89085

The Telecommunications and Data Acquisition Progress Report 42-86

April – June 1986

E. C. Posner
Editor

August 15, 1986

NASA

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Telecommunications and Data Acquisition Progress Report 42-86

April – June 1986

E. C. Posner
Editor

August 15, 1986

NASA

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Preface

This quarterly publication provides archival reports on developments in programs managed by JPL's Office of Telecommunications and Data Acquisition (TDA). In space communications, radio navigation, radio science, and ground-based radio astronomy, it reports on activities of the Deep Space Network (DSN) and its associated Ground Communications Facility (GCF) in planning, in supporting research and technology, in implementation, and in operations. Also included is TDA-funded activity at JPL on data and information systems and reimbursable DSN work performed for other space agencies through NASA. The preceding work is all performed for NASA's Office of Space Tracking and Data Systems (OSTDS).

In geodynamics, the publication reports on the application of radio interferometry at microwave frequencies for geodynamic measurements. In the search for extraterrestrial intelligence (SETI), it reports on implementation and operations for searching the microwave spectrum. The latter two programs are performed for NASA's Office of Space Science and Applications (OSSA).

Finally, tasks funded under the JPL Director's Discretionary Fund and the Caltech President's Fund which involve the TDA Office are included.

This and each succeeding issue of the TDA Progress Report will present material in some, but not necessarily all, of the following categories:

OSTDS Tasks:

    DSN Advanced Systems
        Tracking and Ground-Based Navigation
        Communications, Spacecraft-Ground
        Station Control and System Technology
        Network Data Processing and Productivity
    DSN Systems Implementation
        Capabilities for New Projects
        Networks Consolidation Program
        New Initiatives
        Network Sustaining
    DSN Operations
        Network Operations and Operations Support
        Mission Interface and Support
        TDA Program Management and Analysis
    GCF Implementation and Operations
    Data and Information Systems

OSSA Tasks:

    Search for Extraterrestrial Intelligence
    Geodynamics
        Geodetic Instrument Development
        Geodynamic Science

Discretionary Funded Tasks

Contents

OSTDS TASKS
DSN Advanced Systems
TRACKING AND GROUND-BASED NAVIGATION

COMMUNICATIONS, SPACECRAFT–GROUND

STATION CONTROL AND SYSTEM TECHNOLOGY

NETWORK DATA PROCESSING AND PRODUCTIVITY

DSN Systems Implementation
CAPABILITIES FOR NEW PROJECTS

NETWORK SUSTAINING

DSN Operations
MISSION INTERFACE AND SUPPORT

OSSA TASKS
Search for Extraterrestrial Intelligence

DSN Radar Astronomy

Geodynamics
GEODYNAMIC SCIENCE

DSS 13 Frequency Stability Tests Performed During May 1985 Through March 1986

T. Y. Otoshi and M. M. Franco
Radio Frequency and Microwave Subsystems Section

This article presents results of station frequency stability testing performed at DSS 13 during May 1985 through March 1986. The testing was done on X-band uplink and X- and S-band downlink subsystems as well as on end-to-end systems. The subsystem test data are useful for assessing the frequency stability of various prototype X-band uplink or downlink subsystems for purposes of making design improvements. Information derived from extensive testing at DSS 13 will be useful in the preparation of an X-band Uplink Demonstration Experiment to be conducted at DSS 13, and will also be valuable in the preparations of gravity wave experiments to be conducted at other DSN stations in the future.

I. Introduction

An excellent introduction to the "Gravitational Wave Experiment" was given by Berman (Ref. 1) in 1978. As stated in Ref. 1, "One of the most exciting challenges facing gravitational theoreticians and experimenters in the remaining decades of the 20th century will be the attempted detection and measurement of 'gravitational waves' as predicted by Einstein's General Theory of Relativity." One of the steps recommended by Berman for gravitational wave detection was the development of X-band uplink capability.

The DSN undertook a development program in 1979 to add an X-band uplink capability to DSN Deep Space Stations. Early X-band uplink development work is described in Ref. 2. Descriptions of the overall X-band uplink and X- and S-band downlink systems at DSS 13 and the planned demonstration at DSS 13 with the Galileo spacecraft were given by Meeker

and Timpe in Ref. 3. A theoretical phase stability analysis of this new DSS 13 system was performed by Koerner (Ref. 4).

Frequency stability tests performed at JPL on some prototype subsystems (for the X-band uplink project) were reported by Sosa (Refs. 5 and 6) in 1981. The first actual station stability testing on X-band uplink equipment at DSS 13 was done by Otoshi during March 1984 through March 1985, and the results of those tests were reported in internal JPL documents (Table 4, items 1 through 5). More recently, from May 1985 through March 1986, Otoshi and Franco have performed extensive frequency stability tests at DSS 13 (Table 4, item 6).

The purpose of this current article is to present the results of the recent frequency stability tests performed at DSS 13. It is also the purpose to impart some of the knowledge gained from this work that will be useful to designers of X-band

uplink systems and to users of radio science data. The following sections of this article are organized as follows: (1) description of the current system status at DSS 13 and discussions of problem areas to correct or avoid for future X-band uplink systems, (2) brief descriptions of the test setups and measurement techniques used to obtain the test data, (3) a summary of test results for nine test periods, and (4) concluding remarks, including a list of important knowledge gained as a result of extensive testing at DSS 13.

II. Present System Status

Many problems were encountered during the 11-month period from May 1985 through March 1986. Failures occurred on two different Dana synthesizers, the DSN prototype cable stabilizer, the X-band klystron magnet power supply, the klystron vacuum seal, two S-band translator modules, an X-band translator module, the X-band maser, the S-band maser, and the Block III receiver modules. Some of the problems and failures occurred during testing, while other failures occurred between test periods. Very often, the test plans had to be modified after arriving at the site so that improvised tests could be performed on whatever subsystems were working at the time.

At the present time, most of the hardware is now working and problems are identified. The R&D X-band uplink-downlink hardware at DSS 13 is not up to DSN quality standards, and cannot be expected to perform as reliably as the implementation versions that will be going into 34-m antenna DSN stations in the near future.

Leakage signals corrupted data taken at the Block III receiver doppler extractor ports. The leakage problem was circumvented in subsequent end-to-end tests by putting terminations on the output ports of the doppler reference signal paths at the control room bulkhead, and then operating the Block III receivers in modified open-loop configurations.

The current open- and closed-loop receiver systems at DSS 13 are very difficult to check for leakage or to perform AGC calibrations on because all controls for the step attenuators have to be manually operated by personnel inside the cone on the antenna.

Another important problem area discovered was that the Hydrogen Maser Frequency Distribution system outputs are not adequately isolated. When personnel for other projects run simultaneous tests at DSS 13 (on a noninterference basis) and begin connecting and disconnecting cables from 1 MHz and 5 MHz output ports on the Hydrogen Maser Frequency distribution boxes, the frequency stability of the end-to-end system is affected and degraded.

