsey, W. C., "Some Analytical and Experimental Phase-Locked Loop Results for Low Signal-to-Noise Ratios," *Proc. IEEE*, Vol. 54, No. 9, Sep. 1966.
4. Doob, J. L., *Stochastic Processes*. John Wiley & Sons, Inc., New York, N.Y., 1953.
5. Tikhonov, V. I., "Phase-Locked Automatic Frequency Control Operation in the Presence of Noise," *Avtomatika i Telemekhanika*, Vol. 21, pp. 301–309, Mar. 1960.

E. Decoding and Synchronization Research: Convolutional Codes: The State-Diagram Approach to Optimal Decoding and Performance Analysis for Memoryless Channels, A. J. Viterbi[2]

1. Introduction

In 1967, Viterbi (Ref. 1) proposed a new nonsequential decoding algorithm for convolutional codes. Subsequently, Forney (Ref. 2) showed this algorithm to correspond to maximum likelihood decisions and thus to be optimal for equiprobable messages. Although the computational complexity of the algorithm is proportional to $K2^K$, where K is the constraint length, Heller (SPS 37-54, Vol. III, pp. 171–177) has shown that with small, and consequently practical values of K, significant improvements can be achieved over much longer block codes. Omura (Ref. 3) also considered the algorithm in a state-space context and showed that it corresponds to a dynamic programming solution of the corresponding control problem.

[2]Consultant, School of Engineering and Applied Sciences, UCLA, Los Angeles, California.

In this article, we present a new description, largely motivated by this previous work, of the algorithm in terms of a linear finite-state machine representation of the encoder and its corresponding state diagram.

2. Linear Finite-State Machine Description of the Encoder

Figure 14 shows a three-stage rate ½ binary convolutional encoder. This may be viewed as a linear finite-state machine of four states (generally 2^{K-1} states for a K-stage encoder) corresponding to the last two input data bits in reverse order. The state diagram shown in Fig. 15 is a directed graph indicating the possible transitions between states, and the branches are labelled according to the code symbols (two corresponding to rate ½) generated in the transition. Thus, if the two preceding data bits were 11 and the next bit is a 0, the machine moves from state 11 to state 01, producing the two code symbols 01.

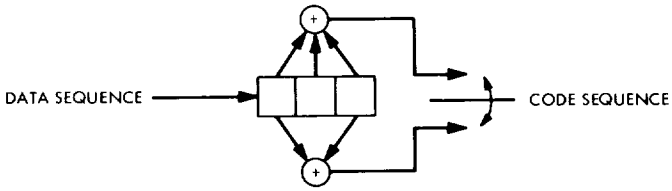


Fig. 14. Binary convolutional encoder ($K = 3$, $R = ½$)

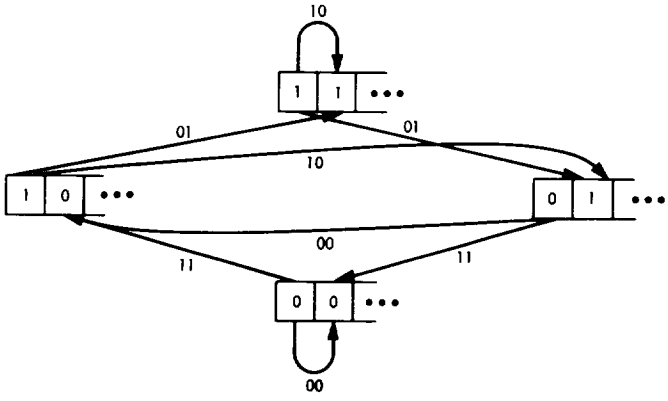


Fig. 15. State diagram of linear finite-state machine

Figure 16 presents the entire communication system employing the convolutional encoder (finite-state machine) just described. We denote in inverse order the input binary data sequence by $\cdots, x_{j+1}, x_j, \cdots, x_2, x_1$, the binary code (channel input) sequence by $\cdots, \mathbf{z}_{j+1}, \mathbf{z}_j, \cdots, \mathbf{z}_2, \mathbf{z}_1$, where $\mathbf{z}_i$ is a two-dimensional vector corresponding to the *two* code symbols generated by the ith data bit, and the code output sequence by $\mathbf{y}_{j+1}, \mathbf{y}_j, \cdots, \mathbf{y}_2, \mathbf{y}_1$, where $\mathbf{y}_i$ is also a two-dimensional vector, but with components from the channel output alphabet, which may even be continuous (e.g., the voltage output of a symbol correlator).

The problem is, of course, to determine the optimal decoder, which produces the data input sequence $\cdots, x_2x_1$ most likely to have generated the received channel output sequence $\cdots, \mathbf{y}_2, \mathbf{y}_1$. To begin, let us suppose hypothetically that the state of the machine at time j (just after the jth input bit) were known by the receiver, and that we sought the input sequence[3] $x_j, \cdots, x_2x_1$ most likely to have produced $\cdots, \mathbf{y}_{j+1}, \mathbf{y}_j, \cdots, \mathbf{y}_2\mathbf{y}_1$. We recognize, first of all, that $\cdots, \mathbf{y}_{j+2}, \mathbf{y}_{j+1}$ are totally irrelevant to determining the necessary likelihood function since they depend only on the state at time j (which is assumed known) and future inputs $\cdots, x_{j+2}, x_{j+1}$, which we are not as yet interested in determining. Thus, for the stated purpose of determining the most likely $x_j, \cdots, x_1$ given the state at time j, we need only compare the 2^{j-2} likelihood functions[3] $p(\mathbf{y}_j, \cdots, \mathbf{y}_2\mathbf{y}_1 | x_j, \cdots x_1)$ for each possible binary sequence $x_{j-2}, \cdots x_1$.

Returning to the actual situation where the state at time j is unknown, we may still proceed as above but repeat the procedure for each of the 2^{K-1} possible states, and thus determine the most likely input data sequence leading to each possible state at time j. We now show that we can reduce the number of likelihood functions which need be compared for each state from 2^{j-2} to merely 2.

[3]Of course, x_j and x_{j-1} are known if the state is known.

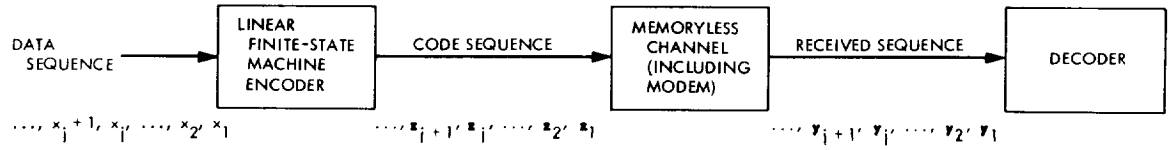


Fig. 16. Communication system employing convolutional coding

For this purpose, we observe first that

$$p(\mathbf{y}_j, \cdots, \mathbf{y}_2\mathbf{y}_1 \mid x_j, \cdots, x_1) = p(\mathbf{y}_{j-1}, \cdots, \mathbf{y}_1 \mid x_j, \cdots, x_1)\, p(\mathbf{y}_j \mid x_j, \cdots, x_1)$$
$$= p(\mathbf{y}_{j-1}, \cdots, \mathbf{y}_1 \mid x_{j-1}, \cdots, x_1)\, p(\mathbf{y}_j \mid x_j, x_{j-1}, x_{j-2}) \quad (1)$$

since the channel is memoryless and $\mathbf{z}_j$, and consequently $\mathbf{y}_j$ depends only on x_j, x_{j-1}, x_{j-2}. Now, if we know, and have stored the most probable input sequence $x_{j-1}^{(i)}, \cdots, x_1^{(i)}$ leading to each possible state $x_{j-1}^{(i)}, x_{j-2}^{(i)}$ at time $j-1$ and its corresponding likelihood function

$$p(\mathbf{y}_{j-1}, \cdots, \mathbf{y}_1 \mid x_{j-1}^{(i)}, \cdots, x_1^{(i)}) \quad (i = 1, 2, \cdots, 2^{K-1})$$

we can find the necessary 2^K likelihood function at time j by using the stored likelihood functions for time $j-1$ and multiplying them by the two single branch likelihood functions, as shown in Eq. (1).

For example, in terms of the code of Fig. 15, to determine the maximum $p(\mathbf{y}_j, \cdots, \mathbf{y}_1 \mid 1, 1, x_{j-2}^{(11)}, \cdots, x_1^{(11)})$ and thus the most probable input sequence leading to state 11 at time j, we need compute and compare the two functions:

$$p(\mathbf{y}_{j-1}, \cdots, \mathbf{y}_1 \mid 1, 0, x_{j-3}^{(10)}, \cdots, x_1^{(10)})\, p(\mathbf{y}_j \mid 110)$$

and

$$p(\mathbf{y}_{j-1}, \cdots, \mathbf{y}_1 \mid 1, 1, x_{j-3}^{(11)}, \cdots, x_1^{(11)})\, p(\mathbf{y}_j \mid 111)$$

Thus, the first term of each product was previously stored (as well as the corresponding data input sequences $1, 0, \cdots, x_1^{(10)}$ and $1, 1, \cdots, x_1^{(11)}$), and the second terms are computed from the just-received channel output vector. Thus, only two computations and one comparison must be made for each state after each new received branch, and the larger likelihood function and its corresponding most probable data input sequence ("survivor" of Ref. 1) must be stored in the corresponding state register (here the open-ended registers in Fig. 15). The algorithm is terminated by $K-1$ "tail" *zeros* which reduce the 2^{K-1} possible states to a single state.

We note particularly, and this is one of the main purposes of this article, that the state diagram serves as a system block diagram for the optimal decoding algorithm, because it indicates the comparisons and transfers required at each step.

3. Performance Analysis by Means of the State Diagram

The second main purpose of this article is to demonstrate that the state diagram can be an effective tool in analyzing performance of a specific convolutional code in any memoryless channel. Since a convolutional code is a group or linear code, we may, without loss of generality, assume that the all *zeros* data input sequence occurred, and thus generated an all *zeros* code sequence. Then an error may occur whenever the likelihood function for any other path in the state diagram which arrives at state $0, \cdots, 0$ at time j is greater than that of the all *zeros* path up to time j.

As a first step, let us determine the Hamming distances of all such incorrect paths from the all *zeros* (correct) path. This will be sufficient to determine a union upper bound for error probability over any memoryless channel, and with a minor modification, an upper bound on bit error probability can also be determined.

In Fig. 17, the state diagram for the convolutional code example pursued above is redrawn with node 00 split open and the code symbols for the given branches replaced by D^k, where k is the weight of the branch symbol sequence ($k = 2$ for 11, $k = 1$ for 01 and 10, and $k = 0$ for 00). The term D^k is followed by (N), when the transition was caused by an input bit which was a "1"; we shall utilize these later to determine bit error probabilities. The transfer function of any given path will be of the form D^i, where i is the total weight (number of code symbols which are "1's") of the path and thus the distance from the all *zeros* (correct) code sequence. For example, the input sequence 00100 gives rise to the path whose transfer function is D^5 and, consequently, whose weight is 5.

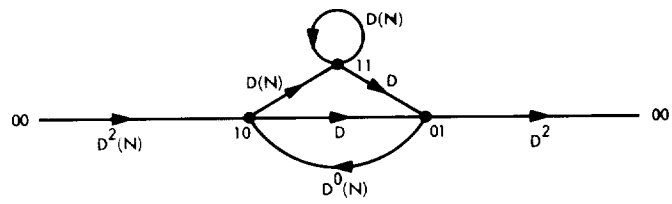


Fig. 17. State diagram of linear finite-state machine encoder relabeled for performance evaluation

The overall transfer function of the directed graph (linear system) is easily obtained:

$$T(D) = \frac{D^5}{1 - 2D}$$
$$= D^5 + 2D^6 + 4D^7 + \cdots + 2^k D^{k+5} + \cdots \quad (2)$$

indicating that there is one incorrect path of weight 5, two of weight 6, four of weight 7, etc. To determine bit error probability, we must also find the number of bit errors resulting from an incorrect path decision. Since the all *zeros* input sequence is assumed, bit errors correspond to "1's" in the incorrect path data sequence. These are indicated by the parenthetical N's; for example, the weight 5 path has a transfer function which is linear in N, corresponding to a single *one* in its data sequence (00100). Now, taking the transfer function in D and N, we have

$$T(D, N) = \frac{D^5 N}{1 - 2DN}$$
$$= D^5 N + 2D^6 N^2 + 4D^7 N^3 + \cdots + 2^k D^{k+5} N^{k+1} + \cdots \quad (3)$$

Thus, there are 2^k paths of weight D^{k+5}, resulting in $k + 1$ bit errors.

From this, we may determine an upper bound on the bit error probability. We shall pursue this example for an additive white gaussian channel and antipodal phase-shift-keyed modulation whereby the code symbol "0" is transmitted as a positive pulse and the code symbol "1" is transmitted as a negative pulse of the same energy. Let the symbol energy be E_s, the bit energy be $E_b = 2E_s$ (since there are two code symbols per data bit for rate ½), and the one-sided noise density be N_0. Then the pairwise error probability for one incorrect path which differs in k symbols from the correct path is upper-bounded (closely) by

$$P \lesssim \exp\left(-k\frac{E_s}{N_0}\right) = \exp\left(-\frac{kE_b}{2N_0}\right)$$

Thus, a union upper bound on the probability of an error at any step is obtained from Eq. (2) to be

$$P_E < D^5 + 2D^6 + 4D^7 + \cdots \Big|_{D=\exp(-E_b/2N_0)}$$
$$= T(D)\big|_{D=\exp(-E_b/N_0)} = \frac{\exp\left(-\frac{5E_b}{2N_0}\right)}{1 - 2\exp\left(-\frac{E_b}{2N_0}\right)} \quad (4)$$

Of greater interest is the bit error probability. Differentiation of Eq. (3) by N, followed by setting $N = 1$, weighs the probability of error by the number of bit errors for the given incorrect path. Thus, the overall bit error is simply

$$P_B = D^5 + 2D^6(2) + 4D^7(3) + \cdots \Big|_{D=\exp(-E_b/2N_0)}$$
$$= \frac{dT(D, N)}{dN}\bigg|_{N=1,\, D=\exp(-E_b/2N_0)}$$
$$= \frac{\exp\left(-\frac{5E_b}{2N_0}\right)}{\left[1 - 2\exp\left(-\frac{E_b}{2N_0}\right)\right]^2} \quad (5)$$

Of course, a union bound is useful only for reasonably high energy-to-noise ratios. In particular, in this case it is clear from Eq. (5) that the bound is useless for $E_b/N_0 < 2 \ln 2$ (1.4 dB), which just happens to correspond to the computational cutoff rate for a white gaussian channel. However, bounds for lower E_b/N_0 (as low as the channel capacity) can be obtained by replacing union bounds by Gallager bounds (Refs. 1 and 4) in the individual terms of Eq. (5).

This same technique can be used for any specific code on any memoryless channel. Unfortunately, as elegant as the method appears, it becomes hopelessly complex to determine the transfer function $T(D, N)$ for an arbitrary code for $K \geq 5$. However, for orthogonal codes (Ref. 5) where all branch weights[4] are equal to $1/2R$, it is easily

[4]This requires $R = 2^{-K}$ and can be instrumented by using an orthogonal block code generator matrix as the convolutional code generator matrix (whose rows are the tap sequences of the $1/R = 2^K$ modulo 2 adders).

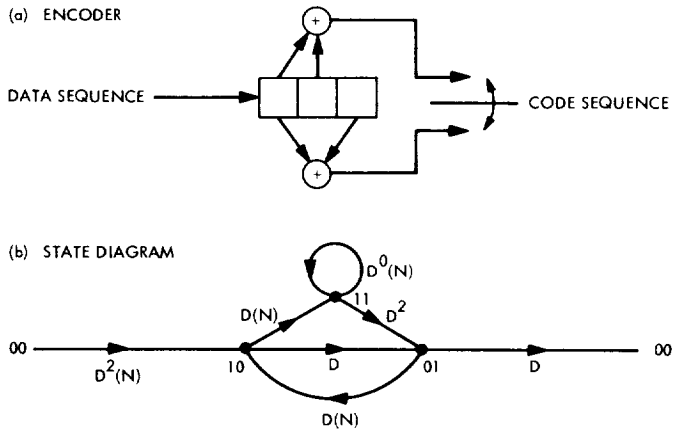


Fig. 18. Encoder displaying catastrophic error propagation

shown that

$$T(D,N) = ND^{K/2R}\frac{1-D^{1/2R}}{1-D^{1/2R}-N(D^{1/2R}-D^{K/2R})} \tag{6}$$

Then, since the symbol energy-to-noise density ratio in this case is RE_b/N_0, we have

$$\begin{aligned} P_B &< \left.\frac{dT(D,N)}{dN}\right|_{N=1,\,D=\exp(-RE_b/N_0)} \\ &= \frac{\exp\left(-\frac{KE_b}{2N_0}\right)\left[1-\exp\left(-\frac{E_b}{2N_0}\right)\right]^2}{\left[1-2\exp\left(-\frac{E_b}{2N_0}\right)+\exp\left(-\frac{KE_b}{2N_0}\right)\right]^2} \\ &\leq \exp\left(-\frac{KE_b}{2N_0}\right)\left[\frac{1-\exp\left(-\frac{E_b}{2N_0}\right)}{1-2\exp\left(-\frac{E_b}{2N_0}\right)}\right]^2 \end{aligned} \tag{7}$$

Note that, for $K = 3$ at high E_b/N_0 ratios, this orthogonal code (which operates at $R = 1/8$) performs considerably worse than the simple $R = 1/2$ code of the preceding example. A biorthogonal convolutional code, which requires a biorthogonal block code generator matrix as its convolutional generator matrix, and hence $R = 2^{-(K-1)}$, has the same state-diagram branch weights except that the initial branch leading from state 0, $\cdots$, 00 to 10, $\cdots$, 0 has weight $1/R$ instead of $1/2R$, resulting in the same transfer function as in Eq. (6) but with the leading term replaced by $ND^{K-1/2R}$. For $K = 3$, this operates at $R = 1/4$ but still yields inferior performance to the preceding $R = 1/2$ code. Regular simplex (or transorthogonal) convolutional codes yield inferior performance to biorthogonal convolutional codes.

4. Catastrophic Error Propagation and Implications for Random Coding Bounds

Massey and Sain (Ref. 6) have defined a catastrophic error as the event that a finite number of code symbol errors cause an infinite number of data bit errors in decoding and have shown that a necessary and sufficient condition for a convolutional code to produce catastrophic errors is that any two rows of the generator matrix, represented as polynomials, have common factors.

In terms of the state diagram and its transfer function, it is clear that catastrophic errors can occur if, and only if, any closed loop has zero weight (i.e., transfer function $D^0 = 1$), because in this case the transfer function will have the factor $1 - D^0 = 0$ in the denominator. More simply, consider, for example, the encoder of Fig. 18 and its corresponding state diagram. The incorrect path

$$\underbrace{001111111,\cdots,100}_{k\ 1\text{'s}}$$

has transfer function N^kD^6 for any value of $k \geq 2$. Thus, it differs from the correct path in only six symbols, no matter how large the value of k. Thus, for a binary symmetric channel, for example, four channel errors could cause an arbitrarily long data error sequence, while for a gaussian channel an arbitrarily large error sequence of any length could occur with probability approximately $\exp(-3E_b/N_0)$.

We observe further that if each row of the generator matrix has even weight (i.e., each modulo -2 adder is connected to an even number of stages) then the all ones self loop will have weight zero for any value of K and, consequently, lead to catastrophic error propagation.

This simple observation has a significant implication on the limitations which must be imposed in deriving upper bounds on the ensemble of convolutional codes of a given constraint length and rate by means of random coding arguments. An apparent limitation on the bounding procedure (Ref. 1) is that we could not consider the ensemble of fixed encoders but had to consider the more general (but less practical) class of variable encoders wherein the generator matrix is modified (randomly reselected) for each new input data bit. In the original proof, it appeared that this was necessary to assure the necessary independence conditions. Using the above observation, we now show that this limitation is fundamental.

In random coding for binary block codes with a uniform measure, the probability of selecting two identical (or nearly identical) codewords for different messages leading to catastrophic error is proportional to 2^{-N}, where N is the (arbitrarily long) code length. Thus, the effect of this form of error on the ensemble error probability decreases exponentially with block length. For convolutional codes, on the other hand, randomly selecting a fixed generator matrix for arbitrary constraint length and rate[5] R leads to catastrophic errors whenever all rows of the matrix have even weight, and this occurs with probability $2^{-1/R}$. Thus, catastrophic errors can occur in a randomly selected fixed convolutional code with a probability

[5]We consider here only rates for which $1/R$ equals an integer, but the argument can be generalized to any rational R.

bounded away from zero *independent* of length. Hence, it follows that the ensemble average error probability for fixed convolutional codes does *not* decrease exponentially with constraint length as does the class of convolutional codes with variable generators (Ref. 1).

References

1. Viterbi, A. J., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inform. Theory*, Vol. IT-13, pp. 260–269, Apr. 1967.
2. Forney, Jr., G. D., *Final Report on a Coding System Design for Advanced Solar Missions*, NASA Contract NAS2-3637. Codex Corporation, Watertown, Mass., Dec. 1967.
3. Omura, J. K., "On the Viterbi Decoding Algorithm," *IEEE Trans. Inform. Theory*, Vol. IT-15, pp. 177–179, Jan. 1969.
4. Gallager, R. G., "A Simple Derivation of the Coding Theorem and Some Applications," *IEEE Trans. Inform. Theory*, Vol. IT-11, pp. 3–18, Jan. 1965.
5. Viterbi, A. J., "Orthogonal Tree Codes for Communication in the Precense of White Gaussian Noise," *IEEE Trans. Commun. Technol.*, Vol. Com-15, pp. 238–242, Apr. 1967.
6. Massey, J. L., and Sain, M. K., "Inverses of Linear Sequential Circuits," *IEEE Trans. Computers*, Vol. C-17, pp. 330–337, Apr. 1968.

F. Coding and Synchronization Studies: The Effect of Amplitude Uncertainty on Estimating Phase of a Square Wave, *S. Butman*

SPS 37-53, Vol. III, pp. 200–209, describes a circuit for estimating the phase of a square wave in the presence of additive white gaussian noise. The circuit is optimum under the restriction that the signal being received must be correlated with no more than two locally generated square waves separated by one quarter of a cycle; and under the assumptions that the amplitude and frequency of the signal, and the spectral density of the noise, are known exactly.

In practice, the effect of a small variation in frequency is equivalent to phase uncertainty and is, therefore, included as part of the phase estimate, without changing the basic structure of the estimator. Likewise, lack of knowledge about the strength of the noise affects only the calculation of the mean-squared error, but does not affect the structure of the estimator. However, exact knowledge of the ampltitude of the received waveform is essential to the implementation. The amplitude must be known in order to set the saturation levels of the soft limiter in the circuit, which is shown in Fig. 19. In SPS 37-53, Vol. III, the amplitude of the incoming signal was assumed to be equal to one, because if the amplitude is known to be A, it is necessary only to pass the received signal through an amplifier of gain $1/A$ and set the soft limiter to saturate at $+1$ and -1. If A is not known and the gain is set at some nominal value, say $1/A_0$ where A_0 is our best estimate of A, then the soft limiter should saturate at $\pm A/A_0$ and not at ± 1. However, since we do not know what A/A_0 is, it is no longer possible to build the circuit. If we insist on keeping the saturation levels at ± 1, the circuit is no longer optimum.

One possibility is to estimate A from the correlation outputs x and y. From SPS 37-53, Vol. III, we know that

$$\left.\begin{aligned} x &= \bar{x}(\tau) + n_x \\ y &= \bar{y}(\tau) + n_y \end{aligned}\right\} \tag{1}$$

where τ is the unknown delay, or phase, n_x and n_y are independent gaussian noises having zero mean and variance σ^2, while $\bar{x}$ and $\bar{y}$ satisfy

$$|\bar{x}| + |\bar{y}| = A$$

We also know that the equation for the best estimate of τ when A is known is (SPS 37-53, Vol. III)

$$\hat{\tau} = \frac{T}{4}\left\{1 - \frac{1}{2}\left[1 + \text{sat}\left(\frac{|x|}{A} - \frac{|y|}{A}\right)\right]\text{sgn}\, x\right\}\text{sgn}\, y \tag{2}$$

If we now use the estimate

$$A_0 = |x| + |y|$$

for the true value of A in Eq. (2), we obtain

$$\hat{\tau} = \frac{T}{4}\left\{1 - \frac{1}{2}\left(1 + \frac{|x| - |y|}{|x| + |y|}\right)\text{sgn}\, x\right\}\text{sgn}\, y$$

due to the fact that

$$\text{sat}\left(\frac{|x| - |y|}{|x| + |y|}\right) = \frac{|x| - |y|}{|x| + |y|}$$

Therefore,

$$\hat{\tau} = \frac{T}{4}\left(1 - \frac{x}{|x| + |y|}\right)\text{sgn}\, y \tag{3}$$

$$= \frac{T}{4}\begin{cases} \dfrac{y}{|x| + |y|} & \text{if } x > 0 \quad (4a) \\ -\dfrac{y}{|x| + |y|} + 2\,\text{sgn}\, y & \text{if } x < 0 \quad (4b) \end{cases}$$

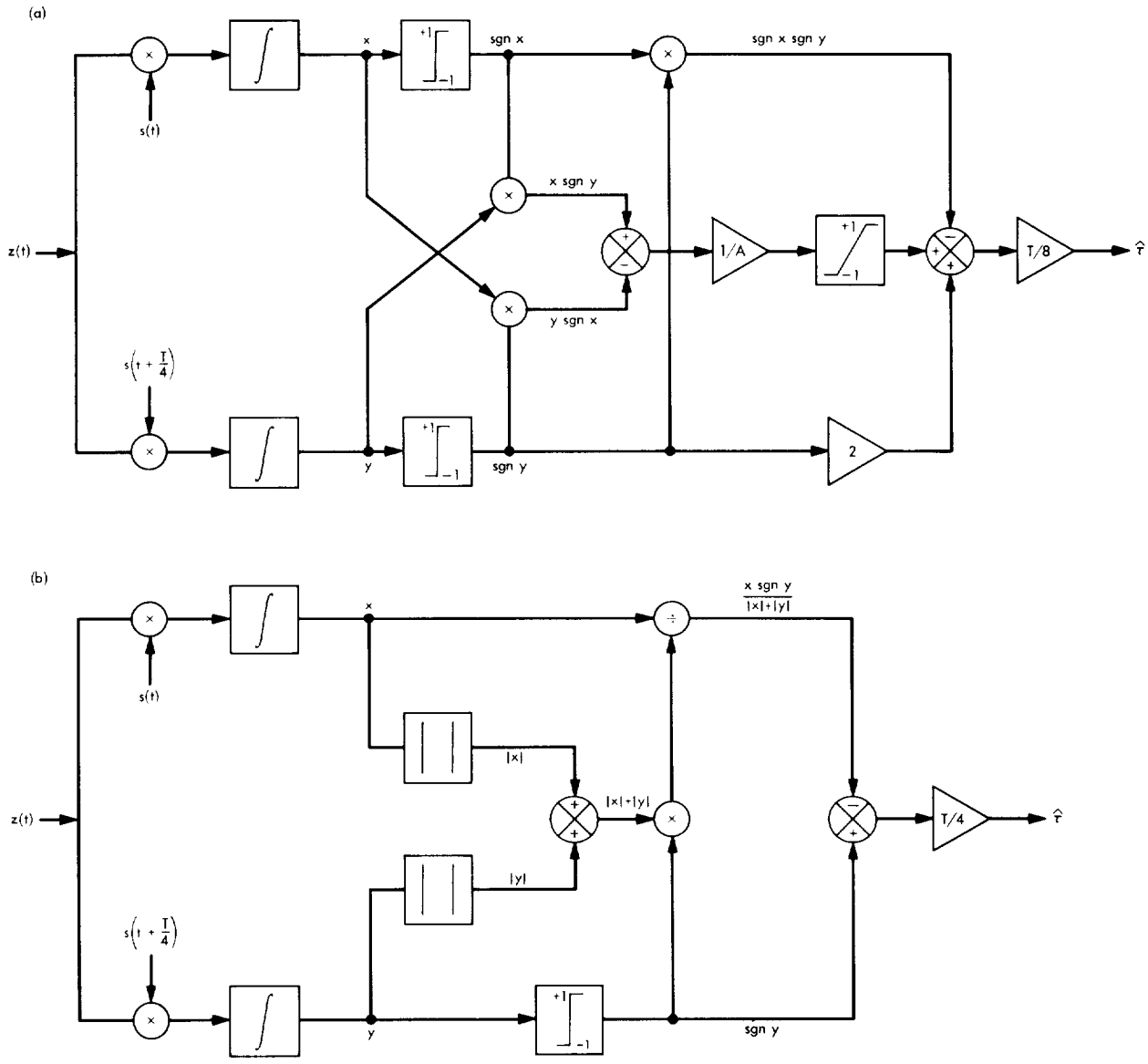


Fig. 19. Square-wave phase estimator mechanization: (a) signal amplitude is known, (b) signal amplitude is unknown

This turns out to be the estimator suggested in SPS 37-52, Vol. II, pp. 46–49. We can use Eq. (3) or Eqs. (4a) and (4b) to estimate the mean-squared error performance of this scheme at high signal-to-noise ratios, that is, when $\sigma^2/A^2 << 1$. We note that during tracking the expected value of τ is zero with high probability. Therefore, $\bar{x} \simeq A$, $\bar{y} \simeq 0$ and we may use Eq. (4a) to write

$$\frac{\hat{\tau} - \tau}{T} = \frac{1}{4}\left(\frac{\bar{y} + n_y}{A + n_x + |\bar{y} + n_y|} - \frac{\bar{y}}{A}\right)$$

$$\simeq \frac{1}{4A}(n_y)$$

Thus, the mean-squared error is

$$\left(\frac{\hat{\tau} - \tau}{T}\right)^2 \simeq \frac{\sigma^2}{16A^2}$$

which is three times greater than the value $\sigma^2/(48A^2)$ predicted in SPS 37-53, Vol. III, for the case when A is known.

The conclusion to be drawn from the above discussion is that it pays to have accurate information about side parameters such as the signal amplitude A. Thus, if there

is some way of determining A more accurately than from the measurements x and y alone, e.g., from previous measurements, it is worthwhile to incorporate this extra knowledge into the structure of the phase estimator.

G. Coding and Synchronization Studies: The Globally Optimal M-ary Noncoherent Digital Communication System, C. L. Weber[6]

Recently, it was shown that with no dimensionality constraint, the noncoherent orthogonal signal structure locally minimizes the probability of error for all signal-to-noise ratios (Refs. 1 and 2). It was also conjectured that the orthogonal signal set was the globally optimum solution to the problem as well. It is herein shown that the orthogonal signal structure is the globally optimum signal set for large signal-to-noise ratios when there is no dimensionality constraint. The result thereby supports the conjecture stated above.

It is assumed, as was done previously, that the transmittable signal set consists of M equipowered, equiprobable narrowband waveforms defined on the time interval $(0, T)$, which are of the following form:

$$s_j(t, \theta) = A_j(t) \cos [\omega_c t + \phi_j(t) + \theta] \tag{1}$$

where $A_j(t)$ and $\phi_j(t)$ are the envelope and the phase of the jth signal, respectively, θ is a random variable uniformly distributed in $(0, 2\pi)$, and $\omega_c/2\pi$ is the carrier frequency. Synchronization jitter is assumed to be negligible.

The received waveform is then

$$y(t) = V s_j(t, \theta) + n(t) \tag{2}$$

where V is the channel scale factor and $n(t)$ is a sample function from a white gaussian noise process with one-sided spectral density N_0 watts/hertz. With the narrowband assumption, the signal energy is given by

$$E = V^2 \int_0^T s_j^2(t, \theta)\, dt = \frac{V^2}{2} \int_0^T A_j^2(t, \theta)\, dt, \qquad j = 1, \cdots, M \tag{3}$$

When the optimum receiver is used with a given signal set, it can be shown (Ref. 2) that the probability of error is a function of the following parameters:

(1) The signal waveform inner products,

$$\alpha_{ij} \triangleq \frac{V^2}{2E} \int_0^T A_i(t) A_j(t) \cos [\phi_i(t) - \phi_j(t)]\, dt$$

$$\beta_{ij} \triangleq \frac{V^2}{2E} \int_0^T A_i(t) A_j(t) \sin [\phi_i(t) - \phi_j(t)]\, dt, \qquad i, j = 1, \cdots, M \tag{4}$$

(2) The resulting signal inner product matrix,

$$\Gamma \triangleq \left[\begin{array}{c|c} A & B' \\ \hline B & A \end{array}\right] \tag{5}$$

where

$$A = \| \alpha_{ij} \|, \alpha_{ii} = 1, \alpha_{ij} = \alpha_{ji}$$

$$B = \| \beta_{ij} \|, \beta_{ii} = 0, \beta_{ij} = -\beta_{ji}$$

and the prime denotes matrix transpose. The range of values that α_{ij} and β_{ij} can take on is specified by the constraint that all admissible Γ matrices must be non-negative definite.

(3) The ratio of signal energy to the noise spectral density, defined as signal-to-noise ratio,

$$\lambda^2 \triangleq 2E/N_0$$

Furthermore, at high signal-to-noise ratios, it can be shown that the probability of error is approximately (Ref. 3)

$$P_E(\lambda, \Gamma) \cong \frac{1}{M} \sum_{\substack{i=1 \\ i \neq j}}^{M} \sum_{j=1}^{M} H(\lambda, \delta_{ij}) \tag{6}$$

where

$$H(\lambda, \delta_{ij}) = Q\left\{\frac{\lambda [1 - (1 - \delta_{ij}^2)^{1/2}]^{1/2}}{2}, \frac{\lambda [1 + (1 - \delta_{ij}^2)^{1/2}]^{1/2}}{2}\right\} - \frac{1}{2} e^{-\lambda^2/4} I_0\left(\frac{\lambda^2 \delta_{ij}}{4}\right) \tag{7}$$

$I_0(x)$ is the modified Bessel function of the first kind of order zero,

$$\delta_{ij} = (\alpha_{ij}^2 + \beta_{ij}^2)^{1/2}, \qquad 0 \leq \delta_{ij} \leq 1 \tag{8}$$

[6]Consultant, Electrical Engineering Department, University of Southern California, Los Angeles, California.

and $Q(a, b)$ is called Marcum's Q-function (Ref. 4), defined as

$$Q(a, b) = \int_b^\infty t\, e^{-(t^2+a^2)/2} I_0(at)\, dt \tag{9}$$

Within the above framework, we now show that the orthogonal signal structure is globally optimum at high signal-to-noise ratios by demonstrating that

$$P_E(\lambda, \Gamma) - P_E(\lambda, \Gamma_0) > 0 \tag{10}$$

for large λ, and for all admissible Γ, where Γ_0 denotes orthogonal signal structure.

$$\Gamma_0 = \left[\begin{array}{cc:cc} 1 & 0 & & \\ 0 & 1 & \multicolumn{2}{c}{0} \\ \hdashline & & 1 & 0 \\ \multicolumn{2}{c:}{0} & 0 & 1 \end{array}\right] \tag{11}$$

Using Eq. (6) in Eq. (10) yields

$$\frac{1}{M} \sum_{\substack{i=1 \\ i \neq j}}^{M} \sum_{j=1}^{M} [H(\lambda, \delta_{ij}) - H(\lambda, 0)] > 0 \tag{12}$$

It can also be shown that (Ref. 3)

$$H(\lambda, 0) < H(\lambda, \delta_{ij}) < \tfrac{1}{2}, \qquad \text{for all } \delta_{ij} \epsilon (0, 1) \tag{13}$$

Therefore, each term in Eq. (12) is positive, thus proving our claim.

References

1. Scholtz, R. A., and Weber, C. L., "Signal Design for Phase-Incoherent Communications," *IEEE Trans. Inform. Theory*, Vol. IT-12, pp. 456–463, Oct. 1966.
2. Weber, C. L., *Elements of Detection and Signal Design*. McGraw-Hill Book Co., Inc., New York, N.Y., 1968.
3. Stone, M. S., *Signal Design for the Noncoherent Channel*, Technical Report 326, Communication Systems Research Laboratory, University of Southern California, Los Angeles, Calif., Apr. 1969.
4. Marcum, J. I., "A Statistical Theory of Target Detection by Pulsed Radar," *IEEE Trans. Inform. Theory*, Vol. IT-6, pp. 145–267, Apr. 1960.

H. Coding and Synchronization Studies: The Performance of Second-Order Loops and Phase-Coherent Communication Systems,

W. C. Lindsey[7] and M. K. Simon[8]

1. Introduction

In a previous article, SPS 37-56, Vol. III, pp. 104–118, the authors discussed the effects which tracking loop stress and noisy phase reference jitter produce on the performance of phase-coherent communication systems. It was shown that, for a second-order loop, two parameters, viz., α and β, served to characterize the error probability; however, any attempt to evaluate link performance as a function of loop signal-to-noise ratio ρ and normalized loop detuning $\Lambda = |\Omega_0|/AK$ required further analysis by means of a digital computer. This article carries the analysis to the point where one may relate ρ and Λ to α and β without the use of a digital computer. Hence, the error probability as a function of three basic communication system parameters ρ, Λ, and ST_b/N_0 may be determined and/or studied.

2. System Model

For a second-order phase-locked loop, it has been shown (Ref. 1) that the steady-state probability distribution $p(\phi)$ phase angle ϕ satisfies

$$p(\phi) = \frac{\exp[\alpha(\cos\phi + \gamma\phi)]}{4\Pi^2 \exp[-\Pi\beta]\, |I_{j\beta}(\alpha)|^2} \times \int_\phi^{\phi+2\Pi} \exp[-\alpha(\cos x + \gamma x)]\, dx \tag{1}$$

when the loop filter $F(s)$ has transfer function

$$F(s) = \frac{1 + \tau_2 s}{1 + \tau_1 s}$$

and

$$\left.\begin{aligned} \alpha &= \left(\frac{r+1}{r}\right)\rho - \frac{1}{r\sigma_G^2}, \qquad \gamma = \frac{\beta}{\alpha} \\ \beta &= \left(\frac{r+1}{r}\right)\frac{\rho}{F_1}\left[\frac{\Omega_0}{AK} - (1 - F_1)\overline{\sin\phi}\right] \end{aligned}\right\} \tag{2}$$

[7]Consultant, University of Southern California, Los Angeles, California.

[8]Member of JPL Spacecraft Telecommunications Systems Section.

for all ϕ in an interval of width 2π centered about any $2n\pi$, where n is any integer. In Eq. (2), the parameters are defined by

$$\left.\begin{aligned}
\Omega_0 &= \omega - \omega_0 \\
r &= AK\frac{\tau_2^2}{\tau_1} \\
F_1 &= \frac{\tau_2}{\tau_1} \\
\sigma_G^2 &= \overline{\sin^2\phi} - (\overline{\sin\phi})^2 \\
AK &= \text{loop gain} \\
\rho &= \frac{2A^2}{N_0 W_L} = \text{signal-to-noise ratio in loop bandwidth} \\
W_L &= \frac{r+1}{2\tau_2} = \text{loop bandwidth} \\
I_\nu(x) &= \text{modified Bessel function of the first kind with complex order } \nu \text{ and argument } x
\end{aligned}\right\} \quad (3)$$

The β factor in Eq. (1) is a measure of the *loop stress*. If the loop is designed such that $\beta \simeq 0$, then we see that $p(\phi)$ is symmetric. Moreover, the *circular moments* of $p(\phi)$ are

$$\overline{\cos n\phi} = \mathrm{Re}\left\{\frac{I_{n-j\beta}(\alpha)}{I_{-j\beta}(\alpha)}\right\}, \qquad \overline{\sin n\phi} = \mathrm{Im}\left\{\frac{I_{n-j\beta}(\alpha)}{I_{-j\beta}(\alpha)}\right\} \quad (4)$$

where $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote, respectively, the "real part of" and the "imaginary part of" the bracketed quantities. The average error probability for a binary phase-shift-keyed system was shown (SPS 37-56, Vol. III) to be represented by

$$P_e = \int_{\pi}^{\pi} p(\phi)\,\mathrm{erfc}\,[(2R)^{1/2}\cos\phi]\,d\phi \quad (5)$$

for the case where ϕ varies slowly over the signal interval T_b and by

$$P_e = \mathrm{erfc}\,[(2R)^{1/2}(\overline{\cos\phi})] \quad (6)$$

when ϕ varies rapidly over the interval T_b. In Eqs. (5) and (6), the parameter $R = ST_b/N_0$ is the signal-to-noise ratio in the data stream and

$$\mathrm{erfc}\,x = \frac{1}{(2\pi)^{1/2}}\int_x^\infty \exp\left(-\frac{z^2}{2}\right)dz \quad (7)$$

is the complementary error function.

3. System Performance Versus β/α for a Fixed α

The basic set of tracking loop performance curves are illustrated in Figs. 20 and 21. Here, we plot ρ versus $\Lambda = |\Omega_0|/AK$ with α and $|\gamma|$ acting as parameters. Notice that $|\gamma| \simeq |\Omega_0|/AK$ for $\rho > 3$ dB. Thus, for a given ρ and Λ one finds that point in the (ρ, Λ) plane corresponding to $(|\gamma|, \alpha)$ in the loop. Upon using these values for $|\gamma|$ and α and entering the grids corresponding to Eqs. (5) and (6) (plotted in SPS 37-56, Vol. III), one easily determines the system error probability.

For illustrative purposes, we plot $\sin\phi$ and σ_G^2 versus $|\gamma|$ for various α in Figs. 22 and 23. Notice that $\sin\phi$ and σ_G^2 are, respectively, monotonic increasing and decreasing for a fixed α up to a point in $|\gamma|$ for which they reverse their trends. Holding α fixed and varying $|\gamma|$ is equivalent to varying the loop stress. It must be pointed out that any attempt to extend the grids of Figs. 20 and 21 beyond those values of Λ shown produces numerical results which become suspect. In this region, the approximations used to estimate the conditional expectation $E\{y_1|\phi\}$ in Ref. 1 undoubtedly break down. Thus, further analysis for other values of Λ can proceed only after one produces a more exact estimate of this function.

Finally, Fig. 24 shows a plot of

$$\overline{\cos\phi} = \mathrm{Re}\left[\frac{I_{1-j\beta}(\alpha)}{I_{-j\beta}(\alpha)}\right] \quad (8)$$

while Fig. 25 illustrates

$$\sigma^2_{\cos\phi} = \overline{\cos^2\phi} - (\overline{\cos\phi})^2 \quad (9)$$

These results are of interest in evaluating Eq. (6), average lock-detector outputs, as well as calibrating automatic gain control measurements in the Deep Space Instrumentation Facility. It is interesting to note in Fig. 24 that there exists a focal point in the neighborhood of $|\gamma| \simeq 0.9$.

Reference

1. Lindsey, W. C., *Nonlinear Analysis and Synthesis of Generalized Tracking Systems*, USCEE 317. University of Southern California, Los Angeles, Calif., Dec. 1968.

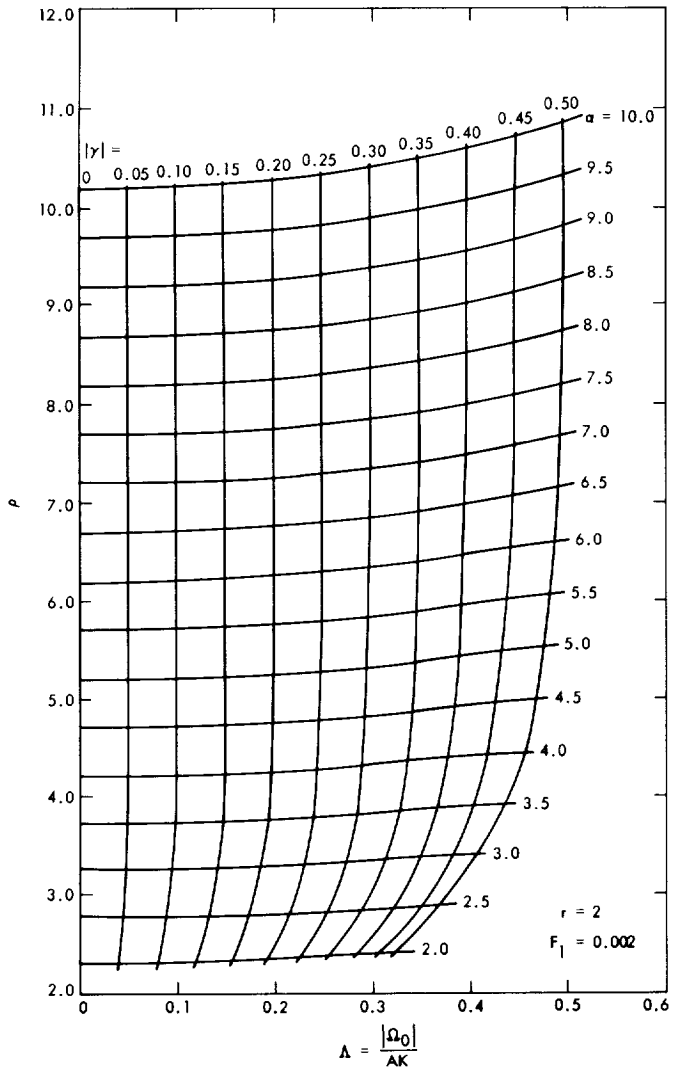


Fig. 20. Loop signal-to-noise ratio ρ versus $|\Omega_0|/AK$ with α and $|\gamma|$ as parameters ($r = 2$)

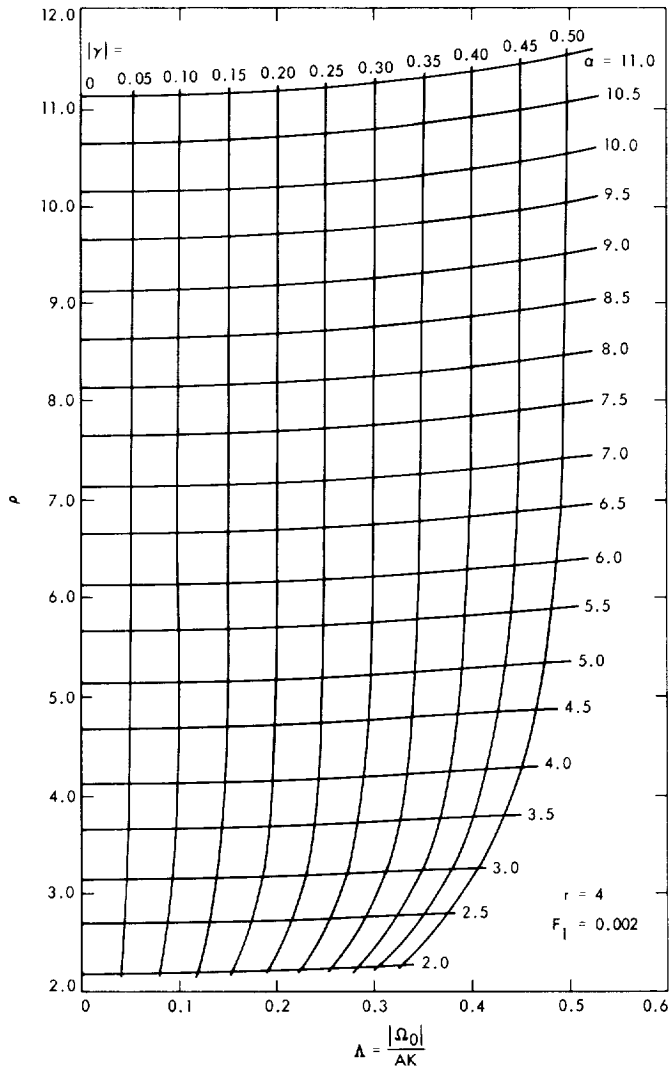


Fig. 21. Loop signal-to-noise ratio ρ versus $|\Omega_0|/AK$ with α and $|\gamma|$ as parameters ($r = 4$)

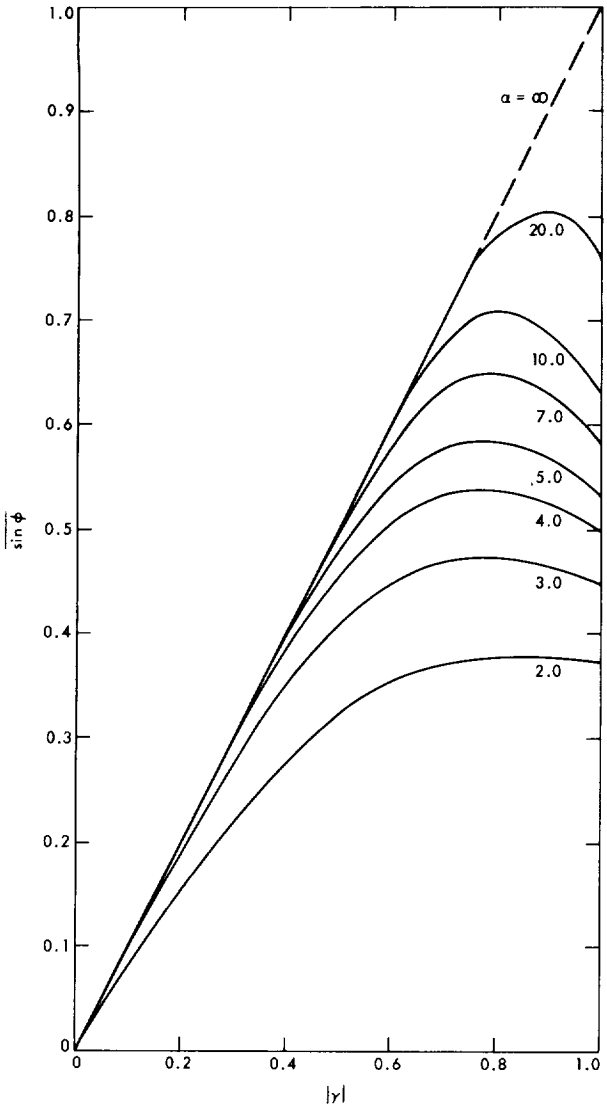


Fig. 22. Sin ϕ vs $|\gamma|$ for various values of α

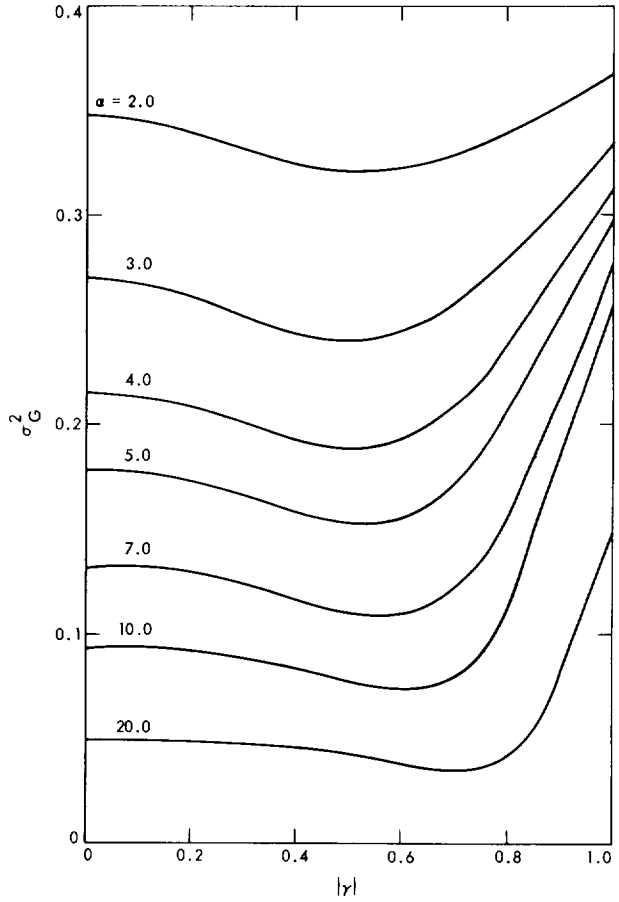


Fig. 23. σ_G^2 vs $|\gamma|$ for various values of α

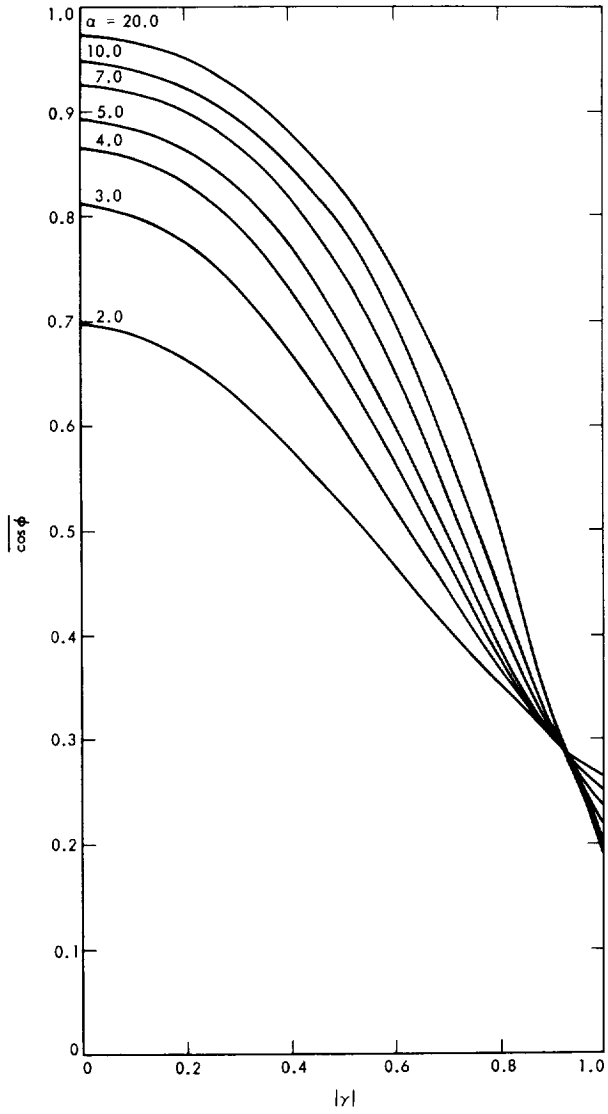


Fig. 24. $\overline{\text{Cos}\ \phi}$ vs $|\gamma|$ for various values of α

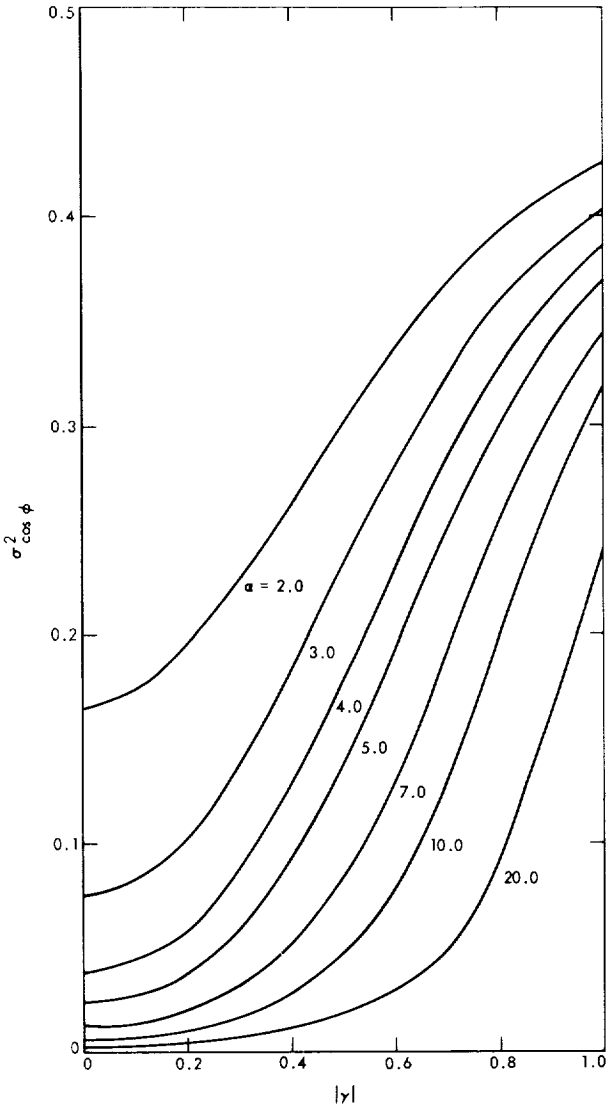


Fig. 25. $\sigma^2_{\cos\phi}$ vs $|\gamma|$ for various values of α

I. Coding and Synchronization Studies: Moments of the First Passage Time in Generalized Tracking Systems, W. C. Lindsey[9]

1. Introduction

This article summarizes the results obtained in Ref. 1 relative to the problem of cycle slipping in generalized tracking systems. In particular, recursive formulas are derived for the moments of the first passage time of the projections of the vector Markov process $\mathbf{y} = (\phi, y_1, \cdots, y_N)$ which arises in a generalized tracking system. Numerical results are given for first- and second-order systems in the presence of loop detuning and initial phase offset. Since the derivation of the results are quite lengthy, we only present the results.

2. Mathematical Model for a Generalized Tracking Loop

As discussed in Ref. 1, the stochastic differential equation of operation of a generalized tracking system is given by

$$\dot{\phi}(t) = \dot{\theta}(t) - KF(p)\,[Ag(\phi) + n(t)] \tag{1}$$

The parameters in this equation are discussed in Ref. 1. We write the loop filter transfer function in the partial fraction expansion

$$F(p) = F_1 + \sum_{k=1}^{N} \frac{1 - F_k}{1 + \tau_k p} \tag{2}$$

Then Eq. (1) can be replaced by the equivalent system of $N + 1$, first-order stochastic differential equations.

$$\left.\begin{aligned}
\dot{y}_0 &= \phi \\
\dot{y}_0 &= \dot{\phi} = \Omega_0 - F_1 K\,[Ag(\phi) + n(t)] + \sum_{k=1}^{N} y_k \\
\dot{y}_1 &= -\frac{y_1}{\tau_1} - \frac{(1 - F_1) K\,[Ag(\phi) + n(t)]}{\tau_1} \\
&\vdots \\
\dot{y}_k &= -\frac{y_k}{\tau_k} - \frac{(1 - F_k) K\,[Ag(\phi) + n(t)]}{\tau_k} \\
&\vdots \\
\dot{y}_N &= -\frac{y_N}{\tau_N} - \frac{(1 - F_N) K\,[Ag(\phi) + n(t)]}{\tau_N}
\end{aligned}\right\} \tag{3}$$

where we have assumed that $\theta(t) = \Omega_0 t + \theta$ and $\Omega_0 = \omega - \omega_0$ is the loop detuning. Written this way, it is clear from Eq. (3) that the projections $y_0, y_1, \cdots, y_N$ form components of an $(N + 1)$-dimensional Markov vector $\mathbf{y} = [y_0, \cdots, y_N]$. Therefore, we have the Fokker–Planck apparatus at our disposal to determine the moments of the first passage time.

3. Moments of the First Passage Time of the kth Projection

In this subsection, we present recursive formulas for the random time $T(y_k)$ it takes for the kth projection of $\mathbf{y}(t)$ to exceed either of the barriers $y_k = \pm y_{k\ell}$ for the first time when the initial position at time $t = t_0$ is $y_k = y_{k0}$. Stated another way, we present expressions for the moments of the expected time $E[T^n(y_k)]$ before one of the barriers is crossed for the first time. This notion is depicted in Fig. 26 for the kth projection. As illustrated, when $y_k(t) = \pm y_{k\ell}$ the trajectory is "absorbed" for all $t > T(y_k)$.

Consider first the projection $y_0(t) = \phi(t)$. In Ref. 1, it is shown that the nth moment $\tau^n(\phi_\ell)$ of the first passage

[9]Consultant, University of Southern California, Los Angeles, California.

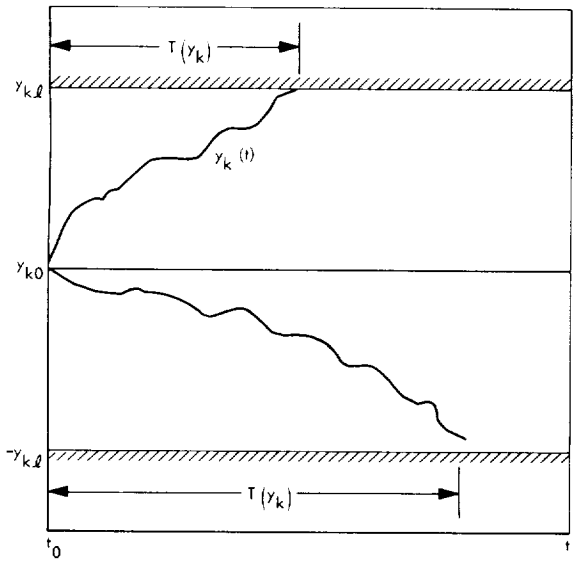


Fig. 26. Trajectory for the kth projection vs time

time for the ϕ process to pass the barriers $y_{kl} = \pm\phi_l$ is given by the recursive formula

$$\tau^n(\phi_l) = \frac{4}{N_0 F_1^2 K^2}\int_{-\phi_l}^{\phi_l} \exp\left[\int^{\phi} h(u)\,du\right] \times \int_{-\phi_l}^{\phi} [C_0'(n-1) - \tau^{n-1}(x)] \times \exp\left[-\int^{x} h(u)\,du\right] dx\,d\phi \tag{4}$$

where

$$\tau^0(x) = u(x - \phi_0)$$

the unit step at ϕ_0, and

$$C_0'(n) = \frac{\int_{-\phi_l}^{\phi_l} \tau^n(x) \exp\left[-\int^{x} h(u)\,du\right] dx}{\int_{-\phi_l}^{\phi_l} \exp\left[-\int^{x} h(u)\,du\right] dx} \tag{5}$$

The function $h(u)$ is defined by

$$h(\phi) = \frac{4}{N_0 F_1^2 K^2}\left[\Omega_0 - F_1 AK g(\phi) + \sum_{k=1}^{N} E(y_k, \bar{t}\,|\,\phi)\right] \tag{6}$$

and $E(y_k, \bar{t}\,|\,\phi)$ is the conditional expectation of y_k given ϕ at time $\bar{t}$. It is clear from Eqs. (4) and (6) that the moments of the first passage time are embedded in a knowledge of the set of expectations $E(y_k, \bar{t}\,|\,\phi), k = 1,2, \cdots, N$.

It is shown in Ref. 1 that the moments of the first passage time of the projections $y_k, k = 1, 2, \cdots, N$ are given by

$$\tau^n(y_{kl}) = \frac{1}{C_k}\int_{y_{kl}}^{y_{kl}} \exp\left[\int^{y_k} g_k(u)\,du\right] \times \int_{-y_{kl}}^{y_k} [C_k'(n-1) - \tau^{n-1}(x)] \times \exp\left[-\int^{x} g_k(u)\,du\right] dx\,dy_k \tag{7}$$

where

$$C_k = (1 - F_k)^2 \frac{N_0 K^2}{4\tau_k^2}$$

$$\tau^0(x) = u(x - y_{k0})$$

and

$$C_k'(n) = \frac{\int_{-y_{kl}}^{y_{kl}} \tau^n(x) \exp\left[-\int^{x} g_k(u)\,du\right] dx}{\int_{-y_{kl}}^{y_{kl}} \exp\left[-\int^{z} g_k(u)\,du\right] dx} \tag{8}$$

In Eq. (7), the $g_k(y_k)$ functions are defined by

$$g_k(y_k) = -\frac{1}{c_k \tau_k}\{y_k + (1 - F_k) AKE[g(\phi), \bar{t}\,|\,y_k]\} \tag{9}$$

To obtain numerical results for the moments, it is necessary to consider specific loop mechanizations.

3. The First-Order Phase-Locked Loop

Here, we set $N = 0$, $W_L = AK/2$ and $g(\phi) = \sin\phi$. From Eq. (6), we then write

$$h(\phi) = \frac{\rho\Omega_0}{AK} - \rho\sin\phi \tag{10}$$

so that Eq. (4) reduces to

$$W_L \tau^n(\phi_l) = \frac{\rho}{2}\int_{\phi_l}^{\phi_l}\int_{-\phi_l}^{\phi} \exp\left\{\rho\left[(\cos\phi - \cos x) + \frac{\Omega_0}{AK}(\phi - x)\right]\right\} [C_0'(n-1) - \tau^{n-1}(x)]\,dx\,d\phi \tag{11}$$

This generalizes Viterbi's result (Ref. 2) for a first-order loop. If we set $n = 1$, $\phi_l = 2\pi$, Eq. (11) reduces to

$$W_L \tau(2\pi) = \frac{\rho}{2}\int_{-2\pi}^{2\pi}\int_{-2\pi}^{\phi} \exp\left\{\rho\left[(\cos\phi - \cos x) + \frac{\Omega_0}{AK}(\phi - x)\right]\right\} [C_0'(0) - u(x - \phi_0)]\,dx\,d\phi \tag{12}$$

where

$$C_0'(0) = \frac{\int_{\phi_0}^{2\pi} \exp\left[-\rho\left(\cos\phi + \frac{\Omega_0\phi}{AK}\right)\right] d\phi}{\int_{-2\pi}^{2\pi} \exp\left[-\rho\left(\cos\phi + \frac{\Omega_0}{AK}\phi\right)\right] d\phi} \tag{13}$$

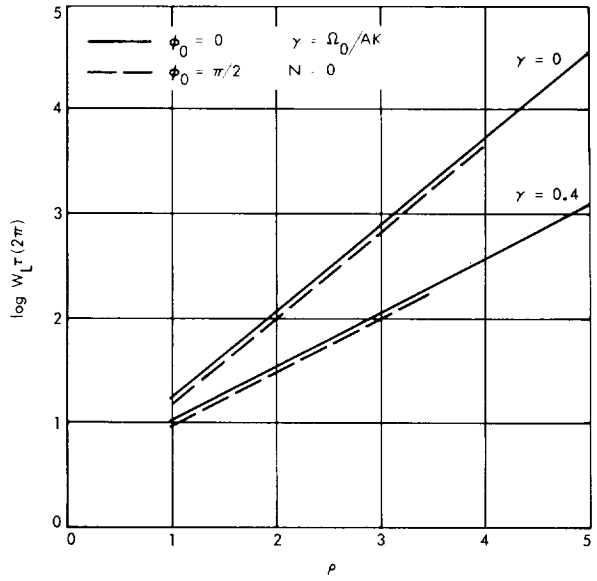


Fig. 27. Plots of Eq. (12) vs ρ for various values of γ and ϕ_0

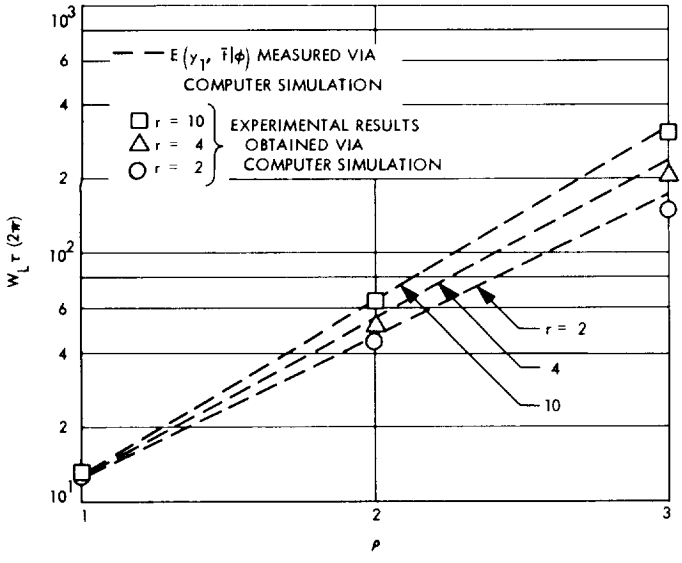


Fig. 28. Mean time to first slip-bandwidth product vs loop signal-to-noise ratio for various values of r

Finally, letting $\phi_0 = \Omega_0 = 0$, we obtain the well-known result

$$W_{L\tau}(2\pi) = \pi^2 \rho I_0^2(\rho) \tag{14}$$

where $\rho = 2A^2/N_0W_L$, $W_L = AK/2$. Plots of Eq. (12) are given in Fig. 27 for various values of ρ, $\gamma = |\Omega_0|/AK$, and ϕ_0.

4. The Second-Order Phase-Locked Loop

Here, we set $N = 1$ and $g(\phi) = \sin\phi$. To evaluate Eq. (4), we need an expression for the conditional expectation $E(y_1, \bar{t}|\phi)$. In SPS 37-43, Vol. III, pp. 76–80, it was shown by the method of computer simulation that certain functions $B(\phi)$ were linear. This suggests that the linear phase-locked loop theory be used to approximate $E(y_1, \bar{t}|\phi)$.

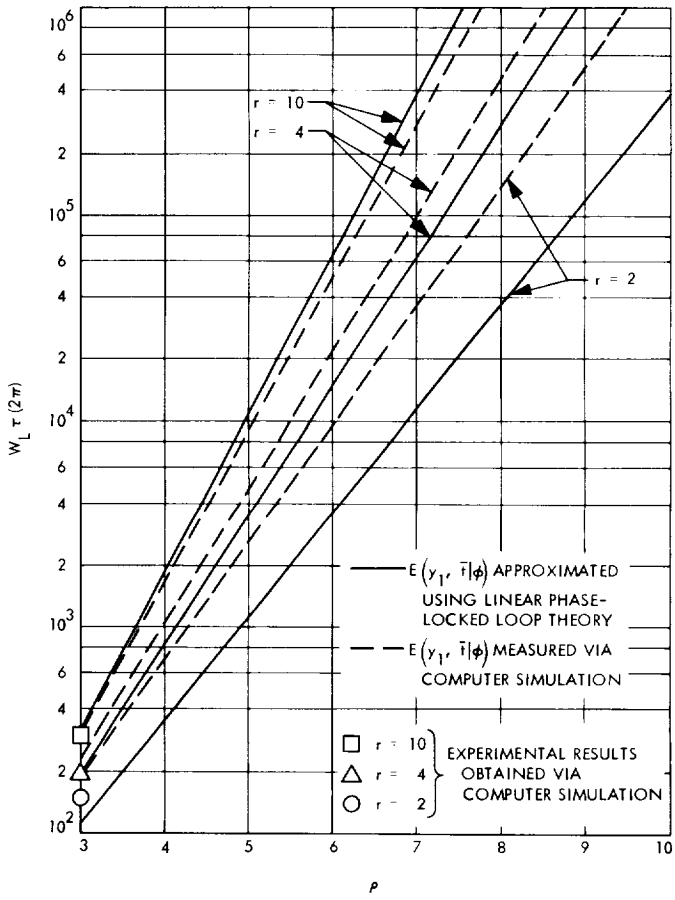


Fig. 29. Mean time to first slip-bandwidth product vs loop signal-to-noise ratio for various values of r [includes $E(y_1, \bar{t}|\phi)$ approximated using linear phase-locked loop theory]

In Ref. 1, it was shown that the linear phase-locked loop theory, $\bar{t} = \infty$, gives

$$E(y_1 \mid \phi) = (1 - F_1)\left[\frac{2rW_L}{(1+r)^2}\phi - \Omega_0\left(1 + \frac{F_1}{1+r}\right)\right] \tag{15}$$

where

$$r = AKF_1\tau_2, \qquad F_1 = \frac{\tau_2}{\tau_1},$$

$$W_L \equiv \frac{r+1}{2\tau_2}, \qquad \tau_1 >> \tau_2 \tag{16}$$

Substitution of Eq. (15) into Eq. (6) yields

$$\int^{\phi} h(u)\,du \simeq \left(\frac{r+1}{r}\right)\rho\cos\phi + \frac{\rho}{2r}\phi^2 + \frac{\rho\Omega_0}{AK}\phi = -U(\phi) \tag{17}$$

If we set $n = 1$ and let $\phi_f = 2\pi$, we find from Eq. (4) that

$$W_{L\tau}(2\pi) = \left(\frac{r+1}{r}\right)^2 \frac{\rho}{2}\int_{-2\pi}^{2\pi}\int_{-\phi_f}^{\phi} [C_0'(0) - u(x - \phi_0)]\exp[U(\phi) - U(x)]\,dx\,d\phi \tag{18}$$

where $U(\phi)$ and $U(x)$ are defined in Eq. (17) and, from Eq. (5),

$$C_0'(0) = \frac{\displaystyle\int_{\phi_0}^{2\pi}\exp[U(x)]\,dx}{\displaystyle\int_{-2\pi}^{2\pi}\exp[U(x)]\,dx} \tag{19}$$

If we let $\phi_0 = 0$ and $\Omega_0 = 0$, then $C_0'(0) = \frac{1}{2}$ and we have a result due to Tausworthe (Ref. 3). Finally, if we let $\phi_0 = 0$, $r = \infty$, we have Viterbi's (Ref. 2) result for a first-order loop.

In Figs. 28 and 29, we present $W_{L\tau}(2\pi)$ versus ρ for $\Omega_0 = \phi_0 = 0$ for various values of r and $\rho = 2A^2/N_0W_L$, the loop signal-to-noise ratio.

References

1. Lindsey, W. C., "Nonlinear Analysis and Synthesis of Generalized Tracking Systems," USCEE Part II. University of Southern California, Los Angeles, Calif., Apr. 1969.
2. Viterbi, A. J., *Principles of Coherent Communication*, McGraw-Hill Book Co., Inc., New York, N. Y., 1966.
3. Tausworthe, R. C., "Cycle Slipping in Phase-Locked Loops," *IEEE Trans. Commun. Technol.*, Vol. Com-15, No. 3, Jun. 1967.

VI. Communications Elements Research

TELECOMMUNICATIONS DIVISION

A. RF Techniques: System Studies for Frequencies Above S-Band for Space Communications,

T. Sato

The 60-in. diam 90-GHz radio telescope (SPS 37-54, Vol. III, p. 205) was converted into a radio sextant by using an optical sun tracker developed for this use. The tracking error was found to be about 1 arc min, keeping the radio boresight accurately on the sun disk which is about 30 arc min in diameter.

The radio sextant was operated for 15 days under varying weather conditions. Stable tracking was obtained during clear sky or light haze conditions. Cloud patches that obscured the sun's limb caused tracking-lock losses. Under these conditions, a manual override mode was used to track at near sidereal rate.

Data (see Table 1) from the observations have been reduced (Ref. 1) to yield zenith atmospheric loss, which ranged from a low of 0.239 dB to a high of 1.157 dB. The average zenith loss for the 15 days was 0.540 dB, which is about half of what is generally accepted for 90 GHz under "dry" conditions. These data were often taken under both "Santa Ana" conditions and after cold frontal passages where the air aloft tends to be dry. These conditions are rather common in the coastal zones of Southern California during the winter months.

Reference

1. Stelzried, C. T., and Rusch, W. V. T., "Improved Determination of Atmospheric Opacity from Radio Astronomy Measurements," *J. Geophys. Res.*, Vol. 72, No. 9, pp. 2445–2447, May 1, 1967.

Table 1. Atmospheric data from 90-GHz solar observations

Date, 1969	Zenith atmospheric loss, dB	Probable error, dB	Ground temperature, °F	Ground humidity, %
1/4	0.513	0.013	83	34
1/11	1.157	0.007	61	61
1/18	0.776	0.111	54	70
1/30	0.619	0.094	57	33
2/1	0.265	0.016	64	38
2/8	0.586	0.005	68	22
2/27	0.239	0.164	62	51
3/1	0.300	0.003	64	31
3/6	0.660	0.018	58	27
3/13	0.410	0.010	55	37
3/15	0.415	0.005	70	22
3/20	0.678	0.014	67	60
3/22	0.567	0.005	72	36
3/27	0.545	0.008	85	10
3/29	0.376	0.024	84	38

B. Spacecraft Antenna Research: S- and X-Band Telemetry and Tracking Feed, *K. Woo*

1. Introduction

An experimental high-gain antenna feed has been developed for the thermoelectric outer-planet spacecraft (TOPS) mission. The feed is capable of telemetering at both 8448 and 2295 MHz and monopulse tracking at 2115 MHz. This article presents the design and the preliminary test results of the feed.

2. Feed Design

The feed is designed to transmit and receive linearly polarized signals. The experimental model (Figs. 1 and 2) is composed of a conical horn (16.8-in.-diam aperture) fed by a series of rectangular waveguide sections. The four vertical probes in the 6.14- × 6.14-in. square guide are for monopulse tracking at 2115 MHz (utilizing the TE_{10}, TE_{11} + TM_{11}, TE_{20} modes). A detailed description of this monopulse operation can be found in SPS 37-56, Vol. III, pp. 91–92. The horizontal probe in the 2.83 × 2.83-in. square guide is for telemetering at 2295 MHz (utilizing the TE_{01} mode). The vertical probe in the X-band guide is for telemetering at 8448 MHz (utilizing the TE_{10} mode). In order to provide a high-efficiency feed for both telemetering frequencies, a step is introduced at the opening of the X-band guide for the purpose of broadening the beamwidth of the 8448-MHz pattern to that of the 2295-MHz pattern. The step generates the TE_{12} + TM_{12} and TE_{30} modes as shown in Fig. 3. These modes, when properly added to the dominant TE_{10} mode, broaden the radiation pattern of the TE_{10} mode.

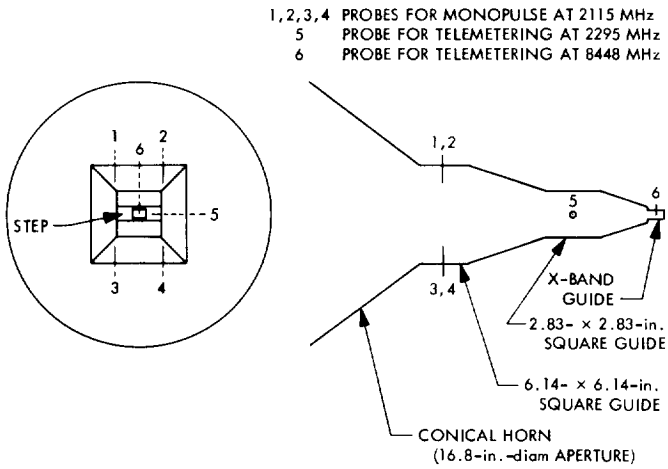


Fig. 1. Schematic of feed design

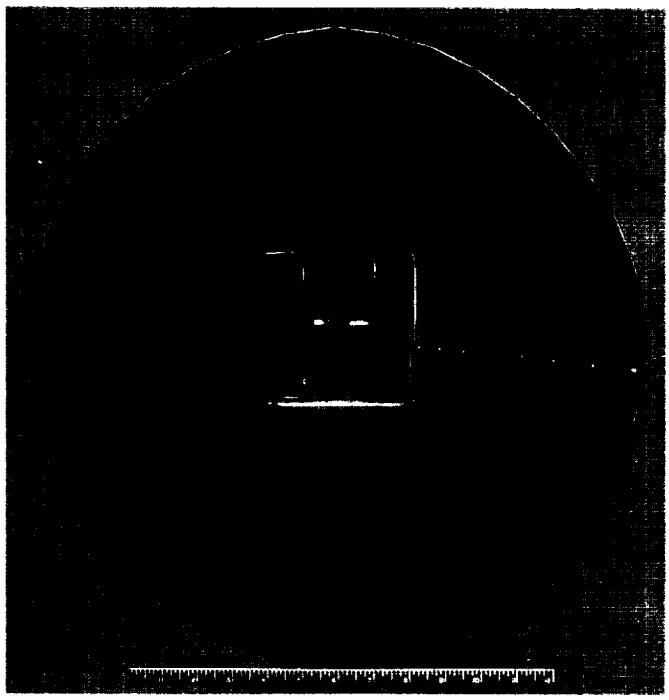


Fig. 2. Experimental model of feed (front view)

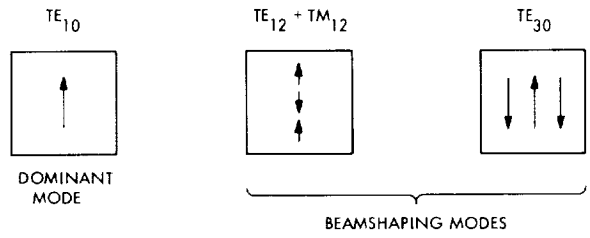


Fig. 3. Beam-forming modes at 8448 MHz

3. Results

Preliminary test results of the experimental feed have been obtained. Figures 4 and 5 show respectively the radiation patterns of the telemetry at both 8448 and 2295 MHz. As can be seen from these figures, the 8448-MHz pattern has been broadened to nearly the beamwidth of the 2295-MHz pattern, thus providing a high-efficiency feed for both frequencies. Figure 6 shows the radiation patterns of the sum (S), vertical difference (ΔV), and horizontal difference (ΔH) channels of the monopulse. Deep nulls have been obtained in the difference channels.

Further work on the feed is in progress. It includes

(1) Refining the 8448-MHz pattern to provide uniform illumination.

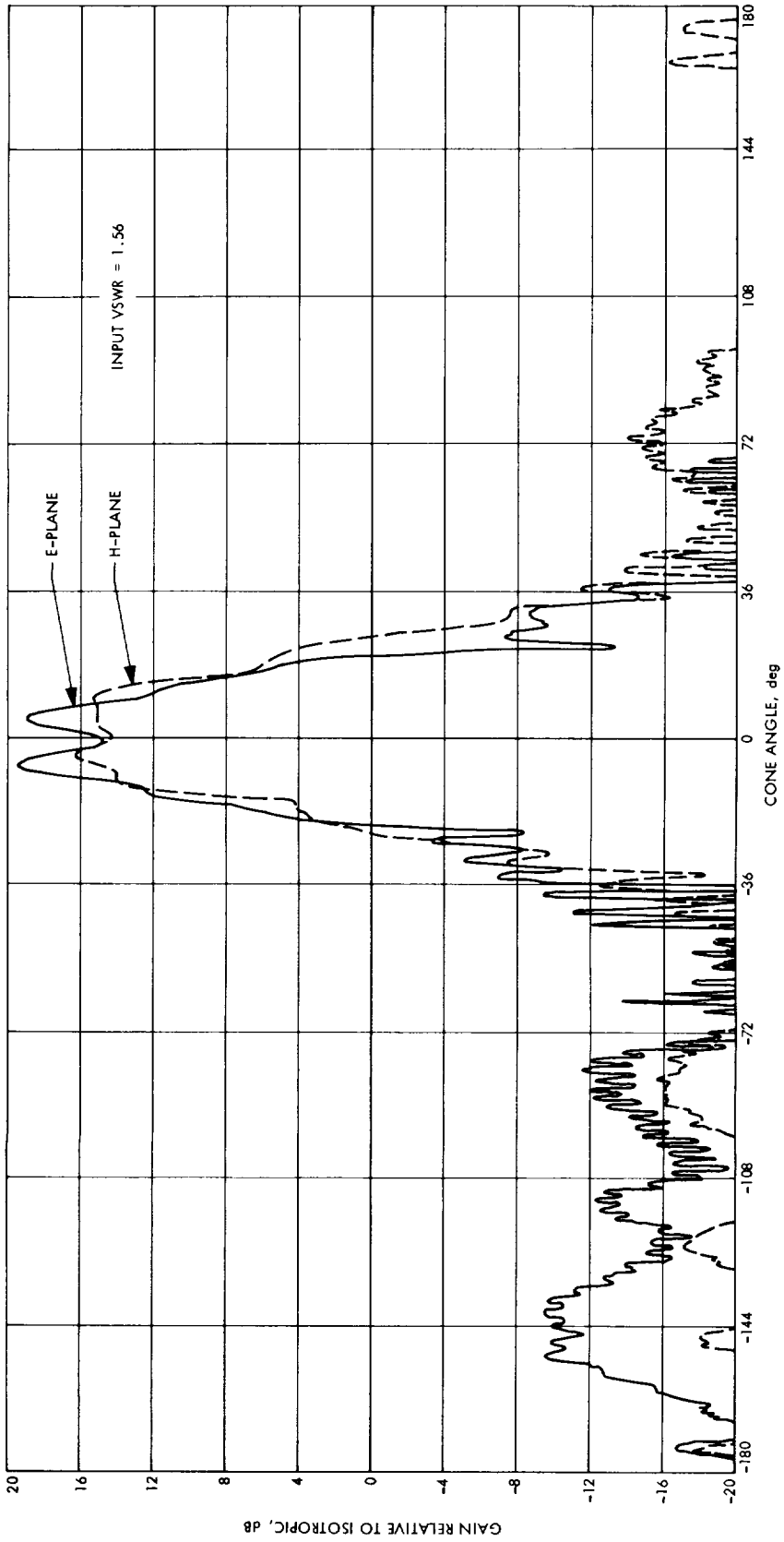


Fig. 4. Radiation patterns of telemetry at 8448 MHz

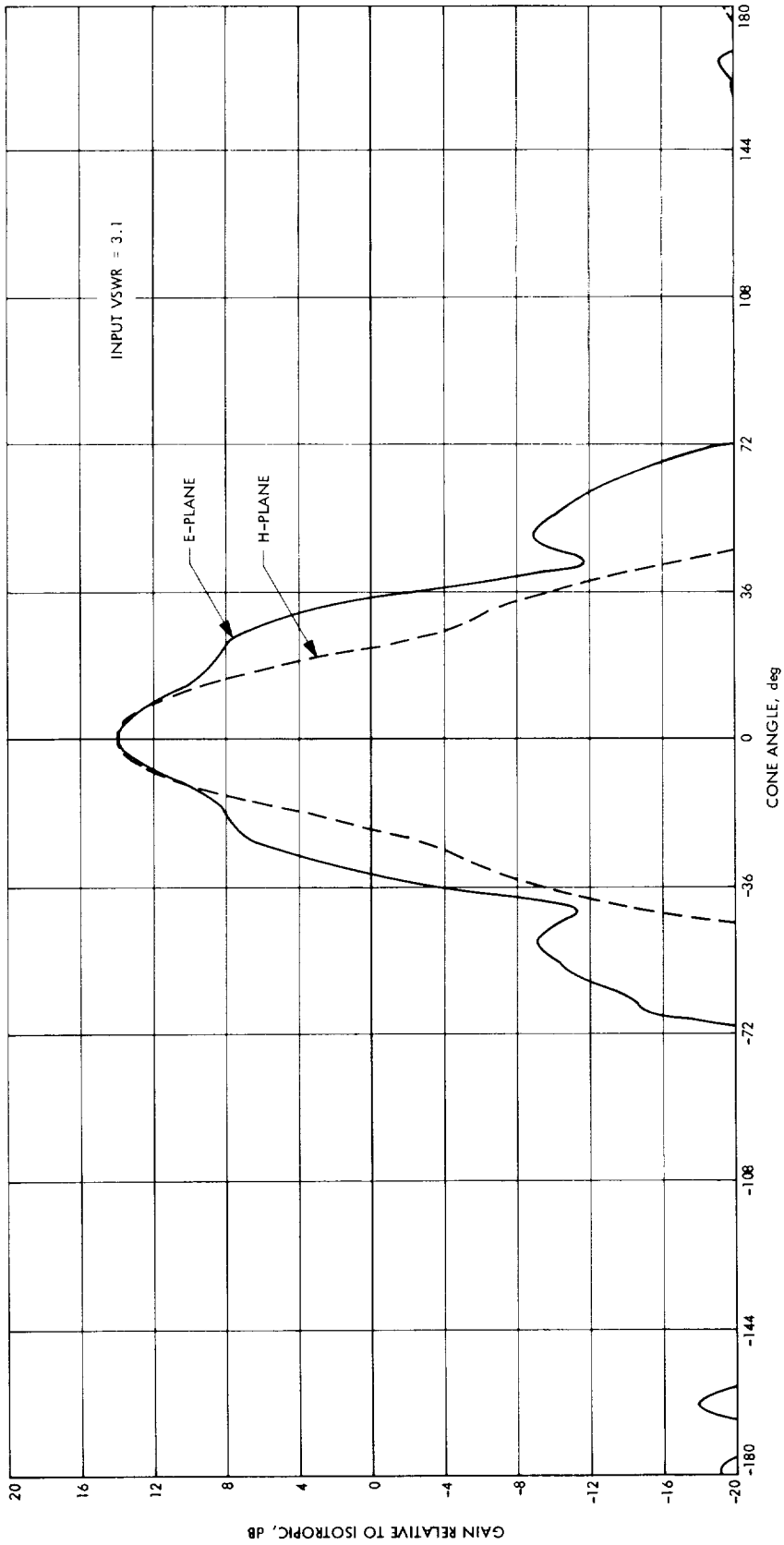


Fig. 5. Radiation patterns of telemetry at 2295 MHz

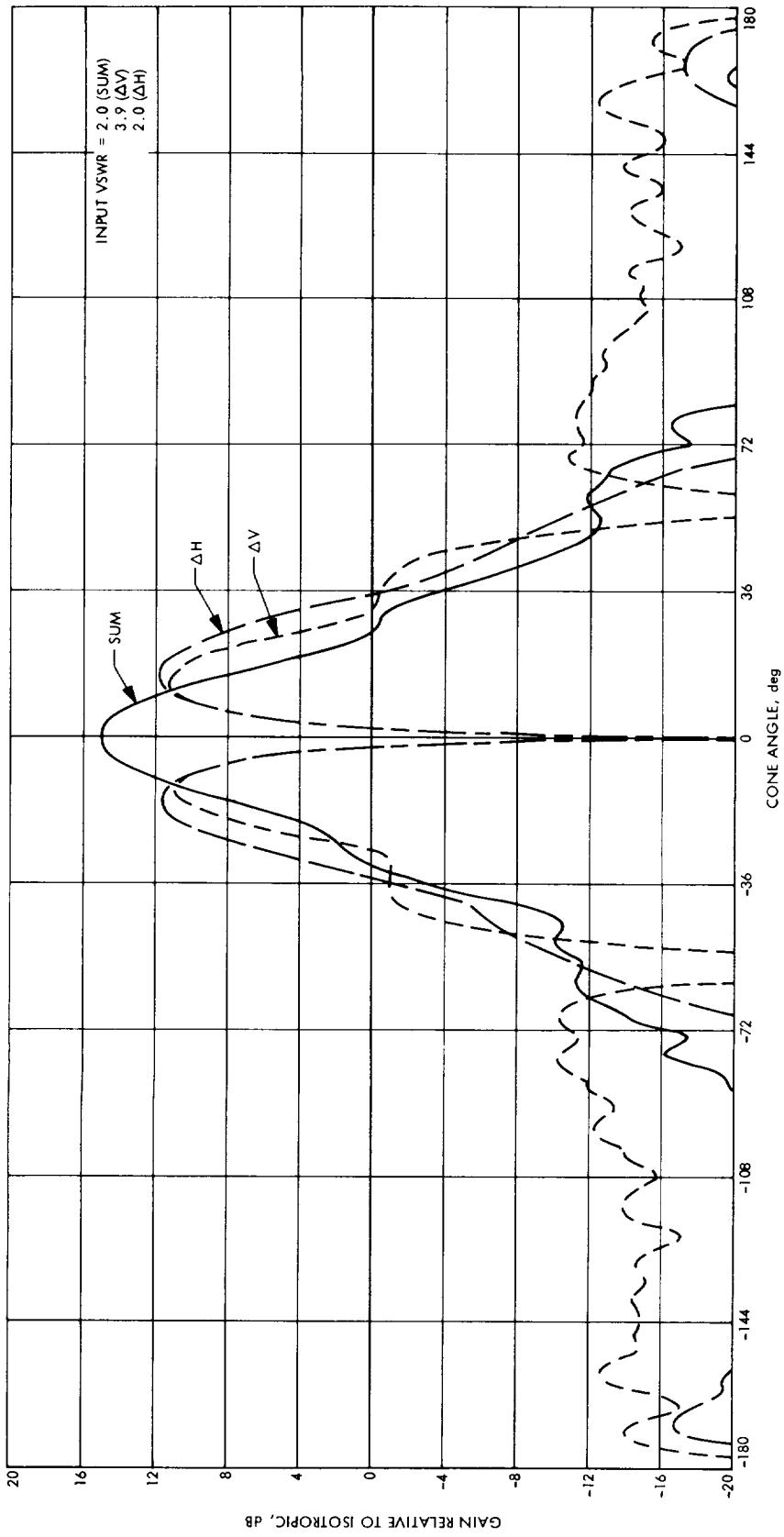


Fig. 6. Radiation patterns of monopulse at 2115 MHz

(2) Equalizing E- and H-plane patterns for both telemetering frequencies.