III. Test Setups

Figure 1 shows the output test ports used for DSS 13 X-band uplink and X-band downlink subsystem and end-to-end system tests. Figure 2 shows the output test ports for the X-band uplink and S-band downlink system. Figure 3 shows the special test equipment used for performing station frequency stability tests, and Fig. 4 is a block diagram of the Data Acquisition System (DAS) used to acquire and collect the test data.

At test output ports 4, 5, and 6, shown in Fig. 1, measurements were made by comparing the output phase of the test signal (coming out the microwave subsystem under test) to the phase of a reference signal. For X-band uplink and X-band translator output tests, respectively, the reference signals were generated by use of oven-stabilized ×72 and ×84 multipliers that were driven by the cable-stabilized 100 MHz reference frequency in the cone. For translator S-band output port 12, shown in Fig. 2, the reference signal was generated by an oven-stabilized ×23 multiplier driven by the cable-stabilized 100 MHz in the cone. The reference signal was fed into one of the ports of a test equipment X- or S-band microwave mixer, while the test signal from the output of the microwave subsystem under test was fed into the other port of the mixer.

The transmitter synthesizer in the control room was set so that the frequency of the test signal would be offset from the frequency of the reference signal by 1 Hz. Then the 1 Hz output from the test equipment X- or S-band mixer was fed into a 1 Hz amplifier-filter (with a passband of DC to 10 Hz) which was followed by the Zero Crossing Detector (ZCD) and the Data Acquisition System (see Fig. 4). The 1 Hz amplifier-filter was not used in all of the tests, but tests showed that negligible degradation of the test data resulted from use of this amplifier-filter.

Connections to mixer(s), subsystem test ports, reference multiplier, and 100 MHz cable stabilizer output port were made with special phase stable test cables. Fixed attenuator pads having the appropriate values were used to obtain desired power levels. For translator output tests, it was necessary to insert an external low-noise amplifier between the translator output port and the mixer input port.

For purposes of testing the microwave subsystems at frequencies that were not integer multiples of 100 MHz, the two-mixer method test configuration shown in Fig. 5 was used. The two-mixer method permits testing at any of the microwave frequencies in the DSN band. This technique is a new development in the specialized field of frequency stability measurements.

For tests at Multi-Mission Receiver (MMR) output ports (see port 7 in Fig. 1 and port 13 in Fig. 2), an HP 8662A synthesizer, driven by a 10 MHz reference frequency, was used as part of the external test equipment. The 10 MHz reference frequency for the synthesizer was derived from the output of a divide-by-10 assembly that was driven by the station's 100 MHz reference frequency from the hydrogen maser. The output frequency of the HP 8662A synthesizer was set to the MMR IF output plus 1 Hz and fed into one port of a test equipment IF mixer. The IF test signal from the MMR was fed into an IF amplifier whose output was fed to the other port of the mixer. The 1 Hz output from the mixer was then fed into the 1 Hz amplifier-filter, followed by the Zero Crossing Detector and Data Acquisition System.

For closed-loop receiver output tests at the X-band doppler mixer output ports (see ports 8 and 9 in Fig. 1) and S-band doppler mixer output ports (see ports 14 and 15 in Fig. 2), an HP 8662A synthesizer was set at 1 MHz + 1 Hz or 5 MHz + 1 Hz and fed into a test equipment IF mixer. The test signal output from the doppler mixer was fed into the other port of the test equipment IF mixer. The 1 Hz output was then fed into the 1 Hz amplifier-filter, followed by the Zero Crossing Detector and Data Acquisition System.

For modified open-loop receiver output tests at ports 11 and 17, it was only necessary to set the transmitter and receiver synthesizers to the appropriate values to obtain a 1 Hz output which was then fed directly into the Zero Crossing Detector, which was followed by the Data Acquisition System. No test equipment mixer or amplifier-filter were necessary in this test configuration and only a single test cable (carrying the 1 Hz signal) was required.

IV. Test Results

The results for the tests performed at DSS 13 during May 1985 through March 1986 are summarized and tabulated in Tables 1 through 3. Tables 1, 2, and 3, respectively, show the results for X-band uplink only tests, X-band uplink and X-band downlink only tests, and X-band uplink and S-band downlink only tests. The test output ports are described in the tables and refer to the test ports shown in Figs. 1 and 2. In order to keep this report concise, only frequency stability results for tau = 1000s will be given.

The following symbols used in the tables are defined as follows:

FFS   = Fractional Frequency Stability (Allan Sigma)

#SDP = Number of second difference points used to determine FFS

The heading "Xmtr Syn" used in column 2 of the tables refers to the particular synthesizer used for the transmitter (or uplink) synthesizer. The symbol

T   refers to the Dana synthesizer, J270(F)-84502, normally used at DSS 13 as the transmitter synthesizer.

R   refers to the Dana synthesizer, J270(F)-79284, normally used at DSS 13 as the receiver synthesizer.

S   refers to the Dana synthesizer, J270(F)-59853, normally used as a DSN spare synthesizer.

It is important to record which synthesizer is used for the tests because the stability of the particular synthesizer used for the transmitter or uplink synthesizer is crucial for obtaining good station stability.

In the tables, the heading "Air Temp." is used to refer to the outside air temperature as monitored by a thermometer placed on the cable tray just outside the control room interface bulkhead. It is important to monitor the outside air temperature because test results seem to improve considerably during periods when the air temperature was 12°C or colder, and also from about 10 p.m. to 3 a.m., when the outside temperature variations are minimal. Good test results were sometimes difficult to obtain during some times of the day and whenever the outside air temperature rose above 25°C.

In the Comments column in the tables, the term "edited data" is used to indicate that the original data set had bad data points in it. For the FFS results corresponding to edited data, only selected good portions of the original data were used. Although the "edited data" results tend to indicate that stabilities are better than what was actually achieved, it was necessary in some cases to edit and salvage the data. In some cases, the causes of bad data points could be isolated and attributed to a non-typical event occurring at the station during the test.

Although much more information is contained in the tables, the discussion of results will be limited to the following general points of interest and comments

(1) For exciter closed-loop tests, the FFS (for tau = 1000s) ranged from 2.27E-16 to 1.58E-15 and was typically better than 1.2E-15.

(2) For transmitter closed-loop tests, the FFS ranged from 3.46E-16 to 2.68E-15. In previous tests (Table 4, items 1 through 5) the FFS for the same configuration was typically better than 1.4E-15, but most of the time, the FFS results were only slightly worse than the exciter closed-loop results.

(3) For transmitter open-loop tests, the FFS ranged from 4.94E-16 to 1.42E-15. It should be noted that the

tests were intentionally done during a period when the air temperature change was small to minimize the klystron heat exchanger cycling on and off during the tests. Previous tests (Table 4, items 1 through 5) showed that the FFS for this configuration varied between 1.2 and 2.0E-15, depending on the air temperature variations occurring during the test.

(4) The X-band translator output tests showed that FFS ranged from 4.99E-16 to 1.56E-15. This is consistent with previously reported data (Table 4, item 5).

(5) The end-to-end tests for X-band uplink and X-band downlink with the receiver in closed loop will not be discussed because the data may be invalid due to leakage signals which were detected after the tests were completed. The effect of the leakage signals is not presently known. The data is shown for reference only.