(3) Obtaining phase patterns for both telemetering frequencies.

(4) Improving input voltage standing-wave ratios (VSWRs).

C. Spacecraft Antenna Research: RF Voltage Breakdown in Uniform Fields, *R. Woo*

1. Introduction

Many studies have been made of RF voltage breakdown in uniform fields in air (Refs. 1–3). Some nomographs (Refs. 4 and 5) exist but these cover limited experimental ranges and a complex computation procedure must be used. A scheme for presenting RF breakdown data for coaxial transmission lines was introduced recently (Ref. 6). The advantage of this scheme is its simplicity and usefulness to the design engineer. Breakdown data for the uniform field geometry may be displayed in a similar fashion. The results are presented in this article.

2. Discussion

The results for the uniform field geometry are shown in Fig. 7 where breakdown voltage is plotted as a function of $p\lambda$ (p is pressure and λ is wavelength) for various values of fd (f is frequency and d is separation distance). Both multipacting and ionization breakdown are included. The solid curves represent experimental data while the dashed curves represent extrapolated data. For the fd range of 20–250 MHz-cm, data were obtained using the experimental setup described in Ref. 7. For other values of fd, existing data summarized in Table 2 were used. The extrapolated curves were drawn using the results of Gould and Roberts (Ref. 2) as a guide. It must be emphasized that the results of Gould and Roberts were computed and that they have not been verified experimentally. The range of experimental parameters in Fig. 7 is extensive. In comparison, for the coaxial configuration, fd varied from 20 to 2630 MHz-cm, and breakdown voltages did not exceed a few hundred volts. Atmospheric breakdown data are also included in Fig. 7. This information is particularly useful when determining, for instance, the power limitations of the RF system at Goldstone. Even where the breakdown voltage is higher than 10^4 V, the curves in Fig. 7 may still be used. For these cases, breakdown conditions are governed by electron attachment and breakdown voltage is simply proportional to pressure. Thus, the curves can be extended by straight lines possessing a slope of 1. It must be mentioned that the fd–$p\lambda$ plane discussed in Ref. 7 can be very helpful in identifying the various breakdown mechanisms.

Table 2. Experimental data

fd, MHz-cm	Frequency f, MHz	Separation distance d, cm	Experimenter
38095	9400	4.05	McDonald, Gaskell, and Gitterman (Ref. 3)
18900	9400	2.01	
11812	9400	1.26	
4706	992	4.74	
3071	9400	0.327	
1966	992	1.98	
994	3125	0.318	Herlin and Brown (Ref. 1)
491	3125	0.157	
250	66.75	3.75	Woo
100	66.75	1.5	
50	66.75	0.75	
20	66.75	0.3	
10	200	0.05	Pim (Ref. 8)

The minimum voltage required for breakdown for a given value of fd is of interest. Shown in Fig. 8 is the minimum breakdown voltage as a function of fd (for fd values greater than 100 MHz-cm) for ionization breakdown. Note that around $fd = 100$ MHz-cm the multipacting voltages are lower than the ionization breakdown voltages.

3. Conclusion

Some experimental data have been obtained for the uniform field geometry. These have been combined with existing experimental data to form a set of design curves. The features of these curves are their simplicity, usefulness, and extensive range of parameters, which cover pressures ranging from atmospheric down to vacuum and frequencies varying from a few megahertz to X-band.

References

1. Herlin, M. A., and Brown, S. C., "Breakdown of a Gas at Microwave Frequencies," *Phys. Rev.*, Vol. 74, No. 3, pp. 291–296, Aug. 1948.
2. Gould, L., and Roberts, L. W., "Breakdown in Air at Microwave Frequencies," *J. Appl. Phys.*, Vol. 27, No. 10, pp. 1162–1170, Oct. 1956.
3. MacDonald, A. D., Gaskell, D. N., and Gitterman, H. N., "Microwave Breakdown in Air, Oxygen, and Nitrogen," *Phys. Rev.*, Vol. 130, No. 5, pp. 1841–1850, June 1963.
4. Gould, L., *Handbook on Breakdown of Air in Waveguide Systems*, Report No. NE 111616. Microwave Associates, Boston, Mass., Apr. 1956.

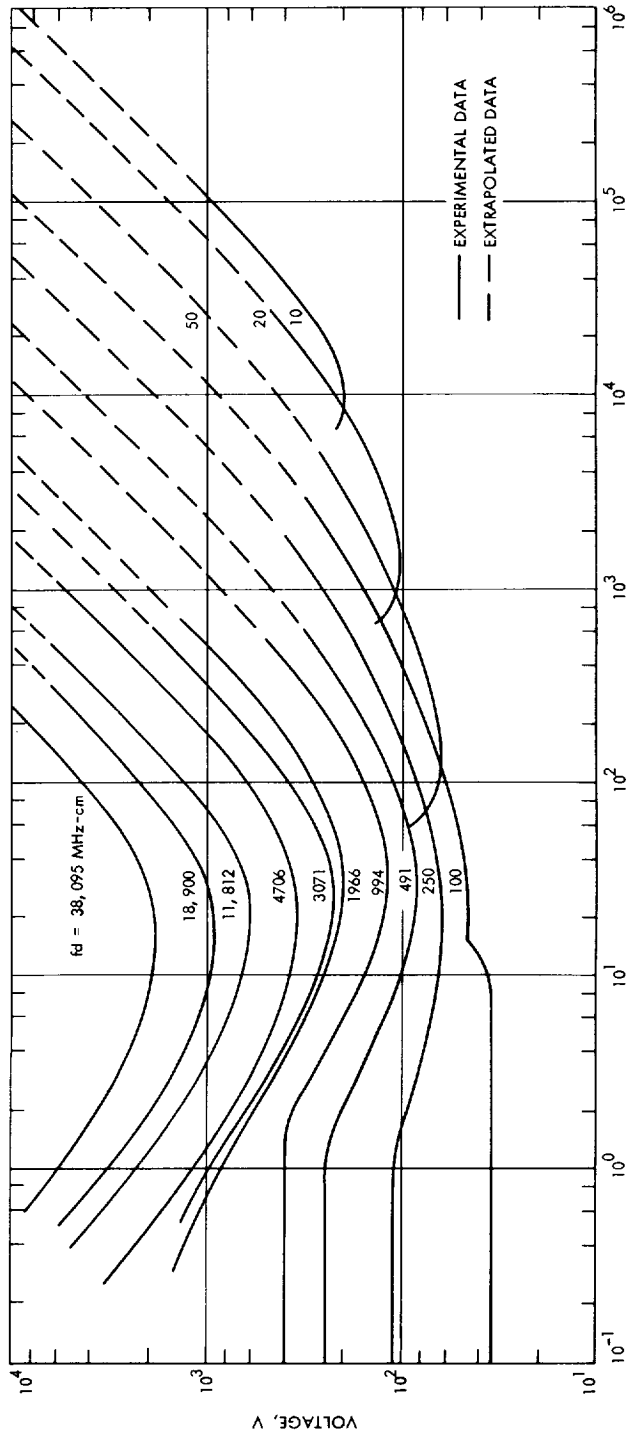


Fig. 7. Breakdown voltage as a function of $p\lambda$

5. Wheeler, H. A., *Nomogram for Some Limitations on High-Frequency Voltage Breakdown in Air*, Report No. 17. Wheeler Laboratories, Inc., Great Neck, N.Y., May 1953.
6. Woo, R., "RF Voltage Breakdown in Coaxial Transmission Lines," *Proc. IEEE*, Vol. 57, No. 2, pp. 254–256, Feb. 1969.
7. Woo, R., "Multipacting Discharges Between Coaxial Electrodes," *J. Appl. Phys.*, Vol. 39, No. 3, pp. 1528–1533, Feb. 1968.
8. Pim, J. A., "The Electrical Breakdown Strength of Air at Ultra-High Frequencies," *Proc. Inst. Elec. Eng. London*, Vol. 96, Part III, pp. 117–129, 1949.

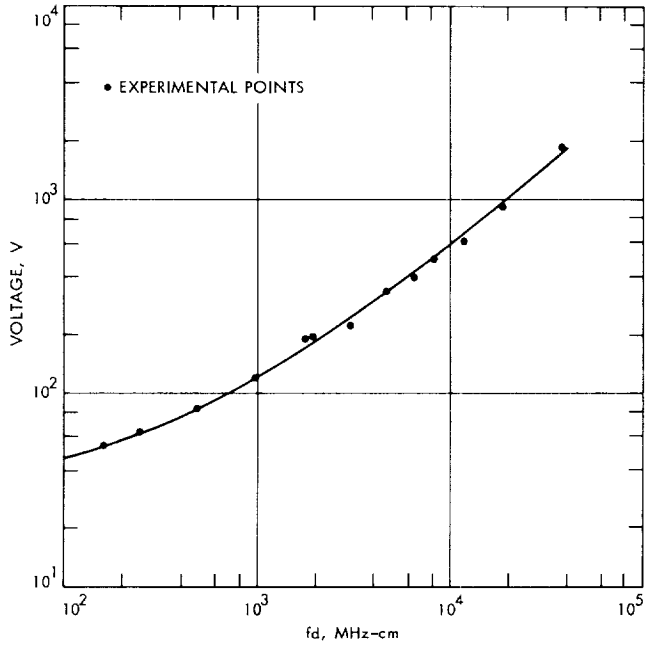


Fig. 8. Minimum ionization breakdown voltage as a function of *fd*

VII. Spacecraft Radio

TELECOMMUNICATIONS DIVISION

A. S-Band Diode Switch Evaluation, *A. W. Kermode*

1. Introduction

Development of S-band solid-state switches that are capable of providing reliable low-loss performance for use in future spacecraft or capsule lander communication systems is being studied. The first phase of a plan to evaluate the capability of a commercially produced diode switch under typical spacecraft or lander environments is reported here.

An S-band diode switch of modified design from the Hyletronics Corporation was procured. A single-pole six-position (SP6T) switch was selected as being representative of the switching complexity of a diode switch that might be required for future missions. The Capsule System Advanced Development requirements for an S-band switch were specifically considered at the time of procurement. The switch was procured to the following specifications:

Frequency	2295 $\pm$15 MHz
Insertion loss	0.6 dB max (goal)
Isolation	25 dB min (goal)
RF power capacity	10 W CW
RF connectors	Omni-Spectra, Inc. OSM type
Size	5 $\times$ 5 $\times$ 0.75 in.
Bias power	To be minimized (i.e., only diodes associated with the RF port in the low loss-transmission state shall be forward-biased. The remaining diodes shall be reverse biased).

The switch design utilized PIN diodes mounted in a 50-Ω stripline configuration, consisting of an etched copper strip center conductor in a symmetrical sandwich structure between two 1/16-in. teflon fiberglass microwave circuit boards. The circuit boards were in turn sandwiched between 1/16-in. aluminum support plates.

The manufacturer determined that the isolation requirement could be met with a single PIN diode mounted in each arm of the switch. The use of a single diode would also help to reduce the insertion loss of each RF port in the low-loss state. A stud-mounted diode was selected for its thermal dissipation qualities at the required 10-W RF level. The diode was connected in a shunt configuration with the RF line, and mounted to the aluminum plate by the threaded stud.

In order to meet the minimum dc bias requirements, the diodes were displaced from the main transmission line of each RF port by an electrical quarter wavelength at the center frequency. Each diode was capacitively coupled to the open end of the quarter-wave stubs. The diode-stub was located an electrical quarter wavelength from the common junction of all RF ports. The bias voltage was coupled to the PIN diodes through low-pass filter networks. A simplified diagram of the switch is shown in Fig. 1.

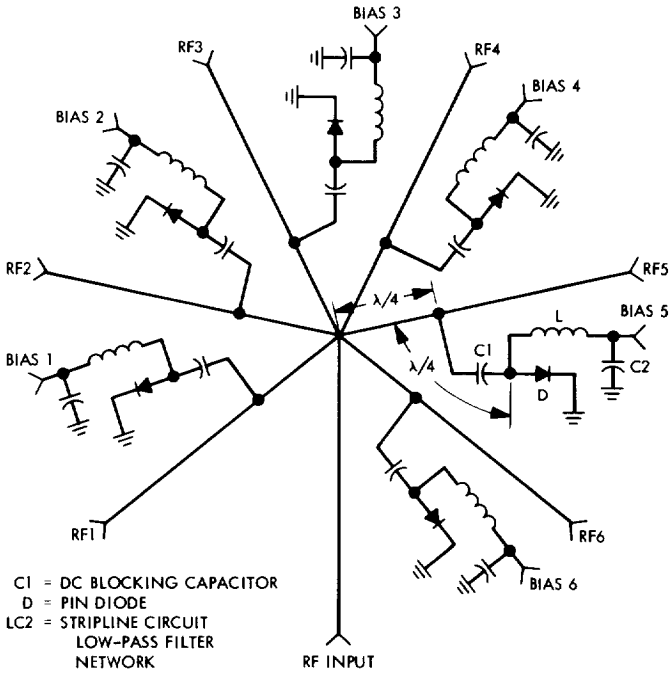


Fig. 1. PIN diode switch diagram

Application of a reverse bias voltage (−30 V) to the PIN diode results in an RF open-circuited stub, which is transformed to a low impedance at the junction with the main line, and is, in turn, transformed to a high impedance at the common junction of all RF ports. The effective high impedance coupling to the common junction provides the isolated state from the RF input port to the various RF output ports.

Application of forward bias (100 mA) to the PIN diode results in an RF short-circuited stub, which is transformed to a high impedance at the coupling point on the main transmission line. The high-impedance shorted stub coupled to the main transmission line establishes a low-loss transmission path between the RF input port and the desired RF output port.

2. Tests and Results

The SP6T diode switch was subjected to the following evaluation sequence:

(1) Bench performance tests at mW and 10-W-CW power levels.

(2) Temperature tests (−10, +25, and +75°C), at a 10-W CW power level.

(3) *Mariner C* type-approval vibration and shock environments.

(4) High-impact shock (10,000 *g*) environments.

Performance tests consisted primarily of insertion loss and isolation measurements. VSWR and phase jitter measurements were included for selected test sequences. The results of the bench performance tests have been summarized in Table 1.

The test data shows that the insertion loss and isolation goals were not attained over the frequency band of interest for all RF ports. This was due to the fact that optimum insertion loss and peak isolation performance were not coincident for all RF ports. Also, the isolation peak (38-dB minimum) of each RF port was not centered at 2295 MHz. The center frequency for peak isolation of each RF port has been tabulated in Table 2. The isolation of each RF port was found to vary up to 1.5 dB maximum across the frequency band of interest, depending upon which RF port was biased to the low-loss state.

Measurements were conducted to determine the effects of bias supply variations on RF performance. It was

Table 1. SP6T diode switch preliminary bench performance data

RF port	Frequency, MHz					
	2280		2295		2310	
	≤ 1 mW	10 W	≤ 1 mW	10 W	≤ 1 mW	10 W
Insertion loss, dB						
1	0.80	0.50	0.85	0.55	0.80	0.60
2	0.90	0.70	0.90	0.65	0.90	0.70
3	1.00	0.85	0.85	0.80	0.80	0.75
4	0.80	0.55	0.70	0.50	0.70	0.50
5	0.80	0.65	0.75	0.65	0.80	0.70
6	0.90	0.60	0.95	0.70	1.10	0.80
Isolation, dB						
1	30.0	28.0	40.0	37.0	34.0	34.0
2	32.0	31.0	40.0	38.0	33.0	30.0
3	33.0	32.0	27.0	26.0	22.0	22.0
4	29.0	28.0	25.0	24.0	22.0	21.0
5	40.0	38.0	35.0	34.0	27.0	27.0
6	25.0	25.0	30.0	29.0	39.0	39.0
RF input port VSWR (RF port in low-loss state with output port terminated)						
1	1.47	1.64	1.52	1.71	1.73	1.83
2	1.60	1.58	1.47	1.55	1.57	1.59
3	1.60	1.66	1.42	1.62	1.47	1.63
4	1.65	1.58	1.46	1.49	1.50	1.50
5	1.52	1.60	1.40	1.57	1.56	1.60
6	1.55	1.29	1.58	1.33	1.82	1.46

Table 2. SP6T diode switch peak isolation versus RF ports

RF port	Frequency for peak isolation (38-dB min), MHz
1	2301
2	2293
3	2268
4	2258
5	2285
6	2315

found that a variation of ±15 V about the nominal reverse bias level of −30 V resulted in a maximum change in isolation of up to 6% at a nominal isolation of 35 dB. The isolation change was a result of the diode capacitance variation versus reverse bias voltage.

The forward bias current was reduced from 100 to 30 mA with a maximum insertion loss increase of 0.05 dB. Increasing the forward bias current to 150 mA resulted in an insignificant insertion loss change.

The switch was subjected to temperature environments (−10 and +75°C) at atmospheric pressure level. Insertion loss and isolation parameters were measured during the temperature tests. The results of the temperature tests have been summarized in Table 3.

The switch was then subjected to *Mariner C* type-approval vibration and shock environments. During the vibration and shock tests, the following parameters were monitored: insertion loss, isolation and phase jitter. In order to avoid overtesting the switch during environmental tests and to reduce the overall environmental test time, the insertion loss was monitored on one of the six RF ports and isolation was monitored on two of the RF ports. Phase jitter was monitored on the three RF ports used for insertion and isolation measurements. During vibration tests, VSWR was monitored on the input RF port. No significant change in insertion loss, isolation, or VSWR resulted during vibration and shock tests. The maximum phase jitter observed was less than 3 deg peak-to-peak (including RF cables).

Table 3. SP6T diode switch temperature performance data (10-W power level)

RF port	Frequency, MHz					
	−10°C			+75°C		
	2280	2295	2310	2280	2295	2310
Insertion loss, dB						
1	0.75	0.95	0.85	0.80	0.80	0.90
2	0.75	0.85	0.80	0.85	0.80	0.90
3	0.60	0.60	0.55	0.70	0.70	0.75
4	0.55	0.70	0.65	0.75	0.80	0.75
5	0.60	0.85	0.80	0.70	0.85	1.00
6	0.80	0.95	1.00	0.70	0.80	1.00
Isolation, dB						
1	31.0	39.0	30.0	27.0	30.0	33.0
2	34.0	37.0	28.0	31.0	34.0	30.0
3	29.0	25.0	21.0	31.0	27.0	23.0
4	26.0	23.0	20.0	28.0	25.0	23.0
5	37.0	32.0	26.0	36.0	37.0	31.0
6	25.0	28.0	36.0	24.0	29.0	32.0

A series of high-impact tests were conducted at the JPL high-impact facility. The tests were conducted at discrete shock levels in both directions along three mutually perpendicular axes. Insertion loss and isolation parameters were measured after each impact test. Impacts were conducted at the following g levels: 1,000, 2,500, 5,000, and 7,500.

During the impact tests, mating right-angle OSM type connectors were placed on five of the seven RF ports. The mating connectors were tightened to a torque in the range of 7 to 9 in.-lb. During the impact test sequence, several of the OSM switch connectors were found to be loose. These connectors were held to the composite sandwich structure by seven 0-80 screws. The screws were checked after each impact, and were tightened as required.

Difficulty was encountered at the 7,500-g impact level, in that two impact tools were fractured. It was necessary to design a special tool for the 7,500-g level. The impact test sequence and evaluation were terminated because of inadequate funding support.

The performance results, after the last 7,500-g impact, (uncalibrated level due to tool fracturing) have been summarized in Table 4. A comparison of the final data in Table 4 to that of the initial data in Table 1 indicates that the overall performance of the switch was degraded very little throughout the stringent test sequence.

Overall performance can be considered better than expected for using standard commercial packaging technology. The tests that are significant, which were not included in the overall evaluation, are: thermal vacuum, repeated temperature cycling, and heat sterilization. A thermal vacuum test was not included in the sequence, due to possible outgassing problems with the use of commercial grade epoxies and paints. The temperature cycling tests could have been accomplished during the heat-sterilization test sequence. However, this portion of the evaluation program was not accomplished, due to inadequate funding support.

3. Conclusions

An evaluation program, which included selected *Mariner C*, type-approval environments, as well as high-impact shock (up to 7,500 g) was successfully conducted on a SP6T S-band PIN diode switch. The results of the test program were encouraging in that a commercial switch was procured and subjected to a stringent evaluation program with no catastrophic failures. Only one sample switch was tested; however, it contained six PIN diodes of a commercial type. A significant amount of knowledge and insight on multiple-port solid-state RF switches was gained through the program. The evaluation has resulted in a multiple-port RF switch design that can be flight-qualified for use in future space flight applications. Two typical applications are: (1) the connection of multiple transmitters to a common antenna, and (2) the connection of various antennas to a single receiver, such as a sequential lobing tracking receiver.

Table 4. SP6T diode switch post environmental test data (milliwatt power level)

RF port	Frequency, MHz		
	2280	2295	2310
Insertion loss, dB			
1	0.75	0.80	0.95
2	0.75	0.75	0.80
3	0.80	0.80	0.85
4	0.80	0.75	0.85
5	0.70	0.70	0.80
6	0.75	0.85	1.00
Isolation, dB			
1	29.0	40.0	33.0
2	32.0	34.0	28.0
3	31.0	25.0	22.0
4	28.0	24.0	21.0
5	40.0	33.0	26.0
6	24.0	30.0	40.0
VSWR			
1	1.48	1.48	1.54
2	1.46	1.42	1.40
3	1.54	1.50	1.45
4	1.60	1.53	1.50
5	1.45	1.40	1.40
6	1.50	1.52	1.57

It has been projected that the basic switch design is capable of operation up to power levels of 20 W CW. Insertion loss performance improvement could be achieved through the following:

(1) Selection and use of higher quality PIN diodes.

(2) Use of a low-loss dielectric material in the circuit boards.

(3) Improved RF matching (lower VSWR) at each RF port.

(4) All ports tuned for optimum performance at the band center.

It is felt that an insertion loss of 0.4 to 0.6 dB maximum and an isolation of 25 dB minimum should be realizable over the ± 15 MHz bandwidth.

VIII. Spacecraft Telecommunications Systems

TELECOMMUNICATIONS DIVISION

A. Pattern Recognition: Invariant Stochastic Feature Extraction and Statistical Classification, *J. P. Hong*

1. Introduction

The extraction of data from two-dimensional images is under study. Most of the work in this area is in reference to characters and the problems associated with recognizing them (Refs. 1–5). However, the fundamentals of feature or information extraction are not limited to this area. A study of the possible use of such techniques in space technology may lead to new data compression and optical guidance systems.

"Pattern recognition" has been used in the literature to mean both the classification of data according to "patterns" and the discovery of "patterns" in data. In this article, "recognition" will be used to mean classification. In the few examples that are included, it will be clear what is meant by "pattern."

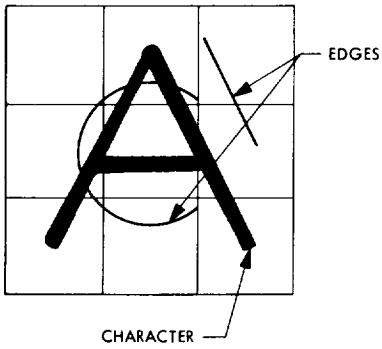
In general, pattern recognition is a two-step process. First, one must take measurements or otherwise obtain the data. Then, he must decide which classification the data belongs in. One has to look at both the measurement process and the decision algorithm together to design an optimum system.

2. Character Pattern Recognition

The taking of the measurements on the data (often called feature extraction) and the classification algorithm are closely related to the problem that one is considering. It is assumed that sampling the data, processing, filtering, etc., are all possible parts of the measurement process. The result of measurement is a set of numbers $\{X_1 \cdots X_n\}$, each X_i giving some pertinent information about the data measured. It is by no means clear what are the "best" measurements. For instance, if one is interested in sorting eggs for marketing according to size, putting them through a grading mesh is sufficient. Suppose, however, one is interested in converting the contents of a medieval handwritten document into digital data. The magnitude of this problem is clear. What are the "best features" to measure? What are the best classification algorithms? Some feature measurement techniques and the decision algorithm associated with character classification are described below.

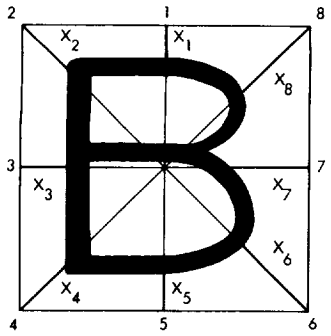
a. Handwritten character classification. Demonstrations have shown that even humans perform rather

poorly in recognizing handwritten characters out of context. The general problem is very difficult. Work has been done by Brain and Hart (Ref. 6) on special types of handwritten characters—printing in confined squares



as on FORTRAN coding sheets. Their feature extraction is in two steps. First, each character is quantized into a 24×24 matrix. The matrix is compared with 84 edges in 9 translated positions. The results of this edge detection is fed into a trainable linear classifier. The training set consisted of 8000 characters from 10 writers. The reported error rate on material obtained from writers not included in the training set is about 20%. This is a rather poor result considering the number of computations—$84 \times 9 = 756$ edge detections!

Fu (Ref. 7) reports a handwritten character classification system that makes 7% error. The number of computations that it needs is far less than required above. The



feature which is extracted is the length that the predetermined lines $1, 2, \cdots, 8$ make with the figure. Therefore, the measurement is $\{X_1 \cdots X_8\}$. It is assumed that $X_1 \cdots X_8$ are gaussian, with a mean and variance depending on the figure. The sequential probability test (Ref. 8) is employed. On a set of two characters (A and B), decisions could be made after six measurements. On a set of four characters (a, b, c, d), a similar experiment showed that the average number of measurements needed to make a 7% significance decision was less than 10 (Ref. 7).

It shall be noted that the test was not run on the complete alphabet. Also, this system is sensitive to alignment and the algorithm falls apart if the centering of the figure exceeds the tolerance. The gaussian assumption is suspect and Fu's sequential results depend on this important assumption. His algorithms are optimum only under the gaussian assumption.

b. Machine-produced block printing. A method widely used for feature extraction of machine-produced block printing is to scan the page in a zig-zag pattern or by a group of parallel lines.

The continuous data stream is matched-filtered (usually digitally) with patterns already stored in a machine. The most sophisticated machine on the market is the IBM 1975 Optical Page Reader (Ref. 9), which operates in the one-error-per-million region. This is due to the fact that the algorithm checks for context when it is "uncertain" about a character's classification, and it has a stored pattern for each font.

The IBM 1975 Optical Page Reader is operational and is used by the Social Security Administration to digitize quarterly employer's reports. It checks context by looking through a dictionary of names. The IBM 1975 is a specialized machine which is prohibitively expensive for most applications.

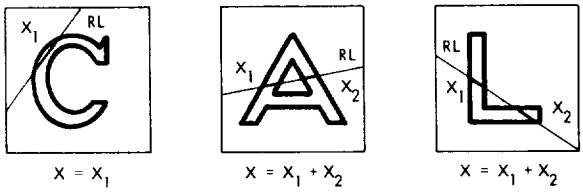
The scan-by-predetermined-lines techniques are subject to alignment constraints. Also, character registration plays an important part in the machine's performance. Skewed or smudged characters play a critical role. It would be desirable if one could find a *feature extraction that is invariant to displacement, rotation, size, and smudges.*

3. Feature Extraction Using Random Line Intersection

A method of classifying two-dimensional shapes that simplifies the measurement and decision processes will now be presented. The early work was done by Ball (Ref. 10) and Wong (Ref. 11). However, the results reported here are those of the author.

We would like to have a feature extraction process that is invariant under any "rigid motion" of the figure in a retina (a confined area of observation). Such movements may be translation or rotation in two dimensions and a third-dimension "motion" manifested by the change of size; i.e., the features that are extracted from the three views below are *invariant*. Such a feature extraction must be independent of coordinate axis.

This consideration rules out Fu's feature extraction by predetermined lines. One could generalize Fu's method and consider *random* lines (*RLs*) and their intersection with a two-dimensional figures. A straight line is chosen for simplicity. The sum of the intersection segments for each random line is added to give a *length* of intersection. The *statistics of this length* is the invariant feature.

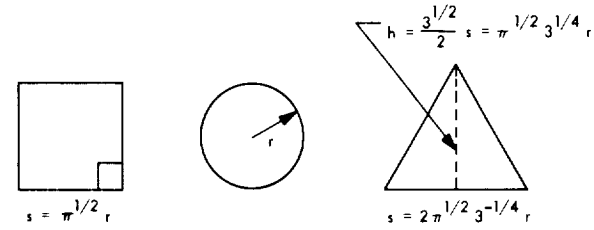


It is of great importance that the random lines have a uniform distribution over the retina. The invariance under two-dimensional rigid motion depends on this. A method of obtaining invariance under the "third-dimensional motion" will be discussed later.

Wong's experiments reported in the summer of 1968 (Ref. 11) consisted of classifying two basic geometric figures: a circle and a square of the same area. It was reported that 4.5 was the mean number of feature extractions required for a classification at a 2% significance level. Wald's sequential probability ratio test was used.

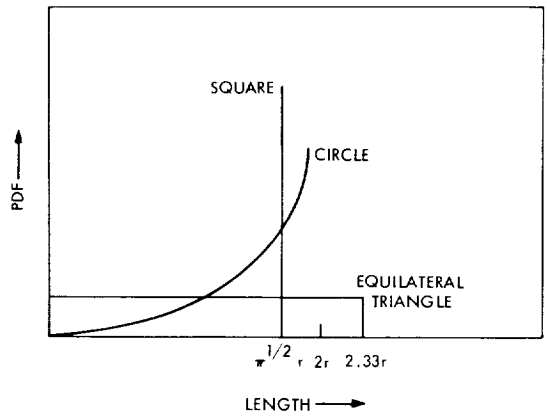
Based on a calculation of the probability density function (PDF) for random line intersection lengths in one dimension and empirical PDF, for the random line intersection lengths in two dimensions, one concludes that the above result is reasonable.

The PDF for three basic geometric figures with equal areas and relative dimensions listed above have been calculated, with the following assumption.



The random line cast is restricted to be vertical and must intersect the figure.

The PDF for a square, an equilateral triangle, and a circle are given below. The three PDFs are so distinctly different that almost any statistical algorithm will give a good separation.



The problem of calculating the PDF for more complex figures, such as the letters of the alphabet, becomes unmanageably difficult. There are good reasons why one need not calculate the PDF for such figures. Type fonts

vary greatly from style to style, moreover direct computation may be near impossible whereas empirical data may be obtained easily.

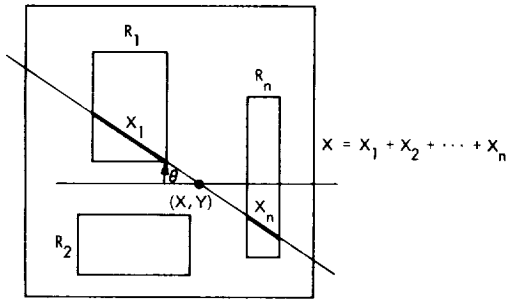
An algorithm for casting random lines and measuring their intersection with figures was coded, checked, and run on an SDS 930 computer. The purposes that the program served are:

(1) Feasibility of obtaining empirical PDF for complex figures was demonstrated.

(2) Invariance to two-dimensional motion was shown.

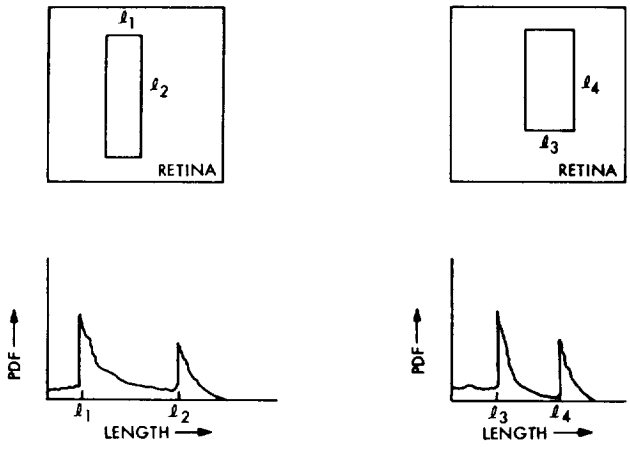
(3) The PDF was distinct for the few figures for which the program was run.

The uniform random line was chosen with the aid of a random number of generator (maximum length shift register). The intersection of this line with every one of

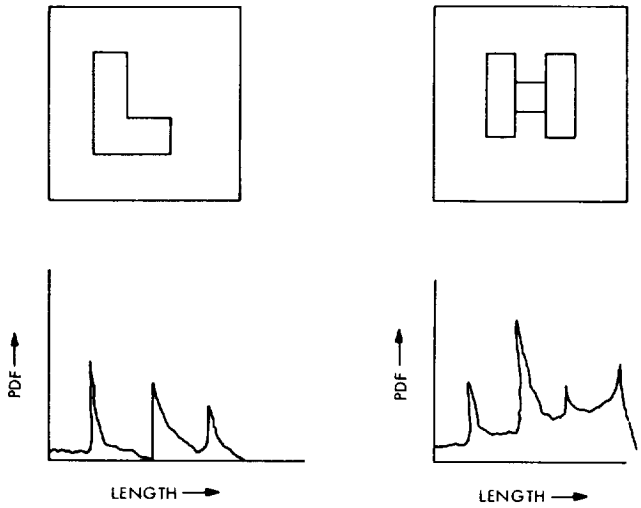


the preassigned rectangles, $R_1, \cdots, R_N$ was determined using the equations of a straight line. The figures of interest were approximated by rectangles.

The PDF gathered for two rectangles from runs of 5000 samples are depicted below.



It was gratifying to see that the peaks of the PDF followed the dimensions of the rectangle. The PDF also satisfies other intuitive properties, as some thought would verify.



To experiment with more complex figures, (i.e., figures with curves) which cannot readily be approximated by rectangles, one would go to a flying spot scanner driven by a random line generator.

From these experiments one can conclude that classification of the above letters with unfailing accuracy is possible with 2000 samples. The object of further study would be to investigate what algorithms could be used for a trade-off of accuracy for speed. This approach of feature extraction opens up the whole field of mathematical statistics, including nonparametric and sequential techniques, for application to the pattern recognition problem.

The *third-dimension invariance* can be obtained by processing the line intersection lengths with an automatic gain control. The line lengths are put through a low-pass filter, thus obtaining its mean, which is used to maintain a constant long-term mean of the intersection lengths.

Invariant feature extraction has some problems associated with it. The PDF for certain characters is the same (i.e., b, p, d, q are rigid transformations of each other in many fonts). For such letters, decision based purely on the PDF will not be conclusive. An optimum system would have to have a secondary algorithm to accommodate such problems.

The maximum likelihood ratio test, Wald's sequential probability ratio test, parametric and nonparametric hypothesis tests and other nonparametric techniques are possible classification algorithms. It is not clear what are the optimum approaches since the PDFs are not analytic. There are many approaches that can be investigated in finding an optimum algorithm.

4. Further Investigation and Experiments

At the moment the author is trying to code a maximum likelihood ratio test. Other investigations and experiments that one could pursue are:

(1) Implement a random line flying spot scanner.

(2) Investigate the field of mathematical statistics to find pertinent useful theory.

(3) Try coding some obvious algorithms, likelihood ratio tests, sequential tests, etc.

(4) Investigate the possibility of using other random lines for invariant feature extraction, such as circles or elipses.

(5) Investigate applications to optical guidance systems.

References

1. Nagy, G., "State of the Art in Pattern Recognition," *Proc. IEEE*, Vol. 56, No. 5, May 1968.
2. Kamentsky, L. A., and Liu, C. N., "Computer Automated Design of Multifont Print Recognition Logic," *IBM. J. Res. and Dev.*, Vol. 1, pp. 2–13, 1963.
3. Novikoff, A. B. J., "On Convergence Proof for Perceptrons," *Proc. Symposium on Math., Theory of Automata,* Brooklyn Polytechnic Institute, 1962.
4. Rosenblatt, F., *Principles of Neurodynamics, Perceptrons, and the Theory of Brain Mechanism.* Sparton Books, Baltimore, Md., 1961.
5. Tsypkin, Y. Z., "Use of the Stochastic Approximation Method in Estimating Unknown Distributions," *Automat. Rem. Cont.*, Vol. 27, pp. 432–434, 1966.
6. Brain, A. E., and Hart, P. E., *Graphical-Data-Processing Research Study and Experimental Investigation,* Technical Report ECOM-01901-25. Stanford Research Institute, Menlo Park, Calif., Dec. 1966.
7. Fu, K. S., *Sequential Methods in Pattern Recognition and Machine Learning.* Academic Press, New York, 1968.
8. Wald, A., *Sequential Analysis.* John Wiley & Sons, Inc., New York, 1947.
9. Hennis, R. B., et al., "IBM 1975 Optical Page Reader," *IBM J. Res. Develop.*, Vol. 12, No. 5, Sept. 1968.
10. Ball, G., Ph.D. thesis. Stanford University, Stanford, Calif., 1962.
11. Wong, E., et al., "Pattern Recognition," *Intensive Short Course in Statistical Communication Theory,* The University of Michigan, Engineering Summer Conferences, 1968.

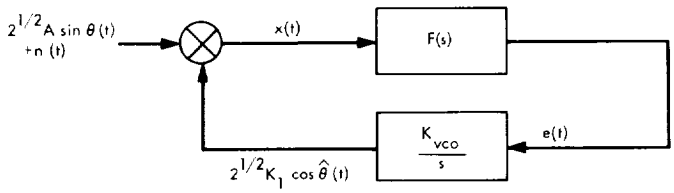
B. On the Equivalence in Performance of Several Phase-Locked Loop Configurations,

M. K. Simon

1. Introduction

Quite often in the practical design of a phase-locked loop to be used as a tracking loop in a phase-coherent receiver, one is required to use a configuration different from the standard model (Fig. 1) whose analysis is well documented (Ref. 1). The question immediately arises as to the equivalence in performance among these various topologies. Consider first the tracking loop block diagram (Ref. 2) illustrated in Fig. 2. For the sake of simplicity, we will ignore the effects of limiter suppression in our comparison. It should be intuitively obvious that if the external reference generator at frequency $\omega_c - \omega_1$ is noise-free, then the additional demodulation done at this frequency should in no way affect the performance of the loop in the presence of noise. Furthermore, the random phase of this generator, $\theta_3(t)$, will be tracked by the loop. Perhaps not so obvious is the configuration of Fig. 3 wherein the additional demodulation reference is derived from a noisy source [i.e., the voltage-controlled oscillator (VCO) output].[1] However, since the two demodulation references (i.e., at frequencies ω_1 and $\omega_c - \omega_1$) are coherent, we will show that the dynamic phase noise performance of this loop is identical to that of Fig. 1 if

[1]This configuration is typical of spacecraft transponder receivers.



$$n(t) = 2^{1/2}\left[n_1(t) \sin \omega_c t + n_2(t) \cos \omega_c t\right]$$
$$\theta(t) = \omega_c t + \theta_1(t)$$
$$\hat{\theta}(t) = \omega_c t + \theta_2(t)$$
$$\phi(t) = \theta_1(t) - \theta_2(t)$$
$$\frac{d\phi(t)}{dt} = \frac{d\theta_1(t)}{dt} - K_1 K_{vco} F(p)\left[A \sin \phi(t) + n'(t)\right]$$
$$n'(t) = -n_1(t) \sin \theta_2(t) + n_2(t) \cos \theta_2(t)$$

Fig. 1. Standard phase-locked loop configuration

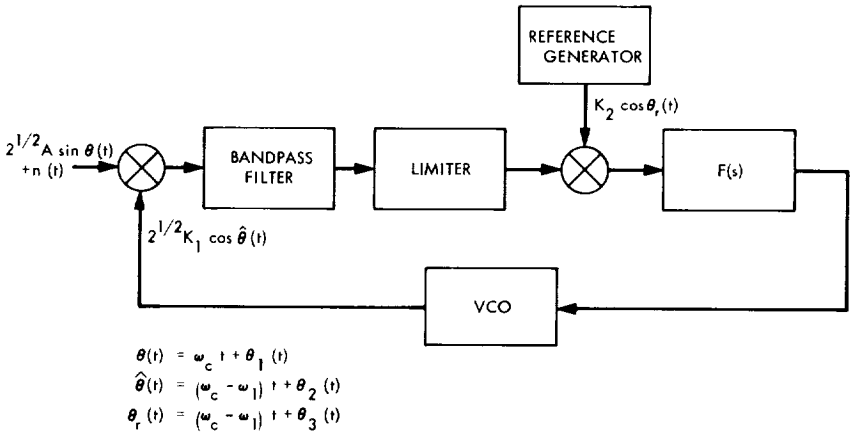


Fig. 2. Phase-locked loop model—variation 1

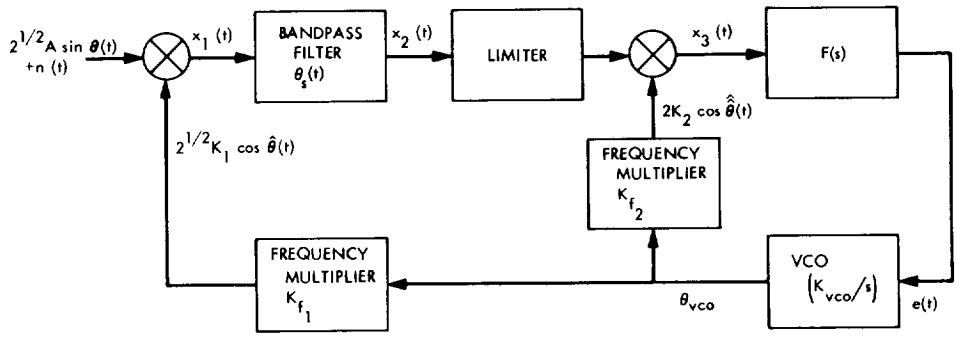


Fig. 3. Phase-locked loop model—variation 2

the loop parameters are defined appropriately. The obvious advantage of the configuration of Fig. 3 over that of Fig. 2 is the saving in an external frequency synthesizer at the expense of a frequency multiplier (step up or step down converter).

The analysis of the loop illustrated in Fig. 3 follows along the lines of the classical approach to the standard configuration of Fig. 1 as documented in Ref. 1. We first look at the noise-free case and investigate the effect on the tracking performance of a possible phase shift, $\theta_s(t)$, in the bandpass filter.

2. Noise-Free Analysis

Let the received signal power be A^2 watts (i.e., a peak voltage of $2^{1/2}A$) and the nominal carrier frequency be denoted by ω_c. Then, the received signal is given by

$$s(t) = 2^{1/2}A \sin\theta(t) \tag{1}$$

where

$$\theta(t) = \omega_c t + \theta_1(t) \tag{2}$$

and $\theta_1(t)$ is the dynamic phase to be tracked by the loop. If the first demodulation is done at a nominal frequency ω_1, then the feedback reference signal at this demodulator can be expressed in the form

$$r_1(t) = 2^{1/2}K_1 \cos\hat{\theta}(t) \tag{3}$$

where

$$\hat{\theta}(t) = \omega_1 t + \theta_2(t) \tag{4}$$

and $\hat{\theta}(t)$ is related to the quiescent VCO frequency, ω_0, and the control signal to the VCO, $e(t)$, by

$$\frac{d\hat{\theta}(t)}{dt} = K_{f_1}\frac{d\theta_{vco}}{dt} = K_{f_1}[\omega_0 + K_{vco}e(t)] \tag{5}$$

Thus, from Eqs. (4) and (5),

$$\left.\begin{aligned}\omega_1 &= K_{f_1}\omega_0\\ \theta_2(t) &= K_{f_1}K_{\text{vco}}e(t)\end{aligned}\right\} \tag{6}$$

The output of the first demodulator contains the sum and difference frequencies $\omega_c - \omega_1$ and $\omega_c + \omega_1$, i.e.,

$$\begin{aligned}x_1(t) = AK_1\{&\sin[(\omega_c - \omega_1)t + \theta_1(t) - \theta_2(t)]\\ &+ \sin[(\omega_c + \omega_1)t + \theta_1(t) + \theta_2(t)]\}\end{aligned} \tag{7}$$

The bandpass filter eliminates the sum frequency $\omega_c + \omega_1$ and introduces a possible phase shift, $\theta_s(t)$. Hence, its output is

$$x_2(t) = AK_1 \sin[(\omega_c - \omega_1)t + \theta_1(t) - \theta_2(t) + \theta_s(t)]$$

Ignoring the effect of the limiter, this signal also represents the input to the second demodulator. The reference signal for the second demodulation, $r_2(t)$, is now derived from the VCO output by an appropriate frequency scaling such that its nominal frequency is $\omega_c - \omega_1$. Thus,

$$r_2(t) = 2K_2 \cos \hat{\hat{\theta}}(t) \tag{8}$$

where

$$\hat{\hat{\theta}}(t) = (\omega_c - \omega_1)t + \theta_3(t)$$

and $\theta(t)$ is related to the VCO parameters by

$$\frac{d\hat{\hat{\theta}}(t)}{dt} = K_{f_2}\frac{d\theta_{\text{vco}}}{dt} = K_{f_2}[\omega_0 + K_{\text{vco}}e(t)] \tag{9}$$

Hence,

$$\left.\begin{aligned}\omega_c - \omega_1 &= K_{f_2}\omega_0\\ \theta_3(t) &= K_{f_2}K_{\text{vco}}e(t)\end{aligned}\right\} \tag{10}$$

From Eqs. (9) and (10),

$$\left.\begin{aligned}\frac{K_{f_2}}{K_{f_1}} &= \frac{\omega_c - \omega_1}{\omega_1}\\ \theta_3(t) &= \frac{K_{f_2}}{K_{f_1}}\theta_2(t)\end{aligned}\right\} \tag{11}$$

The output of the second demodulator, $x_3(t)$, neglecting the sum-frequency term which is eliminated by the filter–VCO combination is

$$\begin{aligned}x_3(t) &= AK_1K_2 \sin[\theta_1(t) - \theta_2(t) - \theta_3(t) + \theta_s(t)]\\ &= AK_1K_2 \sin[\theta'_1(t) - \theta'_2(t)]\end{aligned} \tag{12}$$

where

$$\left.\begin{aligned}\theta'_1(t) &= \theta_1(t) + \theta_s(t)\\ \theta'_2(t) &= \theta_2(t) + \theta_3(t) = \theta_2(t)\left[1 + \frac{K_{f_2}}{K_{f_1}}\right]\end{aligned}\right\} \tag{13}$$

The linear time-invariant filter operates on $x_3(t)$ to produce the input signal to the VCO:

$$e(t) = F(p)x_3(t) \tag{14}$$

where p is the differential operation d/dt.

Defining

$$\phi(t) = \theta'_1(t) - \theta'_2(t) \tag{15}$$

and combining Eqs. (5–14) gives the loop equation

$$\frac{d\phi(t)}{dt} = \frac{d\theta'_1(t)}{dt} - AKF(p) \sin \phi(t) \tag{16}$$

where

$$K = K_1K_2(K_{f_1} + K_{f_2})K_{\text{vco}} \tag{17}$$

Hence, we see that $\theta_s(t)$, the phase shift in the bandpass filter, affects the loop as if it appeared at the input. Also, the equivalent phase estimate, $\theta'_2(t)$, is related to the phase estimate at the first demodulator [e., $\theta_2(t)$] by a constant $(1 + K_{f_2}/K_{f_1}) = \omega_c/\omega_1$. The equivalent mathematical model of this phase-locked loop topology is given in Fig. 4a.

3. Noise Analysis

If the input noise, $n(t)$, is zero mean and bandpass, with a flat, symmetric spectrum of width 2β, centered about $\pm\omega_c$ ($\beta < \omega_c$), then the narrow band expansion of $n(t)$ is valid, i.e.,

$$n(t) = 2^{1/2}[n_1(t) \sin \omega_c t + n_2(t) \cos \omega_c t] \tag{18}$$

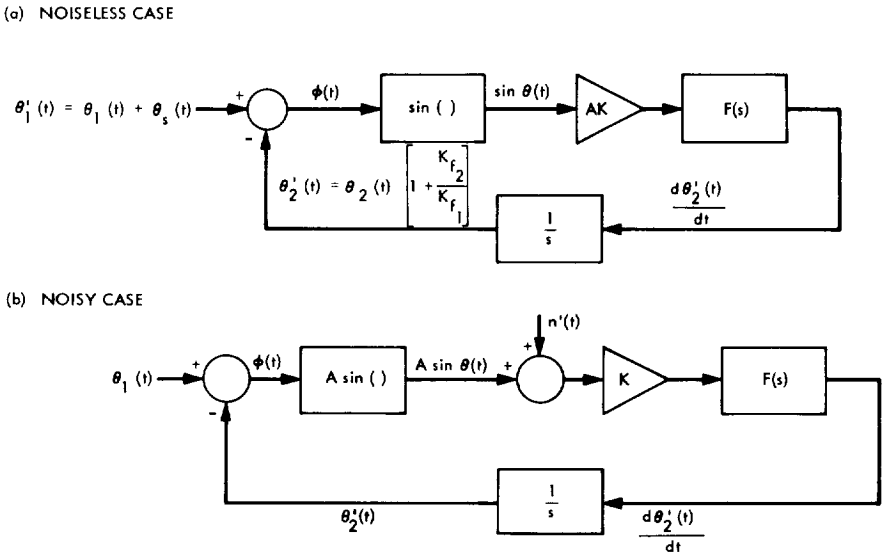


Fig. 4. Equivalent mathematical models of phase-locked loop model—variation 2

where $n_1(t)$ and $n_2(t)$ are independent, zero mean, gaussian processes with identical spectral densities which are the low-pass equivalent of that of $n(t)$. Since $\theta_2(t)$ now also includes the noise, the output of the first demodulator (neglecting sum-frequency terms) is now

$$\begin{aligned} x_1(t) = {} & AK_1 \sin [(\omega_c - \omega_1)t + \theta_1(t) - \theta_2(t)] \\ & + K_1 n_1(t) \sin [(\omega_c - \omega_1)t - \theta_2(t)] \\ & + K_2 n_2(t) \cos [(\omega_c - \omega_1)t - \theta_2(t)] \end{aligned} \quad (19)$$

If we neglect the effect of $\theta_s(t)$ in the noise case for simplicity, we find the output of the second demodulator is given by

$$\begin{aligned} x_3(t) = {} & AK_1K_2 \sin [\theta_1(t) - \theta_2'(t)] \\ & - K_1K_2 n_1(t) \sin [\theta_2'(t)] \\ & + K_1K_2 n_2(t) \cos [\theta_2'(t)] \end{aligned} \quad (20)$$

where $\theta_2'(t)$ is defined by Eq. (13).

Letting $\phi(t) = \theta_1(t) - \theta_2'(t)$, the stochastic differential equation for the loop becomes

$$\frac{d\phi(t)}{dt} = \frac{d\theta_1(t)}{dt} - KF(p)[A \sin \phi(t) + n'(t)] \quad (21)$$

where

$$n'(t) = -n_1(t) \sin \theta_2'(t) + n_2(t) \cos \theta_2'(t) \quad (22)$$

Since $\theta_2'(t)$ is related to the perturbation term in $\theta_{vco}(t)$ by a constant (namely $K_{f_1} + K_{f_2}$), then following similar arguments to those given on pages 30–33 of Ref. 1, $n'(t)$ is essentially a "white" gaussian process of flat spectral density in the bandwidth $|\omega| < \beta$. The equivalent mathematical model is given in Fig. 4b which is the same as for Fig. 1 with θ_2 replaced by θ_2' and K_1K_{vco} replaced by $K_1K_2(K_{f_1} + K_{f_2})K_{vco}$. Hence, with respect to an observer at the input of the filter, the performances of the two configurations are identical if the gains are adjusted to be equal as above.

References

1. Viterbi, A. J., *Principles of Coherent Communication*. McGraw-Hill Book Co., Inc., New York, 1966.
2. Tausworthe, R. C., *Theory and practical Design of Phase-Locked Receivers: Volume I*, Technical Report 32-819. Jet Propulsion Laboratory, Pasadena, Calif., Feb. 15, 1966.

C. On the Probability Density Function of the Output Statistic in an Absolute Value Type Lock Detector, *M. K. Simon*

1. Introduction

A proposed single-channel flight command system[2] for future missions suggests a data-derived bit synchronizer wherein the phase detector is of the absolute value type.

[2]Couvillon, L. A., *MM71 Command System—Technical Considerations*, May 1, 1969 (JPL internal document).

This particular design topology has already been supported by extensive analysis and experimental data.[3,4] The question immediately arises as to whether this absolute value type phase detector configuration can be successfully used as a means of detecting phase lock. The word successfully is herein intended to imply that a lock condition should be detectable within a relatively small number of bit intervals.

This article attempts to shed some light on this question by deriving the first-order probability statistics of the output of such a lock detector from which one may draw some general conclusions concerning performance. One hastens to add that first-order statistics are not sufficient to completely characterize the decision random variable at the lock detector output; nevertheless, they give a reasonable measure of performance if the correlation between successive output samples is weak. In any event, the adjacent output samples are only pairwise dependent, so at worst second-order statistics would be all that is required.

[3]SPS 37-55, Vol. III, pp. 62–70; SPS 37-51, Vol. III, pp. 310–313; SPS 37-50, Vol. III, pp. 326–331.

[4]Simon, M. K., *Nonlinear Analysis of an Absolute Value Type of Early-Late Gate Bit Synchronizer*, Dec. 4, 1968 (JPL internal document).

2. System Description

Consider the bit synchronizer–lock detector combination illustrated in Fig. 5. Assuming low-pass filters of the integrate and dump type, the steady-state nonlinear phase noise performance of the bit synchronizer has been studied in SPS 37-55, Vol. III, and Footnote 4. It was shown therein that the bit synchronizer could be replaced by an equivalent phase-locked loop with an S-curve and additive noise spectrum which in general are functions of the input signal-to-noise ratio. Of particular interest is the case where the early–late interval is 1/4 of a bit interval T. The feedback reference signals in the lock detector (i.e., A′ and B′) are then 1/4 of a bit interval advanced with respect to the corresponding

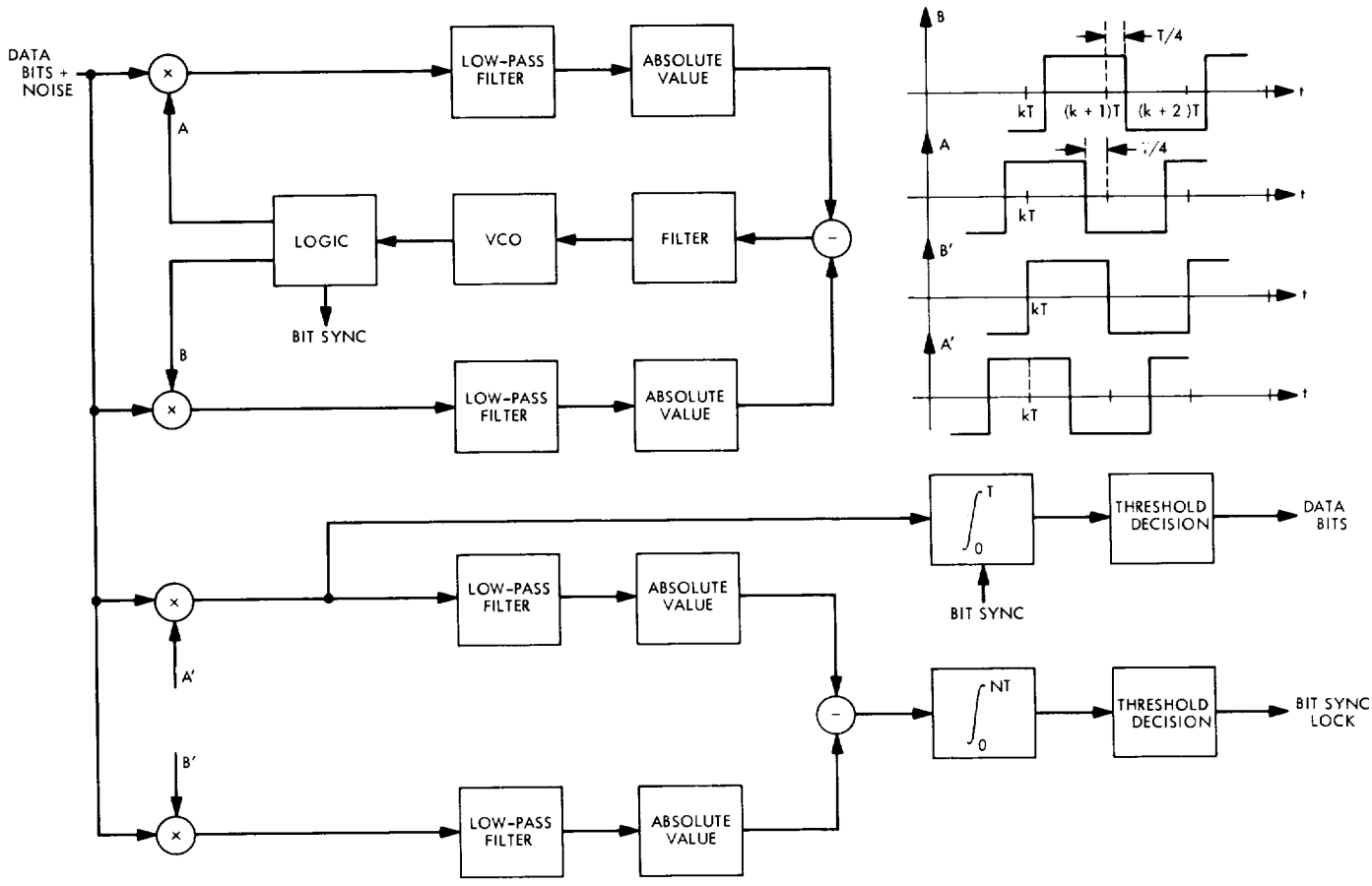


Fig. 5. Bit synchronizer–lock detector block diagram

cross-correlating signals in the bit synchronizer (A and B) as illustrated in Fig. 5. As a result, it is simple to show that the moments of the error random variable, e'_k, conditioned on the normalized offset, λ, are obtained from a $T/4$ shift of those associated with the random variable, e_k. Using Eqs. (10), (11), and (B-25) through (B-35) of Footnote 4,

$$\frac{E_{n,s}\{e'_k|\lambda\}}{2AT} = \begin{cases} \dfrac{\alpha_0}{4}\operatorname{erf}\alpha_0(R_s)^{1/2} - \dfrac{\beta_0}{4}\operatorname{erf}\beta_0(R_s)^{1/2} \\ + \dfrac{1}{4(R_s\pi)^{1/2}}\{\exp(-\alpha_0^2 R_s) \\ - \exp(-\beta_0^2 R_s)\}, \qquad -\dfrac{1}{2} \leqq \lambda \leqq 0 \\ \dfrac{\alpha_1}{4}\operatorname{erf}\alpha_1(R_s)^{1/2} - \dfrac{\beta_1}{4}\operatorname{erf}\beta_1(R_s)^{1/2} \\ + \dfrac{1}{4(R_s\pi)^{1/2}}\{\exp(-\alpha_1^2 R_s) \\ - \exp(-\beta_1^2 R_s)\}, \qquad 0 \leqq \lambda \leqq \dfrac{1}{2} \end{cases} \tag{1}$$

where A is the amplitude of the ± 1 random input signal pulse train, $R_s = A^2T/N_0$ is the input signal-to-noise ratio of the data, and

$$\left.\begin{aligned} \alpha_0 &= 1 + 2\lambda \\ \beta_0 &= \beta_1 = -2\lambda \\ \alpha_1 &= 1 - 2\lambda \end{aligned}\right\} \tag{2}$$

The subscripts n and s on the expectation denote averaging over both the noise and the random signal sequence.

$$\frac{E_{n,s}\{e'_k e'_{k+m}|\lambda\}^*}{2\sigma^2} \triangleq \frac{R(m,\lambda)}{2\sigma^2} = \begin{cases} 1 + \dfrac{5}{4}R_s - f\left[\dfrac{1}{2}, (2R_s)^{1/2}, (2R_s)^{1/2}\right] \\ - f\left[\dfrac{1}{2}, \dfrac{1}{2}(2R_s)^{1/2}, -\dfrac{1}{2}(2R_s)^{1/2}\right], \qquad m = 0 \\ \dfrac{1}{4}\left\{(R_s)^{1/2}\operatorname{erf}(R_s)^{1/2} + \dfrac{1}{2}(R_s)^{1/2}\operatorname{erf}\left[\dfrac{1}{2}(R_s)^{1/2}\right]\right. \\ \quad \left. + \dfrac{1}{\pi^{1/2}}\exp(-R_s) + \dfrac{1}{\pi^{1/2}}\exp\left(-\dfrac{R_s}{4}\right)\right\}^2 \\ - \dfrac{1}{4}\left\{f\left[\dfrac{1}{2}, (2R_s)^{1/2}, (2R_s)^{1/2}\right] + f\left[\dfrac{1}{2}, \dfrac{1}{2}(2R_s)^{1/2}, \dfrac{1}{2}(2R_s)^{1/2}\right]\right\}, \\ \qquad + 2f\left[\dfrac{1}{2}, \dfrac{1}{2}(2R_s)^{1/2}, (2R_s)^{1/2}\right]\Big\}, \qquad m = 1 \\ 0, \qquad\qquad m \geqq 2 \end{cases} \tag{3}$$

*Actually this result is exact only for $\lambda = 0$. However, for sufficiently large signal-to-noise ratio in the loop bandwidth, the variation of this quantity with λ may be considered negligible.

where

$$\left.\begin{aligned}\sigma^2 &= \frac{N_0 T}{2}\\ f(\rho, b, a) &= \frac{1}{(2\pi)^{1/2}}\int_0^\infty r(\rho, b, a, x)dx\end{aligned}\right\} \tag{4}$$

and

$$r(\rho, b, a, x) =$$

$$(1-\rho^2)^{1/2}x \exp\left[-\frac{x^2+a^2}{2}\right]$$

$$\times\left\{\left(\frac{2}{\pi}\right)^{1/2}\exp\left[-\frac{(\rho x)^2+(b-\rho a)^2}{2(1-\rho^2)}\right]\right.$$

$$\times\cosh\left[ax-\frac{\rho(b-\rho a)x}{1-\rho^2}\right]$$

$$+\left[\frac{\rho x+(b-\rho a)}{(1-\rho^2)^{1/2}}\right]\frac{\exp(ax)}{2}\operatorname{erf}\left[\frac{\rho x+(b-\rho a)}{[2(1-\rho^2)]^{1/2}}\right]$$

$$\left.+\left[\frac{\rho x-(b-\rho a)}{(1-\rho^2)^{1/2}}\right]\frac{\exp(-ax)}{2}\operatorname{erf}\left[\frac{\rho x-(b-\rho a)}{[2(1-\rho^2)]^{1/2}}\right]\right\} \tag{5}$$

We note from Eqs. (1) and (2) that when the bit synchronizer is phase-locked (i.e., $\lambda \approx 0$), the lock detector operates on the average in the neighborhood of the peak of the S-curve. What is more important, however, is the probability density function of e'_k since this gives the relative frequency of occurrence of the mean. Since the threshold at the lock detector output is ordinarily set based upon accumulation of the mean of e'_k, it is desirable that this first-order moment occur near a point of maximum probability.

3. Derivation of the First-Order Probability Density Function of e'_k

An equivalent functional block diagram for the lock detector is illustrated in Fig. 6. The error voltage, e'_k, can then be expressed as

$$e'_k = (c_k + \nu_k)\operatorname{sgn}(c_k + \nu_k) - (b_k + \mu_k)\operatorname{sgn}(b_k + \mu_k)$$

where

$$\left.\begin{aligned}c_k &= \int_{(k-1)T}^{kT} s(t)dt\\ \nu_k &= \int_{(k-1)T}^{kT} n(t)dt\\ b_k &= \int_{(k-1/2)T}^{(k+1/2)T} s(t)dt\\ \mu_k &= \int_{(k-1/2)T}^{(k+1/2)T} n(t)dt\end{aligned}\right\} \tag{6}$$

Letting d_k denote the kth input digit which takes on values $\pm A$ with equal probability,

$$\left.\begin{aligned}\frac{c_k}{T} &= d_{k-1}[1+\lambda] - d_k\lambda\\ \frac{b_k}{T} &= d_{k-1}\left[\frac{1}{2}+\lambda\right] + d_k\left[\frac{1}{2}-\lambda\right],\\ &\qquad -\frac{1}{2}\leq\lambda\leq 0\\ \frac{c_k}{T} &= d_{k-2}[\lambda] + d_{k-1}[1-\lambda]\\ \frac{b_k}{T} &= d_{k-1}\left[\frac{1}{2}+\lambda\right] + d_k\left[\frac{1}{2}-\lambda\right],\\ &\qquad 0\leq\lambda\leq\frac{1}{2}\end{aligned}\right\} \tag{7}$$

μ_k and ν_k are correlated gaussian zero random variables with

$$\sigma^2_{\nu_k} = \sigma^2_{\mu_k} = \frac{N_0 T}{2} = \sigma^2 \tag{8}$$

and

$$E\{\nu_k \mu_k\} = \frac{1}{2}\sigma^2$$

Thus,

$$p(\nu_k, \mu_k) = \frac{1}{2\pi\sigma^2\left(\frac{3}{4}\right)^{1/2}}\exp\left[-\frac{\nu_k^2+\mu_k^2-\nu_k\mu_k}{2\sigma^2\left(\frac{3}{4}\right)}\right] \tag{9}$$

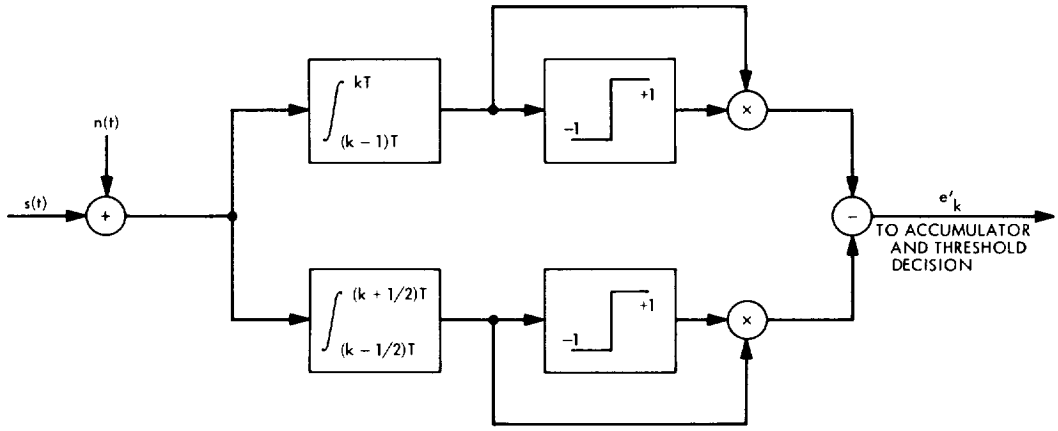


Fig. 6. Functional block diagram of lock detector

We begin by finding the conditional distribution of e'_k, namely,

$$F\{V \mid \lambda\} = 1 - \text{Prob}\{e'_k > V \mid \lambda\} \tag{10}$$

From Eq. (6),

$$\begin{aligned}\text{Prob}\{e'_k > V \mid \lambda\} = {} & \text{Prob}\{c_k + \nu_k > 0;\ b_k + \mu_k > 0;\ (c_k + \nu_k) - (b_k + \mu_k) > V\} \\ & + \text{Prob}\{c_k + \nu_k < 0;\ b_k + \mu_k < 0;\ (c_k + \nu_k) - (b_k + \mu_k) < -V\} \\ & + \text{Prob}\{c_k + \nu_k > 0;\ b_k + \mu_k < 0;\ (c_k + \nu_k) + (b_k + \mu_k) > V\} \\ & + \text{Prob}\{c_k + \nu_k < 0;\ b_k + \mu_k > 0;\ (c_k + \nu_k) + (b_k + \mu_k) < -V\}\end{aligned} \tag{11}$$

Equation (11) may be rewritten in integral form as

$$\begin{aligned}\text{Prob}\{e'_k > V \mid \lambda\} = {} & \int_{-b_k}^{\infty}\int_{\max\{V - c_k + \mu_k + b_k,\ -c_k\}}^{\infty} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-\infty}^{-b_k}\int_{-\infty}^{\min\{-V - c_k + \mu_k + b_k,\ -c_k\}} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-\infty}^{-b_k}\int_{\max\{V - c_k - b_k - \mu_k,\ -c_k\}}^{\infty} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-b_k}^{\infty}\int_{-\infty}^{\min\{-V - c_k - b_k - \mu_k,\ -c_k\}} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k\end{aligned} \tag{12}$$

For $V \geqq 0$, Eq. (12) becomes

$$\begin{aligned}\text{Prob}\{e'_k > V \mid \lambda\} = {} & \int_{-b_k}^{\infty}\int_{V - c_k + b_k + \mu_k}^{\infty} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-\infty}^{-b_k}\int_{-\infty}^{-V - c_k + \mu_k + b_k} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-\infty}^{-b_k}\int_{V - c_k - b_k - \mu_k}^{\infty} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k \\ & + \int_{-b_k}^{\infty}\int_{-\infty}^{-V - c_k - \mu_k - b_k} p(\mu_k, \nu_k)\, d\nu_k\, d\mu_k\end{aligned} \tag{13}$$

or after a change of variables,

$$\begin{aligned}\text{Prob}\{e'_k > V \mid \lambda\} = &\int_{-b_k}^{\infty}\int_{V-c_k+b_k+\mu_k}^{\infty} p(\nu_k, \mu_k)\, d\nu_k\, d\mu_k \\ &+ \int_{b_k}^{\infty}\int_{V+c_k-b_k+\mu_k}^{\infty} p(\nu_k, \mu_k)\, d\nu_k\, d\mu_k \\ &+ \int_{b_k}^{\infty}\int_{V-c_k-b_k+\mu_k}^{\infty} q(\nu_k, \mu_k)\, d\nu_k\, d\mu_k \\ &+ \int_{-b_k}^{\infty}\int_{V+c_k+b_k+\mu_k}^{\infty} q(\nu_k, \mu_k)\, d\nu_k\, d\mu_k \end{aligned} \tag{14}$$

where $p(\nu_k, \mu_k)$ is given by Eq. (9) and

$$q(\nu_k, \mu_k) = \frac{1}{2\pi\sigma^2\left(\frac{3}{4}\right)^{1/2}} \exp\left[-\frac{\nu_k^2 + \mu_k^2 + \mu_k\nu_k}{2\sigma^2\left(\frac{3}{4}\right)}\right] \tag{15}$$

The double integral of a two-dimensional gaussian function, $p(x,y)$, in a semi-infinite region may be evaluated as

$$\begin{aligned} I &= \int_a^{\infty}\int_{b(y)}^{\infty} p(x,y)\, dx\, dy \\ &= \frac{1}{4}\int_{a/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp[-z^2]\, \text{erfc}\left[\frac{\frac{b(2^{1/2}\sigma z)}{2^{1/2}\sigma} - \rho z}{(1-\rho^2)^{1/2}}\right] dz \end{aligned} \tag{16}$$

where

$$\rho = \frac{E\{xy\}}{\sigma^2} \tag{17}$$

and erfc[r] is the complementary error function defined by

$$\text{erfc}[r] = \frac{2}{\pi^{1/2}}\int_r^{\infty} \exp[-t^2]\, dt \tag{18}$$

For a two-dimensional function of the type defined by Eq. (15), the equivalent double integral is obtained by replacing ρ by $-\rho$ in Eq. (16), namely,

$$\begin{aligned} J &= \int_a^{\infty}\int_{b(y)}^{\infty} q(x,y) dx\, dy \\ &= \frac{1}{4}\int_{a/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp(-z^2)\, \text{erfc}\left[\frac{\frac{b(2^{1/2}\sigma z)}{2^{1/2}\sigma} + \rho z}{(1-\rho^2)^{1/2}}\right] dz \end{aligned} \tag{19}$$

where ρ is defined by Eq. (17).

Applying Eqs. (16–19) to Eq. (14), with $\rho = 1/2$, gives

$$\text{Prob}\{e'_k > V \mid \lambda\} = \frac{1}{4}\int_{-b_k/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp[-z^2] \operatorname{erfc}\left[\frac{\frac{V - c_k + b_k}{2^{1/2}\sigma} + \frac{z}{2}}{\left(\frac{3}{4}\right)^{1/2}}\right] dz$$

$$+ \frac{1}{4}\int_{b_k/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp[-z^2] \operatorname{erfc}\left[\frac{\frac{V + c_k - b_k}{2^{1/2}\sigma} + \frac{z}{2}}{\left(\frac{3}{4}\right)^{1/2}}\right] dz$$

$$+ \frac{1}{4}\int_{b_k/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp[-z^2] \operatorname{erfc}\left[\frac{\frac{V - c_k - b_k}{2^{1/2}\sigma} + \frac{3z}{2}}{\left(\frac{3}{4}\right)^{1/2}}\right] dz$$

$$+ \frac{1}{4}\int_{-b_k/2^{1/2}\sigma}^{\infty} \frac{2}{\pi^{1/2}} \exp[-z^2] \operatorname{erfc}\left[\frac{\frac{V + c_k + b_k}{2^{1/2}\sigma} + \frac{3z}{2}}{\left(\frac{3}{4}\right)^{1/2}}\right] dz \tag{20}$$

Substituting Eq. (7) in Eq. (20) and averaging over all equally likely sequences formed from d_{k-2}, d_{k-1}, d_k gives the final result

$$\text{Prob}\{e'_k > V \mid \lambda\} = 1 - F(V \mid \lambda)$$

$$= \frac{1}{8}\int_{(R_s)^{1/2}}^{\infty} \left[g\left(\frac{1}{2}, (R_s)^{1/2} W, z\right) + g\left(\frac{3}{2}, (R_s)^{1/2}(W-2), z\right)\right] dz$$

$$+ \frac{1}{8}\int_{-(R_s)^{1/2}}^{\infty} \left[g\left(\frac{1}{2}, (R_s)^{1/2} W, z\right) + g\left(\frac{3}{2}, (R_s)^{1/2}(W+2), z\right)\right] dz$$

$$+ \frac{1}{8}\int_{2\lambda(R_s)^{1/2}}^{\infty} \left[g\left(\frac{1}{2}, (R_s)^{1/2}(W+1), z\right) + g\left(\frac{3}{2}, (R_s)^{1/2}(W-1-4\lambda), z\right)\right] dz$$

$$+ \frac{1}{8}\int_{-2\lambda(R_s)^{1/2}}^{\infty} \left[g\left(\frac{1}{2}, (R_s)^{1/2}(W-1), z\right) + g\left(\frac{3}{2}, (R_s)^{1/2}(W+1+4\lambda), z\right)\right] dz \tag{21}$$

where

$$g(a, b, z) = \frac{2}{\pi^{1/2}} \exp(-z^2) \operatorname{erfc}\left[\frac{az + b}{\left(\frac{3}{4}\right)^{1/2}}\right] \tag{22}$$

and

$$W = V/AT \tag{23}$$

The conditional probability density of e'_k is found by differentiating $F(V \mid \lambda)$, i.e.,

$$p(e'_k \mid \lambda) = -\frac{d}{dV}\left[1 - F(V \mid \lambda)\right]\Big|_{V=e'_k} \tag{24}$$

If we normalize e'_k by AT, i.e.,

$$\epsilon'_k = \frac{e'_k}{AT} \tag{25}$$

then,

$$p(\epsilon'_k \mid \lambda) = -\frac{d}{dW}\left[1 - F(W \mid \lambda)\right]\Big|_{W=\epsilon'_k} \tag{26}$$

where W is defined by Eq. (23).

Since

$$\frac{d}{dx}\operatorname{erfc}\left[ax + b\right] = -a\frac{2}{\pi^{1/2}}\exp\left[-(ax+b)^2\right] \tag{27}$$

$p(e'_k \mid \lambda)$ may be expressed in the form of Eq. (21) with $g(a,b,z)$ replaced by $h(a,b,z)$, and $h(a,b,z)$defined by

$$h(a,b,z) = \frac{4}{\pi}\left(\frac{R_s}{\frac{3}{4}}\right)^{1/2}\exp\left(-z^2\right)\exp\left[-\left(\frac{az+b}{\left(\frac{3}{4}\right)^{1/2}}\right)^2\right] \tag{28}$$

After completing the square, term by term, the final result is in closed form:

$p(\epsilon'_k \mid \lambda) =$

$$\begin{aligned}
&\frac{1}{4}\left(\frac{R_s}{\pi}\right)^{1/2}\Bigg\{\exp\left[-R_s(\epsilon'_k)^2\right]\operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(1+\frac{\epsilon'_k}{2}\right)\right] + \exp\left[-R_s(\epsilon'_k)^2\right]\operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(-1+\frac{\epsilon'_k}{2}\right)\right] \\
&+ \exp\left[-R_s(\epsilon'_k+1)^2\right]\operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(2\lambda+\frac{\epsilon'_k+1}{2}\right)\right] + \exp\left[-R_s(\epsilon'_k-1)^2\right]\operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(-2\lambda+\frac{\epsilon'_k-1}{2}\right)\right]\Bigg\} \\
&+ \frac{1}{4}\left(\frac{R_s}{3\pi}\right)^{1/2}\Bigg\{\exp\left[\frac{-R_s(\epsilon'_k+2)^2}{3}\right]\operatorname{erfc}\left[(4R_s)^{1/2}\left(\frac{\epsilon'_k}{2}\right)\right] + \exp\left[\frac{-R_s(\epsilon'_k-2)^2}{3}\right]\operatorname{erfc}\left[(4R_s)^{1/2}\left(\frac{\epsilon'_k}{2}\right)\right] \\
&+ \exp\left[\frac{-R_s(\epsilon'_k+1+4\lambda)^2}{3}\right]\operatorname{erfc}\left[(4R_s)^{1/2}\left(\frac{\epsilon'_k+1}{2}\right)\right] + \exp\left[\frac{-R_s(\epsilon'_k-1-4\lambda)^2}{3}\right]\operatorname{erfc}\left[(4R_s)^2\left(\frac{\epsilon'_k-1}{2}\right)\right]\Bigg\},
\end{aligned}$$

$$\epsilon'_k \geqq 0;\ -\frac{1}{2} \leqq \lambda \leqq 0 \tag{29}$$

By a similar procedure to that just outlined, the conditional density function of ϵ_k' in the region, $0 \leq \lambda \leq 1/2$ ($\epsilon_k' \geq 0$), can be determined as:

$$
\begin{aligned}
p(\epsilon_k' \mid \lambda) = \\
&\frac{1}{8}\left(\frac{R_s}{\pi}\right)^{1/2} \Bigg\{ \exp\left[-R_s(\epsilon_k')^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(1+\frac{\epsilon_k'}{2}\right)\right] + \exp\left[-R_s(\epsilon_k')^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(-1+\frac{\epsilon_k'}{2}\right)\right] \\
&+ \exp\left[-R_s(\epsilon_k'-2\lambda)^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(1-\lambda+\frac{\epsilon_k'}{2}\right)\right] + \exp\left[-R_s(\epsilon_k'+2\lambda)^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(-1+\lambda+\frac{\epsilon_k'}{2}\right)\right] \\
&+ \exp\left[-R_s(\epsilon_k'+1-2\lambda)^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(\lambda+\frac{\epsilon_k'+1}{2}\right)\right] + \exp\left[-R_s(\epsilon_k'-1+2\lambda)^2\right] \\
&\times \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(-\lambda+\frac{\epsilon_k'-1}{2}\right)\right] \\
&+ \exp\left[-R_s(\epsilon_k'+1-4\lambda)^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(\frac{\epsilon_k'+1}{2}\right)\right] + \exp\left[-R_s(\epsilon_k'-1+4\lambda)^2\right] \operatorname{erfc}\left[\left(\frac{4R_s}{3}\right)^{1/2}\left(\frac{\epsilon_k'-1}{2}\right)\right]\Bigg\} \\
&+ \frac{1}{8}\left(\frac{R_s}{3\pi}\right)^{1/2} \Bigg\{ \exp\left[\frac{-R_s(\epsilon_k'+2)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(\frac{\epsilon_k'}{2}\right)\right] + \exp\left[\frac{-R_s(\epsilon_k'-2)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(\frac{\epsilon_k'}{2}\right)\right] \\
&+ \exp\left[\frac{-R_s(\epsilon_k'+2-2\lambda)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(-\lambda+\frac{\epsilon_k'}{2}\right)\right] + \exp\left[\frac{-R_s(\epsilon_k'-2+2\lambda)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(\lambda+\frac{\epsilon_k'}{2}\right)\right] \\
&+ \exp\left[\frac{-R_s(\epsilon_k'+1+2\lambda)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(-\lambda+\frac{\epsilon_k'+1}{2}\right)\right] + \exp\left[\frac{-R_s(\epsilon_k'-1-2\lambda)^2}{3}\right] \\
&\times \operatorname{erfc}\left[(4R_s)^{1/2}\left(\lambda+\frac{\epsilon_k'-1}{2}\right)\right] \\
&+ \exp\left[\frac{-R_s(\epsilon_k'+1)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(-2\lambda+\frac{\epsilon_k'+1}{2}\right)\right] \times \exp\left[\frac{-R_s(\epsilon_k'-1)^2}{3}\right] \operatorname{erfc}\left[(4R_s)^{1/2}\left(2\lambda+\frac{\epsilon_k'-1}{2}\right)\right]\Bigg\},
\end{aligned}
$$

$$\epsilon_k' \geq 0;\ 0 \leq \lambda \leq \frac{1}{2} \tag{30}$$

Finally, for negative ϵ_k', one simply replaces ϵ_k' by $-\epsilon_k'$ in the arguments of the erfc functions of Eqs. (29) and (30).

The unconditional density function of ϵ_k' is found by averaging $p(\epsilon_k' \mid \lambda)$ over the density function of the phase error, $p(\lambda)$, as determined by the bit synchronizer [Eq. (17) of Footnote 4]. As previously mentioned, for a large signal-to-noise ratio in the loop bandwidth, the bit synchronizer phase error distribution will be narrow and centered around $\lambda = 0$. Hence, to a good approximation,

$$p(\epsilon_k') \approx p(\epsilon_k' \mid \lambda)\big|_{\lambda=0} \tag{31}$$

Figures 7–12 plot $p(\epsilon_k' \mid \lambda)$ versus ϵ_k' with λ as a parameter (negative and positive) and $R_s = 3, 5, 10$. It is interesting to note that although the conditional mean of ϵ_k' is an even function of phase offset, λ, the first-order conditional density function, $p(\epsilon_k' \mid \lambda)$, is not. For increasing

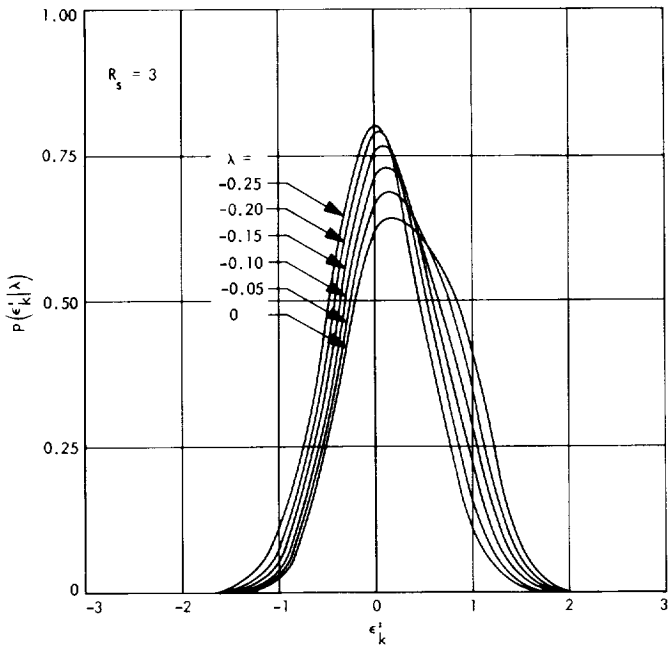


Fig. 7. Conditional probability density function of ϵ'_k for $R_s = 3$, $\lambda = -0.25$ to 0

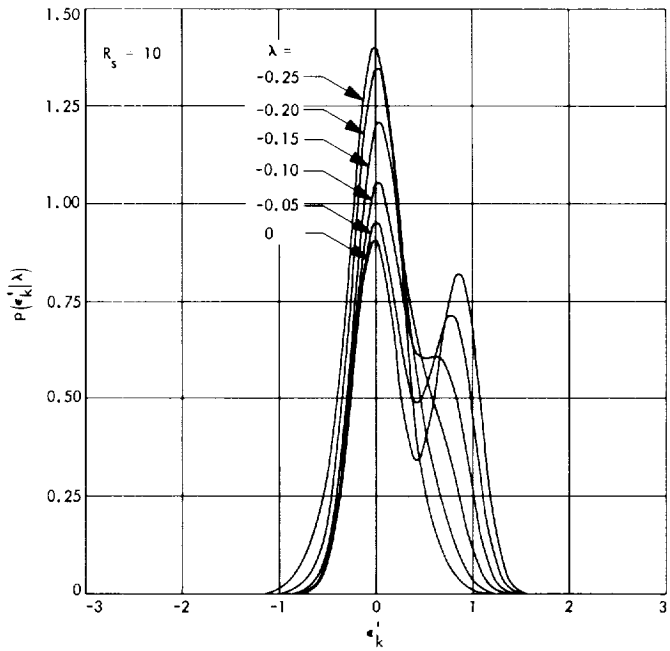


Fig. 9. Conditional probability density function of ϵ'_k for $R_s = 10$, $\lambda = -0.25$ to 0

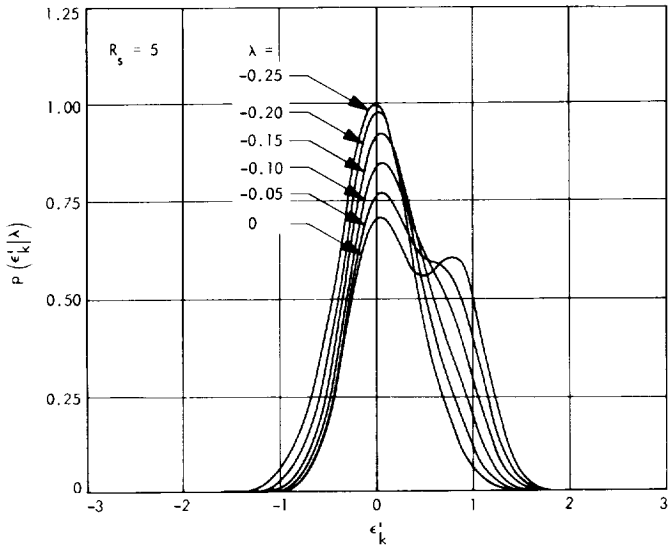


Fig. 8. Conditional probability density function of ϵ'_k for $R_s = 5$, $\lambda = -0.25$ to 0

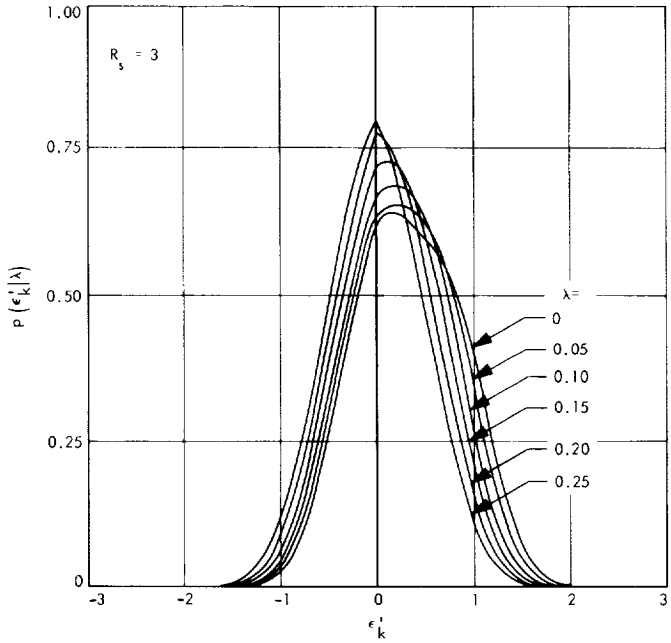


Fig. 10. Conditional probability density function of ϵ'_k for $R_s = 3$, $\lambda = 0$ to 0.25

R_s, the particular density function, $p(\epsilon'_k \mid \lambda) \mid_{\lambda=0}$, becomes bimodal. Furthermore, it appears that the corresponding mean is located approximately midway between the probability peaks, and thus at a point of *minimum* probability. In fact, in the limit for large enough input signal-to-noise ratio, the value of the probability density function at the mean goes to zero. The conclusion to be drawn from this is that for relatively small accumulation intervals before threshold decision, the random variable, ϵ'_k, will tend to

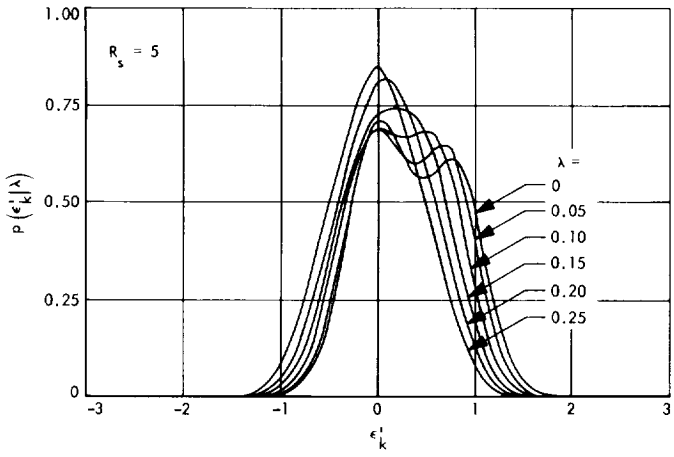


Fig. 11. Conditional probability density function of ϵ'_k for $R_s = 5$, $\lambda = 0$ to 0.25

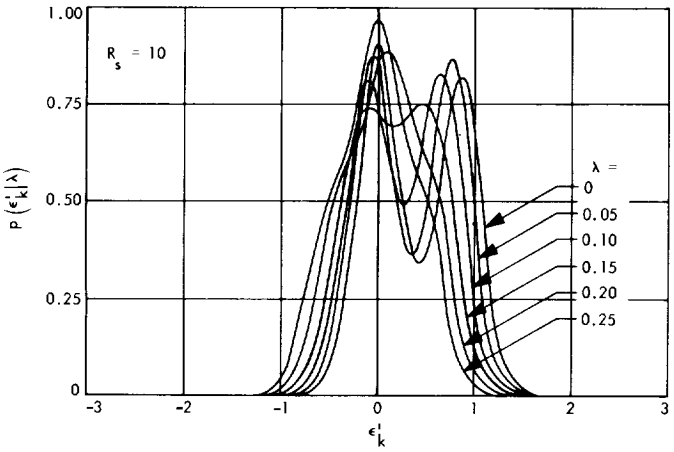


Fig. 12. Conditional probability density function of ϵ'_k for $R_s = 10$, $\lambda = 0$ to 0.25

sit at one of the two probability peaks; hence, a lock decision based upon accumulation of the mean will be in error. In fact, for a one-bit decision interval, and a large input signal-to-noise ratio, there is essentially a 50-50 chance of being at either of the two peaks. As the accumulation interval becomes large, the input to the threshold decision device approaches the average of the two probability peaks which is close to the mean. The moral of the story is: a phase detector topology which yields good performance as a bit synchronizer may not be desirable when used as a lock detector particularly because of its nonlinear behavior in the neighborhood of the peak of the loop S-curve. Other lock detector configurations are presently under investigation as possible alternates.