(6) The end-to-end tests for X-band uplink and X-band downlink with the receiver in modified open loop showed that the stabilities were far better than expected. For the system with the transmitter bypassed, the FFS ranged from 6.72E-16 to 3.05E-15. Most of the time the FFS was better than 1.25E-15. For the system with the transmitter included, the FFS values ranged from 9.51E-16 to 2.09E-15.

(7) The S-band translator output tests showed that FFS ranged from 4.71E-16 to 7.07E-15. The FFS was typically about 3.0E-15 or better, which agrees with some earlier test results that were not reported.

(8) The S-band MMR output tests showed that the FFS ranged from 1.15E-15 to 5.04E-15. Some problems were encountered in these measurements due to noisy 295 MHz IF signals coming down to the control room. The IF signal had to be amplified with an external amplifier and mixed with a 295 MHz plus 1 Hz signal from an HP 8662A synthesizer.

(9) The end-to-end tests for X-band uplink and S-band downlink with the receiver in closed loop may not be valid because leakage signals might have existed at the time of these particular S-band downlink tests. The results, therefore, should be considered preliminary and are shown for reference purposes only. This data will be compared to data taken at some future time when equipment leakage problems are eliminated.

(10) The end-to-end test for X-band uplink and S-band downlink with the receiver in modified open-loop configuration showed that the stability was not good. However, only one test was made and the results may not be indicative of true system performance.

Based upon the test results obtained at various subsystem output ports, the total end-to-end system stabilities were better than expected. It should be pointed out that subsystem test results were degraded by the use of external test setup equipment such as reference path multipliers, mixers, an auxiliary synthesizer, amplifiers, and test cables. In the final end-to-end system tests for the modified open-loop configurations, the test setup did not include any of this external equipment, except for a test cable that was used to carry the 1 Hz output signal from the system to the Zero Crossing Detector.

It should be pointed out that none of the FFS test results presented in this report included the instability of the hydrogen maser. The reason for this is that, for the various test configurations and test methods employed to make the measurements, the instability of the hydrogen maser (or station frequency standard) cancels out to a first order.

V. Summary and Conclusions

In conclusion, it is appropriate to pass on knowledge gained from extensive DSS 13 frequency stability testing. Not all of this information can be determined from the test results given in this report, but most of the information can be derived from internal documents (Table 4).

(1) The stability of the Cable Stabilizer is extremely critical to obtaining good system stability results (Table 4, item 5).

(2) The best synthesizer should be used as the transmitter synthesizer. The receiver synthesizer should also have good stability, but its stability did not seem to be as critical as that of the transmitter synthesizer.

(3) The outside air temperature change is definitely a factor in station stability degradation. There is currently 1200 ft of uncompensated cable from the Dana synthesizer in the control room to the exciter subassembly in the cone. Phase changes of this uncompensated cable can be considerable during some time periods of the day when the outside air temperature is high.

(4) Poor isolation of 1 MHz and 5 MHz output ports on the frequency distribution system at DSS 13 caused degradation of test results on some of the end-to-end system tests. It is important that while stability tests are being performed at the station, other experimenters using the station on a noninterference basis do not connect and disconnect any of the cables carrying station reference frequencies into their equipment (Table 4, item 6).

(5) The test results show that the two-mixer method is a valid technique for measuring the stability of a microwave subsystem whose output frequency is not an exact integer multiple of 100 MHz. The innovative development of this two-mixer method is an important technological breakthrough for station stability measurement work.

(6) It was discovered that the 1 Hz signal at the output of the test equipment mixer can be amplified and piped down 1200 ft of cable to the control room without degrading the test data. This allows most of the test equipment (Zero Crossing Detector, oscilloscope, computer, disk drives, keyboard, and monitor) to be located in the control room and to be operated from inside the control room rather than inside the cone on the antenna (Table 4, item 6).

(7) No significant differences in results were observed when special tests were made using the DSS 13 hydrogen maser and then the cesium frequency standard as the reference frequency standard (Table 4, item 6). These experimental results gave verification to theoretical predictions that, for the particular measurement techniques employed, most of the instability of the reference frequency source cancels out.

(8) For performing station stability tests, it is important to have a data collection software program that enables the quality of every data point to be examined. If a current data point is bad, the experimenters should be immediately alerted by an audio beep note and also by a visual message on the computer monitor. Oftentimes, someone in the station inadvertently disconnects cables, or slams doors on the racks of critical equipment, or moves critical test cables inside the cone. Also, power glitches occur. With the data collection software currently being used at DSS 13, any degradation of data due to these causes could be seen immediately.

(9) For station stability testing, it is important to have a data collection software program that enables all data points to be stored and saved for postprocessing. Oftentimes, a glitch occurs in the data. If the cause of the glitch is known, and is a unique isolated phenomenon, it is valid to edit out the data point and thereby save most of the test data. It is also important to save this data to be used later to plot stability versus time for diagnostic purposes.

This article has presented results of tests performed on subsystem and end-to-end test systems at DSS 13 during May 1985 through March 1986. The test result of particular interest to participants of the Galileo Gravity Wave Experiment and members of the X-band Uplink Demonstration Team are the FFS values (for tau = 1000s) of between 9.51E-16 to 2.09E-15 obtained for an end-to-end system consisting of the exciter, transmitter in closed loop, X-band translator, X-band maser, MMR, and a modified Block III open-loop receiver. It is important to reemphasize that the FFS values do not include the stability of the hydrogen maser station frequency standard because in the measurement technique used, the instability of the frequency source tends to cancel out.

Due to the fact that only a few valid tests could be performed thus far on complete end-to-end system configurations, the end-to-end system test results are limited and not conclusive. It is recommended that more testing be done on the end-to-end systems at DSS 13. It is also recommended that a complete analysis be made of the measurement technique. Furthermore, cross-comparisons need to be made between the field-use test equipment (that was used to obtain the results of this article) and the test equipment being used by the JPL Frequency and Timing Group to make stability measurements on hydrogen masers. It is further recommended that the DSS 13 system be reanalyzed to see if account can be taken of common frequency source stabilities. Perhaps the method of taking the square root of the sum of individual subsystem Allan Variances is too conservative an approach. It is possible that some common-mode cancellations might be taking place in the system, thereby making the actual system stability better than was predicted.

Acknowledgments

References

1. Berman, A. L., "The Gravitational Wave Detection Experiment: Description and Anticipated Requirements," *TDA Progress Report 42-46*, pp. 100–108, Jet Propulsion Laboratory, Pasadena, Ca., Aug. 15, 1978.

2. Hartop, R., John, C., and Kolbly, R., "X-Band Uplink Ground Systems Development," *DSN Progress Report 42-56*, pp. 48–58, Jet Propulsion Laboratory, Pasadena, Ca., April 15, 1980.

3. Meeker, J., and Timpe, C., "X-Band Uplink Technology Demonstration at DSS-13," *TDA Progress Report 42-77*, pp. 24–32, Jet Propulsion Laboratory, Pasadena, Ca., May 15, 1984.

4. Koerner, M. A., "Doppler System Phase Transfer Functions for a System with an X-Band Uplink and X-Band and S-Band Downlinks," *TDA Progress Report 42-76*, pp. 58–69, Jet Propulsion Laboratory, Pasadena, Ca., Feb. 15, 1984.

5. Sosa, E. N., "Station Stability Measurement," *TDA Progress Report 42-62*, pp. 35–42, Jet Propulsion Laboratory, Pasadena, Ca., May 15, 1981.

6. Sosa, E. N., and Tyner, D. A., "Synthesizer Stability Evaluation," *TDA Progress Report 42-66*, pp. 319–330, Jet Propulsion Laboratory, Pasadena, Ca., December 15, 1981.

Table 1. DSS 13 X-band uplink only test results during May 1985 — March 1986

Part A. Exciter output tests with the exciter operating closed loop.
The output test port is EXC in Fig. 1.

Date of test	Xmtr syn[a]	Uplink frequency, MHz	Downlink frequency, MHz	FFS for tau = 1000s	#SDP	Air temp., °C	Comments
MAY 85	R	7200.00	NA[b]	1.58E-15	18	21.0–15.0	5-hour test
JUN 85	S	7200.00	NA	3.27E-16	13	37.0–36.5	
JUL 85	S	7200.00	NA	3.86E-16	11	31.9–30.0	
JUL 85	S	7200.00	NA	4.98E-16	9	29.0–26.0	
JUL 85	S	7195.00	NA	3.08E-16	15	24.0	Two-mixer method
JUL 85	S	7162.30	NA	1.09E-15	13	23–26.5–22	Two-mixer method
JUL 85	S	7200.00	NA	2.27E-16	3	26.0–28.0	Short test
SEP 85	S	7200.00	NA	8.77E-16	10	21.3–21.0	
SEP 85	T	7200.00	NA	9.49E-16	6	17.5–13.0	
SEP 85	T	7200.00	NA	9.46E-16	13	20.5–19.5	
NOV 85	S	7200.00	NA	4.77E-16	13	10–12–9.6	
DEC 85	T	7200.00	NA	8.84E-16	3	12.0	Short test
JAN 86	T	7200.00	NA	1.15E-15	6	21.0–20.0	
JAN 86	S	7200.00	NA	2.80E-16	6	19.8–18.5	
FEB 86	S	7180.00	NA	1.32E-15	10	15.5–14.5	Two-mixer method
FEB 86	S	7180.00	NA	9.06E-16	10	14.0–14.5	Two-mixer method
FEB 86	S	7166.94	NA	2.51E-16	3	17.0–15.0	Two-mixer method
FEB 86	S	7200.00	NA	7.65E-16	10	15.0–16.0	

Part B. Transmitter output tests with the transmitter operating closed loop.
The output test port is XMT in Fig. 1.

MAY 85	R	7200.00	NA	2.68E-15	6	29.0–28.0	Edited data
MAY 85	R	7200.00	NA	1.76E-15	5	29.0–28.0	Edited data
MAY 85	R	7200.00	NA	3.46E-16	8	26.5–21.0	
DEC 85	T	7200.00	NA	5.13E-16	8	11.5–10.3	

Part C. Transmitter output tests with the transmitter operating open loop.
The output test port is XMT in Fig. 1.

MAY 85	R	7200.00	NA	1.42E-15	13	20.0–18.5	
DEC 85	T	7200.00	NA	4.94E-16	13	10.0–10.5	

[a]The symbols T, R, S in column 2, respectively, refer to the DSS 13 Transmitter, DSS 13 Receiver, and DSN Spare Dana synthesizers used as the transmitter (or uplink) synthesizer.
[b]NA = not applicable.

Table 2. DSS 13 X-band uplink and X-band downlink only test results during May 1985 — March 1986

Part A. Translator X-band output tests with transmitter bypassed. The output
test port is XLTRX in Fig. 1.

Date of test	Xmtr syn[a]	Uplink frequency, MHz	Downlink frequency, MHz	FFS for tau = 1000s	#SDP	Air temp.,°C	Comments
DEC 85	T	7162.30	8415.00	1.56E-15	13	13.5–8.5	Two-mixer method
DEC 85	T	7162.30	8415.00	6.08E-16	13	9–10–9	Two-mixer method
DEC 85	T	7162.30	8415.00	4.99E-16	13	8.5–9.0	Two-mixer method

Part B. End-to-end system test with transmitter bypassed and with the Block III
receiver in closed loop. The output port is EECLX (Port 8) in Fig. 1. The
results for Part B were corrupted by leakage signals and are shown for
reference purposes only.

JAN 86	S	7177.92	8433.33	6.00E-16	9	17.0–9.5	
JAN 86	S	7177.92	8333.33	9.65E-16	2	5.5–6.0	Antenna moving
JAN 86	S	7177.92	8433.33	9.60E-16	11	5.5–10.8	
JAN 86	S	7177.92	8433.33	2.96E-15	6	9.5–8.5	Edited data

Part C. End-to-end system test with closed-loop transmitter included and with
the Block III receiver in closed loop. The output port is EECLX (Port 8)
in Fig. 1. The results for Part C were corrupted by leakage signals and are
shown for reference purposes only.

JAN 86	S	7177.92	8433.33	5.93E-16	13	8.0–7.0	
JAN 86	S	7177.92	8433.33	1.62E-15	2	7.0–5.5	Antenna moving
JAN 86	S	7177.92	8433.33	4.03E-15	4	10.0–9.5	Edited data

Part D. End-to-end system test with transmitter bypassed and with the Block III
receiver in modified open loop. The output port is EEOLX (Port 11) in Fig. 1.

JAN 86	S	7177.92	8433.33	7.95E-16	7	10.8–14.2	
FEB 86	S	7180.00	8435.78	9.04E-16	9	24.5–16.0	Note air temp.
FEB 86	S	7166.94	8420.43	6.72E-16	10	14.5–15.0	
MAR 86	S	7166.94	8420.43	3.05E-15	13	21.5–14.0	Edited data
MAR 86	S	7166.94	8420.43	1.19E-15	20	12.0–17.0	6-hour test
MAR 86	S	7166.94	8420.43	6.94E-16	16	24.8–12.5	Note air temp.

Part E. End-to-end system test with closed-loop transmitter included and with
the Block III receiver in modified open loop. The output port is EEOLX
(Port 11) in Fig. 1.

FEB 86	S	7180.00	8435.78	1.25E-15	6	25.5–24.8	
FEB 86	S	7180.00	8435.78	2.09E-15	17	25.2–28.5	
MAR 86	S	7166.94	8420.43	9.51E-16	11	25.0–25.5	

[a]See footnote a, Table 1.

Table 3. DSS 13 X-band uplink and S-band downlink only test results during May 1985 — March 1986

Part A. Translator S-band output tests with transmitter bypassed. The output
test port is XLTRS in Fig. 2.