D. The Performance of Suppressed Carrier Tracking Loops in the Presence of Frequency Detuning, *W. C. Lindsey and M. K. Simon*

1. Introduction

This article establishes the performance of two widely used circuits viz., the squaring loop and the Costas loop, used in communications engineering for purposes of: (1) demodulation of double-sideband suppressed carrier (DSB-SC) analog signals, and (2) extraction of a carrier or subcarrier reference for use in a phase-coherent receiver where digital modulation is transmitted. Previous work (Ref. 1) has established the nonlinear performance of these two circuits shown in Figs. 13 and 14 for the case where the loop filters have unit gain, i.e., first-order loops. This article extends these results to the case of higher-order referencing extracting loops and, in particular, treats the case of greatest practical interest, viz., the second-order loop with frequency detuning. The results are sufficiently general to establish the so-called "squaring loss" for the various practical prefiltering selectivity characteristics, e.g., filters whose impulse response can be modeled as the Butterworth, Bessel, Laguerre, or Chebyshev type. This loss is evaluated analytically for several typical examples.

In the case of coherent demodulation of analog signals, the output low-pass filter characteristics of Figs. 13 and 14 are determined using the Wiener filtering theory. This problem has been treated (Ref. 2) for the case of linear modulation at the transmitter by one of two classes of modulation processes, viz., the Butterworth and asymptotically gaussian processes. In Figs. 13 and 14, $\hat{m}(t)$ represents the receiver's reconstruction of the transmitted message, $m(t)$.

If the modulation consists of a sequence of equiprobable ± 1's, as in pulse-code-modulated (PCM) telemetry, the output low-pass filter becomes a cross-correlator which is implemented by an integrate and dump circuit. The coherent reference for this filter is obtained via the squaring loop or Costas loop. This article is primarily concerned with the problem of establishing a coherent reference for purposes of extracting digital data as well as establishing the performance (error probability) of the data detector as a function of system parameters, e.g., frequency detuning, loop signal-to-noise ratio, etc. As previously mentioned, for the case where the modulation is a zero mean stochastic process, the results given here and those presented in Ref. 2 can be used to establish system performance. In what follows, we draw heavily upon the notation introduced in Refs. 1, 2, and 3.

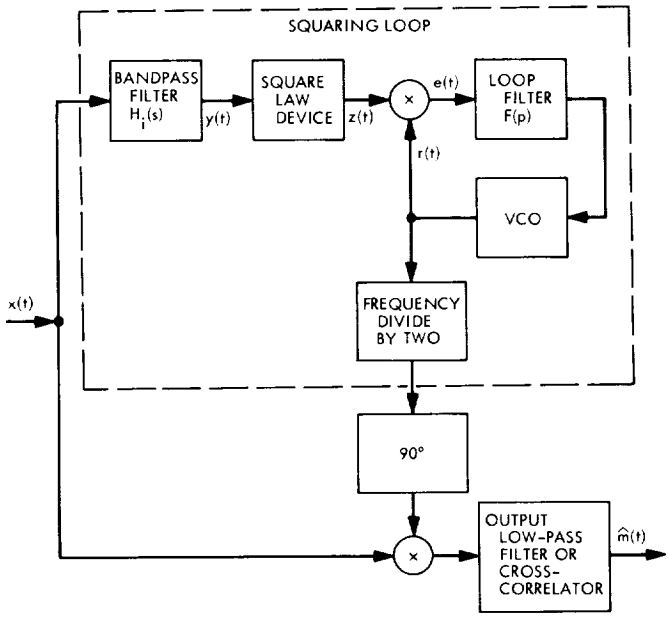


Fig. 13. Coherent demodulator with squaring loop reference extractor for DSB-SC signals

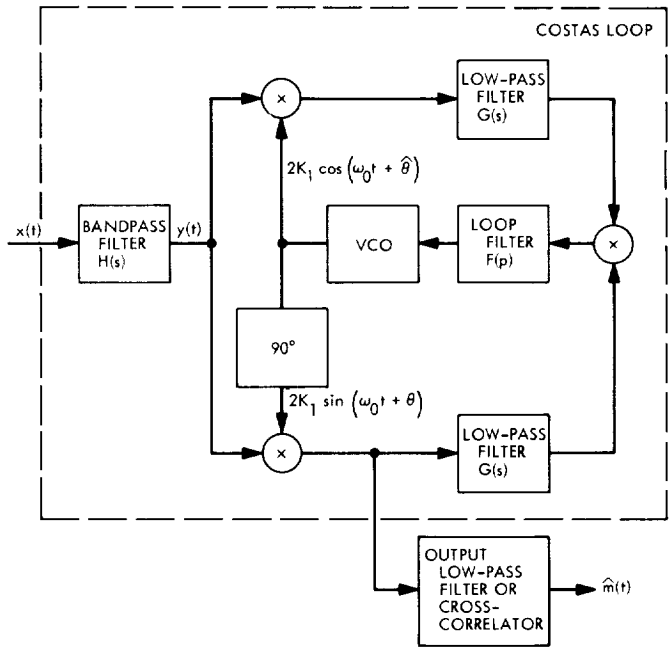


Fig. 14. Coherent demodulator with Costas loop reference extractor for DSB-SC signals

2. Coherent Demodulation of DSB Signals With a Squaring Loop or a Costas Loop

In this subsection, we establish the stochastic differential equation of operation of the squaring loop outlined in Fig. 13 and introduce notation which will be necessary in later analyses. As has been previously shown (Ref. 1), coherent demodulation of DSB-SC signals with a Costas loop (see Fig. 14) is mathematically equivalent to the circuit performance of Fig. 13, i.e., the circuits of Figs. 13 and 14 have the same stochastic differential equation of operation.[5] Thus, the choice of circuit to be used in a particular system design becomes entirely a problem of implementation.

Let the observed data $x(t)$ be given by

$$x(t) = 2^{1/2} Am(t) \sin \Theta(t) + n_i(t) \tag{1}$$

where $\Theta(t) = 2\pi f_0 t + \theta(t)$, $m(t)$ is the signal modulation, f_0 is the nominal carrier frequency, and $\theta(t)$ is the random phase to be tracked by the squaring loop. The noise, $n_i(t)$, is assumed white gaussian with one-sided spectral density N_0 W/Hz. The received signal is then bandpass filtered by $H_i(s)$ with the resultant output, $y(t)$, in the form

$$y(t) = 2^{1/2} Am(t) \sin \Theta(t) + n(t) \tag{2}$$

where the bandwidth B_i of $H_i(s)$ has been assumed wide with respect to the bandwidth of the modulation, and its transfer function introduces no additional phase shift. For the case of digital data transmission, the modulation will be in the form of a ± 1 pulse train with period T. Hence, we are equivalently assuming that $T \gg 1/B_i$. Furthermore, if $B_i < f_0$, then the noise $n(t)$ can be expressed in the form of a narrow band process (Ref. 4) about the nominal center frequency of the bandpass filter, i.e.,

$$n(t) = 2^{1/2}[n_A(t) \cos 2\pi f_0 t + n_B(t) \sin 2\pi f_0 t] \tag{3}$$

where $n_A(t)$ and $n_B(t)$ are statistically independent gaussian noise process with identical one-sided spectra corresponding to the low-pass equivalent of $n(t)$ (i.e., $N_0 \left| H_i(j\omega) \right|^2$). Alternately, one can expand the noise $n(t)$ about the actual frequency of the input observed data by

$$n(t) = 2^{1/2}[n_1(t) \cos \Theta(t) + n_2(t) \sin \Theta(t)] \tag{4}$$

[5]We assume here that the shaping of the noise spectrum produced by the cascade of $G(s)$ and $H(s)$ in Fig. 14 is equivalent to that produced alone by $H_i(s)$ in Fig. 13.

with $n_1(t)$ and $n_2(t)$ given by

$$\left.\begin{aligned} n_1(t) &\triangleq n_A(t)\cos\theta(t) - n_B\sin\theta(t) \\ n_2(t) &\triangleq n_A(t)\sin\theta(t) + n_B\cos\theta(t) \end{aligned}\right\} \tag{5}$$

Assuming $\theta(t)$ is narrow band relative to B_i (e.g., $\theta(t) = \theta_0 + \Omega_0 t$; $\Omega_0 \ll 2\pi B_i$), then following arguments similar to those given in Viterbi (Ref. 5), it can be concluded that $n_1(t)$ and $n_2(t)$ are approximately statistically independent gaussian variables with spectra equivalent to those of $n_A(t)$ and $n_B(t)$, respectively. Hereafter, we use the expansion of $n(t)$ given in Eq. (3) under the above assumptions.

Assuming a perfect square-law device, the input to the phase-locked loop (PLL) part of the overall squaring loop is (keeping only terms around $2f_0$)[6]

$$\begin{aligned} z(t) = y^2(t) = [&-A^2m(t) + n_1^2(t) - n_2^2(t) \\ &- 2Am(t)n_2(t)]\cos 2\Theta(t) \\ &+ [2Am(t)n_1(t) + 2n_1(t)n_2(t)]\sin 2\Theta(t) \end{aligned} \tag{6}$$

The reference signal, $r(t)$, in Fig. 13 is conveniently represented by

$$r(t) = 2K_1\sin 2\hat{\Theta}(t) = 2K_1\sin[2(2\pi f_0 t) + 2\hat{\theta}(t)] \tag{7}$$

where $\hat{\theta}(t)$ is the PLL estimate of $\theta(t)$. The dynamic error signal then becomes (keeping only baseband frequency components)

$$\begin{aligned} e(t) &= K_m z(t) r(t) \\ &= K_m K_1\{[A^2m^2(t) - n_1^2(t) + n_2^2(t) \\ &\quad + 2Am(t)n_2(t)]\sin 2\phi(t) \\ &\quad + [2Am(t)n_1(t) + 2n_1(t)]\cos 2\phi(t)\} \end{aligned} \tag{8}$$

where $\phi(t) \triangleq \theta(t) - \hat{\theta}(t)$, and K_m is the multiplier gain.

[6]This assumes a one-sided loop bandwidth $B_L \ll 1/T \ll B_i$ (i.e., $1/T$ is the effective measure of bandwidth for the data modulation).

The instantaneous frequency at the voltage-controlled oscillator (VCO) output is related to $e(t)$ by

$$\frac{2d\hat{\Theta}(t)}{dt} = K_V[F(p)e(t) + n_V(t)] + 2(2\pi f_0) \tag{9}$$

where K_V is the VCO gain constant in rad/s/V, $F(p)$ is the loop filter with p the differential operator d/dt, and $n_V(t)$ is a noise process associated with the VCO. If the input phase process is now characterized by $\theta(t) = \theta_0 + \Omega_0 t$, then the stochastic differential equation of operation of Fig. 13 becomes

$$\begin{aligned} \frac{d\Phi(t)}{dt} = 2\Omega_0 &- KF(p)[A^2m^2(t)\sin\Phi(t) \\ &+ v(t,\Phi(t))] - K_V n_V(t) \end{aligned} \tag{10}$$

where $K = K_1K_mK_V$, $\Phi(t) = 2\phi(t)$, and

$$\begin{aligned} v(t,\Phi(t)) = [&-n_1^2(t) + n_2^2(t) + 2Am(t)n_2(t)]\sin\Phi(t) \\ &+ [2Am(t)n_1(t) + 2n_1(t)n_2(t)]\cos\Phi(t) \end{aligned} \tag{11}$$

$\Phi(t)$ represents the actual phase being tracked by the loop. In the case of digital modulation, $m(t) = \pm 1$ so that $m^2(t) = 1$. For the case where the message $m(t)$ is a zero mean random process, one replaces $m^2(t)$ in Eq. (10) by its mean-squared value. Assuming that the process is normalized such that $\overline{m^2} = 1$ we have

$$\dot{\Phi}(t) = 2\Omega_0 - KF(p)[A^2\sin\Phi(t) + v(t,\Phi(t))] - K_V n_V(t) \tag{12}$$

To determine the steady-state probability density function of $\Phi(t)$, we apply a procedure based upon the Fokker–Planck equation.

3. The Fokker–Planck Equation for Approximate Markov Processes

The method used in determining the probability density function, $p(\Phi)$, is that given in Ref. 6 which is based upon the "fluctuation equation" approach. Basically, the method acknowledges the fact that under certain circumstances, a non-Markov process can be replaced or

approximated by a Markov process. The validity of this replacement is discussed in detail in Ref. 6.

To begin the development, the loop filter $F(p)$ is expressed in terms of its Heaviside expansion, viz.,

$$F(p) = F_0 + \sum_{k=1}^{N} \frac{1 - F_k}{1 + \tau_k p} \tag{13}$$

Substituting Eq. (13) into Eq. (12) and introducing the coordinates

$$\left.\begin{aligned} y_0(t) &\triangleq \Phi(t) \\ y_k(t) &\triangleq -\frac{(1 - F_k)K}{1 + \tau_k p}\left[A^2 \sin \Phi(t) + v(t, \Phi(t))\right], \\ & \qquad k = 1, 2, \cdots, N \end{aligned}\right\} \tag{14}$$

yields a set of $N + 1$ first-order stochastic differential equations, i.e.,

$$\left.\begin{aligned} \dot{\Phi}(t) &= 2\Omega_0 - KF_0\left[A^2 \sin \Phi(t) \right. \\ &\quad \left. + v(t, \Phi(t))\right] - K_V n_V(t) + \sum_{k=1}^{N} y_k(t) \\ \dot{y}_k(t) &= -\frac{y_k(t)}{\tau_k} - \frac{(1 - F_k)K}{\tau_k} \\ &\quad \times \left[A^2 \sin \Phi(t) + v(t, \Phi(t))\right], \\ & \qquad k = 1, 2, \cdots, N \end{aligned}\right\} \tag{15}$$

The vector $\mathbf{y} = [y_0, y_1, \cdots, y_N]$ represents, in an appropriate $N + 1$ dimensional space, the state of the system at time t. If, in fact, $v(t, \Phi(t))$ and $n_V(t)$ were white gaussian noise processes, then $\mathbf{y}$ would be a Markov vector (Ref. 6), and the Fokker–Planck equation could be applied directly. However, if the correlation time (Ref. 6) of the processes $n_V(t)$ and $v(t, \Phi(t))$ is small in relation to the correlation time of the system, one can replace $\mathbf{y}$ by an approximate vector Markov process and apply the "fluctuation equation" approach discussed by Stratonovich (Ref. 6).

Each component of $\mathbf{y}$ can be expressed as a generalized nonlinear function of $\mathbf{y}$, $v(t, \Phi(t))$, and $n_V(t)$, viz.,

$$y_k(t) = G_k\left[\mathbf{y}, v(t, y_0(t))\right], \qquad k = 0, 1, \cdots, N \tag{16}$$

so that the steady-state solution for the joint probability density function $p(\mathbf{y})$ satisfies, in the diffusion approximation, the equation of probability flow per unit time, viz.,

$$\nabla \cdot \mathcal{J}(\mathbf{y}) = 0 \tag{17}$$

where

$$\nabla \triangleq \left[\frac{\partial}{\partial y_0}, \cdots, \frac{\partial}{\partial y_N}\right]$$

is the del operator. Here $\mathcal{J} = (\mathcal{J}_0, \mathcal{J}_1, \cdots, \mathcal{J}_N)$ is the probability current density (Ref. 3) with the projections $\mathcal{J}_k = \boldsymbol{\varepsilon}_k \cdot \mathcal{J}$; $\boldsymbol{\varepsilon}_k$ being a unit vector pointed in the positive direction of the kth coordinate axis. In the steady-state, the kth projection of the probability current is defined by

$$\mathcal{J}_k(\mathbf{y}) \triangleq \left[K_k(\mathbf{y}) - \frac{1}{2}\sum_{l=0}^{N} \frac{\partial}{\partial y_l} K_{lk}(\mathbf{y})\right] p(\mathbf{y}) \tag{18}$$

with intensity coefficients $K_k(\mathbf{y})$ and $K_{lk}(\mathbf{y})$ given by

$$\left.\begin{aligned} K_l(\mathbf{y}) &\triangleq \overline{G_l[\mathbf{y}, v(t, y_0)]} \\ K_{lk}(\mathbf{y}) &\triangleq \int_{-\infty}^{\infty} \overline{\{G_l[\mathbf{y}, v(t, y_0)]\, G_k[\mathbf{y}, v(t + \tau, y_0)] - K_l(\mathbf{y})K_k(\mathbf{y})\}}\, d\tau \end{aligned}\right\} \tag{19}$$

In Eq. (19), the overbar denotes statistical average conditioned on a fixed **y**. Using Eqs. (15) and (16) the intensity coefficients in Eq. (19) are evaluated as

$$\left.\begin{aligned} K_0(\mathbf{y}) &= 2\Omega_0 - F_0A^2K\sin\Phi + \sum_{k=1}^{N} y_k \\ K_l(\mathbf{y}) &= -\frac{y_l}{\tau_l} - \frac{(1-F_l)}{\tau_l}A^2K\sin\Phi, \\ &\qquad l = 1, 2, \cdots, N \\ K_{00}(\mathbf{y}) &= K^2F_0^2\int_{-\infty}^{\infty}\left[R_v(\tau) + \frac{K_V^2}{K^2F_0^2}R_V(\tau)\right]d\tau, \\ &\qquad l = 1, 2, \cdots, N \\ K_{lk}(\mathbf{y}) &= \frac{K^2(1-F_l)(1-F_m)}{\tau_l\tau_m}\int_{-\infty}^{\infty}R_v(\tau)\,d\tau, \\ &\qquad l, k = 1, 2, \cdots, N \end{aligned}\right\} \tag{20}$$

where $n_v(t)$ is assumed independent of the input noise $n_i(t)$ and

$$\left.\begin{aligned} R_v(\tau) &\triangleq \overline{v(t,\Phi)\,v(t+\tau,\Phi)} \\ R_V(\tau) &\triangleq \overline{n_V(t)\,n_V(t+\tau)} \end{aligned}\right\} \tag{21}$$

a. The probability density of the phase error. Upon substitution of Eq. (20) into Eqs. (17) and (18), the problem of finding the probability density function of $p(\Phi)$ reduces to a special case of the solution given in Ref. 3. Without belaboring the details, the probability density function of the phase-error can be approximated by

$$p(\Phi) = \frac{\exp[\beta\Phi + \alpha\cos\Phi]}{4\pi^2\exp(-\pi\beta)\,|\,I_{j\beta}(\alpha)\,|^2}\int_{\Phi}^{\Phi+2\pi}\exp[-\beta x - \alpha\cos x]\,dx, \qquad |\Phi| \leqq \pi \tag{22}$$

where

$$\left.\begin{aligned} \beta &\triangleq \frac{4}{N_{sq}F_0^2K^2}\left[2\Omega_0 - A^2K\overline{\sin\Phi}\sum_{k=1}^{N}(1-F_k)\left(1 + \frac{N_{sq}}{4A^4\tau_k\sigma_G^2}\right)\right] \\ \alpha &\triangleq \frac{4}{N_{sq}F_0^2K^2}\left[A^2KF_0 - \frac{KN_{sq}}{4A^2\sigma_G^2}\sum_{k=1}^{N}\left(\frac{1-F_k}{\tau_k}\right)\right] \\ \sigma_G^2 &= \overline{\sin^2\Phi} - (\overline{\sin\Phi})^2 \\ N_{sq} &\triangleq 2\int_{-\infty}^{\infty}\left[R_v(\tau) + \frac{K_V^2}{K^2F_0^2}R_V(\tau)\right]d\tau \end{aligned}\right\} \tag{23}$$

and $I_\nu(x)$ is the modified Bessel function of imaginary order and of argument x. The circular moments of this probability density function are given by (Ref. 3 and SPS 37-56, Vol. III, pp. 104–118)

$$\left.\begin{aligned}\overline{\cos n\Phi} &= \mathrm{RE}\left[\frac{I_{n-j\beta}(\alpha)}{I_{-j\beta}(\alpha)}\right]\\ \overline{\sin n\Phi} &= \mathrm{IM}\left[\frac{I_{n-j\beta}(\alpha)}{I_{-j\beta}(\alpha)}\right]\end{aligned}\right\} \tag{24}$$

where RE[•] and IM[•] denote respectively the "real-part-of" and the "imaginary-part-of" the bracketed quantities. These moments are of interest when the circuits of Figs. 13 and 14 are used to demodulate analog signals.

With $v(t, \Phi)$ as defined in Eq. (11), it is relatively straightforward to show that

$$R_V(\tau) = 4[A^2 R_{n_1}(\tau) + R^2_{n_1}(\tau)] \tag{25}$$

where

$$R_{n_1}(\tau) = \overline{n_1(t)n_1(t+\tau)} \simeq \frac{N_0}{2}\int_{-\infty}^{\infty} |H_i[j2\pi(f-f_0)]|^2 \times \exp[j2\pi(f-f_0)\tau]\, df \tag{26}$$

In the derivation of Eq. (25), it is necessary to have $m(t) = m(t+\tau)$. This assumption is valid for all τ of interest since $T \gg \tau_n$ with τ_n the decorrelation time of the noise. Substitution of Eq. (25) into Eq. (23) gives[7]

$$N_{sq} = 4A^2 N_0 \mathcal{S}_L^{-1} \tag{27}$$

where

$$\mathcal{S}_L \triangleq \left[1 + \frac{2}{A^2 N_0}\left\{\int_{-\infty}^{\infty}\left[\mathrm{R}^2_{n_1}(\tau) + \frac{1}{4}\left(\frac{K_V}{KF_0}\right)^2 R_V(\tau)\right]d\tau\right\}\right]^{-1} \tag{28}$$

is defined to be the circuit "squaring loss."

b. The nth moment of the first passage time. The nth moment of the first passage time is of interest since this parameter is directly related to the nth moment of the time of first loss of phase synchronization. Denoting the nth moment of the first passage time to barrier Φ_l by $\tau^n(\Phi_l)$, and using a result given in Ref. 3, we have

$$\tau^n(\Phi_l) = \frac{4}{N_{sq}K^2F_0^2}\int_{-\Phi_l}^{\Phi_l}\int_{-\Phi_l}^{\Phi}[C_0'(n-1) - \tau^{n-1}(x)] \cdot \exp[U_0(x,\bar{t}) - U_0(\Phi,\bar{t})]\, dx\, d\Phi \tag{29}$$

where $r^0(x) = u(x-\Phi_0)$, $u(x)$ is the unit step, and $\Phi = \Phi_0$ at the initial time, $t = t_0$. The constant $C_0'(n)$ is determined from

$$C_0'(n) = \frac{\displaystyle\int_{-\Phi_l}^{\Phi_l}\tau^n(x)\exp[U_0(x,\bar{t})]\,dx}{\displaystyle\int_{-\Phi_l}^{\Phi_l}\exp[U_0(x,\bar{t})]\,dx} \tag{30}$$

with

$$U_0(x,\bar{t}) = -\frac{4}{N_{sq}K^2F_0^2}\int^{x}\left[2\Omega_0 - A^2KF_0\sin\Phi + \sum_{k=1}^{N}E(y_k,\bar{t}\,|\,\Phi)\right]d\Phi \tag{31}$$

where $E(y_k, \bar{t}\,|\,\Phi)$ is the conditional expectation of y_k at $\bar{t}$ given Φ and $\bar{t}$ is a point such that $\bar{t}\epsilon[0, \infty]$. If the orthogonality principle is used to evaluate $E(y_k, \bar{t}\,|\,\Phi)$, then (Ref. 3)

$$U_0(x,\bar{t}) = \beta x + \alpha\cos x \tag{32}$$

and $\bar{t}$ is taken as ∞.

At this point, our results are extremely general in that they hold for a broad class of loop filters as well as prefilter characteristics, $H_i[j2\pi(f-f_0)]$. In the next subsection, several special cases of practical interest are considered.

4. Tracking Performance of Second-Order Loop; $N = 1, (F_0 = F_1 = \tau_2/\tau_1)$

a. The phase error density $p(\Phi)$. The quantities β and α which characterize the probability distribution of Φ

[7]It is assumed here that $H_i(j\omega)$ *is* normalized such that $H_i(0) = 1$.

can now be related to an equivalent set of system parameters ρ', B_L, and r by

$$\beta = \left(\frac{r+1}{r}\right)^2 \frac{\rho'}{4B_L}\left[2\Omega_0 - A^2K\overline{\sin\Phi}\,(1-F_0) \times\left(1+\frac{F_0}{(r+1)\rho'\sigma_G^2}\right)\right]$$

$$\alpha = \left(\frac{r+1}{r}\right)\rho' - \frac{1-F_0}{r\sigma_G^2} \tag{33}$$

where

r = loop damping coefficient = $A^2KF_1\tau_2$ ($r = 2$ for 0.707 damping; $r = 4$ for critical damping)

B_L = one-sided loop bandwidth = $(r+1)/(4\tau_2)$; $r\tau_1 \gg \tau_2$

ρ' = effective signal-to-noise ratio in loop bandwidth = $(\rho/4)\,\mathcal{S}_L$

ρ = input signal-to-noise ratio in loop bandwidth = $A^2/(N_0B_L)$

b. The mean-time to first slip. The nth moment of the first slip time is found from Eqs. (29) and (30) with $N = 1$ and $\Phi_t = 2\pi$. Making use of the linear tracking theory to establish a relationship for $E(y_1, \bar{t} \mid \Phi)$ (Ref. 3), Eq. (31) becomes[8]

$$U_0(\Phi, \bar{t}) \approx -\left(\frac{r+1}{r}\right)\rho'\cos\Phi - \frac{\rho'}{2r}\Phi^2 - \frac{2\rho'\Omega_0}{A^2K}\Phi \tag{35}$$

and

$$W_L\tau^n(2\pi) \approx \left(\frac{r+1}{r}\right)^2 \frac{\rho'}{2}\int_{-2\pi}^{2\pi}\int_{-2\pi}^{\Phi}\left[C_0'(n-1) - \tau^{(n-1)}(x)\right] \times \exp\left[\rho'\left(\frac{r+1}{r}\right)[\cos\Phi - \cos x] + \frac{\rho'}{2r}(\Phi^2 - x^2) + \frac{2\rho'\Omega_0}{A^2K}(\Phi - x)\right] dx\, d\Phi \tag{36}$$

[8]As discussed in Ref. 3, the linear PLL theory with $t = \infty$ appears to give a better estimate of $E(y_1, t \mid \Phi)$ than the estimate obtained using the orthogonality principle which produced Eq. (31). This fact has been justified on the basis of experimental work (Ref. 3).

c. Evaluation of the squaring loss for various prefiltering characteristics $H_i(s)$. As a first example, consider a bandpass filter with a resistance–capacitance (RC) transfer function. Then the equivalent low-pass spectrum for $n_1(t)$ or $n_2(t)$ has a correlation function:

$$R_{n_1}(\tau) = R_{n_2}(\tau) = \frac{N_0}{4}(2\pi B_i)\exp(-2\pi B_i \mid \tau \mid) \tag{37}$$

where $B_i = 1/(2\pi RC)$ Hz/s is the 3-dB frequency of the filter. If we neglect the effect of $n_V(t)$, then from Eqs. (37) and (28),

$$\mathcal{S}_L = \frac{1}{1+\frac{\pi}{4\rho\gamma}} \tag{38}$$

where

$$\gamma = \frac{B_L}{B_i} \tag{39}$$

For an ideal bandpass filter (i.e., $H_i(j\omega) = 1$; $|f| < B_i$, and zero otherwise), then

$$\mathcal{S}_L = \frac{1}{1+\frac{1}{\rho\gamma}} \tag{40}$$

We note that these are the two limiting cases of the Butterworth spectra (Ref. 2). Various other examples of prefiltering characteristics can easily be carried out. (See Table 1 for a partial list.)

5. Receiver Performance for Digital Modulation

The output of the squaring loop [i.e., $r(t)$] is frequency divided by two to provide a noisy reference for demodulation of the data off the carrier. Thus, the effective phase error for purposes of computing the error probability performance of the data detector is $\phi = \Phi/2$. By simple transformation of variables, the density function for ϕ becomes

$$q(\phi) = 2p(\Phi)\Big|_{\Phi=2\phi}, \qquad |\phi| \leq \frac{\pi}{2} \tag{41}$$

Table 1. Squaring loss for various prefiltering characteristics, $\mathcal{S}_L = [1 + K_L/\rho\gamma]^{-1}$

Prefilter type	Equivalent low-pass transfer characteristic* $\lvert H(j\omega_m)\rvert^2$	K_L
nth order Butterworth	$\dfrac{1}{1+\left(\dfrac{\omega_m}{\omega_i}\right)2n}$	$\dfrac{\Gamma\left(\frac{1}{2n}\right)\Gamma\left(2-\frac{1}{2n}\right)}{2n}$
Gaussian	$\exp\left[-2\left(\dfrac{\omega_m}{\omega_i}\right)^2\right]$	$\dfrac{\pi^{1/2}}{4}$
Sinusoidal roll-off $(0 \le \xi \le 1)$	$\frac{1}{4}\left[1-\sin\frac{\pi}{2}\left(\frac{\omega_m-\omega_i}{\xi\omega_i}\right)\right]^2$ for $\lvert\omega_m-\omega_i\rvert \le \xi\omega_i$; 0 for $\omega_m-\omega_i \ge \xi\omega_i$; 1 for $-\omega_i \le \omega_m-\omega_i \le -\xi\omega_i$	$1-\frac{37}{64}\xi$

* $\omega_m = 2\pi(f-f_0)$, $\omega_i \triangleq 2\pi\delta_i$

If $P_e(\phi)$ is the conditional probability that the data detector will commit a bit (or word) error, then the average bit (or word) error probability is

$$P_e = \int_{-\pi/2}^{\pi/2} P_e(\phi)q(\phi)\,d\phi = \int_{-\pi}^{\pi} P_e\left(\frac{\Phi}{2}\right)p(\Phi)\,d\Phi \qquad (42)$$

For a correlation type receiver with perfect word and bit sync and an orthogonally coded transmitted signal set, the word error probability is given by (Ref. 5)

$$P_{e_W}(\phi) = 1 - P_{c_W}(\phi) = 1 - \int_{-\infty}^{\infty} \frac{e^{-z^2}}{\pi^{1/2}}\left\{\frac{1}{2}\operatorname{erfc}\left[-z-(nR)^{1/2}\cos\phi\right]\right\}^{2^n-1}dz \qquad (43)$$

where n is the number of symbols (binits) per code word, $R = A^2T_b/N_0$, where T_b is the bit time, and erfc (y) is the complementary error function defined by

$$\operatorname{erfc}(y) = \frac{2}{\pi^{1/2}}\int_y^{\infty} e^{-t^2}\,dt \qquad (44)$$

Note that R can be related to ρ by $\rho = R\delta$, where $\delta = 2/W_LT_b$. The corresponding conditional bit error probability, $P_{e_B}(\phi)$, for a ± 1 transmitted signal set is given by

$$P_{e_B}(\phi) = 1 - \int_{-\infty}^{\infty} \frac{e^{-z^2}}{\pi^{1/2}}\left\{\frac{1}{2}\operatorname{erfc}\left[-z-(2R)^{1/2}\cos\phi\right]\right\}dz = \frac{1}{2}\operatorname{erfc}\left(R^{1/2}\cos\phi\right) \qquad (45)$$

Also note that $P_{e_B}(\phi)$ is not obtained from $P_{e_W}(\phi)$ for orthogonal codes by letting $n = 1$. This is so since in an orthogonally coded system the words are 90 deg apart whereas in a one-bit system (± 1) they are separated by 180 deg.

Figures 15 and 16 plot P_{e_B} as defined by Eqs. (45) and (42) vs Ω_0/AK with ρ' and R_{RC} or R_{ideal} as parameters for $F_0 = 0.002$, $\delta = 25$, $\gamma = 0.002$, and $r = 2$ (typical of Deep Space Network receivers). These curves can also be used for other combinations of values γ and δ. To determine

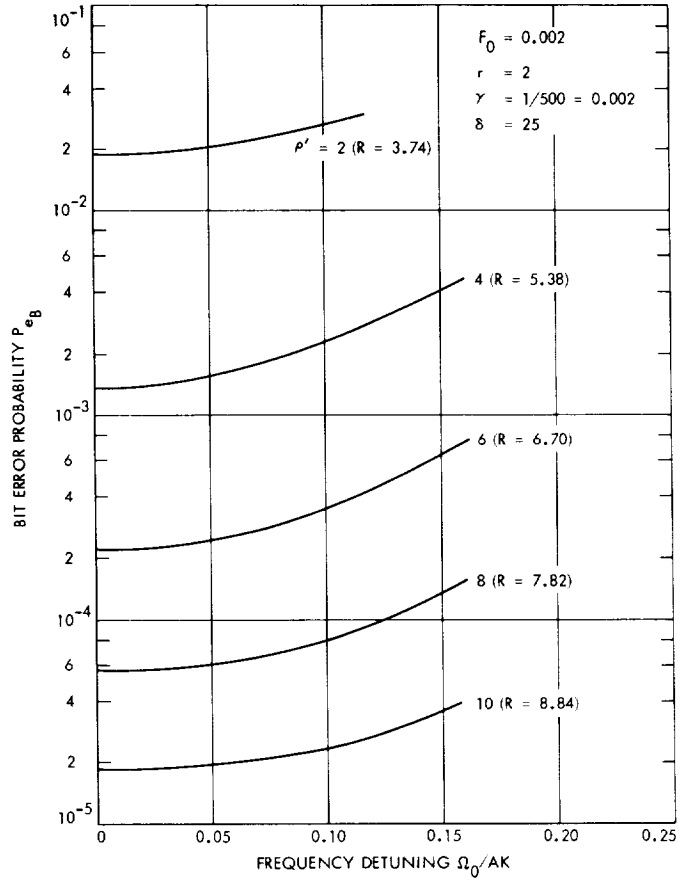


Fig. 15. Bit error probability as a function of frequency detuning for ideal bandpass filter

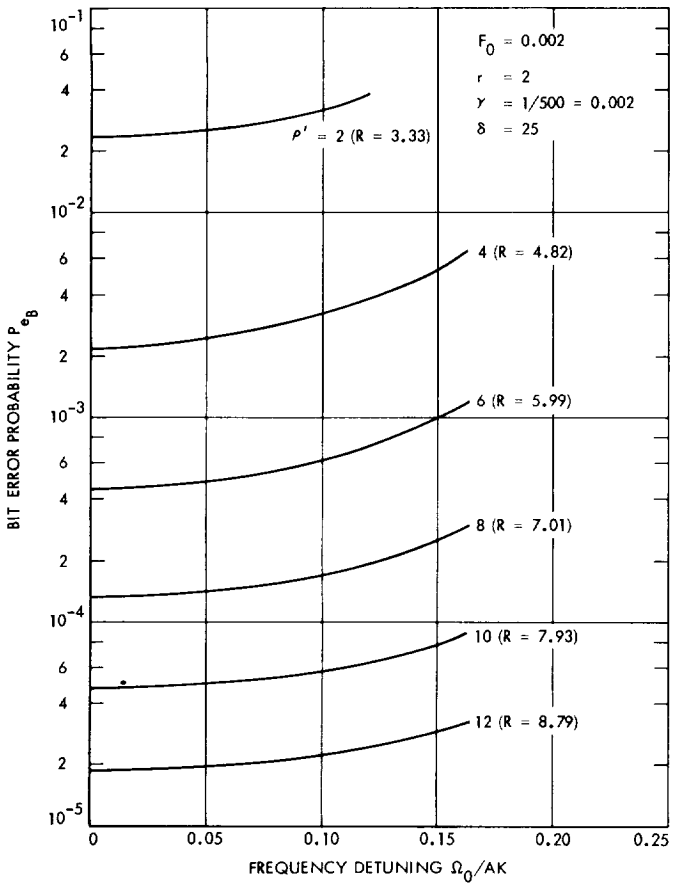


Fig. 16. Bit error probability as a function of frequency detuning for single-pole RC bandpass filter

these, Fig. 17 plots $R\delta$ vs ρ' with γ as a parameter for ideal and RC bandpass filters.

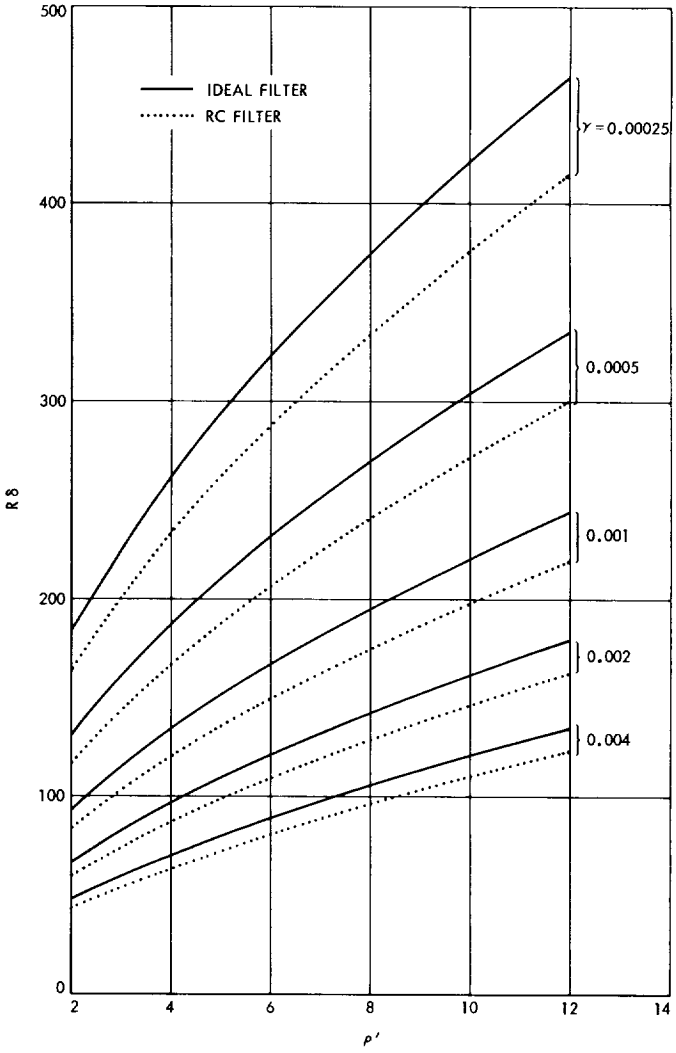


Fig. 17. Input signal-to-noise ratio in loop bandwidth as a function of effective signal-to-noise ratio in loop bandwidth for ideal and RC bandpass filters

6. Concluding Remarks

This article has presented results that can be applied to the problem of design and planning of coherent communication systems which transmit digital or analog data as double-sideband suppressed carrier signals. The demodulation of digital data has been discussed in detail. Coherent demodulation of analog signals can be treated using the results presented here in combination with those given in Ref. 2.

References

1. Didday, R. L., and Lindsey, W. C., "Subcarrier Tracking Methods and Communication System Design," *IEEE Trans. Commun. Technol.*, Vol. COM-16, No. 4, pp. 541–550, Aug. 1968.

2. Lindsey, W. C., "Optimum Coherent Linear Demodulation," *IEEE Trans. Commun. Technol.*, Vol. COM-13, No. 2, June, 1965.

3. Lindsey, W. C., "Nonlinear Analysis of Generalized Tracking Systems," *Proc. IEEE*, Oct. 1969. Also USC EE 317, University of Southern California, Los Angeles, Calif., Dec. 1968.

4. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*. McGraw-Hill Book Co., Inc., New York, 1958.

5. Viterbi, A. J., *Principles of Coherent Communication*. McGraw-Hill Book Co., Inc., New York, 1966.

6. Stratonovitch, R. L., *Topics in the Theory of Random Noise*, Vol. 1, pp. 83–103. Gordon and Breach, Inc., New York, 1963.

IX. Flight Computers and Sequencers

GUIDANCE AND CONTROL DIVISION

A. Reliability Study of Fault-Tolerant Computers,

F. P. Mathur

1. Introduction

The general objective of the reliability study is to investigate the problem of estimating and quantifying the reliability of fault-tolerant computer configurations. Present-day interplanetary mission requirements of aerospace digital computers designed to perform in a space environment is in the order of ten or more years.[1] The success of a whole mission is often dependent on the success of the onboard digital computer. Several design approaches employed to satisfy stringent reliability requirements for assuring error-free computation in aerospace computers include the selection of highly reliable components, the use of proven techniques of interconnection and packaging, extensive simulation to verify the logic design, and both functional and diagnostic testing.

This study addresses itself to the following three interrelated tasks:

(1) The development of reliability equations that model the reliability of fault-tolerant organizations.

(2) The analysis, with respect to significant parameters, of various redundancy techniques utilizable in the design of such ultra-reliable computers.

(3) To critically compare the various redundancy techniques available and determine figures of merit and guidelines for their optimum usage.

These tasks lead to the development of a mathematical reliability model of generalized self-testing and repairing computers that will permit the analytic evaluation of the significant reliability parameters. Concurrent with this analytic effort, a Computer Aided Reliability Estimation (CARE) software program is being written. The CARE program is being developed on the UNIVAC 1108 multiprocessor system; its purpose is to serve as a computer-aided reliability design tool to designers of ultra-reliable systems.

[1]Avizienis, A.A., F. Mathur, D. Rennels, J. Rohr, "Automatic Maintenance of Aerospace Computers and Spacecraft Information and Control Systems," to be presented at the *AIAA Aerospace Computer Systems Conference*, Los Angeles, Sept 8-10, 1969.

2. Methodology

This reliability study complements the self-testing and repairing fault-tolerant computers design effort. The task of designing for fault-tolerance requires taking into account all anticipated classes of faults. The task of reliability analysis is to estimate and predict quantitatively the reliability (i.e., the probability of survival for a given mission) of the resulting design and, thus, to give a measure of the effectiveness of the fault-tolerance capabilities. (In general, the procedure of designing a fault-tolerant configuration followed by its reliability evaluation is not a two-step procedure but an iterative one.) The insight obtained by the formulation of the mathematical model and the derivation of its representative reliability equation followed by the analysis of the behavior of the reliability as a function of the various variables of design leads to refinements in the original design and suggests new approaches to the problem of designing for ultra reliability.

3. Reliability Parameters

The significant reliability parameters besides the probability of surviving for the length of the mission are the mean life of the system, the reliability at the mean life, the maximum mission duration of a system at a given reliability, and the reliability gain, which may be with respect to either the nonredundant design or competitive designs. These reliability parameters are evaluated under the simplifying assumption that the underlying failure-law is exponential; the exponential failure-law is justifiable on the basis of equipment complexity and the high degree of replication and replacement utilized (Ref. 1). The exponential distribution indicates that the failure rates are constant for a specific mode of usage, although different failure rates apply depending on whether the units are active, dormant, or inert. A spare unit is said to be *active* if its failure rate (λ) is identical to the failure rate of the powered unit, *inert* if the failure rate of the unit is zero, and *dormant* if its failure rate (μ) lies between the active and inert values.

When the standby units are considered active, the reliability estimation yields a conservative estimate and gives a lower bound on the survival probability. The assumption of inert standby units yields an optimistic estimate and provides an absolute upper bound of reliability. In actuality, the failure rates lie between these two limits and the failure rates of the redundant standby units are expressed in terms of *dormancy factors* (δ). The dormancy factor is expressed as a percentage and is defined as the ratio of the failure rate of the standby unit to the failure rate of the active unit multiplied by one hundred $[(\mu/\lambda) \times 100]$. Thus, $\delta = 100\%$ implies that the standby unit is active, while $\delta = 0\%$ states that the standby unit is inert. Very little published statistical data on dormancy factors are available from electronic component manufacturers. Some currently quoted figures lie between 10 and 20% for electronic equipment. This scarcity of accurate statistical data on parameters such as failure rates and dormancy factors limits absolute reliability estimates, but does not affect the relative reliability comparison of competitive redundancy configurations using identical technologies. The improvement, or gain, in reliability due to application of protective redundancy to an initially nonredundant design and related parameters may be estimated without recourse to statistical data.

4. Reliability Models

The well-known reliability concepts and models of self-repair were looked into and their applicability to the modeling of the STAR computer were investigated. Specifically, the triple-modular redundancy (TMR) class of systems were investigated in depth and various new results pertaining to the asymptotic behavior of such systems under limiting or boundary conditions have been established. Replacement systems using selective or dynamic redundancy have also been investigated and a comparison between these and the TMR form of massive or static redundancy is being made. A hybrid configuration of the two systems was developed using the TMR concept along with the spares as standby replacements. The mathematical model of the TMR/Spare system has been derived along with the representative reliability equations and expressions for mean time to failure. A first-cut reliability model and reliability estimate of the STAR computer were also obtained.

5. The TMR/Spare System

The basic TMR/Spare concept is shown in Fig. 1. The simplex unit is triplicated and its outputs voted upon (in general, one of many restoring organs may be used) to yield the system output. A bank of S dormant spare units are provided such that when one of the TMR units fails, it is replaced by a spare unit. The operating units have a failure rate of λ while the spare units have a failure rate of μ. When a spare replaces a failed operating unit, it experiences a sudden change in failure rate from μ to λ. The characteristic reliability equation describing the

behavior of such a system was derived and proved to be as follows:

$$R(\text{TMR/Spare}) = 3R^2 \left\{ \prod_{i=0}^{S-1} \left(\frac{3K+S-i}{K+S-i} \right) - \frac{2RK^2}{S!} \left[\prod_{i=0}^{S-1} (3K+S-i) \right] \sum_{j=0}^{S} \binom{S}{j} \frac{(-1)^{S-j} R_s^{S-j}}{(K+S-j)(3K+S-j)} \right\}$$

where

R = reliability in operating mode

S = number of spares for each functional unit

R_s = reliability of unit in spare mode

$R(T) = \exp(-\lambda T)$

$R_s(T) = \exp(-\mu T)$

$K = \lambda/\mu$

T = time

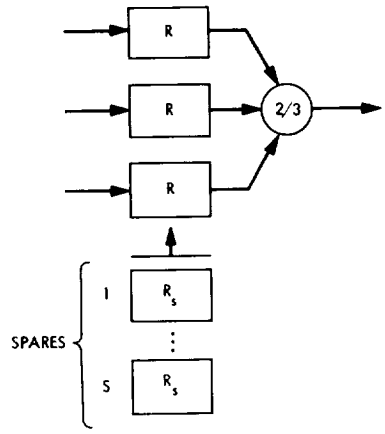


Fig. 1. TMR/Spare system concept

The behavior of this equation is illustrated in Fig. 2 where the reliability curves are compared and contrasted against both that of a simplex unit and those of generalized TMR systems. For simplicity of comparison, the unreliability of the voting units was not taken into account in the examples shown. The generalized TMR system, abbreviated as the NMR system, utilizes $2n+1$ units all of whose outputs are majority voted upon to produce the system output. The significant reliability and cost advantage arises from the fact that the NMR system can only tolerate n faults where the TMR/Spare system for the same number of total units can tolerate almost twice as many faults. The reliability analysis of this TMR/Spare system is totally new and these analytic results were informally presented at a recent workshop (Ref. 2).

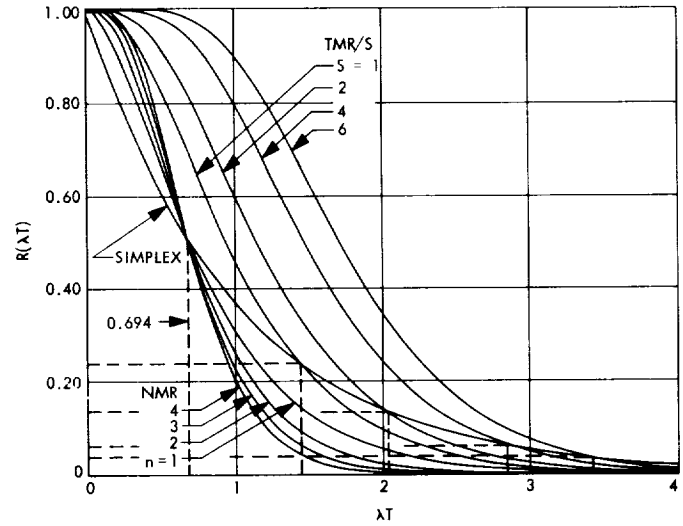


Fig. 2. Reliability comparison of a TMR/Spare system vs normalized time with and without spares

6. The STAR Computer

The STAR computer (Ref. 3) block diagram is shown in Fig. 3. The STAR computer consists of a number of functional units [i.e., the main arithmetic processor (MAP), the control processor (COP), etc.] the spares of which are all permanently connected to the bus lines. Diagnosis is achieved by coding all information. All data and instruction words are coded by means of arithmetic codes. Repair is achieved by switching off power to the faulty unit and energizing a spare. The functions of testing and repairing reside in the test and repair pro-

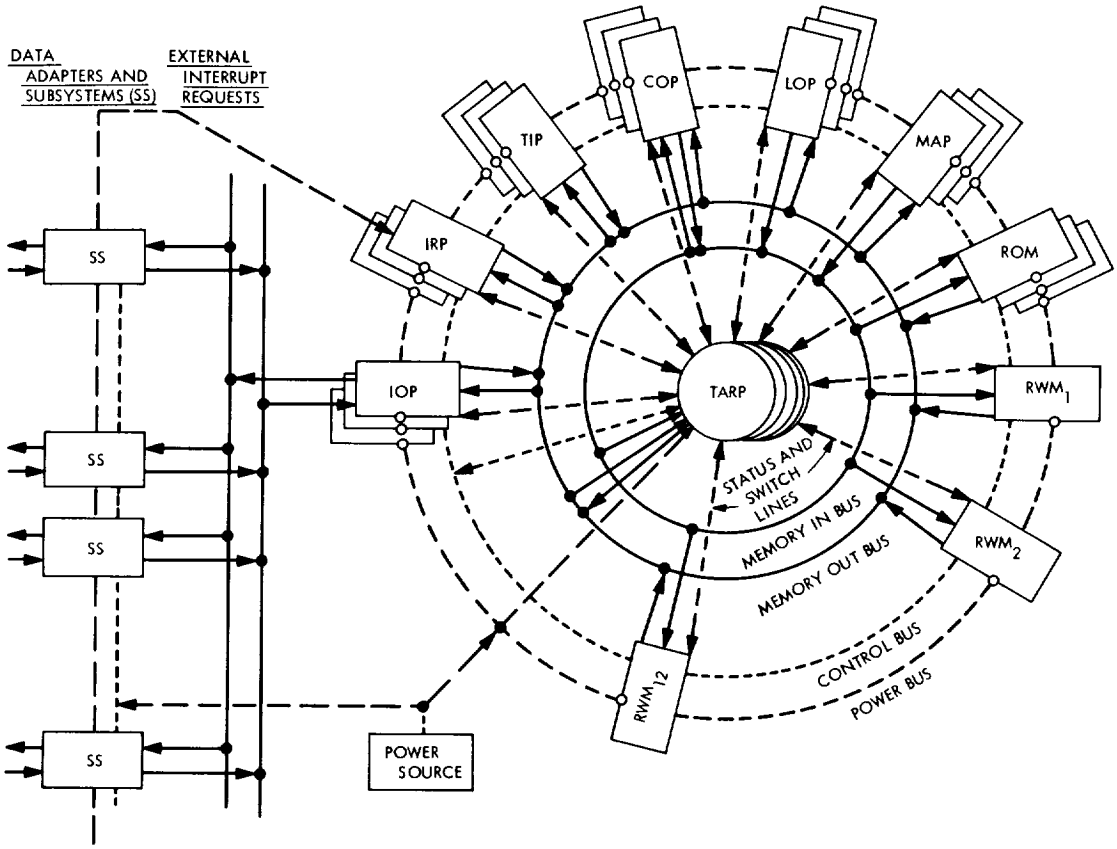


Fig. 3. STAR computer block diagram

cessor (TARP). The implementation of the TMR/Spare model mentioned in *Subsection 5* is being studied to the application of the TARP unit.

A first-cut reliability estimate of the STAR computer was derived based on the following reliability model. The functional processors comprising the STAR computer are considered to be in series reliability. The TARP unit, which is the hard-core of the system, is in majority triplicated configuration with a number of spares.

The upper and lower bounds on the reliability of the STAR computer were obtained relative to the reliability of the *Mariner* Mars 1969 computer, which was used as a basis for comparing the STAR computer reliability under various simplifying assumptions. While comparing the reliabilities, it must be emphasized that these differing systems have different computational capabilities and functional characteristics. In contrast with the STAR computer, the computational capabilities of the *Mariner* Mars 1969 computer are exceedingly limited, since the *Mariner* Mars 1969 computer is a bit-serial machine with a bit rate of 2-4 kHz and an instruction set consisting of sixteen ok-codes, whereas the STAR is a byte-serial machine with a 1-MHz clock and an instruction set of approximately eighty ok-codes of which approximately half are indexable.

The mean time between failures (MTBF) for the complete *Mariner* Mars 1969 computer was evaluated by others to be 58,600 h. This calculation was based on the piece parts and components failure-rate data obtained mainly from MIL-HDBK-217A (Dec. 1, 1965). Using this MTBF and assuming an exponential failure-law, the survival probability of the *Mariner* Mars 1969 computer was obtained for various mission times. Knowing the reliability of the *Mariner* Mars 1969 computer, and by postulating comparative complexities in equipment, a rough estimate of the STAR computer reliability can be derived.

The reliability models of (1) the *Mariner* Mars 1969 computer, (2) a simplex computer having the same computational capability as the STAR computer, and (3) the STAR computer are shown in Fig. 4. The *Mariner* Mars

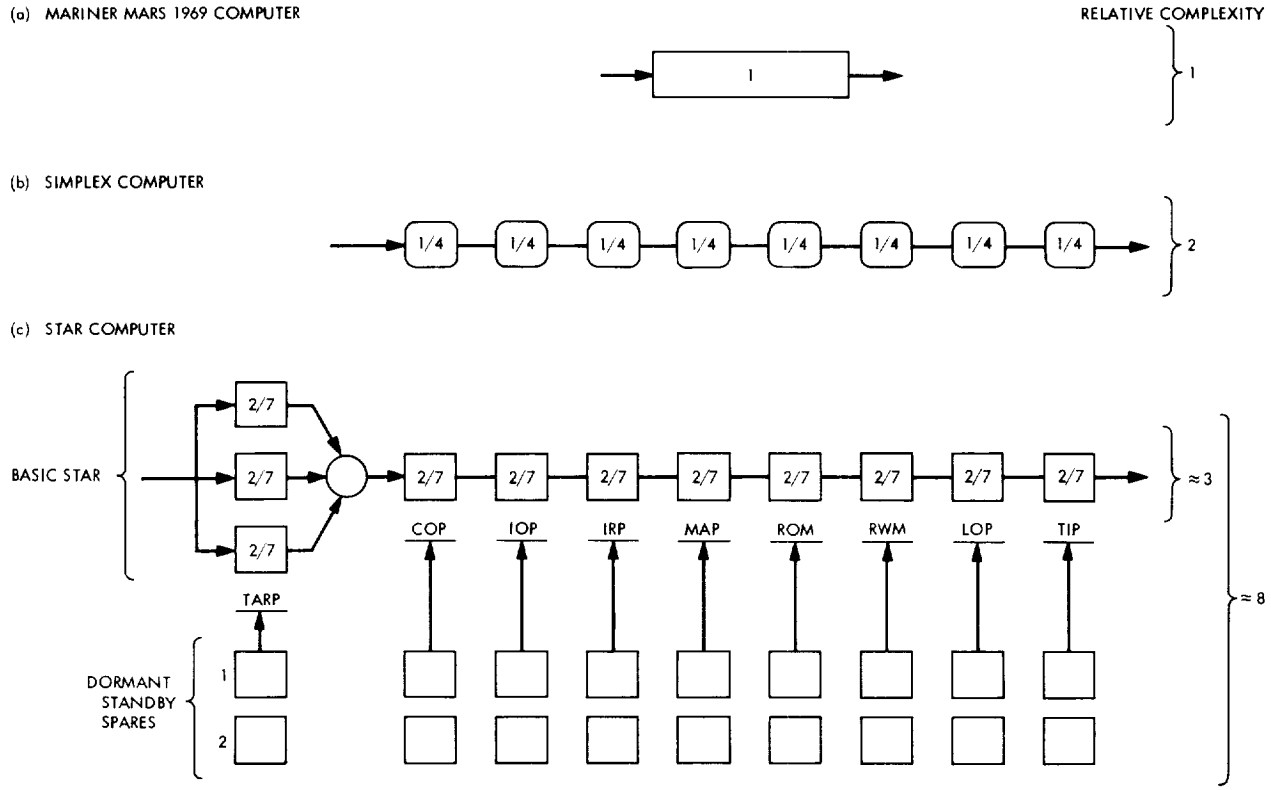


Fig. 4. Computer reliability models

1969 computer (Fig. 4a) is assigned a complexity of unity. It is then assumed that the simplex computer (Fig. 4b) is twice as complex as the *Mariner* Mars 1969 computer and consists of eight uniform functional units; thus, each of the simplex functional units is one-fourth as complex as the *Mariner* Mars 1969 computer. The relative complexity of 2/7 for the STAR functional units (Fig. 4c) is then arrived at by applying a factor of 8/7 to each of the simplex functional units. This factor takes into account the overhead due to the self-testing and repairing features within each functional unit of the STAR computer (e.g., 1 byte of the 8 bytes in a STAR word is dedicated to error detection). Also, as compared to the simplex organization, the STAR computer requires the TARP. The TARP unit is made triple-modular redundant and is allocated a number of spares (a TMR/Spare configuration). The other STAR units are made selectively redundant by the use of dormant standby spares.

The bounds on the reliability of the STAR computer are evaluated for an allocation of two spares for each functional unit as well as for an allocation of three spares. Some typical numeric values of reliability versus mission time for the various on-board configurations are shown in Table 1. Results shown in Table 1 assume a complexity factor (CF) of 1/4. The CF is here defined as the ratio of complexity of a single STAR functional unit to the complexity of the complete *Mariner* Mars 1969 computer. Table 2 presents similar results with CF = 1/3. Tables 3 and 4 show some typical numeric values of the gain in reliability for CF = 1/4 and 1/3, respectively. The gain in reliability is here defined as the reliability of the STAR computer divided by the reliability of the *Mariner* Mars 1969 computer.

Tables 5 and 6 show some typical values of maximum mission duration versus the desired mission reliability for CF = 1/4 and 1/3, respectively. For example, if a mission reliability of, say, 0.8 is desired, then Table 5 states that the maximum mission duration that may be planned using the *Mariner* Mars 1969 computer is 1.5 yr, 0.7 yr using the simplex computer, and between 8 and 13 years using the STAR computer with two spares. A computer program, written to evaluate and plot these results, generated the plots of reliability versus mission time and reliability gain versus mission time shown in Figs. 5–8.

Table 1. Reliability versus mission time for various configuration (CF = ¼)

Planned mission time, h	*Mariner* Mars 1969 complete computer	Simplex computer	STAR computer with S spares (STF[a] = 8/7)			
			Upper bound[b]		Lower bound[b]	
			S = 3	S = 2	S = 3	S = 2
4368 (≈6 mo)	0.928	0.861	0.999999927	0.999987	0.9999984	0.999924
43,680.0 (≈5.0 yr)	0.475	0.225	0.9991	0.986	0.988	0.94
87,360.0 (≈10.0 yr)	0.225	0.051	0.985	0.899	0.871	0.674

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.
[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

Table 2. Reliability versus mission time for various configurations (CF = ⅓)

Planned mission time, h	*Mariner* Mars 1969 complete computer	Simplex computer	STAR computer with S spares (STF[a] = 8/7)			
			Upper bound[b]		Lower bound[b]	
			S = 3	S = 2	S = 3	S = 2
4368 (≈6 mo)	0.928	0.82	0.9999998	0.99997	0.999995	0.99982
43,680.0 (≈5.0 yr)	0.475	0.14	0.997	0.97	0.966	0.87
87,360.0 (≈10.0 yr)	0.225	0.19	0.96	0.79	0.71	0.45

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.
[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

Table 3. Typical gain in reliability numeric values (CF = 1/4)

Planned mission time, h	STAR computer with S spares (STF[a] = 8/7)			
	Upper bound[b]		Lower bound[b]	
	S = 3	S = 2	S = 3	S = 2
4368 (≈6 mo)	1.077388	1.07737	1.077386	1.07731
43680.0 (≈5.0 yr)	2.105	2.078	2.082	1.98
87360.0 (≈10.0 yr)	4.375	3.996	3.87	2.993

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.
[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

Table 4. Typical gain in reliability numeric values (CF = 1/3)

Planned mission time, h	STAR computer with S spares (STF[a] = 8/7)			
	Upper bound[b]		Lower bound[b]	
	S = 3	S = 2	S = 3	S = 2
4368 (≈6 mo)	1.077387	1.07735	1.077382	1.07719
43680.0 (≈5.0 yr)	2.101	2.039	2.036	1.84
87360.0 (≈10.0 yr)	4.25	3.51	3.14	1.993

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.
[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

Table 5. Typical values of maximum mission duration versus reliability (CF = 1/4)

Desired mission reliability	Maximum mission duration, yr					
	Mariner Mars 1969 complete computer	Simplex computer	STAR computer with S spares (STF[a] = 8/7)			
			Upper bound[b]		Lower bound[b]	
			S = 3	S = 2	S = 3	S = 2
0.9	0.7	0.4	17.0	10.0	9.0	6.0
0.8	1.5	0.7	21.0	13.0	11.5	8.0
0.7	2.4	1.2	24.5	15.5	13.5	9.5
0.6	3.5	1.8	27.5	18.0	15.0	11.0

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.

[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

Table 6. Typical values of maximum mission duration versus reliability (CF = 1/3)

Desired mission reliability	Maximum mission duration, yr					
	Mariner Mars 1969 complete computer	Simplex computer	STAR computer with S spares (STF[a] = 8/7)			
			Upper bound[b]		Lower bound[b]	
			S = 3	S = 2	S = 3	S = 2
0.9	0.7	0.3	12.5	7.5	6.7	4.5
0.8	1.5	0.6	16.0	9.7	8.5	6.0
0.7	2.4	0.9	18.5	11.7	10.0	7.0
0.6	3.5	1.3	20.5	13.5	11.3	8.3

[a]STF = ratio of complexity of the basic STAR with zero spares to an equivalent (with respect to computational capability) nonself-testing and repairing (simplex) computer.

[b]The upper bound on reliability is computed by considering the failure rate of the spare units to be zero. The lower bound is computed by considering the failure rate of the spare units to be identical with the powered units.

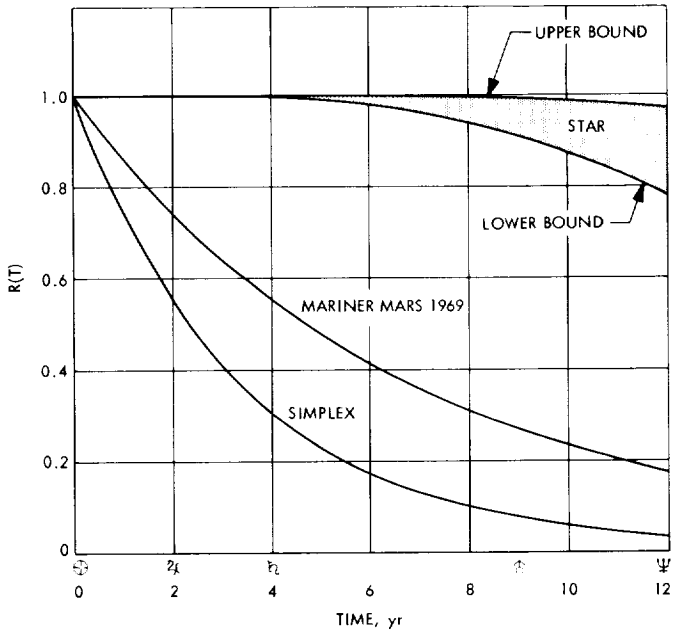


Fig. 5. Upper and lower bounds on the STAR computer reliability with three spares (CF = 1/4)

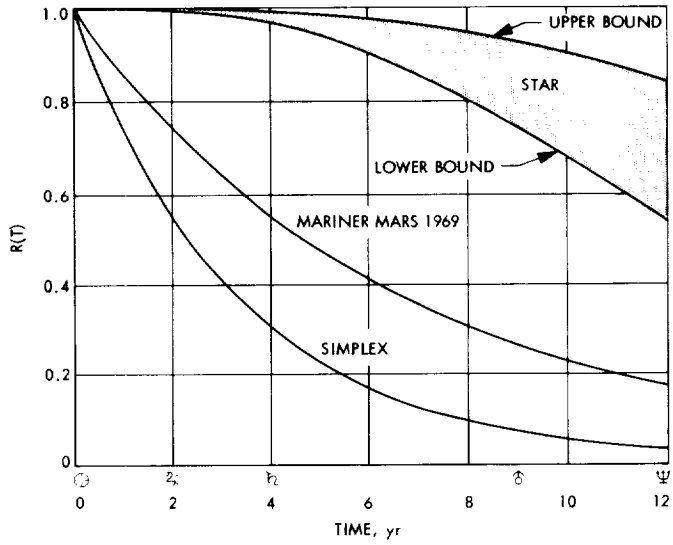


Fig. 6. Upper and lower bounds on the STAR computer reliability with two spares (CF = 1/4)

7. CARE Program

The CARE program is a software design-tool used to facilitate reliability analysis; it may be interactively accessed by a designer from a teletype console to calculate his reliability problem in "real time." The input is in the form of system configuration description followed by queries on the various reliability parameters of interest and their behavior with respect to mission time, fault-coverage, failure rates, dormancy factors, allocated spares, and cost in hardware. The CARE subroutines are periodically updated to reflect refinements made on existing reliability models and enlarged as further data becomes available.

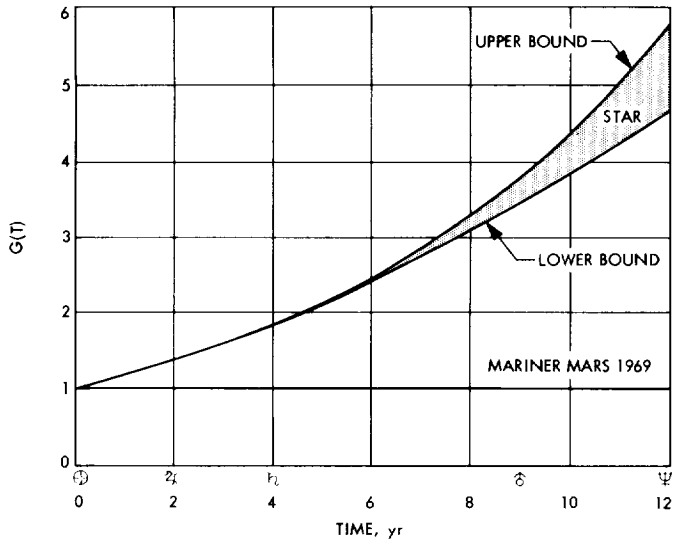


Fig. 7. Upper and lower bounds on reliability gain with three spares (CF = ¼)

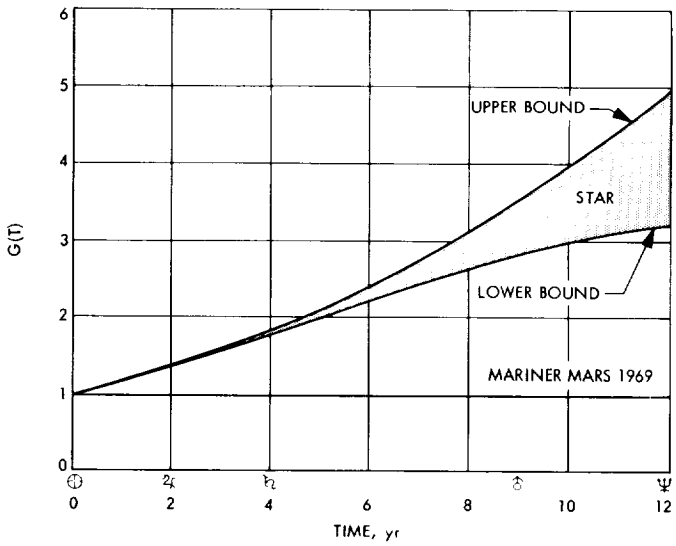


Fig. 8. Upper and lower bounds on reliability gain with two spares (CF = ¼)

8. Conclusion

The large number of interrelated tasks described herein are all being worked on concurrently. It is anticipated that the effort on developing new mathematical models and equations and refining the well-known ones will continue, along with the task of making critical comparisons and developing figures of merit. The effort on the CARE program has just been initiated and a considerable amount of work has yet to be done before it becomes fully operational. Further work will also be done in estimating the reliability of the STAR computer with greater exactitude.

References

1. Drenick, R. F., "The Failure Law of Complex Equipment," *J. Soc. Ind. Appl. Math.*, Vol. 8, No. 4, pp. 680–690, Dec. 1960.

2. Mathur, F. P., *Design and reliability analysis of "hard cores" of fault-tolerant computers*, Paper presented at the ACM Workshop on Hardware-Software Interaction for System Reliability and Recovery in Fault-Tolerant Computers, Pacific Palisades, Calif., July 14-15, 1969.

3. Avizienis, A. A., *An Experimental Self-repairing Computer*, Technical Report 32-1356. Jet Propulsion Laboratory, Pasadena, Calif., Aug. 1968.

X. Guidance and Control Analysis and Integration

GUIDANCE AND CONTROL DIVISION

A. Support Equipment for a Strapdown Navigator, *R. E. Williamson*

1. Objectives

The objectives of this effort are to provide a test van, data acquisition equipment, prime power, test support, and system engineering for a Strapdown Electrostatic Gyro Aerospace Navigator (SEAN). The SEAN System and support equipment are shown in Figs. 1 and 2.

2. Progress

a. Van system. The van will be used for over-the-road testing of the SEAN System to provide a dynamic environment prior to flight testing. The van has been procured, road-tested to determine vibration levels and shock intensities, and modified in preparation for equipment installation. The modifications include: wiring the van for 60- and 400-Hz power; installation of air conditioning, lighting, seating, and intercom; and fabrication of structures for mounting the equipment in the van.

b. Power system. Power for road operations will be provided by trailer-mounted engine generators. A 30-kV-A, 400-Hz generator will be used for SEAN System power and two 5-kV-A, 60-Hz generators for air conditioning power.

A noninterruptable power conversion unit, operating from a 400-Hz source and a battery pack will provide constant, transient-free, 400-Hz power to the SEAN System. The unit, while operating from the battery pack alone, will provide system power during the transfer from facility power to the engine generators, and will also provide power for up to 20 min in the event of failure of the primary 400-Hz power.

The power system, including generators, trailer, power conversion unit, battery pack, and associated equipment, has been completed and was integrated with the SEAN System in April 1969.

c. Data acquisition. The data acquisition system will provide the capabilities required to evaluate and troubleshoot the SEAN System.

Digital recording system. A digital magnetic-tape recording system will be provided to record computer system outputs (position, velocity, and navigation time),

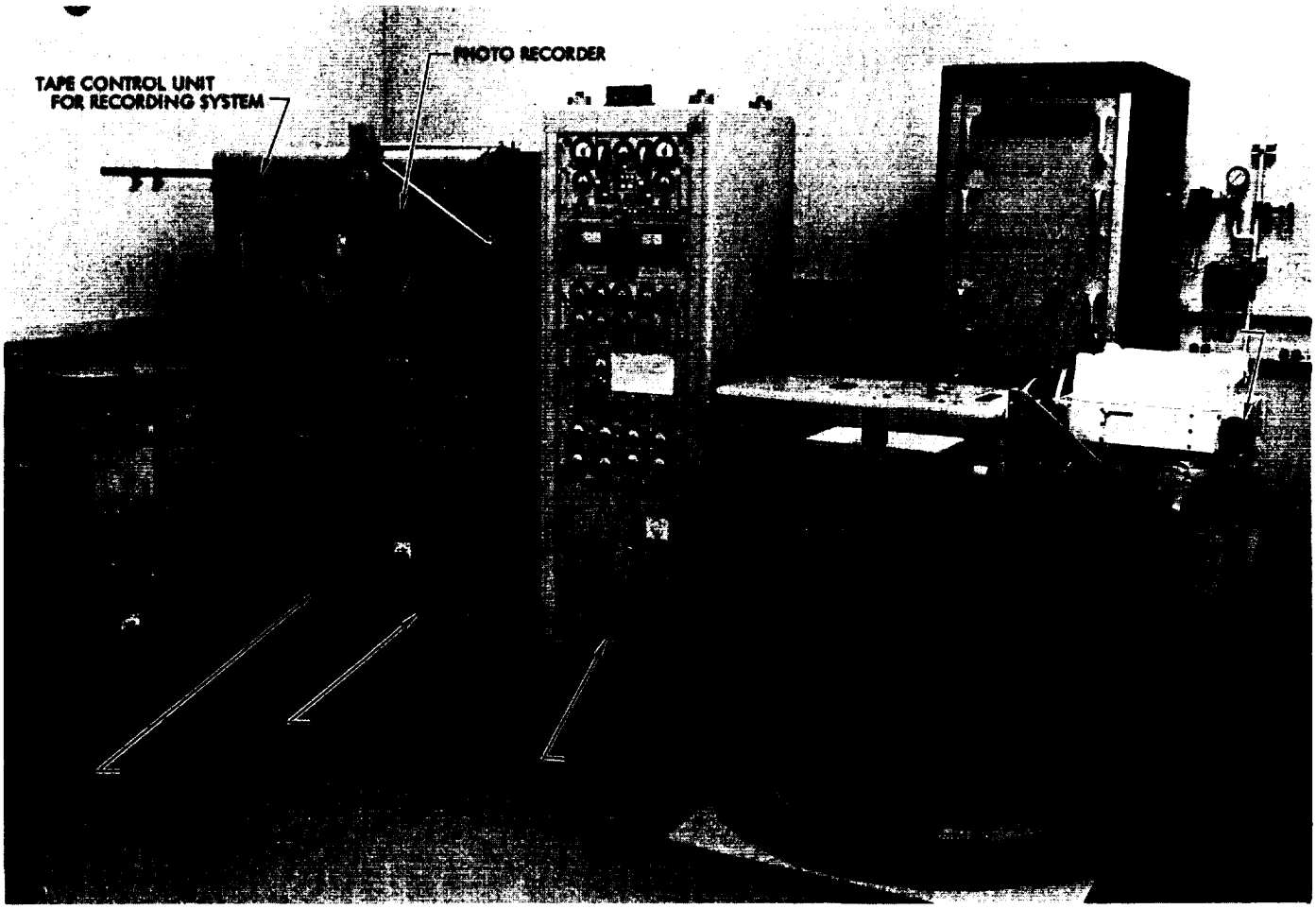


Fig. 1. SEAN System and associated equipment

total inertial measurement unit (IMU) data (gyro and velocity meter counts), and other parameters as required.

The recording system consists of two flight-environment-qualified digital magnetic-tape recorders and a tape control unit. The control unit interfaces with the SEAN digital computer memory where it obtains the system data and formats it for recording on the tapes in an IBM-compatible format.

The recording system has been completed and integrated with the SEAN computer and is presently being used to support laboratory navigation tests.

Oscillographic recording system. An oscillographic recording system will be provided to monitor analog functions such as gyro suspension "g" loading, IMU temperatures, gyro status, and battery and power supply voltages.

A Brush, Model Mark 848, eight-channel thermal-writing recorder has been purchased, and a Mid-Western, Model 602, 36-channel optical-writing recorder has been obtained from the JPL loan pool. The two recorders and associated instrumentation have been mounted in one console and are being used to support SEAN laboratory testing.

Photo recorder. A 35 mm, remotely controlled, photo recorder designed for aircraft use is being provided to photographically record the computer adapter display output information (latitude, longitude, speed, azimuth, and time). The camera is controlled by the SEAN computer to photograph the display panel every 5 s. The camera and its control unit were integrated with the SEAN System in March 1969.

d. Altimeter. A CPU-46A altitude computer, Bendix Type 31101, has been procured to provide digitized

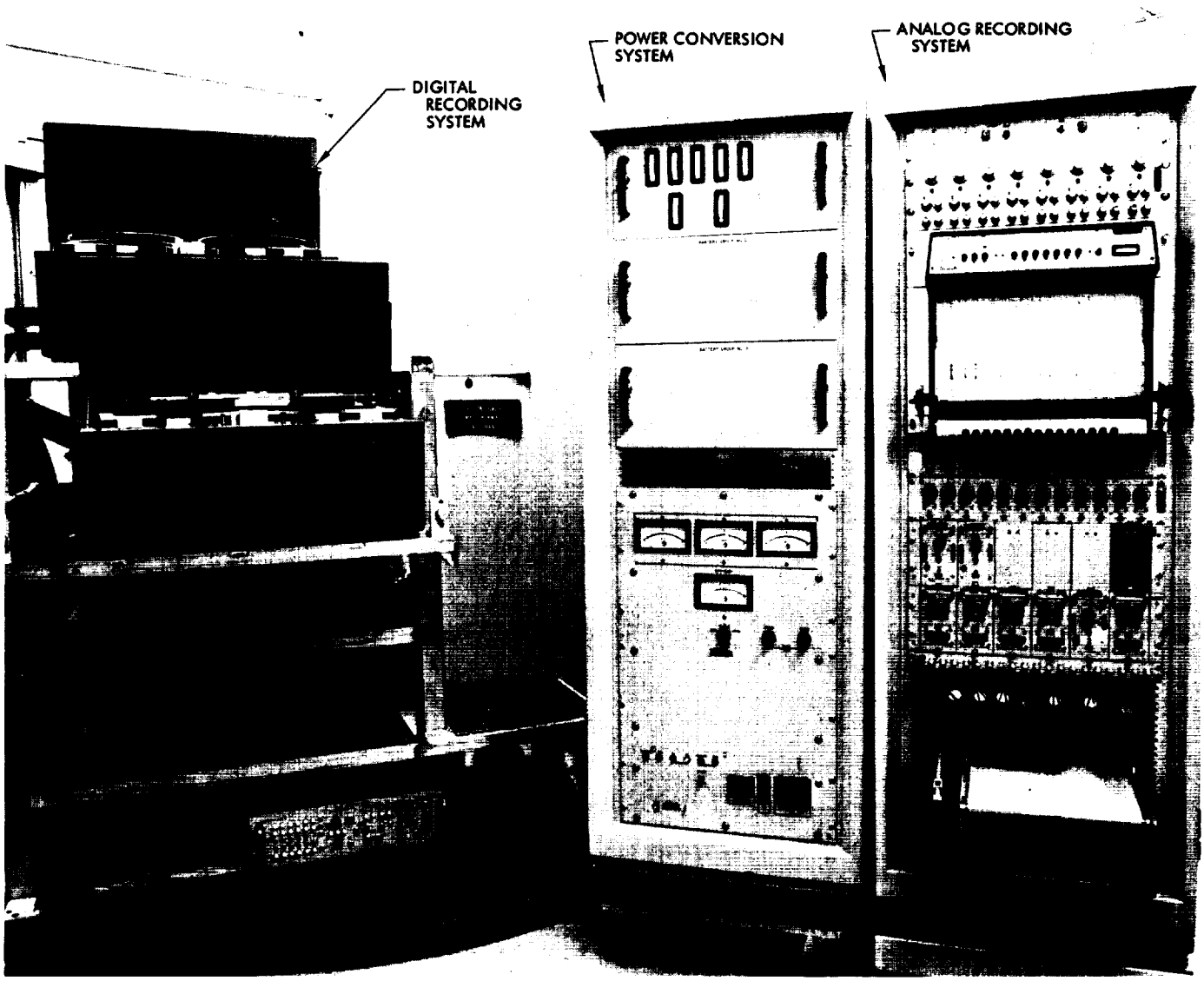


Fig. 2. SEAN System support equipment

altitude information. The altimeter was integrated with the SEAN System in December, 1968.

e. System integration and test support. The integration, testing, installation, maintenance, and documentation support effort for the SEAN System has included the following tasks: procurement of system cables, preparation of block and cabling diagrams and detail interface schematics, and checkout and integration of the support equipment. Work yet to be done in this area includes test and maintenance support of the SEAN System, installation and checkout of the system in the van, and documentation updating.

XI. Spacecraft Control

GUIDANCE AND CONTROL DIVISION

A. Baseline Attitude-Control Subsystem for the Thermoelectric Outer-Planet Spacecraft,

W. E. Dorroh, Jr.

1. Introduction

The design of the Thermoelectric Outer Planet Spacecraft (TOPS) has evolved from previous studies of the system and subsystem requirements for a flyby mission to the outer planets. These earlier studies considered dual-spin, solar electric, and several types of three-axis-controlled spacecraft as design options. The Advanced Systems Technology (AST) program has selected the RTG[1]-powered three-axis-controlled spacecraft option for a more detailed study of the requirements of such a mission.

A description of the baseline attitude-control subsystem for TOPS is presented along with some of the major design considerations.

[1]RTG = radioisotope thermoelectric generator.

2. Functional Requirements

The general attitude-control subsystem functional requirements for the TOPS spacecraft are as follows:

(1) Provide a stable three-axis attitude-controlled spacecraft to enable accurate high-gain antenna pointing during the communications period from L + 250 days to end of mission (9 to 12 yr). Earth tracking within a 0.05-deg (1 σ) error cone from Jupiter and beyond is required.

(2) Provide a capability for orienting the midcourse motor thrust vector in any preprogrammed direction for trajectory correction. Velocity angular dispersions within 7 mrad (1 σ) are desired. Nine corrections are required for the four-planet trajectory.

(3) Provide a stabilized spacecraft to permit accurate pointing of a planet-oriented science platform and an approach guidance platform from approximately E − 30 days until E + 1 day, for all planets.

(4) Provide commanded turn rates of 0.1 deg/sec about the yaw and roll axes.

(5) Provide sufficient acceleration about each of the axes to stop the commanded turn rate within a 4-deg field-of-view.

3. Major Constraints Imposed on Attitude-Control Subsystem

Many additional constraints are imposed on the attitude-control subsystem by configuration and mission-related requirements. Those constraints which affected subsystem design approaches are as follows:

(1) A general mission-oriented constraint is that the attitude-control subsystem be capable of performing its function during a nominal mission without the aid of ground command. Thus, all of the requirements listed above must be met via on-board logic and control systems.

(2) The selected TOPS spacecraft configuration has several large flexible appendages (including the unfurlable high-gain antenna and the science, RTG, and magnetometer booms) which interact with the control system.

(3) The high-gain antenna is rigidly attached to the spacecraft bus, requiring earth pointing of the spacecraft roll axis to maintain communications over that link.

4. Control Torque Source

Both mass expulsion and momentum exchange systems were considered as possible torque sources for the three-axis control system. Mass expulsion systems for long missions requiring low pointing errors suffer from high propellant weights and large numbers of valve cycles.

A momentum-wheel control system with mass expulsion unloading has several advantages:

(1) Since electrical energy is used as the prime source of power, there is no inherent requirement for deadbands.

(2) The wheels provide a recoverable energy source.

(3) Although the wheels eventually saturate and require mass expulsion to remove the momentum, this momentum exchange can be performed at convenient times.

(4) No mass expulsion is required to alter spacecraft attitude.

5. Momentum Wheel Sizing

Momentum storage requirements H are based on the maximum spacecraft angular rates required:

$$H = I_w \cdot \omega_w = I_s \cdot \omega_s$$

where I_w and ω_w are the wheel moment of inertia and maximum speed, and I_s and ω_s are the spacecraft moment of inertia and maximum turn rate (0.1 deg/sec) about that axis.

$$H_y = H_z = 2.62 \text{ lb-ft-s.}$$

$$H_x = 0.52 \text{ lb-ft-s}$$

Maximum spacecraft angular acceleration is determined by the requirement that the wheels reduce the spacecraft rate to zero (from the maximum) within 4 deg of angular position.

$$\alpha_s = \frac{(\theta)^2}{2\theta} = 21.8 \times 10^{-6} \text{ rad/s}^2$$

The ac motor torque requirements N are based on the maximum spacecraft angular acceleration required:

$$N = I_s \alpha_s$$

Thus

$$N_y = N_z = 0.0327 \text{ ft-lb} = 6.3 \text{ oz-in.}$$

$$N_x = 0.0065 \text{ ft-lb} = 1.26 \text{ oz-in.}$$

The following empirical expression shows the relationship between synchronous speed and power at stall condition for a given stall torque:

$$\text{Stall power (W)} = \frac{\text{stall torque, oz-in.}}{663} \times \text{synchronous speed, rpm}$$

The synchronous speed of the ac drive motor has been chosen on other programs on the basis of reliability, life, and power to be 1200 rpm, using 400-Hz power.

$$P_y = P_z = 11.4 \text{ W}$$

$$P_x = 2.27 \text{ W}$$

To increase the system reliability and to reduce the power requirements during cruise, the total angular

momentum storage required for each axis will be shared equally by two wheels. Both wheels will be used for commanded turns, but one will be placed in standby during cruise. Thus if one wheel fails, commanded turns can still be performed at half speed.

Table 1 summarizes the momentum wheel characteristics, utilizing off-the-shelf wheels. Some weight and power savings can be achieved by designing wheels tailored to the TOPS requirements.

Table 1. Momentum wheel characteristics (per wheel)

Parameter	Yaw and roll	Pitch
Momentum wheel storage, lb-ft-s	1.46	0.5
Inertia, slug-ft^2	0.014	0.0047
Weight, lb	10.1	4.8
Dimensions (diameter × height, in.)	7.5 × 3.5	6 × 3
Motor stall torque, oz-in.	4	2
Stall power, W	9.5	4.6
Synchronous speed, rpm	1200	1200
Maximum operating speed, rpm	1000	1000

6. Mass Expulsion System

Mass expulsion is required for unloading the momentum wheels, initial rate reduction, and third-axis (roll) control during midcourse motor burn. The estimated total torque impulse requirement, excluding leakage, is approximately 850 lb-ft-s.

Two methods of unloading the momentum wheels were considered:

First, the gas jet is turned on for a fixed length of time, the torque-time product being equal to the momentum stored in the wheel. This requires a variable on-time and a method of calculating the required on-time based on wheel speed. If the jet torque is equal to or less than the motor torque, no transient occurs during unloading, since the momentum wheels can keep up with the unloading impulse. For jet torques T_D greater than wheel torques T_W the attitude error θ_E is given by

$$\theta_E = \frac{H^2\left(1 - \frac{T_W}{T_D}\right)}{2I_S T_W}$$

Since the error resulting from unloading a fully saturated wheel would exceed the allowable pointing error, multiple pulses would be required.

Second, full voltage is applied to the wheel to drive it to zero speed, allowing the gas jets to maintain attitude control. This method is simpler to implement (although it requires more valve actuations) and has tentatively been selected for use on TOPS. The resulting transient error is limited to a small excursion out of the gas system deadbands.

7. Antenna Pointing

High-gain antenna pointing is required at approximately $L+250$ days. Figure 1 shows the apparent earth movement for the earth–Jupiter leg of the mission for a Canopus–sun-oriented spacecraft. Several methods of antenna pointing were investigated:

(1) Single mechanical degree-of-freedom antenna, plus biased Canopus sensor (SPS 37-51, Vol. III, pp. 63–65).

(2) Biased pitch and yaw sun sensors.

(3) Two-degrees-of-freedom antenna.

The first of these has inherent high pointing errors when the sun and earth are approximately aligned, which is the case for most of the encounters.

The third system has the undesirable feature of mechanically gimballing a large antenna, and precludes the use of the antenna feed for mounting sun sensors and gas jets.

Two-degrees-of-freedom pointing can be accomplished by biasing both pitch and yaw sun sensors, at the cost of some increase in weight and power, which has no inherent pointing error as the earth track passes near the sun. In addition, the mechanical gimballing of the antenna is avoided.

The primary sun sensor provides a field-of-view of ± 3 deg about the pitch axis and ± 14 deg about the yaw axis. The output signal for each axis is produced by sunlight passing through a narrow slit and impinging on a coded mask detector.

The inherent resolution of this basic system is limited by the angular diameter of the sun, which directly affects the width of the sunlit area at the detector plane. As this width increases, a particular detector channel functions correctly until this width exceeds the width of two slots in the coded patterns.

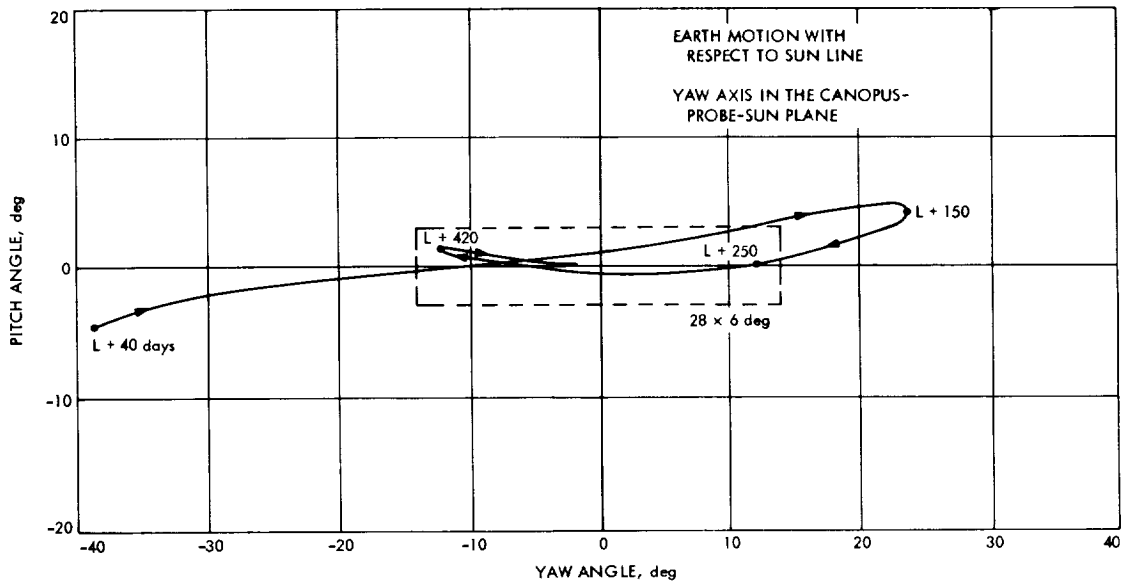


Fig. 1. Apparent earth movement for earth–Jupiter leg of mission with a Canopus–sun-oriented spacecraft

8. Attitude-Control Configuration

A block diagram for a single-axis attitude-control system is shown in Fig. 2. The basic control loop is very similar to the *Mariner* spacecraft mechanization, consisting of a deadband with inputs from optical and inertial sensors. Unlike the *Mariner* system, deadband size is not limited by gas consumption or valve cycles but only by sensor noise, and values of ± 0.8 mrad (compared to ± 4 mrad for *Mariner*) can be mechanized to achieve high accuracy antenna pointing. Passive rate compensation (derived rate) will be employed during cruise. Wheel unloading control is shown as method (2) in *Subsection 6.*

Provision is made in the logic circuit to unload the wheels based on wheel speed or CC&S commands, inhibit unloadings during critical communications periods, and switch to the standby wheel, if required.