Date of test	Xmtr syn[a]	Uplink frequency, MHz	Downlink frequency, MHz	FFS for tau = 1000s	#SDP	Air temp., °C	Comments
MAY 85	R	7177.92	2300.00	3.36E-15	13	15.0–16.0	Reasonable
MAY 85	R	7177.92	2300.00	7.07E-15	5	17.0–25.0	Edited data
JUL 85	S	7162.30	2295.00	2.05E-15	13	26.0–21.0	Two-mixer method
NOV 85	S	7162.30	2295.00	1.17E-15	6	3.5–1.0	Edited data
NOV 85	S	7162.30	2295.00	4.71E-16	4	3.5–1.0	Edited data

Part B. Multi-mission receiver S-band output tests with transmitter bypassed.
The output test port is MMRS in Fig. 2.

JUL 85	S	7162.30	2295.00	2.76E-15	4	20.5–19.5	Edited data
JUL 85	S	7162.30	2295.00	1.15E-15	7	20.5–19.5	Edited data
SEP 85	T	7162.30	2295.00	5.04E-15	8	22.0–24.2	Edited data

Part C. End-to-end system test with transmitter bypassed and with the Block III
receiver in closed loop. The output port is EECLS (Port 14) in Fig. 2. The
results for Part C might have been corrupted by leakage signals and are
shown for reference purposes only

NOV 85	S	7162.30	2295.00	6.45E-15	13	0.8–<0.0	
NOV 85	S	7180.00	2300.70	4.40E-15	2	2.0–1.5	
NOV 85	S	7180.00	2300.70	4.60E-15	14	<0	
JAN 86	S	7177.92	2300.00	1.23E-15	9	14.0–11.5	

Part D. End-to-end system test with closed loop transmitter included and with
the Block III receiver in closed loop. The output port is EECLS (Port 14) in
Fig. 2. The results for Part D might have been corrupted by leakage signals
and are shown for reference purposes only.

NOV 85	S	7180.00	2300.70	6.94E-15	3	5.0–4.0	Xmtr output 6 kW
NOV 85	S	7180.00	2300.70	4.26E-15	3	0.0–1.0	Edited data

Part E. End-to-end system test with transmitter bypassed and with the Block III
receiver in modified open loop. The output port is EEOLS (Port 17) in Fig. 2.

NOV 85	S	7162.30	2295.00	1.52E-14	8	7.0–8.0	Bad result

[a]See footnote a, Table 1.

Table 4. JPL internal documents

1. Otoshi, T. Y., "Report on First Station Stability X-Band Measurements at DSS 13,", *IOM 3331-XX-XXX* (not numbered), Jet Propulsion Laboratory, Pasadena, Ca., April 4, 1984.

2. Otoshi, T. Y., "Final Report on Second Station Stability X-Band Measurements at DSS 13 During April 1984," *IOM 3331-84-027*, Jet Propulsion Laboratory, Pasadena, Ca., May 5, 1984.

3. Otoshi, T. Y., "Report on Third Station Stability X-Band Measurements at DSS 13," *IOM 3331-84-034*, Jet Propulsion Laboratory, Pasadena, Ca., June 18, 1984.

4. Otoshi, T. Y., "Current Status Report and Distribution of Preliminary Reports for Dec 1984 and Feb 1985 Stability Tests at DSS 13," *IOM 3331-085-012*, Jet Propulsion Laboratory, Pasadena, Ca., March 6, 1985.

5. Otoshi, T. Y., "Report on DSS 13 Station Stability Tests Performed during March 7–13, 1985," *IOM 3331-85-027*, Jet Propulsion Laboratory, Pasadena, Ca., May 8, 1985.

6. Otoshi, T. Y., and Franco, M. M., "Interim Progress Report on DSS 13 Station Stability Tests Performed During May 1985–March 1986," *IOM 3331-086-026*, Jet Propulsion Laboratory, Pasadena, Ca., May 20, 1986.

Fig. 1. Output test ports for frequency stability testing on X-band uplink and X-band downlink subsystems at DSS 13 (modified figure from Ref. 3). Example frequencies are for channel 14.

FOR OPEN-LOOP RECEIVER OPERATION, THE VCO IS AUTOMATICALLY SWITCHED OUT, AND THE MIXER AND ×50 ASSEMBLY ARE REPLACED BY A ×31.25 MULTIPLIER

NEW DESIGN

f_{XT} (7162.3125)

XPNDR

$f_{SR} = \frac{240}{749} f_{XT}$ (2295)

S-BAND COMBINER
X-BAND POLARIZER
X-BAND FEED

XLTRS ⑫

MODIFIED BLK III RCR

X/S TEST SIG TRANSLATOR

X-TRANSMITTER

$f_{SR} = \frac{240}{749} f_{XT}$ (2295)

CPLR X

CPLR S

f_{XT} (7162.3125)

f_{SR}

$\frac{509}{749} f_{XT}$ (4867.3125)

MIXER –

f_{XT} (7162.3125)

DPLXR

TO X-BAND RCR

S-BAND DPLXR

f_{SR} (2295)

LONG FB LOOP

XMTR P.A.

FROM S-BAND XMTR

X/S DOPPLER EXTRACTOR

SHORT FB LOOP

S-BAND TWM

× 5

$\frac{8144}{749}$

÷ 5

MIXER +

× $\frac{13}{16}$

φ-DET AND φ-COMP

$\frac{509}{749} f_{XT}$ (4867.3125)

MIXER –

f_{XT} (7162.3125)

× 16

MIXER +

× 65

MULTIMISSION RCVR

f_{SR} (2295)

× 20

MIXER –

MIXER –

(2000)

100

CS RECEIVER

ANTENNA

1200 ft

CONTROL ROOM

20

MMRS ⑬

$f_{SR}-2000$ (295)

$f_{SR}-2000$ (REF) (295)

$f_{SR}-2000 + 1050$ (1345)

DCO #1 (41.39453125)

CS XMTR

U.C./D.C.

MIXER –

MIXER –

X-EXCITER

× 50

1050

(1050)

MIXER –

(50)

100 MHz REF

5 MHz REF

100 MHz REF

DCO #3

(47.90)

MIXER –

(26.90)

MIXER –

1 GHz

1 GHz

50

49/51

MIXER –

MIXER –

45 MHz

BLK III RCVR

21 MHz VCO

MIXER –

60 MHz

⑭

EECLS

⑮

49/
5 10 45 51 60 1 GHz

10

EEOLS

⑯

1 MHz

5 MHz

100 MHz H-MASER REF

REF FREQ'S DERIVED FROM H-MASER

LOOP FILTER

MIXER –

10 MHz

EEOLS

⑰

X/S DOPPLER MIXER

Fig. 2. Output test ports for frequency stability testing on X-band uplink and S-band downlink subsystems at DSS 13 (modified figure from Ref. 3). Example frequencies are for channel 14.