An inertial reference is provided by strapdown rate gyros, with their outputs electronically integrated to produce inertial position.

Celestial sensors (sun and Canopus) provide optical references. Stored pitch and yaw sun sensor bias voltages (in the CC&S) are used to orient the spacecraft roll axis to the earth while maintaining lock on celestial references.

Provision is made for utilizing position error signals from the radio subsystem for closed RF loop pointing in pitch and yaw, in order to meet the pointing requirements during high bit rate data transmission. If RF lock is lost, the attitude-control will switch to the celestial sensors.

9. Functional Sequence

Following separation from the launch vehicle, the mass expulsion system reduces the initial tumbling rates to within a controlled rate deadband (rate gyros are used as sensors). Due to the high angular rates (~ 3 deg/s), momentum wheels would quickly saturate in this mode, and hence are not used. The sun is acquired in pitch and yaw, and the resulting sun acquisition signal turns on the wheels and initiates Canopus acquisition.

During the cruise phase, gyros are turned off, and rate damping is provided by passive compensation networks. Periodically during the mission, disturbance torques will cause the wheels to saturate, and the gas jets will be fired to unload the wheels. This will occur only about 1000 times per axis, during a 12-yr mission.

At $L + 250$ days, earth pointing of the high-gain antenna is required. The sensor output is digitally compared to stored earth-pointing bias angles in CC&S and the result is used to operate the momentum wheels. The pitch sun sensor biases cause an apparent change in the Canopus cone angle. This angle will also be stored in CC&S, updated as required, and provided to the Canopus tracker.

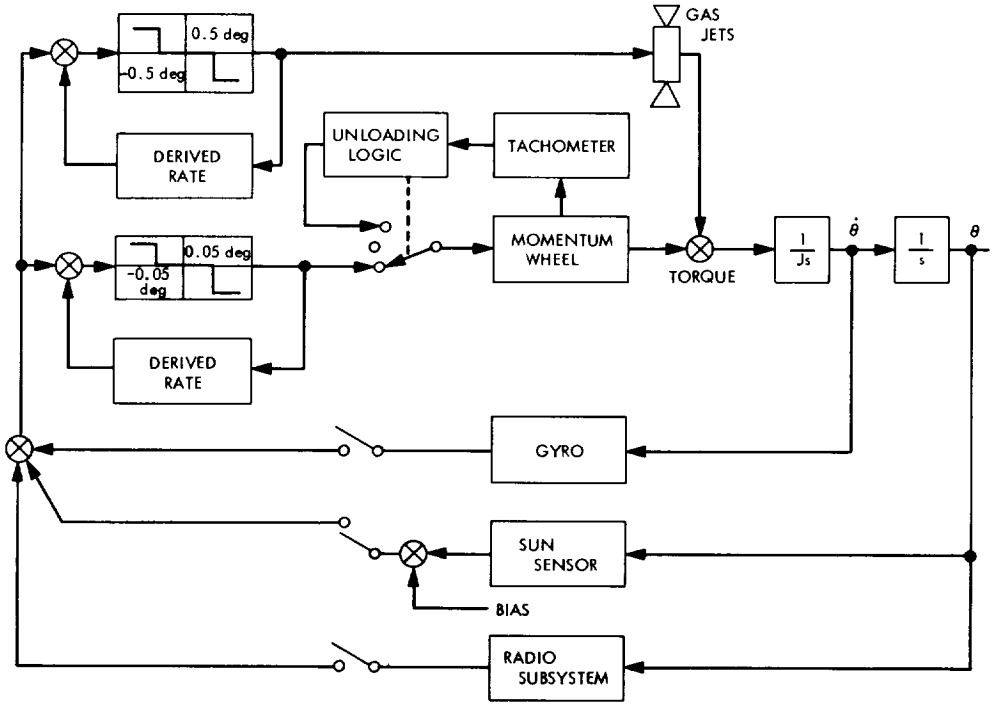


Fig. 2. Single-axis attitude-control system block diagram

In order to allow single degree-of-freedom actuators in the medium-gain antenna and cruise science, the spacecraft cruise orientation will be such that the negative-yaw axis will point toward the South Ecliptic Pole, rather than Canopus. This requires that additional bias angles from the CC&S be provided to the tracker.

For trajectory corrections, commanded turns are performed using the gyros as an inertial reference. Momentum wheels are used, and thus no mass expulsion is required.

A *Mariner* Mars 1971 spacecraft type of gimballed autopilot is implemented for thrust vector control during midcourse motor burn, and a positive feedback gimbal angle loop is employed to reduce thrust vector error. Since a gimballed autopilot can provide only two-axis torque control, the attitude-control gas system will be employed for third-axis control. Digital encoders will be required for the gimbal angles, with pre-aim angles provided by the CC&S.

At approximately $E-30$ days, the scan platform control system will be activated. It consists of two stepper motor driven axes, with shaft encoder feedback. Desired platform position angles will be provided by the CC&S. Due to the slow scanning rates required in far encounter mode, capacitor energy storage will be used to decrease peak stepper motor power consumption. At approximately $E-4$ h, the gyros will be turned on, in anticipation of temporary degradation (low signal-to-noise ratio) of the celestial sensors. Control will be switched to the inertial references by stored command. The scan platform slew rates are sufficiently high during this period to preclude the use of energy storage for the scan actuators.

The spacecraft will remain on inertial control until post-encounter and post solar occultation. When the planet has receded from the optical sensor fields of view, the celestial references will be reacquired, and the spacecraft will return to the cruise mode.

To increase system reliability, standby redundant sun sensors, Canopus sensors, gyros and attitude-control electronics are employed.

B. Effects of Ion Engine Thrust Disturbances on an Attitude-Control System, *L. L. Schumacher*

1. Introduction

The Solar Electric Propulsion System Technology program SEPST III is preparing to make a sustained life test of the mercury electron-bombardment ion thrusters

and their control system. Past systems evaluations (SEPST II) have manifested two problems: (1) values originally selected for the stepper motor stepping rate and step size must be modified to more realizable values for hardware applications; and (2) it was discovered that ion engines similar to the ones to be life-tested exhibit a periodic thrust interruption. The effects of these two problems on stepper motor life were investigated in a detailed simulation of the spacecraft and attitude-control system. It was determined that the new stepping rate and step size are acceptable from a control point of view, that the periodic thrust interruption does not cause more stepper motor steps to be taken, and a more exacting estimate of stepper motor life may be necessary.

2. Determination of Disturbance Torque Profile and Spacecraft Model

Attitude control during the cruise phase of the flight will be effected by thrust vector pointing. Thrust vector pointing is accomplished with two-axis translation of the engine array, with gimballed pairs of engines within the array providing the third axis of control. The control positioning is done in discrete amounts with stepper motors. As a result, the thrust vector can be assumed to be offset from the CG by some fraction of one step at all times, thereby creating a disturbance torque. The thrust interruptions also cause impulsive disturbance torques.

The proposed ion engines develop about 0.01 lb of thrust each, and each interruption lasts about ⅓ s. It can be shown that this disturbance causes the largest momentum change in the axes controlled by linear translation of the thruster array. The interaxis coupling torques due to this disturbance are small and can be neglected. The flexible body dynamics also tend to reduce the number of corrective motor steps taken, so they are also neglected. For the above reasons, a single-axis rigid body model was used to determine what effects the thrust interruptions have on the number of steps taken by the control system stepper motor.

3. Solar Electric Ion Engines Attitude-Control System

The block diagram of Fig. 3 shows the proposed attitude-control system. The control loop simulation used DSL/90 simulation language. The number of steps necessary to correct for the periodic thrust interruptions were determined in the following manner: The thrust vector was assumed to be misaligned by various fractions of one step, and the limit cycle was observed with and without the periodic thrust interruptions. It was

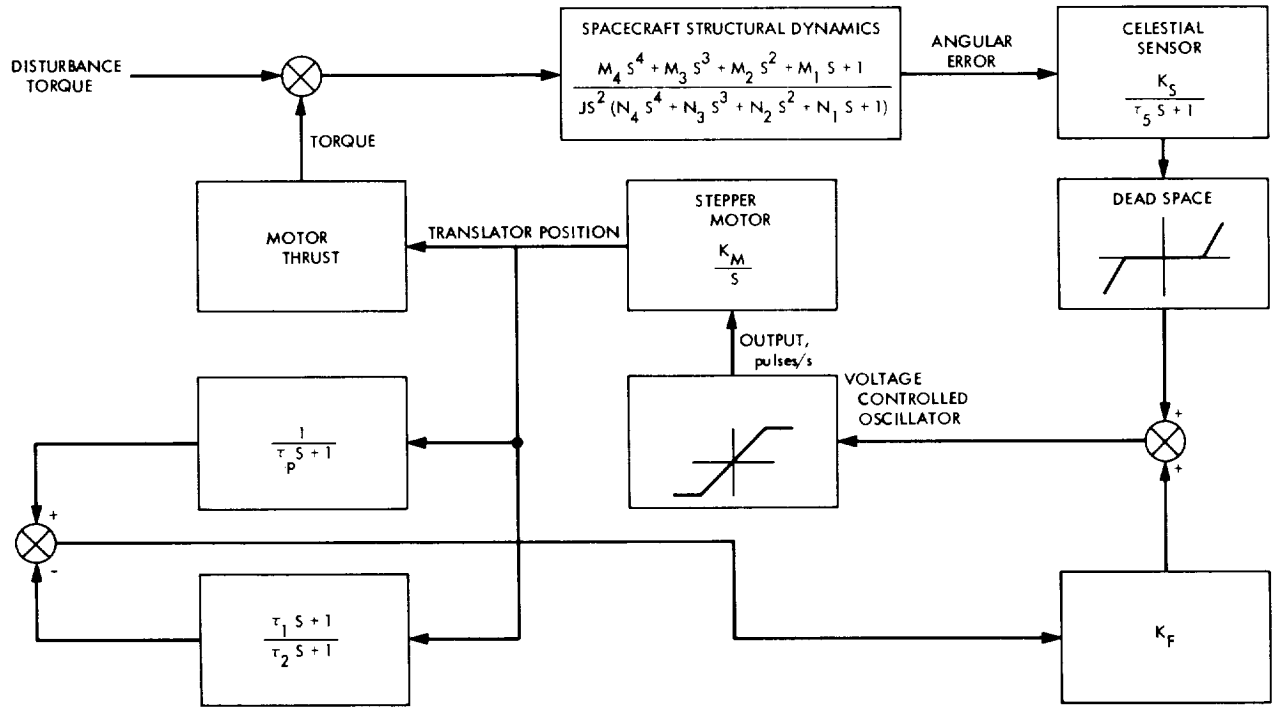


Fig. 3. Translator control system block diagram

observed that the number of steps taken with the disturbance was at least equal to, and in most cases less than, the number without the disturbance.

The explanation of this seemingly unlikely condition lies in the nature of stepper motor and system compensation. In the case where the disturbance torque is due entirely to the inherent CG misalignment, the steady-state position remains on the edge of the deadband, as in Fig. 4. The position sensor output during this time is very near, but not quite, zero. Also, each time the translator steps, an error pulse is produced by the feedback loop causing the translator to step right back. This stepping action is produced by a voltage-controlled oscillator. This device is basically an integrator which integrates an error signal until a certain level is reached, at which time it discharges this error in the form of a signal which steps the stepper motor. The error is the sum of the position sensor error signal and the feedback signal (Fig. 3). The total number of stepper motor steps is therefore a function of the feedback compensation and the integral of the position error signal.

When a small impulsive disturbance torque occurs in the spacecraft and the angular position is very near the deadband, there is a large probability that the position will be moved within the deadband with the proper rate polarity to remain there (Fig. 5). The translator continues to step at a rate which is a function of the feedback compensation parameters only. While within the deadband, the translator takes fewer steps than outside the deadband. The difference is the integral of the position error that normally occurs when the spacecraft is oscillating at the edge of the deadband. Since the rates are very small, this time within the deadband becomes large, and the integral of the small position error for long periods of time becomes significant. For this reason the overall effect of the periodic thrust interruptions on the stepper motor life was concluded to be negligible. In the case cited above, 590 steps were used in the undisturbed case as opposed to 560 steps for a periodically disturbed case.

4. Conclusion

It is concluded that the periodic thrust interruptions will have no adverse effect on stepper motor life. However, former studies estimated that time between steps in the steady state to be 32 s. The average time between steps in the recent simulations was 15 s. For a powered flight of 500 days (4.32×10^7 s) this brings the total number of steps to 2.55×10^6, as opposed to 1.35×10^6 previously estimated. If 10^7 steps are assumed to be the performance limit of a stepper motor, then the operational safety factor is reduced from 8 to 4. If the future ground tests uphold the above estimates, a more exacting estimate of stepper motor lifetime may be necessary.

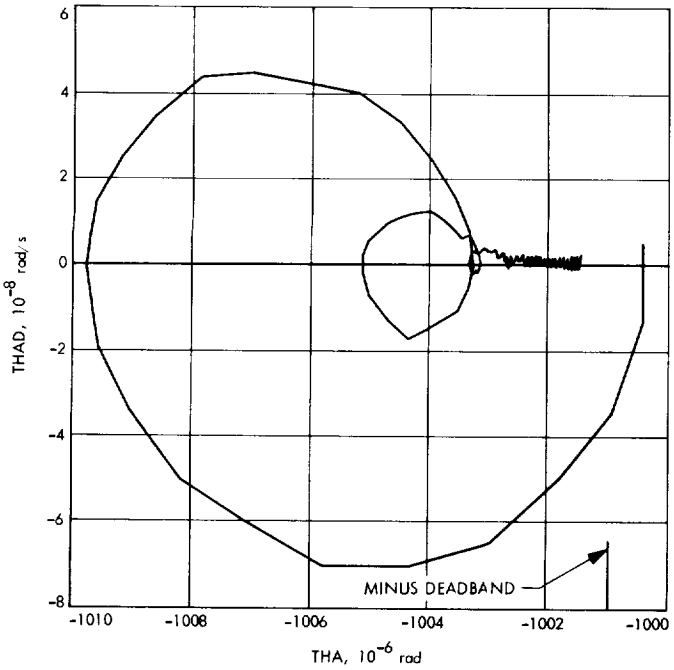


Fig. 4. Phase plane with no disturbance

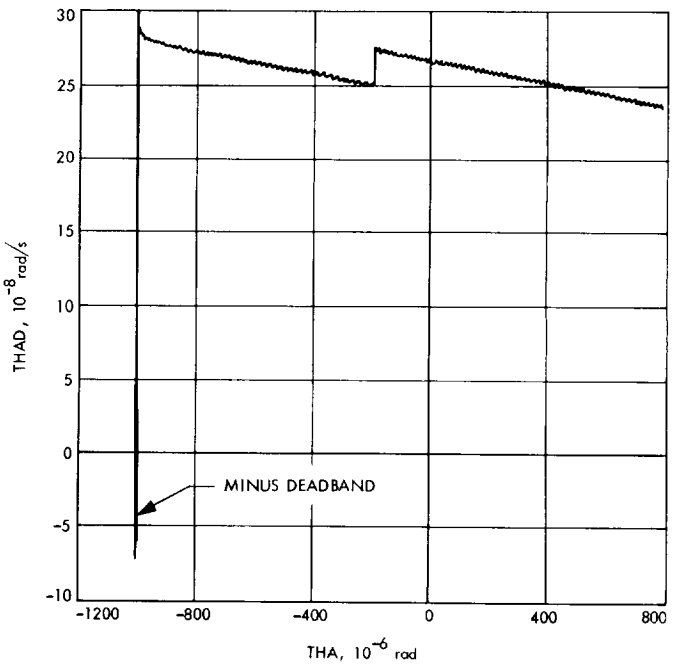


Fig. 5. Phase plane with disturbance

C. Design and Development of the SCR dc to dc Voltage Converters for the Minimum Energy Controller, *Y. E. Sahinkaya*

1. Introduction

This article contains a brief description of the theoretical and practical concepts utilized in the design, development, and laboratory testing of the SCR dc to dc voltage converters for the minimum energy controller. (SPS 37-56, Vol. III, pp. 138–144; SPS 37-56, Vol. III, pp. 99–103 and Footnote 2.) Figure 6 shows the configuration of the SCR dc to dc voltage converters. The function of each component in Fig. 6 is as follows:

SCR1 is the main SCR of the dc to dc step-down voltage converter which supplies the required electric current from the battery into the armature circuit whenever a signal v_{T1} of appropriate duration and amplitude is applied from the pulse-width modulator to its gate circuit. SCR1 is turned off by SCR2, which applies a voltage pulse across SCR1 in the reverse direction. The amplitude of the voltage pulse with respect to ground potential is approximately equal to twice the value of the battery voltage. This commutation or turn-off action of SCR1 is accomplished by the operation of a resonant commutation circuit consisting of $L1$, D1, C1, and SCR2. At the end of commutation action, SCR1 regains its blocking ability and is ready to be turned on by v_{T1} in the next cycle. By adjusting the duration of on and off times of SCR1 within the pre-specified period of operation, it is possible to vary the average voltage across the motor terminal A and B shown in Fig. 6, according to a control law. The average voltage between terminals A and B is always less than the battery voltage E_b.

The gate signals v_{T3} and v_{T4} are turned off by the automatic action of the armature current direction sensor. SCR3 is the main SCR of the dc to dc step-up voltage converter which transfers the required current from the motor armature circuit into the battery whenever a signal v_{T3} is applied from the pulse-width modulator to its gate circuit. The gate signals v_{T1} and v_{T2} are turned off by the automatic action of the armature current direction sensor. SCR3 is turned off by SCR4 which applies a voltage pulse across SCR3 in the reverse direction. The amplitude of the voltage pulse with respect to ground potential is

[2]Sahinkaya, Y. E., *An Optimal Electric Drive System*, May 26, 1969 (JPL internal document).

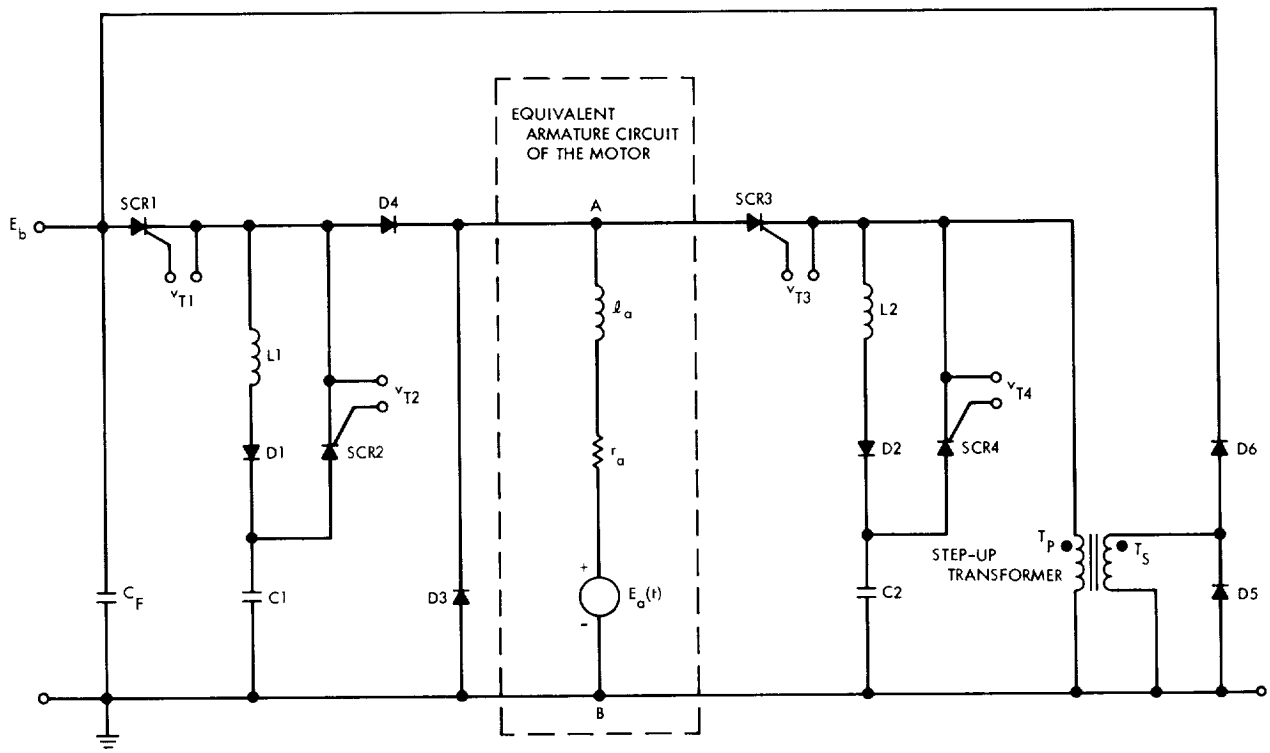


Fig. 6. The dc to dc voltage converters of the minimum energy controller

approximately equal to twice the value of the motor back emf voltage. This commutation action is accomplished by the action of a resonant commutation circuit consisting of $L2$, D2, C2, and SCR4. At the end of commutation action, SCR3 regains its blocking ability and is ready to be turned on by v_{T3} in the next cycle. By adjusting the duration of on and off times of SCR3 within pre-specified period of operation, it is possible to vary the average voltage across the primary winding T_p of the step-up transformer (Fig. 6) according to a control law. The voltage across T_p is always less than the back emf voltage of the motor.

Capacitor C_F is the filter capacitor which also maintains a stiff source voltage for the control circuits.

Diode D4 prevents the charging of capacitor C1 by the back emf voltage of the motor during the regenerative braking control action.

Diode D3 provides a continuous current flow to the armature circuit during the motoring mode of operation. Note that the battery voltage E_b is switched on and off across the armature terminals A and B in such a way that the average value of the armature voltage approximates to the value of voltage given by the optimal control law. This diode is called the free-wheeling diode for the motor armature circuit.

The equivalent armature circuit of the motor is assumed to have a constant dc resistance r_a, a constant inductance ℓ_a and a voltage generator E_a or back emf voltage which is linearly proportional to the speed of the motor.

A step-up transformer whose secondary winding T_s contains more turns than its primary winding T_p, charges the battery from a pre-specified minimum value of back emf voltage during the regenerative braking. Note also that the voltage level across T_s for the maximum value of back emf voltage during the regenerative braking must also be taken into account in the choice of the ratio of turns between the primary and secondary windings.

Diode D5 is the free-wheeling diode for the step-up transformer circuit. Its primary function is to prevent the occurrence of extremely high voltages whenever SCR3 is turned off during the regenerative control action. During the time interval over which SCR3 is off, the energy stored in the transformer windings is dissipated in the resistance of D5.

Diode D6 provides a unidirectional current flow from the motor armature circuit into the battery during the time interval over which SCR3 is in conduction.

2. Design of the SCR dc to dc Voltage Converters

a. Design of the SCR dc to dc step-down voltage converter. The design parameters of the SCR dc to dc step-down voltage converter are obtained as follows:

Turn-on circuit. SCR1 is turned on by the application of a gate signal v_{T1}. By assuming zero initial conditions on the current in L_1 and on the voltage across C1 and applying standard methods to the commutating circuit of SCR1, yield the following equations:

$$i_{C1}(t) = \left(\frac{E_b}{L_1\omega}\right) e^{-\alpha t} \sin \omega t \tag{1}$$

$$V_{C1}(t) = E_b\,[1 - e^{-2\alpha t} \cos \omega t] - E_b \left(\frac{\alpha}{\omega}\right) e^{-\alpha t} \sin \omega t \tag{2}$$

where

$i_{C1}(t) \triangleq$ instantaneous value of the charging current of capacitor C1

$V_{C1}(t) \triangleq$ instantaneous value of the voltage across capacitor C1

$r_c \triangleq$ dc resistance of diode D1 which is assumed to be constant

$$\alpha \triangleq \frac{r_c}{2L_1}$$

$$\omega^2 \triangleq \left(\frac{1}{L_1C_1} - \frac{r_c^2}{4L_1^2}\right)$$

Note that in the derivation of Eqs. (1) and (2) it has been assumed that when SCR1 is turned on, the battery current flows mostly through L_1, and D1 to C1, over the charging interval, which is on the order of 100–200 μs for practical reasons. This is justified since $\ell_a \gg L_1$. Furthermore, if α is also small, then $e^{-\alpha t} \approx 1.0$. Hence Eqs. (1) and (2) become

$$i_{C1}(t) \cong \left(\frac{E_b}{L_1\omega}\right) \sin \omega t \tag{3}$$

$$V_{C1}(t) \cong E_b(1 - \cos \omega t) \tag{4}$$

From Eq. (4) it is seen that

$$v_{C1}(t_c) \cong 2E_b$$

if, and only if,

$$t = t_c = \frac{\pi}{\omega} = \pi(L_1C_1)^{1/2} \tag{5}$$

where $t_c \triangleq$ time at which the capacitor charging current ceases. In practice, in general

$$t_w(\min) > t_c$$

where

$t_w(\min) \triangleq$ minimum allowable pulse-width for the conduction of SCR1

When capacitor C1 is charged to a voltage value which is equal to twice the value of the battery voltage with respect to ground potential, the battery current flows into the motor circuit according to the following equation:

$$i_a(t) \cong i_a(t_c)\left\{\exp\left[-\left(\frac{r_a}{l_a}\right)(t-t_c)\right]\right\} + \left(\frac{E_b - E_a(t)}{r_a}\right)\left\{1 - \exp\left[-\left(\frac{r_a}{l_a}\right)(t-t_c)\right]\right\}, \quad t_c \le t \le t_p \tag{6}$$

where

$i_a(t) \triangleq$ instantaneous value of the motor armature current

$t_p \triangleq$ period of pulse-width modulation operation

In Eq. (6), the back emf voltage $E_a(t)$ is assumed to be constant during each cycle of operation. Therefore, Eq. (6) describes the variation of armature current in a given cycle of operation.

Turn-off circuit. SCR1 is turned off by the commutation action of SCR2. The current through SCR1 in the reverse direction must at least be equal to the motor armature current. For the worst case, the following equation must be satisfied:

$$C_1\frac{dv}{dt} = i_a\,(\max) \tag{7}$$

From Eq. (7)

$$v(t) = E_b + \left(\frac{i_a\,(\max)}{c}\right)t_{off} \tag{8}$$

t_{off} = turn-off time of SCR1
if $v(t) \approx 2E_b$ during the turn-off of SCR1, then

$$C_1 \approx \left(\frac{t_{off}\, i_a(\max)}{E_b}\right) \tag{9}$$

Letting $E_b = 25.0$ V and using Eqs. (5) and (9), the following parameter values for C_1 and L_1 are obtained:

$$C_1 = 5.0\ \mu\text{F}$$

$$L_1 = 0.5\ \text{mH}$$

where

$$t_c = 100.0\ \mu\text{s}$$

$$t_{off} = 12.0\ \mu\text{s}$$

$$i_a(\max) = 10.0\ \text{A}$$

A detailed analysis has indicated that a pulse frequency of 250.0 Hz is adequate for the proper variation of the armature voltage.

Components. The following components are selected:

SCR1 = C 12 C (GE), mounted on NC401 heat sink

SCR2 = C 9 A (GE)

D1 = IN3881R(GE)

D3, D4 = IN388R(GE), mounted on NC 301 heat sinks

C1 = 28 F 952 (GE)

CF = 1.0 μf, 200.0 V

L_1 = 500.0 μF

SCR1 and SCR2 must have turn-off times which are less than and equal to 12.0 μs and good *di/dt* and *dv/dt* capabilities.

b. Design of the SCR dc to dc step-up voltage converter. The design parameters of the SCR dc to dc step-up voltage converter are obtained as follows:

Turn-on circuit. The values of inductance L_2 and commutation capacitor C2 are identical to those of L_1 and C1, respectively.

Turn-off circuit. The ratio of turns of T_s/T_p is assumed to be 5.0. This in turn requires that regenerative braking is possible for back emf voltages which are more than 5.0 V. Note that the maximum back emf voltage is always less than 25.0 V, hence the maximum battery charging voltage is always less than 125.0 V. Since the duration of regenerative braking is not large at high voltages, the battery terminals are not expected to be damaged under these circumstances.

Components. The following components are selected:

SCR3 = C12A(GE), mounted on NC 301 heat sink

SCR4 = C9A(GE)

D2 = IN3881R(GE)

D5, D6 = IN3883R(GE), mounted on NC 301 heat sinks

C2 = 28F952(GE)

L2 = 500.0 μF

The same considerations as given for SCR1 and SCR2 are used in the selection of SCR3 and SCR4.

3. Test Results

The dc to dc SCR step-down voltage converter has been tested with the pulse-width modulator. The preliminary test results indicate that the actual operation of the system agrees very closely with its predicted operation based on the theoretical results. The test results for the dc to dc SCR step-up voltage converter will be given in a future article.

Reference

1. *Silicon-Controlled Rectifier Manual,* Third Edition, General Electric Co., Auburn, N.Y., 1969.

D. Effects of Inertia Cross-Products on TOPS Attitude-Control, *L. F. McGlinchey*

1. Introduction

Several of the recent configurations for the Thermal-electric Outer Planet Spacecraft (TOPS) have inertial properties with at least one inertia cross-product that is a significant percentage of the principal moments of inertia. This, in general, is an undesirable condition from the standpoint of the attitude control. The major reasons are as follows:

(1) A large inertia cross-product can produce cross-coupling torques that are sufficient in magnitude to ensure momentum wheel saturation during acquisitions and commanded turns. This condition degrades the attitude-control system performance.

(2) During motor burns, inertia cross-products can produce severe attitude transients at motor ignition. This effect can degrade pointing accuracy, particularly for short motor-burn durations.

To illustrate the effects of cross-products on the TOPS attitude control, a brief analysis was made in which an allowable upper bound for the magnitude of inertia cross-products is established. In addition, if large cross-products cannot be avoided, recommended changes to the attitude-control system mechanization are described.

2. Analysis

The basic equations of motion for the spacecraft can be derived from the fundamental equation relating torque and angular momentum.

$$\mathbf{T}_{\text{space}} = \dot{\mathbf{H}}_{\text{space}} = \dot{\mathbf{H}}_{\text{spacecraft}} + \boldsymbol{\omega} \times \mathbf{H} \tag{1}$$

$$= \frac{d}{dt}[\Pi \cdot \boldsymbol{\omega}] + \boldsymbol{\omega} \times \mathbf{H} \tag{2}$$

where

$\mathbf{H}$ = total angular momentum

$\boldsymbol{\omega}$ = spacecraft angular velocity

Π = spacecraft inertia dyadic

Equation (2) can be expanded and solved for $\dot{\omega}_x$, $\dot{\omega}_y$, and $\dot{\omega}_z$ to yield equations for the spacecraft x, y, and z torques in the following form:

$$N_x = I_{xx}\dot{\omega}_x = A_{11}L_x + A_{12}L_y + A_{13}L_z \tag{3}$$

$$N_y = I_{yy}\dot{\omega}_y = A_{21}L_x + A_{22}L_y + A_{23}L_z \tag{4}$$

$$N_z = I_{zz}\dot{\omega}_z = A_{31}L_x + A_{32}L_y + A_{33}L_z \tag{5}$$

In Eqs. (3), (4), and (5), the coefficients (A_{ij}) are constants and are determined from the elements of the

inertia dyadic. The torques L_x, L_y, and L_z are given by

$$L_x = (I_{zz} - I_{yy})\omega_y\omega_z + I_{yz}(\omega_z^2 - \omega_y^2) + I_{xy}\omega_x\omega_z - I_{zx}\omega_x - T_x \quad (6)$$

$$L_y = (I_{xx} - I_{zz})\omega_x\omega_z + I_{zx}(\omega_x^2 - \omega_z^2) + I_{xy}\omega_y\omega_z - I_{yz}\omega_x\omega_y - T_y \quad (7)$$

$$L_z = (I_{yy} - I_{xx})\omega_x\omega_y + I_{xy}(\omega_y^2 - \omega_x^2) - I_{yz}\omega_x\omega_z + I_{zx}\omega_y\omega_z - T_z \quad (8)$$

Inspection of Eqs. (6), (7), and (8) shows that the spacecraft torques given by Eqs. (3), (4), and (5) are made up of external torques (control torques), gyroscopic cross-coupling torques, and inertia cross-product torques. During all the various attitude-control modes, with the exception of initial rate reduction and acquisition of the sun, the gyroscopic and other terms containing cross rate products ($\omega_i\omega_j$) are essentially zero. Initial rate reduction and acquisition, however, are accomplished using the mass expulsion system. The presence of these additional cross-coupling torques creates a small effect on gas consumption and the time required to accomplish these events. The areas of concern, however, are the effects of the inertia cross-product terms on momentum wheel control during commanded turns and searches, initial transient behavior at midcourse motor ignition, and autopilot stability. With regard to commanded turns and roll searches, a large cross-product of inertia can produce wheel saturation in an axis normal to the turn axis. This will give rise to wheel unloading during the turn which produces undesirable attitude transients in this mode. Referring to Eqs. (3), (4), and (5), the resultant torque about each control axis must be zero for proper attitude control. Therefore

$$A_{11}L_x + A_{12}L_y + A_{13}L_z = 0 \quad (9)$$

$$A_{21}L_x + A_{22}L_y + A_{23}L_z = 0 \quad (10)$$

$$A_{31}L_x + A_{32}L_y + A_{33}L_z = 0 \quad (11)$$

Also in Eqs. (6), (7), and (8) the $\omega_i\omega_j$ cross rate terms are essentially zero.

$$L_x \approx L'_x = I_{yz}(\omega_z^2 - \omega_y^2) - T_x \quad (12)$$

$$L_y \approx L'_y = I_{zx}(\omega_x^2 - \omega_z^2) - T_y \quad (13)$$

$$L_z \approx L'_z = I_{xy}(\omega_y^2 - \omega_x^2) - T_z \quad (14)$$

Equations (9) through (14), in general, can be used to determine the effect of cross-products on attitude control. However to facilitate the analysis, and to provide some indication of allowable upper bounds on inertia cross-products for the TOPS spacecraft, it is assumed that only one large cross-product will exist and that the other two can be neglected. Under this assumption, Eqs. (9), (10), and (11) reduce to

$$I_{xy}(\omega_y^2 - \omega_x^2) - T_z = 0, \qquad \text{if } I_{xy} \text{ is large} \quad (15)$$

$$I_{yz}(\omega_z^2 - \omega_y^2) - T_x = 0, \qquad \text{if } I_{yz} \text{ is large} \quad (16)$$

$$I_{zx}(\omega_x^2 - \omega_z^2) - T_y = 0, \qquad \text{if } I_{zx} \text{ is large} \quad (17)$$

or, in general, for any set of control axes i, j, k:

$$I_{ij}(\omega_j^2 - \omega_i^2) - T_k = 0 \quad (18)$$

In terms of cross-coupled momentum

$$H_k = T_k \cdot t = I_{ij}(\omega_j^2 - \omega_i^2) \cdot t \quad (19)$$

For a commanded turn or search about the j-axis

$$t = \frac{\Delta\theta_j}{\omega j}$$

$$\omega_i = 0$$

$$H_k = I_{ij}\omega_j\Delta\theta_j$$

Under these conditions, it is desirable to maintain $H_k < (H_w)_k$ = momentum storage of wheel for k-axis. Hence an upper bound on I_{ij} is given by

$$(I_{ij})_{\max} < \frac{(H_w)_k}{\omega_j(\Delta\theta_j)_{\max}}, \qquad \Delta\theta_{\max} = 180 \text{ deg} \quad (20)$$

For the recent TOPS configurations, the momentum wheels are sized for $(H_w)_x = (H_w)_z = 1.46$ ft-lb-s and $(H_w)_y = 0.5$ ft-lb-s. For a turn rate of 0.1 deg/s,

$$|I_{xy}| \text{ or } |I_{yz}| = 266.3 \text{ slug-ft}^2$$

$$|I_{zx}| = 91.2 \text{ slug-ft}^2$$

If more than one inertia cross-product is significant, then their effects must be evaluated using Eqs. (9) through (14).

3. Conclusion

The forgoing analysis was made to develop a set of useful equations for evaluating the effects of inertia cross-products; its magnitude should be less than the value obtained using Eq. (33).

If large cross-products of inertia cannot be avoided, i.e., the vehicle control axes cannot be closely aligned with the principal axes, then the following modifications can be made to the attitude-control system:

First, to eliminate cross-coupling in the momentum wheel control system, particularly during turns, the axes of the momentum wheels and gyros can be aligned with the principal axes instead of the reaction jet control axes. Commanded turns would now be performed about the principal axes. Depending on the angular misalignment between the celestial sensor axes and the principal axes, appropriate mixing of the celestial sensor error signals may be required.

Second, with regard to the autopilot control system, inertia cross-products can create severe start-up transients and be destabilizing to the control system. Two possible mechanizations exist which can eliminate this problem:

(1) If possible, locate the two engine gimbal axes and the engine thrust vector axis with the principal axes. Since the gyros can also be aligned with these axes, no mixing is required. A minimum requirement for this mechanization would be to at least locate the gimbal axes along principal axes.

(2) If (1) cannot be mechanized, then the second alternative is to electrically mix out the cross-products in the autopilot electronics. However this can be risky, since this type of mechanization must presume that the cross-products of inertia are very accurately known.

E. Actuator Endurance Testing for a Clustered Ion Engine Array, *J. D. Ferrera and E. V. Pawlik*

1. Introduction

In the interest of demonstrating the systems technology associated with solar-powered electric propulsion systems for interplanetary spacecraft applications, a system breadboard which models all the subsystems necessary to perform a spacecraft mission was developed and tested. This breadboard includes gimbal actuators on two engines for roll attitude control and a translator actuator to translate the two engine structure to align the resultant thrust vector with the center of gravity of the spacecraft for pitch and yaw control. A complete description of the test system and functional description and test data of the two types of actuators can be found in SPS 37-54, Vol. III, p. 60. This report describes the results of the system endurance testing with regard to the actuators.

2. System Test

The complete system was subjected to a total of approximately 550 h of vacuum testing during which time at least one (but never both) of the engines were operating at full power. The actuators were run intermittently throughout the test. After the first 250 h, the system was taken out of the chamber for modification to the ion engines. No changes were made to any of the actuators. Figure 7 is a photograph of the system after 250 h. It can be seen that backsputter material coating the experimental setup was a problem in the small vacuum chamber that was used. This would not be a problem in a flight application.

3. Pre-endurance Test Problems

During assembly and initial shakedown test, two minor problems developed relative to the actuators. One was a slight buckling problem associated with the gimbal actuator, resulting in failure to rotate the engine. This problem was solved by bolting an aluminum bar across the two unsupported ends of the lead screw support bracket. No failure to function was encountered with the gimbals during or after the 550 h.

The second problem was a significantly high leak rate in the gimbal cover seal where the O-ring was butted together and epoxied. To solve this problem, the faulty O-ring was replaced with an O-ring specially made with a vulcanized butt end joint. No further leakage problems were encountered during the test.

4. Endurance Test Results

Leak rate tests were performed on the gimbal actuator prior to the test (after the O-ring replacement) and after 250 and 550 h. The results are summarized in Table 2. The results indicate a leak rate within specification with no significant increase with time. No leak-rate check has been performed to date, after the test on the translator actuator. However, a pre-test leak check over a 50-h period indicated a constant leak rate within specification.

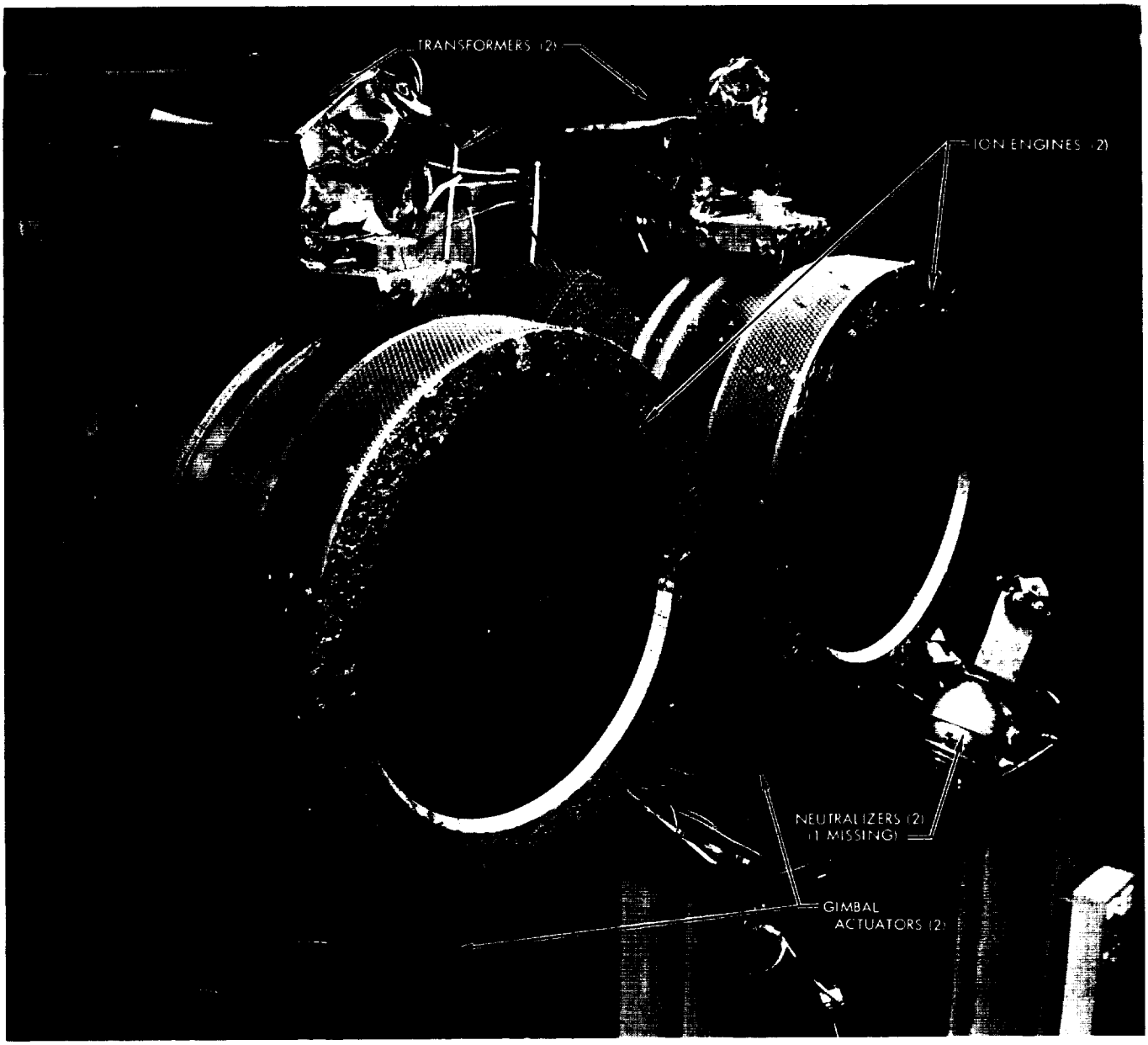


Fig. 7. SEPST-II breadboard system after 250 h of operation

Table 2. Gimbal actuator leak rates

Type	Leak rate of 90% N_2 — 10% He, std cm^3/h		
	Prior to test	After 250 h	After 550 h
Gimbal actuator 1	5.5×10^{-3}	2.7×10^{-3}	7.4×10^{-3}
Gimbal actuator 2	4.0×10^{-3}	3.8×10^{-3}	4.5×10^{-3}

Table 3 summarizes the cycle test results for all three actuators. Table 3 also includes the calculations of expected number of cycles required based on mission requirements. As can be seen in that table, all actuators met, and in two cases significantly surpassed, the total cycle requirement. No failures occurred. The translator actuator would have been further tested had not the endurance test been stopped after 550 h, due to ion engine malfunction.

Table 3. Endurance test—total steps accumulated/actuator

Type	Mission requirements, No. of steps	Accumulated in test, No. of steps
Translation actuator	$<6.35 \times 10^6$	8.30×10^6
Gimbal actuator 1	$<6.35 \times 10^5$	121.25×10^5
Gimbal actuator 2	$<6.35 \times 10^5$	24.0×10^5

Thermistors were mounted in all three actuators to monitor internal temperature. This was primarily due to a concern that the high operating temperatures of the ion engine (~250°C) might cause overheating and resultant malfunction of the stepper motor (<165°C) and pickoff electronics (<135°C), particularly inside the gimbal actuators. A typical temperature profile based on test data is shown in Fig. 8. The thermal analysis as it pertains to the detail design of the actuators is presented in the SPS 37-54, Vol. III. As can be seen in this figure, sufficient thermal protection was achieved to protect the electrical components.

The mechanical end stop pickoff voltages in both directions of travel were monitored for all three actuators periodically during the test. This was used as an indication of actuator null shifts as a function of temperature and time. No such shifts were observed.

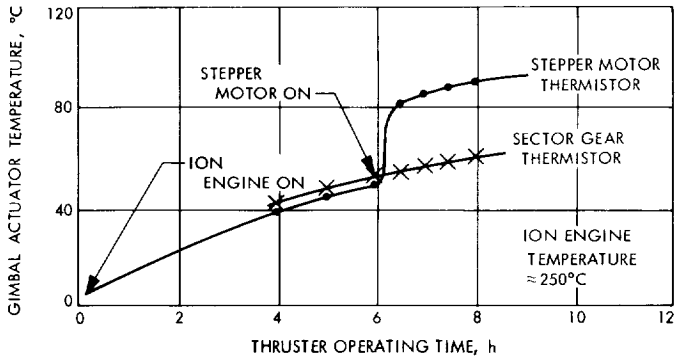


Fig. 8. Gimbal actuator temperature versus time

A draw bar pull test was performed on the platform before the test began and after 250 and 550 h to determine the amount of force necessary to translate the moving platform in the horizontal direction. A force of 3 lb was measured in both the positive and negative direction. This force remained constant throughout the duration of the test indicating that there was negligible degradation in performance of the moving platform roller bearings (which were exposed to the chamber vacuum). Also, since the translator actuator was designed to provide 11-lb pull capability, sufficient force margin existed.

5. Conclusions

The above information summarizes, in brief, the results of the endurance testing. With the completion of these tests, it can be stated that all design objectives were met and in some cases exceeded. A prototyping activity is currently under way on the actuators, concerned primarily with weight reduction and launch vibration compatibility.

Reference

1. Pawlik, E. V., Macie, T., Ferrera, J. D., Paper 69-236. Presented at the AIAA 7th Electric Propulsion Conference, Williamsburg, Va., March 3-6, 1969.

F. Sterilizable Inertial Sensors: High-Performance Accelerometer, *P. J. Hand*

1. Introduction

The objective of this task is to develop a miniature high-performance accelerometer which is capable of surviving the thermal sterilization temperature of 135°C without catastrophic failure or significant performance degradation. This instrument is one of a number of inertial sensors that have been developed by JPL and

JPL contractors for possible application in spacecraft guidance systems which require thermal sterilization. In addition to the thermal sterilization environment, these devices must be capable of surviving the difficult environments of 14-g rms vibration and 200-g peak shock.

The accelerometer chosen for this development is the Bell Aerosystems Company Model VII. This is a flexure-suspended force-balance type, utilizing a capacitance pickoff and employing eddy current damping of the pendulous mass. It can be used with either analog or digital force-balance servo loops.

2. Status

As reported in SPS 37-55, Vol. III, pp. 119–121, the first sterilizable Model VII accelerometer was damaged during bench testing prior to the 200-g shock environment. The second unit, after surviving 6 thermal sterilization cycles at the manufacturers facility, was shipped to JPL for further testing. This unit was subjected to 15 total shocks of 200 g peak, each with a duration of 0.5 $\pm$0.2 ms. The pendulum was constrained by means of an analog capture loop capable of supplying a sustained current equivalent to 180 g of input. During the brief fast-rising shock wave, this capture loop, plus the accelerometer eddy current damping, is effective enough to keep the pendulous mass from striking the mechanical stops. The result of these 15 shocks was a total bias shift of 30 μg. This is quite satisfactory performance for an instrument of this type.

So far, a total of four accelerometers have passed the 6 thermal sterilization cycles without catastrophic failure. The degradation of the bias error term has been well beyond specification limits in all cases, however. Worst-case bias shifts on the first unit were 1098 μg as reported in SPS 37-55, Vol. III. Subsequent shifts in the other three units, which were sterilized at the vendor's facility, were 2792 μg worst case and 86 μg at a minimum.

The results of these tests are shown in Table 4. Only the change from before to after each sterilization cycle is shown. Scale factor stability, which was marginal on the first unit, has been well within specification on all subsequent units. Worst-case scale factor shift was 134 ppm on one unit. The specification limits were a bias shift of 300 μg maximum and scale factor shifts no larger than 500 ppm.

The fifth and last accelerometer of this series is in the final processes of assembly at the vendor. In an effort to eliminate the thermal incompatibility between the coefficient of expansion of the aluminum proof mass assembly and the stainless-steel mounting structure, a proof mass assembly of beryllium is being built into the fifth unit. This should reduce the flexure joint stresses by a factor of 40. Also, a different heat-treating method will be used on the flexures themselves. This unit will also be sterilized at the vendor's facility and shipped to JPL in the near future.

Table 4. Model VII accelerometer sterilization results

Accelerometer	Sterilization cycle					
	1	2	3	4	5	6
SN 659						
Second sterilized unit						
Bias shift, μg	+859	+86	+325	+2320	+1200	+448
Scale factor shift, ppm	−132	+44	−31	−77	+02	+25
SN 660						
Third sterilized unit						
Bias shift, μg	−2792	+505	+847	−641	+2526	−1621
Scale factor shift, ppm	+122	+119	−04	−36	+89	38
SN 657						
Fourth sterilized unit						
Bias shift, μg	+1192	+1769	+384	−1654	+1483	−1408
Scale factor shift, ppm	+63	+134	−27	+12	+74	+42

XII. Guidance and Control Research

GUIDANCE AND CONTROL DIVISION

A. Temperature Control Below Room Temperature With a Commercial Proportional Controller, *A. R. Johnston*

The use of thermistor-controlled proportional-temperature regulators is well known, and many such instruments are available on the commercial market. This article describes a simple alteration to such an instrument which was used in our laboratory to control the temperature of a small chamber in the range from 20 to −40°C. Cooling was provided by cold nitrogen gas boiled off from a reservoir of liquid nitrogen by a resistor connected as the load of the temperature controller, and submerged in the liquid nitrogen. Alternatively, the same controller–thermistor combination could be connected to a heater if an elevated temperature were desired.

The cooling of a specimen by vapor from liquid nitrogen or helium has been a widely used technique (Refs. 1 and 2). In fact, the system described by Angelo (Ref. 1) is very similar to ours. The method described below offers a simple way to accurately control the temperature of a specimen below room temperature with a common proportional-temperature controller which is likely to be already available on the laboratory shelf. Only a minor alteration to the controller is necessary.

Our technique is illustrated in Fig. 1. The thermally controlled chamber was a cylinder approximately 7 cm in diameter and 5 cm in length, insulated by 1 cm of asbestos board. The cold nitrogen gas from the Dewar was led to the chamber through a short piece of ⅜-in. tubing. Our particular controller[1] used a thermistor sensor connected in a transformer-excited ac bridge circuit. The error signal from the bridge was amplified and used to control the power delivered to a load. The value of one of the bridge resistors could be varied with a front-panel control to select the desired temperature setting. The only alteration made to the controller was to provide a

[1]YSI Model 72, manufactured by Yellow Springs Instrument Co., Inc., Yellow Springs, Ohio.

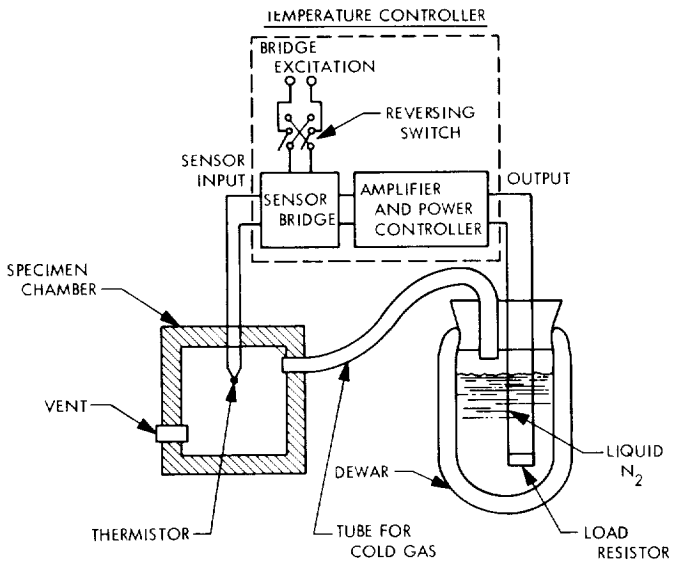


Fig. 1. The apparatus as used for cooling

reversing switch in the thermistor bridge excitation circuit. When cooling, the gain of the temperature-control loop must be inverted. An analogous change, such that power would be delivered as the temperature at the sensor rose above a specified point (the reverse of the situation found in normal operation), could be made in any other type of controller.

It was not necessary to change the thermistor sensor itself although, in general, some trimming of the sensor resistance might be needed. A new calibration curve was of course necessary for setting the controller to a predetermined temperature. The value of load resistance placed in the liquid nitrogen Dewar was chosen so that the controller delivered about 1 W when full on. With the setup shown it was possible to reach −40°C. We obtained a much lower temperature using a different specimen chamber which was designed to fit in the Dewar over the liquid nitrogen, thus obtaining much more effective thermal isolation. Typical temperature stability was ½°C after the initial transient, a stability which is roughly what was observed with the same specimen chamber and controller in the normal heating mode. The time required to stabilize at a newly set temperature was also similar to that required when heating. About ¼ lb/h of liquid nitrogen was consumed at 0°C.

The merit of this method was that it permitted both heating and cooling by the same controller. Only the simple modification to the sensor bridge excitation was required, and the Dewar and liquid nitrogen are cheap and readily obtained.

References

1. Angelo, P. M., Palma, M. U., and Vaiana, G. S., *Rev. Sci. Instr.*, Vol. 38, p. 415, 1967.

2. Huber, J. C., *J. Sci. Instr.*, Vol. 2, Series 2, p. 294, 1969.

B. Photocurrents in MIM Structures, *G. Lewicki and J. Maserjian*

1. Introduction

Photoresponse measurements on thin-film structures consisting of two metal electrodes separated by a thin insulating layer (MIM structures) are currently in progress. This work was undertaken to determine whether the energy-band representation of MIM structures could be defined by such measurements and to gain a quantitative feeling for the photoemission yields obtainable with these structures.

In this article the basic physics behind photoresponse will be described in order that it may be used as a reference in future reports. Some illustrative experimental results will be presented along with a brief discussion of the possible use of MIM structures as photodetectors.

2. Photoexitation Processes in MIM Structures

MIM structures can be discussed in terms of energy-band representations such as that shown in Fig. 2. The

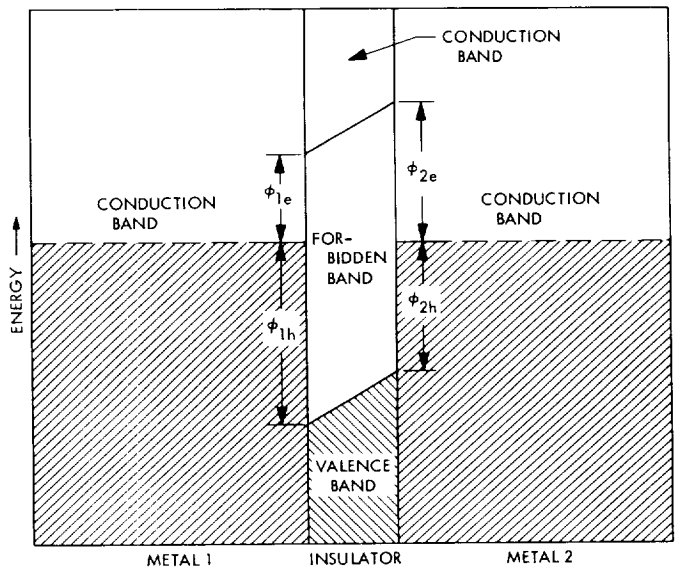


Fig. 2. MIM energy-band representation (hatched areas represent filled electron states)

forbidden band of the insulator acts as an energy barrier to electrons within the metals. Electrons can pass from metal to metal only via the conduction band or the valence band of the insulator. Normally, very few electrons wthin the metals have sufficient energy to propagate through the insulator conduction band while electrons having sufficiently low energies to propagate freely through the insulator valence band find very few unoccupied states (holes) within the opposite metal.

Light incident on the metals and the insulator can excite electrons to sufficiently high energies (or from sufficiently low energies for hole excitation) to allow passage of electrons from metal to metal. That is, light can give rise to photocurrents in MIM structures. The threshold energies for photons giving rise to photocurrents can give information about the energy-band representation of an MIM structure. Photocurrents can be used to define the barrier energies ϕ_{1e}, ϕ_{1h}, ϕ_{2e}, ϕ_{2h} and the insulator forbidden energy gap E_g inasmuch as $\phi_{1e} + \phi_{1h} = \phi_{2e} + \phi_{2h} = E_g$. However, extreme care must be used to interpret photoresponse measurements since, as will be shown, photoresponses can be quite complex.

There are six photoexcitation processes which can take place. Each one of these gives rise to the photocurrent with the dependence

$$Y = I/eS \propto (h\nu - \phi)^2 \tag{1}$$

where photoyield Y is defined as the ratio of photocurrent I to the product of the incident photon flux S and the electron charge e. The quantity $h\nu$ is the photon energy and ϕ is the threshold energy for the process.

All six processes are illustrated in Fig. 3. Electrons can be photoexcited in metal 1 to cross over into metal 2. Conversely, electrons can be photoexcited in metal 2 to cross over into metal 1. The two processes give photocurrents of opposite polarities; the photon threshold energy is the greater of ϕ_{1e} and ϕ_{2e}.

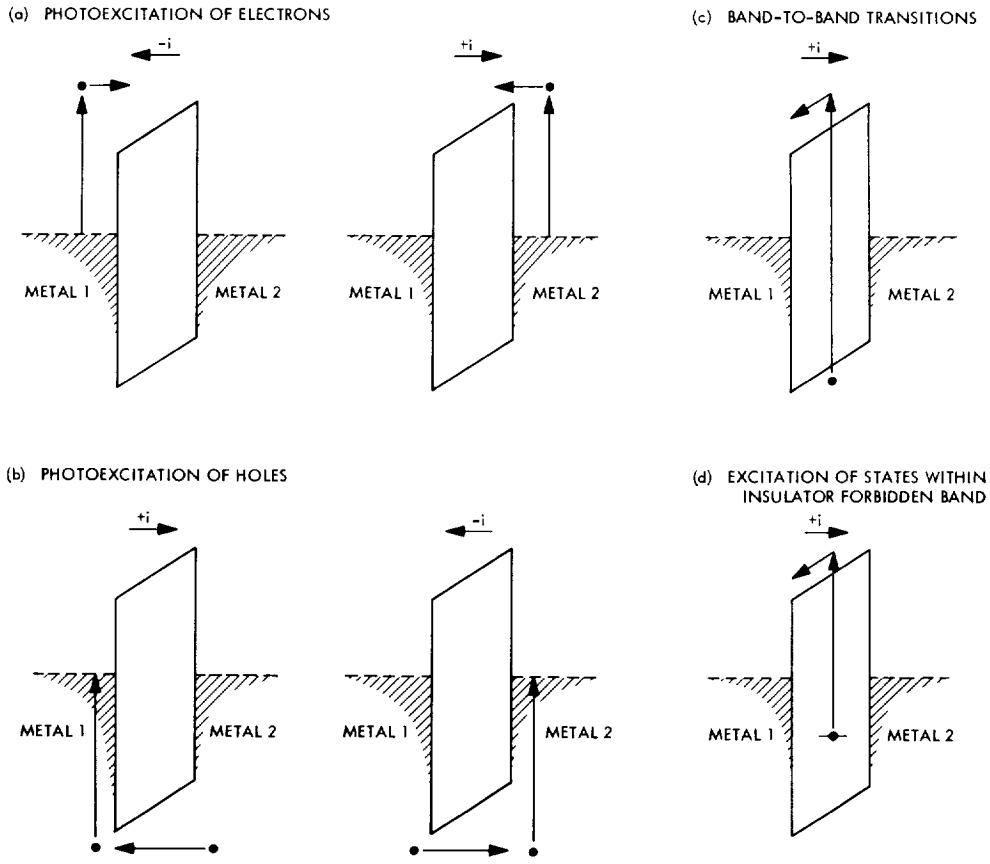


Fig. 3. Various processes giving rise to photocurrents in MIM structures

Electrons can be excited from deep levels on metal 1 leaving behind holes. These holes are filled by electrons originating in metal 2 crossing through the valence band of the insulator (Fig. 3b). This process can be considered as photoexcitation of holes in metal 1, the holes crossing over to metal 2. The inverse process involves holes being photoexcited in metal 2, the holes crossing over into metal 1. These last two processes give rise to photocurrents with opposite polarities with a photon threshold energy equal to the greater of ϕ_{1h} and ϕ_{2h}.

Finally, there is the possibility of electrons being excited from the valence to the conduction band of the insulator, the resulting holes and electrons moving into the metals under the influence of the built-in field within the insulator (Fig. 3c). Electrons could also be excited from discrete states within the insulator (Fig. 3d).

With this brief description, several questions come to mind. How can currents arising from photoexcitation of electrons or holes in the metals be differentiated from those arising from photoexcitation of electrons within the insulator? How can the lower of ϕ_{1e} and ϕ_{2e} be determined in the case of photoexcitation of electrons within the metals and the lower of ϕ_{1h} and ϕ_{2h} in the case of photoexcitation of holes within the metals? Finally, how can currents arising from photoexcitation of holes be distinguished from those arising from photoexcitation of electrons?

Photoexcitation of holes can be distinguished from photoexcitation of electrons in the following manner. Photoresponse measurements can be made on MIM structures having metals with different electronegativities for one of the electrodes. The barrier height ϕ_{2e} increases and ϕ_{2h} decreases as the electronegativity of metal 2 increases. The two values of ϕ that are obtained for each electrode (as described below) must be assigned in a manner consistent with this condition.

Photoexcitation of holes and electrons within the metals can be differentiated from photoexcitation of holes and electrons within the insulator by the fact that threshold photon energies for the former processes vary with voltage applied to the MIM structure while the photon threshold energies for the latter processes do not. This effect can be seen in Fig. 4, where the excitation processes are illustrated with voltages applied to the MIM structures.

Application of voltage to the structure also allows the determination of *both* ϕ_{1e} and ϕ_{2e} in the case of electron

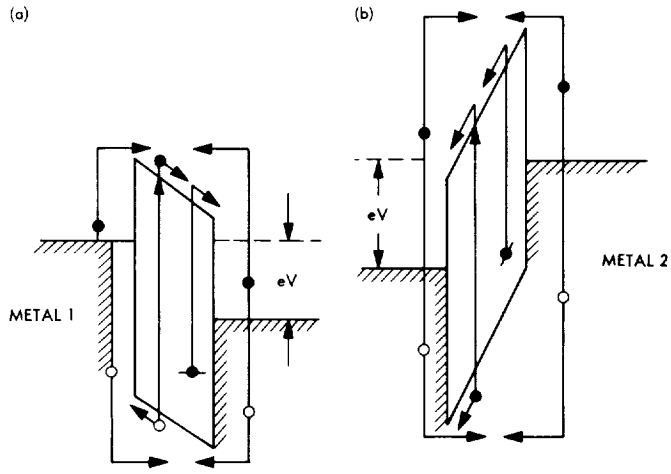


Fig. 4. Excitation processes with: (a) positive voltage applied to metal 2; (b) positive voltage applied to metal 1

excitations in the metals, and *both* ϕ_{1h} and ϕ_{2h} in the case of photoexcitation of holes in the metals. The variation of photon threshold energies with voltage for all processes is shown in Fig. 5. When both electrons and holes are simultaneously involved, the determination can be obtained as follows: For large positive voltages applied to metal 2, photocurrents resulting from photoexcitations of electrons in metal 2 and holes in metal 1 can be essentially eliminated by moving their photon threshold energies to very large values, leaving behind photocurrents defining ϕ_{1e} and ϕ_{2h}. From measurements of the total yield Y vs $h\nu$, the relative contributions can be determined for these two remaining processes. For large negative voltages applied to metal 2, photocurrents resulting from photoexcitation of electrons in metal 1 and holes in metal 2 can be removed by moving their threshold photon

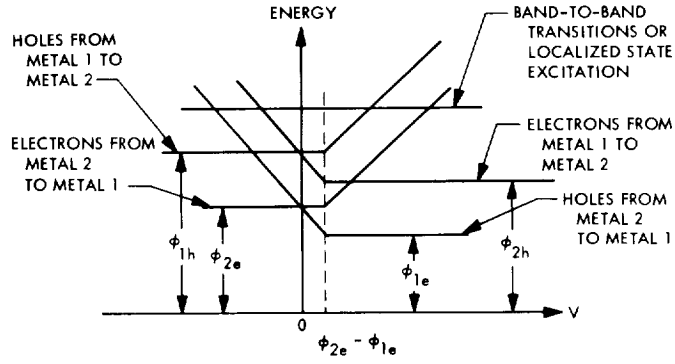


Fig. 5. Threshold energies for various excitations as a function of voltage V applied to metal

energies to very high values, leaving behind photocurrents defining ϕ_{2e} and ϕ_{1h}. Again, these two remaining processes can be separated from the Y-vs-$h\nu$ data.

An unambiguous determination of barrier energies in MIM structures requires that photocurrents be measured as a function of $h\nu$, applied voltage, and also electronegativity of one of the metals used as an electrode. An additional consistency check is given by the condition

$$E_g = \phi_{1e} + \phi_{1h}$$
$$= \phi_{2e} + \phi_{2h}$$

The extensive net of measurements that is required is justified by the information that can be obtained from it.

3. Some Results of Current Photoresponse Investigations

Photocurrents in thin films of Al–AlN–Mg, Al–AlN–Al, Al–AlN–Pb, and Al–AlN–Au are currently being investigated. All photoexcitation processes described above have been observed except band-to-band excitations which were above the range of our monochromator. An unambiguous determination of the energy-band representation for this structure is almost complete. This representation will be presented for publication in the open literature. The result of a typical experiment is shown in Fig. 6, where the square root of the photoyield is plotted as a function of the energy of incident photons (to obtain a linear dependence) for different voltages applied to the gold electrode of an Al–AlN–Au structure. Two excitation processes are present. Both processes correspond to a 3.025-eV threshold at zero voltage. These data, along with data from structures with different electrodes, identify the two processes as photoexcitations of holes in the gold electrode for the positive response (corresponding to barrier ϕ_{2h} in Fig. 2), and photoexcitations of holes in the aluminum for the negative response (corresponding to barrier ϕ_{1h} in Fig. 2).

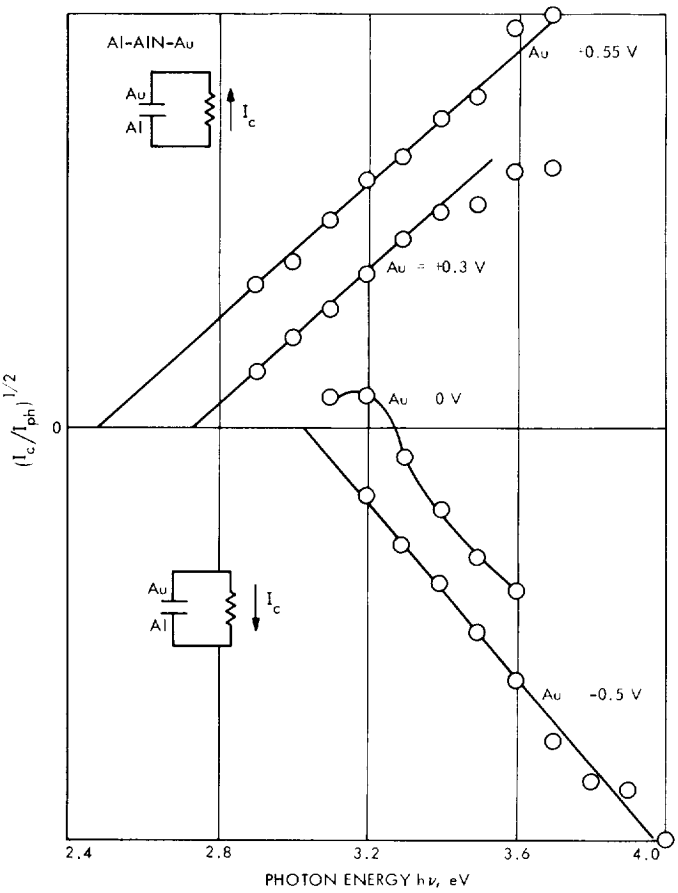


Fig. 6. Photoresponse of Al–AlN–Au structure for different voltages applied to Au electrode

4. Possible Application of MIM Structures

Photoyields approaching 1% at 5.0-eV photon energies have been observed in structures based on AlN as the insulator. This observation raises the question as to whether these structures could be used as photodetectors. On the negative side is the fact that most quantum detectors have quantum efficiencies or photoyields on the order of 20%. It is doubtful that photoyields with MIM structures would ever reach this efficiency. Nevertheless, different counterelectrodes are being placed on the Al–AlN structures to see if there are metals which give larger yields. It is also possible that higher efficiencies can be obtained with other insulating films.

On the positive side, the structures can be made as high-impedance devices with impedances greater than 10^8 Ω. It must be remembered that the signal is proportional to the photocurrent times impedance while noise is proportional to the square root of the impedance. Furthermore, the spectral response of the structures can be varied widely by application of voltage; consequently, they could be used as spectrometers. This characteristic is not available with any other kind of quantum detector.

MIM structures based on insulating films of AlN could only be used as UV detectors because of the high threshold energies involved for photoexcitation processes. In the future, MIM structures based on different insulating films will be investigated. These structures will have lower-energy thresholds for photoexcitation processes.

C. Metal–Cadmium Sulfide Work Functions at Higher Current Densities, *R. J. Stirn*

1. Introduction

Recently, the work function between certain metals and photoconducting cadmium sulfide (CdS) has been found to have quite different values and properties from the work function as measured on semiconducting CdS (Refs. 1 and 2). The results were obtained from an analysis of stationary high-field domains in the range of negative differential conductivity (Ref. 3). The theory involved and some preliminary experimental results were presented in SPS 37-54, Vol. III, pp. 73–82.

It was found that the measured currents were orders of magnitude larger than the thermionic currents calculated for a work function (i.e., barrier height) as determined for a zero-current condition. The analysis of the high-field domains in CdS indicated that tunneling through the potential barrier could not be of major importance. Indeed, although we will show results which were obtained directly from the domain analysis (as explained in SPS 37-54, Vol. III), we will also show results obtained from current–voltage characteristics and current kinetic measurements which are consistent with the model used in the domain analysis. The presence of the high-field domain is not really required then, except to show that fields high enough to allow for tunneling through the barrier at the cathode are not possible.

2. High-Field Domain Results

The experimental setup and the method of analysis of the data at high applied voltages were given in SPS 37-54, Vol. III. A comparison of the barrier heights so obtained was made with published values of the barrier heights obtained from conventional photoresponse and differential capacitance techniques at zero-current flow in SPS 37-53, Vol. III, pp. 68–71. Briefly, the barrier heights measured on photoconducting CdS were about 40–50% lower than those on normal semiconducting CdS, and instead of the situation found in the latter, we found a large temperature and current-density dependence. For convenience, a figure showing the dependence on photon flux density (i.e., current density) is repeated here (Fig. 7).[2] In this figure, which includes only Au and Cu for purposes of clarity, one can also see the large changes in barrier height with temperature (at a constant flux density). For comparison, the barrier height for gold on semiconducting CdS at zero-current flow at room temperature is 0.80 eV, while for copper it is 0.35 eV. In addition, these latter values are essentially independent of temperature.

The large temperature dependence of the barrier height ψ^* is required because of our equating the measured current with the thermionic current over the barrier (thus ruling out any tunneling contribution). The saturated current density, which is proportional to $\exp(-\psi^*/kT)$, is found to be nearly independent of temperature. Thus, because of the thermionic current model used, ψ^* must be nearly proportional to the temperature T.

The temperature dependence of the current density for five metal cathodes is shown in Fig. 8 for a constant photon flux density of 5×10^{13} $cm^{-2}s^{-1}$. The high-temperature dropoff of the saturated current densities is due to thermal quenching in the photoconductor (thermal release of holes from traps which then recombine with free electrons). This can be shown by comparing the curves with the dashed curve showing the temperature dependence of the carrier concentration n_I in the bulk crystal *outside* of the high-field domain. This curve (averaged for five samples) was obtained at low applied voltages from the relation

$$j = n_I e \mu E_I \qquad (1)$$

where j is the current density in the *forward* direction,[3] e is the electrode charge, μ is the electron mobility in CdS, and E_I is the field in the crystal taken to equal the applied voltage divided by the electrode separation. It is seen that the n_I curve has a similar convex nature, thus showing that the slight dropoff is due to a bulk effect. Thermal quenching is the only bulk effect of importance in photoconductors.

The low-temperature dropoff is due to the normal freeze-out of carriers. The unexpected rise at higher temperatures for the sample with a gold cathode is unexplained at this time. Thermal quenching does not seem to be present in that sample.

The temperature dependence of the barrier height is shown more explicitly in Fig. 9 for the six metals investigated. Also, to show the individuality of the metals

[2]The power-of-ten factor in SPS 37-54, Vol. III, p. 82 (Fig. 12), should read 10^{12} instead of 10^{14}.

[3]All samples had an ohmic contact of indium on one side of the crystal. Thus, the forward direction (for the opposite contact) is defined with a negative voltage on the indium contact.

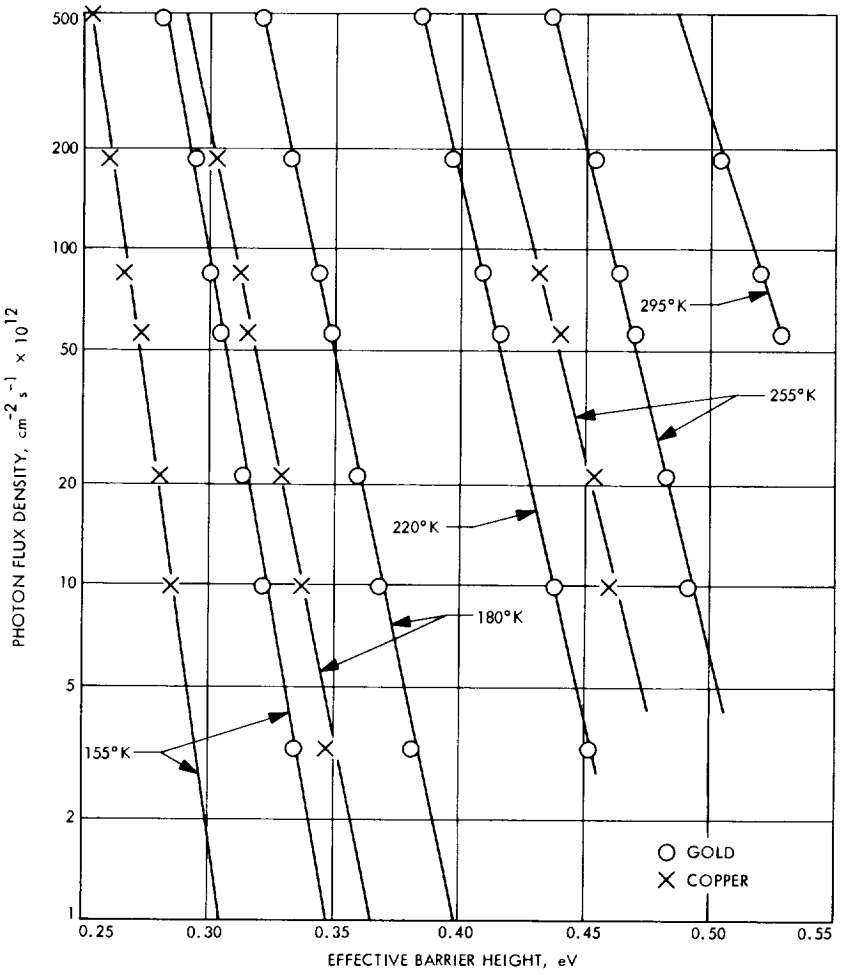


Fig. 7. Barrier-height light dependence for Au and Cu on photoconducting CdS

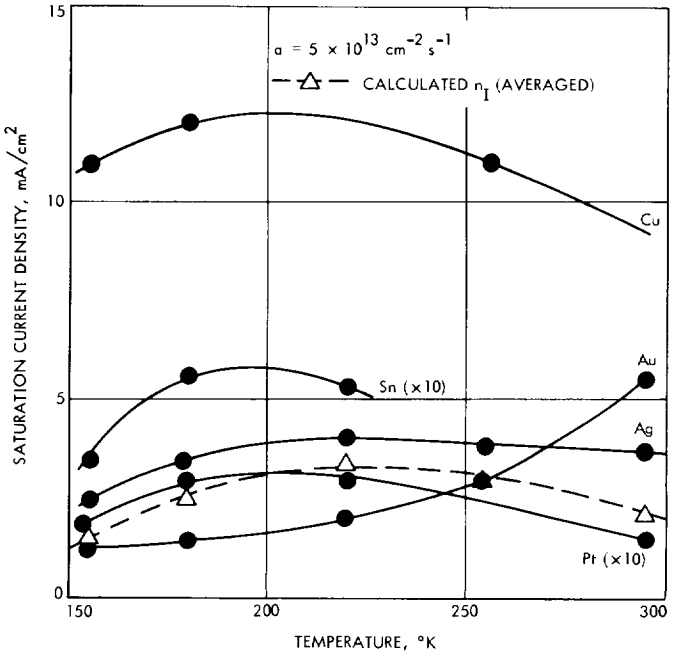


Fig. 8. Temperature dependence of saturation current densities and averaged low-field bulk carrier concentration

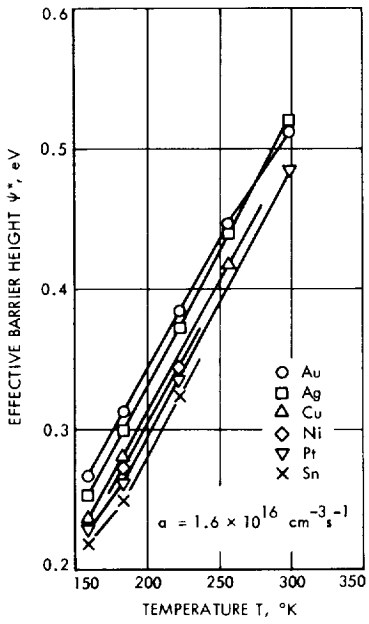


Fig. 9. Barrier height for different metal cathodes (In anode) as a function of temperature

more clearly, the variation of ψ^* with light intensity (photocurrent) is shown in Fig. 10 for one temperature. The differences in values of ψ^* between metals are only hundredths of an electron volt, however, as compared to tenths of an electron volt for the zero-current case. The significance of the behavior is not understood at this time.

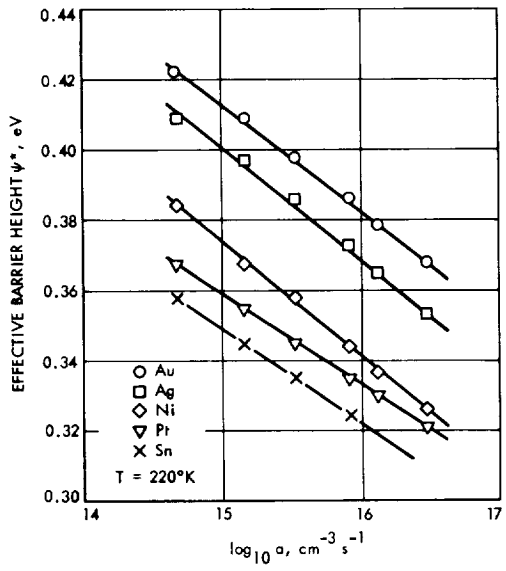


Fig. 10. Barrier height for different metal cathodes (In anode) as a function of optical excitation density

Several other figures can be presented now which show quite well the consistency in the theory of the high-field-domain analysis. First, it is necessary to repeat one figure from SPS 37-54, Vol. III, p. 77. This figure (Fig. 11) shows how the electric field (Fig. 11c) and carrier-density (Fig. 11b) distributions and parts of the potential barrier (Fig. 11d) tie in with critical points of Fig. 11a. In Fig. 11a, the linear curve n_2 with a slope of minus unity ($n \propto 1/E$) represents the current transport equation, given by Eq. (1), since diffusion current is negligible in CdS. The curve n_1 represents Poisson's equation and shows the decrease in carrier concentration with field due to enhanced hole recombination. The critical solution of the two equations at low fields (point I) gives the intrinsic bulk concentration n_i (referred to as n_I above) at some field strength E_I, which is the field outside of the high-field-domain region. The second critical point (II) gives the concentration n_{II} within the domain (equal to n_c, the concentration at the barrier top, within a factor of 2 to 14, depending on the temperature) at the value of the field E_{II} within the domain.[4] SPS 37-54, Vol. III, should be consulted for details. Figures 12–14 present data which are very consistent with the model represented by Fig. 11a. These figures show the measured values of n_I, E_I, n_{II} and E_{II}—data which are readily accessible. A linear line of minus unity slope is drawn in to connect the pairs of data, except that in these figures, the mobility is taken to be independent of electric field only up to 40 kV/cm. Above this value, μ is assumed to be proportional to $1/E$ as expected for hot-carrier scattering from optical phonons. This mobility dependence has been experimentally verified (Ref. 4), and the effect is to make the n_2 curve in Fig. 11 horizontal above 40 kV/cm. Figure 12 gives the critical solution points at I and II for the five temperatures investigated for the case of a gold contact and a constant light intensity. Figure 13 shows the differences for the six metals investigated at a constant temperature and light intensity, while Fig. 14 compares different photon flux densities for the case of a silver contact at room temperature. In the latter, the points at I (low field) fall somewhat below the lines representing the n_2 curves. An explanation for this might be that, at the higher temperature used here, thermal quenching has more effect on the concentration in the bulk than in the narrow region at the cathode barrier because of the smaller available number of hole traps in the latter region. The dashed curves indicate apparent but unexplained trends.

[4]The third solution point (III) involves anode-adjacent high-field domains at even higher applied voltages which were purposely not formed in this investigation.

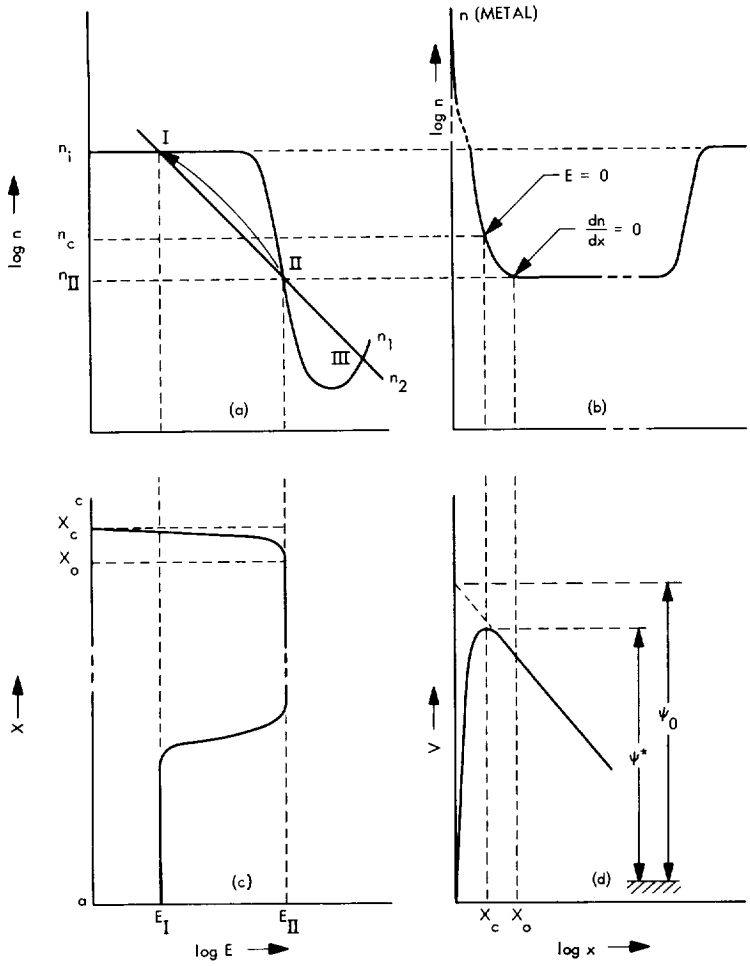


Fig. 11. Carrier concentration vs x and E, electric field vs x, and the potential barrier near the cathode

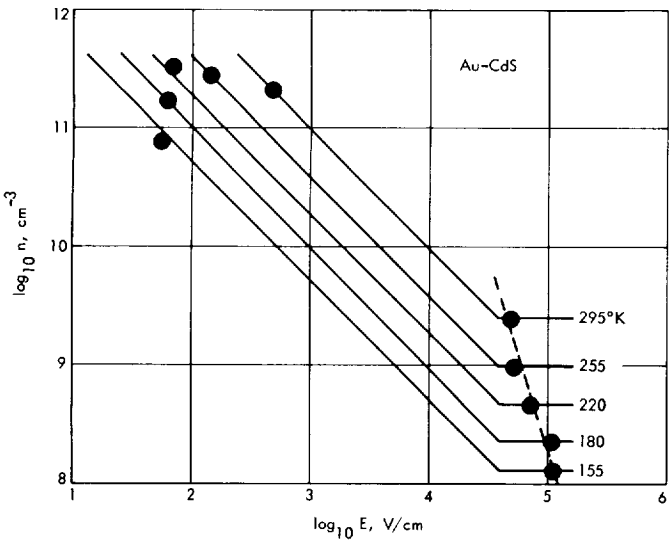


Fig. 12. Temperature dependence of (n_I, E_I) and (n_{II}, E_{II}) singular points (see Fig. 11a)

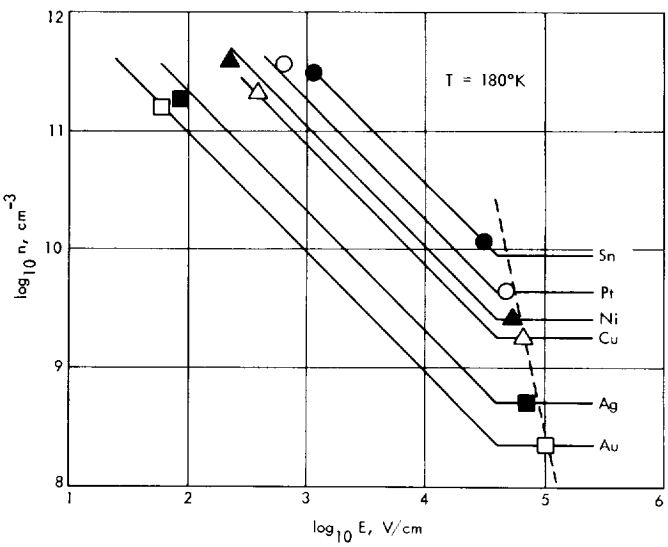


Fig. 13. Variation of (n_I, E_I) and (n_{II}, E_{II}) singular points with metal cathode (see Fig. 11a)

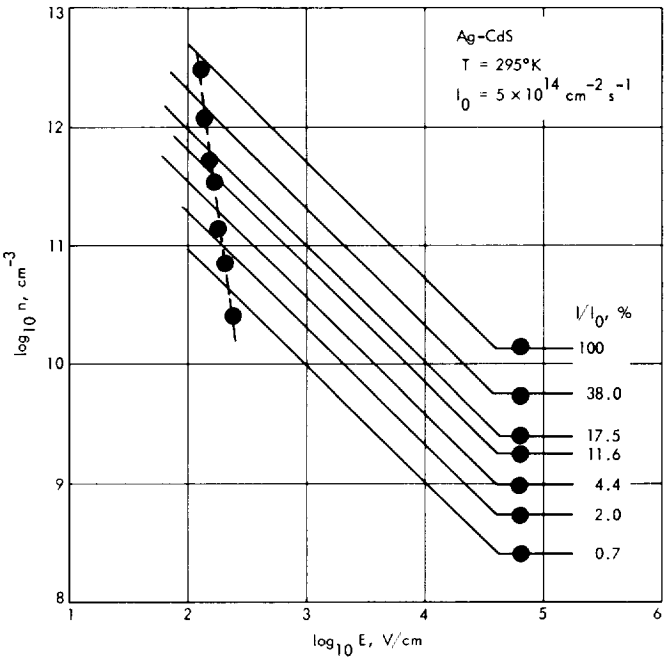


Fig. 14. Light excitation dependence of (n_{II}, E_I) and (n_{II}, E_{II}) singular points (see Fig. 11a)

3. Results at Low Applied Voltages

High-field-domain formation, and hence current saturation, normally occur above about 100–200 V. Below this voltage, values of the work function as a function of voltage between some of the metals and CdS were estimated from the current–voltage characteristics by assuming the current to be thermionic current over the reduced barrier. This must be so at higher voltages because the domain analysis does not allow fields greater than E_{II} (30–120 kV/cm) to be present at the barrier, and thus does not allow for tunneling. At low applied voltages ($V < 0.1$ V), the available voltage drop is too small to allow sufficient tunneling through the barrier for the observed current density. This can be seen from the Poisson equation and the availability of space charge (most probably from electron traps between the Fermi and quasi-Fermi level). Sufficient tunneling in the intermediate voltage range is highly improbable because of the otherwise non-monotonic behavior of the current mechanism.

The assumption of inhomogeneous contacts, and thus current patches, cannot explain the observed high-current densities since a well-reproducible ratio of "patchiness" for the different metals is highly improbable. Thus, a real lowering of the barrier must be assumed.

The measured current density, at not too low an applied voltage, is orders of magnitude larger than the thermionic current given by

$$j_{th} = \frac{e\, v_{th}}{(6\pi)^{1/2}} N_c \exp\left(-\frac{\psi^*}{kT}\right) \qquad (2)$$

where v_{th} is the thermal velocity of electrons in the CdS, N_c is the effective density-of-states at the conduction band edge, k is Boltzmann's constant, T the temperature in degrees Kelvin, and the barrier height ψ^* is the value obtained in the zero-current-density limit, e.g., $\psi_0^* = 0.8$ eV for Au. Thus, as long as our measured current density is larger than $j_{th}(\psi_0^*)$, we will formally equate the thermionic current with the measured current and thereby obtain a *current-dependent* effective barrier height. This barrier height is plotted in Fig. 15 as a function of applied voltages for three of the metals at a temperature of 155°K. The light intensity used was the highest available at 492 mμ from the monochromator so as to measure the current at as low a voltage as possible (~ 20 mV). The high voltage values do extrapolate to the values of barrier heights determined from the high-field-domain analysis. Presumably, in the limit of zero voltage, the low-voltage values would approach the zero-current values as measured by photoresponse. An attempt is now being made to measure the barrier heights on one of these samples as a function of low applied voltages by the technique of photoresponse. Very serious problems to be overcome are the high impedance of these samples and the contribution of photocurrents from traps in the bulk. The latter response, greatly larger in magnitude, should not

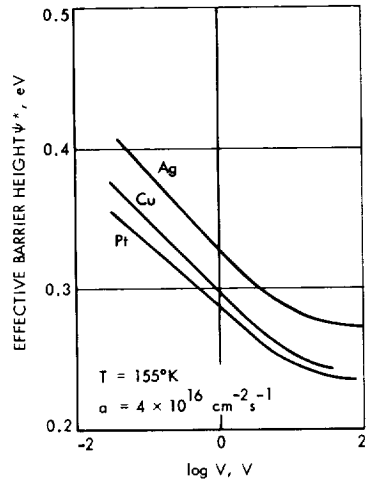


Fig. 15. Barrier height for different metal cathodes (In anode) as a function of applied voltage

show a voltage dependence as would the response from the metal contact.

4. Current Kinetics

The high-field domain and the intermediate voltage results force one to accept a marked lowering of the work function from about 0.8 eV for zero-current densities (for Au and Pt) to below 0.3 eV (at 155°K). Such lowering can be explained by a change in the dipole part of the interface, due to redistribution of electrons and holes very close to the metal contact. Steepening this distribution by carrying holes from the CdS side and electrons from the metal side into this region will lower the work function.

To check this possibility, the kinetics of the current flow upon sudden application of a voltage step were investigated. A voltage step was applied to the samples using a mercury-wetted relay switch free of jitter. The current through the sample was monitored by measuring the voltage drop across a 100-kΩ series resistor with a Tektronix storage oscilloscope having a 1-$M\Omega$ input resistance. The circuit had a total capacity less than 80 pF. The current value at any given time could be monitored by varying the sweep rate. The results for an applied voltage of 86 V (below that required for domain formation) at 155°K are shown in Fig. 16 for a relatively low incident light intensity. Since the indium contact cannot be expected to be a perfectly injecting electrode, it probably has a slight barrier associated with it causing the structure in the forward current (indium negative). Note that changes in current in the reverse direction (when the silver contact is negative, and thus blocking) are more than *three orders of magnitude* as compared to a factor of only three in the forward direction. A possible explanation for the shape of the curves in Fig. 16, particularly the unexpected overshoot between 0.1 to 10 ms, will now be presented.

From the equilibrium current density in the forward direction, we can estimate a bulk conductivity σ of $5 \times 10^{-5}\ \Omega^{-1}\text{cm}^{-1}$ and a carrier density of $n = 3 \times 10^{11}\ \text{cm}^{-3}$, while for an initial barrier height of 0.55 eV (for silver), the carrier density in the barrier region is about 10 orders of magnitude smaller. Initially, the current must decrease essentially with the dielectric relaxation time $\tau_1 = \epsilon\,\epsilon_0/\sigma$, since the barrier is completely blocking. Here, ϵ is the dielectric constant and ϵ_0 the permittivity of free space. For the conductivity estimated above, τ_1 is on the order of 2 μs and, thus, is masked by the RC constant of the circuit which is about that magnitude. This decrease in

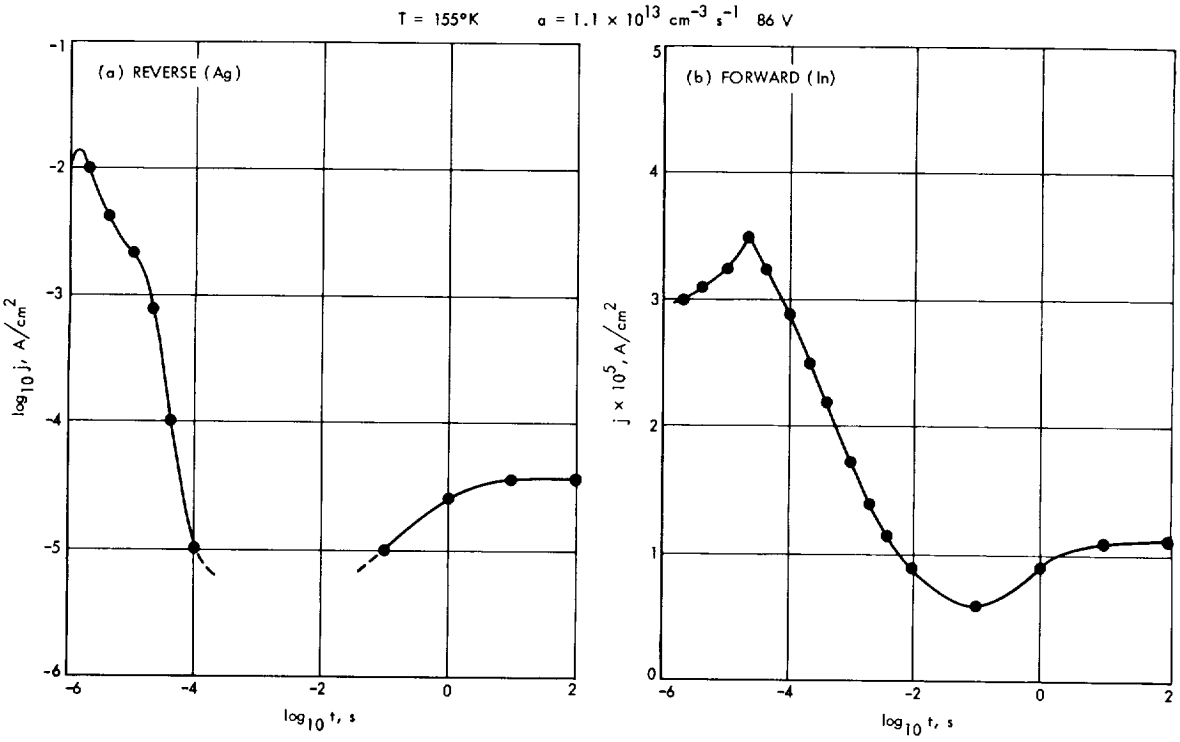


Fig. 16. Time-dependent current for silver and indium cathodes

current forces the field at the barrier to increase as electrons are swept out from the conduction band and electron traps, causing a positive space-charge region to build up. While the contact is completely blocking, the current is determined by this release of electrons in the region of developing space charge of width δx, and is given by

$$j = e\,\delta x\, n_t/\tau_t \tag{3}$$

where n_t is the density of trapped electrons which can be released with a relaxation time τ_t.

The space-charge density, and hence the field at the cathode, increases with time. The increase in field is required because, in the region of total depletion (width Δx), the current is given entirely by displacement current. The corresponding increase in voltage drop across $\Delta x + \delta x$ forces the field in the bulk, and therefore the current, to decrease. The field at the cathode can only increase until it is large enough to release holes from hole-traps. The exhaustion of holes reduces the space charge, limiting the field at the cathode while forming, temporarily, what we could call a quasi-domain. Since the field is no longer increasing, the displacement current vanishes in Δx, and the current must be carried by holes. The current density observed in our sample with an Ag cathode decreases from 10^{-1} A/cm^2 (obtained from the maximum observed current in the forward direction) to $\leq 10^{-5}$ A/cm^2.

With the passage of time, the holes moving toward the cathode lower the barrier, thereby increasing the thermionic contribution of electrons moving in the opposite direction. These electrons produce a negative space-charge density close to the cathode and thereby reduce the field there. The total current increases due to the increased field in the bulk. This increase is experimentally observed.[5]

The final current density is controlled by the field distribution in the crystal, which is given by the electrical properties of the CdS crystal and the chemical nature of the metal.

We can estimate the times involved in the model proposed and compare them with the observed kinetics by the following considerations.

[5]That portion of the curve not drawn in represents current which could not be measured because of the noise level in the system.

In the displacement-current range, the current density is given by

$$j = \epsilon\,\epsilon_0\,\Delta E/\Delta t \tag{4}$$

which yields a field change $\Delta E \simeq 10^5$ V/cm in the space-charge layer of width δx from the measured time constant $\Delta t \gtrsim 10^{-4}$ s and for $j \leq 10^{-3}$ A/cm^2. The space-charge layer moves with a velocity $v = \delta x/\Delta t$. The width δx can be estimated to be $\approx 10^{-4}$ cm from the Poisson equation and the assumption of a reasonable space-charge density of 10^{16} cm^{-3}. This value yields a velocity of 1 cm/s for expanding the quasi-domain. From the width of the final quasi-domain ($\Delta x \approx V_{\text{applied}}/\Delta E$), one estimates a time constant after which the quasi-domain is fully developed and the current must start to increase again of $\tau_D = \Delta x/v \approx 10^{-3}$ s. Although the beginning of the current increase could not be observed, the order of magnitude of τ_D as calculated is commensurate with the experiment.

Also, with the estimated value of δx and the use of Eq. (3), we obtain $n_t/\tau_t \lesssim 6 \times 10^{18}$ cm^{-3}s^{-1}, yielding $\tau_t \gtrsim 10^{-3}$ s. The trap relaxation time τ_t is related to α_t, the thermal release probability of electrons from traps, by

$$\tau_t = \alpha_t^{-1} = (\alpha_t^*)^{-1} \exp\left(\frac{E_c - E_t}{kT}\right) \tag{5}$$

where α_t^* is the frequency factor $\approx 10^{13}$s^{-1}. With the value of τ_t obtained from the estimated value of δx, we get a trap depth of $E_c - E_t \approx 0.28$ eV. This is in good agreement with thermally stimulated current curves, which show a pronounced electron trap at about 0.38 eV with a density of about 10^{16} cm^{-3}, as assumed earlier. The latter value would be expected to be lowered by about a tenth of an electron volt because of the high fields at the barrier (Poole–Frenkel effect).

5. Conclusions

An analysis of the high-field domains shows that tunneling cannot be dominant for the CdS crystals investigated here. Rather, a marked lowering of the barrier height between the metal and CdS must have taken place upon application of the applied voltage in order to account for the large values of photocurrent and large gain factors. This lowering could be caused by a sufficient transport of holes into the barrier region, which in turn causes a substantial redistribution of charges between the top of the barrier and the actual metal–CdS interface. Such redistribution of carriers close to the boundary could cause a substantial change in the dipole part of

the metal–CdS work function. Evidence of such redistribution has been obtained by observing the current kinetics upon sudden application of the applied voltage.

It should be pointed out that in most other materials, specifically semiconductors of higher conductivity, conditions will not allow barrier-height lowering. In that case, doping or defect level distribution near the Fermi level is sufficient to allow enough storage of positive space charge for creation of fields large enough to cause tunneling ($\gtrsim 10^6$ V/cm).

References

1. Böer, K. W., Dussel, G. A., and Voss, P., "Experimental Evidence for a Reduction of the Work Function of Blocking Gold Contacts with Increasing Photocurrents in CdS," *Phys. Rev.*, Vol. 179, p. 703, 1969.
2. Stirn, R. J., Böer, K. W., Dussell, G. A., and Voss, P., "Metal–CdS Work Functions at Higher Current Densities" Third International Conference on Photoconductivity, Stanford University, Aug. 12–15, 1969; Proceedings to appear in *J. Phys. Chem. Solids*, Dec. 1969.
3. Böer, K. W., and Voss, P., "Stationary High-Field Domains in the Range of Negative Differential Conductivity in CdS Single Crystals," *Phys. Rev.*, Vol. 171, p. 899, 1968.
4. Böer, K. W., and Bogus, K., "Electron Mobility of CdS at High Electric Fields" *Phys. Rev.*, Vol. 176, p. 899, 1968.

D. Thermal Noise in Space-Charge-Limited Solid-State Diodes, *A. Shumka*

A theoretical analysis was made for thermal noise in space-charge-limited (SCL) solid-state diodes by using Langevin's equation to calculate the spectral intensity S_V for the open-circuit noise voltage and the spectral intensity S_i for the short-circuit noise current. Our analysis has made it possible to resolve the differences between two existing theories (Refs. 1 and 2) by uncovering a discrepancy in one of them (Ref. 2).

It is generally accepted that the high-frequency noise in an SCL solid-state diode is of thermal origin and is usually referred to as limiting noise. But the theoretical description of this noise has been a subject of controversy as evidenced by the existence of several differing theories (Refs. 1–3). In particular, if we relate these various theories to $S_V = 4\,kT\,R_n$, where R_n is the equivalent noise resistance, then we obtain for R_n the different expressions $r(\omega)$, $2\,r(\omega)$, and $(2/3)\,r(\omega)$, where $r(\omega)$ is the differential resistance. Published measurements on the limiting noise in various solid-state diodes have been marginal, largely because of an appreciable $1/f$ component of noise and have failed to resolve the controversy.

Since a preliminary review of each theory did not disclose any major flaw which would invalidate it, we decided to perform a careful and detailed treatment of the noise theory in an attempt to establish a basis on which a comparison between the various theories could be made. To do this, we solved Langevin's equation to determine S_V and S_i and their interrelationship. Our results, which will be published in detail elsewhere, are summarized here:

$$S_V = 4\,kT\,2\,r(\omega) \tag{1}$$

$$S_i = 4\,kT\,2\,r(\omega)/\,|\,Z(\omega)\,|^2 \tag{2}$$

and

$$S_i = S_V/\,|\,Z(\omega)\,|^2 \tag{3}$$

where $|\,Z(\omega)\,|$ is the magnitude of the impedance of the solid-state diode. Equation (3) clearly shows the coupling between S_V and S_i, and is in general agreement with the fluctuation–dissipation theory (Ref. 4). Equation (1) agrees with van der Ziel's result (Ref. 1), while Eq. (2) disagrees with Klaassen's result (Ref. 2), which has $S_i = 4\,kT\,(2/3)\,r(\omega)/\,|\,Z(\omega)\,|^2$. A comparative analysis reveals that Klaassen in his theory arbitrarily assumes the fluctuating voltage to have a direct dependence on the fluctuating charge density within the solid-state diode. This assumption is in conflict with Poisson's equation, according to which the second derivative of the fluctuating voltage with respect to position is directly dependent on the fluctuating charge density. On this basis we can rule out Klaassen's result.

The result of Webb and Wright (Ref. 3), according to which $S_V = 4\ kT\ r(\omega)$, was arrived at by applying Nyquist's theorem directly. This result does not consider the nonlinearities of the solid-state diode, and, on this basis, can be discarded (Ref. 5).

References

1. van der Ziel, A., *Solid-State Electron*, Vol. 9, p. 1139, 1966.
2. Klaassen, F. M., *Solid-State Electron*, Vol. 11, p. 377, 1968.
3. Webb, P. W., and Wright, G. T., *J. Brit. Inst. Radio Eng.*, Vol. 23, p. 111, 1962.
4. Callen, H. B., and Welton, T. A., *Phys. Rev.*, Vol. 83, p. 34, 1951.
5. Grafov, B. M., and Levich, V. G., *Sov. Phys.–JETP*, Vol. 27, p. 507, 1968.

XIII. Materials

ENGINEERING MECHANICS DIVISION

A. Jupiter Entry Heat Shield, *W. Jaworski*

1. Introduction

Recent studies on Jupiter entry probes (Refs. 1, 2) point out considerable difficulties in providing an adequate thermal protection for scientific payloads due to excessive heat shield weight requirements. This is because the Jupiter entry velocities are on the order of 60 km/s, resulting in an extremely high radiative heating flux (up to 2,000,000 Btu/ft^2-s) generated by the aerodynamic shock, which would cause a rather rapid consumption of heat shield material produced by intense surface sublimation.

A closer examination of the methods of analysis used in the forementioned studies indicates that certain important factors which might have significant bearing on the results have not been accounted for—namely, the absorption and reradiation of radiative heating flux by evolved gaseous products of sublimation.

The objectives of this investigation were to bring these pertinent factors into focus and to evaluate their influence on heat shield performance for a simplified Jupiter entry case.

2. Analytical Approach and Procedure

The available ablation computer programs within NASA and the aerospace industries were found inadequate to handle this type of radiative heat transfer, specifically where multilayer gas mixtures of different origin are involved. The Sputter computer program, operated by Gulf General Atomic, Inc., (Ref. 3) used in this analysis fulfills adequately these needs. However, since it was not designed to handle planetary entry cases, some modification was necessary.

Through a contractual arrangement with Gulf General Atomic, Inc., a two-phase plan was drawn up to accomplish the stipulated objectives. The first phase represented a test case to ensure that the program could accept given altitude, ambient density, and velocity profiles as a function of time, and to produce corresponding radiative fluxes and pressures produced by the presence of a body in the gas stream. The hydrodynamic shocks were allowed to reflect freely from a flat graphite wall. The second phase provided for additional constraints. It allowed only a single stationary shock to be formed at some distance from the flat graphite surface, and the atmospheric as well as sublimation product gases (between the shock and the body) were depleted sidewise

so that a time-dependent mass flow equilibrium was maintained. These constraints were considered to adequately simulate stagnation point conditions.

The input trajectory data, used for both phases, is given in Table 1, and the Jupiter atmosphere was assumed to consist of 61% hydrogen, 36% helium, and 3% neon. This information was taken from Ref. 4 in order that a comparison could be made of the radiative fluxes (at some initial point) calculated by the Sputter program method as opposed to the method used in Ref. 4. The said trajectory represents a normal entry ($\gamma_E = -90$ deg) at a ballistic coefficient M/C_DA of 1.98 slugs/ft.[2] The shock stand-off for the second phase of the effort was assumed to be 5.5 cm (an approximation to values given in Ref. 4). The graphite plate[1] used is of pyrolytic variety having a density of 140 lb/ft.[3]

Table 1. Jupiter trajectory for an entry velocity V_E of 60 km/s at entry angle γ_E of −90 deg

Time, s	Altitude, km	Flight velocity,[a] km/s	Atmospheric density,[a] g/cm³
0	250.0	60.0	4.53×10^{-10}
1.668	150.0	59.8	1.17×10^{-7}
2.005	130.0	59.4	3.56×10^{-7}
2.344	110.0	58.2	1.08×10^{-6}
2.698	90.0	54.6	3.28×10^{-6}
3.095	70.0	45.0	1.00×10^{-5}
3.341	60.0	36.28	1.74×10^{-5}
3.669	50.0	24.98	3.03×10^{-5}
4.217	40.0	13.02	5.38×10^{-5}
5.000	32.0	5.55	8.10×10^{-5}

[a]Velocity and density are free-stream values.

The first phase confirmed the ability of the Sputter program to generate the necessary data from the trajectory input information given in Table 1. The required output from the second phase called for incoming radiation heat flux generated by the shock, radiative heat flux absorbed by compressed atmospheric gas mixture, radiative heat flux absorbed and reradiated by carbon vapor, and surface recession rate—all as a function of time.

[1]This particular type of graphite was used by Gulf General Atomic, Inc., because of the availability of reliable thermo-ablation properties.

3. Results and Discussion

The results obtained from the entire effort are meaningful and very promising. Figure 1 presents radiative heat fluxes as they approach or cross gas–gas and gas–solid interfaces, assuming there is none or little diffusion between atmospheric and sublimation (carbon) gases. Curve ① gives values of gross radiative heating flux flowing from the shocked atmosphere toward the solid; curve ② is the net radiative heating flux, which is the gross flux less the flux flowing from carbon vapor away from the solid; and curve ③ is the net radiative heating flux flowing from carbon vapor into the solid. It is the latter flux that causes sublimation of graphite, assuming relatively negligible value for heat going into the solid due to thermal conduction.

By inspection, it is seen that carbon vapor is very much opaque, since a substantial amount of radiative heat flux

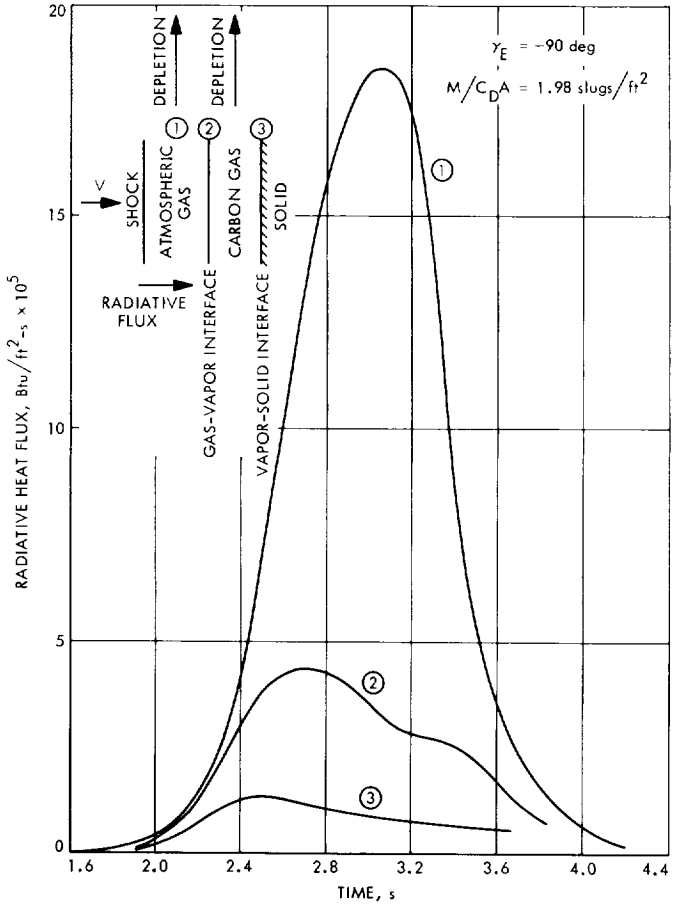


Fig. 1. Effects of sublimation products on radiative heat flux

is being reradiated at the gas–vapor interface. Furthermore, the absorption of radiative heat flux that penetrates carbon vapor layer is also very significant.

Figure 2 shows a plot of overall attenuation coefficient obtained by dividing values of curve ③ by corresponding values of curve ①. It will be observed that maximum attenuation occurs at the maximum incident radiative heat flux, and at this point the net radiative heat flux causing sublimation is only about 5% of the incident value.

A comparison of the results given by curve ① in Fig. 1 with the calculations performed in Ref. 4 shows that there is a good agreement regarding the maximum radiative heat flux incident upon gas–vapor interface, thus validating the somewhat approximate method used in the Sputter program (grey gas approximation).

The maximum surface recession due to sublimation is shown in Fig. 3. It is seen that under conditions investigated, the material surface consumption amounts to approximately 2.5 in. The penetration of bond line temperature (600°F) in the post-ablation period turns out to be only of the order of 0.01 in., due to rather low thermal conductivity of pyrolytic graphite and high rate of surface recession during ablation. These results, however, must be factored to cover the uncertainties in the heat transfer and surface recession calculation methods as well as uncertainties in material thermo-ablative properties. From Venus entry study results,[2] the uncertainty factor of 1.5 is assumed.

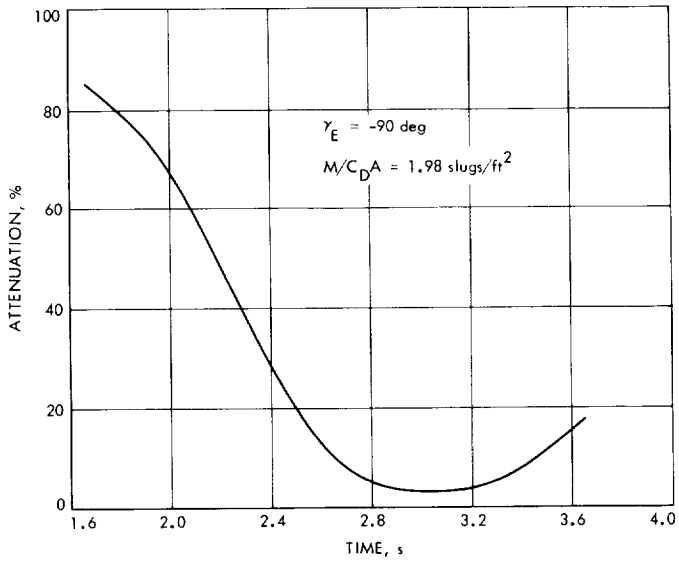


Fig. 2. Shock radiative flux attenuation for Jupiter entry

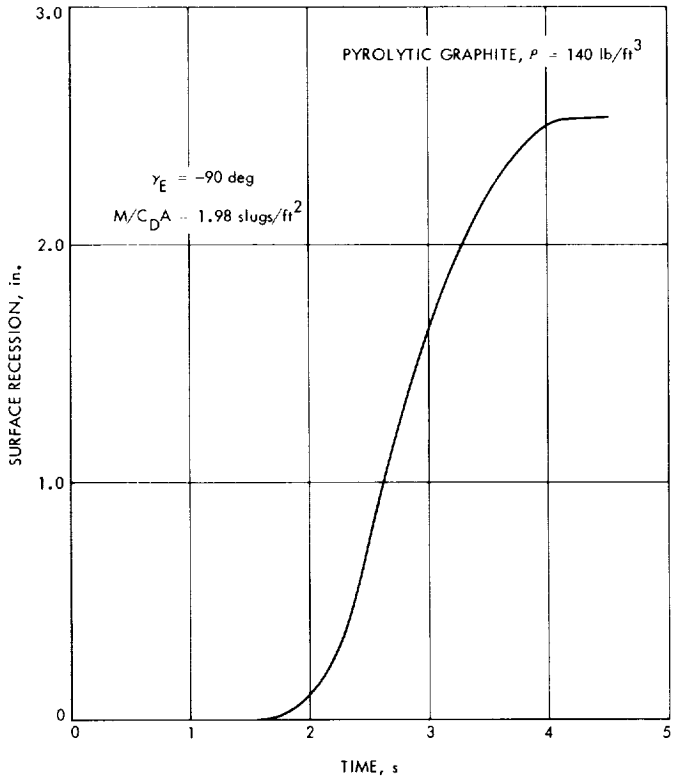


Fig. 3. Stagnation point surface recession for Jupiter entry

Table 2 lists preliminary design specifications, based on the foregoing information, and assuming an entry probe shape to be the blunted cone having a 60-deg half-cone angle, a base diameter of 4 ft, and a nose radius of 1 ft.

The estimates in Table 2 indicate that approximately 53.5% of entry vehicle total weight could be available for payload and structure. More accurate calculations would probably show lower heat shield weight than the figure given above.

4. Conclusions and Recommendations

The foregoing analysis has shown that a Jupiter entry probe with reasonable allowance for payload and structure appears to be feasible. However, the example case

[2]Jaworski, W., and Nagler, R. G., *A Parametric Analysis of Venus Entry Heat Shield Requirements* (forthcoming JPL publication).

Table 2. Preliminary design specifications for Jupiter entry heat shield

Parameter	Value
Entry velocity, km/s	60
Entry angle, deg	−90
Ballisitic coefficent, slugs/ft^2	1.98
Drag coefficent C_D	1.52
Entry vehicle total weight, lb	1214
Average heat shield thickness (85% of stagnation point value), in.	3.35
Forward surface side area, ft^2	14.46
Heat shield unit weight, lb/ft^2	39
Heat shield weight, lb	565
Heat shield weight fraction	0.465

considered may not satisfy all the requirements related to overall mission constraints. In view of this, a spectrum of different entry conditions should be investigated to provide a complete entry information. In addition, the following efforts are recommended:

(1) Modification of Sputter program to allow for more accurate mass depletion mechanism and determination of shock stand-off distance.

(2) Repetition of the present analysis using Sputter program as modified in (1) to assess the difference in heat shield performance.

(3) Further modification of Sputter program to allow for spherical and conical shock approximations.

(4) Repetition of the present analysis to obtain radiative heat flux distribution and corresponding surface recession values over the forward part of the blunted cone body.

(5) Search for the best applicable graphite material that would have reproducible thermo-ablation properties.

References

1. Tauber, M. E., *Atmospheric Entry Into the Major Planets With Emphasis on Jupiter,* Paper No. 68-1150 presented at the AIAA Conference on Entry Vehicle Systems Technology, Williamsburg, Va., December 3–5, 1968.

2. Gilligan, J. E., *Thermophysical Aspects and Feasibility of a Jupiter Atmospheric Entry,* Report S-4, IIT Research Institute, Chicago, Ill., January, 1968.

3. Triplett, J. R., *Sputter, A General Purpose One-Dimensional Radiation and Fluid Mechanics Computer Program,* GA-4820, Part I, General Atomic, Inc., Division of General Dynamics, John Jay Hopkins Lab., San Diego, Calif., Mar. 1965.

4. Stickford, G. H., Jr., and Menard, W. A., *Bow Shock Composition and Radiation Intensity Calculations for a Ballistic Entry Into the Jovian Atmosphere,* Paper 68-787 presented at the AIAA Third Thermophysics Conference, Los Angeles, Calif., June 1968.

XIV. Applied Mechanics

ENGINEERING MECHANICS DIVISION

A. Heat Pipe Performance Map, *J. Schwartz*

1. Introduction

A heat pipe with ammonia as its working fluid was tested and its performance mapped. The data was compared to a similar heat pipe with water as its working fluid.

The accumulation of free hydrogen, a noncondensible gas, was observed to occur in the water heat pipe. A postulation of this phenomenon is presented and conclusions drawn.

2. Experimental Apparatus

The ammonia heat pipe test schematic is shown in Fig. 1. It consists of the instrumented heat pipe with a calorimeter-type evaporator and condenser section. The circulating fluid (water) in each of these heat exchangers was driven by a pump and its mass flow rate was measured with a flow meter. The water entering the evaporator section was heated by means of a hot water bath, and that entering the condenser section was cooled by means of a chiller.

The heat pipe is 15.5 in. long with an OD of 7/16 in. and a wall thickness of 0.020 in. A 100-mesh, double-layer screen wick lines its interior. An end cap and fill tube complete the assembly as shown in Fig. 2. All of the heat pipe components were made of Type 304 stainless steel.

The flow-through calorimeter-type heat exchangers (evaporator and condenser) were made of plexiglass. Both are in a shape of a 3-in. cube which fit over the heat pipe in two sections secured with four screws. Both evaporator and condenser sections were provided with a thermocouple which measured a differential temperature between the inlet and outlet water, as well as one which measured the absolute water temperature.

Fifteen copper–constantan (5-mil OD) thermocouples were spot-welded to the heat pipe in approximately 1-in. intervals. Three thermocouples were under the condenser and three under the evaporator sections. One thermocouple was spot-welded to the end cap and another to the fill tube base outside the condenser. In addition, a thermocouple was inserted to a depth of 1 in. into the container at each of its ends. These thermocouples (vapor-core thermocouples) were designed to measure the vapor

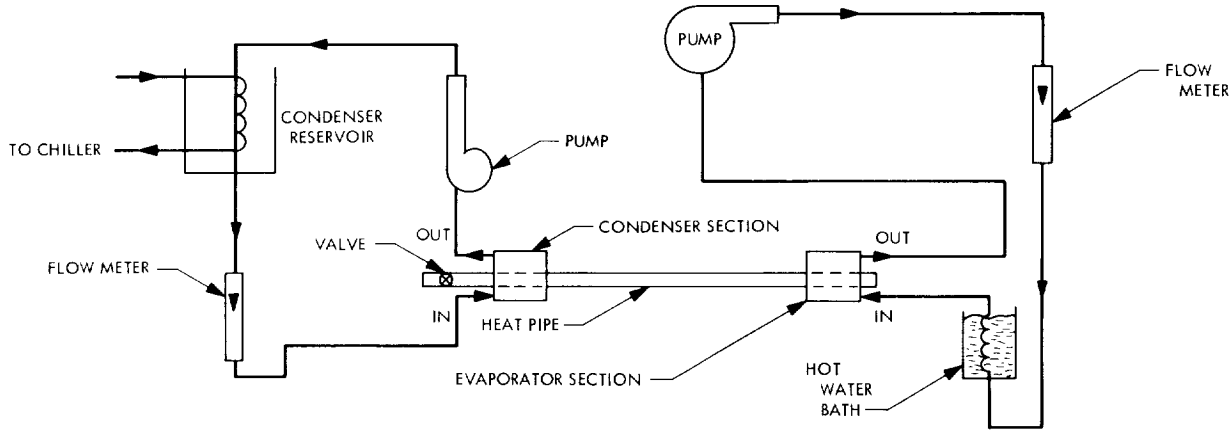


Fig. 1. Heat pipe test schematic

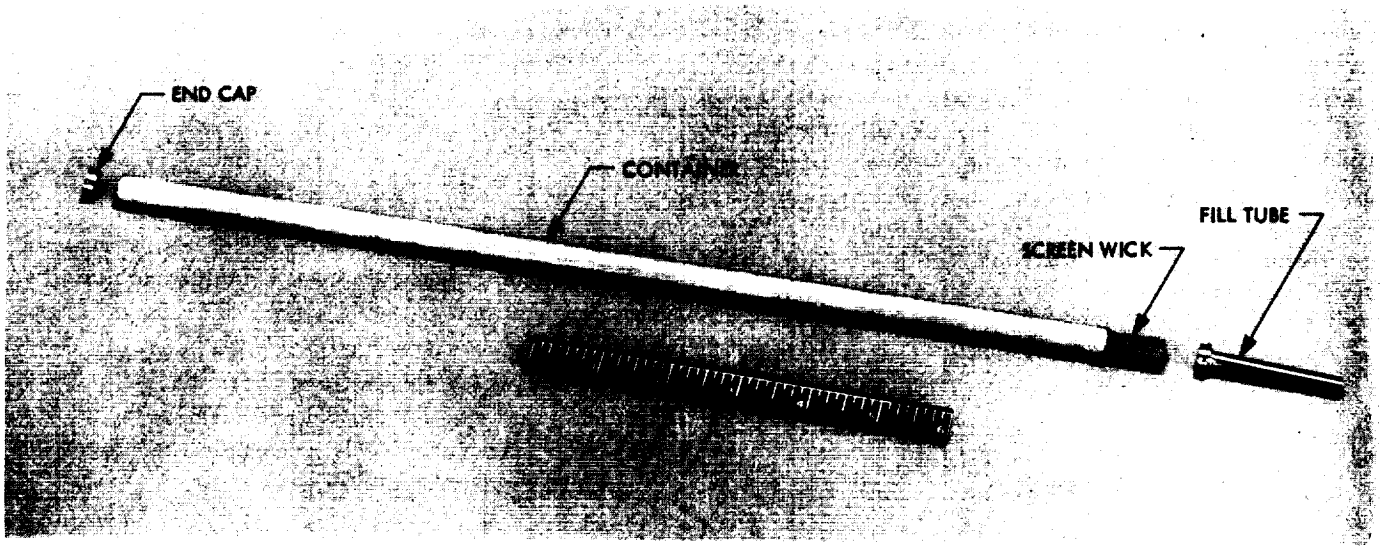


Fig. 2. Heat pipe assembly

temperature. After the spot-welding operation, the thermocouples were wrapped twice around the container and their beads covered with a silastic compound for protection as well as thermal insulation from the surrounding environment.

All of the heat pipe components were vapor degreased and passivated prior to assembly. The wick was then inserted into the container using a clean stainless-steel mandrel. Finally, the end cap and fill tube were electron-beam welded to the container at opposite ends.

The heat pipe was evacuated for 24 h prior to the fill operation. It was then isolated from the vacuum source and a predetermined amount of ammonia introduced into the container. A dewer of liquid nitrogen (LN_2) was inserted over the heat pipe to provide a suitable heat sink. When the transfer was completed, a valve attached to the fill tube was closed and the heat pipe assembly disconnected.

3. Test Procedure

The heat pipe assembly was wrapped in glass wool for insulation purposes and placed in a suitable enclosure. All of the thermocouples were connected to a recorder and monitored continuously throughout the test.

A hot water bath, through which the evaporator loop passed, provided the thermal power loads to the heat

pipe. The condenser loop was fed from a reservoir whose temperature was maintained by a chiller. The heat pipe was subjected to thermal power loads in steps while the condenser reservoir was maintained at a constant temperature. The power load was increased from a nominal value (see Figs. 3 and 4) until the beginning of a dryout condition was reached. The heat pipe was then left to cool down for further tests. Stabilization occurred after approximately 30 min from the time of the initial power input. The onset of a dryout condition was detected by a rapid increase of the heat pipe surface and vapor temperatures in the evaporator region compared to those in the condenser section of the heat pipe.

4. Discussion of Results

The equilibrium temperature distributions of the ammonia heat pipe are shown in Figs. 3 and 4 for various thermal power loads. Figure 4 is a comparison between data from the ammonia heat pipe and a water heat pipe of the same dimensions and for the same test conditions. The data indicate that the lower the condenser temperature, i.e., the sink temperature, the more thermal power is transported by the heat pipe, and the lower its surface (and vapor) temperature at dryout. In addition, the ammonia heat pipe is capable of transporting more than twice the thermal load than the water heat pipe up to an operating temperature of approximately 100°F. Above

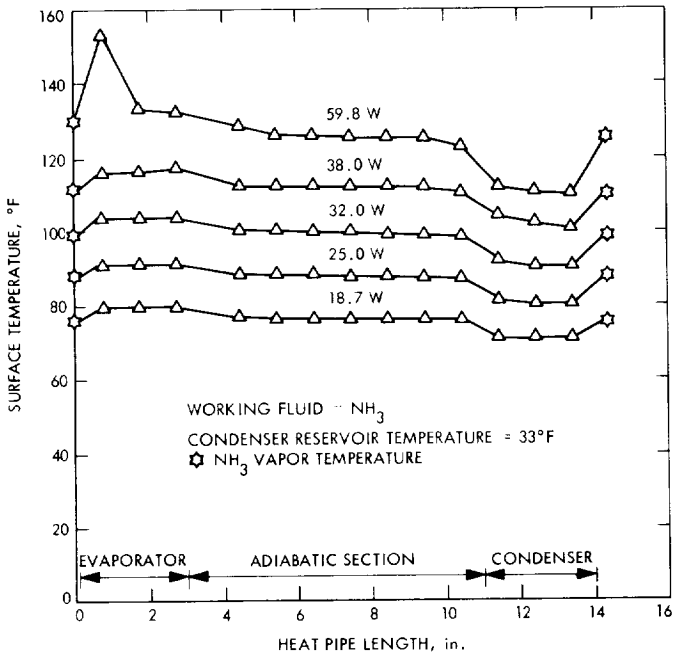


Fig. 3. Stabilization temperature distributions of ammonia heat pipe for various heat loads

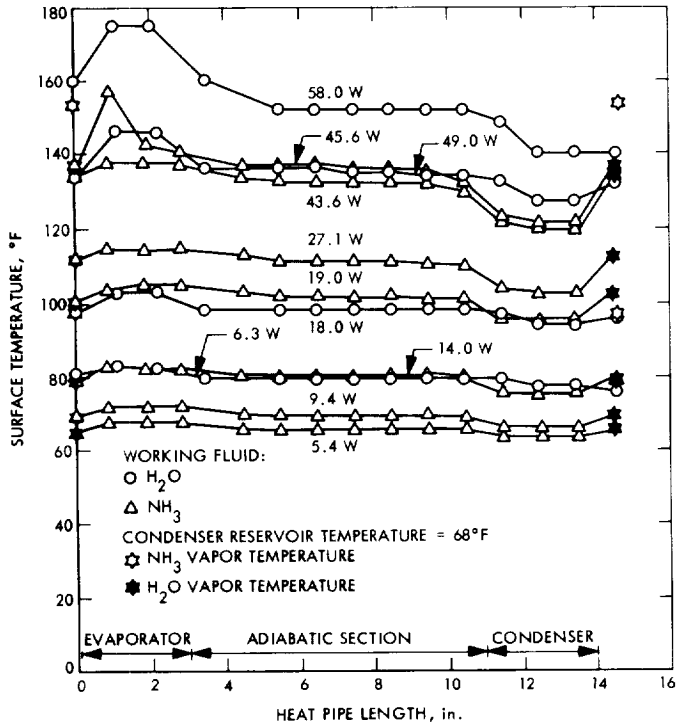


Fig. 4. Stabilization temperature distributions of water and ammonia heat pipes for various heat loads

this temperature, the water heat pipe is slightly more efficient than the ammonia heat pipe, whose capability decreases rapidly with increasing operating temperatures. Just before dryout, the water heat pipe is much more efficient than its ammonia counterpart.

Figure 5 illustrates the average temperature drop across the saturated wick and container wall of two heat pipes with different working fluids but otherwise of the same construction. The largest temperature drop occurred across the saturated wick in both heat pipes. At the beginning of dryout, the temperature drop increased rapidly, which was probably due to vapor formation in the space between the wick and container wall as well as in the wick itself. The vapor blanket, an effective thermal barrier, caused the temperature differential to increase with a very small increase in the heat load.

The coefficient of effective thermal conductivity for the saturated wick was calculated as a function of heat load. In the experimental temperature range, the maximum value of this coefficient is only seven times that of liquid ammonia. The data also indicate that the coefficient of effective thermal conductivity for the ammonia-saturated wick is approximately three times that of a similar water-saturated wick. This is primarily due to the fact that the

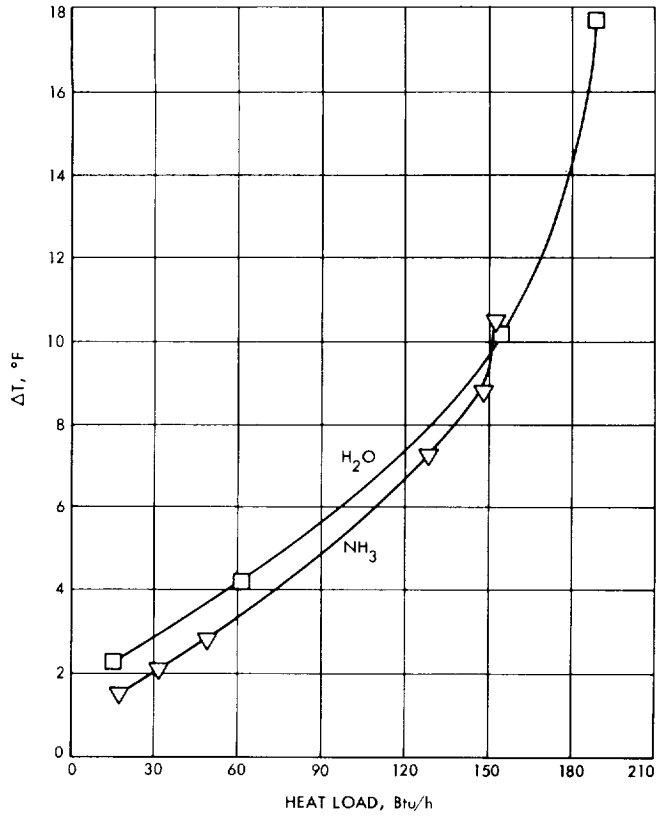


Fig. 5. Average temperature drop across container wall and saturated wick as a function of heat load

ammonia-saturated wick became much more depleted than the water-saturated wick as heat loads were applied, which resulted in a thinner liquid layer and a shorter heat path.

In both cases, however, the coefficient of effective thermal conductivity for the saturated wick was much lower than that for the container and wick materials. This indicates that the metal strands of the screen wick were not very effective in the heat transfer process, and that the condensate in the wick played a dominant thermal role.

The effective coefficient of thermal conductivity of the entire heat pipe was also calculated from the test data. Its highest value is approximately 9250 Btu/h ft °F compared to a water heat pipe with a value of 14,750 Btu/h ft °F. Nevertheless, this represents a value of 42 times that of pure copper in the corresponding temperature range.

5. Noncondensible Gas Phenomenon

Early in the test program, it was observed that those heat pipes which are composed of stainless steel and water displayed a large temperature differential between the ends after some lapsed time. When this phenomenon was first observed, it was not clear whether noncondensible gases were the sole cause, or whether another mechanism was also responsible.

The gases present in a number of test heat pipes were analyzed with a mass spectrometer. The data indicate that hydrogen H_2 was generated within the heat pipe, which was not gas leaking in from the external source. It was perhaps surprising that 98% of the gas was hydrogen.

To learn more about noncondensible gas generation and accumulation rates, a life test was undertaken utilizing one water–ammonia, two ammonia, and three water heat pipes. The working fluid of the first unit was composed of the required volume of water to which was added 0.5% ammonia, by weight. Eight evenly distributed copper–constantan thermocouples were spot-welded to each heat pipe. A resistance heater assembly and flow-through calorimeter-type heat exchanger utilizing water were attached to each unit at opposite ends. The system was left to operate continuously and all temperatures were monitored periodically.

Deterioration of the water and water–ammonia heat pipes was observed to occur in the first few days of operation. The criterion used for determining deterioration was the temperature distribution along the heat pipe. All of the units started the test with an isothermal temperature distribution.

However, after a few days of operation, all but the ammonia heat pipes displayed a significant temperature differential between the ends which increased with time. After approximately 5 mo of continuous operation, the gases in one of the water heat pipes were analyzed with a mass spectrometer. The results confirm the previous findings concerning the identity of the noncondensible gas. In addition, the test data indicate that the generation rate of hydrogen is a strong function of temperature.

The H_2 generation rate followed a similar pattern for all heat pipes. After approximately 2 mo of operation, H_2 generation appeared to have ceased or its rate was too small to be noted.

Attention is now focused on the mechanism of H_2 generation and the reactions involved within the heat pipe.

Hydrogen generation within the heat pipe must have been the result of a chemical reaction rather than a thermal degradation mechanism because of the relatively low

operating temperatures. Water is known to be very stable thermally and begins to decompose only at temperatures above 1000°C. For example, when water is heated in a closed system to a temperature of 1327°C, only 0.0446% decomposition occurs. It is, therefore, most likely that the water in the heat pipe reacted with some (or all) of the metal constituents in the stainless-steel container which came in contact with it.

The most likely chemical reaction between iron and water, for example, is

$$Fe + 2H_2O \rightleftharpoons Fe(OH)_2 + H_2 \quad (1)$$

or in the ionic form,

$$Fe^0 + 2H^+ \rightleftharpoons Fe^{++} + H_2^0 \quad (2)$$

The assumption here is based on the condition that iron in the container wall reacted with the hydrogen ion H^+ in the water to form hydrogen gas, H_2. The same reaction may have involved nickel, chromium, or other metals present in smaller quantities. The concentration of H^+ is very low in water. This is shown in the following known relationship:

$$(H^+)(OH^-) = 10^{-14} \text{ (per liter)} \quad (3)$$

For pure water, the numerical value of H^+ is 1×10^{-7} normal (N); therefore, $(OH)^-$ must have the same value. However, if $(OH)^-$ were to be increased to 0.1 N, then

$$(H^+)(0.1) = 10^{-14}$$

or

$$H^+ = 10^{-13} N \quad (4)$$

This is very desirable because the lower the H^+ concentration, the less is its availability to react. Ammonia (NH_3), for example, is more attractive since

$$(H^+)(NH_2^-) = 10^{-16} \quad (5)$$

Although $(NH_2)^-$ can ionize further, ionization is extremely low, and can be ignored. Thus,

$$H^+ = 10^{-8} N \quad (6)$$

in ammonia is smaller than in pure water.

In light of this discussion, the production of hydrogen gas within the heat pipe followed a process described by Eq. (1) and similar reactions involving other metals. The reaction proceeded toward formation of H_2 and continued until some inhibiting mechanism took place, at which point it appeared to have ceased, at least on a detectable level. The nature of the inhibiting mechanism is not known at present. Likewise, its influence on the formation of hydrogen is not understood other than what was postulated above, based on temperature patterns' behavior with time.

As to the temperature differential along the test heat pipes which was attributed to the presence of hydrogen gas, it appears that various interrelated phenomena may be involved. The amount of hydrogen gas collected from the heat pipe was so small that it takes more than 40 times that volume to fill a 4.5-in. section of the heat pipe. This fact immediately rules out the possibility that the entire condenser region was filled with hydrogen gas. Accumulation of hydrogen in the condenser region must also be ruled out *a priori* because of the tendency of the gas to diffuse throughout the heat pipe.

One possible explanation which may be offered is that hydrogen gas was transported with water vapor from the evaporator to the condenser sections of the heat pipe. In the condenser region, hydrogen somehow became trapped in the wick matrix which interfered with the normal process of gas diffusion. As more hydrogen was trapped by the wick, a thin annular layer began to form within it. This layer apparently hindered normal vapor condensation, effectively removing the entire region from the heat pipe loop. The core in the condenser region may have been filled with stagnant water vapor in which conduction was the only mode of heat transfer.

6. Conclusions

The performance map of the heat pipe provides a useful tool for preliminary design analysis of its capabilities. Of primary importance are temperature distribution and transverse temperature drop curves as a function of heat loads. Likewise, the input power flux curve is very important. Of secondary importance are the condensate velocity, vapor density, and mass flow rate.

The ammonia heat pipe is more efficient in the temperature region below 100°F. However, above this point, the water heat pipe is increasingly more efficient than

the ammonia heat pipe as the operating temperature increases.

The coefficient of effective thermal conductivity for the ammonia-saturated wick is approximately three times that of the water-saturated wick. This value, however, is only one-tenth of the value for the screen wick and container materials.

The beginning of dryout conditions occurred at a different value of heat load (and vapor temperature) for different condenser reservoir temperatures. As the condenser reservoir temperature was decreased, dryout occurred at higher heat loads but at lower vapor temperatures. In all of the test runs, the vapor-core thermocouples served as the criteria for satisfactory operating conditions.

From mass-spectrometer analysis of the gases present in various heat pipes, it was established that water heat pipes generated hydrogen, a noncondensible gas. Apparently, iron with other metals, such as nickel and chromium, in the stainless-steel container reacted chemically with water and generated the hydrogen. Based on preliminary data, it appears that the gas-generation rate is a strong function of temperature, the volume increasing with increasing operating temperatures. The reaction, however, seemed to taper off with time, which suggests a decrease in H_2 generation rate.

The life-test water heat pipes show that a gas-generation rate of 4.5 in. along the length of the heat pipe per 65 days was obtained where the average operational temperature level was 85°F. The generation rate was not linear with time, and the reaction evidently ceased at this point. The nature of the inhibiting mechanism is not known at present. The ammonia heat pipes indicated no noncondensible gas generation during the entire test duration.

Based on the life-test data, it is evident that the volume of hydrogen gas was too small to fill the entire condenser region (4.5 in. in length) of the heat pipe. The temperature differential which was observed must have been due to various interacting causes. It seems plausible that hydrogen gas was trapped by the wick in the condenser region of the heat pipe, eventually forming an annular layer which effectively blocked condensation from taking place. The reason why hydrogen gas preferred to accumulate in the cold region in the first place is not fully understood. A great deal of empirical data, as well as theoretical analysis, are lacking to offer a satisfactory explanation for the above phenomenon.

The problem of noncondensible gas generation may be avoided simply by choosing a heat pipe whose metal components are below hydrogen in the electromotive series. For example, the material combination of water and copper is not likely to react and produce a noncondensible gas. Similarly, if a metal above hydrogen in the electromotive series is chosen, it may be used in conjunction with a nonreacting working fluid other than water. An example is stainless steel and ammonia in the heat pipe, which was discussed earlier.

B. Optimum Pressure Vessel Design Based on Fracture Mechanics and Reliability Criteria,

E. Heer and J. N. Yang

1. Introduction

In view of long-duration space missions in which the weight of pressure vessels constitutes a significant portion of the spacecraft weight, and in which long-time (up to 10 yr) reliable performance of pressure vessels is expected in a relatively unknown space environment, the application of rather arbitrary safety factors is less than satisfactory. A rational answer regarding the safety of pressure vessels should therefore be based on the concepts of probability theory and should be expressed in terms of the probability of survival or reliability, consistent with the reliability requirements of the overall spacecraft system.

For brittle materials and for ductile materials with sufficiently large existing flaws (no sharp line exists between these cases), fracture mechanics concepts can be employed to establish failure criteria for design based on the presence of existing flaws in the material. If the maximum existing flaw in a vessel is known, techniques have been developed to determine the burst pressure with reasonable accuracy. However, the presence of flaws being a fact, their size and growth behavior under sustained and repeated loads have considerable statistical variations, thus requiring the application of probability theory for a rational approach to design.

2. Statistical Properties of Pressure Vessels

Considerable attention has been given (e.g., Ref. 1) to the subject of the statistical representation of the failure properties of materials. Extensive use has been made of the Weibull distribution (Ref. 2), which has its basis in the "weakest link hypothesis": Failure of the entire component occurs when any one of its material elements is subjected to its critical failure stress.

The entire component with total material volume $V = \Sigma V_j$ is subdivided into suitably small material volume elements V_j. The fracture strength of the jth volume element can then be represented by the well-known Griffith–Irwin equation

$$R_j = A a_j^{-1/2} \tag{1}$$

where A is assumed to be deterministic constant, since for a given component its dispersion is small compared to the dispersion of the flaw size a_j contained in V_j.

The statistical distribution of R_j can be obtained from the corresponding distribution of a_j, and vice versa, using Eq. (1). Present state of technology does not, in most cases, allow the measurement of the distributions of a_j. The distributions of R_j which characterize the material are therefore measured directly by "coupon tests," and the statistical parameters are determined as shown in a JPL Technical Report.[1] Experience has shown that for uniaxial coupon tests the distribution function of R_j can be well represented by the Weibull distribution

$$F_{R_j}(x) = 1 - \exp\left[-\frac{V_j}{v}\left(\frac{x - x_\mu}{x_0}\right)^k\right], \qquad x \geqq x_\mu \tag{2}$$

where v is the unit volume and x_μ, x_0, and k are parameters.

Assuming that every flaw orientation is equally likely and that the mean value of these orientations in a two-dimensional stress field (characteristic of pressure vessels) is a good approximation (Footnote 1 gives a detailed discussion of the flaw angle distribution aspect), it can be shown that the strength distribution function of the entire pressure vessel can be written as

$$F_R(x) = 1 - \exp\left\{-\frac{1}{v}\int_V \left[\frac{x(\phi_1 + \phi_2) - x_\mu}{x_0}\right]^k dV\right\}, \qquad x(\phi_1 + \phi_2) \geqq x_\mu \tag{3}$$

in which $S\phi_1$ and $S\phi_2$ are the analyzed principal stresses at a point due to the applied pressure load S.

If the vessel is subjected to cyclic and/or sustained pressures which are smaller than the failure pressure, subcritical flaw growth is expected in general, thus reducing in time the fracture stress for a given flaw. If S_c and S_s are respectively the magnitudes of cyclic and sustained pressure loading and R is the strength of the vessel before S_c and S_s are applied, then the loss of strength in time can be expressed as

$$S_c/R = c(n) \tag{4}$$

$$S_s/R = s(t) = b + (1 - b)\, e^{-at} \tag{5}$$

where $c(n)$ and $s(t)$ are equal to one at $n = 0$ and $t = 0$ and are monotonically decreasing functions; n and t represent, respectively, cycles and time to failure after the application of S_c or S_s. It is assumed here that $c(n)$ and $s(t)$ are deterministic.

A typical loading history for pressure vessels is shown in Fig. 6. Let the vessel strength be given by R_0 after the application of the proof load S_0, by $R(N)$ after the subsequent application of N cycles of S_c and by $R(T)$ after S_s has been applied for a period T following the cyclic loading. It can then be shown (Footnote 1) that

$$\begin{aligned} R(N) &= S_c/c\,[c^{-1}(S_c/R_0) - N], && \text{for } R_0 > S_c \\ &= 0, && \text{for } R_0 \leqq S_c \end{aligned} \tag{6}$$

$$\begin{aligned} R(T) &= S_s/s\,\{s^{-1}[S_s/R(N)] - T\}, && \text{for } 1 > S_s/R(N) > b \\ &= R(N), && \text{for } b \geqq S_s/R(N) \\ &= 0, && \text{for } S_s/R(N) \geqq 1 \end{aligned} \tag{7}$$

where c^{-1} and s^{-1} are the inverse of functions c and s, respectively.

3. The Probability of Vessel Failure

After the pressure vessel has been subjected to the deterministic proof load S_0, the original strength distribution (Eq. 3) is truncated at S_0 and becomes (Footnote 1)

$$F_{R_0}(x) = H(x - S_0)\left(1 - \exp\left\{-\frac{1}{v}\int_V \left[\left(\frac{x(\phi_1 + \phi_2) - x_\mu}{x_0}\right)^k - \left(\frac{S_0(\phi_1 + \phi_2) - x_\mu}{x_0}\right)^k\right] dV\right\}\right), \qquad S_0(\phi_1 + \phi_2) \geqq x_\mu \tag{8}$$

[1]Heer, E., and Yang, J. N., *Optimum Structural Design Based on Fracture Mechanics and Reliability Criteria*, JPL Technical Report in preparation.

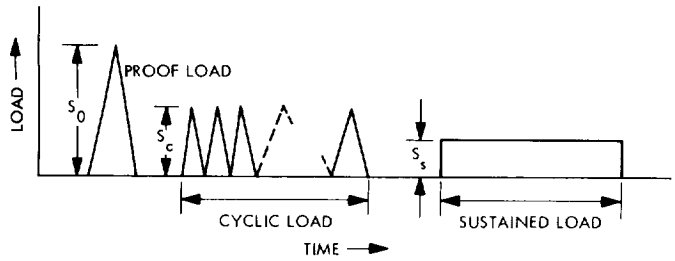


Fig. 6. Typical loading history for pressure vessels

where $H(x - S_0)$ is the Heaviside unit step function.

If the density functions of S_c and S_s are $p_{S_c}(x)$ and $p_{S_s}(x)$, then the probability of failure due to cyclic loading after the proof load is

$$P_c = \int_0^\infty F_{R_0}\left(\frac{x}{c(N-1)}\right) p_{S_c}(x)\,dx \tag{9}$$

and the probability of failure due to sustained loading, given that the vessel survived the cyclic loading, is

$$P_{cs} = \int_0^\infty \int_0^\infty F_{R_0}\left(\frac{x}{c\,[c^{-1}(xs(T)/y) + N]}\right) \times p_{S_c}(x)\,P_{S_s}(y)\,dx\,dy \tag{10}$$

The probability of failure of sustained loading after proof load, i.e., the cyclic loading in Fig. 6 is not applied, is

$$P_s = \int_0^\infty F_{R_0}\left(\frac{x}{s(T)}\right) p_{S_s}(x)\,dx \tag{11}$$

4. Optimum Design

A general formulation and discussion of optimum structural design based on reliability and proof-load test taking into account overall cost constraints has been given in Refs. 3 and 4. General developments of the subsequent discussion are given in Ref. 3.

It is assumed here that a pressure vessel representing one component is to be designed optionally for a sustained load S_s after it has been proof-tested at a level S_0. The optimization problem is stated as follows: Minimize the weight W of the vessel subject to the constraint of a maximum expected cost EC_a relative to total project failure cost C_F,

$$EC_a \geqq EC = [B + \beta(\alpha|\epsilon - 1|)] + \gamma\frac{P_0(\epsilon)}{1 - P_0(\epsilon)} + P_f(\nu, \epsilon) \tag{12}$$

where on the right-hand side the first term represents a first order approximation to the relative cost of coupon tests necessary to establish the truncated element strength distribution, the second term represents the relative expected cost of proof-load testing and the third term represents the probability of project failure or the relative cost of project failure, which in this case is given by P_s in Eq. (11). In Eq. (12), B is the minimum relative basic cost of coupon tests if proof-load testing is conducted at a level $\epsilon = S_0/\bar{R} = 1$, α gives the rate of cost increase for $\epsilon < 1$ and $\epsilon > 1$ for which it may be different, while β indicates relative cost of coupon testing. γ is the relative cost of failure of a vessel due to proof load, $P_0(\epsilon)$ is the probability of failure during proof-load application, and $P_0(\epsilon)/[1 - P_0(\epsilon)]$ is the expected number of vessels lost during proof loading before the one that survives is obtained. ν is the central safety factor defined by $\nu = \bar{R}/\bar{S}_s$ and which is proportional to the vessel wall thickness h and hence to the weight of the pressure vessel.

5. Numerical Example and Discussion

For a representative numerical example, a spherical 20-in.-diam spacecraft pressure vessel is chosen which must sustain for 360 h an assumed gaussian distributed operating pressure S_s with mean $\bar{S}_s = 1000$ psi and coefficient of variation of 2%. The vessel material is titanium Ti-61A-4V having a Weibull strength distribution with mean $\bar{R} = 160{,}000$ psi, coefficient of variation of 10%, and $x_\mu = 0$. The Weibull strength distribution has been determined using coupons 8 in. long, 0.5 in. wide, and 0.25 in. thick. In Eq. (12), α is chosen to be equal to 1 for $\epsilon < 1$ and equal to 2 for $\epsilon > 1$. B is left arbitrary since it does not influence the optimization process. The vessel is to be designed for room temperatures at which in Eq. (5) the parameter values are $b = 0.5$ and $a = 0.01$. It is assumed that the stress field is homogeneous so that $\phi_1 = \phi_2 =$ constant in Eqs. (3) and (8).

Typical numerical results for the indicated parameters are presented in Table 1 and Figs. 7 and 8, where the load distribution is normal with coefficient of variation 0.02 and the strength distribution is truncated Weibull with coefficient of variation 0.10. Figures 7a and 7b represent EC as functions of ϵ for given ν and β. The main conclusion to be drawn from these figures is that the optimum relative proof-load level ϵ^* should be chosen from the locus indicated by curve 1. The parameter β has, in this case, a slight tendency to move ϵ^* closer to unity. This can also be seen in Table 1. This effect of β can be substantial depending on the particular combination of ν and γ. It is due to this fact that, under reasonable relative expected cost constraints, the optimum proof-load level

Table 1. Optimum design of pressure vessel

γ	h, in.	ϵ^*	S_0, psi	ν	P_f
$EC_a = 0.1 \times 10^{-4}$; $\beta = 0$					
10^{-7}	0.17457	1.179	2113	1.7925	0.174×10^{-5}
10^{-6}	0.1852	1.114	2108	1.8926	0.177×10^{-5}
10^{-5}	0.2065	1.0091	2110	2.091	0.472×10^{-6}
10^{-4}	0.2447	0.850	2088	2.444	0.72×10^{-6}
Standard optimum design	0.5134	0	0	4.8238	0.1×10^{-4}
$EC_a = 0.1 \times 10^{-4}$; $\beta = 10^{-5}$					
10^{-7}	0.1765	1.169	2117	1.81	0.1042×10^{-5}
10^{-6}	0.1870	1.096	2093	1.91	0.8236×10^{-6}
10^{-5}	0.2066	1.0086	2110	2.092	0.4639×10^{-6}
10^{-4}	0.2484	0.851	2089	2.477	0.6941×10^{-6}

will fall with great likelihood within the range of two standard deviations around the mean given by $\bar{R}$. From the designers' point of view this is very desirable, since, in general, a considerably greater number of coupon tests is required for characterizing the truncated element strength distribution within a certain level of statistical confidence, if ϵ^* falls outside this region.

In Figs. 8a and 8b, EC is shown as a function of ϵ^* for different ν, γ, and β. Here it is clearly seen that the degree to which the optimum test level ϵ^* is shifted toward unity depends on β, i.e., the cost of coupon tests. The effect of parameters β, ν, and γ on the optimum design has been discussed in detail in Ref. 4.

In Table 1, the results of a standard optimum design in which the proof-load test has not been considered (or has not been performed) are compared with the results

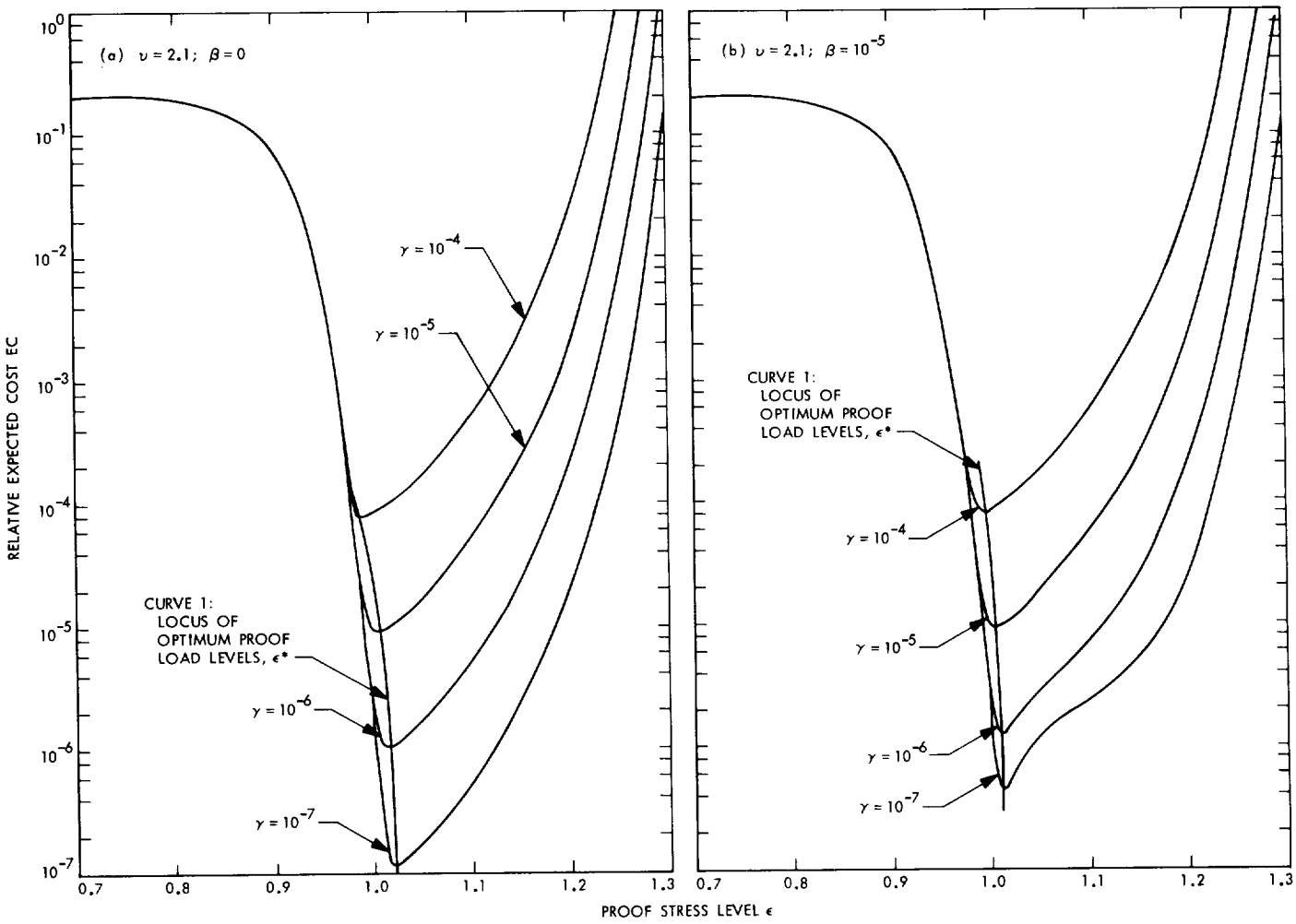


Fig. 7. Relative expected cost EC as a function of proof stress level ϵ

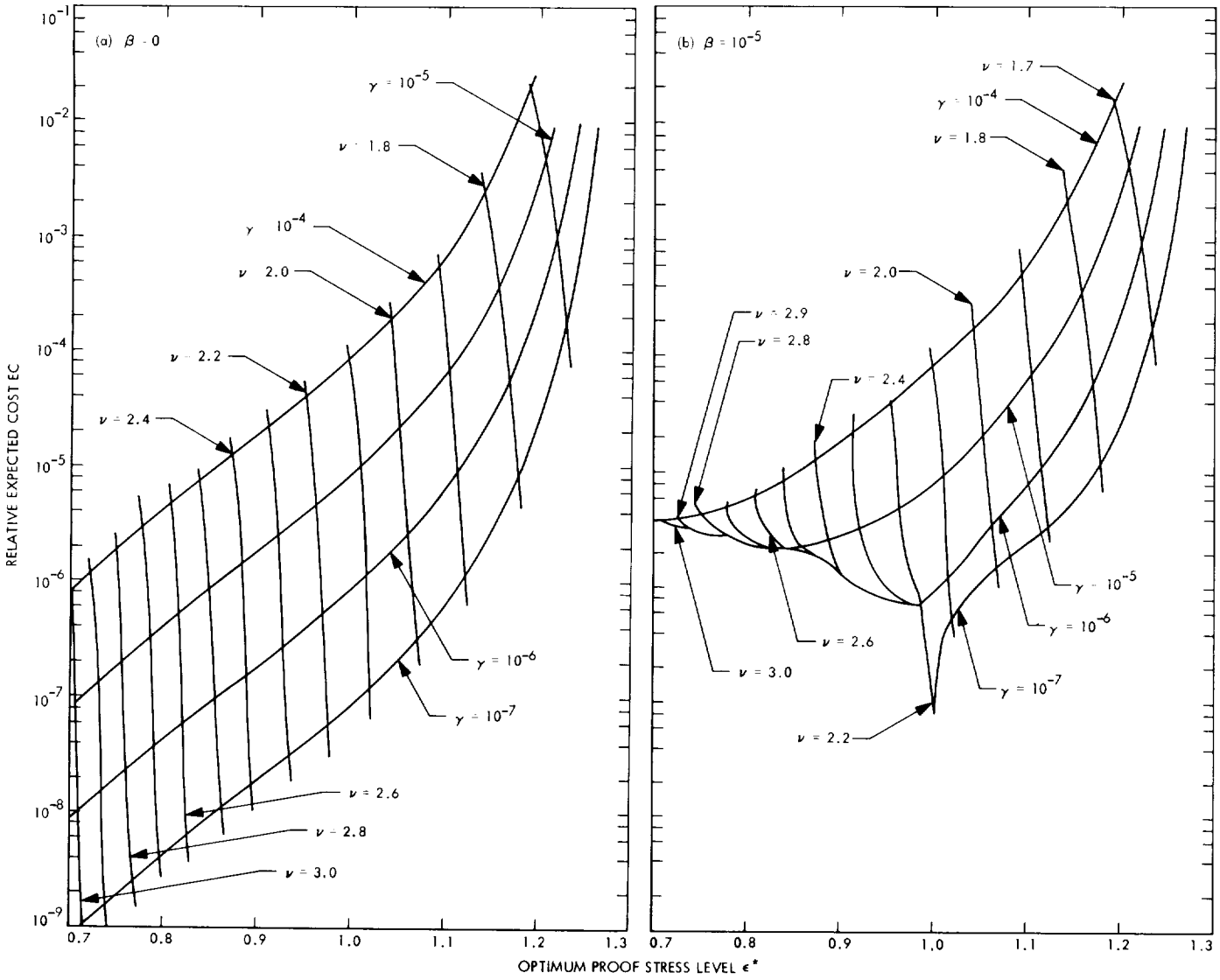


Fig. 8. Relative expected cost EC as a function of optimum proof stress level ϵ^*

including the proof-load test. It is seen that for the same expected cost constraint a considerable weight saving together with a considerable decrease in the probability of failure P_f has been realized.

References

1. Freudenthal, A. M., "Statistical Approach to Brittle Fracture," *Fracture*, Vol. II, Chap. 6. Edited by H. Liebowitz. Academic Press, Inc., New York, 1968.
2. Weibull, W., "A Statistical Distribution Function of Wide Applicability," *J. Appl. Mech.*, pp. 293–297, 1951.
3. Shinozuka, M., and Yang, J. N., "Optimum Structural Design Based on Reliability and Proof-Load Testing," *Annals of Reliability and Maintainability*, Vol. 8, 1969. Also available as Technical Report 32-1402, Jet Propulsion Laboratory, Pasadena, Calif., June 15, 1969.
4. Shinozuka, M., Yang, J. N., and Heer, E., "Optimum Structural Design Based on Reliability Analysis," presented at the Eighth International Symposium on Space Technology and Science, Tokyo, Japan, 1969.

XV. Instrumentation

ENVIRONMENTAL SCIENCES DIVISION

A. An Experimental Determination of the Stefan-Boltzmann Constant, *J. M. Kendall, Sr.*

1. Introduction

The standard total radiation absolute radiometer provides an absolute standard of high accuracy (<0.5% error in measuring intensities between 0.045 and 0.16 W/cm^2). It is usable throughout the UV, visible, and IR ranges. The high accuracy is mostly attributable to the use of a black internally coated cavity receptor. Use is made of the electrical heating equivalence for radiation heating (or loss) through the aperture. Since the cavity radiometer has an accurate hemispherical response, it is suitable for making accurate measurements of isotropic thermal radiation and can, therefore, be used to make an experimental measurement of the Stefan–Boltzmann constant. The maximum operating temperature of the radiometer of about 200°C (set by the inability of the internal coating of the cavity to withstand higher temperatures) limits measurements of intensities to a maximum of about 0.28 W/cm^2.

A description of the radiometer is given in considerably more detail in Ref. 1.

2. Description of the Radiometer

The radiometer (Fig. 1) has overall dimensions of 1¼-in. diameter by 1½-in. length. The aperture area is very nearly 1 cm^2. The copper thermal guard, gold-plated inside and out, is quite massive in order to assure isothermality. On the outside surface of the thermal guard

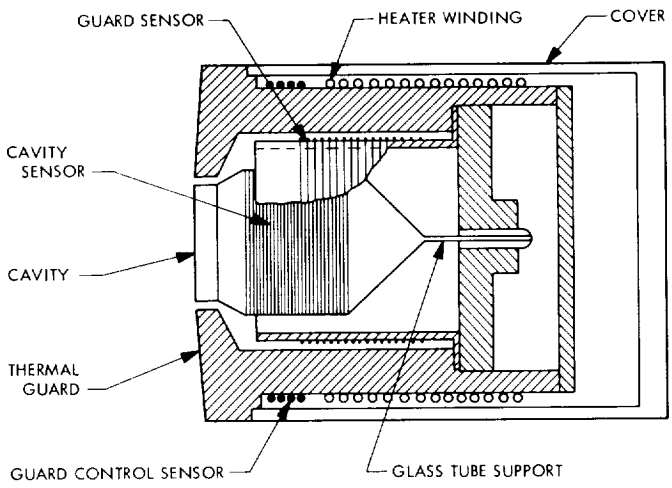


Fig. 1. Standard radiometer

is a heater winding which provides a means of raising the radiometer temperature to the desired operating value, which is maintained by the guard control sensor winding of Pt wire.

The cavity is fabricated from 5-mil silver sheet. On its cylindrical portion is located a combined heating and sensing winding of Pt wire. The cylinder (9/16-in. diameter) is spun down to a smaller cylinder to make the aperture. Because of its design and because of the high thermal conductivity of silver in the cavity, the cavity deviates from isothermality by no more than 0.1°C, for which deviation allowance is made in reduction of measurment data.

The heater–sensor winding on the cavity provides a highly accurate means of measuring the cavity temperature, and also supplies the necessary heat to maintain the cavity temperature accurately equal to the temperature of the thermal guard. The thermal coupling between the cavity and the thermal guard is made as small as possible to avoid radiative heat transfer and to avoid degraded sensitivity and accuracy of the radiometer. In addition, the surface of the heater–sensor winding is covered by a layer of ½-mil aluminum foil of low emissivity. Furthermore, the inside surface of the guard sensor is polished silver. These provisions all go to decrease the radiative heat exchange between cavity and thermal guard. The glass supporting the cavity in the thermal guard is of negligible conductance.

Located inside the thermal guard, and in good thermal contact with the guard, is the guard sensor referred to above. This sensor consists of a cylinder of 20-mil-thick silver on which is wound yet another Pt winding for accurate measurement of the guard temperature.

The black coating inside the cavity has thermal resistance of about 2.5°C drop/W/cm² (Ref. 2), the adverse effect of which is greatly reduced by the cavity enhancement of emissivity ϵ and absorptivity α. The residual effect is allowed for in data reduction.

3. Electronic Circuit

The electronic circuit must:

(1) Maintain the thermal guard at a known preset constant temperature.

(2) Supply exactly enough electrical heat to the cavity to maintain it at the thermal guard temperature and to make up for heat radiated out of the aperture.

(3) Accurately measure the electrical power required to fulfill the requirement of item 2.

Figure 2 shows the electronic circuit. The three independent systems are:

(1) Thermal guard control system (temperature control sensor and servo control) for maintaining the radiometer at the set temperature.

(2) Thermal guard sensor system for accurately determining thermal guard temperature.

(3) Cavity sensor system and associated bridge circuit for accurately measuring cavity temperature and for supplying an exact amount of heat to maintain the cavity at the exact temperature of the thermal guard.

The time constant of the radiometer, if it had no automatic controls, would be inconveniently long (about 30 min to make a measurement). Amplification and manual control, however, make it possible to get full accuracy in about 1 min from a sudden large change in incoming intensity level.

There are Pt resistance windings for the cavity sensor and the thermal guard sensor, which must be very accurately calibrated (to about 0.01°C accuracy) for resistance

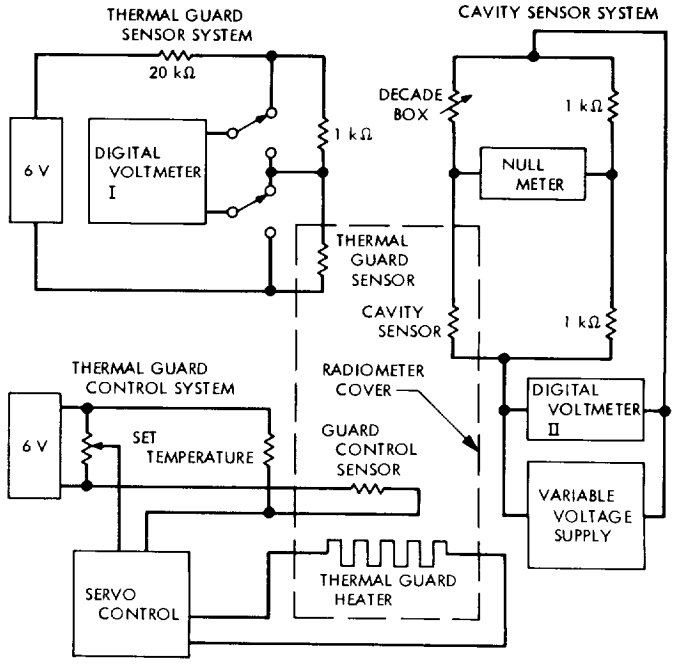


Fig. 2. Radiometer electronic circuit

versus temperature. A special oven is used whose temperature is determined with a Leeds and Northrup Pt resistance thermometer.

4. Measurement of Stefan–Boltzmann Constant

For making measurements from which the Stefan–Boltzmann constant was determined, a test configuration (Fig. 3) with a cold cavity in a vacuum chamber (pressure held $<1 \times 10^{-5}$ torr) was used with the radiometer positioned in the aperture of the cold chamber as shown. The cold cavity is a double-walled vessel filled with LN_2 to keep the inside surface at 77°K. The radiometer sees nothing but the black cold walls of the interior of the chamber over its hemispherical viewing angle. Since the radiometer is operated at 26°C and higher, its aperture is emitting radiation characteristic of its absolute temperature, with almost no radiation coming back into the radiometer aperture from the cold walls.

Table 1 lists correction quantities involved in measuring the Stefan–Boltzmann constant in the cold cavity. Table 2 shows the results of the error analysis made for the setup of Fig. 3. Table 2 lists estimated values of errors in all the various parameters recognized as significant. These errors remain after all possible corrections have been made. If these errors are added up as a simple sum, they

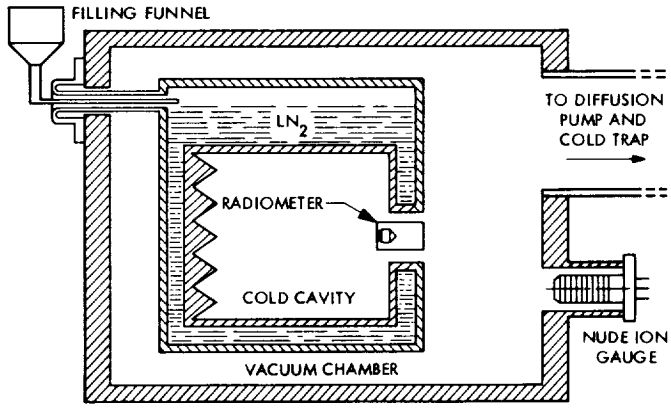


Fig. 3. Radiometer in cold cavity and vacuum chamber

Table 1. Correction quantities for measurement of Stefan–Boltzmann constant

Effect	26°C	83°C	115°C	135°C
Surround-radiation entering through annulus reflected into radiometer, μW	95	95	95	95
Emitted radiation off radiometer reflected back into radiometer, μW	13	35	40	49
Emitted radiation off cold cavity walls at 77°K into radiometer, μW	200	200	200	200
Radiative and conductive heat transfer from thermal guard to radiometer cavity, μW	32	65	93	114
Variation of aperture area with temperature, cm^2	1.0181	1.02025	1.02145	1.0222
Temperature drop between Pt winding and Ag in cavity (correction factor for radiation out of aperture), °C	0.9995	0.9991	0.9988	0.9984

Table 2. Error analysis results

Parameter	Method of determining error	Nominal value	Estimated error	Resulting error in σ_{meas} after making corrections, %
Area A, cm^2	Optical comparator	1.0181 at 26°C	0.1%	±0.1
Emissivity of coating	Calorimetric	0.945 in infrared	1%	±0.16
Temperature	Pt thermometer	492–684 Ω	0.04°C	±0.05
Electric power	E^2/r	0.045–0.16 W	0.000015	±0.02
Temperature drop (heater winding to cavity coating)	Measured, calculated	0.14°C drop at 0.16 W	0.01°C	±0.01
Heating of cavity by electrical leads	Calculated	46 μW at 0.16 W	5 μW	±0.003
Heat transfer by thermal coupling	Measured, calculated	62 μW	10 μW	±0.007
Cold cavity radiation at 77°K	Calculated	0.0002 W	20 μW	±0.01
Reflection back into radiometer	Calculated	0.00014 W	14 μW	±0.01
				Total 0.37

Table 3. Measurements of Stefan–Boltzmann constant[a]

Temperature, °C	λ_{max}, μm	ϵ_{cal}	$\epsilon_{cav_{eff}}$	σ_{meas}, W cm^{-2} °K^{-4} × 10^{-12}	$\sigma_{meas}/\sigma_{th}$
136	7.1	0.946	0.9911	5.6735	1.0007
115	7.45	0.941	0.9902	5.6850	1.0027
83	8.1	0.935	0.9891	5.6888	1.0034
26	9.6	0.925	0.9873	5.6992	1.0052
				Av 5.6866	Av 1.0030

[a]Theoretical value $\sigma_{th} = 5.6697$.

amount to a little over ⅛%. Actually, the probable error is almost certain to be less, since surely some of the errors would cancel out some other errors. One might guess that the overall error is about <¼%.

After making all necessary (small) corrections, the measured Stefan–Boltzmann constant is

$$\sigma_{meas} = \frac{W_c}{A\epsilon_{eff}\, T_c^4}$$

where

W_c = watts radiated out of the aperture

A = area of aperture, cm^2

ϵ_{eff} = effective emissivity of aperture

T_c = temperature, °K

Measurements of the Stefan–Boltzmann constant σ_{meas} made at four different temperatures are shown in Table 3. The second column gives the wavelength for maximum spectral intensity λ_{max} characteristic of the radiating temperature. The third column gives the coating emissivity ϵ_{cal} and the effective cavity emissivity $\epsilon_{cav_{eff}}$ calculated by methods of Ref. 3 and 4.

The trend of decreasing disagreement as the temperature is increased is probably due to small, as yet unrecognized, systematic effects, or possibly due to under- or over-corrections. The measured quantities from which the Stefan–Boltzmann constant is obtained are smaller at the lower temperatures, hence, are less accurate than the larger quantities resulting from the higher temperature measurements. Nevertheless, all values of the Stefan–Boltzmann constant obtained here are in closer agreement with the theoretical value than any heretofore published. Table 4 gives representative values of σ obtained by several experimenters in previous years, all of which values are higher than the theoretical value.

Table 4. Some previous measurements of the Stefan–Boltzmann constant

Experimenter	Year	σ_{meas}, W cm^{-2} °K^{-4} × 10^{-12}	$\sigma_{meas}/\sigma_{th}$
Coblentz (Ref. 5)	1915	5.722	1.009
Wachsmuth	1921	5.73	1.011
Hoffmann	1923	5.764	1.017
Kussmann	1924	5.695	1.022
Hoare	1928	5.736	1.012
Mendenhall	1929	5.79	1.021
Muller	1933	5.771	1.017
Eppley and Karoli (Ref. 6)	1957	5.772	1.017
Present work	1968	5.6866	1.003

While the value obtained in this work is also higher, it approaches the theoretical value more closely.

The theoretical value of the constant is

$$\sigma_{th} = \frac{2\pi^5 k}{15h^3c^2}$$

Substituting the presently accepted values[1] of k, h, and c gives 5.6697×10^{-12} W cm^{-2} °K^{-4}. The more closely the measured value σ_{meas} agrees with σ_{th} the greater the presumed accuracy of the radiometer, and the greater the confidence that can be placed in measurements made with it.

References

1. Kendall, Sr., J. M., *The JPL Standard Total-Radiation Absolute Radiometer*, Technical Report 32-1263. Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1968.
2. Blevin, W. R., and Brown, W. J., "Black Coatings for Absolute Radiometers," *Metrologia*, No. 2, pp. 139–143, 1966.

[1]The value of $\sigma = 5.6697 \times 10^{-12}$ was adopted by the National Bureau of Standards from general physical constants recommended by the National Academy of Science–National Research Council (Ref. 7).

3. Truenfels, E. W., "Emissivity of Isothermal Cavities," *J. Opt. Soc. Am.*, Vol. 52, pp. 1162–1171, 1963.

4. Sparrow, E. M., and Jonsson, V. K., "Radiant Emission Characteristics of Diffuse Conical Cavities," *J. Opt. Soc. Am.*, Vol. 53, pp. 816–821, 1963.

5. Coblentz, W. W., *Present Status of the Determination of the Constant of Total Radiation From a Black Body*, Bulletin of the National Bureau of Standards, Washington, D.C., Vol. 12, p. 553, 1915–1916.

6. Eppley, M., and Karoli, A. R., "Absolute Radiometry Based on Change of Electrical Resistance," *J. Opt. Soc. Am.*, Vol. 47, pp. 748–755, 1957.

7. National Bureau of Standards Technical News Bulletin, U.S. Government Printing Office, Washington, D.C., Oct. 1963.

B. Primary Absolute Cavity Radiometer of Wide Spectral Range, *J. M. Kendall, Sr., and C. M. Berdahl*

1. Introduction

In developing the primary absolute cavity radiometer (PACRAD) for measurement of total radiation in UV, visible, and IR ranges, a design was arrived at, which not only minimizes unwanted effects from internal heat transfer, but also lends itself to straightforward computation of the magnitudes of the effects which affect accuracy. These effects are either compensated out, calibrated out, or are eliminated by applying computed correction factors. The correction factors, nine in number, are all small, altogether amounting to $<0.25\%$. The total uncertainty of the correction factors is $<0.2\%$, with overall indicated error $<0.3\%$.

When considered as a radiometric standard, the PACRAD is a receiving instrument for measuring radiation intensity. In contrast, a standard source of radiation generates a radiant output of accurately known intensity at a particular position with respect to the source.

The radiometer and associated electronics are described; how it functions is explained; and how the correction factors were arrived at is discussed. Data from a comparison of PACRAD's and Eppley Angstrom pyrheliometers are given.

2. Description of the Radiometer

Figure 4 shows the radiometer and associated equipment, while Fig. 5 shows a schematic diagram of the arrangement of the parts of the radiometer, and gives nomenclature. Briefly, the radiometer consists of a view limiting aperture, a view limiting tube, a "muffler," a massive thermal guard (which is also a heatsink), and a cavity receptor.

The cavity receptor, internally coated with Parsons' black lacquer (Ref. 1), converts incoming radiation into heat which is conducted to the heatsink by the thermal resistor. A symmetrically arranged cavity and thermal resistor provide compensation for time rate of change of temperature of the heatsink. The heatsink is supported in a Dewar flask to provide an isothermal environment.

Cavities and thermal resistors are fabricated from 0.005-in.-thick pure silver by spinning and soldering. The cavity assembly is thermally joined to the heatsink by the thermal joint ring.

A thermopile measures the temperature difference across the thermal resistors, to indicate a measure of incoming radiation intensity. The heater winding located on the conical portion of the cavity is completely surrounded by the cone shield; all heat produced electrically is transferred to the cavity cone, thus providing near-perfect equivalence of electrical heating to radiation heating.

Figure 6 is a detailed diagram of the cavity assembly, and Fig. 7 is a simplified schematic diagram of the basic electronic circuit. The precision potentiometer shown in Fig. 4 measures the electrical quantities E_p, E_r, and e or e_{cal}.

The incoming radiation passes through the aperture, enters the cavity, is absorbed by the internal absorptive coating, and is converted into heat in the irradiated area of the cavity cone. This heat, inflowing through the thermal resistor into the heatsink, causes a temperature drop proportional to the intensity of the irradiance. The temperature drop in the thermal resistor causes the thermopile to generate an electromotive force (emf) e which is proportional to the incoming intensity of radiation, W/cm^2. The net magnitude of the intensity is $W/cm^2 = Ke$, where K is the calibration constant of the radiometer. To determine K, the radiometer is capped so that no radiation enters. Power ($W_{elec} = E_pI$, where $I = E_r/r$) is then applied to the heater winding on the cavity cone to produce heating accurately equivalent to some convenient intensity of irradiance. From the resulting thermopile emf, e_{cal}, the calibration constant is $K = W_{elec}C_f/e_{cal}$. C_f is a correction factor calculated from miscellaneous effects (discussed later) and differs from unity by $<¼\%$; for the model shown in Fig. 4, $C_f = 0.99981$.