Fig. 3. Instrumentation used for station stability testing
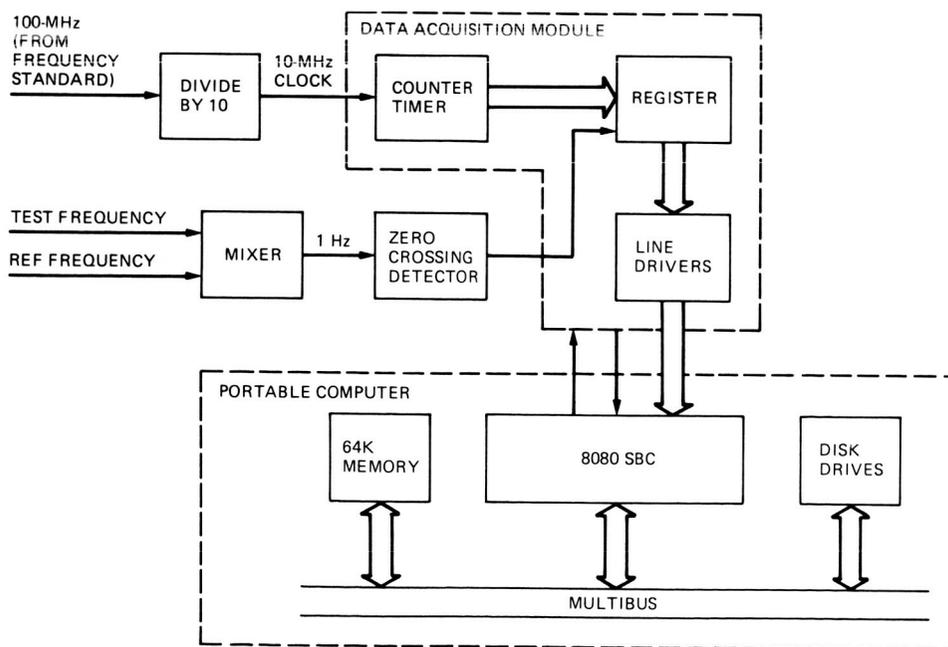
Fig. 4. Block diagram of Data Acquisition System (inside the dashed lines)

100 MHz

MICROWAVE
SUBSYSTEM
UNDER TEST

f_{TEST}

AMPLIFIER
(FOR XLTR
TESTS ONLY)

H-MASER
REFERENCE
FREQUENCY

100 MHz

x M

f_M

MIXER
No. 1

ISOLATOR

f_{REF}

MIXER
No. 2

1 Hz

1 Hz
AMPLIFIER
AND FILTER
(DC TO 10 Hz
LOW PASS)

f_{SYN}

HP3325A
FREQUENCY
SYNTHESIZER

OSCILLOSCOPE

ZERO
CROSSING
DETECTOR

100 MHz

DIVIDE
BY 10

DATA
ACQUISITION
SYSTEM
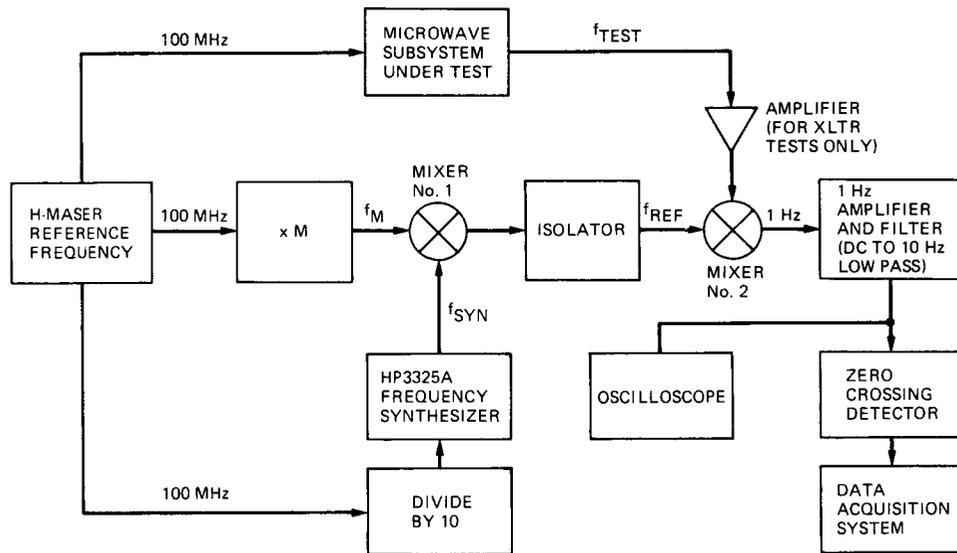
NOMINAL VALUES FOR MICROWAVE SUBSYSTEM TESTING AT CHANNEL 18					
DSN MICROWAVE BAND	M	f_M, MHz	f_{SYN},[a] MHz	f_{REF}, MHz	f_{SYN}/f_M
X-BAND UPLINK	72	7200	33.06	7166.94	0.0046
X-BAND DOWNLINK	84	8400	20.43	8420.43	0.0024
S-BAND DOWNLINK	23	2300	3.52	2296.48	0.0015

[a]SET f_{SYN} SO THAT $f_{REF} = f_{TEST}$ (+/-) 1 Hz

Fig. 5. Two-mixer method for microwave subsystem frequency stability testing

New Short-Time Alignment Technique for 70-Meter Antenna Surface Panels

M. S. Katow

Ground Antenna and Facilities Engineering Section

With severely limited field modification time for upgrading the 64-m antenna to 70-m diameter, a new shorter time method for aligning the surface panels of the main reflector was needed. For each target on the surface panel, both distance (or range) and elevation angle measurements are made. A new technique for setting the surface panels at zenith look has been devised. This article describes the software required to convert the computed target distortions obtained from the JPL-IDEAS structural analysis computer program (defining the gravity load change from a 45-deg elevation angle to zenith look) into the theodolite reading at zenith look. The technique results in a perfectly shaped reflector at the 45-deg rigging elevation angle, with acceptable surface error tolerance.

I. Introduction

The normal procedures used to set the surface panels of the main reflector are as follows:

(1) Assemble the surface panels at zenith look with the mounting screws near the midpoints of their adjustment ranges with the edges of the surface panels aligned. A drill tape, holding drill bushings located at predetermined fixed arc distances (measured from the vertex of the paraboloid surface), is positioned on the surface panels. The mounting holes for the targets (for theodolite readings) are drilled as located by the bushings. The tape is laid on the surface radially and is moved circumferentially to cover all target positions.

(2) The elevation angular positions of the targets are read by the theodolite, with the antenna tilted at the rigging angle (usually 45-deg elevation). The theodolite

bearings are preloaded to suit the 45-deg tilt prior to taking angle measurements. The difference between the actual angle read and the ideal angle (computed from the ideal reflector surface equation) is converted into a target translation (either up or down), which in turn is then converted into the number of turns needed for the mounting screws.

(3) The reflector is rotated to zenith look to correct the target's position by turning the surface panel mounting screws, accessible through holes in the surface panels.

The new panel setting method uses an advanced elevation-angle-measuring theodolite (Kern Model E2) attached to a distance or ranging device (Kern Model DM 503). Optical mirrors (prisms) are used as targets mounted on the surface panel corners. By this new combination of target angle and range (measured from the center) plus the computed distor-

tion angle of the target (due to the change in gravity loading caused by the rotation of the reflector structure from the 45-deg rigging elevation angle to zenith look), the surface panels will be set directly at zenith look. In this case, the theodolite bearing preloading will not be required.

The theodolite zenith angle and the range measurements (using the ranging device) are simultaneously read and then fed into a portable microcomputer so that the target position corrections can be computed instantly and printed out.