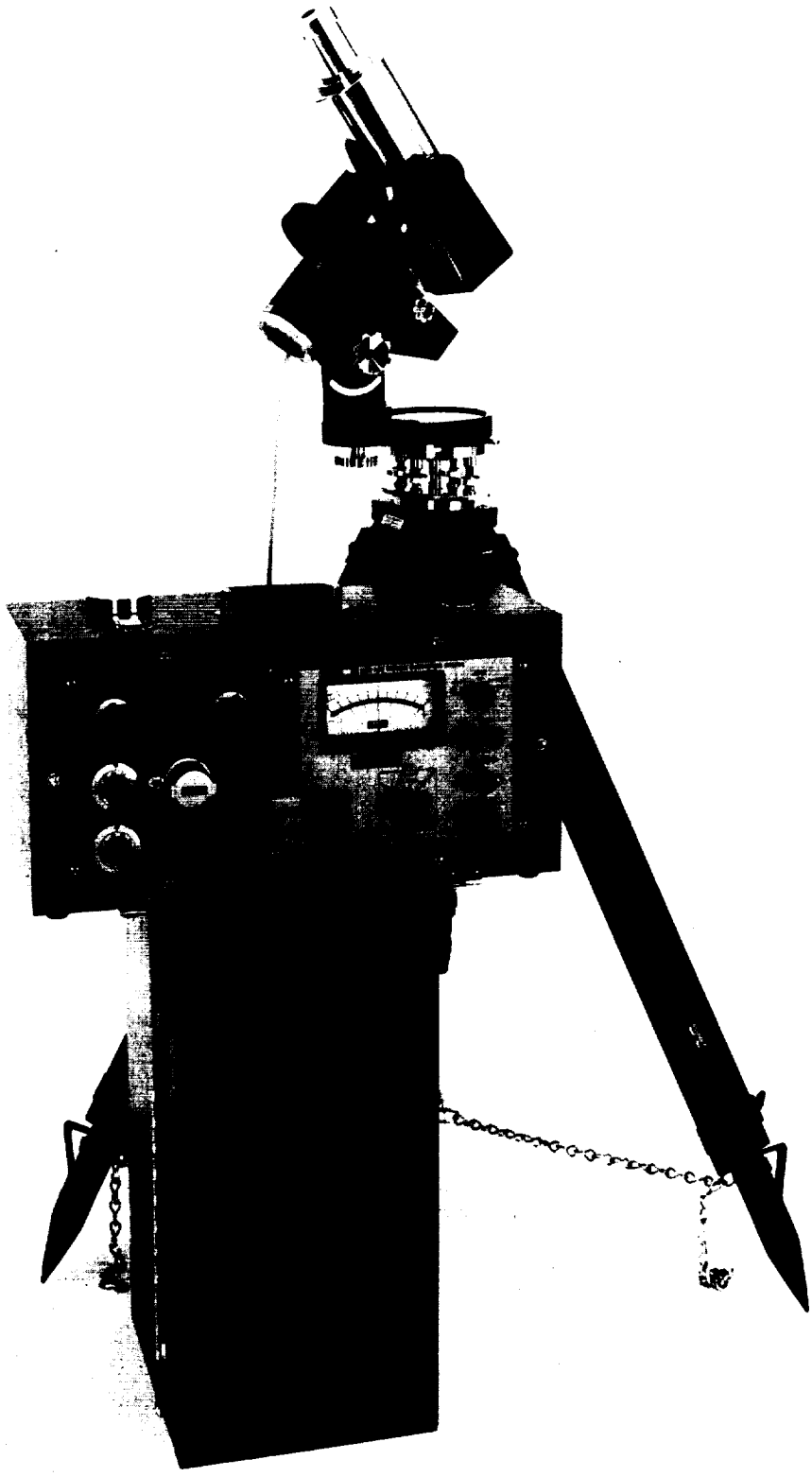


Fig. 4. Radiometer and associated equipment

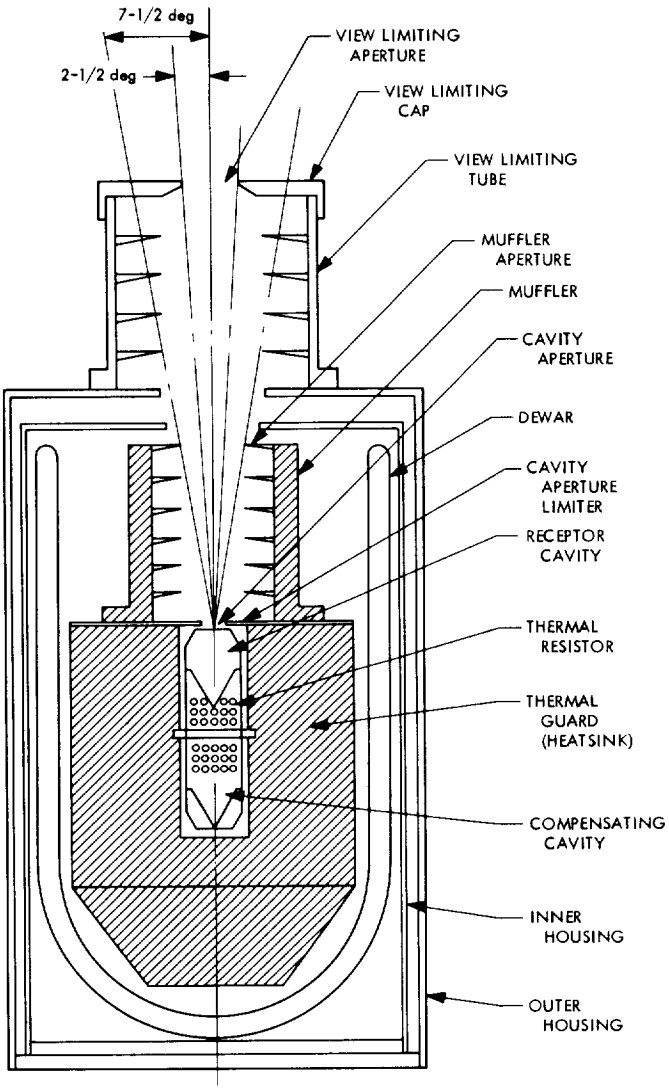


Fig. 5. Radiometer schematic

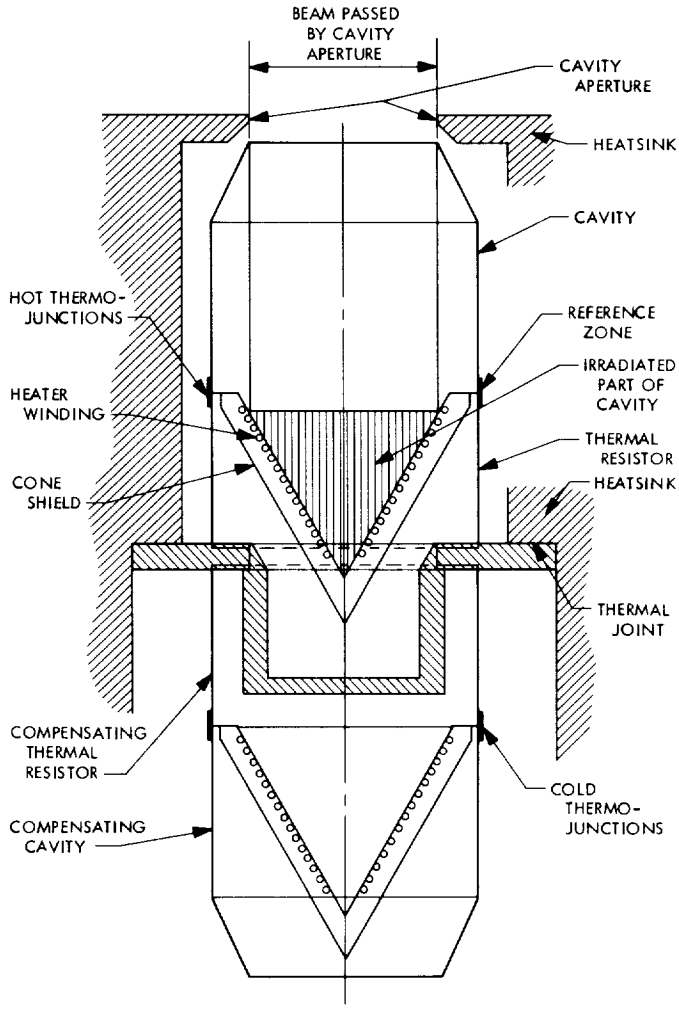


Fig. 6. Cavity assembly diagram

During measurement when the cap is off the radiometer, a certain amount of radiation W_L is emitted through the aperture by the cavity to escape into external space. The amount escaping depends on the temperature of the cavity, on aperture area, and on the size of the radiometer acceptance angle. Quantitatively, the escaping radiant power is given by $W_L = A\epsilon\sigma T^4 F_{1\text{-}2}$, where

A = aperture area, cm^2

ϵ = effective emissivity of the cavity (0.999)

σ = Stefan–Boltzmann constant (5.6697×10^{-12} W cm^{-2} °K^{-4})

T = cavity temperature, °K

$F_{1\text{-}2}$ = a factor determined by the radiometer acceptance angle (for a 5-deg acceptance angle, $F_{1\text{-}2} = 0.002$; for 2π steradians acceptance, $F_{1\text{-}2} = 1$).

The measured value of irradiance is

$$\text{W/cm}^2 = Ke + W_L/A$$

For the model shown in Fig. 4, $K = 0.1193$ W/cm^2/mW and $W_L/A = 0.00008$ W/cm^2.

3. Heat Transfers–Wanted and Unwanted

With the radiometer capped and with no electric heating, the radiometer comes to an overall isothermal state (to within <0.005°C) within a minute or two. There is no heat transfer anywhere in the radiometer; the thermopile generates no emf.

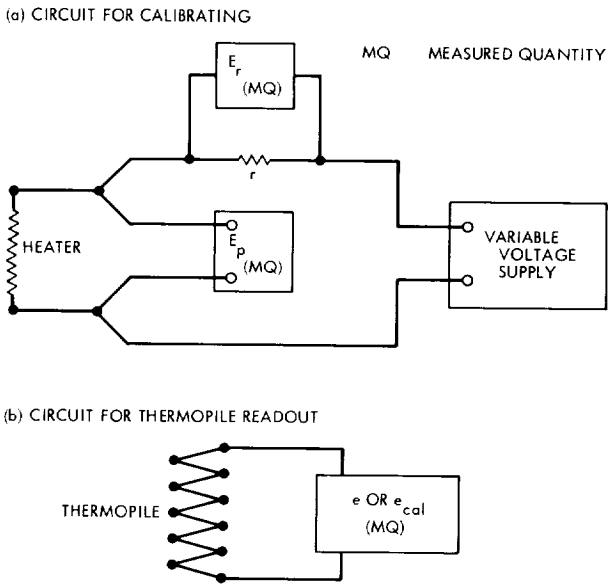


Fig. 7. Electronic circuit schematic

On the other hand, with the radiometer uncapped and with radiation coming into the cavity cone, most of the heat so produced flows through the metallic thermal resistor to the heatsink and causes the thermopile to generate an emf.

In addition to this wanted heat transfer through the metallic thermal resistor, air conductive and radiative transfer (unwanted) from the cavity to the heatsink occurs, and bypasses the thermal resistor. The thermopile then produces no output emf from these flows. The effect on accuracy of the radiometer is not as great as might be imagined. In fact, there is no loss of accuracy. The calibrating heater winding is located exactly in the same place on the cavity cone illuminated by incoming radiation. The unwanted air conductive and radiative heat transfer are equally effective with electric heating in bypassing exactly the same proportion of the heat generated in the cavity. In other words, heating at calibration is accurately equivalent to radiative heating by incoming radiation.

The above statements need some qualification. When the fine details of unwanted heat transfers are searched out, about nine different kinds of small unwanted flows are recognized. Figure 8 shows a schematic representation of the heat transfers. The lack of perfect absorptivity of the Parsons' black coating, and the effect of thermal resistance of it are considered below, but the remaining seven flows, omitted here, are considered in detail in Ref. 2. The correction factors and their estimated uncertainties are given in Table 5. The correction factors are necessitated by non-equivalence of electric heating to radiative heating. Deviations from exact equivalence occur when

(1) Difference of temperature distributions occur between radiative heating and electric heating.

(2) A part of the incoming radiation is reflected out, or is emitted as IR, through the aperture.

(3) A thermal resistance temperature drop occurs in the coating.

(4) IR radiation is emitted into external space due to cavity temperature.

Table 5. Summary of correction factors and estimates of associated uncertainties

Parameter	Correction factor	Uncertainty
Absorptivity of cavity	1.00115	±0.00050
Thermal resistance of cavity coating	1.00007	±0.00005
Difference of temperature distributions in cavity cone between radiation heating and electric heating	1.00000	±0.00005
Reflected radiation out of cavity cone absorbed by cavity cylinder	1.00029	±0.00010
Air conduction to cavity cylinder from radiation heating of cavity aperture	0.99990	±0.00005
Re-reflected radiation from muffler into cavity	0.99991	±0.00003
Non-equivalent heat flow from cone shield to heatsink with electric heating	0.99996	±0.00003
Uncertainty in electronic measurements	1.00000	±0.00050
Area of aperture, difference from 1 cm^2	0.99853	±0.00050
	Overall factor 0.99981	Simple sum ±0.00181
Radiation lost out of view limiting aperture for 5-deg acceptance angle, W	0.0000834	±0.00001

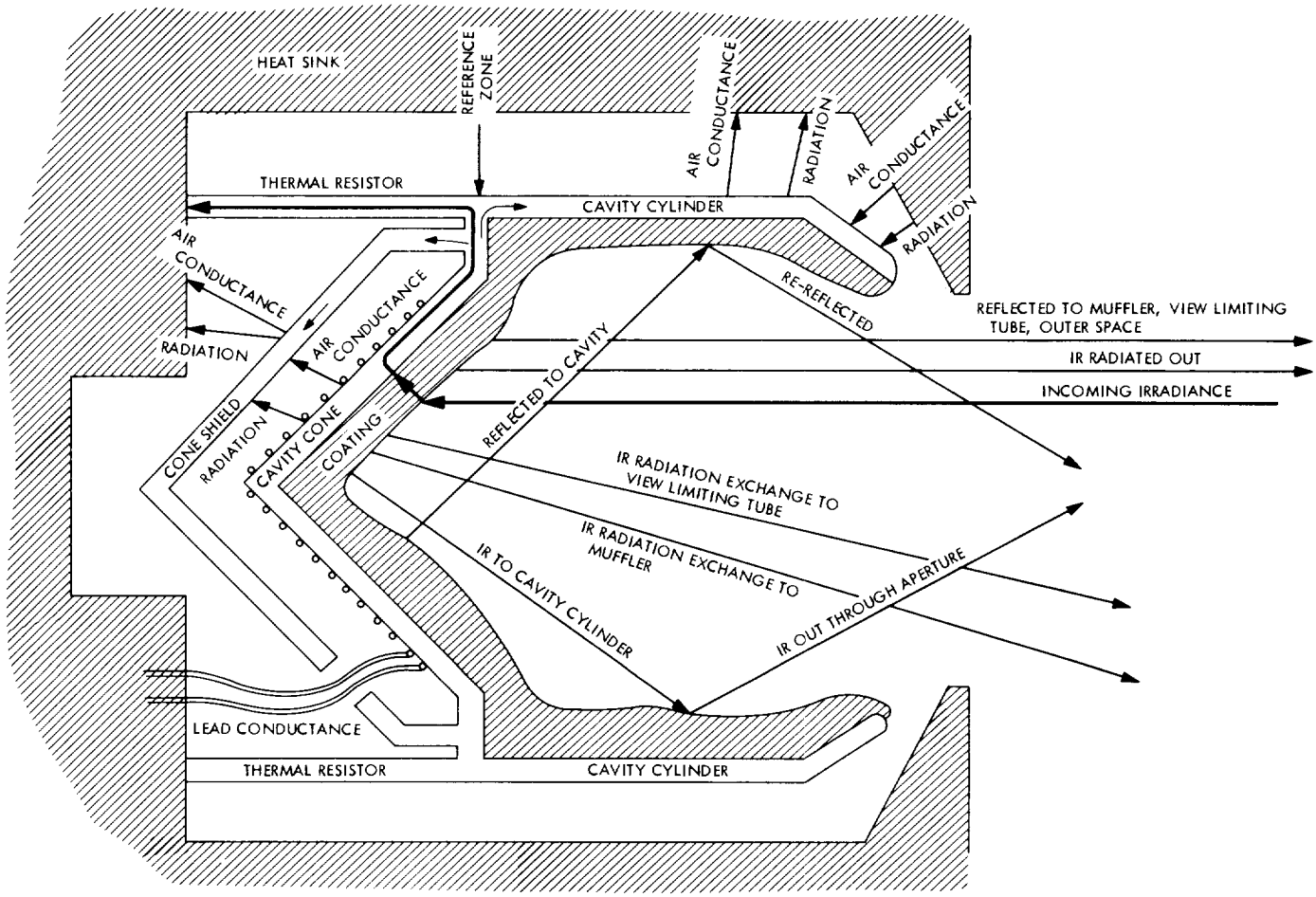


Fig. 8. Pattern of heat transfers

4. Effective Absorptivity of the Cavity

The effective absorptivity of the irradiated cavity cone to incoming radiation was calculated for incoming collimated radiation by using a modified form of Sparrow and Jonsson's method for cones (Ref. 3). The coating absorptivity of Parsons' black lacquer is given as 0.98 (Ref. 4). The effective absorptivity of the conical portion of the cavity was computed to be 0.98884 for collimated irradiance.

The cavity as a whole, however, has an absorptivity considerably greater than the absorptivity of the conical portion: 90% of the radiation reflected out of the conical portion is absorbed by the unilluminated portion of the cavity. The overall absorptivity of the cavity is 0.99885, which requires a $C_f = 1.00115$. Figure 9 shows a plot of effective absorptivity α of both the conical portion of the cavity and the overall cavity as a function of the coating alone. It is to be noted that the cavity enhancement of absorptivity is quite appreciable. The enhancement not only suppresses the effect of the thermal resistance of the coating, it also extends the spectral range which can be accurately measured.

5. Effect of Thermal Resistance of Coating Inside Cavity

The Parsons' black lacquer coating the inside surface of the cavity has thermal resistance of about 2.5°C drop through the surface for 1 W/cm² irradiance (Ref. 1). Carrying through the computations which allow for the cavity enhancement of absorptivity (Ref. 2) gives a correction factor $C_f = 1.00007$.

6. Summary of Correction Factors

Table 5 gives a summary of correction factors and estimates of the associated uncertainties for all non-equivalent effects which have been recognized. The first two items in this summary were just considered above.

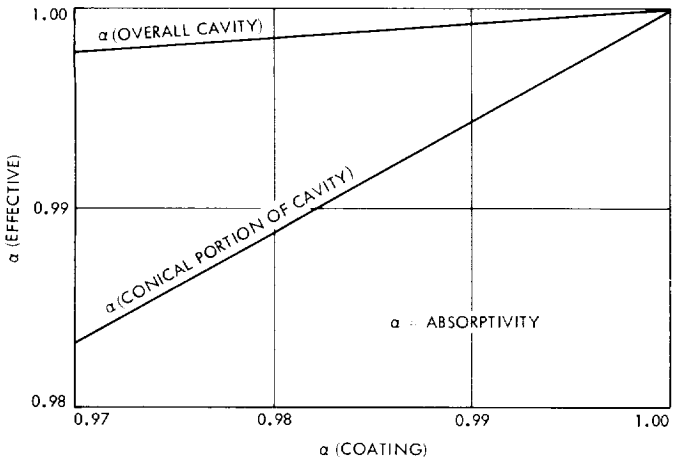


Fig. 9. Effective absorptivity of conical portion of cavity and of overall cavity vs absorptivity of coating alone for collimated incoming radiation

The remaining items are considered in Ref. 2. The overall indicated uncertainty appears to be $<0.2\%$.

7. Time Response

The effective thermal resistance associated with the cavity is composed of the metallic thermal resistor, air conduction between cavity and heatsink, and radiative coupling of the cavity with the heatsink. The cavity has heat capacity, which with the effective thermal resistance determines the time constant ($1/e$) of the cavity–thermal resistance system. This time constant determines the time response to, say, a step-function of either electric heating, or of incoming radiation heating, and indicates how long one must wait before an accurate reading of intensity (or electric heating) can be obtained.

The time constant was both calculated and measured. It was found to be 7 s for $1/e$ response. For a measurement reading of the thermopile output to settle to the final value within 0.05% requires about 8 time constants, or 1 min.

8. Concluding Remarks

Three models of PACRAD have been built. PACRAD I and PACRAD II are virtually identical; both have heavy copper heatsinks. PACRAD III has a magnesium heatsink and no Dewar flask. It weighs 2.16 lb. Tests have shown that all three models agree with one another within 0.15%.

Several tests have been made between PACRAD's and Eppley Angstrom pyrheliometers. Typical results of such comparisons are shown in Table 6 for a run at Table Mountain, California, on April 23, 1969. PACRAD's II and III pyrheliometers serial 9000 and 7257 were compared. It is to be noted that both PACRAD's and pyrheliometers gave very consistent results, with the pyrheliometers giving measured values of solar intensity 2.3% lower. Work is being done to resolve the discrepancy.

References

1. *Parsons' Black Lacquer.* The Eppley Laboratory, Newport, R.I.
2. Kendall, J. M., Sr., *Primary Absolute Cavity Radiometer PACRAD*, Technical Report 32-1396. Jet Propulsion Laboratory, Pasadena, Calif., July 15, 1969.
3. Sparrow, E. M., and Jonsson, V. K., "Radiant Emission Characteristics of Diffuse Conical Cavities," *J. Opt. Soc. Am.*, Vol. 53, pp. 816–821, 1963.
4. Blevin, W. R., and Brown, W. J., "Black Coatings for Absolute Radiometers," *Metrologia*, No. 2, pp. 139–143, 1966.

Table 6. Radiometer comparison at Table Mountain (April 23, 1969)

PST	PACRAD				Angstrom pyrheliometer				Ratio of averages Angstrom/PACRAD
	II, mW	III, mW	Average, mW	Ratio II/III	A serial 9000, mW	B serial 7257, mW	Average, mW	Ratio A/B	
11:12	104.57	104.38	104.47	1.0018	101.76	102.03	101.89	0.997	0.975
11:15	104.71	104.52	104.60	1.0018	101.90	102.03	101.96	0.999	0.975
11:37–11:39	103.55	103.51	103.53	1.0004	101.01	101.22	101.11	0.998	0.977
11:42	104.02	103.95	103.98	1.0007	101.59	101.79	101.69	0.998	0.978
11:44–11:46	103.31	103.21	103.26	1.0010	100.96	101.04	101.00	0.999	0.978
	Averages of ratios			1.00114				0.9982	0.9766[a]

[a]This average value indicates that, as a group, the measured values of the PACRADs were 2.3% higher than the Angstrom pyrheliometer values.

XVI. Aerodynamic Facilities

ENVIRONMENTAL SCIENCES DIVISION

A. Terminal Dynamics Drop Tests Performed in the Vertical Assembly Building at the Kennedy Space Flight Center, *P. Jaffe*

During September 1968, a test was performed in the vertical spin tunnel of the Langley Research Center to determine the stability characteristics of rolling blunt planetary entry configurations in the slow-speed terminal flight regime. Theoretical research has indicated that these blunt configurations can become unstable at very small roll rates in this regime and that the degree of instability increases with increasing roll rate (Refs. 1 and 2). The data from that test are still being reduced; however, the few runs that were reduced indicate results that deviate from the predicted motion in a somewhat erratic manner. The α-β motion was, in general, circular as anticipated, but superimposed on it were spikes and bumps. The question has been raised—are these perturbations due to turbulence or non-uniform flow, which could invalidate the data, or are they real? In an attempt to substantiate the data and also obtain terminal information in still air, a drop test was performed in the vertical assembly building (VAB) at the Kennedy Space Flight Center.

The VAB is an excellent facility for performing drop tests. Within the building, there is a clear drop area almost 470 ft high and substantially better than 60 × 60 ft in cross section. The air currents within the building can be controlled so that there is virtually no movement. There is quick and easy access to the drop platform and all manner of test support within the building.

To insure that meaningful information would be obtained, 6-deg-of-freedom trajectories were computed for the principal model, a 60-deg (half-angle) blunted cone weighing 4.45 lb and having a base diameter of 14 in. One of the ground rules followed in the test was that the same models used for the Langley test would be used for the VAB test. Figure 1 contains the time and velocity versus drop distance computation. The simulated drop took about 8.5 s. After 3 s, the model had dropped 110 ft and the velocity had grown to about 88% of its terminal value of 65 ft/s. It is during the last 5.5 s, when the velocity is close to terminal, that the most valuable information is obtained.

Simulated total angle-of-attack histories of the same model for three dynamic stability coefficients ($C_{m_q} + C_{\dot{m}_\alpha}$) at roll rates of 4 and 8 rad/s, respectively, are shown in Figs. 2a and 2b. In general, the α-β motion was circular, either damped or divergent. These figures demonstrate the dependence and sensitivity of the motion on the dynamic stability coefficient. For the case of maximum sensitivity (8 rad/s), the amplitudes corresponding

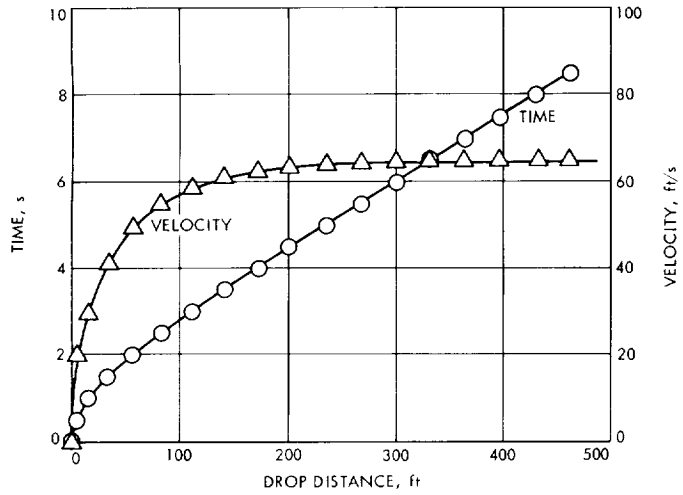


Fig. 1. Time and velocity vs drop distance

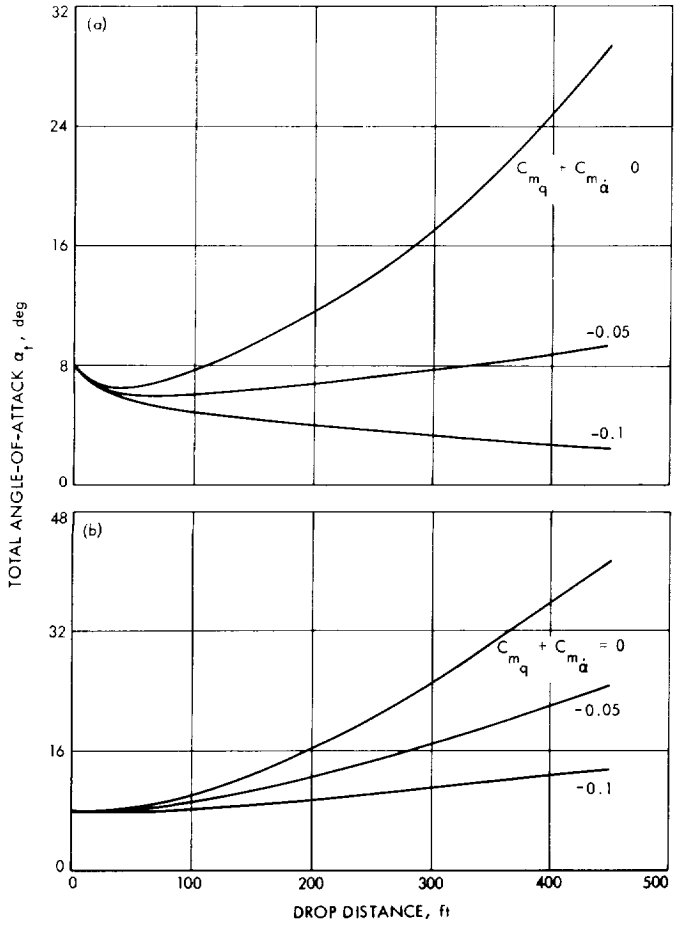


Fig. 2. Total angle-of-attack vs drop distance: (a) roll rate = 4 rad/s, (b) roll rate = 8 rad/s

to dynamic stability coefficients of 0.0 and −0.05 differed by 20 deg by the end of the drop; and, for the least sensitive case (dynamic stability coefficients between −0.05 and −0.1 at 4 rad/s), the difference was 6.5 deg. Therefore, if the total angle-of-attack could be measured during the test to 3 deg or better, the dynamic stability coefficient could be determined to at least 0.05, and probably better.

The following scheme was settled upon, after considering several alternative methods, for determining the angle-of-attack. A small 3-V light was mounted inside the model as far forward from the base as possible on the axis of symmetry. The original base was replaced by white cardboard with an 8-in.-diameter section removed from its center. From the back, an observer would see a fine light against a dark background, which is framed by a white annulus. As the model's angle-of-attack is increased, the light would appear to move closer to the inner edge of the annulus, and the angle-of-attack could be determined from the apparent displacement of the light from the center, as shown in Fig. 3. Checks were made at JPL with a variety of telephoto lenses, and it was concluded that by using the proper combination of lenses the angle-of-attack could be determined to ±3 deg.

The actual test was performed during the three-day period June 23 through June 25, 1969. It was carried out at the north end of the transfer area between bays 3 and 4. The models were dropped from the catwalk of the 250-ton crane, which is a distance of 466 ft from the floor. A 40 × 40-ft net was suspended about 7 ft above the floor to recover the models. Fortunately, all but one of the models landed in the net, undamaged. To minimize disruption of the activities in the VAB, the test was performed during the second shift (4 p.m. to midnight). Five different configurations were tested: 60-deg (half-angle) sharp and rounded edge blunted cones; 45-deg, 51.5 deg, and 70-deg sharp edge blunted cones; and a tension shell configuration. In addition, there were two different moment-of-inertia models in the 60-deg sharp edge class. Thirty-one drops were made during the test at roll rates ranging from 0 to 31 rad/s.

A countdown procedure was used during the test. After getting a clearance from the safety and photographic people, a count of −30 s was initiated. Prior to the countdown, the model was mounted on the combination support and spin-up mechanism which grips the model at the base with three fingers. At −20 s, the model light was

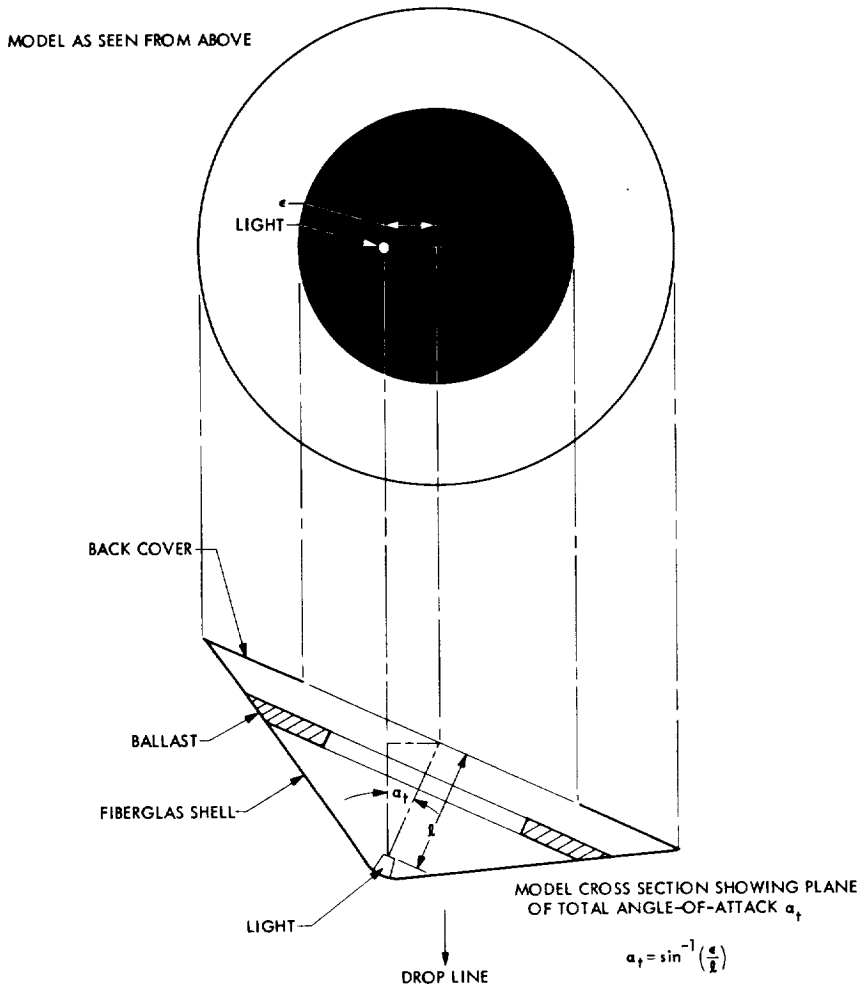


Fig. 3. Model cross section and angle-of-attack determination

turned on, the model was spun up, and placed into position for the drop. Between −5 and −10 s, the cameras were turned on. Simultaneously, at the count of MARK (zero) the model was released and a light on the floor turned off. Walkie-talkies provided the communications link. Six 16-mm cameras were used for each drop: four were clustered around the drop point looking at the base of the model with fixed telephoto lenses ranging from 3 to 10 in., and two were perpendicular to the drop line, one 154 ft from the floor and the other 356 ft from the floor. The upper cameras nominally ran at 100 frames/s and the side cameras operated at 400 frames/s.

Lighting the drop chambers was one of the major problems. Additional lights were used to fill in some of the key dim areas, and some of the permanent lights were redirected into the drop area. In addition, a bank of searchlights placed at the far south end of the transfer area (about 500 ft away) aided in increasing the overall light level. Even with the additional lighting, there were still dimly lit areas. However, by tailoring the *f* stop of each camera to the drop region it covered, adequate movies were obtained. In addition to the lighting problem, the timing electronics which were to place a chronological statement and a timing pulse on the edge of the film continually caused us difficulty. Only about 20% of the drops had timing on all cameras. High humidity added another problem by causing the film to swell and stick, making it very difficult and timeconsuming to load film.

In spite of these and other difficulties, the data from a large percentage of the drops are reduceable. A sequence of photographs of one of the 60-deg sharp edge cone drops taken with a 6-in. lens is shown in Fig. 4; the time between photographs is 0.04 s. The completed reduction of this drop showing the total angle-of-attack as a function of time is shown in Fig. 5. As can be seen,

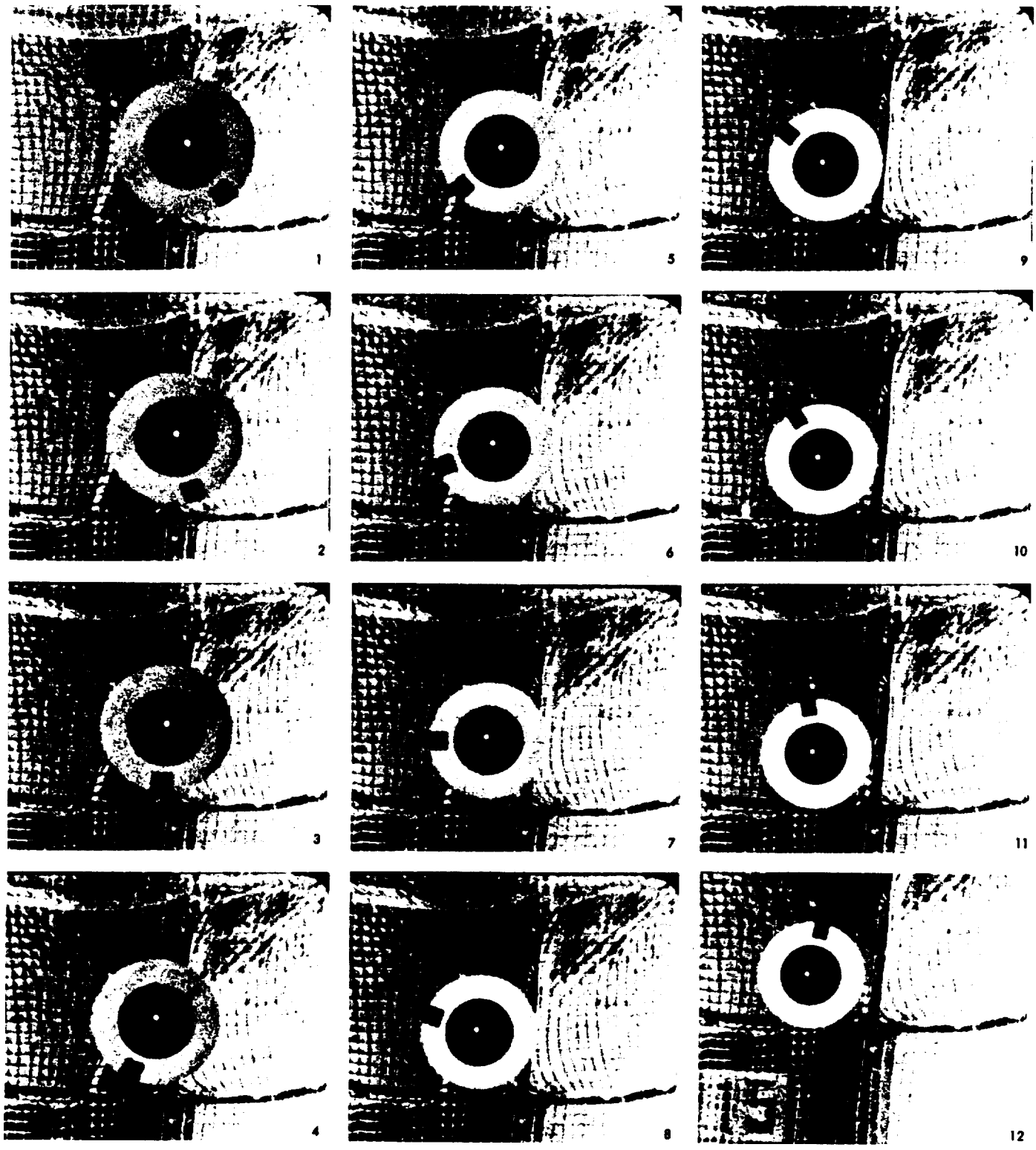


Fig. 4. Photographic sequence from drop 15 taken with a 6-in. lens

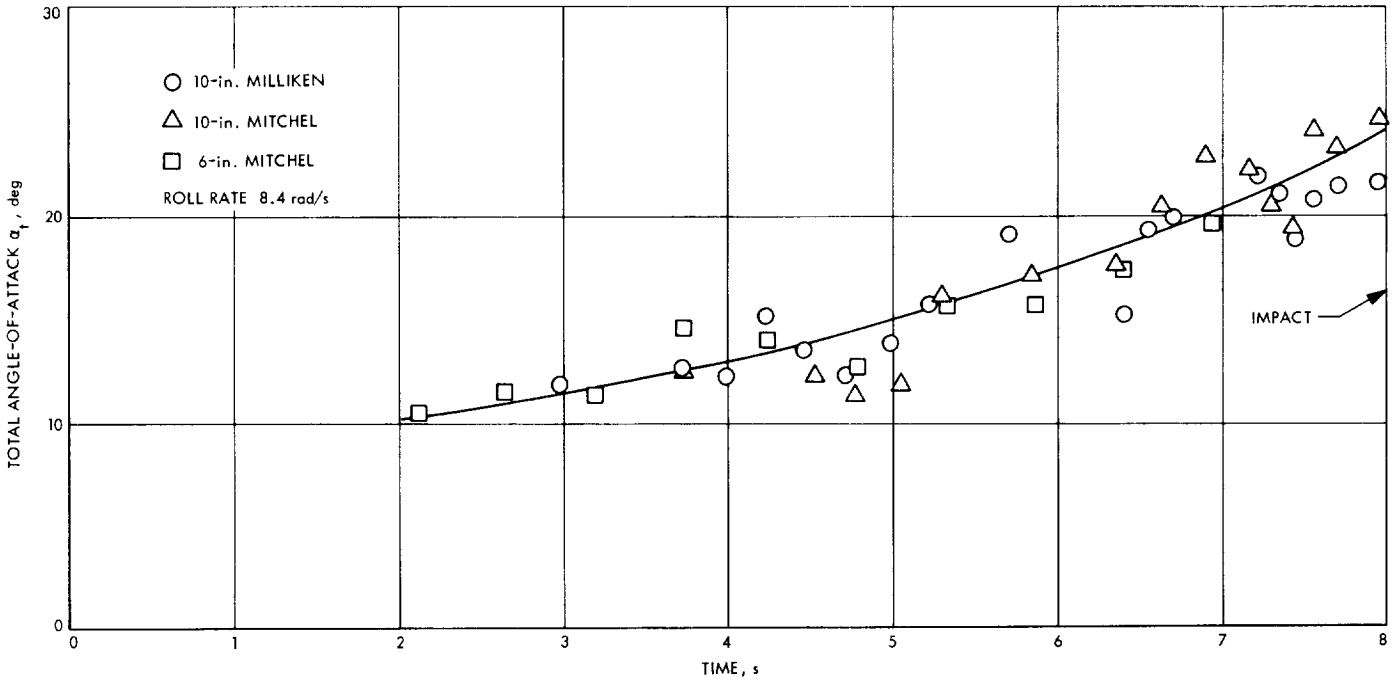


Fig. 5. Total angle-of-attack history for drop 15

the data from the three cameras do fall in a band of ± 3 deg, as expected. Also, the non-fluctuation in amplitude indicates that the model's α-β motion was either circular or very near circular, as predicted. Upon comparing the results with 6-deg-of-freedom computations, the dynamic stability coefficient for the drop was determined to be -0.1 ± 0.02. At the current time, only one drop has been reduced; further conclusions as well as comparisons with the Langley data must wait for additional reductions, which are underway. Regardless of what the data indicates, both the technique for determining the angle-of-attack and the utilization of the VAB as a drop test facility have been demonstrated.

References

1. Shirley, D. H., and Misselhorn, J. E., "Instability of High-Drag Planetary Entry Vehicles at Subsonic Speeds," *J. Spacecraft Rockets*, Vol. 5, No. 10, Oct. 1968.
2. Jaffe, P. "Terminal Dynamics of Atmospheric Entry Capsules," *AIAA J.*, Vol. 7, No. 6, Jun. 1969.

B. Shock Tube Thermochemistry Program,

W. A. Menard

A computer program for calculating the chemical equilibrium properties associated with moving, standing, and reflected normal shocks has been developed (Ref. 1). The program calculates thermodynamic properties and chemical composition from basic spectroscopic data. Both dissociation and ionization over an unlimited temperature range can be considered. The program can treat initial mixtures consisting of up to ten gases. At present, input data have been compiled for sixty-six species (listed in Table 1).

An analysis of the computational procedures and accuracy of the results is being documented.[1] Partition functions are calculated using approximate methods which give the thermodynamic properties and the composition of principal species with only a few percent uncertainty, and the composition of trace species to about 10% uncertainty, for high-temperature gaseous mixtures encountered in shock-tube research.

To make the thermochemistry data available to the scientific community, computer results are being tabulated in report form.[2] Each volume of the series is for a

[1]Horton, T. E., *The Computation of Partition Functions and Thermochemistry Data for Atomic, Ionic, Diatomic, and Polyatomic Species*, Technical Report 32-1425. Jet Propulsion Laboratory, Pasadena, Calif. (to be published).

[2]Menard, W. A., and Horton, T. E., *Shock-Tube Thermochemistry Tables for High-Temperature Gases, Vol. I: Air, Vol. II: 90% CO_2–10% N_2, etc.*, Technical Report 32-1408. Jet Propulsion Laboratory, Pasadena, Calif. (to be published).

Table 1. Molecular and atomic species considered in program

C_2	CO_2	O^-
N_2	C_2H	O^+
O_2	C_2H_2	O^{++}
H_2	N_2O	O^{+++}
CN	CH_2	H^-
CO	CH_3	H^+
CH	CH_4	Ar^+
NO	CHO^+	Ar^{++}
NH	CH_2O	Ar^{+++}
OH	NO_2	Ne^+
CO^+	NH_3	Ne^{++}
CH^+	O_3	Ne^{+++}
N_2^+	H_2O	He^+
NO^-	HCO	He^{++}
NO^+	C^-	C
O_2^+	C^+	N
O_2^-	C^{++}	O
OH^+	C^{+++}	H
OH^-	N^-	Ar
H_2^+	N^+	Ne
C_3	N^{++}	He
C_2N_2	N^{+++}	e^-

particular gas, set of gases, or gas mixture. Volumes completed or near completion are:

Vol. I. Air (78.08% N_2–20.95% O_2–0.97% Ar)
Vol. II. 90% CO_2–10% N_2
Vol. III. He, Ne, Ar
Vol. IV. N_2
Vol. V. CO_2

Other volumes will be prepared as the need for new gases arises.

In each of the volumes, the thermodynamic properties and species concentrations are tabulated for moving, standing, and reflected shock waves. Initial pressures range from 0.05 to 50 torr (which give pressures behind the shock waves from 0.001 to 100 atm), and temperatures range from 2000 to 100,000°K.

Reference

1. Horton, T. E., and Menard, W. A., *A Program for Computing Shock-Tube Gasdynamic Properties*, Technical Report 32-1350. Jet Propulsion Laboratory, Pasadena, Calif., Jan. 15, 1969.

XVII. Solid Propellant Engineering

PROPULSION DIVISION

A. Flame Spreading in Solid Propellant Rocket Motors, *R. L. Klaus*

1. Introduction

This is the first reporting of a program whose goal is to develop a mathematical model and computer program to describe the ignition and flame spreading phase of a burning solid propellant rocket motor and to compare this model with experimental results to be obtained at JPL. The overall program will consist of five major steps: (1) The development of a mathematical model and computer program that allows the description of ignition and flame spreading, allowing for spatial variation of temperature and pressure throughout the rocket motor chamber and nozzle. It is formulated to accommodate the refinements introduced by the succeeding steps. (2) An investigation of an incorporation into the model of the most modern treatment of heat transfer as applicable to a rocket motor grain, and in particular, the investigation of heat transfer through a turbulent boundary layer with mass addition. (3) The development of a heat transfer correlation that includes the effects of the impingement of hot particles on the propellant surface, which is characteristic of many igniters. (4) The development and incorporation into the model of a proper criterion of ignition. (5) A more detailed analysis of igniter plume impingement on the propellant surface with particular application to aft-end ignition.

Reported here is the first phase of Step 1, namely the formulation of the overall concept, presentation of the governing differential equations, and a suggested approach to their numerical solution.

Present understanding of flame spreading is largely based on the concepts developed at Princeton University (Ref. 1). Probably the most serious limitation of this approach is its failure to account for spatial variation of temperature and pressure in the rocket chamber, even though it is admitted that this is a serious consideration. Moreover, this also clouds the heat transfer analysis, since it must be based on a global temperature and pressure in the chamber rather than on local temperatures and pressures. Nevertheless, the approach has been promising enough to warrant the present refinements.

Attempts have been made at the United Technology Center (Ref. 2) and Aerojet-General (Ref. 3) to refine the heat transfer analysis by recourse to new experimental data. This work is to be further evaluated in Step 2 of this

program in the light of the information on local temperature and pressure in the chamber. Recent work (Ref. 4) has focused on the development of a model of the igniter plume impingement with special application to aft-end ignition. This work is to be incorporated into the present model in Step (5) of the program.

2. Mathematical Model

The mathematical model is schematically represented in Fig. 1. The x-coordinate is taken to be distance along the centerline of the motor. The description is taken to have one dimension x in addition to time t. Thus all quantities are taken to be functions of x and t, and the equations are partial differential equations in x and t. The contour of the motor is specified by giving A_p (port or cross-sectional area) and the perimeter of the motor, both as functions of x and possibly t. The contour of both the chamber and nozzle are specified in this manner all the way to the exit plane. For part of the contour, propellant is specified; the rest of the contour consists of inert material.

The geometry of the igniter plume is specified as input information. Its point of impingement on the propellant is specified as part of this information. This data is supplied from experimental observations.[1] The cross-sectional area of the igniter plume is specified as a function of x. The presence of the igniter and other inert hardware in the chamber may be accounted for by suitably modifying A_p in the region where such hardware is located.

The igniter is modeled by specifying various conditions at the beginning of the plume as functions of time. If the igniter flow is subsonic, the gas stagnation temperature and either pressure or mass flux are specified. If the igniter gases are supersonic, all three of these conditions are specified and some sort of shock structure must exist through which the velocity of the igniter gases is reduced to the subsonic values which exist through the bulk of the chamber. This shock structure can, under certain circumstances, be quite complex. For the purpose of this analysis, however, it will be assumed that one normal shock exists in the plume of the igniter, upstream of its point of impingement.

The three conservation equations may be applied across the shock and lead to the following three relationships:

$$\frac{M_2^2}{M_1^2}\left[\frac{\gamma M_1^2 + 1}{\gamma M_2^2 + 1}\right] = \frac{1 + \left(\frac{\gamma - 1}{2}\right) M_1^2}{1 + \left(\frac{\gamma - 1}{2}\right) M_2^2}$$

$$\frac{T_2^*}{T_1^*} = \frac{1 + \left(\frac{\gamma - 1}{2}\right) M_1^2}{1 + \left(\frac{\gamma - 1}{2}\right) M_2^2}$$

$$\frac{p_2^*}{p_1^*} = 1 + (M_1^2 - M_2^2)\left[\frac{\gamma}{\gamma M_2^2 + 1}\right]$$

[1]There is experimental evidence to indicate that under certain circumstances the igniter plume does not actually impinge on the surface. This most frequently occurs for supersonic igniters, in which case the point of maximum heat transfer is usually adjacent to the shock disk in the igniter plume. This case can be treated by artificially constructing an igniter plume such that the shock is properly located and by considering the maximum heat transfer to take place adjacent to that point.

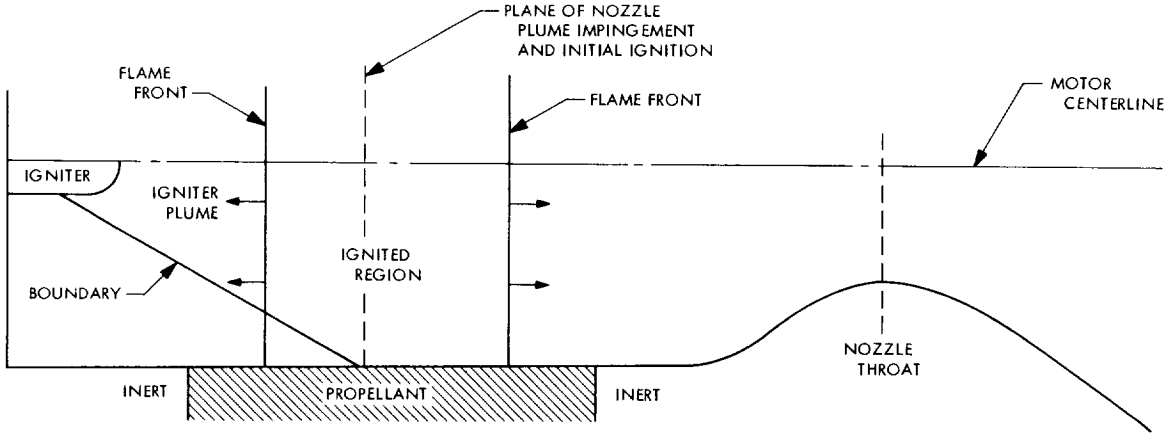


Fig. 1. Schematic of flame spreading in a solid propellant motor

where M, T^* and p^* are the Mach number, nondimensionalized temperature, and nondimensionalized pressure, respectively, and the subscripts 1 and 2 refer to locations immediately upstream and downstream of the shock. Since the problem is unsteady, the location of the shock may change with time and must be determined at each step in the calculation as the point at which these relationships are satisfied.

At the point at which the igniter plume impinges on the surface of the propellant, the surface is assumed to reach the temperature of the plume gas. Downstream of that point the surface temperature is calculated through the use of some specified heat transfer correlation, which is also an input to the program. As a first step, the correlation for Nusselt number as a function of Reynolds number used in Ref. 1 will be used. It is

$$Nu_x = \frac{hx}{kg} = 0.0734\, Re_x^{0.8}$$

The program will be written so that other, more accurate correlations may be inserted at a later date.

The igniter gases are considered to be confined to the region of the igniter plume. They displace and mix with the chamber gases initially in the region of the plume, but it is assumed that no mixing occurs across the plume boundary. There may be motion of gases toward the head end of the motor due to pressurization, but this motion takes place outside the area of the plume. Such motion is important in that it causes the motion of one of the flame fronts toward the head end of the motor.

The mathematical description of the gas dynamics of the chamber and nozzle regions is based on the unsteady, one-dimensional conservation equations, allowing for possible momentum, energy, and mass fluxes perpendicular to the direction of the main flow. The derivation of these equations as well as a discussion of the boundary conditions is given in SPS 37-56, Vol. III, pp. 179–185. There result three simultaneous partial differential equations in temperature, pressure, and Mach number. An additional complication is due to the split boundary conditions. However, it has been shown that the calculation may begin at an initial time, and by substitution differences in time for derivatives, the equations may be reduced to ordinary differential equations which then may be partially uncoupled, formally integrated, and written in a manner which lends itself to direct iterative solution.

The gas dynamics equations are coupled to the heat conduction in the solid because they necessitate a knowledge of the heat transfer to the propellant grain and the area over which the propellant is ignited.

The surface temperature and heat transfer to the propellant, on the other hand, are calculated from a knowledge of the local bulk gas temperature and the specified heat transfer correlation. This is done through solution of the one-dimensional heat conduction equation with convection boundary conditions and a time-varying bulk gas temperature and heat transfer coefficient. The solution to this problem is discussed in SPS 37-57, Vol. III, pp. 133–139. Thus both the gas dynamics and heat conduction problems must be solved simultaneously and over small time increments. This is not an insurmountable problem, however, and it is being programmed for a digital computer.

A proper criterion of ignition must be incorporated into the program. Though both experimental and theoretical considerations have led to the conclusion that the ignition temperature of a propellant is a strong function of the heat transfer history; for present purposes it is considered that the propellant ignites whenever a certain ignition temperature is reached. The inclusion of a better ignition criterion is envisioned in Step (4).

Once the first ignition occurs, there is the beginning of the movement of two flame fronts, one toward the head end and one toward the nozzle. Ignition is taken to occur as successive areas of the propellant satisfy the ignition criterion, in this case when the surface reaches the ignition temperature. Once ignition is achieved, it is assumed that steady burning at the local pressure is instantaneously achieved at that point, following the normal burning rate law. The hot gases produced by this combustion serve, of course, to accelerate the flame-spreading process.

3. Ignition and Flame-Spreading Transient

The ignition and flame-spreading transient described in this model falls into five distinct periods.

The first is the initial impingement, which begins with the firing of the igniter and ends with initial ignition of the propellant. Hot gases begin to move down the plume and mix with the gases initially in the chamber. At the point of impingement of the plume there is both motion toward the head-end (which begins the head-end pressurization) and the aft-end. The shock, if present, begins at the beginning of the igniter plume and may begin to move downstream as the flow field develops. Eventually the propellant surface at the point of impingement satisfies the ignition criterion, which terminates this phase.

The second period is initial flame spreading. Initial ignition has taken place, and under the stimulus of both the igniter and the hot gases produced by the burning propellant, heat is transferred to the propellant surface which eventually causes it to ignite. This is the mechanism of flame spreading. The flame spreads in both directions from the point of initial ignition. This period ends with the cut-off of the igniter.

The third period is a brief period of collapse of the igniter plume. For the purposes of this model the mechanism is taken to be as follows: beginning at the head end of the motor, the plume volume becomes available for the chamber gases as the final volume of igniter gases vacate that portion of the plume. The effect is to make the entire chamber finally available for the chamber gases. This period ends when that state is achieved.

The fourth period is the completion of flame spreading, where, under the stimulus of the hot gases from the partially ignited surface, the two flame fronts spread until the entire surface is ignited.

The fifth period occurs after the entire surface has been ignited and the motor begins to pressurize until the steady-state pressure- and temperature-distributions are achieved.

This summarizes the concept of the flame spreading model to be used as the basis for a computer program to describe this effect. This program is presently being written.

References

1. Most, W. J., et al., *Thrust Transient Prediction and Control of Solid Rocket Engines*, Paper 68-33, Western States Section, The Combustion Institute, Menlo Park, Calif., Oct. 29–31, 1968.
2. Kilgroe, J. D., *Studies on Ignition and Flame Propagation of Solid Propellants*, Project 2229, Final Report, United Technology Center, August 1967.
3. Micheli, P. L., *The Empirical and Analytic Modeling of the Role of the Propellant in Ignition*, Paper 68-34, The Combustion Institute, Western States Section, Menlo Park, Calif., Oct. 29–31, 1968.
4. Kilgroe, J. D., Fitch, R. E., Guenther, J. L., *An Evaluation of Aft-End Ignition for Solid Propellant Motors*, Consolidated Engineering Technology Corp., Final Report, Contract NAS 3-10297, September 1968.

XVIII. Polymer Research

PROPULSION DIVISION

A. A Relationship Between Network Chain Concentration and Time Dependence of Rupture for SBR Vulcanizates, *R. F. Fedors and R. F. Landel*

1. Introduction

In a previous study (SPS 37-36, Vol. IV, pp. 137–146) it was shown that the failure behavior of amorphous gum elastomers can be conveniently described in terms of a family of physical property surfaces in stress–strain–time (σ, ϵ, t) space. Fracture in reference to these surfaces represents some limiting boundary or discontinuity such that a distinct space curve is associated with each surface.

The projection of the family of space curves to the stress-strain plane gives rise to a family of failure envelopes. It has been shown that the high temperature portions of these envelopes are independent of the network chain concentration ν_e, provided that the reduced variable $(\sigma_b/\nu_e)(T_0/T)$, where T_0 is an arbitrary reference temperature and T, the test temperature, is used instead of σ_b itself, i.e., a plot of log $(\sigma_b/\nu_e)(T_0/T)$ versus log ϵ_b yields a superposition of the high-temperature portions of individual failure envelopes to a master curve for rupture that is essentially independent of variables such as test rate, test temperature, chemical nature of the elastomer ν_e, and the presence of small amounts of diluent (Ref. 1).

The projection of the family of space curves to the stress–time or strain–time planes gives rise to a family of curves which represent the time dependence of rupture. In view of the fact that the failure envelopes could be superposed, it was natural to wonder whether the same could be carried out with the time dependence of rupture as well. Several years ago, using published data for several elastomers, we began such a study. It was found that some semblance of superposition could be obtained by shifting the response curves along the time scale by a factor proportional to ν_e. We now wish to provide evidence, using data on a single elastomer, that such a shift does lead to at least partial superposition of the time to failure response.

2. Discussion

a. Background. In a study of the effects of statistical variability and network chain concentration on the failure envelopes of amorphous gum elastomers, it was noted that changes in ν_e corresponded, qualitatively at least, to

changes in the time scale. Thus, it was reported that a decrease in ν_e, at constant test conditions, had the same effect on the ultimate properties as if the test rate had been increased or the test temperature had been decreased, and conversely (Ref. 2). This observation suggests that ν_e might be a suitable parameter for reducing or normalizing the time scale and hence for generating master curves for the time dependence of rupture.

Subsequently, Plazek demonstrated that the long time or terminal region of the creep compliance curves for both natural rubber (NR) and styrene butadiene rubber (SBR) gum vulcanizates can be superposed to yield a master curve by using ν_e as a reduction variable for the time scale (Ref. 3).

The superposed creep compliance response is defined by Plazek in the following way:

$$J_x\left(\frac{t}{a_T a_x}\right) = J_p\left(\frac{t}{a_T a_x}\right)\frac{J_e(R)}{J_e} \tag{1}$$

where J_x is the universal creep compliance response; J_p is the response for an elastomer with a given ν_e value; $J_e(R)$ and J_e are the compliances at mechanical equilibrium for an arbitrarily chosen reference system and for a system with a given ν_e value, respectively; a_T is the usual WLF[1] time–temperature shift factor; and a_x is an empirical shift factor that relates ν_e to the time scale of the experiment. For both NR and SBR, the factor a_x was found to be related to ν_e by means of the following equation:

$$a_x = \left(\frac{C}{\nu_e}\right)^{15.4} \tag{2}$$

where C is a constant. Plazek found that NR and SBR had different C values, and this was tentatively ascribed to differences in the monomeric friction coefficients for these two materials. Equation (2) implies that for a given time scale, the greater the ν_e value, the closer the system is to mechanical equilibrium; it also indicates that small changes in ν_e have an enormous influence on the effective time scale.

It is of interest then, to ascertain whether the ν_e-time equivalence found by Plazek for small strain creep response is also applicable to the time dependence of rupture. It has already been observed that the WLF a_T factor, which was first used to relate time and temperature in the small strain region, is also applicable to large strain rupture (Ref. 4).

[1]WLF = Williams–Landel–Ferry.

To this end, we have assembled published data for the rupture behavior of SBR which will be used to test the applicability of ν_e-time equivalence to rupture.

b. Data. Several studies have been carried out on the rupture behavior of SBR. Smith investigated ring specimens of a sulfur-cured SBR gum having a constant value of ν_e (106×10^{-6} moles/cm^3), at strain rates between 0.158×10^{-3} and 0.158 sec^{-1} and at fourteen temperatures between $-67.8°$ and 98.3°C. He found that the rupture data obtained at a given temperature as a function of the strain rate could be shifted horizontally, and the result was a set of master curves relating both σ_b and ϵ_b to the reduced strain rate $(Ra_T)^{-1}$ (Ref. 4). The magnitude of the horizontal shift was equal to a_T as obtained from the WLF equation in the form

$$\log a_T = -\frac{8.86\,(T - 263)}{101.6 + T - 263} \tag{3}$$

Baranwal investigated the failure behavior of SBR, in the form of ring specimens, at strain rates between 2.10×10^{-4} and 0.526 sec^{-1}, and at eight temperatures between -30 and 70°C. The vulcanizate was sulfur-cured and had a ν_e value as estimated for equilibrium swelling of 204×10^{-6} moles/cm^3. In agreement with the findings of Smith, Baranwal observed that the rupture data could be shifted to yield a superposed curve of log σ_b and ϵ_b as a function of reduced strain rate. The magnitude of the horizontal shifts was equal to a_T, as calculated from Eq. (3) (Ref. 5).

Stress–strain measurements to failure were also carried out by Healy on a sulfur-cured SBR gum at strain rates between 2×10^{-3} and 0.2 sec^{-1} and at temperatures between -35°C and 70°C. The ν_e value, estimated from equilibrium swelling measurements, is reported as 94×10^{-6} moles/cm^3. Stress-at-break and strain-at-break were plotted against t_b. Master curves were obtained by shifting isothermal data along the time axis. The a_T factors obtained were in fairly good agreement with the values calculated from Equation (2), except at the higher test temperatures (Ref. 6).

Whereas, Smith, Baranwal, and Healy carried out their studies with vulcanizates having constant ν_e while the test temperature and test rate were allowed to vary, other studies of the failure behavior of SBR have been carried out at constant rate and constant temperature while ν_e was allowed to vary.

For example, Taylor and Darin investigated the effect of chain concentration on the ultimate properties of SBR

at 30°C employing dumbbell-shaped specimens run at constant crosshead speed. The chain concentration was varied by changing the content of the quantitative crosslinking agent decamethylene bis (methylazocarboxylate) (Ref. 7). Their data indicated that σ_b first increases, then passes through a maximum, and finally decreases as the content of curing agent is increased. Strain-at-break, on the other hand, decreases monotomically under the same conditions. In this study, the concentration of chemically crosslinked units in the network was taken equal to the concentration of curing agent present. No work was carried out to estimate the effective number of network chains, ν_e, which is generally 2 to 3 times greater than that predicted on the basis of curing agent concentration alone.

Epstein and Smith carried out a similar study using the same polymer and curing agent. However, instead of relying on the concentration of curing agent as an adequate measure of ν_e, Epstein and Smith estimated ν_e independently, using both swelling and modulus measurements. Their data obtained at room temperature on dumbbell specimens using a constant crosshead speed also demonstrates that σ_b passes through a maximum at a relatively low value of ν_e. In addition, these workers were among the first to report that the σ_b, ν_e, and the ϵ_b, ν_e response was dependent on test rate and conditions of measurement (Ref. 8).

Fedors and Landel studied the rupture behavior of sulfur-cured SBR vulcanizates as a function of ν_e at fixed temperature (28°C) and rate (4.25 min^{-1}) using ring specimens. In order to arrive at more reliable estimates for σ_b and ϵ_b, a study was made of the statistical variability of both σ_b and ϵ_b, using 46 specimens at each ν_e value and from these data, the most probable values for σ_b and ϵ_b were obtained. In this investigation also, it was found that σ_b passes through a maximum at a relatively low value of ν_e, while ϵ_b decreases monotonically with ν_e (Ref. 9).

The studies cited above constitute six independent sets of data on the failure behavior SBR gum vulcanizates having two types of crosslinks, viz., sulfur crosslinks and decamethylene bis (methylazocarbonylate) crosslinks. In the former case, the crosslinked units contain polysulfidic sulfur, while in the latter case, the crosslinked units contain carbon, oxygen, and nitrogen. In addition, three sets of data were obtained by allowing the rate and temperature to vary over wide ranges at constant ν_e. For the remaining three sets of data, ν_e was allowed to vary at constant test rate and temperature. We shall attempt to superpose these failure data, using ν_e as a reduction variable for the time scale in the manner suggested by Plazek.

In our use of these failure data, the following procedures were employed:

(1) The data as reported by Smith were reduced to a standard temperature of 263°K while most of the other data were determined at about 300°K. Hence, a_T was calculated from Eq. (2) and was used to shift Smith's data to a reference temperature of 300°K. His data are reported as log $\sigma_b\,(263/T)$ versus log $(1/Ra_T)$ and as log ϵ_b versus log $(1/Ra_T)$, and these were converted to log $(\sigma_b/\nu_e)/(298/T)$ versus log (t_b/a_T) and log ϵ_b versus log (t_b/a_T) plots. The value of ν_e used (106×10^{-6} moles/cm^3) was calculated from an equilibrium modulus value obtained at 87.8°C from a log–log plot of reduced stress versus reduced strain.

(2) The data of Baranwal were reduced to a reference temperature of 264°K. The appropriate a_T needed to shift the data to a reference temperature of 300°K was calculated from Eq. (3). The data were reported in graphical form as log $\sigma_b\,(273/T)$ and ϵ_b versus log $(1/Ra_T)$. The reduced strain rate was converted to reduced time-to-break.

(3) The data of Healy were reduced to a reference temperature of 273°K. The appropriate a_T needed to shift the data to a reference temperature of 300°K was obtained from a tabulated list of experimentally determined values.

(4) The data of Epstein and Smith obtained at a crosshead speed of 2 in./min were used. Plots of log $(\sigma_b/\nu_e)/(298/T)$ and log ϵ_b versus log t_b were constructed from the tabulated data. The ν_e values used were those estimated from equilibrium modulus measurements. They also estimated ν_e from equilibrium swelling measurements but had to assume a solvent-polymer interaction coefficient χ_1 in order to calculate ν_e using the Flory-Rehner swelling equation. For this reason, the ν_e values obtained from modulus measurements were considered more reliable. Dumbbell-shaped specimens were used, and we estimated the strain rate using the reported crosshead speed and the specimen gage length. In addition, we had to assume the strain rate to be constant even though it is well known that the strain rate experienced by a dumbbell-shaped specimen decreases with strain. This effect is negligible for our purposes.

(5) For the data of Taylor and Darin, plots of lot $(\sigma_b/\nu_e)/(298/T)$ and log ϵ_b versus log t_b were constructed from the tabulated data. For these data ν_e values were not independently measured; rather, they report only the concentration of crosslink agent. Since Epstein and Smith worked on essentially the same kinds of vulcanizates, we

assumed that both sets of data could be characterized by the same ν_e when the concentration of crosslinking agent is the same. Dumbbell-shaped specimens were used, and the strain rate was assumed constant and was estimated from the reported specimen gage length and the crosshead speed.

(6) The data of Fedors and Landel were obtained using ring specimens. Plots of lot $(\sigma_b/\nu_e)/(298/T)$ and log ϵ_b versus log t_b were constructed from tabulated data. Values of ν_e were estimated from equilibrium swelling.

In the initial attempt to effect superposition, a_x was calculated from Eq. (2) taking $C = 106 \times 10^{-6}$ moles/cm^3, i.e., the data of Smith was chosen to be the reference state and the data was plotted as log $(\sigma_b/\nu_e)/(T_0/T)$ versus log $t_b/a_T a_x$. However, it was determined, by trial and error, that much better overall superposition could be obtained by taking the exponent in Eq. (2) as 7.7 rather than 15.4 as found by Plazek for small strain creep behavior. It may be added that we have been able to superpose time-to-break data obtained by Smith (Ref. 10) for a series of Viton A-HV gum vulcanizates having different ν_e values. The value of the exponent required for superposition is also about 7.7.[2]

The results for the time dependence of σ_b are shown in Fig. 1. The data are plotted in the form, log $(\sigma_b/\nu_e)/(T_0/T)$ versus log $t_b/a_T a_x$. We normalized σ_b to unit ν_e on the basis of Eq. (1). The full curve represents the average response, while the dashed curves represents the approximate band of scatter for the data of Smith. The dotted curve corresponds to the data of Healy while the dot-dash curve corresponds to Baranwal's data. These three sets of data were obtained at constant ν_e but varying test rate and temperature. The open triangles represent the data of Taylor and Darin, the open circles the data of Epstein and Smith, while the filled circles represent the data of Fedors and Landel. These three sets represent data obtained at constant temperature and rate but varying ν_e. It can be seen that most of the data fall neatly within the band of scatter characteristic of Smith's data.

Scatter in data such as these is due primarily to statistical variability of rupture and, to a smaller extent, to variability in ν_e from specimen to specimen. When ν_e is kept constant, unique values of σ_b and ϵ_b are not obtained when replicate specimens are tested. Variability of both σ_b and ϵ_b can be represented by a distribution of the double exponential type and the parameters which characterize the distribution have been found to vary with

[2]Fedors, R. F., and Landel, R. F., unpublished results.

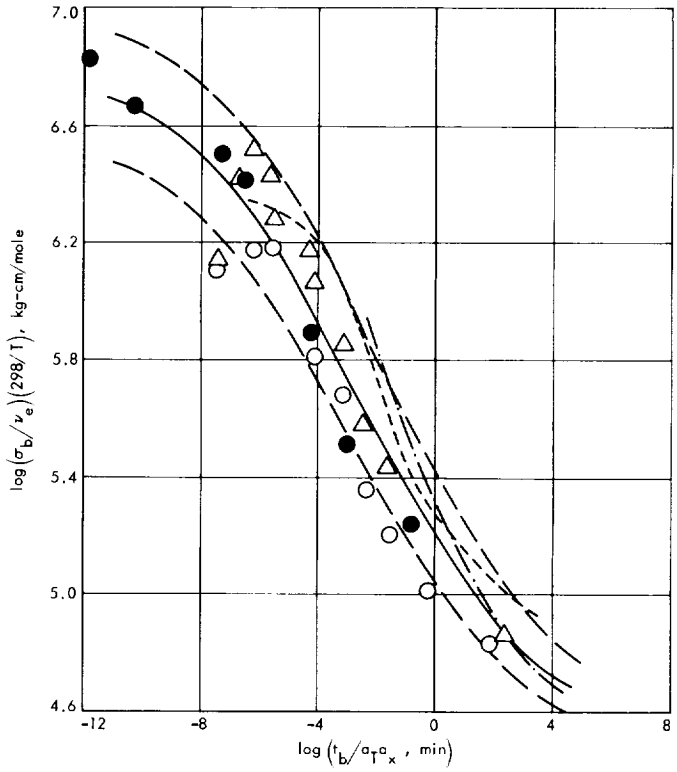


Fig. 1. Dependence of reduced stress-at-break on the reduced-time-to-break for SBR

both ν_e and with test temperature such that the distribution broadens as either ν_e on test temperature is decreased (Refs. 2 and 4).

Hence, we should expect the scatter is greater at short times (low ν_e, low temperature) then at longer times as is observed. In addition, the shape of the "average" curve obtained with a given set of data depends in part on the number of specimens tested, since this determines the degree of scatter observed which, in turn, affects the way in which the average curve is drawn in. Hence, the slight divergence in curve shape for the various sets of data is not considered significant.

Figure 2 shows the log ϵ_b, log $(t_b/a_T a_x)$ response employing the same symbols and the same exponent 7.7 used in Fig. 1. At long times, the data superpose to within the band of scatter characteristic of Smith's data. The differences at short times between the data of Fedors on the one hand and of Taylor and Epstein on the other may be due, in part, to the fact that the latter two investigations employed dumbbell-shaped specimens rather than rings. Epstein reported that ring specimens provided lower values for both σ_b and ϵ_b than did dumbbell speci-

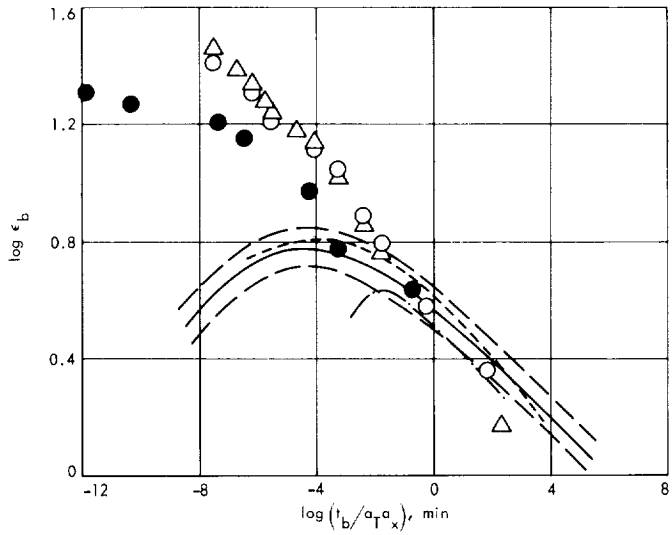


Fig. 2. Dependence of reduced strain-at-break on the reduced-time-to-break for SBK

mens, especially at low v_e values. The data obtained with dumbbell specimens are in close agreement but fall above the data obtained on rings. However, as v_e is increased, and hence as the effective time scale is increased, the difference due to the specimen shape decreases and the data converge.

The data of Smith obtained at constant v_e, first increase, pass through a maximum, then finally decrease with decrease in time scale to yield a bell-shaped curve. This is also apparent in Baranwal's data, but the time scale at which the maximum occurs is higher than is the case for Smith's data; in addition the maximum value of ϵ_b, $\epsilon_{b,\max}$ is lower for Baranwal's vulcanizate. These effects are due to finite chain extensibility. It has been shown that $\epsilon_{b,\max} = 1 + k/v_e^{1/2}$ where k is a constant that depends on the chemical nature of the network chain and on the detailed structure of network (Ref. 2 and Footnote 3). Assuming k to be constant for the SBR vulcanizates considered here, it is evident that as v_e increases $\epsilon_{b,\max}$ must decrease and, according to Eq. (2), the time scale at which the maximum occurs must also increase. The v_e value of Baranwal's vulcanizate is about twice the value for Smith's sample. It thus appears that there exists a limiting envelope for the $\log \epsilon_b$, $\log (t_b/a_T a_x)$ response, which represents the maximum ϵ_b that any system can attain. As a result of finite chain extensibility, however, individual curves representing the response of a system at constant v_e will peel off the envelope at ϵ_b values which decrease, and at time scales which increase as v_e increases.

It is interesting to note that Figs. 1 and 2 constitute failure criteria which can be used to obtain rough estimates for both σ_b and ϵ_b. For example, if v_e, test rate, and test temperature are known, these can be used to estimate a_x (Eq. 2 with exponent equal to 7.7) and a_T (Eq. 3). Further, since t_b is taken equal to ϵ_b/R where R is the test rate, all quantities are known except ϵ_b. But since $\log \epsilon_b$ on the ordinate must equal $\log \epsilon_b$ on the abscissa, one traverses the curve in Fig. 2 until equal values of $\log \epsilon_b$ are obtained from both axes. Once ϵ_b is obtained, the factor $\log (t_b/a_T a_x)$ is also known, and the appropriate value of $\log (\sigma_b T_0/v_e T)$ is obtained from Fig. 1, from which the value of σ_b is obtained. In the region where finite extensibility effects come into play, this procedure will yield maximum estimates for ϵ_b and hence minimum estimates for σ_b.

c. Implications. According to Fig. 1, the reduced stress-at-break can be taken as a universal function of the reduced time-to-break, at least to an approximation representative of scatter in typical data of this type, and hence we can write

$$\frac{\sigma_b}{v_e}\frac{T_0}{T} = k\left(\frac{t_b}{a_T a_x}\right) \tag{4}$$

Likewise, Fig. 2 indicates that ϵ_b for strains not too close to $\epsilon_{b,\max}$ can also be taken as approximately a universal function of the reduced time-to-break, and hence

$$\epsilon_b = h\left(\frac{t_b}{a_T a_x}\right) \qquad (\epsilon_b < \epsilon_{b,\max};\ \text{long time region}) \tag{5}$$

It is known that the stress experienced by an elastomer at finite strains can be expressed as the product of two factors, one of which depends on time alone and the other of which depends only on the strain. Thus, we may write

$$\sigma = G(t)\, f(\epsilon) \qquad (\epsilon < \epsilon_{b,\max};\ \text{long time region}) \tag{6}$$

or

$$\frac{\sigma T_0}{v_e T} = g(t)\, f(\epsilon) \tag{7}$$

[3]Landel, R. F., and Fedors, R. F., in *Rubber Chemistry and Technology* (in press).

where $g(t) \equiv (G(t)\, T_0/\nu_e T)$. As $t \to \infty$, $G(t) \to G(\infty)$, the equilibrium value of the modulus. According to kinetic theory, $G(\infty) = \nu_e RT$, and hence $g(\infty) = RT_0$.

Since the rupture point is the terminus of the stress strain response, Eq. (7) is also applicable to failure. Hence

$$\frac{\sigma_b T_0}{\nu_e T} = g(t_b)\, f(\epsilon_b) \qquad (\epsilon_b < \epsilon_{b,\max};\ \text{long time region}) \tag{8}$$

Substituting Eq. (4) and (5) for the appropriate quantities in Eq. (8) yields

$$k\left(\frac{t_b}{a_T a_x}\right) = g(t_b)\, f\left[h\left(\frac{t_b}{a_T a_x}\right)\right] \tag{9}$$

To the extent that the strain function f can be assumed to be a universal function, then Eq. (9) implies that g will be a universal function as well. Thus we can finally write

$$\frac{\sigma_b T_0}{\nu_e T} = g\left(\frac{t_b}{a_T a_x}\right) f(\epsilon_b) \qquad (\epsilon_b < \epsilon_{b,\max};\ \text{long time region}) \tag{10}$$

Equation (10) indicates that a plot of $\log(\sigma_b T_0/\nu_e T)$ versus $\log \epsilon_b$, i.e., the reduced failure envelope, will also be a universal function. It has already been shown that this is indeed the case (Ref. 1). In addition, the fact that the reduced failure envelope is independent of the chemical structure of an elastomer, again to within an approximation represented by the scatter in such data, indicates that g and f are not very sensitive to chemical structure.

For the case where ϵ_b is in the long time region but close in magnitude to $\epsilon_{b,\max}$ the function f becomes dependent on parameters other than ϵ_b. The most important parameter seems to be n, the number of statistical units per network chain (Footnote 3). For this more general case, Eq. (8) can be written

$$\frac{\sigma_b T_0}{\nu_e T} = g(t_b)\, f_1(\epsilon_b, n) \qquad (\text{long time region}) \tag{11}$$

However, Fig. 1 indicates that Eq. (4) is still valid, and hence

$$k\left(\frac{t_b}{a_T a_x}\right) = g(t_b)\, f_1(\epsilon_b, n) \qquad (\text{long time region}) \tag{12}$$

A possible explicit expression for f_1 is provided by the inverse Langevin approximation, which has the form (Ref. 11)

$$f_1 = f_1(\epsilon, n) = n^{1/2}\left[L^{-1}\left(\frac{\lambda}{n}\right) - \frac{1}{\lambda^{3/2}}\, L^{-1}\left(\frac{1}{\lambda^{1/2} n^{1/2}}\right)\right] \tag{13}$$

where $\lambda = 1 + \epsilon$ and L^{-1} is the inverse Langevin function. Equation (13) has the property that $f_1 \to f$ as $\epsilon \to 0$ or as $n \to \infty$. This equation has been shown to provide a good fit to experimental failure data on a wide variety of elastomer types for ϵ_b values up to $\epsilon_{b,\max}$. The value of n required for fit were in qualitative agreement with n estimated independently from other kinds of experiments (Ref. 1 and SPS 37-45, Vol. IV, p. 99).

Hence, for values near $\epsilon_{b,\max}$, ϵ_b is no longer a universal function of the reduced time-to-break, since f_1 depends on n as well. In this region, instead of a universal curve plot of ϵ_b against reduced time-to-break should yield a family of curves each one of which is related to the value of the parameter n. However, by plotting $f_1(\epsilon_b, n)$ instead of ϵ_b a universal curve should be obtained, i.e.,

$$f_1(\epsilon_b, n) = r\left(\frac{t_b}{a_T a_x}\right)$$

Assuming this to be the case Eq. (12) becomes

$$k\left(\frac{t_b}{a_T a_x}\right) = g(t_b)\, r\left(\frac{t_b}{a_T a_x}\right) \tag{14}$$

or $g(t_b)$ must be universal. Hence

$$\frac{\sigma_b T_0}{\nu_e T} = g\left(\frac{t_b}{a_T a_x}\right) f_1(\epsilon_b, n) \tag{15}$$

Now, a plot of $\log(\sigma_b T_0)/(\nu_e T)$ versus $\log \epsilon_b$, i.e., the reduced failure envelope, will consist of a family of curves each one of which is characterized by its n value. This family will converge to a single master curve for those conditions for which $f_1 \to f$, i.e., when ϵ_b is small and or n is large.

References

1. Landel, R. F., and Fedors, R. F., "Fracture of Amorphous Polymers," *Rubber Chemistry and Technology*, Vol. 40, p. 1049, 1967.

2. Landel, R. F., and Fedors, R. F., *Proceedings of the Fourth International Congress on Rheology*, Brown University, Providence, R.I., August 26, 1963. Part 2, p. 543, E. H. Lee, ed. Interscience Publishers, New York, 1965.
3. Plazek, D. J., *J. Polymer Sci.*, Vol. 4, p. 745, 1966.
4. Smith, T. L., *J. Polymer Sci.*, Vol 32, p. 99, 1958.
5. Baranwal, K. C., *Mechanism of Reinforcement of SBR by a Spherical Polystyrene Filler*. Ph.D. Thesis, University of Akron, 1965.
6. Healy, J. D., *Mechanism of Reinforcement of Amorphous Elastomers by Particulate Fillers*. Ph.D. Thesis, University of Akron, 1967.
7. Taylor, G. R., and Darin, S. R., *J. Polymer Sci.*, Vol 17, p. 511, 1955.
8. Epstein, L. M., and Smith, R. P., *Trans. Soc. Rheology*, Vol. 2, p. 219, 1958.
9. Fedors, R. F., and Landel, R. F., *Trans. Soc. Rheology*, Vol. 9, p. 195, 1965.
10. Smith, T. L., Technical Documentary Report No. ASD-7DR 63-430, May 1963, Wright-Patterson Air Force Base, Ohio.
11. Treloar, L. R. G., "The Physics of Rubber Elasticity," Oxford University Press, 1958.

B. An Empirical Method of Estimating the Void Fraction in Mixtures of Uniform Particles of Different Size, *R. F. Fedors and R. F. Landel*

1. Introduction

The ordering and packing of particles plays a major role in a very large number of diverse technological fields (Ref. 1). For example, the conditions which lead to the attainment of the most dense packing of particles is of intense interest in the fields of ceramics, concrete, solid propellants, asphalts, and powder metallurgy as well as in disciplines such as soil mechanics and geology. In addition, random packing of uniform spheres has been used as a model for the simple liquid state (Ref. 2) and in cytology as a model for biological cell shape (Ref. 3). In spite of this wide applicability of packing, relatively few studies have been carried out and reported in the literature concerning the characteristics of packed particle beds.

In his classic 1931 paper, Furnas considered the important problem of estimating the upper bound to the maximum density of a packed bed formed by blending uniform particles of different size (Ref. 4). Although Furnas' equation is widely quoted (Ref. 1), it should be pointed out that Westman and Hugill arrived at an equivalent result, but in a different form, a year earlier (Ref. 5).

In the approach used by these investigators, one begins by considering the void volume associated with a unit volume of packed bed formed when the particles of largest size are permitted to pack. If particles of sufficiently small size are now added, it is evident that these could be placed in the original voids associated with the large particles. The net effect will be an increase in the number of particles present without any overall increase in volume of the packed bed. In principle this process can be repeated by adding a third, fourth, etc., component of sufficiently small size that the particle will fit into the void space associated with the next larger particle. Thus, the blending of particles of appropriate size will, in general, lead to an increase in the packing density of the bed. Furthermore, it is also evident that at some ratio of large to small particle size as well as at some fraction of the large to small component, the packing density will reach a maximum.

For many applications, it is desirable to have a method of estimating, to a first approximation at least, both the volume fraction of each of the components at which the maximum density is achieved as well as the volume fraction of particles in the packed bed at maximum density.

In this report, it is pointed out that the equation developed for calculating the packing fraction (Refs. 4 and 5), can provide useful estimates of the *mix ratio* at which maximum packing occurs. Further, an empirical equation, free of adjustable parameters, is proposed which seems capable of providing estimates of the packing fraction at maximum density as a function of the size ratio (ratio of the size of the small particle to that of the large particle).

2. Discussion

a. Composition of packed bed at maximum density. Since the result is of interest and is apparently not available in the literature, the general expression for the volume fraction of particles in a packed bed of maximum density produced by blending n components, each of uniform size, but not all of which, however, necessarily have the same void volume will be derived.

The n components are arranged in order of decreasing particle size such that the component of largest size carries the subscript 1 and the component of smallest size carries the subscript n. Consider now a packed bed which is made up of component 1 only. The void volume in this packed bed can be expressed as

$$v_1 = V_T(1 - \phi_1) \tag{1}$$

where v_1 is the void volume, V_T is the total volume, and ϕ_1 is the volume fraction of particles in the bed. If the size of component 2 is sufficiently small, these particles can be inserted into the existing void spaces associated with component 1 without requiring any increase in overall volume. A rough estimate of this size will now be obtained.

For large particles, the void fraction for a single component is about 0.37 for a *randomly* packed bed. This void fraction is equivalent to that for an *ordered* packed bed containing equal fractions of cubic packing and of hexagonal close-packing. If this mixture of ordered packing can be taken, to a first approximation at least, as a model for the randomly packed bed, then it is easy to show that the maximum size ratio of sphere then can be accommodated in the interstitial void of that fraction of the bed in cubic packing is about 0.41 while the size ratio for the corresponding hexagonal close pack is 0.15. The average value of the size ratio is thus about 0.3. On the basis of this estimate, which is probably too high, the size of component 2 should be about four times smaller than that of component 1.

If the volume fraction of particles in a packed bed of component 2 is ϕ_2, the maximum volume of component 2 that can be accommodated in the void volume of component 1 is

$$v_2 = v_1\phi_2 = V_T(1-\phi_1)\phi_2 \tag{2}$$

Using the same arguments, the volume fraction of the nth component can be written as

$$v_n = v_{n-1}\phi_n = V_T(1-\phi_1)(1-\phi_2)\cdots(1-\phi_{n-1})\phi_n \tag{3}$$

Hence, the volume fraction of particles at maximum density, ϕ^*_{max}, in a packed bed prepared by blending n components of appropriate particle size is

$$\phi^*_{max} = \phi_1 + (1-\phi_1)\phi_2 + (1-\phi_1)(1-\phi_2)\phi_3 + \cdots + (1-\phi_1)(1-\phi_2)\cdots(1-\phi_{n-1})\phi_n \tag{4}$$

When $\phi_1 = \phi_2 = \cdots = \phi_{n_1}$ this expression is equivalent to the equation or Furnas (Ref. 4), and to the expression proposed earlier by Westman and Hugill (Ref. 5).

According to Eq. (4) then, only a knowledge of the individual ϕ_i values are necessary to estimate ϕ^*_{max} when the particle sizes of the individual component are sufficiently graded, since there is no explicit dependence of ϕ^*_{max} on particle size. There is, however, an implicit dependence on size since ϕ_i itself is a function of particle size (Ref. 6 and SPS 37-48, Vol. III, pp. 109–116). Since each factor on the right-hand side of Eq. (4) is less than unity, it is evident that the relative contribution of each term, representing contributions from components of successively smaller size, decreases very rapidly, and the major contributions to ϕ^*_{max} are provided by the first few terms.

The expected composition of the mixture at which the maximum density occurs can be calculated from Eq. (4). If each term is divided by ϕ^*_{max}, there is obtained

$$\frac{\phi_1}{\phi^*_{max}} + \frac{(1-\phi_1)\phi_2}{\phi^*_{max}} + \frac{(1-\phi_1)(1-\phi_2)\phi_3}{\phi^*_{max}} + \cdots + \frac{(1-\phi_1)(1-\phi_2)\cdots(1-\phi_{n-1})\phi_n}{\phi^*_{max}} = 1 \tag{5}$$

The first term represents the volume fraction of component 1 at maximum density, $\bar{X}^*_1$, the second the volume fraction of component 2, $\bar{X}^*_2$, etc. Thus, the composition of the packed bed at maximum density can be estimated using the following equations:

$$\begin{aligned}
\bar{X}^*_1 &= \frac{\phi_1}{\phi^*_{max}} \\
\bar{X}^*_2 &= \frac{(1-\phi_1)(\phi_2)}{\phi^*_{max}} \\
&\vdots \\
\bar{X}^*_n &= \frac{(1-\phi_1)(1-\phi_2)\cdots(1-\phi_{n-1})\phi_n}{\phi^*_{max}}
\end{aligned} \tag{6}$$

Here also, only a knowledge of the individual ϕ_i values are required to calculate the composition of maximum density.

If it is assumed that $\phi_1 = \phi_2 = \cdots = \phi_n = 0.63$, then the ϕ^*_{max} values as well as the composition $\bar{X}^*_1, \cdots, \bar{X}^*_n$ predicted using Eqs. (4) and (6) are listed in Table 1. It is apparent from these calculations that the increase in packing obtained by the addition of other components rapidly diminishes as the number of components increases beyond 3.

Data on a binary system having a size ratio essentially equal to zero has been published by Furnas (Ref. 7). For this system, $\phi_1 = \phi_2 = 0.50$, and the measured parameters at maximum density are: $\phi^*_{max} = 0.76$, $\bar{X}^*_1 = 0.67$ and

Table 1. Calculation of packed bed characteristics

Number of components	ϕ^*_{max}	$\bar{X}^*_1$	$\bar{X}^*_2$	$\bar{X}^*_3$	$\bar{X}^*_4$	$\bar{X}^*_5$
1	0.63	1.0	—	—	—	—
2	0.86	0.73	0.27	—	—	—
3	0.95	0.67	0.25	0.09	—	—
4	0.98	0.64	0.24	0.09	0.04	—
5	0.99	0.64	0.23	0.08	0.03	0.01

$\bar{X}^*_2 = 0.33$. The calculated values are: $\phi^*_{max} = 0.75$, $\bar{X}^*_1 = 0.67$ and $\bar{X}^*_2 = 0.33$. Westman and Hugill (Ref. 5) working with sized sand with a size ratio $R = 0.02$, and characterized by $\phi_1 = 0.63$ and $\phi_2 = 0.58$, obtained the following values from direct experimental measurements: $\phi^*_{max} = 0.82$, $\bar{X}^*_1 = 0.70$ and $\bar{X}^*_2 = 0.30$. The predicted values are $\phi^*_{max} = 0.84$, $\bar{X}^*_1 = 0.74$, and $\bar{X}^*_2 = 0.26$. McGeary (Ref. 8) working this steel shot having an R ratio equal to 0.058 and $\phi_1 = \phi_2 = 0.625$ finds $\phi^*_{max} = 0.84$, $\bar{X}^*_1 = 0.70$, and $\bar{X}^*_2 = 0.30$. The predicted values are: $\phi^*_{max} = 0.81$, $\bar{X}^*_1 = 0.72$ and $\bar{X}^*_2 = 0.28$.

b. Effect of particle size on maximum density. If two or more components each having the same particle size are mixed, it is clear that no volume change would be expected to occur. Hence, a "lower bound" of zero volume change occurs when the size ratio $R = 1$. On the other hand, when the particle sizes are ordered for each component, $r_1 > r_2 > r_3 \cdots > r_n$ and sufficiently small so that Eq. (4) applies, an upper bound to the volume change will occur at ϕ^*_{max}. It is thus clear that ϕ^*_{max} must vary inversely with R.

As an example of the magnitude of the effect of the size ratio R on ϕ^*_{max}, data of Westman and Hugill on lead shot are listed in Table 2 (Ref. 5). Each binary mixture contains 30 vol. % of the fine component. It is evident from these data that ϕ^*_{max} is rather strongly dependent on R at least for this system.

Table 2. Effect of size ratio on maximum packing of lead shot

Size ratio, R	$\phi^*_{max}(R)$
1.0	0.63
0.796	0.63
0.685	0.64
0.584	0.65
0.474	0.65
0.369	0.68
0.242	0.71
0.147	0.76

There are at least two effects which can give rise to a dependence of ϕ^*_{max} on R. In the first instance, if the particle size of the second component is too large, it will not fit into an existing interstitial void volume without requiring an increase in overall volume. An increase in volume would imply a lower value of ϕ^*_{max}. This effect would be expected to predominate at large R values. For the crude estimate made previously, the R value would be roughly greater than 0.3. For these R values, neither Eq. (4) nor Eq. (6) would be expected to be valid, since the underlying assumptions implicit in these equations are no longer true.

In the second instance, even if the size of component 2 is sufficiently small to be accommodated, not all of the available void space would be usable because of what may be termed the "wall effect." It is well known that the measured ϕ value depends strongly on the container size because the space near the container wall cannot be used as efficiently as the space within the container for packing of particles. It is found that the measured ϕ value increases as container size increases, i.e., as the surface area to volume ratio of the container decreases to zero.

This effect should predominate for R roughly less than 0.3. However, here Eq. (6) is still expected to be valid, since the equation merely defines the composition at which the volume available for occupancy by component 2 is a maximum. The value of ϕ^*_{max} itself will be less than that predicted on the basis of Eq. (4) due to the "wall effect." It has been found that an empirical equation similar in form to Eq. (4) can be used to obtain first order approximation values for the dependence of ϕ^*_{max} on R. The equation is

$$\begin{aligned}\phi^*_{max}(R) = {} & \phi_1 + (1 - a_{1,2})(1 - \phi_1)\phi_2 \\ & + [(1 - a_{1,3})(1 - a_{2,3})]^{1/2} \\ & \times (1 - \phi_1)(1 - \phi_2)\phi_3 + \cdots \\ & + [(1 - a_{1,n})(1 - a_{2,n}) \cdots (1 - a_{n-1,n})]^{1/n-1} \\ & \times (1 - \phi_1)(1 - \phi_2) \cdots (1 - \phi_{n-1})\phi_n \end{aligned} \tag{7}$$

where

$$a_{i-1,i} = \left(\frac{r_i}{r_{i-1}}\right)^{1/2}$$

and r is some measure of the particle size, such as the radius. Since $a_{i-1,i} = 1$ when $r_i/r_{i-1} = 1$, Eq. (7) reduced to $\phi^*_{max}(R) = \phi_1$ when the components all have the same size. When $r_i/r_{i-1} = 0$, $a_{i-1,i} = 0$ and Eq. (4) is obtained.

Furnas has derived an equation for the effect of the size ratio R on ϕ^*_{max} for a system where

$$\phi_1 = \phi_2 = \cdots = \phi_n = \phi_1 ,$$

which is given by

$$\frac{(1-\phi)^n \ln \phi (1-\phi)}{(1-\phi^{n+1})[1-(1-\phi^n]} = \frac{(2.62\,K^{1/n} - 3.24\,K^{2/n}) \ln K}{(1.0 - 2.62\,K^{1/n} + 1.62\,K^{2/n})\,n^2} \tag{8}$$

where n is one less than the total number of components for a system of maximum density, and K is the ratio of the smallest to the largest particle size. The numerical values appearing in this equation were obtained by means of a curve fit to experimental data (Ref. 7). For comparison with the prediction of Furnas' Eq. (8), Eq. (7) is shown in Fig. 3 as the dotted curve calculated for a two-component system with $\phi_1 = \phi_2 = 0.60$. For $R < 0.5$, the difference between the two equations is small.

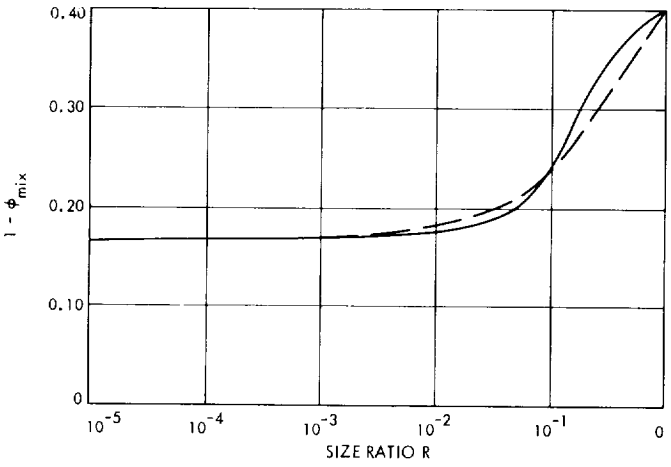


Fig. 3. Dependence of minimum void volume on size ratio R. The full curve is Eq. (8) and the dotted curve corresponds to Eq. (7)

Very limited data are available in the literature on the void content of packed beds as a function of both size ratio and composition. Such data as are available are shown in Figs. 4 to 7 plotted as void volume of the mixture $(1 - \phi_{mix})$, where ϕ_{mix} is the volume fraction of solids in the packed bed, as a function of composition. Figure 4 shows data of Furnas (Ref. 7) as the full curves, Fig. 5 the data of Westman and Hugill (Ref. 5) on sized sand as the filled points, Fig. 6 the data of McGeary (Ref. 8) on mixtures of steel shot and tungsten powder as the filled points, and finally Fig. 7 the data (also of McGeary) on steel shot. These data show very clearly the occurrence of a minimum in the void volume (or a maximum in packing density). The value of the maximum density as calculated from Eq. (7) as well as the position of the maximum density calculated from Eq. (4) is represented in each figure as the point of intersection of the two dotted lines. It is evident that both the magnitude and position of the calculated $\phi^*_{max}(R)$ values is in good agreement with experimental values except for large R values, i.e., R values greater than about 0.3.

The equations for the dotted lines in the figures are given by

$$\phi_{mix} = [(\phi_1 - \phi_2) + (1-a)(1-\phi_1)\phi_2] \times [\phi_1 + (1-\phi_1)\phi_2]\frac{\bar{X}_1}{\phi_1} + \phi_2 \tag{9}$$

when

$$\bar{X}_1 \leq \frac{\phi_1}{\phi_1 + (1-\phi_1)\phi_2}$$

and

$$\phi_{mix} = (1-a)[\phi_1 + (1-\phi_1)\phi_2]\bar{X}_2 + \phi_1 \tag{10}$$

when

$$\bar{X}_1 \geq \frac{(1-\phi_1)\phi_2}{\phi_1 + (1-\phi_1)\phi_2}$$

As is evident in both figures, a linear dependence of ϕ_{mix} on composition actually provides a good first approximation representative of the experimental data. For these data at least, Eqs. (9) and (10) provide values of ϕ_{mix} which differ from the experimentally measured values by only about 5% or less. This was also found to be true for unpublished data of Cokelet for two component blends of glass beads.[4]

To the extent that Eqs. (9) and (10) provide reasonable estimates of ϕ_{mix}, it should be possible to estimate the viscosity of slurries containing blends rather than a single component of uniform size as a function of both the size ratio and the composition. Experimental studies have demonstrated that the viscosity of a slurry is related to the composition of the slurry by means of (Ref. 9):

$$\frac{\eta}{\eta_0} = \left(\frac{\phi_{mix}}{\phi_{mix} - \phi}\right)^{2.5} \tag{11}$$

[4]Cokelet, G. R., California Institute of Technology (unpublished results).

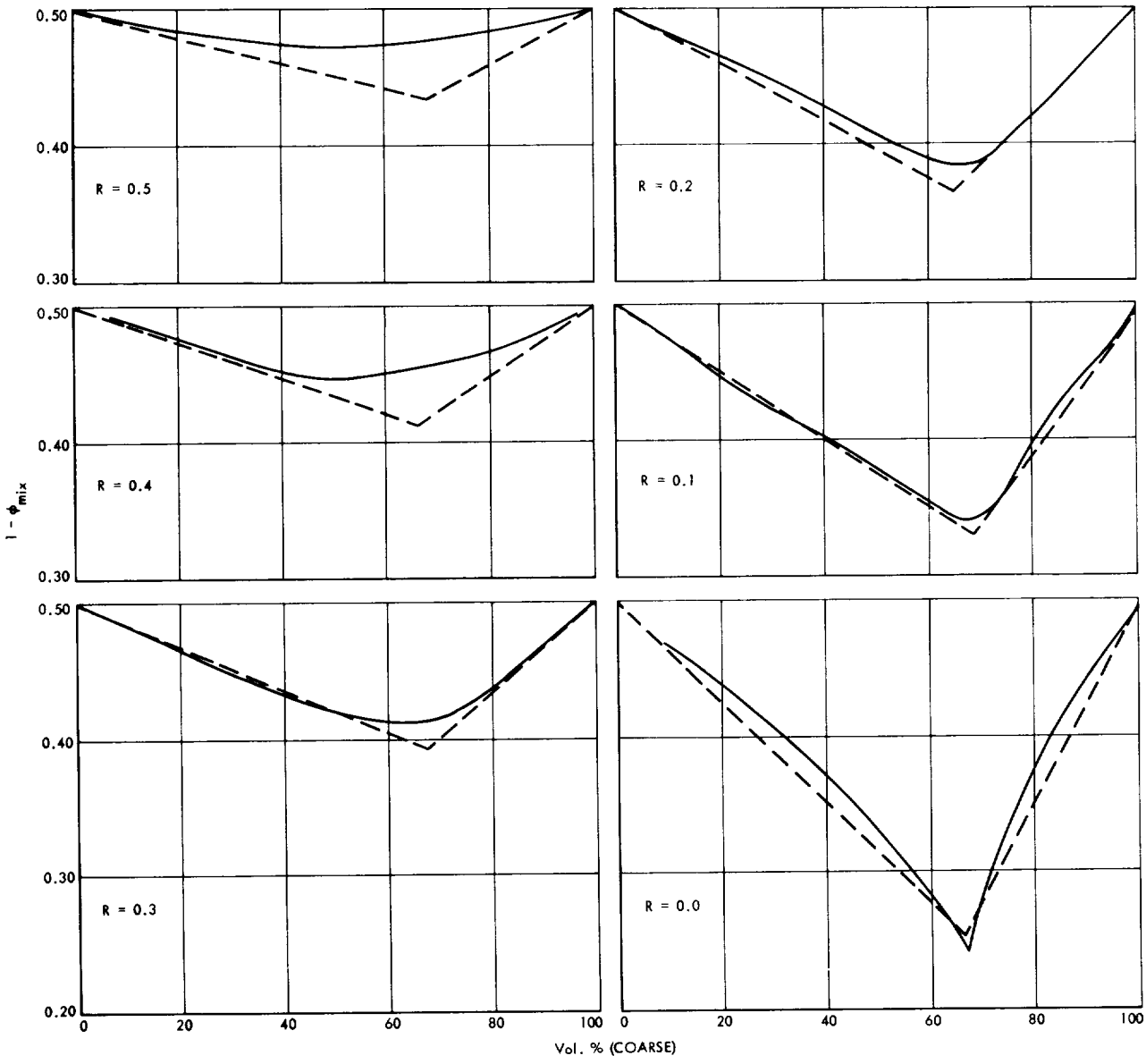


Fig. 4. Dependence of minimum void volume on size ratio. Data of Furnas (Ref. 7)

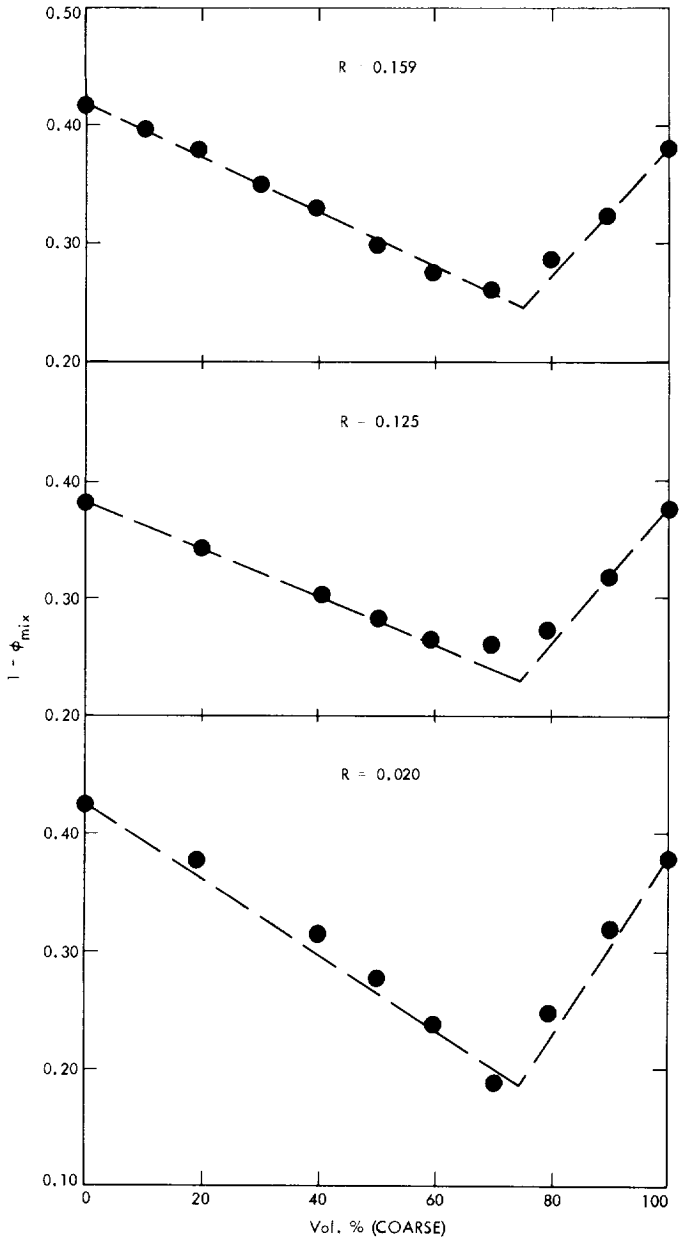


Fig. 5. Dependence of minimum void volume on size ratio. Data of Westman and Hugill (Ref. 5)

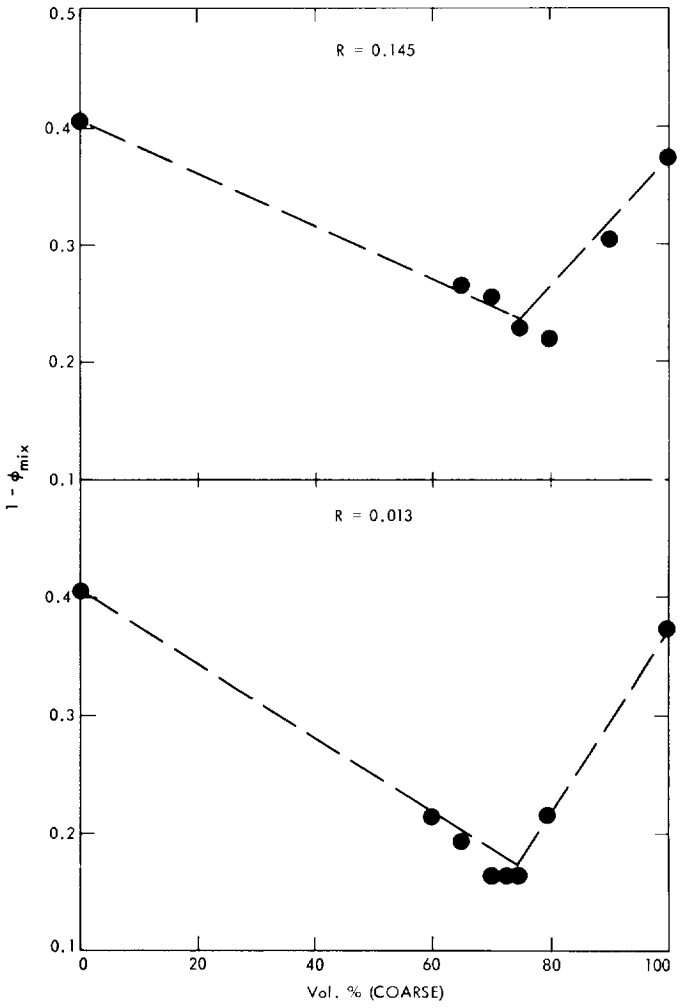


Fig. 6. Dependence of minimum void volume on size ratio. Data of McGeary (Ref. 8) (steel shot and tungsten powder)

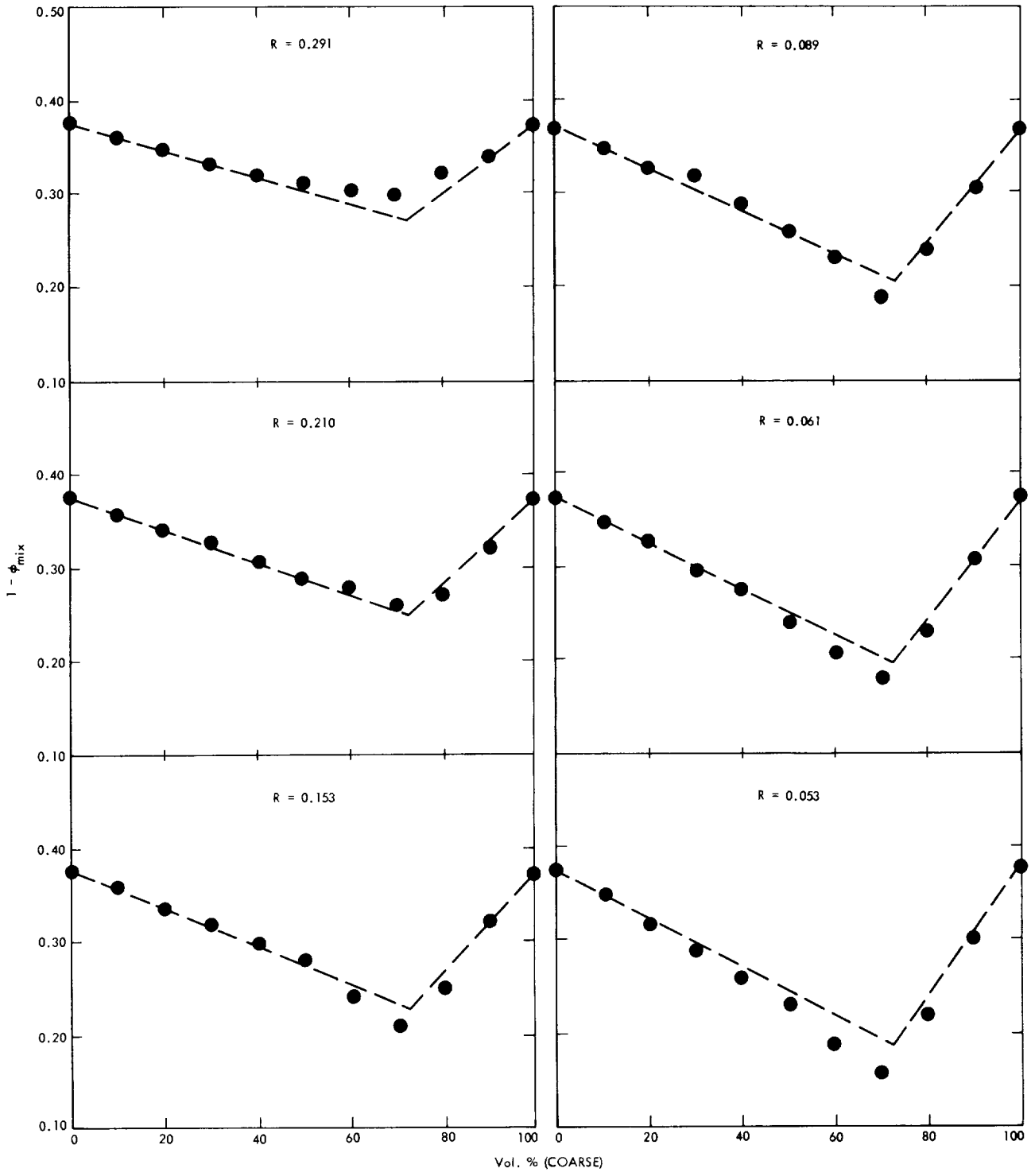


Fig. 7. Dependence of minimum void volume on size ratio. Data of McGeary (Ref. 8) (steel shot)

where η and η_0 are the viscosities of the filled and unfilled liquid, respectively, ϕ_{mix} is the volume fraction (maximum) of particles in a packed bed, and ϕ is the volume fraction of filler present in the slurry. Note that ϕ is always less than ϕ_{mix}. If Eqs. (9) and (10) can be used to estimate ϕ_{mix} as a function of both composition and size ratio, then Eq. (11) can be used to estimate the viscosity. Note also that equations for estimating the ϕ_i values required in Eqs. (9) and (10) have been proposed. Since the viscosity in Eq. (11) can be replaced by the small strain modulus in filled crosslinked elastomers, the equation also provides a way to estimate the effect of composition and size ratio of filler on the small strain modulus of the composite.

References

1. Dallavalle, J. M., "Micrometrics, the Technology of Fine Particles," Pitman Publishing Corp., New York, 1948.
2. Bernal, J. D., and Mason, J., *Nature*, Vol. 188, p. 910, 1960.
3. Marvin, J. W., *Am. J. Botany*, Vol. 26, p. 280, 1939.
4. Furnas, C. C., *Ind. Eng. Chem.*, Vol. 23, p. 1052, 1931.
5. Westman, A. E. R., and Hugill, H. R., *J. Am. Ceramic Soc.*, Vol. 13, p. 767, 1930.
6. Moser, B. G., and Landel, R. F., "Rheology of Slurries III: A Theory of the Sedimentation Volume for Systems Containing Uniform Spheres," paper presented at the Pacific Conference on Chemistry and Spectroscopy, Anaheim, Calif., Nov. 1967.
7. Furnas, C. C., *Bureau Mines Report Investigations*, Vol. 2894, p. 7, 1928; *Bureau Mines Bull.*, Vol. 307, p. 74, 1929.
8. McGeary, R. K., *J. Am. Ceramic Soc.*, Vol. 44, p. 513, 1961.
9. Landel, R. F., Moser, B. G., and Bauman, A. J., *Proceeding of the Fourth International Congress on Rheology*, Brown University, Providence, R.I., Ed. E. H. Lee, Vol. II, p. 663, Interscience Publishers, New York, 1965.

C. Viscosity of Hardened Red Blood Cells,

R. F. Landel and R. F. Fedors

1. Introduction

As part of a program of research on the mechanical properties of composite solid propellants, a series of investigation have been carried out on the viscosity of slurries (Ref. 1). It was found that the viscosity of a slurry η was related in a simple manner to the viscosity of the fluid η_0, to the volume fraction of filler present ϕ, and to the maximum packing ϕ_m; i.e., to the maximum volume fraction of filler which can be incorporated into the fluid. The explicit dependence is given by

$$\eta = \eta_0 \left(1 - \frac{\phi}{\phi_m}\right)^{-2.5} \qquad (1)$$

This equation reduces to the Brinkman equation (Ref. 2) when ϕ_m is taken to be unity and to the Roscoe equation (Ref. 3) when ϕ_m is taken as 0.74, but differs from them in a very important respect. It was shown (Ref. 1) that ϕ_m is, in general, not constant but rather a variable whose magnitude depends on factors such as particle size, particle size distribution, and surface energy. Nevertheless, Eq. (1) was found to be valid for slurries consisting of a wide variety of solid spheroidal fillers suspended in nonaqueous liquids.

When the filler is a liquid, as in an emulsion, the relative viscosity departs from the prediction of Eq. (1) if the droplets deform. Blood can be considered as a suspension of deformable droplets consisting of red blood cells dispersed in plasma as the medium. In normal blood, ϕ is about 0.45. The extent of deformation of the cells depends, of course, on the shear rate so that blood is strongly non-Newtonian in its flow behavior. Even at relatively low shear rates, the viscosity of blood is less than that of a slurry containing the same volume fraction of nondeformable particles.

2. Discussion

Recently, Chien, et al., evaluated the effect of particle deformability by making a comparison between the viscosity of suspensions of normal deformable cells and cells which had been hardened by treatment with acetaldehyde (Ref. 4). This hardening process does not alter the biconcave shape of the cell. Thus, they were able to work with both deformable and nondeformable particles of the same size and shape. They also measured ϕ_m by a sedimentation technique. Normal cells deform and pack until they occupy 0.95 to 0.97 of the volume of sediment, while the hardened cells occupy only 0.60 to 0.62 of the sediment volume (15,000 g for 5 or 30 min, respectively).

In our earlier work we derived a theory which predicted that ϕ_m should be 0.63 for monodisperse spheres (SPS 37-40, Vol. IV, p. 84), in agreement with experimental observations of Scott (Ref. 5). Scott also observed that ϕ_m decreases as the container size used for measurement is decreased.

Having both ϕ and ϕ_m available, it is of interest to calculate the viscosity of suspensions containing the hardened cells using Eq. (1). The results are shown in Fig. 8 for both canine and human cells. The curves and open circles are the data as measured and reported by Chien; the filled circles represent the calculated values using Eq. (1).

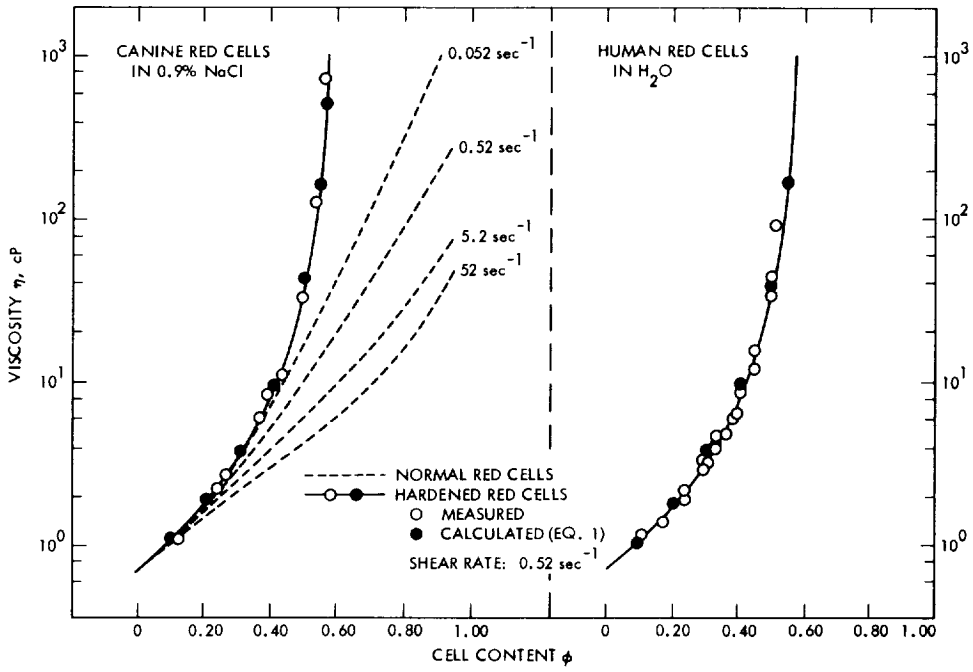


Fig. 8. Dependence of the viscosity of suspensions of normal and hardened red blood cells on the cell content

For these calculations, a value of 0.7 cP for η_0 was estimated from Chien's plot as the viscosity when $\phi = 0$; and ϕ_m was taken as 0.62 rather than 0.60, since ϕ_m represents the maximum volume fraction of filler that can be incorporated into the liquid. The agreement between calculated and experimental results is excellent, especially in view of the fact that no adjustable parameters were used. Thus, the applicability of Eq. (1) has been extended to yet another system.

Moreover, it is important to note that in the low shear rate region, ($\sim$0.052 sec^{-1}, Fig. 8), the response of normal cells in blood is scarcely distinguishable from that of the hardened erythrocytes, which means that little cell deformation is taking place under the conditions of the experiment (Couette flow, at 37°C). This implies that the same must be true *in vivo*, when these shear rates obtain, for flow which is taking place in the larger arteries and veins. Correspondingly, and more importantly, small changes in the RBC content of blood or in the way in which the cells aggregate, both of which can give rise to small changes in ϕ_m, will produce very large changes in the viscosity of blood even at low shear rates.

References

1. Landel, R. F., Moser, B. G., and Bauman, A. J., *Proceedings of the Fourth International Congress on Rheology*, Part 2, p. 663. Edited by E. H. Lee, Interscience Publishers, New York, 1965.
2. Brinkman, H. C., *J. Chem. Phys.*, Vol. 32, p. 571, 1952.
3. Roscoe, R., Brit. *J. Appl. Phys.*, Vol. 3, p. 267, 1952.
4. Chien, S., et al., *Science*, Vol. 157, p. 829, 1967.
5. Scott, G. D., *Nature*, Vol. 188, p. 908, 1960.

D. Isobutylene Prepolymers, *J. A. Miller, Jr.*

1. Introduction

Molecular sieves are crystalline metal alumino-silicates belonging to a class of solids called zeolites. Widely used as drying agents and in the fields of adsorption and reaction catalysis, they are available in several forms (Ref. 1). A brief review of their structure and properties will serve to introduce their new-found importance in the solid propellant prepolymer program.

The general formula of molecular sieve is

$$X \cdot \left[\frac{M^2}{n} \cdot Al_2O_3\right] \cdot Y\,SiO_2 \cdot Z\,H_2O$$

where M is the associated cation and n its valence. Table 3 summarizes the types available.

The sieve crystal is composed of silicate and aluminate tetrahedra, the excess negative charge of each aluminate moiety being neutralized by the associated cation. These

Table 3. Molecular sieves

Type[a]	Unit cell	Pore diameter, Å
3A	K_9Na_3 $[(AlO_2)_{12}(SiO_2)_{12}] \cdot XH_2O$	3
4A	Na_{12} $[(AlO_2)_{12}(SiO_2)_{12}] \cdot XH_2O$	4.2
5A	$Ca_{4.5}Na_3$ $[(AlO_2)_{12}(SiO_2)_{12}] \cdot XH_2O$	5
10X	$Ca_{32}Na_{22}$ $[(AlO_2)_{86}(SiO_2)_{106}] \cdot XH_2O$	8
13X	Na_{86} $[(AlO_2)_{86}(SiO_2)_{106}] \cdot XH_2O$	10

[a]Available from Union Carbide Corporation, Linde Division.

tetrahedra, along with the water of hydration, are arranged to yield a crystal interlaced with large cavities connected by smaller ports. These cavities and apertures, whose sizes are a function of Si/Al ratio and type of cation, account for greater than 50% of the sieve volume. Figure 9 represents the cavity of a 4Å Linde molecular sieve crystal, a truncated octahedron with internal dimensions 11Å in diameter joined by apertures 4Å in diameter.[5] A molecule thus must have an effective diameter approximating 4Å in order to enter the sieve's central cavity. Dehydration does not alter the structure of the sieve and activates the internal cavities for adsorption and reaction catalysis.

Adsorption and separation may be accomplished according to molecular size or polarity, e.g., a 5A sieve will separate *n*-octane from iso-octane and propylene from propane.

The subject of catalysis is generally reserved for a multivalent cationic-type *X* and *Y* molecular sieve. These species, due to the increased distances between cation centers, exhibit carbonium ion and ionic activity (Ref. 2). These type sieves are also desirable since their large internal cavities can accommodate most molecules.

An interesting development has been that of molecular selectivity catalysis (Refs. 3 and 4). In these cases, reactions catalyzed by molecular sieve have been performed on similar type molecules of different diameters. It is the above development which directly applies to the solid propellant prepolymer program. Our main interest was the use of molecular sieve as a drying agent; however, polymerization occurred when used with isobutylene to yield a novel polymer. The polymer, structurally a head-to-tail poly(isobutylene), was terminally difunctional.

[5]Al and Si cations occupy the corners, each surrounded by four oxygen anions. The excess negative charges are neutralized by metal ions. The water of hydration is contained within the cavity.

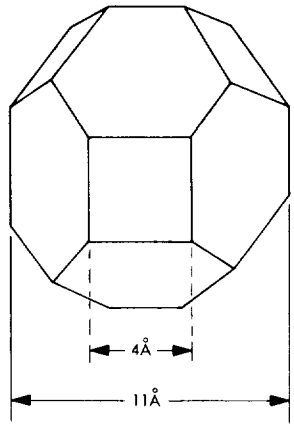


Fig. 9. Cavity of a 4Å Linde molecular sieve crystal

2. Experimental Methods

All operations were performed on the vacuum manifold described previously. A 300-ml flask equipped with a magnetic stirrer and side-arm was sealed onto the vacuum line. The desired type and quantity of molecular sieve was added through the side-arm, which was then sealed. The flask was evacuated and, with periodic heating to 100°C, pumped to a pressure of 10^{-6} mm Hg. The flask was then cooled to −196°C and untreated Phillip's research grade isobutylene was condensed onto the sieve. The isobutylene and sieve were degassed to 10^{-6} mm Hg, quickly warmed to reaction temperature (at which isobutylene was liquid), and finally stirred at autogenous pressure. The flask did not warm to the touch, indicating there was little or no polymerization exotherm. After stirring for the desired period of time, all low boiling fractions were distilled from the mixture at room temperature and collected at −196°C. The low boiling fraction was stored and used for other polymerizations. The reaction flask containing the sieve and reaction residue was removed from the vacuum line. The sieve was washed several times with pentane. All pentane washings were combined, washed with water, and dried over anhydrous magnesium sulfate. The pentane was decanted from the $MgSO_4$ and evaporated at reduced pressure until polymer appeared. The polymer was finally dried under vacuum at 50°C. The final product was a clear, colorless, viscous liquid.

Number average molecular weights and molecular weight distributions were determined on a Waters gel permeation chromatography unit calibrated with a sample of poly(isobutylene) whose molecular weight was determined by viscosity methods. Infrared spectra were recorded on a Perkin–Elmer 421 spectrometer.

Average functionalities were determined on a Varian 100-MHz nuclear magnetic resonance (NMR) spectrometer. Spectra were recorded on CCL_4 solutions of the polymer and separated into vinyl and saturated proton regions. Each section was integrated to give the relative ratio of both types of protons. The vinyl region was divided into its components and each component assigned a structure. Each vinyl integral was then divided by the total number of vinyl protons in its assumed structure to yield its equivalent in unsaturation. The average functionality $\bar{f}$ was then determined by solving the following equations:

$$M_N = 56\bar{X}_n + 55\bar{f} + 57(2 - \bar{f})$$

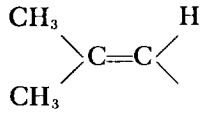
$$R = \frac{\bar{f}}{8\bar{X}_n + 6\bar{f}x + 5\bar{f}y + 7\bar{f}z}$$

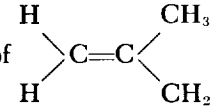
where

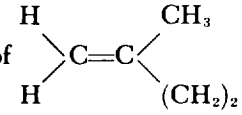
M_N = number average molecular weight

$\bar{X}_n$ = average degree of polymerization

R = ratio of double bonds to number of protons from saturated sources

x = mole fraction of  groupings

y = mole fraction of  groupings

z = mole fraction of  groupings

3. Experimental Results

a. Polymerizations. Table 4 summarizes the polymerizations performed to this time. The following important features should be noted:

(1) The number average functionality of the polymers approximates two. Polymerization of isobutylene by conventional catalysts yields a monofunctional polymer.

(2) Ozonolysis of polymer 382-153-34 resulted in no molecular weight reduction. This is a strong indication that the functional unsaturation is at both ends of the polymer.

(3) The lowest molecular weight obtained was 1150. The bulk viscosity of this polymer, although not determined, seemed to be low enough for successful propellant processing.

(4) The highest molecular weight obtained was 3600. The highest molecular weight poly(isobutylene) previously obtained using molecular sieve was 240 (Refs. 5 and 6).

(5) Stirring has a pronounced effect on the final molecular weight.

(6) Polymer molecular weight and polymerization efficiency are directly proportional to contact time.

(7) The 5A sieve, which excludes isobutylene from its central cavity, is a polymerization catalyst, whereas 13X, which accommodates isobutylene, is not a catalyst.

The problem of preparing a low-bulk viscosity, maximum molecular weight, terminally difunctional poly(isobutylene) for solid propellant applications seems to have

Table 4. Summary of polymerizations

Run	Sieve type	Amount of sieve, g	Amount of isobutylene, ml	Temperature, °C	Time, h	Stirring	M_N	M_W/M_N	f^a	Efficiency, %
382-153-34[b]	5A	30	200	RT[c]	24	No	1150	2.9	1.8	—
382-153-44	5A	25	75	RT	32	Yes	2020	3.5	—	18
382-153-45	5A	25	75	0	96	Yes	2900	2.6	2.1	37
382-153-53	5A	25	75	RT	264	Yes	3600	3.4	2.1	30
382-153-66	4A	25	75	RT	96	Yes		No polymer formed		
382-153-67	13X	25	75	RT	120	Yes		No polymer formed		

[a]Number of double bonds/average molecule.

[b]Ozonolysis resulted in no molecular weight reduction.

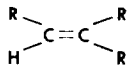
[c]RT = room temperature.

been partially solved. The next step would be to determine the chain extension capabilities of the olefinic endgroups.

b. Structural determination.

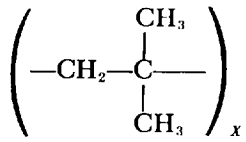
Infrared data. Spectra of all samples exhibited doublets at 1360 and 1385 cm^{-1}. These doublets are generally attributed to gem-dimethyl vibrations and are common to poly(isobutylenes) prepared by conventional cationic techniques (Refs. 7 and 8). All samples also exhibit the important bonds given in Table 5.

Table 5. Bonds exhibited in samples

Frequency, cm^{-1}	Source
3080	C—H stretch in $=CH_2$
1780	Overtone from 890 band
1637	C=C stretch
888	C—H bending in $=CH_2$
821	C—H bending in 

NMR data. Figures 10 and 11 are the important sections of the 100-MHz NMR spectrum of sample 382-153-53. All other samples exhibited similar spectra.

In Fig. 10, the areas of peaks A and B are in the ratio of 6:2. This result, along with the observed gem-dimethyl IR vibration, indicates the polymer backbone is the normal head-to-tail structure:



The fine structure surrounding peaks A and B is due to changes in environment of the above methyl and methylene experience at the polymer ends. Another source is the methyl groupings attached to the unsaturated endgroups.

Figure 11 indicates that three and possibly four types of vinylic protons are present on the polymer. The position of peak C would indicate the presence of a trisubstituted double bond. This corresponds to the 821 cm^{-1} absorption in the IR. Peaks D and E result from at least two sources of unsaturation. Peak E and an equal area

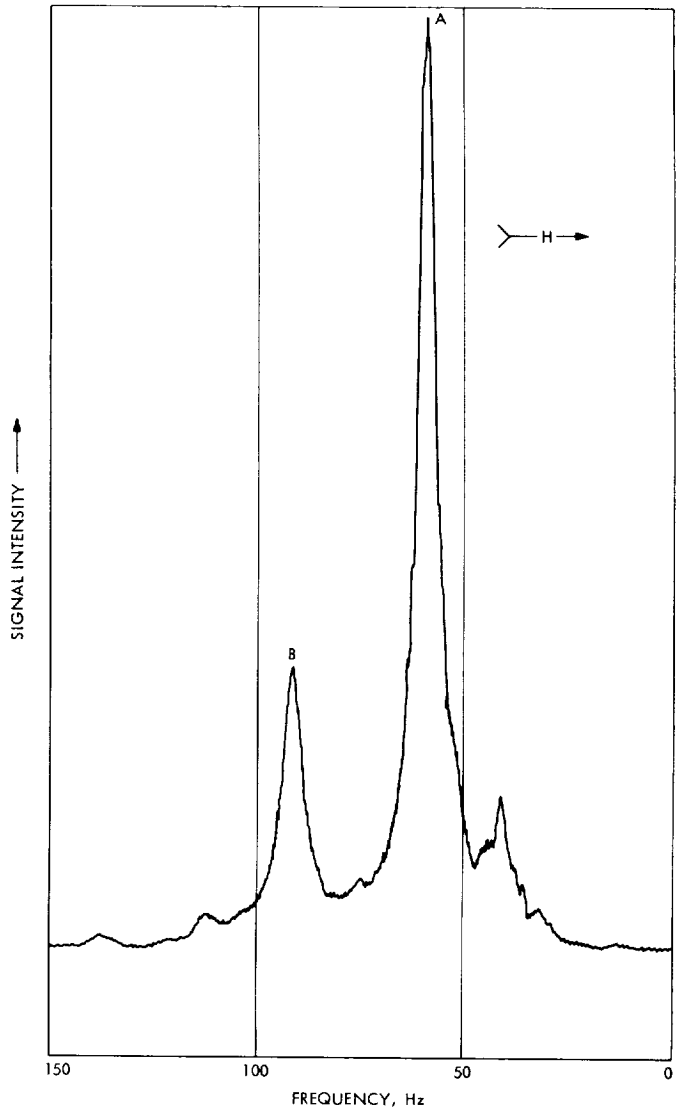


Fig. 10. Saturated hydrocarbon proton region of polymer 382-153-53

of D is a doublet which would result from a disubstituted exomethylene structure. This observation corresponds to the 3080 and 888 cm^{-1} absorptions in the IR.

The source of the hidden portion of peak D is a mystery at this time. A trisubstituted structure in a different environment than the structure responsible for peak A is a possibility. More work on the 220-MHz instrument would clear up some of this confusion.

From the above information, it can be stated with some confidence that the poly(isobutylenes) obtained in this study were head-to-tail in structure and unsaturated at

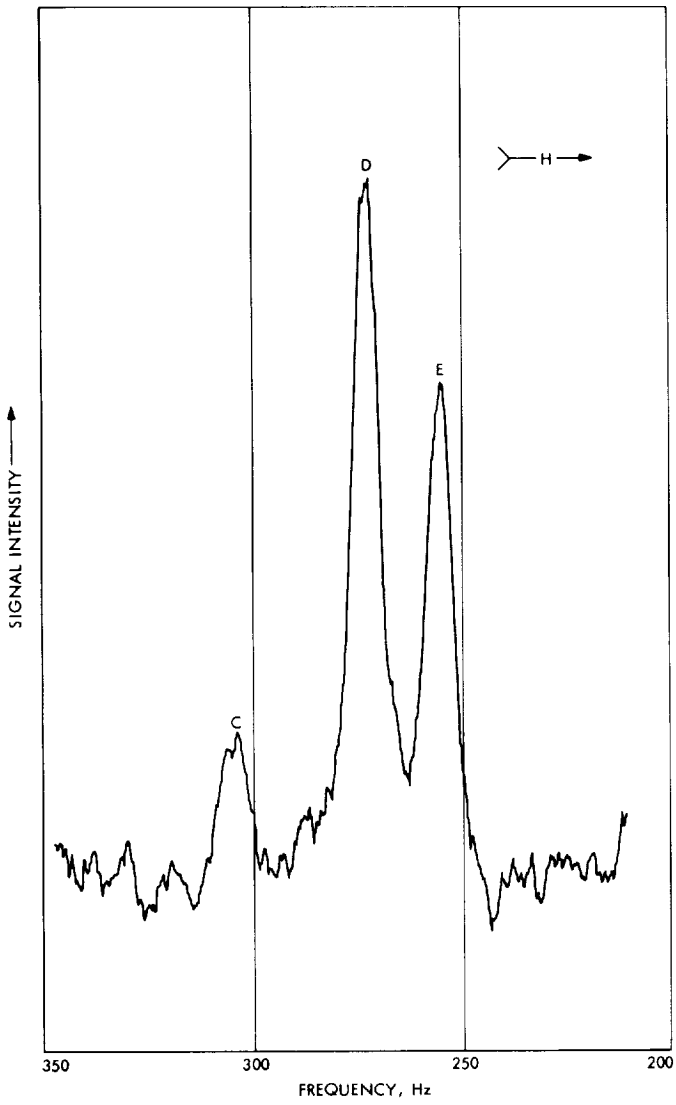
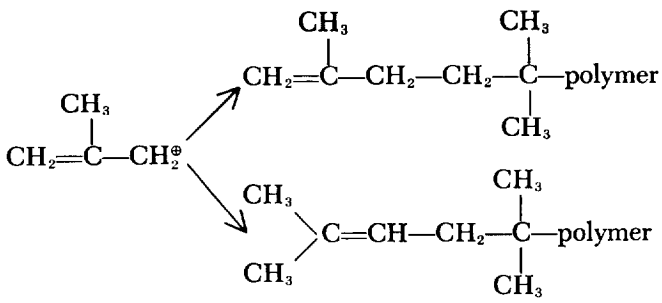


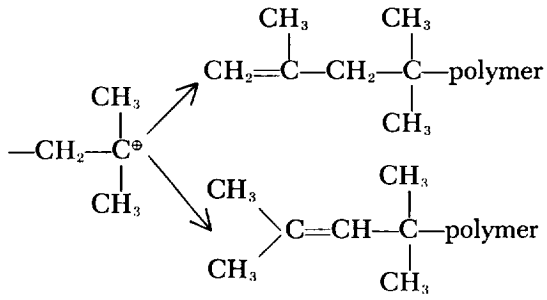
Fig. 11. Vinyl proton region of polymer 382-153-53

both ends by a combination of four types of groups. These types and their sources are summarized below.

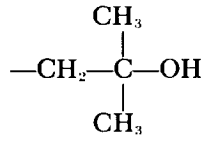
Initiation by



Termination from



or dehydration of

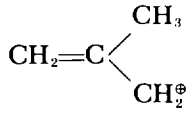


4. Polymerization Mechanism

More experimental details are needed before a total mechanistic description of this reaction can be presented. Based on the information available, the following statements can be made.

In view of the assumptions made to determine their values, the ratios of weight average to number average molecular weight (M_W/M_N) are reasonably close to two, indicating the reaction is stepwise in nature. The observed increase of M_N with time adds credence to this possibility.

As mentioned in *Subsection 3*, the fact that the poly(isobutylenes) are difunctional indicates initiation is performed by an unsaturated species. The methallylic cation

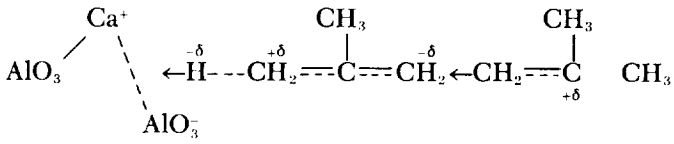


formed by hydride abstraction from isobutylene is a likely candidate. The thermodynamic stabilities of the methallylic and *t*-butyl cations are approximately equal, thus initiation would be an isoenergetic process.

A clue to the how of hydride abstraction is that molecular sieve in its divalent form is necessary in order for polymerization to proceed. Strong dipolar fields created by the incompletely neutralized calcium and aluminate ions are present in divalent sieve. These fields are capable

of polarizing the C—H band (Ref. 2). Possibly, they are sufficiently strong to effect hydride abstraction. Since the 5A sieve excludes isobutylene, the following picture can be painted:

(1) Hydride is abstracted by the sieve at the surface or in the activated channel to form the methallylic cation which initiates polymerization by attacking another isobutylene molecule. A push–pull mechanism can be visualized.



(2) The polymer would grow on the surface of the sieve, its positive end neutralized by an incompletely neutralized aluminate anion.

(3) Termination by proton elimination would occur upon depletion of the monomer supply.

Such a process, hydride abstraction followed by proton elimination, would be expected to produce hydrogen.

Another possibility presents itself if the *t*-butyl cation is formed by protonation of isobutylene in the 5Å hole. The initiating capabilities of this species, suppressed because of steric restrictions, would then serve as the hydride abstraction agent. The expected by-product of this reaction would be 2-methyl propane.

References

1. *Linde Molecular Sieves*, Technical Bulletin F-1979B. Union Carbide Corp., Linde Div.
2. Pickert, P. E., Bolton, A. P., and Lanewala, M. A., *Molecular Sieve Zeolites: Trendsetters in Heterogeneous Catalysis*, Technical Bulletin F-3187. Union Carbide Corp., Linde Div.
3. Weisz, P. B., et al., *J. Catalysis*, Vol. 1, p. 307, 1962.
4. Tetterly, L. C., and Koetitz, K. F., *Abnormal Hydro-Bromination of Substituted Olefins by Molecular Sieve Catalysis*, Conference at University of London, London, England, Apr. 4, 1967.
5. Skarstrom, C. W., U.S. Patent 3,155,741, U.S. Department of Commerce, Washington D.C., 1964.
6. Fleck, R. N., U.S. Patent 3,132,186, U.S. Department of Commerce, Washington, D.C., 1964.
7. Kozyreva, M. S., *Opt. Spektrosk.*, Vol. 6, p. 303, 1959.
8. Bacskai, R., and Lapporte, S. J., *J. Polym. Sci.*, Series A, Vol. 1, p. 2225, 1963.

E. Viscoelastic Behavior of Elastomers Undergoing Scission Reactions,

J. Moacanin, J. J. Aklonis,[6] *and R. F. Landel*

Although behavior of elastomers when viscoelastic relaxation and chemical reactions occur simultaneously is of enormous practical importance, there is a distinct lack of data on such systems in the literature. We were able to find only one paper which gives results suitable for quantitative analysis (Ref. 1). Even these data are limited in scope since the main interest of the authors was to determine conditions where viscoelastic response was not perturbed by chemical reactions. Of necessity, the data from this paper were used to test our theory (SPS 37-57, Vol. III, pp. 169–172).

In our previous work, the following function was used to simulate the relaxation behavior of an elastomer:

$$G(t) = \frac{G_g + G_e\left(\frac{t}{\tau}\right)^m}{1 + \left(\frac{t}{\tau}\right)^m} \tag{1}$$

This equation evidences a glassy plateau at short times as well as a pseudo-equilibrium rubbery plateau at long times. On the other hand, Thirion suggested that the following function can be used to fit his experimental results under conditions where only relaxation occurs:

$$f(t) = f_e + ct^{-n} \tag{2}$$

where $f(t)$ is the force at time t necessary to keep the sample extended at some fixed strain and c and n are parameters. It is easy to show that Eq. (1) becomes identical to Eq. (2) at long times. When $t >> \tau$, unity can be neglected in the denominator of Eq. (1), which can then be written as

$$G(t) = G_e\left[1 + \left(\frac{G_g}{G_e}\right)\left(\frac{t}{\tau}\right)^{-m}\right] \tag{3}$$

Multiplication through by the constant strain γ gives

$$f(t) = \gamma G_e\left[1 + \left(\frac{\tau^m G_g}{G_e}\right)t^{-m}\right] \tag{4}$$

[6]Department of Chemistry, University of Southern California, Los Angeles, California.

which is identical to Eq. (2), with the proper identification of constants. It is at short times that the two equations differ drastically. In the limit as $t \to 0$, Eq. (1) approaches G_g, as it should, whereas Eq. (2) goes to infinity. The latter behavior is trivial, however, since we are concerned with elastomeric properties at times much longer than a few seconds, i.e., $t >> \tau$. Therefore, for the sake of simplicity we use Eq. (2) in the following form:

$$G(t) = G_e\left[1 + \left(\frac{t}{\tau}\right)^{-m}\right] \tag{5}$$

In two papers dealing with the viscoelastic behavior of crosslinked elastomers in absence of degradation, Thirion tabulated values of m and τ for natural rubber preparations of varying crosslink density and for a range of temperatures (Refs. 2 and 3). As expected, τ was a strong function of both temperature and crosslink density. Values of m also varied somewhat. However, using the same rubber preparations, Plazek (Ref. 4) could fit both his creep data and Thirion's stress relaxation results on a single reduced master curve covering 15 decades of log-time. Crosslink density was used as a reduction variable. We were able to reproduce Plazek's experimental master curve with only three parameters, as prescribed by Eq. (5):

$$[J_r(t)]^{-1} = G_r(t) = 1.66 \times 10^6\left[1 + \left(\frac{t}{0.01}\right)^{-0.113}\right] \tag{6}$$

(Note: Here the reciprocal relationship between the creep compliance $J_r(t)$ and $G_r(t)$ is virtually exact since their rate of change is very small, i.e., $m \approx 0.1$.) The parameter $\tau = 0.01$ is for the reference molecular weight between crosslinks of $M_c = 20{,}500$. The value of $m = 0.113$ is independent of M_c. The constancy of m does not necessarily disprove Thirion's observation of a range of values. It merely indicates that when averaged over the extended time scale a single value adequately represents all the data. Furthermore, due to the relative simplicity of Eq. (6), the dependence of τ on M_c can be handled analytically. This dependence was previously neglected (SPS 37-57, Vol. III).

Proceeding, as previously, we write for a network undergoing an incremental scission ν to $\nu + \Delta\nu$ at time t (i.e., equate the stress at the instant of change).

$$f_\nu(t + \bar{m}) = f_{\nu+\Delta\nu}(t + \bar{m} + \Delta\bar{m}) \tag{7}$$

and

$$\gamma G_e(\nu)\left[1 + \left(\frac{t + \bar{m}}{\tau a_\nu}\right)^{-m}\right] = \gamma G_e(\nu + \Delta\nu) \times \left[1 + \left(\frac{t + \bar{m} + \Delta\bar{m}}{\tau a_{\nu+\Delta\nu}}\right)^{-m}\right] \tag{8}$$

where γ is the constant strain, τ the relaxation time characteristic of the initial crosslink density ν_0, i.e., at $t = 0$. The product τa_ν gives the relaxation time characteristic of the particular ν, and $\bar{m}$ is the amount by which the experimental time has to be shifted to satisfy Eq. (4). (This is the same quantity as a_t from SPS 37-57, Vol. III.) Remembering that G is proportional to ν, and γ is constant for a stress relaxation experiment, we can write

$$\left[1 + \left(\frac{t + \bar{m}}{\tau a_\nu}\right)^{-m}\right] = \left(1 + \frac{\Delta\nu}{\nu}\right)\left[1 + \left(\frac{t + \bar{m} + \Delta\bar{m}}{\tau a_{\nu+\Delta\nu}}\right)^{-m}\right] \tag{9}$$

Plazek (Ref. 4) showed that

$$a_\nu = \left(\frac{\nu_0}{\nu}\right)^b \tag{10}$$

with b constant for a given polymer system. Thus,

$$a_{\nu+\Delta\nu} = \left(\frac{\nu_0}{\nu}\right)^b\left(1 + \frac{\Delta\nu}{\nu}\right)^{-b} \approx a_\nu\left(1 - \frac{b\Delta\nu}{\nu}\right) \tag{11}$$

Substitution of this result into Eq. (9) gives, after considerable simplification and going to the limit of an infinitesimal change,

$$\frac{d\bar{m}}{d\nu} = (m\nu)^{-1}\left[\frac{(t + \bar{m})^{1+m}}{(\tau a_\nu)^m} + 1 - mb\right] \tag{12}$$

For short times, this expression is virtually zero. For sufficiently long times, however, the first term in brackets becomes much greater than $(1 - mb)$ and, therefore, the latter can be neglected. With this simplification, one obtains finally

$$\frac{d\bar{m}}{d\nu} = (m\nu)^{-1}\frac{(t + \bar{m})^{1+m}}{\tau^m}\left(\frac{\nu}{\nu_0}\right)^{mb} \tag{13}$$

But ν is determined by scission kinetics. In general, ν will vary as some function of the rate constant k and time, i.e.,

$$\nu = \nu_0 g(kt) \tag{14}$$

Hence, variable transformation gives

$$\frac{d\bar{m}}{dt} = \left(\frac{dg}{dt}\right)\frac{[g(kt)]^{mb-1}}{m\tau^m}t^{1+m}\left(1+\frac{\bar{m}}{t}\right)^{1+m} \quad (15)$$

Thirion (Ref. 1) has shown that the chemical degradation of natural rubber proceeds according to zero-order kinetics, at least for small extents of degradation. Thus, we may write

$$v = v_0(1-kt) \quad (16)$$

i.e.,

$$g(kt) = 1 - kt$$

Then the differential equation determining $\bar{m}$ becomes

$$-\frac{d\bar{m}}{dt} = \frac{k}{m\tau^m}t^{1+m}\left(1+\frac{\bar{m}}{t}\right)^{1+m}(1-kt)^{mb-1} \quad (17)$$

We have integrated the above differential equation subject to several approximations which are suitable for small extents of reaction, i.e., <5%, the range of available experimental data. Compared to unity, kt is small and thus it is neglected; $\bar{m}$ has to be smaller than t, which allows for a binomial expansion of the last term in Eq. (17); and for $\bar{m}/t < 0.5$, the higher-order terms in the expansion can be neglected, resulting in

$$\frac{d\bar{m}}{dt} + (1+m)\frac{k}{m\tau^m}t^m\bar{m} = -\frac{k}{m\tau^m}t^{1+m} \quad (18)$$

The integrating factor

$$\exp\left(\frac{k}{m\tau^m}t^{1+m}\right)$$

permits the direct integration of Eq. (18) between the limit $t = \bar{m} = 0$ and $t = t$, $\bar{m} = \bar{m}$, giving

$$-\bar{m} = (1+m)^{-1}\left(\frac{m\tau^m}{k}\right)^{1/(1+m)}\exp\left(-\frac{k}{m\tau^m}t^{1+m}\right) \times \int_0^{(k/m\tau^m)t^{1+m}} Z^{1/1+m}e^Z\,dZ \quad (19)$$

which may be simplified to

$$-\bar{m} = (1+m)^{-1}\left(\frac{m\tau^m}{k}\right)^{1/(1+m)}Z^{*(2+m)/(1+m)} \times \int_0^1 Y^{1/(1+m)}\exp Z^*(Y-1)\,dY \quad (20)$$

where

$$Z^* = \frac{k}{m\tau^m}t^{1+m}$$

Equation (20) offers an explicit expression for $\bar{m}$ in terms of the experimentally available values for m, τ, and k. It should be noticed that although the parameter b is not present in this equation, its influence is still felt in the overall outcome of the calculation since it was used to calculate the various values of the relaxation time as a function of the polymer crosslink density. These values are used to calculate the appropriate stress relaxation modulus.

We are presently using Eq. (20) to analyze experimental data from Ref. 1 and are making an effort to obtain a general solution for Eq. (17). In particular, we are interested in a solution which is good for $\bar{m}$ approaching t, a situation which initial calculations indicate will be reached for fairly low extents of reaction.

References

1. Thirion, P., and Chasset, R., *Rubber Chem. and Tech.*, Vol. 37, p. 617, 1964.
2. Thirion, P., "Viscoelastic Relaxation of Rubber Vulcanizates," *Proceedings of the International Conference on the Physics of Non-Crystal Solids*, Amsterdam, North Holland, p. 345, 1965.
3. Thirion, P., *Propriétées Viscoélastiques des Vulcanisats*, Rapport de Recherches No. 49, Institut Francais Caoutchouc, Paris, France, 1963.
4. Plazek, D. J., *J. Poly. Sci.*, Series A-2, Vol. 4, p. 745, 1966.

F. Studies on Polymeric Materials Intended for Use in the Venus Environment, *E. F. Cuddihy and J. Moacanin*

1. Introduction

The high temperature on Venus—believed to be in excess of 550°F—imposes severe limitations on the selection of materials for the construction of a lander. With very few exceptions, this temperature exceeds the upper limit of thermal stability of almost all polymeric materials.

Previously, a screening program for evaluating commercially available candidate polymer materials was carried out. The test procedure consisted in exposing materials to a simulated Venus environment (Table 6) for 6, 24, and 72 h and noting afterwards on the basis of simple physical test whether the materials passed, survived marginally, or failed. The specific details of this screening test, the materials tested, and their performance are discussed by S. Kalfayan and R. Silver in SPS 37-57, Vol. III, pp. 157–163.

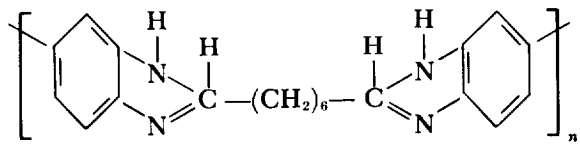
The present study was complementary to the screening program. Its purpose was to provide an in-depth analysis of materials that exhibited marginal properties in order to identify the mode of failure and propose practical approaches for making them passable. This effort will also result in acquiring fundamental knowledge needed to guide in preparing classes of new materials or selecting or modifying commercial polymers.

Initial emphasis of this study was on a marginally performing moldable benzimidazole (EXP-820, Whittaker Corporation) and a failed glass-fiber-filled poly(phenylene oxide) (General Electric Company).

2. Moldable Benzimidazole

Polybenzimidazoles are a new class of aromatic-based polymers (Ref. 1) having good high-temperature properties. But these materials usually degrade at the excessively high temperatures needed to soften them sufficiently to permit molding. In order to lower the molding temperature, the Whittaker Corporation has incorporated aliphatic linkages into the aromatic backbone.

A sample of an experimental polymer obtained from the Whittaker Corporation, poly-2,2′-hexamethylene-5,5′-bibenzimidazole



designated EXP-820, was subjected to the Venus screening test. Although the material noticeably discolored (compare A and B in Fig. 12) through the first 6 h of exposure, it experienced a weight loss of only 0.7%. But because the material had apparently flowed resulting in excessive dimensional changes, the polymer was designated marginal. It was later discovered that the discoloration was only on the surface (about 1 mil).

Table 6. Simulated Venus environment

Parameter	Environment
Temperature, °F (°C)	550 (288)
Pressure, psig	216
Composition of atmosphere, %	
Water	0.1
N_2	0.5
O_2	0.8
Ar	0.01
CO_2	Remainder supplied by tank 99.99% pure
Exposure	Fresh samples are exposed for 6-, 24-, and 72-h periods

a. Dynamic mechanical properties. The observation of flow suggested that the temperatures at which this material softened must be below the test temperature of 550°F (288°C). This can be readily measured for a polymer from the maxima in the temperature dependence of the loss tangent from dynamic measurements. Also, a comparison of the dynamic modulus and loss tangent curves for the material before and after the test provides a sensitive method for assessing changes in properties.

These curves are shown in Fig. 13, from which it can be seen that EXP-820 starts to soften at 200°C and has essentially become rubbery by 280°C, explaining the observed flow and dimensional changes exhibited by the sample. More important, however, is the observation from the dynamic properties that the material is unchanged after the exposure, which points up the excellent thermal stability of EXP-820. Comparing A and C in Fig. 12 shows that the thermal exposure acted to clear the polymer, producing a more homogeneous product as compared to the initially very heterogeneous polymer. Perhaps the material as received is undercured and the thermal exposure continued the cure. This process would buy additional exposure time at the elevated temperature before degradation would dominate. This continuation of the cure could also explain the increase in modulus observed for EXP-820 at around 200°C.

In all, then, the problem for EXP-820 is not stability or degradation but softening, which at present prohibits its use. Two modifications immediately present themselves for improving EXP-820. One is to recognize that it is the aliphatic component of hexamethylene, six contiguous methylene units, which controls the temperature range of softening. Either lowering the concentration of hexamethylene, or decreasing the length of the aliphatic chain to less than six continguous methylene units, or

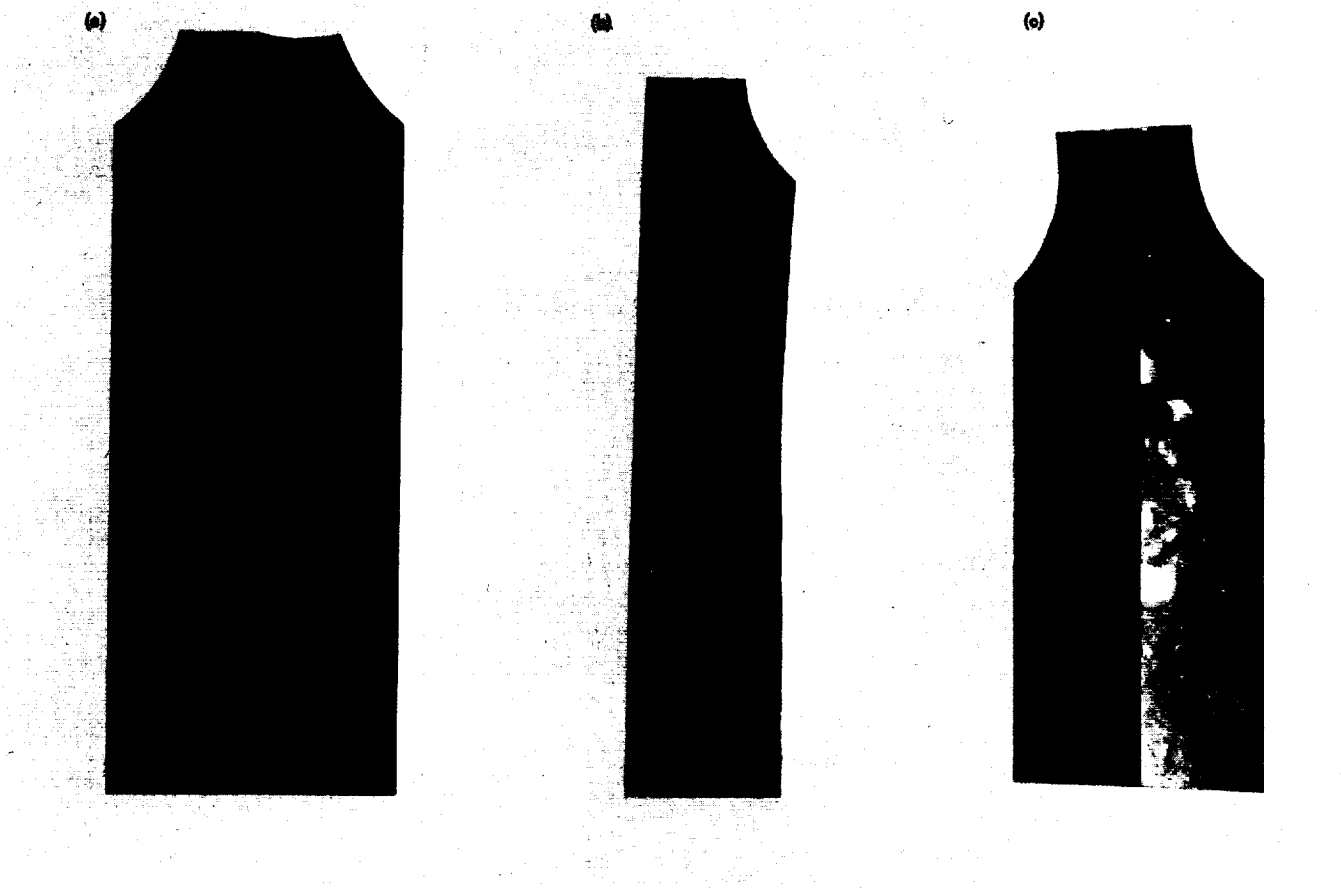


Fig. 12. Physical condition of poly-2,2′-hexamethylene-5,5′-bibenzimidazole: (a) before exposure to simulated Venus environment, (b) after 6-h exposure to simulated Venus environment, (c) after 6-h exposure to simulated Venus environment with portion of surface char removed to expose remainder of sample (surface char approximately 1-mil thick)

both, could result in raising the softening temperature range. This route, of course, would require some fundamental chemistry.

b. Filled EXP-820. The second and simpler modification is to fill EXP-820 with a reinforcing filler. Since the binder, in this case EXP-820, is stable, the solid composite should survive in excellent shape. Samples of a composite prepared from a glass-fiber weave and poly-2,2′-hexamethylene-5,5′-bibenzimidazole (EXP-820A) were received from the Whittaker Corporation and subjected to the Venus screening test. In addition, a composite prepared from unidirectional graphite fibers was tested but fell apart along the direction of the graphite fibers when an attempt was made to measure its Rockwell hardness. The glass-fiber-weave composite, on the other hand, survived in excellent shape after 72-h exposure in the simulated Venus environment. The material underwent no dimensional changes and only experienced a weight loss of 2.3%. A summary of physical properties is given in Table 7. The modulus curves of the composite given in Fig. 14 show that the high-temperature modulus of the system was increasing with increasing exposure time; thus the system's rigidty and dimensional stability were constantly improving. The reinforcing action of the glass-fiber-weave filler on EXP-820 can be dramatically observed by comparing Figs. 13 and 14.

3. Poly (Phenylene Oxide)

A chopped-glass-fiber-filled poly(phenylene oxide) (PPO, General Electric Company) intended for use as an impact-resistant battery case failed the Venus screening test. After 6 h, the material had extensively degraded,

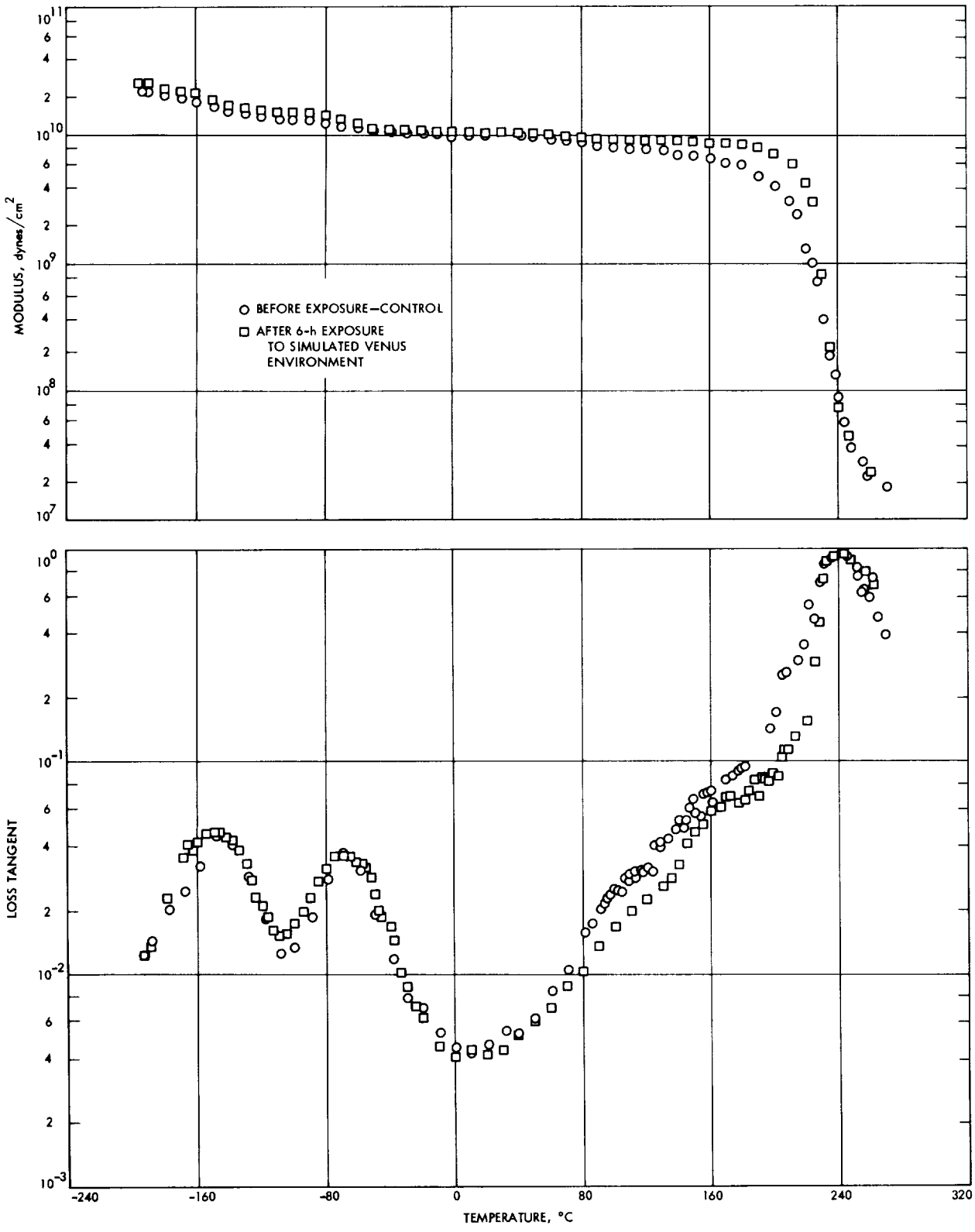


Fig. 13. Dynamic mechanical properties of poly-2,2′-hexamethylene-5,5′-bibenzimidazole

Table 7. Summary of properties of glass-fiber-weave composite with poly-2,2′-hexamethylene-5,5′-bibenzimidazole and the unfilled polymer

Exposure time simulated Venus environment	Tensile strength at break,[a] psi	Elongation at break,[a] %	Rockwell hardness[a, b]	Density,[a] g/cm³	Weight loss, %
Control	26,880	6.5	80	1.793	0.0
6 h	31,990	6.5	83	1.870	0.7
24 h	[c]	[c]	85	1.862	1.2
72 h	31,250	5.5	85	1.874	2.3
Unfilled—control	[c]	[c]	80	1.172	0.0
Unfilled—6 h	8,080	4.0	[d]	1.156	0.7

[a]Measured at room temperature.

[b]Rockwell hardness scale E, 15-s wait.

[c]Sample broke prior to test.

[d]Not measured.

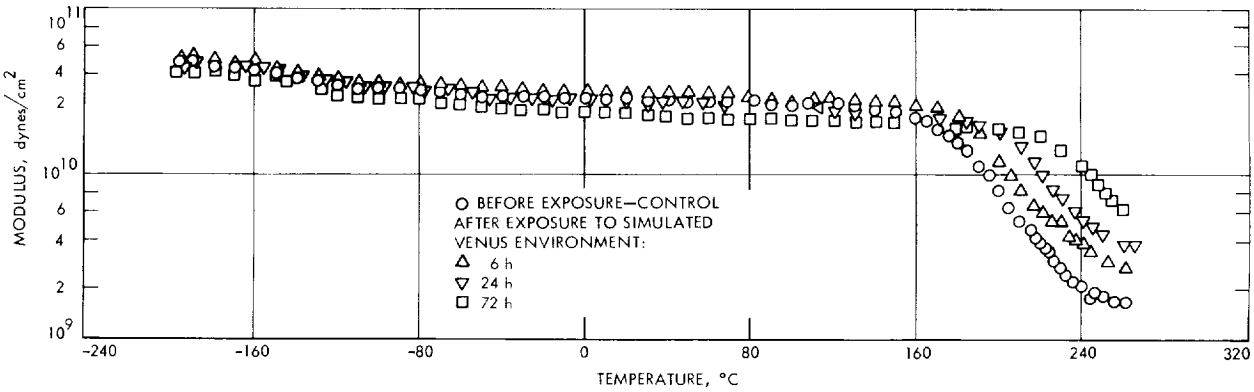


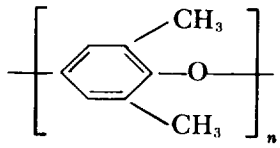
Fig. 14. Modulus curves of glass-fiber-weave-filled poly-2,2′-hexamethylene-5,5′-bibenzimidazole

leaving as residue a black, puffy char easily crushed by hand.

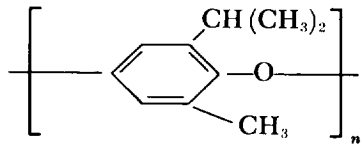
The PPO which softens near 190°C was reinforced with the chopped-glass fiber to prepare a composite having rigidity and high modulus at the simulated Venus temperature of 288°C. Though in principle the reinforcement can work as seen for the EXP-820, it failed here because the PPO binder degraded.

Thermal gravimetric analysis (TGA) of PPO has found this material to start undergoing degradation in air around 110–120°C and in inert atmospheres around its softening temperature of 190°C (Ref. 1). In either environment at 288°C, TGA revealed that PPO degrades rapidly. This, of course, substantiates the current finding in the Venus simulator and would ordinarily be cause for no further interest. However, it has been reported in one case that a chemical modification of PPO resulted in a significant improvement in thermal stability.

The structure of PPO, chemically poly(2,6-dimethyl-1,4-phenylene oxide), is



Conley and Alvino (Ref. 2) reported that substituting isopropyl for one of the methyl groups yielded poly(2-methyl-6-isopropyl-1,4-phenylene oxide)



which achieved improvements in the high-temperature stability. After 90 min in oxygen at 250°C, the dimethyl

polymer charred badly and had lost 46% of its initial weight, whereas the isopropyl product, although darkened, had lost only 1.8% of its initial weight. It was suggested by the authors that the material was preserved through a crosslinking reaction involving the isopropyl group.

In future efforts directed toward developing acceptable Venus polymeric materials, or a general list of thermally stable polymers, a systematic investigation of the chemical modification of PPO should be included.

References

1. Lee, H., Stoffey, D., and Neville, K., *New Linear Polymers*. McGraw-Hill Book Co., Inc., New York, N. Y., 1967.
2. Conley, R. T., and Alvino, W. M., "The Thermo-oxidative Degradation of Poly(phenylene) Ether Polymers," *ACS Organic Coatings and Plastics Chemistry*, Vol. 25, No. 2, 1965.

XIX. Research and Advanced Concepts

PROPULSION DIVISION

A. 70 kWe (Net) Thermionic Reactor Plant Arrangement, *J. P. Davis*

Thermionic-reactor systems are candidates for unmanned electric propulsion application. A representative mission of interest is an unmanned Jupiter orbiter, the general characteristics of which have been studied previously.[1] At present, General Electric is conducting an overall system study based on a thermionic reactor energy source under contract to JPL. The emphasis is on a 300 kWe unit utilizing thermionic reactor design evolved under AEC contracts. Such a system would be launched to 700-nmi orbit by a *Titan IIIX* (1207) booster and spiral out from earth orbit. The other mode of injection often considered is direct launch to escape utilizing a *Titan* or *Titan-Centaur* booster. The optimum nuclear power plant size for this mode of launch is 50–100 kWe, depending on the electric propulsion system specific weight.

The *Titan IIIX* allowable payload envelopes are shown in Fig. 1. Values for the dimensions shown in Fig. 1 are given in Table 1 along with launch probability estimates.

[1]Stearns, J .W., *et al, Briefing Prepared for a Joint Meeting of the National Aeronautics and Space Administration and the Atomic Energy Commission on Nuclear–Electric Propulsion for Unmanned Space Exploration* (JPL internal document).

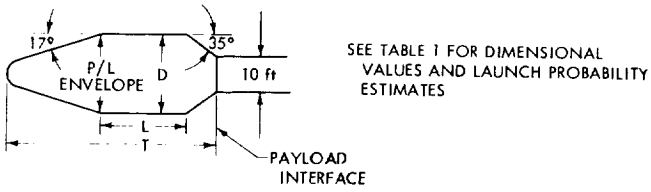


Fig. 1. *Titan III* payload envelopes

Table 1. *Titan III* payload envelopes dimension values and launch probability estimate (Fig. 1)

Diameter (D), ft	Cylindrical length (L), ft	Total length (T), ft	Launch probability, %
10	42.0	50	99
13	14.5	31	99
13	23.5	40	99
15	16.5	38	99
15	22.5	44	99
18	20	46	99
18	24	50	99
20	22	50	89
20	24	53	86

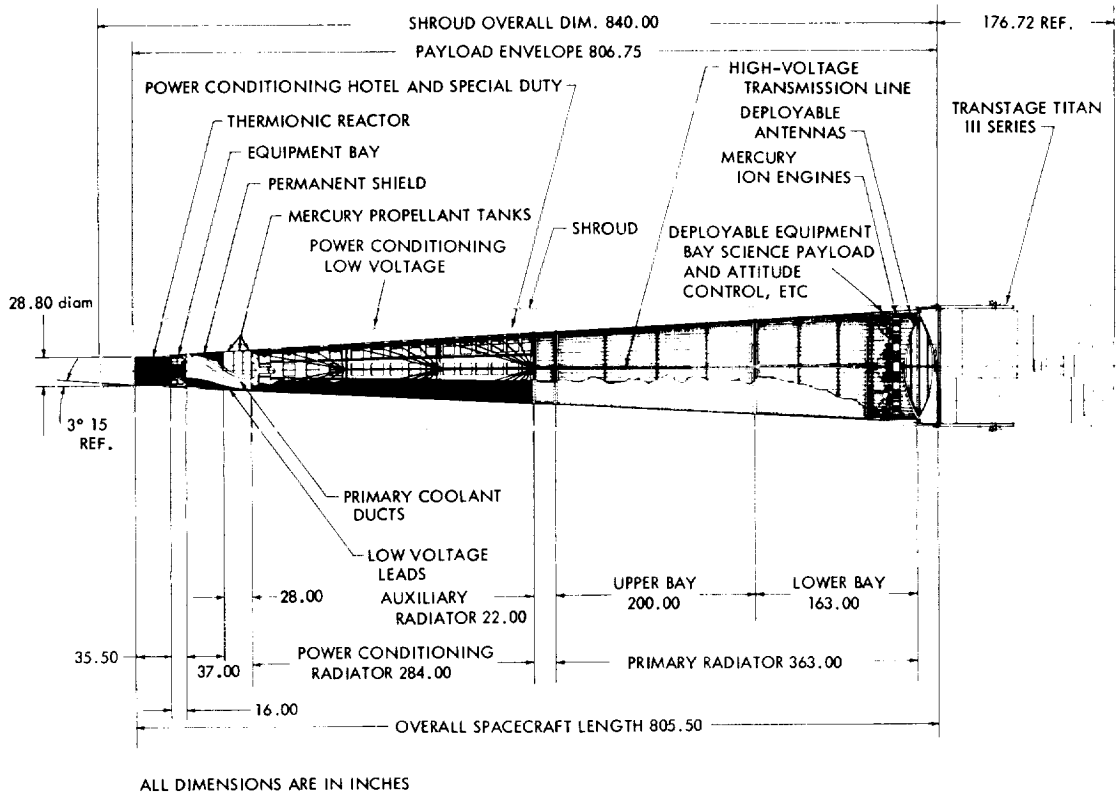


Fig. 2. 300-kWe (gross) thermionic plant arrangement

A typical 300 kWe plant arrangement developed by General Electric is shown in Fig. 2. This arrangement is fairly typical of those proposed for nuclear–electric propulsion application, which are characterized by having the reactor at one end of the spacecraft and radiation-sensitive electronics at the other. Since radiation scattering from the radiators can be a controlling source of dose to the payload, it is necessary to shield these radiators from direct reactor flux. The shielding weight is a very strong function of nose-cone angle as shown in Fig. 3.

Reference 1 describes the basic thermionic-reactor types presently under investigation. For thermionic-reactor systems of the flashlight type, voltage output is in the 5–20 V neighborhood and it is mandatory to locate power conditioning close to the reactor. Thus, this power conditioning becomes the controlling item determining shield weight, and advantages of large separation distances potentially afforded by a "nose-end" reactor location are nullified. A further disadvantage of the nose-end reactor location is that the shield interferes with coolant piping to the radiator, which must either penetrate or be routed around the shield periphery. The requirement in low-voltage output thermionic systems for the close

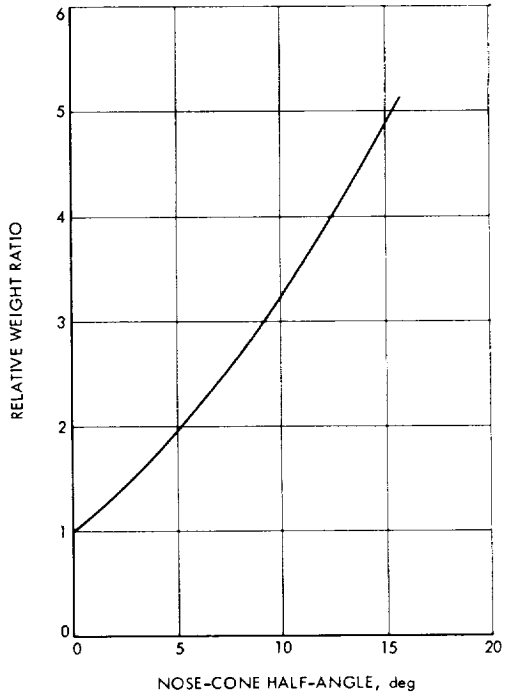


Fig. 3. Typical effect of nose-cone half-angle on shield weight

proximity of power conditioning and its associated low-temperature radiator is an additional thermal-control problem requiring that a low-temperature region be sandwiched between two high-temperature regions, i.e., the reactor and the primary radiator, with high-temperature plumbing passing through the cooler region.

At the 300-kWe power level where overall plant length must be restricted to that launchable on a particular booster, there appears to be little choice other than the configuration of Fig. 1 since a significantly larger radiator diameter than reactor diameter is required simply to afford the required heat rejection area. To arrange the primary radiator on the same side of the shield as the reactor appears, at first glance, to require prohibitively heavy shielding weights to adequately shield for the radiation backscattered from the radiator toward sensitive electronic components (Fig. 4). This has not yet been verified quantitatively and there is some possibility that the weight penalties may not be as high as presently estimated. General Electric will be investigating the possibility of this arrangement at the 300-kWe level.

At the 70-kWe level, the picture changes dramatically and the primary radiator can definitely be located on the reactor side of the shield within the total length constraint of the *Titan III*, with a zero cone-angle requirement. A simple schematic of this arrangement is shown in Fig. 5. This arrangement will be pursued in a detailed study at JPL in which the initial considerations will be the structural support requirements for the launch environment where bracing of the very high length-to-diameter ratio spacecraft, either internally or from the shroud, will undoubtedly be required. The potential merits of this arrangement, however, seem worth a detailed investigation.

With respect to the reactor, the basis for this study is the JPL uninsulated, externally-fueled arrangement (SPS 37-46, Vol. IV, pp. 142–147) shown in hexagonal cluster form in Fig. 6 and in cylindrical fuel element form in Fig. 7. This reactor utilizes the multi-ducted electromagnetic pump principle (SPS 37-49, Vol. III, pp. 201–207, Fig. 4). A thermal and electrical distribution schematic is shown in Fig. 8.

In brief, the advantages accruing from an arrangement having the primary radiator adjacent to and on the same side of the shield as the reactor are:

(1) Elimination of shield penetrations by coolant piping.

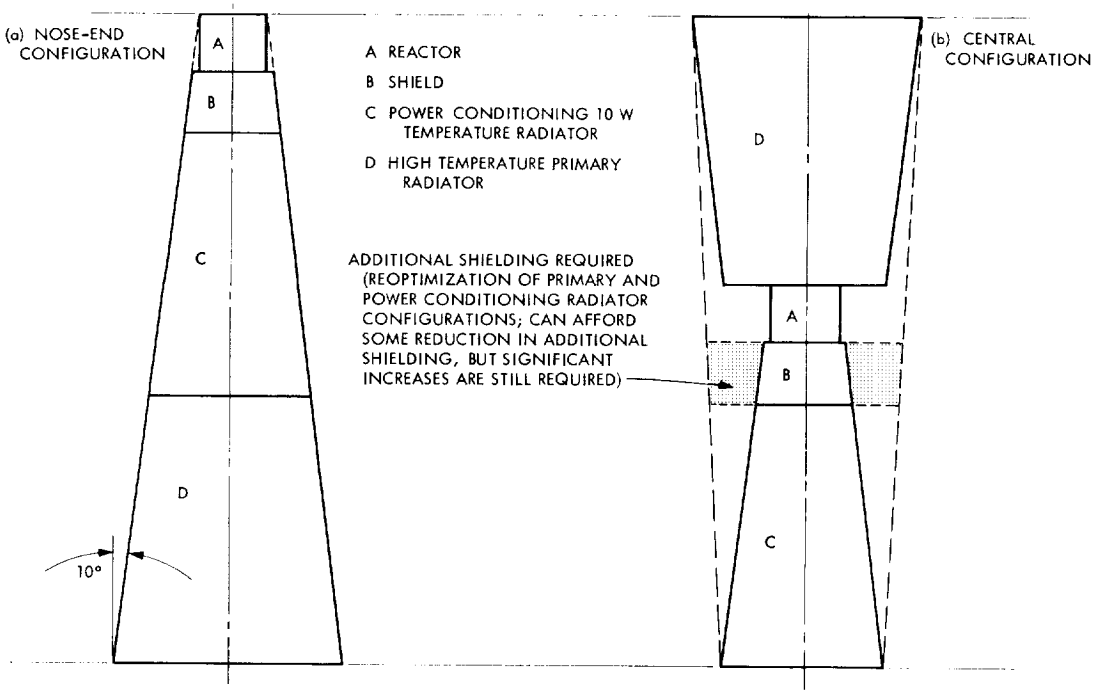


Fig. 4. Effect of nose-end vs central reactor location on shielding for length-limited spacecraft

(2) Elimination of radiator as significant nuclear radiation source, either by scattering or coolant activation.

(3) Reduction in shield weight if reduction in radiation source cone-angle can be accomplished.

(4) Elimination of hot coolant transversing cold regions, and cold busbars and cabling transversing hot regions, and improved thermal control.

(5) Reactor/shield masses closer to booster base-plate.

Potential disadvantages are:

(1) Structural support requirements for a high length-to-diameter ratio spacecraft.

(2) Lack of esthetic appeal.

Reference

1. Davis, J. P., *et al, Review of Proposed In-Pile Thermionic Space Reactors, Part I—General,* Technical Memorandum 33-262. Jet Propulsion Laboratory, Pasadena, Calif., Oct. 15, 1965.