The new setting method eliminates several time-consuming processes, including the use of the drill tape to locate the targets and instead, using semiautomated target range readout. Also, corrections to the adjustment screw of a target can immediately follow a reading by the theodolite. The description and calibration of the new angle/range measuring instruments will be given in a separate future reporting by others. In this article, only the algorithms used in the computer software are described.

II. Target Identification

The primary objective of the 70-m antenna surface panel setting alignment procedure is to obtain the newly shaped surface profile and geometry, as shown in Fig. 1 of JPL drawing No. 9486339 (see footnote 1) at the 45-deg elevation rigging angle. Because this shaped surface is axis-symmetric, only a radial profile is needed to completely describe the surface. A table of radial distances and heights, as measured from, and above, the vertex of the replaced 64-m surface, is also shown in the drawing. Figure 1 of this article gives the typical height dimensions of the theodolite and each optical target (prism or corner cube).

The targets are located on the individual surface panels (as shown in JPL drawing No. 9487694; JPL internal document). Figure 2 shows the target numbering scheme. The circumferential rows are numbered sequentially, starting from the reflector center up to row 21 (the outermost row). However, to account for the targets that were intentionally moved to miss hatches, etc., row numbers 22 through 25 are used as shown in Fig. 2. The radial columns are numbered clockwise from top to bottom. A target is numbered as the sum of the 100X column number plus the row number.

The first and fourth quadrant targets shown in Fig. 2 conform closely to the locations of the work points of the

half-structural finite element model with the Y-Z plane of symmetry. The target numbers in the second and third quadrants are given negative signs. The distortion vectors of the targets in the second and third quadrants have only sign changes in their X components relative to the first and fourth quadrant distortion-vectors of the corresponding mirror-image targets. When the theodolite reads or points to a target, the inputs of range, elevation angle, and azimuth angle in subroutine (TNODE) define the target number.

III. Obtaining the Undistorted, Shaped Surface Profile

When the surface panels are initially assembled on the reflector, the targets on the panels are arranged in rows or circles, with the radial distances measured from the centerline of the theodolite (as shown in Fig. 1) to within ±1 cm (0.4 in.) of nominal, using measuring tapes or other ranging devices.

A table giving nominal radial distances of each target, including radial distances of plus and minus 2.54 cm (1 in.) and 5.08 cm (2 in.) for each row of targets, was generated. This table was then used to compute respective heights vs range, defining the undistorted, shaped surface profile by interpolation, using the slightly modified software described in JPL internal document D-1843 (see footnote 1). By geometry, the range, elevation angle, and slope, as defined by Fig. 1, were also computed (in run-stream-ST-MGLBCHK/70M) for each radial distance.

Because the new ranging theodolite is most precise with the azimuth axis vertical, the setting of the surface panels will be restricted to the zenith look of the reflector. As will be described later, the distortion component at each target will be reduced to an equivalent change (Δ) in the ideal elevation angle of the theodolite-to-target line. The change (Δ) is added to the elevation angle that defines the undistorted shaped profile. The result is that target-setting measurements and corrections can be done directly with the antenna at the zenith-look altitude.

The 70-m antenna shaped-surface profile was designed by JPL to generate a uniform radiation pattern on the main reflector and satisfying other microwave requirements. A Physical Optics (PO) analysis[2] was performed to generate the shaped-surface contour that closely follows the existing 64-m antenna paraboloid. The focal length for the 64-m antenna is 27.109 m (1067.294 in.), and the final design focal length of the best-

[1] A. G. Cha, and W. A. Imbriale, *Computer Programs for the Synthesis and Interpolation of 70M Antenna Reflector Surface*, JPL Report D-1843, internal document, Jet Propulsion Laboratory, Pasadena, Calif., November 1984.

[2] A. G. Cha, *Physical Optics Analysis of NASA/JPL Deep Space Network 70M Antennas*, JPL Report D-1853, internal document, Jet Propulsion Laboratory, Pasadena, Calif., November 1984.

fit paraboloid to the new shaped profile of the 70-m antenna is 27.237 m (1072.329 in.).

IV. Gravity Loading Distortion Analysis

The gravity loading components of the reflector structure at zenith look (shown in Fig. 3) are computed as changes from the gravity loading at the 45-deg rigging elevation angle. From the structural analysis program (IDEAS, Ref. 1), the distortion vector obtained at zenith look will equal the sum of: (a) the unit (1.0 g) gravity loading (off/on) distortion vector in the direction +z multiplied by -0.293, plus (b) the unit (1.0 g) gravity loading (off/on) distortion vector in the direction +y multiplied by +0.707.

With a negligible loss in accuracy, the Z coordinates of the top chords of the reflector structure were assumed to be on a close-fit paraboloid to the shaped profile. Thus, the paraboloid RMS (root mean square) best-fit program (described in Ref. 2) with its many computing options, can be advantageously used.

The two outputs (file 81) of the IDEAS program (with the JRMFIL option) are for the unit +Z gravity-loading distortion vectors assigned to the top nodes of the reflector structure, followed by the unit +Y gravity loading distortion vectors. The format of the IDEAS output is compatible with the RMS program input requirements consisting of the coordinates of the grid nodes, the three-component distortion vectors, (u, v, w) the area weighting factors (A) and the identifying number of the grid nodes.

A special run-stream (ST-RMS-ZENLK/70MA-HALF) adds the +Z and +Y gravity loadings, as proportioned by the above gravity vector changes (outlined in Fig. 3), and outputs the summed zenith-look distortion vector (in file 80). At the same time, a constrained best-fit paraboloid was fitted to these distortion vectors (at the zenith look), forming a paraboloid having a 27.229 m (1072.0 in.) focal length. A 1/2 contour map of the best-fit paraboloid is shown in Fig. 4.

The vertex of the best-fit paraboloid, for a gravity loading in the +Y lateral direction, has typically shifted in the -Y direction by -17.752 cm (-6.969 in.) with a rotation of the Z axis about the X axis of -0.003881 radians. The RMS program computes the surface normal errors at each grid node of the best-fit-paraboloid. The resulting contour level lines for the right half are plotted in Fig. 4.

The above best-fit results obtained by the RMS program are of interest also for predicting the performance loss of the 70-m antenna due to the distortions of the reflector structure from the ideal shape. However, in this article the constrained

best-fit paraboloid is assumed to be the baseline shape to which the (target) distortion vectors are added to define the distorted shape.

By conforming or constraining the axis of the best-fit paraboloid to coincide with the axis of the measuring theodolite of the undistorted, shaped profile (by translation, rotation and focal length change), the resultant target distortion vectors can be converted to a distortion angle (ideally, to be read by the theodolite), as shown in Fig. 5.

The distortion data at zenith look are first sequenced (by run-stream ST-SEQ-RMS/70MA) to provide grid node numbers for the midpoint adjustment screws for panel rows No. 4 and 5 and the targets moved to clear hatches and other openings, as shown in Fig. 2. Two inner nodes for columns No. 3, 7, 11, 15, etc. are deleted to match the targets actually used on the surface panels. Next, the sequenced distortion data at zenith look is best-fit with the RMS program, with constraints described above (by run-stream ST-EL90545/FFITT-70-A).