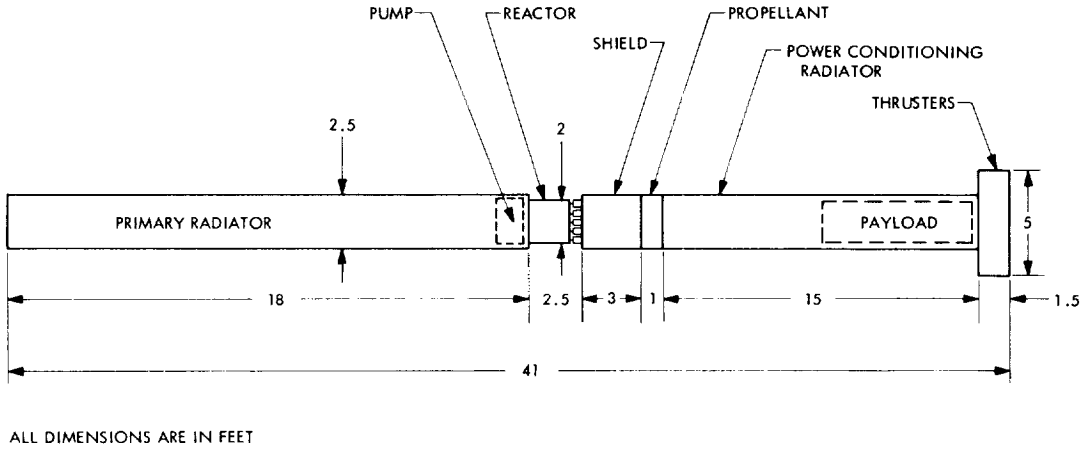


Fig. 5. 70-kWe thermionic reactor plant

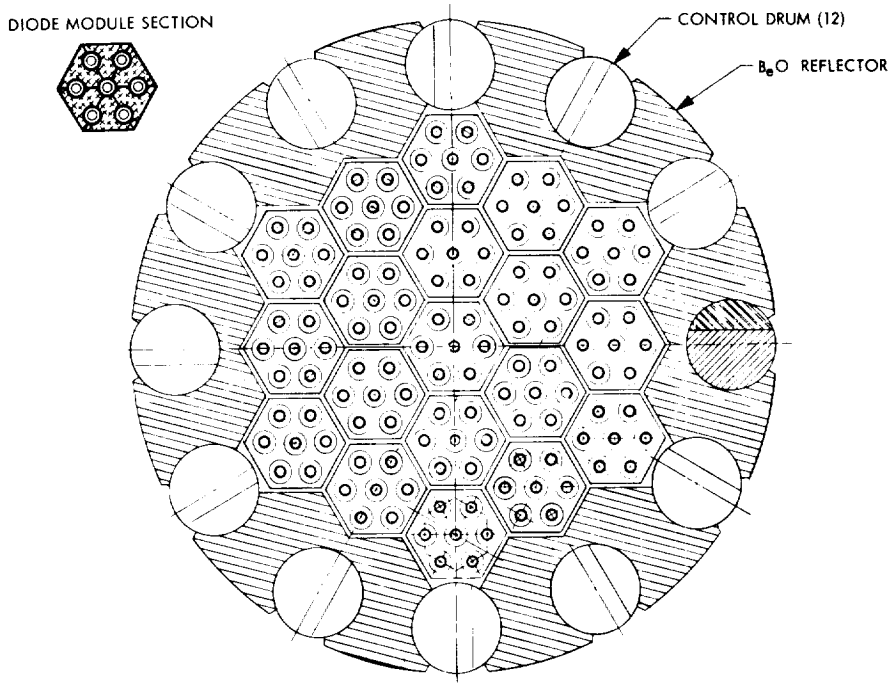


Fig. 6. 70-kWe thermionic reactor-hexagonal clusters arrangement

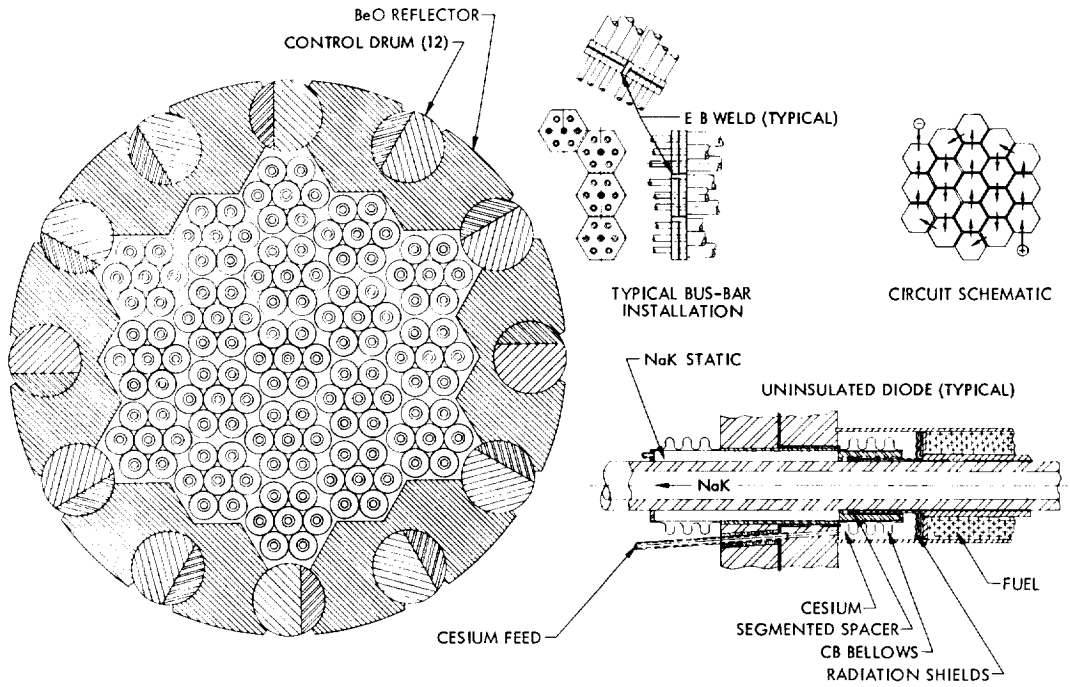


Fig. 7. 70-kWe thermionic reactor–cylindrical fuel elements arrangement

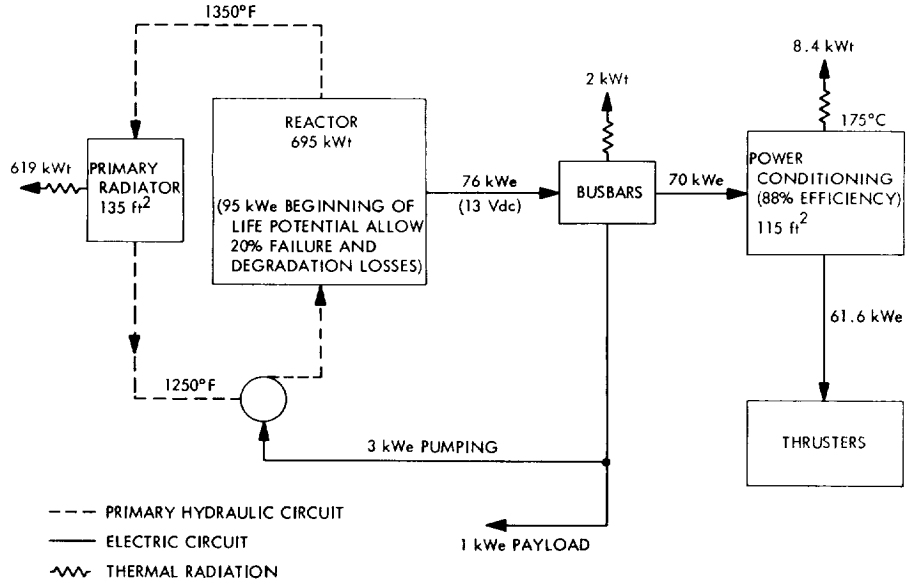


Fig. 8. 70-kWe thermionic power plant electrical distribution

B. Neutralization of a Movable Ion Thruster Exhaust Beam, *E. V. Pawlik*

1. Introduction

An ion thruster system containing many electric propulsion elements that would be necessary for a mission such as described in Ref. 1 is currently being investigated experimentally. Some results obtained from a study of this system are presented in Refs. 2 and 3. An essential component is the cesium plasma-bridge neutralizer, mounted at a fixed location, which supplies electrons across a varying distance to electrically neutralize the positively charged ion beam of a gimballed thruster. Experimental data has been obtained to determine the penalties associated with this mode of operation.

2. Test Setup

A photograph of the mechanical portion of the system mounted on a vacuum chamber end dome is shown in Fig. 9. The ion thrusters are the 20-cm diam electron-bombardment type that utilize a plasma source and employ mercury as the propellant. Each thruster can be gimballed $\pm$ 10 deg about a single axis. Two cesium neutralizers of the type described in Ref. 4 were purchased from Electro-Optical Systems (EOS) for use in the system. The thrusters, gimbal actuators, and neutralizers are mounted on a common structure that can be translated along a single axis. The neutralizer is essentially a small-orifice (0.005-in. diam), hollow cathode

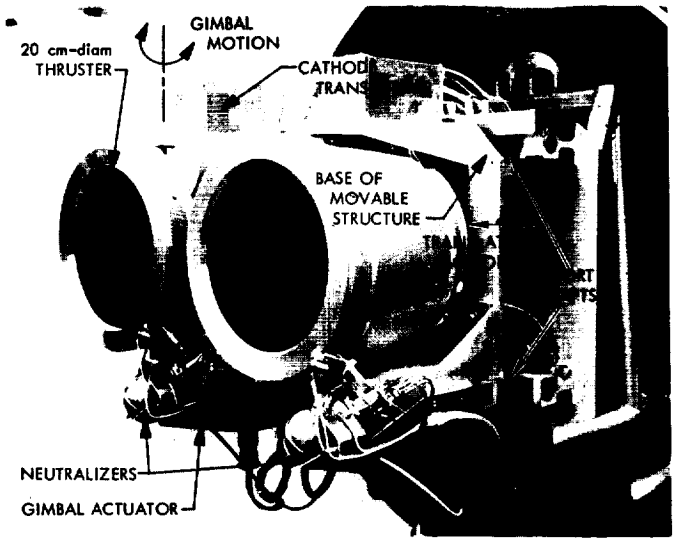


Fig. 9. Vacuum portion of an electric propulsion system

using cesium to provide a plasma that bridges the gap between the electron source and the ion beam. A keeper electrode is located slightly downstream of the cathode and used to electrostatically initiate a plasma discharge and maintain the flow of cesium vapor at a desired rate by a closed-loop feed back control. The keeper voltage, necessary to maintain a constant current, serves as a flow-rate sensor and provides an error signal to the current-regulated vaporizer power supply. A vaporizer thermocouple provides an independent indication of this cesium flow. A schematic diagram of the neutralizer electrical connections is shown in Fig. 10.

A second loop sets the maximum permissible current to the vaporizer and cathode heaters, which are connected in series. This loop is included to prevent thermal runaway of the neutralizer during initial turn-on and restart after thruster arcing, and provides no controlling function during normal operation.

A third control loop is required to control the neutralizer emission current. The neutralizer must be biased negatively with respect to the beam to allow electrons to flow from the neutralizer into the beam. The magnitude of this coupling voltage is a direct measure of the efficacy of the neutralizer. In the test system, the ion beam is directly coupled to the vacuum chamber wall, and its potential is essentially zero. A bias supply on the neutralizer provides the coupling voltage, which is a current-regulated supply with the regulated current set to equal the beam current. The bias voltage then varies as necessary to maintain the required emission current.

The exact placement of the neutralizer with respect to the thruster is shown in Fig. 11. A half-cone angle of 40 deg was assumed for the spread of the ion beam after leaving the thruster. This angle was determined for a similar thruster described in Ref. 5. The neutralizer was directed 45 deg downstream and placed so as to be slightly removed from the beam edge when the thruster was gimballed to the outermost position.

The ion thruster and neutralizer were both operated from a flight-type power conditioning unit designed to function with a solar array. The system was designed to run over a 1:2 range in output power corresponding to a 0.5–1.0-A variation in ion beam current. One variable input signal was used to specify the power level. An internal set point could be used to set a cesium flow that would be adequate over the entire output power range.

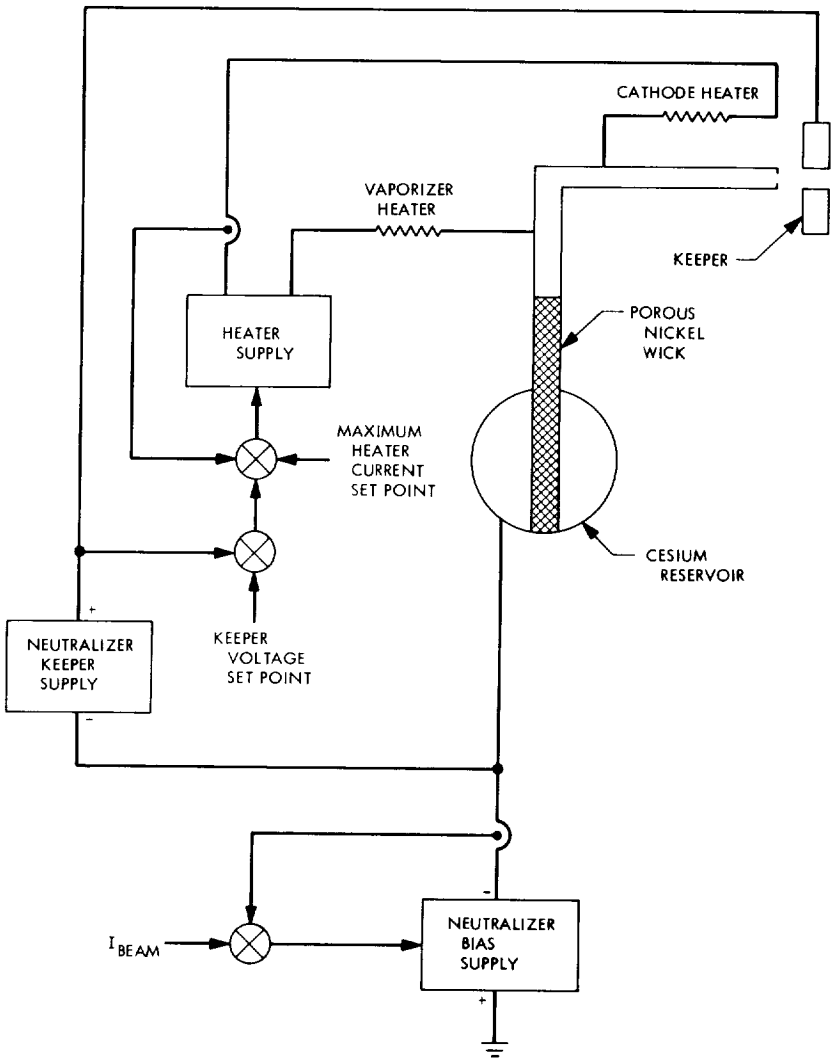


Fig. 10. Neutralizer wiring diagram

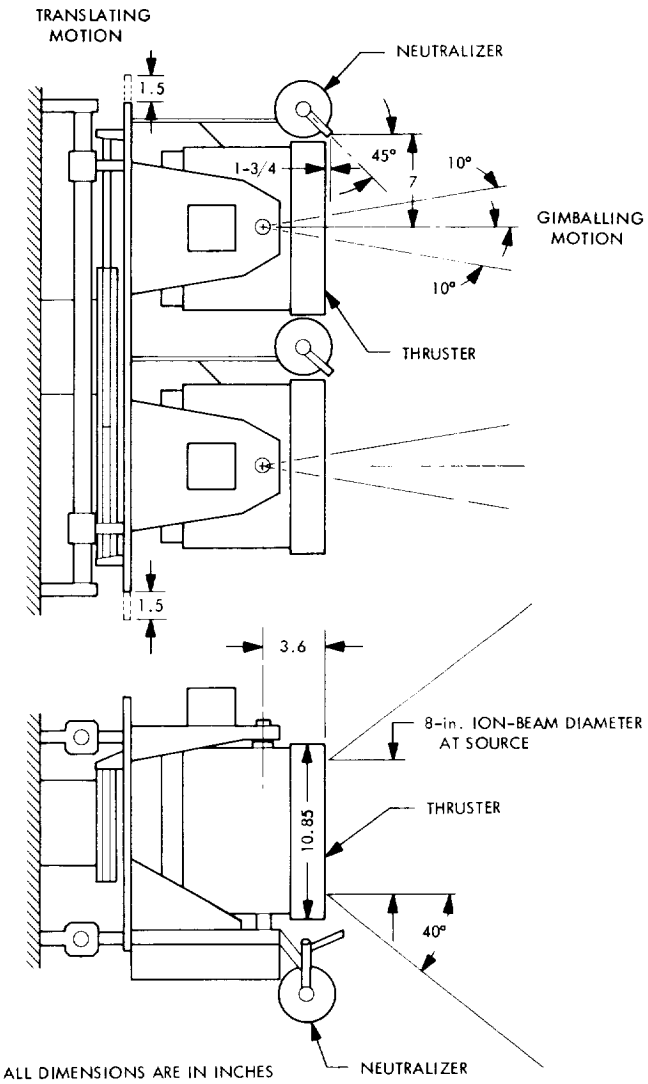


Fig. 11. Neutralizer placement with respect to thruster

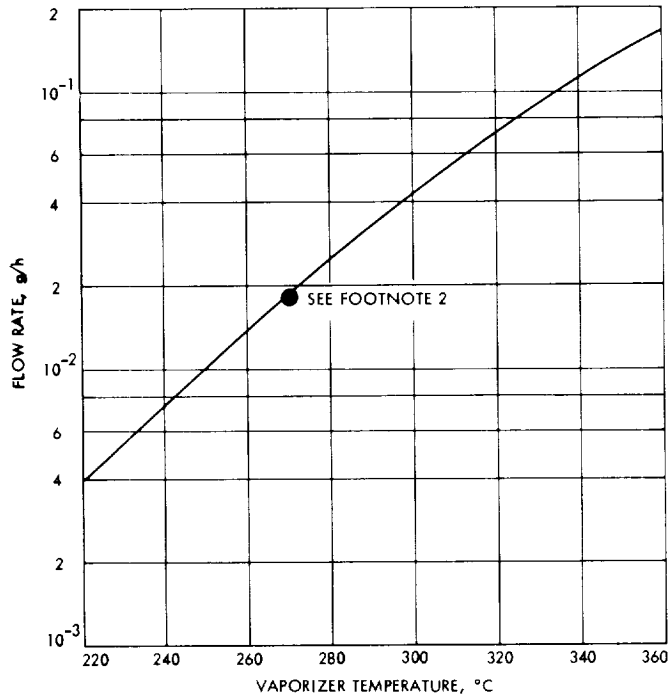


Fig. 12. Estimated neutralizer flow rate vs vaporizer temperature

3. Results

The keeper voltage was maintained at a set point of 5.2 V. The cesium flow for this set point was derived from an extrapolation as presented in Fig. 12. The solid data point represents the flow rate as measured during an endurance test on this type of neutralizer at EOS[2]. A curve fitted through this point with the flow proportional to the vapor pressure divided by the square root of the vaporizer temperature. Flow rates derived on this basis were on the order of 0.09 g/h. The bias voltage necessary to maintain the neutralizer emission equal to the beam current as the thruster was gimballed is presented in Fig. 13 for two values of beam current. Data point values for Fig. 13 are as follows:

Data point symbol	Beam current, A	Neutralizer vaporizer temperature, °C	Neutralizer cesium flow, g/h	Thruster mercury flow, g/h
○	0.480	328	0.087	4.00
□	0.913	331	0.091	7.61

Low voltage levels of neutralizer to beam coupling were found to exist for all values of gimbal position. These voltages were considerably below the 30-V level where detrimental wear has been observed (Ref. 5). The cesium flow was calculated to be 1.2–2.2% of the ion beam. No changes in the neutralizer set point were attempted at the low beam current operation. Operation of this neutralizer type with a moving beam interface yields acceptable coupling voltages and only a small penalty (~ 1%) in cesium flow. This performance might be improved if the set point were adjusted proportionately to the power output level.

[2]Personal communication from Mr. E. James of EOS.

References

1. *1975 Jupiter Fly By Mission Using Electric Spacecraft*, ASD 760-18, March 1968.
2. Pawlik, E. V., Macie, T., and Ferrera, J., "Electric Propulsion System Performance Evaluation," Paper 69-236, presented at the Seventh AIAA Electric Propulsion Conference, Williamsburg, Va., Mar., 1969.
3. Macie, T. W., Pawlik, E. V., Ferrera, J. D., and Costogue, E. N., "Solar Electric Propulsion System Evaluation," Paper 69-498, presented at the Fifth AIAA Propulsion Joint Specialist Conference, U. S. Air Force Academy, Colorado Springs, Colo., June, 1969.
4. Sohl, G., Fosnight, V. V., and Goldner, S. J., *Electron Bombardment Cesium Ion Engine System*, Report 6954, Electro-Optical Systems, Pasadena, Calif., Apr., 1967.
5. Rawlin, V. K., and Pawlik, E. V., *A Mercury Plasma-Bridge Neutralizer*, NASA TM X-52335, National Aeronautics and Space Administration, Washington, 1967.

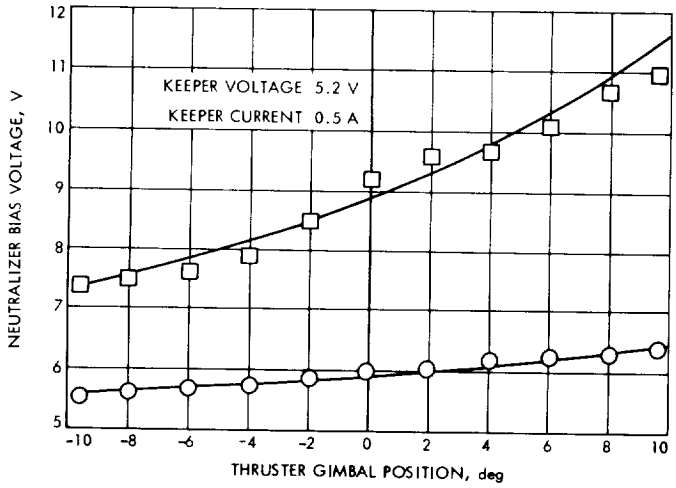


Fig. 13. Neutralizer bias voltage vs thruster gimbal position

C. Liquid–Metal MHD Power Conversion,

L. G. Hays

Liquid-metal magnetohydrodynamic (MHD) power conversion is being investigated as a power source for nuclear-electric propulsion. A liquid-metal MHD system has no moving mechanical parts and operates at heat-source temperatures between 1200 and 1400°K. Thus, the system has the potential of high reliability and long lifetime using readily available containment materials such as Nb–1%Zr.

Fabrication of a 50-kWe NaK–nitrogen conversion system and of a 100-kWt Cs–Li loop is continuing, and a new type of separator with no wall friction is being tested.

Evaluation of the flow resulting from the impingement of two-phase jets was continued. Substantial concentration of the liquid phase had been shown to occur (SPS 37-57, Vol. III, pp. 177–179) and tests were initiated to provide quantitative characterization of the flow.

The experimental apparatus is shown in Fig. 14. Nitrogen and water are supplied to the nozzles through flexible lines. The two nozzles are mounted so that the angle of impingement of their exit flows can be varied from 5 to 30-deg half angle. The resultant flow exits from the apparatus through a gap formed by two knife-edge blocks and through secondary outlets. The flow leaving through the knife-edge gap produces a thrust that is measured by a load cell. This force measurement enables determination of exit velocity and volume ratio of gas to liquid when the exit mass flow is known. The exit mass flow is determined by measuring the flow from the secondary outlets and subtracting from the total inlet flow, which is measured by turbine-type flowmeters. The width of the gap and its axial location are varied by changing the knife-edge blocks, enabling measurement of the axial and transverse variation in flow properties.

A preliminary determination of the distribution of liquid at the best axial locations (determined experimentally) is shown in Fig. 15 for nozzle impingement half angles of 5 and 10 deg. Data point values for Fig. 15 are given in Table 2. The water flow was varied from 38 to 77 kg/s and the nitrogen from 1.3 to 0.47 kg/s to produce the variation in mass ratio from 15 to 82. The 10-deg angle clearly provides greater concentration of the liquid phase than does the 5-deg angle. For example, for the 10-deg angle, 82% of the flow is contained within a gap of 2.5 cm ($y = 1.25$ cm) while only 61% is contained within that gap for the 5-deg angle. These values can be compared to the value of 31% that would occur at that gap setting if no impingement occurred. While further

Table 2. Data point values for Fig. 15

Symbol	Mass ratio	Nozzle angle, deg
▽	82.1	10
○	56.1	10
□	25.6	10
△	15.0	10
●	56.1	5
■	25.6	5
▲	15.0	5

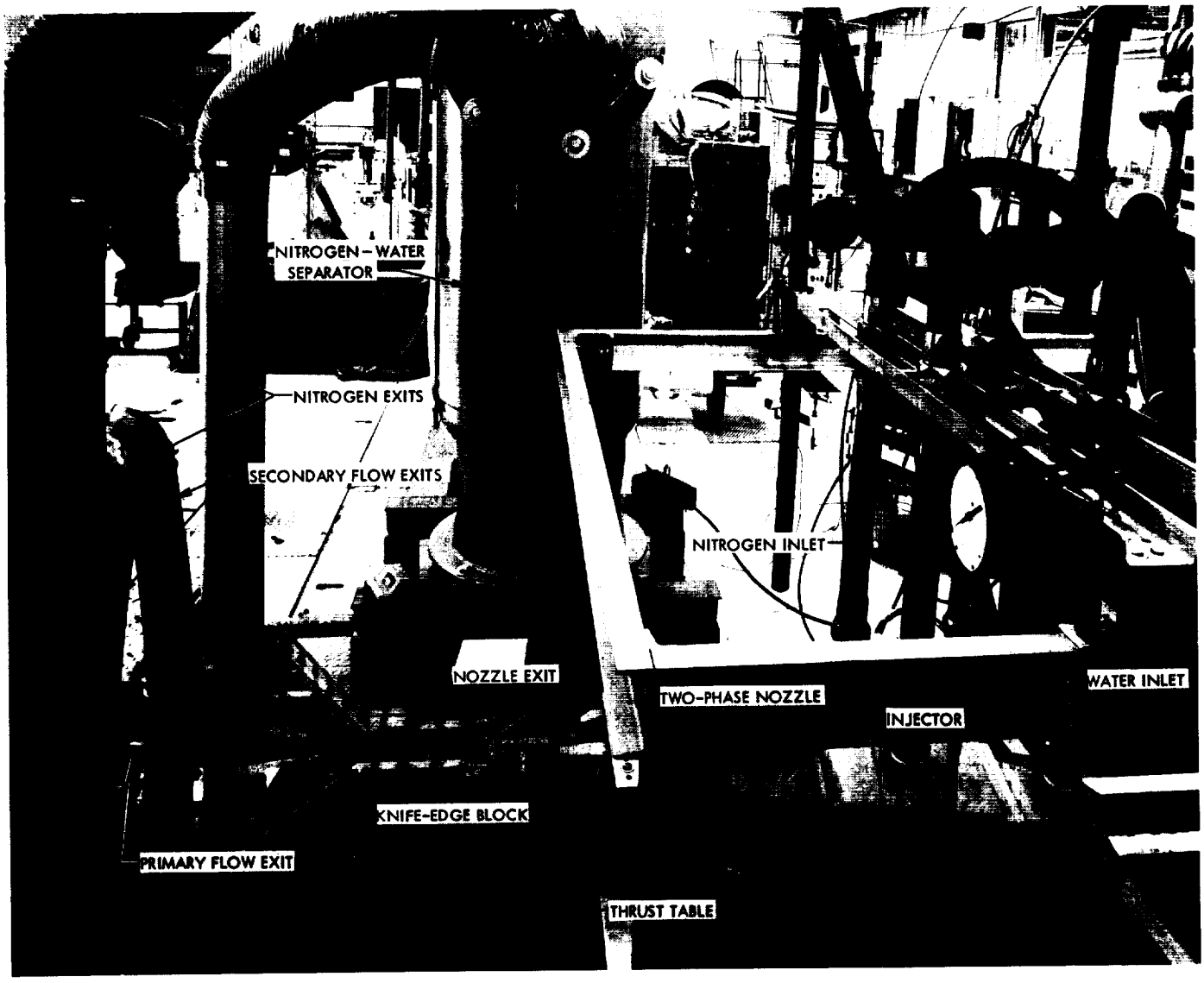


Fig. 14. Impinging nozzle test setup

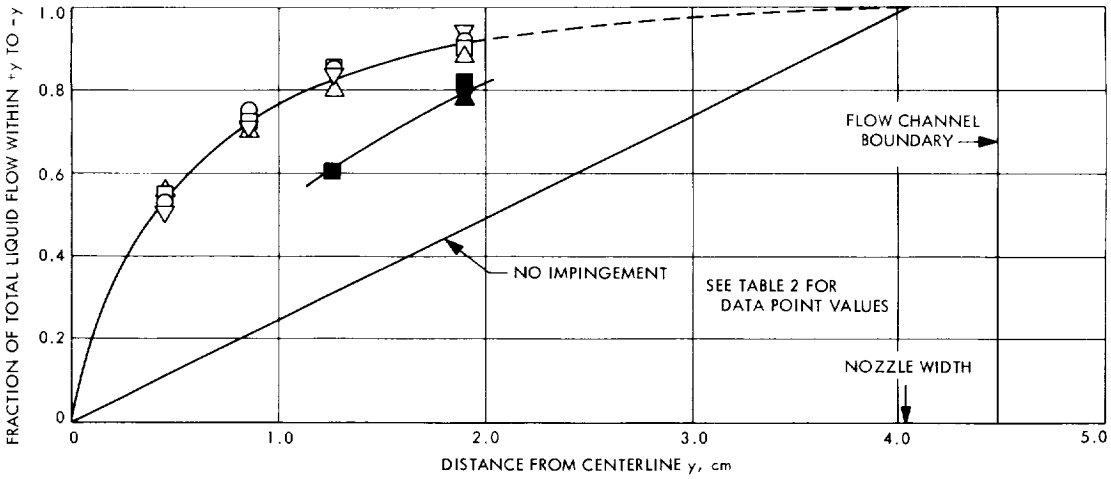


Fig. 15. Liquid flow distribution for impingment at 5 and 10-deg half angles

separation would be required to utilize these flows in an MHD generator, the degree of concentration attained is enough to enable a significant reduction in separator area and, hence, friction loss. At the lowest gap width tested (0.90 cm), the liquid concentration is five times as great as that at the nozzle exit.

The degree of concentration of the liquid phase (the volume ratio at the nozzle exits divided by the volume ratio of the resultant jet) appears to be independent of the absolute value of the volume ratio. That is, the resultant jets were geometrically identical for variations in nozzle exit volume ratio from about 7 to 38; or, in terms of Fig. 15, the fraction of the total liquid flow within the gap, $+y$ to $-y$, is nearly constant for mass ratios varying from 15 to 82.

Impingement half-angles of 15, 20 and 30 deg will be tested for the same range of mass ratios and gap settings. Previous qualitative tests at 15 and 20 deg indicate that a higher degree of liquid concentration is achieved for these angles than at the 5 and 10-deg angles.

D. Design and Evaluation of Propellant Tankage for SEPST Program, *J. R. Womack*

1. Introduction

The goal of the Solar Electric Propulsion System Technology (SEPST) Project is to develop and demonstrate the operation of a complete breadboard thrust subsystem applicable to future interplanetary spacecraft. The elements comprising this system are presented in Ref. 1. The development status of various components of this system is described in Ref. 1; SPS 37-35, Vol. IV, pp. 169–174; and SPS 37-48, Vol. III, pp. 131–134.

The propellant tankage design and the operational results of using this hardware in an experimental test of a clustered mercury–electron bombardment ion engine system are described in detail in SPS 37-35, Vol. IV. Also included are the results of tests to verify compatibility of the bladder material (neoprene) with the propellant (mercury) and pressurizer (freon 113). The results, along with those reported in SPS 37-48, Vol. III, lead to the conclusion that the bladder and tank design for the SEPST Propulsion System are satisfactory.

Subsequent testing to evaluate the bladders resulted in a number of tearing failures that indicated that the neoprene was not sufficiently strong. Consequently, a higher strength neoprene was specified and additional tests were scheduled to qualitatively evaluate its strength characteristics. This article describes the results of these tests and discusses the development status of the SEPST propellant tank design.

2. Bladder Evaluation Tests

The bladders tested were of the following composition:

(1) Neoprene GN: 100 parts.

(2) Stearic acid: 1 part.

(3) Heozone A: 4 parts.

(4) Maglite D: 4 parts.

(5) EPC: 30 parts.

(6) Zinc oxide: 5 parts.

The bladder curing requirement was specified to be 20 min at 307°F. With these specifications, it is estimated that the higher strength neoprene has a tensile strength of 4000 lb/in.^{2}, an elongation value of 7000%, a tear strength of 450 lb/in., and a hardness of 61 Shore A.

For those bladders in which failures had been observed, the material was specified only as "sulphur-free neoprene." The more detailed specification listed above was suggested by Robert F. Fedors of the JPL Polymer Research Section.

The experimental setup and execution of the tests were similar to those described in SPS 37-35, Vol. IV and SPS 37-48, Vol. III. Figure 16 shows the tank mounted on the test stand. The tank design is similar to that described in Ref. 1 except that a lucite, rather than stainless steel, hemisphere was used for the mercury expulsion side of the tank. Each of the bladders used in the tests were 0.060-in. thick and had 0.030-in. ridges (see Fig. 16) on the mercury side to avoid trapping liquid against the tank.

In three of the six runs performed, 35 psia GN_2 was used to expel the mercury. The tankage (including mercury) was at ambient temperature for two of the runs and at approximately 150° for the other. Freon 113 was used in the remaining three runs with the tankage temperature at approximately 150°F. The freon pressure was measured to be approximately 25 psia, which agreed with the vapor pressure expected at this temperature. The tank was oriented alternately in two positions to satisfactorily test the bladder: (1) with the mercury outlet 90 deg from the vertical, and (2) with the mercury outlet in the top vertical position.

After each of the six tests, the bladder was removed from the tank and carefully examined. Special attention was directed to the portions of the bladder near the location where the bladder formed the seal between the two hemispheres. This is the location at which the tear failures noted in *Subsection 1* occurred. No failures were observed from any of the runs, and it was concluded that the present bladder design is satisfactory.

Testing is continuing and, at present, the tankage is going through a long-term soak test at approximately 150°F. After several months, the bladder will be examined to determine if a long soak period in contact with both mercury and freon could have an adverse effect on its strength not indicated in the short term tests.

3. Tank Design

The final design of the propellant tank for the SEPST propulsion system, shown in Fig. 17, is a modification of the design presented in SPS 37-48, Vol. III. In the present design, the tank has a 9-in. diameter and a wall thickness of 0.020 in. The storage volume required for the freon 113 is integral with the propellant tank and separated from the propellant storage volume by an internal baffle. Nine 0.060-in.-diam holes allow freon

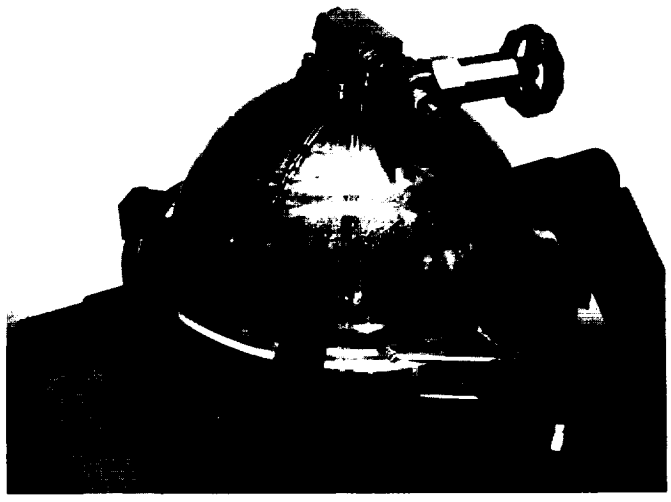


Fig. 16. Bladder evaluation test setup (post expulsion)

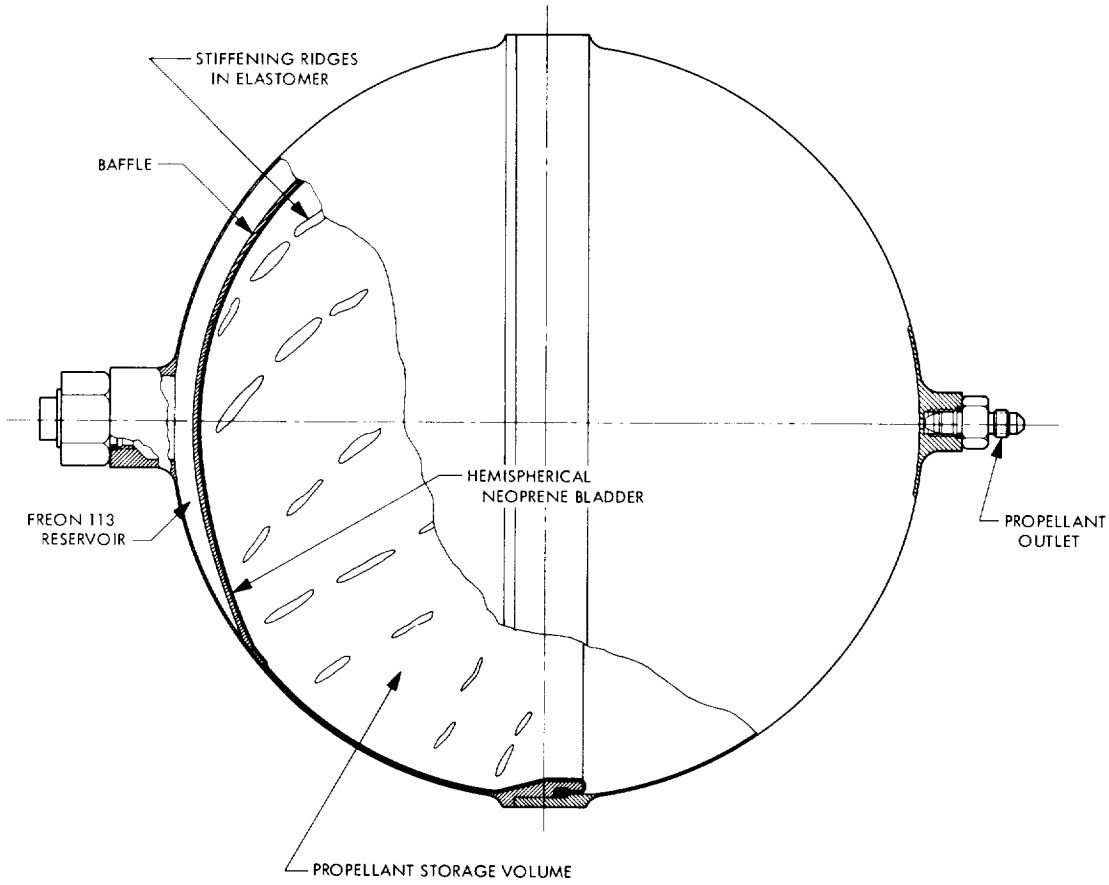


Fig. 17. SEPST program propellant tank design (propellant-loaded position)

vapor to pass through the baffle. All tank material is titanium 6A-4V.

At present, the tank design is being analyzed to determine if the design could satisfactorily meet the g forces and vibration levels imposed during launch by the *Atlas* SLV-3C/*Centaur*. Upon completion of this analysis, any appropriate design changes will be made and two tanks fabricated. It is planned to test one of these tanks (fully loaded with mercury) on the JPL shake table after which a final evaluation of tank and bladder design can be made.

Reference

1. Masek, T. D., and Womack, J. R., *Experimental Studies with a Clustered Ion Engine System*, Paper 67-698. Presented at the AIAA Electric Propulsion and Plasmadynamic Conference, Colorado Springs, Colo., Sept. 1967.

XX. Liquid Propulsion

PROPULSION DIVISION

A. Metallic Expulsion Devices, *H. B. Stanford*

1. Introduction

As spacecraft using liquid propellant rocket systems are sent on longer missions with a consequent increase in flight time, the need for propellant expulsion devices having long-term compatibility and impermeability to the propellants involved is imperative. Experience indicates that metallic devices are the most likely to satisfy these long-term space storage demands, particularly so if the added requirement of spacecraft sterilization is imposed. Two such expulsion devices that are considered to be capable of meeting these stringent requirements have been developed, and test hardware has been produced, under separate contracts.

2. Toroidal Tank–Bellows Expulsion Unit

One device capable of sterilization and extended storage with many of the practical spacecraft propellants is the toroidal tank–stainless steel bellows expulsion unit. The concept of this device derives from the combination of a toroidal tank proposed at JPL for a unitized spacecraft propulsion package and the thin-walled, formed bellows with large extension capability that evolved from industrial development. The detail design, and fabrication of flight-type test hardware to be used for evaluating the feasibility of this device, was done by the Solar Division of International Harvester Co., San Diego, California.

Two units were fabricated. For practical reasons, including the simplicity of a monopropellant system for initial investigation, these units were designed to *Mariner* Mars 1969 requirements and specifications for propellant utilization, pressure capability, etc. In line with the original idea of a unitized propulsion package, the design included the installation of *Mariner* Mars 1969 components about the toroidal opening with the rocket engine mounted in the toroid center (Fig. 1).

The three basic components of the propellant expulsion system are the bellows, the pressurization gas container, and the outer container (Fig. 2). The propellant, hydrazine, is stored inside the bellows at nominal pressure with the bellows extended. The pressurization gas, nitrogen, is stored inside the toroidal gas container at 3000 psia. Both the bellows and gas toroid are contained within an outer tank shell. This outer container serves as a structural support for the remaining propulsion system components, as a pressure container during propellant

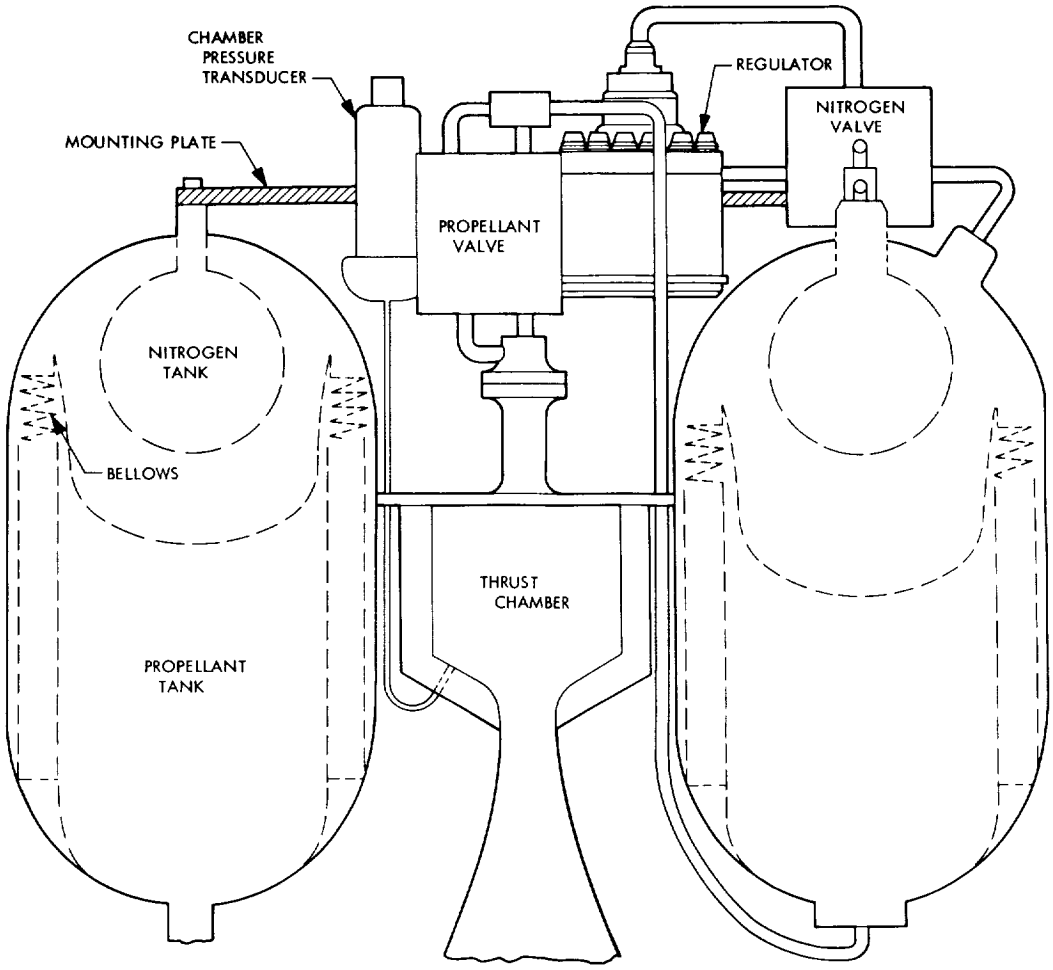


Fig. 1. Cross section of propulsion package with toroidal tank and welded bellows assembly

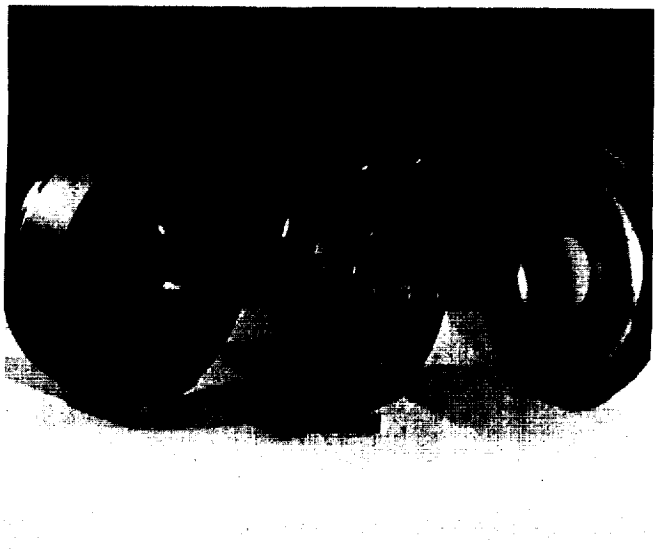


Fig. 2. Toroidal tank–bellows component parts

expulsion, and can become an integral part of the spacecraft structure as a weight-saving technique. The propellant is expelled by directing the pressurization gas out of the gas toroid, through the control valves and regulator, and then to the outer container at operating pressure. The resulting force from this gas pressure surrounds and compresses the bellows, which in turn pressurizes the propellant within the bellows, causing it to flow out through the expulsion fitting. Controls on both the pressurization and expulsion piping will regulate the rate and duration of propellant flow.

Because the end application for this hardware is to determine feasibility for space flight missions, minimum weight was a primary consideration for material selection. Titanium was used for the pressurization gas container and the outer assembly container. The material for the bellows convolutions was type 321 stainless steel since this was the material that Solar had previously

used to successfully develop their near zero-radius formed bellows.

Material for the bellows end domes was initially type AM 350 stainless steel for minimum weight, ease of welding to the bellows convolutions, and for propellant compatibility. Because of forming problems encountered during fabrication, the AM 350 stainless steel was replaced by type 321 stainless steel.[1]

The total weight of the final assembly is 26.2 lb. If the bellows end domes had been fabricated from type AM 350 stainless steel instead of type 321 stainless steel, the final assembly weight would have been 19.0 lb.

The assembly (Fig. 3) is a compact, short, cylindrical package that measures 16 in. in diameter and 12 in. in length. With the *Mariner* Mars 1969 components attached, the overall length is increased to approximately 16 in. (Fig. 4).

To ensure acceptability, the assembly was leak-tested and subjected to 10 expulsion cycles before delivery to JPL. Future testing at JPL will include slosh and expulsion testing, expulsion and volumetric efficiency measurement, eventual long-term storage with propellant, and hot-firing sequences after extended storage time.

3. Reinforced Stainless Steel Expulsion Diaphragm

The wire-reinforced stainless steel expulsion diaphragm is another device having excellent long-term storability characteristics with most spacecraft propellants. It is generally compatible and impermeable to the practical propellants and, in most cases, is capable of being sterilized. Although cycle life is limited, these devices are generally considered reliable for more than one complete cycle of operation. Three to four complete cycles before failure have been achieved in test operations.

This device has been in development for several years at Arde, Inc., Mahwah, New Jersey, under the sponsorship of various aerospace activities. JPL initially sponsored work on the development of an 18-in.-diameter wire-reinforced diaphragm in 1965. Recently, under JPL contract, Arde, Inc. has done development work to incorporate the use of gold braze for the attachment of the reinforcing wires to the reversing diaphragm for the

[1]Once the problem of forming the AM 350 stainless steel end domes was recognized, the additional effort required to develop the technique of forming them was considered to be beyond the immediate scope of this contract.

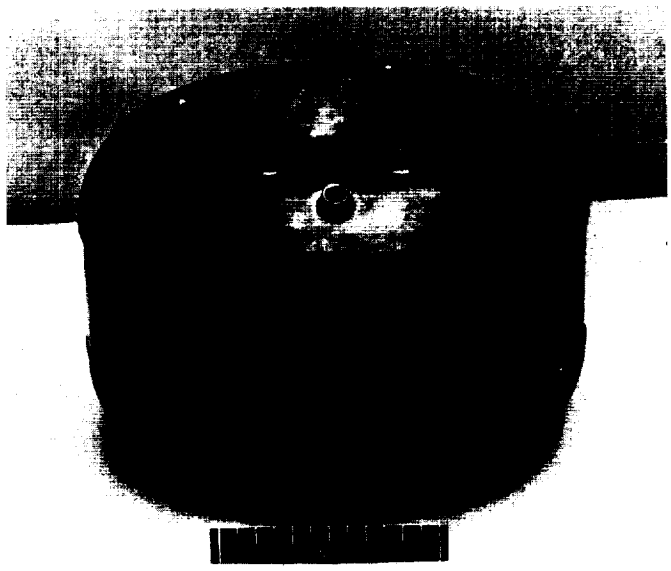


Fig. 3. Toroidal tank–bellows finished assembly

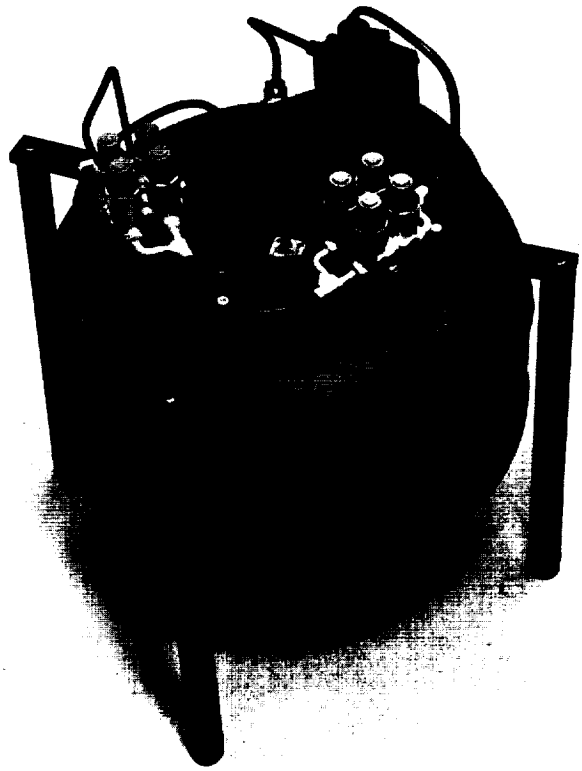


Fig. 4. *Mariner* Mars 1969 components mounted on toroidal tank assembly

purpose of making a more universally compatible assembly than was possible with copper alloy. Three gold braze alloys were investigated:

(1) Engaloy 255 (82 Au 18 Ni).

(2) Engaloy 238 (80 Au 20 Cu).

(3) Nicoro 80 (81.5 Au 16.5 Cu 20 Ni).

Engaloy 255 was selected as having the most acceptable characteristics from the standpoint of compatibility to hydrazine and nitrogen tetroxide, fillet formation and flow characteristics, strength and ductility of the braze joint, and low diffusion into the parent material (Ref. 1).

The second part of this contract was to incorporate the wire-reinforced stainless steel diaphragm into an Ardeform[2] stainless steel tank of flight weight and configuration. Original design requirements specified an 18-in. spherical tank with bosses at either pole for inlet and outlet fittings.

Proof pressure was set at 800 psig, although this was reduced to 500 psig by mutual agreement because of the added expense of making cryogenically forged girth rings required for the higher pressure.

The weight goal for the tank alone, not including the diaphragm, bosses, and fittings, was 8 lb, which was intended to be competitive with a titanium tank of comparable size and capability. The assembly as designed by Arde included a hemispherical diaphragm (Fig. 5) with a toroconical section from the girth attachment point about 30 deg upward. This design slightly reduces volumetric efficiency, but tends to improve cycle life because of reduced interference of the reinforcing wires during cycling and better hinging of the diaphragm. The tank girth weld joint, the hinge area for the diaphragm, was offset by 0.070 in. to allow for the 0.125-in.-diameter reinforcing wires attached to one side only of the diaphragm. This offset allows the diaphragm to "bottom" at the end of an expulsion cycle, thus increasing expulsion efficiency without permanently deforming the diaphragm.

The tank wall thickness is nominally 0.020 in., which is determined by the practicality of fabrication rather than the strength of the material. A heat-affected zone at the girth weld is thickened to a maximum 0.115 in. over a distance of approximately 2 in. on either side of the weld (Fig. 5). This girth ring is part of the preform and is cryogenically formed with the rest of the tank. To complete the assembly, the cryoformed tank must be bisected in the girth ring then rewelded with the diaphragm in place. The heat of welding locally anneals the girth ring, necessitating the extra thickness of material in this area.

A total of seven 18-in.-diameter wire-reinforced diaphragms were made. The first one, with 0.125-in. reinforcing wires on the conical section and 0.090-in. wires on the spherical portion, weighed 4.3 lb, and in test achieved three reversals before failure. It was determined that the failure was due to cocking and tipping of the spherical section of the diaphragm which the 0.090-in. wires were unable to control. The next diaphragm was reinforced entirely with 0.125-in. wires, weighed 5.2 lb, and achieved seven complete reversals before a leak occurred. Since this was more than the five reversals specified as a goal, the design change was accepted although it added 0.9 lb in weight.

A third diaphragm was welded into an 18-in-diameter Ardeform tank for verification testing. The tank was proofed at 500 psig with no yielding of the tank material or evident leakage. The tank was later burst at 920 psi. Since the tank had a minimum wall thickness of 0.018 in., it exhibited an equivalent uniaxial ultimate tensile strength of 225,000 psi. The failure was circumferential and outside of the bladder-to-tank-girth joint weld area, proving this joint not to be the limiting factor in the tank design.

Two diaphragms reinforced with 0.125-in. wire were installed in 18-in.-diameter Ardeform tanks. These tanks were proof-tested to 500 psig, leak-tested, and, together with two additional diaphragms having flanges intact, were delivered to JPL for further testing. The design weight goal of 11–12 lb for the tank diaphragm assembly was exceeded by about 25%. It is felt, however, that by the use of lighter bosses and girth rings, optimized tank wall thickness, and hollow reinforcing wires on the reversing diaphragm that a total weight of 12 lb for an 18-in.-diameter tank, diaphragm assembly of equal capability is feasible.

[2]Ardeform is a trade name for a cryogenic stretch forming process used to obtain super-high tensile strength from some steel alloys, particularly 301 stainless steel. The process is considered proprietary by Arde, Inc.

Reference

1. Gleich, D., *Interim and Final Report: Development of Gold Brazing Technique and Design and Supply of 18" Diameter Positive Expulsion Tank Assembly,* Report 56001-2. Arde, Inc., Mahwah, N. J., Feb. 1969.

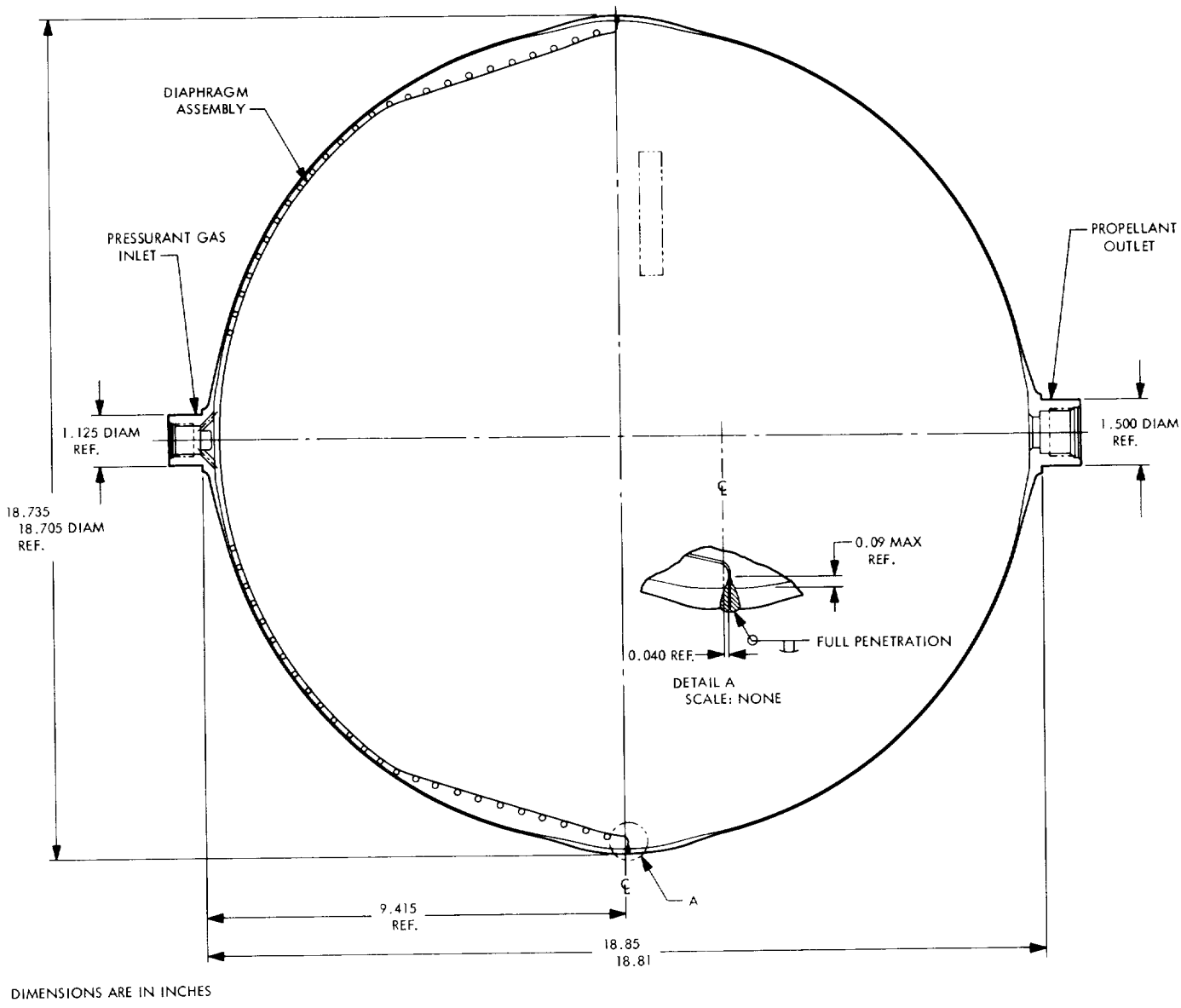


Fig. 5. Wire-reinforced stainless steel reversing diaphragm and stainless steel tank assembly

B. Resonant Combustion—Location of the Initial Disturbance in Spontaneously Resonant Rocket Engines, *R. Kushida*

1. Introduction

The 18-in.-diameter resonant combustion research rocket engine, using an injector designated as RC-1 (SPS 37-30, Vol. IV, pp. 123–130; SPS 37-45, Vol. IV, pp. 184–192), frequently exhibits a spontaneous transition from smooth running to a high-amplitude oscillation when not equipped with baffles. The growth of a wave from its first appearance to a very high amplitude wave (typically 200 to 1000 psi) takes only the time for the wave to travel once or twice across the diameter of the chamber.

The RC-1 injector has its unlike impinging doublet elements arrayed in a circularly symmetric pattern so that mixture ratio and mass flux are approximately uniform over the injector face. Figure 6 represents the composite flux distribution, as determined in cold flow for an individual element and superimposed to create the overall pattern. The outer row of elements, denoted as the boundary flow, are separately manifolded. Normally, 10% of the total propellant flow is directed through this portion of the injector. The N_2O_4–50/50 (N_2H_4/UDMH) propellant combination is used.[3]

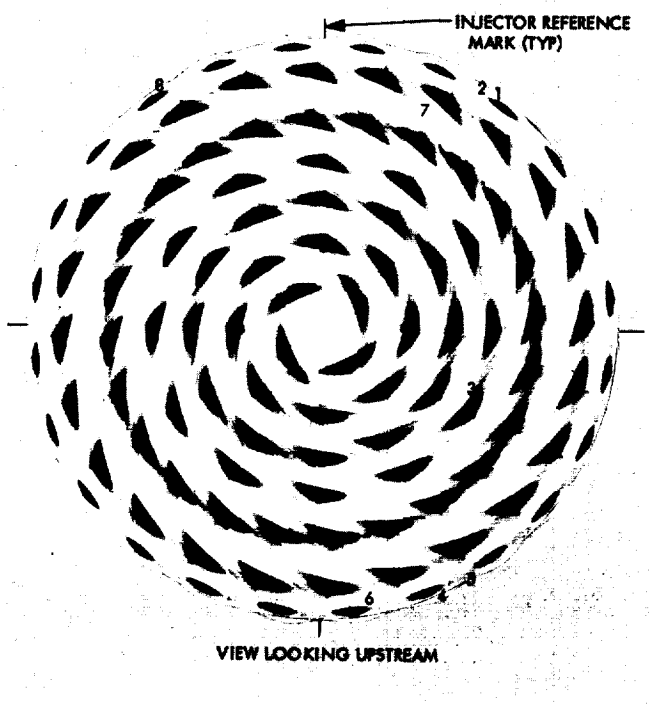


Fig. 6. Display of injector spray patterns for RC-1 injector face (numbered spots, identified in Table 2, are locations of initial disturbances)

The occurrence of random combustion roughness or popping, which is exhibited by the engine when equipped with baffles of adequate length to prevent transitions to resonance, recently has been found to correlate with certain combinations of flows and propellant temperatures in the boundary injection system (SPS 37-56, Vol. III, pp. 204–212). This correlation was achieved as a sequel to the results reported in SPS 37-45, Vol. IV and it establishes the fact that pops are produced when the dynamic pressure ratio of the boundary element streams is near unity with the boundary injectant temperatures above 60°F.

When baffles are not utilized, the engine nearly always is spontaneously resonant when operated under boundary flow conditions producing pops. The transition to resonance normally occurs after periods of varying length of smooth operation. The precipitating disturbance is usually the first pop of the engine firing.

It would be desirable to verify that the spatial origin of the pops was located in the boundary flow region of the chamber volume. To this end, a method of inferring the location of the origin from an array of simultaneously recorded pressure measurements was devised. This article presents that method, along with several results obtained with it.

2. Method

High-response ($\sim$ 100-kHz response) pressure transducers are located at known positions on the walls of the rocket chamber (SPS 37-49, Vol. III, pp. 223–236). These are used to determine the time of arrival of pressure waves. The differences in the time of arrival at the various locations are used to infer the origin of the disturbance and the velocity of propagation of the wave. To simplify calculations, we assume that the disturbance originates at a point in space and time, and that the pressure wave from that disturbance expands spherically at a constant velocity.

The time of origin of the disturbance is denoted by t_0, the position by x_0, y_0, z_0, and the velocity by c. The position of the ith transducer is given by x_i, y_i, z_i and

[3]UDMH = unsymmetric dimethylhydrazine.

the time of arrival of the pressure wave by t_i, where the subscript i identifies the transducer. We can write an expression equating the distance between the source and a transducer S_i to the distance traveled by the pressure pulse $c^*(t_i - t_0)$ in the time between initiation and detection. One such equation can be written for each transducer. In the 18-in.-diameter engine, there are generally nine transducers operating simultaneously; hence, there are nine equations. There will be more equations than unknowns (c, t_0, x_0, y_0, z_0), so we shall try to obtain the best fit values for the unknown quantities. We write the equation for the variance ϵ of the position as

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}[S_i - c(t_i - t_0)]^2 \tag{1}$$

where the distance S_i is given by

$$S_i = [(x_i - x_0)^2 + (y_i - y_0)^2 + (z_i - z_0)^2]^{1/2} \tag{2}$$

The values of c, t_0, x_0, y_0, and z_0 which minimize the ϵ are the best fit.

Trial values for the position of the disturbance (x_0, y_0, z_0) are assumed. By differentiating ϵ with respect to c and t_0, while holding position constant, and setting the result equal to zero (i.e., the necessary condition for a minimum), we can derive the following expressions:

$$c = \frac{\overline{St} - \bar{S}\,\bar{t}}{\overline{t^2} - \bar{t}^2} \tag{3}$$

and

$$t_0 = \bar{t} - \frac{\bar{S}}{c} \tag{4}$$

where we have defined the mean values

$$\overline{St} \equiv \sum_{i=1}^{N}\frac{S_i t_i}{N}, \qquad \bar{S} \equiv \sum_{i=1}^{N}\frac{S_i}{N}$$

$$\bar{t} \equiv \sum_{i=1}^{N}\frac{t_i}{N}, \qquad \overline{t^2} \equiv \sum_{i=1}^{N}\frac{t_i^2}{N}$$

and N is set equal to the number of data points.

Note that the times t_0 and t_i are measured from a common, although arbitrary, zero in time. Using Eqs. (2), (3), (4), and then (1), the variance ϵ can be calculated. The assumed location of the origin is shifted systematically through the volume of the combustion chamber until the position of minimum ϵ is found. This position is taken as the origin of the disturbance.

3. Results

The explosion of a bomb located at a known initial position in the rocket engine was used to assess the accuracy of the analysis. The results (Table 1) in the rows labeled "Analysis" are obtained by the method discussed above, while the values for the "Actual" row are those of the physical location of the center of the bomb. The angular location is from a fixed fiducial mark on the injector plate. The axial location is the distance from the flat injector face. The full chamber length is 16.06 in., and the radius from the axis of the chamber to the wall is 9.03 in. (SPS 37-30, Vol. IV, pp. 123–130). There are thirteen locations in the wall at which pressure transducers could be mounted; nine of the thirteen could be used for recording at any one time.

The bomb is a Micarta tube ½ in. in diameter which protrudes inward to a distance of 2 in. from the chamber wall. It contains 13.5 grains of high explosive.

The error in locating the bomb using pressure data and the analysis was the smallest for the angular position, relatively close for the radial position, and poor for the axial location. The angular position of the bomb

Table 1. Bomb location by pressure wave analysis

Run	Method	Location			$\sigma = (\epsilon)^{1/2}$, in.	c, ft/s
		Radial, in.	Angle, deg	Axial, in.		
B1089	Analysis	7.0	50	6.7	0.18	2440
	Actual	8.1[a]	53	4.0		
B1090	Analysis	6.9	55	6.8	0.16	2110
	Actual	8.1[a]	53	4.0		
B1091	Analysis	9.0	55	9.3	0.28	2490
	Actual	8.1[a]	53	4.0		
B1093	Analysis	1.0	260	10.0	0.37	1720
	Actual	0.0[a]	—	0.9		
B1102	Analysis	7.2	53	4.5	0.32	2940
	Actual	8.1[a]	53	4.0	—	
					Av 0.26	

[a]Geometric center of bomb.

was within 3 deg, which is comparable to the 3.5-deg displacement required to encompass the bomb jacket diameter. The radial position of the bomb was found with rather larger scatter. The computed axial location of the bomb was somewhat further downstream from the injector face than the actual bomb in all the test examples of Table 1. It is suspected that the assumptions of spherically symmetric wave propagation and a constant propagation velocity are in sufficient error to cause this discrepancy.

The magnitude of the differences in the case of the angular and radial location of a bomb are small enough to promise reasonable precision in the location of the unknown resonance-initiating disturbance. As for the axial location, all that can be concluded is that the disturbance will probably be located closer to the injector face than the computed location.

The results for several spontaneously resonant runs are given in Table 2. The angular and radial positions which were used to locate the disturbances on a view of the RC-1 injector are shown in Fig. 6.

Identifying the first detectable wave on the pressure records was not always as definite as it was for a bomb explosion. However, where the first wave was not too clear, the second or third cycle of the wave was nearly always of high amplitude. Hence, by tracing backwards in time, the presence of the initial wave could be distinguished from that of a different random pressure excursion.

Table 2. Location of initial disturbance in spontaneous instability

Case	Run	Location			$\sigma = (\epsilon)^{1/2}$, in.	c, ft/s
		Radial, in.	Angle, deg	Axis, in.		
1	B1089	9.0	40	4.3	0.08	3200
2	B1090	9.0	38	9.1	0.33	1940
3	B1093	5.0	115	14.5	0.44	2340
4	B1095	9.0	157	9.5	0.44	516
5	B1096	9.0	150	11.0	0.34	2240
6	B1097	8.5	169	7.5	0.03	1480
7	B1099	7.5	22	11.0	0.32	1580
8	B1101	9.0	327	3.0	0.23	1070
—	B1103	No definite location			—	—
					Av 0.28	

The measure of the accuracy of the assumption of a localized source for the initial disturbance can be assessed crudely by comparing the standard deviation [$\sigma = (\epsilon)^{1/2}$] for the location of the bomb ($\sigma = 0.26''$) with σ for the location of the initial disturbances ($\sigma = 0.28''$). The comparative closeness of these figures indicates that our hypothesis that the first pressure wave is a result of a point disturbance is justified.

The best-fit values of the propogation velocity c are included in Tables 1 and 2. Since pressure waves travel at sonic or slightly supersonic speeds, the computed velocities can be taken as measures of the upper bounds of the sound speed in the reacting mixture. In comparison, the equilibrium sound speed at the core flow mixture ratio of 1.90 is 3810 ft/s, and at the boundary flow mixture ratio of 1.30, it is 3990 ft/s. The significantly lower speeds computed for the best-fit values are attributed to the effect of spray droplets in the gas, incomplete chemical reaction, and gross density gradients existing in the mixture. The scatter in the values is believed to be random so that not much significance is attached to variations observed from run to run.

In one run, B1103, the resonance developed as a gradually increasing amplitude pressure oscillation which had no defined first wave on any of the transducers. This type of transition was definitely different from those noted for the other runs, in that the rate of growth was much slower. No localized origin could be calculated in this case.

4. Discussion

Of the eight runs for which definite origins could be computed, six origins were located in the boundary flow region. The computed axial locations are not given much credence, as discussed in the bomb location comparison, but the radial and angular positions are believed to be capable of centering on individual injector elements. Disturbances for cases 1 and 2 appear to have originated from the same element, while a different (but common) element may have been responsible for the pops in cases 4 and 5 and possibly 6. We conclude, however, from the distribution of origins displayed in Fig. 6 that no one single element is responsible for all *initial* disturbances. It is clear that the pop disturbances generally do originate from the region of the chamber boundary.

The slow rate of growth of the oscillation in run B1103 compared to the explosively rapid development of a pressure wave in all other cases of spontaneous transition to instability caused us to look for the possible source of the difference. The major difference was that there was no boundary flow during run B1103. The absence of a sharp initial disturbance in the absence of boundary flow confirms the observations in SPS 37-56, Vol. III that roughness and popping is correlated with certain flow conditions in the boundary flow.

XXI. System Analysis Research

MISSION ANALYSIS DIVISION

A. *Apollo 10* Range Data Provide Additional Support For Lunar Ephemeris LE 16, *J. D. Mulholland*

JPL Lunar Ephemeris LE 16 (SPS 37-57, Vol. II, pp. 51–53) is a long-span tabulation of the lunar motion, produced by a combination of numerical integration and analytic techniques. It may be regarded as a gravitationally consistent ephemeris which approximates a Van Flandern theory and is referred to the FK4 coordinate frame. As yet, there are only fragmentary results on the performance of this ephemeris in the analysis of real observations. Those that have been reported are generally favorable, but not unambiguously. Further evidence is required to establish a confidence level.

Another fragment is now provided by the preliminary tracking analysis from *Apollo 10* (Ref. 1). The ephemeris used in this analysis was DE 19/LE 4 (Ref. 2), although this important datum was omitted from the report. A "range bias" was derived for each pass at each station; I assume this to be the mean value about which the range residuals were distributed. These range biases exhibit a rather large scatter. Reference 1 states them to be "generally less than 50 meters," but 10 of the 25 points taken ½ day or more distant from lunar orbit exceed 50 m, one very early in the mission being 291 m. Nonetheless, a distinct difference may be noted between the biases corresponding to near-lunar operations and those corresponding to the other parts of the mission. In particular, the former group scatter about −220 m, while the latter scatter (roughly) about zero. The text of Ref. 1 identifies this as an artifact of the lunar ephemeris used in the mission, and Fig. 1 seems to verify this belief.

The figure shows the plotted range biases superimposed on a graph of the difference LE 16 − LE 4. The interpretation of this figure is as follows: During the intervals up to ½ day before entering lunar orbit and beyond ½ day after leaving lunar orbit, the vehicle is sufficiently far from the moon to be essentially unaffected by small ephemeris errors. Thus, one should expect the data up to about May 21.5 and after about May 24.5 to scatter about zero unless other systematic biases are present. That is roughly what is observed. During lunar orbit operations, the

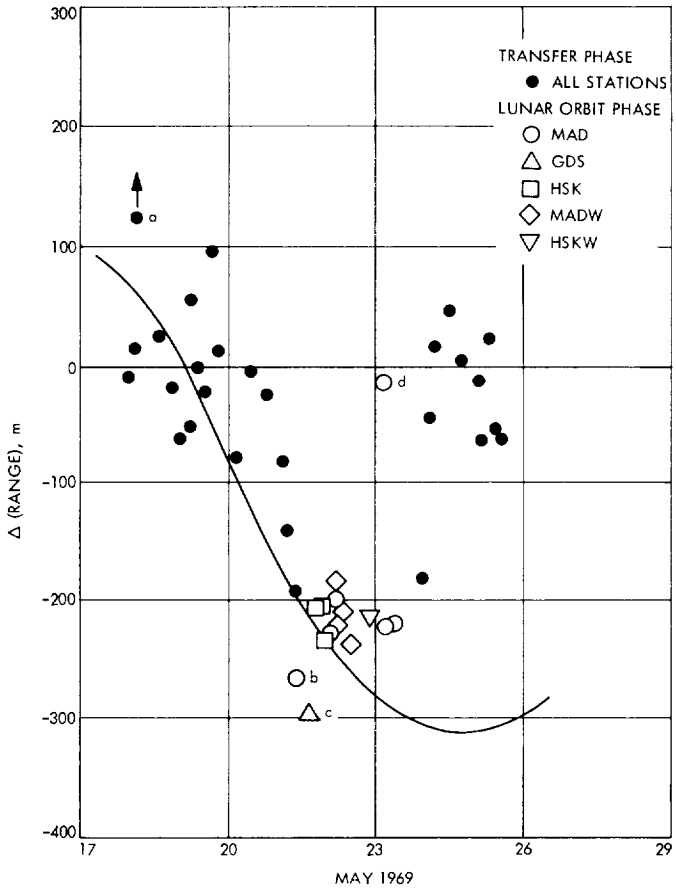


Fig. 1. Range biases for the *Apollo 10* command module tracking (symbols during lunar orbit phase refer to individual tracking stations in the *Apollo* network)

vehicle should reflect the true position of the moon, rather than the ephemeris position. Finally, there should be two transition regions, ½ day on either side of the orbital phase, when the biases exhibit a more-or-less smooth transition between the three major groups. If one postulates that LE 16 represents the true position of the moon better than does LE 4, then exactly this behavior can be seen in Fig. 1. Under this interpretation, the apparent clustering of points about the curve prior to May 21.5 is quite coincidental and meaningless. I believe this interpretation to be correct, and that these data support the belief that LE 16 is more accurate than LE 4. Certainly this was the case on May 22, 1969.

Because of the nature of the comparison, it is not possible to say with any confidence why the lunar orbit biases do not scatter randomly about the curve. It may or may not be indicative of the error in LE 16 at that point in time. Nonetheless, their deviation from the curve, though not random, is not worse than the deviation of the other points from zero. The points labelled a and d are obviously blunder points. Reference 1 identifies point c as a blunder point also, but it deviates from the curve by the same amount as point b. If the interpretation given above is correct, point c is no more suspect than several of the points that are regarded as valid, but noisy.

If LE 16 is a reasonable representation of the lunar motion, one may expect range biases of −300 to −380 m during the orbital phase of *Apollo 11*.

No significant part of the observed range rate biases or residual scatter can be attributed to the ephemeris. The trans-lunar biases were on the order of 3–4 mm/s, with 3σ noise "consistently less than 10 mm/s." During lunar orbit operations, the amplitude of the range rate residuals was 40 mm/s for the command module, while excursions to 900 mm/s were noted for the lunar excursion module. During this period, the LE 16 − LE 4 range rate differences were −0.3 to −0.8 mm/s.

References

1. Thomas, D. A., et al., "*Apollo 10*/AS-505 MSFN Metric Tracking Performance (Preliminary)," NASA X-832-69-224. National Aeronautics and Space Administration, Washington, D.C., May 1969.
2. Devine, C. J., "JPL Development Ephemeris Number 19," Technical Report 32-1181. Jet Propulsion Laboratory, Pasadena, Calif., Nov. 15, 1967.

B. Perturbations in Geometrical Optics, *H. Lass*

The differential equations for the deflection of light, in a field whose index of refraction is approximately spherically symmetric, are derived from Fermat's principle. The choice of the radial distance as the independent variable complicates the perturbed equations of motion for the light rays. By an appropriate choice of the independent variable (regularization), these equations can be simplified considerably.

Let

$$ds = (g_{\alpha\beta}(x)\, dx^{\alpha}\, dx^{\beta})^{1/2}$$

represent arc length for a coordinate system (x^1, x^2, x^3) of a Euclidean space, and let $n(x^1, x^2, x^3)$ be the index of

refraction. Fermat's principle states that the light ray will travel along a path such that

$$\int n\,ds = \int n(x)\,(g_{\alpha\beta}\,\dot{x}^\alpha\,\dot{x}^\beta)^{1/2}\,ds \tag{1}$$

is an extremal, with $\dot{x}^\alpha = dx^\alpha/ds$.

The Lagrangian is $L = nB$, with $B = (g_{\alpha\beta}\,\dot{x}^\alpha\,\dot{x}^\beta)^{1/2}$, and the Euler–Lagrange equations are

$$\frac{d}{ds}\left(\frac{\partial L}{\partial \dot{x}^i}\right) = \frac{\partial L}{\partial x^i} \tag{2}$$

A simpler form for the Lagrangian can be obtained by the following analysis. From

$$\frac{\partial L}{\partial \dot{x}^i} = n\frac{\partial B}{\partial \dot{x}^i}, \qquad \frac{\partial L}{\partial x^i} = n\frac{\partial B}{\partial x^i} + \frac{\partial n}{\partial x^i}B$$

we obtain

$$\frac{d}{ds}\left(n\frac{\partial B}{\partial \dot{x}^i}\right) = n\frac{\partial B}{\partial x^i} + \frac{\partial n}{\partial x^i}B \tag{3}$$

Multiplying Eq. (3) by B, and using the fact that along any path $B \equiv 1$ ($dB/ds = 0$), yields

$$\begin{aligned}\frac{d}{ds}\left(\frac{n}{2}\frac{\partial B^2}{\partial \dot{x}^i}\right) &= n\frac{\partial}{\partial x^i}\left(\frac{B^2}{2}\right) + \frac{\partial n}{\partial x^i}\\ \frac{d}{ds}\left[n\frac{\partial}{\partial \dot{x}^i}(B^2+1)\right] &= \frac{\partial}{\partial x^i}[n(B^2+1)]\\ &\quad - (B^2+1)\frac{\partial n}{\partial x^i} + 2\frac{\partial n}{\partial x^i}\\ &= \frac{\partial}{\partial x^i}[n(B^2+1)]\end{aligned} \tag{4}$$

Thus, we may use

$$\bar{L} = \frac{n(x)}{2}(B^2+1)$$

as our new Lagrangian. The special case of Cartesian coordinates yields

$$\bar{L} = \frac{1}{2}n(x,y,z)\,(\dot{x}^2+\dot{y}^2+\dot{z}^2+1)$$

$$\frac{\partial \bar{L}}{\partial \dot{x}} = n\dot{x}, \qquad \frac{\partial L}{\partial x} = \frac{\partial n}{\partial x}$$

since $\dot{x}^2+\dot{y}^2+\dot{z}^2 \equiv 1$. Thus,

$$\left.\begin{aligned}\dot{n}\dot{x} + \frac{\partial n}{\partial x}\dot{x}^2 + \frac{\partial n}{\partial y}\dot{x}\dot{y} + \frac{\partial n}{\partial z}\dot{x}\dot{z} &= \frac{\partial n}{\partial x}\\ \ddot{x} = \frac{1}{n}\frac{\partial n}{\partial x} - \frac{\dot{x}}{n}\left(\frac{\partial n}{\partial x}\dot{x} + \frac{\partial n}{\partial y}\dot{y} + \frac{\partial n}{\partial z}\dot{z}\right)\end{aligned}\right\} \tag{5}$$

with similar equations for $\ddot{y}$, $\ddot{z}$.

The generalization of Eq. (5) for any coordinate system is

$$\frac{d^2x^i}{ds^2} + \Gamma^i_{\alpha\beta}(x)\frac{dx^\alpha}{ds}\frac{dx^\beta}{ds} = \frac{1}{n}\frac{\partial n}{\partial x^\alpha}\left(g^{i\alpha} - \frac{dx^i}{ds}\frac{dx^\alpha}{ds}\right) \tag{6}$$

for $i = 1, 2, 3$, with $ds^2 = g_{\alpha\beta}(x)\,dx^\alpha\,dx^\beta$, and $\{\Gamma^i_{\alpha\beta}(x)\}$ the Christoffel symbols.

The special case $n \equiv$ constant yields $\ddot{x}^i + \Gamma^i_{\alpha\beta}\dot{x}^\alpha\dot{x}^\beta = 0$, the differential equations for the geodesics (linear paths for a Euclidean space).

Now Eq. (6) can be simplified by means of the regularization $ds = n\,d\tau$, so that

$$\left.\begin{aligned}\frac{dx^i}{ds} &= \frac{1}{n}\frac{dx^i}{d\tau}\\ \frac{d^2x^i}{ds^2} &= \frac{1}{n^2}\frac{d^2x^i}{d\tau^2} - \frac{1}{n^3}\frac{dx^i}{d\tau}\frac{\partial n}{\partial x^\alpha}\frac{dx^\alpha}{d\tau}\end{aligned}\right\} \tag{7}$$

Equation (6) becomes

$$\frac{d^2x^i}{d\tau^2} + \Gamma^i_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = g^{i\alpha}\frac{\partial}{\partial x^\alpha}\left(\frac{n^2}{2}\right) \tag{8}$$

Equation (8) is the equation of motion of a particle of unit mass moving in a potential field $V = -n^2/2$, a well-known result. An energy integral exists in the form

$$\frac{1}{2}g_{\alpha\beta}\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} - \frac{n^2}{2} = E = \text{constant} \tag{9}$$

It is a trivial fact that $E = 0$, since $dx^\alpha/d\tau = n\,dx^\alpha/ds$, so that

$$E = \frac{n^2}{2}(g_{\alpha\beta}\,\dot{x}^\alpha\,\dot{x}^\beta - 1) \equiv 0$$

Now suppose $n(x) = n_0(x)[1 + \epsilon n_1(x)], \epsilon << 1$. We let $x^i = y^i + \epsilon z^i$, and neglect higher-order terms. Thus,

$$\left.\begin{aligned}
\frac{dx^i}{d\tau} &= \frac{dy^i}{d\tau} + \epsilon\frac{dz^i}{d\tau} \\
\frac{d^2x^i}{d\tau^2} &= \frac{d^2y^i}{d\tau^2} + \epsilon\frac{d^2z^i}{d\tau^2} \\
\Gamma^i_{\alpha\beta}(x) &= \Gamma^i_{\alpha\beta}(y + \epsilon z) = \Gamma^i_{\alpha\beta}(y) + \epsilon\frac{\partial\Gamma^i_{\alpha\beta}(y)}{\partial y^\gamma} z^\gamma \\
\frac{1}{2}n^2(x) &= \frac{1}{2}n_0^2(x) + \epsilon n_0^2(x)\, n_1(x) \\
&= \frac{1}{2}n_0^2(y) + \epsilon n_0(y)\frac{\partial n_0}{\partial y^\alpha} z^\alpha + \epsilon n_0^2(y)\, n_1(y)\frac{\partial}{\partial x^\alpha}\left[\frac{1}{2}n^2(x)\right] \\
&= \frac{\partial}{\partial y^\alpha}\left[\frac{n_0^2(y)}{2}\right] + \frac{\epsilon}{2}\frac{\partial^2 n_0^2}{\partial y^\alpha \partial y^\gamma} z^\gamma + \epsilon\frac{\partial}{\partial y^\alpha}\left[n_0^2(y)\, n_1(y)\right] \\
g^{i\alpha}(x) &= g^{i\alpha}(y) + \epsilon\frac{\partial g^{i\alpha}}{\partial y^\gamma} z^\gamma
\end{aligned}\right\} \tag{10}$$

Equation (8) yields for the unperturbed motion ($\epsilon = 0$)

$$\frac{d^2y^i}{d\tau^2} + \Gamma^i_{\alpha\beta}(y)\frac{dy^\alpha}{d\tau}\frac{dy^\beta}{d\tau} = g^{i\alpha}(y)\frac{\partial}{\partial y^\alpha}\left[\frac{n_0^2(y)}{2}\right] \tag{11}$$

with $i = 1, 2, 3$.

The perturbed equations are

$$\begin{aligned}
&\frac{d^2z^i}{d\tau^2} + 2\Gamma^i_{\alpha\beta}(y)\frac{dy^\alpha}{d\tau}\frac{dz^\beta}{d\tau} + \frac{\partial\Gamma^i_{\alpha\beta}(y)}{\partial y^\gamma}\frac{dy^\alpha}{d\tau}\frac{dy^\beta}{d\tau} z^\gamma = \\
&\qquad g^{i\alpha}(y)\left\{\frac{1}{2}\frac{\partial n_0^2(y)}{\partial y^\alpha \partial y^\gamma} z^\gamma + \frac{\partial}{\partial y^\alpha}\left[n_0^2(y)\, n_1(y)\right]\right\} \\
&\qquad + \frac{\partial g^{i\alpha}(y)}{\partial y^\gamma}\frac{\partial}{\partial y^\alpha}\left[\frac{n_0^2(y)}{2}\right] z^\gamma
\end{aligned} \tag{12}$$

so that the perturbation terms z^i satisfy linear differential equations.

We consider now a spherical coordinate system (r, θ, ϕ), with $n_0 = F(r)$, $n_1 = f(\theta, \phi)$. For our initial conditions at $s = 0$, or $\tau = 0$, we choose $\theta = \pi/2, d\theta/d\tau = 0$, by an appropriate orientation of the axes. The angle θ is the co-latitude, r the radial distance, and ϕ the azimuth.

The Christoffel symbols are

$$\left.\begin{aligned}
\Gamma^1_{22} &= -r \\
\Gamma^1_{33} &= -r\sin^2\theta \\
\Gamma^2_{12} &= \Gamma^2_{21} = \frac{1}{r} \\
\Gamma^2_{33} &= -\sin\theta\cos\theta \\
\Gamma^3_{13} &= \Gamma^3_{31} = \frac{1}{r} \\
\Gamma^3_{23} &= \Gamma^3_{32} = \cot\theta
\end{aligned}\right\} \tag{13}$$

All other $\Gamma^i_{jk} \equiv 0$, with $x^1 = r$, $x^2 = \theta$, $x^3 = \phi$.

Substituting into Eq. (11) with $y^1 = r$, $y^2 = \theta$, $y^3 = \phi$ yields

$$\left.\begin{aligned}
\frac{d^2r}{d\tau^2} - r\left(\frac{d\theta}{d\tau}\right)^2 - r\sin^2\theta\left(\frac{d\phi}{d\tau}\right)^2 &= F(r)\, F'(r) \\
\frac{d^2\theta}{d\tau^2} + \frac{2}{r}\frac{dr}{d\tau}\frac{d\theta}{d\tau} - \sin\theta\cos\theta\left(\frac{d\phi}{d\tau}\right)^2 &= 0 \\
\frac{d^2\phi}{d\tau^2} + \frac{2}{r}\frac{dr}{d\tau}\frac{d\phi}{d\tau} + 2\cot\theta\frac{d\theta}{d\tau}\frac{d\phi}{d\tau} &= 0
\end{aligned}\right\} \tag{14}$$

We note that $\theta \equiv \pi/2$ satisfies the middle equation of Eq. (14) as well as the initial conditions. Hence, the unperturbed motion is planar for the case $n_0 = n_0(r)$. The last equation of Eq. (14) yields a first integral (conservation of angular momentum), $r^2\, d\phi/d\tau = h =$ constant. An integration of Eq. (14) yields

$$\left.\begin{aligned} r &= r(\tau) \\ \phi &= \phi(\tau) \\ \theta &\equiv \frac{\pi}{2} \end{aligned}\right\} \tag{15}$$

with $r(0) = r_0$, $\phi(0) = \phi_0$, $\dot{r}(0) = \dot{r}_0$, $\dot{\phi}(0) = \dot{\phi}_0$.

The reader can verify that the special case $F(r) = (2\mu/r)^{1/2}$ yields the equation of motion of a particle under the action of a central inverse square law force.

Let us examine now the perturbed motion given by Eq. (12). Let $z^1 = R$, $z^2 = \Theta$, $z^3 = \Phi$, recalling that $\theta \equiv \pi/2$. For $i = 1$, we obtain

$$\frac{d^2R}{d\tau^2} - 2r\frac{d\phi}{d\tau}\frac{d\Phi}{d\tau} - \left(\frac{d\phi}{d\tau}\right)^2 R = \frac{1}{2}\frac{d^2F^2}{dr^2}R + 2f\left(\theta = \frac{\pi}{2}, \phi\right)FF'(r) \tag{16}$$

The case $i = 2$ yields

$$\frac{d^2\Theta}{d\tau^2} + \frac{2}{r}\frac{dr}{d\tau}\frac{d\Theta}{d\tau} + \left(\frac{d\phi}{d\tau}\right)^2 \Theta = \frac{F^2(r)}{r^2}\frac{\partial f}{\partial\theta}\left(\theta = \frac{\pi}{2}, \phi\right) \tag{17}$$

Finally, for $i = 3$, we obtain

$$\frac{d^2\Phi}{d\tau^2} + \frac{2}{r}\frac{dr}{d\tau}\frac{d\Phi}{d\tau} - \frac{2}{r^2}\frac{dr}{d\tau}\frac{d\phi}{d\tau}R + \frac{2}{r}\frac{d\phi}{d\tau}\frac{dR}{d\tau} = \frac{F^2}{r^2}\frac{\partial f}{\partial\phi}\left(\frac{\pi}{2}, \phi\right) \tag{18}$$

There is a simple check to verify Eqs. (14), (16), (17), and (18). A first integral of the motion is

$$g_{\alpha\beta}(x)\frac{dx^\alpha}{d\tau}\frac{dx^\beta}{d\tau} = n^2$$

or

$$g_{\alpha\beta}(y + \epsilon z)\left(\frac{dy^\alpha}{d\tau} + \epsilon\frac{dz^\alpha}{d\tau}\right)\left(\frac{dy^\beta}{d\tau} + \epsilon\frac{dz^\beta}{d\tau}\right) = n^2(y + \epsilon z)$$

which yield

$$\left.\begin{aligned} g_{\alpha\beta}(y)\frac{dy^\alpha}{d\tau}\frac{dy^\beta}{d\tau} &= F^2(r) \\ 2g_{\alpha\beta}(y)\frac{dy^\alpha}{d\tau}\frac{dz^\beta}{d\tau} + \frac{\partial g_{\alpha\beta}}{\partial y^\gamma}\frac{dy^\alpha}{d\tau}\frac{dy^\beta}{d\tau}z^\gamma &= 2F^2(y)f\left(\frac{\pi}{2}, \phi\right) + \frac{dF^2}{dr}R \end{aligned}\right\} \tag{19}$$

A simple differentiation with respect to τ, followed by replacing the second derivatives from Eqs. (14), (16), (17), and (18), leads to two identities.

Thus, a useful check in the numerical integration for the perturbation equations is, from Eq. (19),

$$\frac{dr}{d\tau}\frac{dR}{d\tau} + r^2\frac{d\phi}{d\tau}\frac{d\Phi}{d\tau} + r\left(\frac{d\phi}{d\tau}\right)^2 R = F^2(r)f\left(\frac{\pi}{2}, \phi\right) + F(r)F'(r)R \tag{20}$$

The final integration of the equations yields

$$\left.\begin{aligned} \bar{r}(\tau) &= r(\tau) + \epsilon R(\tau) \\ \bar{\theta}(\tau) &= \frac{\pi}{2} + \epsilon\Theta(\tau) \\ \bar{\phi}(\tau) &= \phi(\tau) + \epsilon\Phi(\tau) \end{aligned}\right\} 0 \leq \tau < \infty \tag{21}$$

with

$$s = \int_0^\tau n\, d\tau$$

the distance traveled by a photon.