The results from this constrained best-fit are plotted in Fig. 6 for only the YZ plane. The coordinate system for this constrained best-fit paraboloid is shown by the rotated baseline marked with 0.001066 radians rotation. In the Z direction, the normal distortion vectors are plotted vs the radial distances from the center for the YZ plane (for the change from the 45-deg elevation angle to zenith look) and the result after the best-fit procedure.

V. Panel Midpoint Data Interpolation

Distortion data for the midpoints of panel rows No. 4 and 5 and the targets relocated to miss hatches and other openings were not output by the IDEAS program because no finite-number grid numbers existed for these nodes. The normal distortion vectors for these adjustment targets were generated by straight-line interpolation from the distortions of the targets on the same column on adjacent rows. (Run-stream ST-GEN-INTM/70M-DIST computes and writes the interpolated distortion vectors in file 8).

VI. Surface Panels Installation

The assembly of the reflector structure at zenith look to match the engineering drawing specifications and to the axis-symmetric configuration presents no special problems. Some additional compensation for gravity-induced bending deflections may be required as more truss members are added radially during field assembly.

If the surface panels along the YZ plane of symmetry are installed to compensate for the "as computed" gravity dis-

17

placement, the top edge of the uppermost panels must be set about 46 mm (1.80 in.) below the nominal surface, and the bottom edge set about +37 mm (1.47 in.) above the nominal surface, as shown in Fig. 6. However, surface panel supports for this wide range of adjustments will be impractical for the present simple design scheme of the adjustment device.

The adjustment ranges can be decreased if the rotational components of the gravity displacements are removed and compensated by a small change in the elevation angle. This is accomplished by the best-fit described above, and with the alignment of the theodolite axis with the constrained-fit paraboloid axis. As shown in Fig. 6, the negative normal setting value reduced to about 4.0 mm (0.2 in.) and 7.0 mm (0.3 in.) for the positive-signed offsets along the YZ plane. The normal setting values for the first and fourth quadrant are shown in the contour map of Fig. 7.

The actual setting position (or theodolite angle) of a target will be the sum of the angle for the undistorted, shaped profile plus the equivalent angle for the distortion as described above with the axis of the reflector at 89.939 deg (90.000–0.061) elevation angle.

VII. Theodolite Computer Software

The theodolite readings (elevation and azimuth angle plus range data to be stored and processed in the computer) will be entered automatically at the same time by a pushbutton switch on the theodolite. The software sequence of steps is as follows:

First, after run-stream (ST-THEOX/70A) is activated, a table of range, elevation angle, and slope of the targets (70M-RANGE/TABLE) located on the undistorted shaped profile will be read in. Second, the gravity distortion data (EL90545-FFIT/70MA-SIM-FDR + INTM-ANGD1/70MA) for the zenith-look altitude are read in as normal distortion vectors for all target nodes. Third, the previous day's accumulated target readings are read into the computer core (from file 11). Finally, new target readings of zenith angle, range, and azimuth angle to a new target are read in by the theodolite. From the elevation angle (converted from the zenith angle theodolite reading), range and azimuth angle, subroutine TNODE decodes the node number (column X 100 + row).

When the measured range is input to the program THEOX/70M the elevation angle (ELG) for a target on the undistorted shaped profile (Fig. 7) is determined for this range by interpolating the range table (70M-RANGE/TABLE) generated by run-stream ST-MGLBCHK/70M, as previously discussed.

The distortion vector, in its normalized form (FER in Fig. 7) is converted to an equivalent theodolite elevation angle (ERELA in Fig. 7) which, when added to angle ELG, should result in the correct setting elevation angle of ELGR. The program THEOX/70M, for each reading input from the theodolite, will give the following outputs: (1) the elevation angle for a target on the undistorted shaped profile for the input range reading, (2) the difference between the input elevation angle and the above computed angle, (3) the target number, and (4) the distortion angle, linear error, and number of corrective turns of the adjustment screw.

After one day's target reading, the program will list a summary of the nodes read and their setting values, followed by a listing of 50 nodes without any theodolite input data. An overall RMS of the normal errors will be output, followed by the RMS values per each row of targets. File 11 will be overwritten with all new input data, including the corrected reading.

VIII. Quadripod and Subreflector Alignment

If the quadripod is assembled to align with the same theodolite used to set the surface panels, and to its datum targets, the quadripod will be erected with a built-in compensation of 0.061 deg for the resulting deflection from zenith to the 45-deg rigging angle. However, it will be necessary to add more compensation than 0.061 deg for gravity displacements indicated by the displacement numbers shown in Fig. 6.

The datum targets on the main reflector, used at zenith look by the panel's target setting theodolite, will not be normal to the reflector axis at the 45-deg rigging elevation angle where the final setting of the subreflector will be necessary. Another set of datum targets can be installed for use at 45 deg elevation.

IX. Results of Algorithms Verification

A check on the overall accuracy of the algorithms used was made by first generating (from the run-stream ST-MGLBCHK/70M) a series of simulated theodolite readings of: (1) zenith angle, (2) range, and (3) azimuth angle (TEST-DATA/70M-PERFECT) for the targets located on the undistorted, shaped profile at radial distance of the 70-m targets (file 9). To reduce the number of data points, targets at every odd numbered column and row for only the first and fourth quadrants were selected to compare with results from the distortion data.

The simulated theodolite reading of each target on the undistorted, shaped profile was generated. These readings should produce correction angle or normal length equal to the computed normal distortion vector, as described in section IV.

These simulated theodolite readings were input to the basic theodolite/computer operating program (THEOX/70M-ABS) by run-stream (ST-THEOX/TEST-70MA), incorporating the distortion vectors from an initial IDEAS analysis (designated as 70MA in this article). For each theodolite reading on the undistorted, shaped profile, the computed normal error is in agreement with the normal error computed by the run-stream (ST-RMS-ZENLK/70MA-HALF, described in section IV). The overall normal RMS distortion was 2.89 mm (0.114 in.), which also is in agreement with the normal RMS error of 2.88 (0.1135 in.) of ST-RMS-ZENLK/70A-HALF, where the RMS error was computed for the odd-numbered columns and rows of targets.

To check the generation of the undistorted, shaped profile described in section III, a series of even-numbered radial distances was input to the run-stream (ST-MGLBCHK/70M-CHECK). The resulting height calculations verified the values in Table 1 (JPL drawing No. 9486339, Sheet 3 of 3; JPL internal document) to four decimal places.

X. Summary

With a limited time for antenna construction, a new scheme was required for setting the surface panels, using a ranging and angle-measuring theodolite connected to a computer, with the reflector positioned at zenith look.

The software used in the computer is described in this article. To comply with the polar coordinate system of the theodolite, the dimensions of the undistorted, shaped profile of the reflector were converted to range and elevation-angle values. By constrained best-fit, the computed distortion vector (by the IDEAS program) with the RMS program adjustments for the surface panel supports were minimized and added to the shaped-profile dimensions.

With the computer-operated theodolite, theodolite-target field readings were converted in real time into point corrections. The latter are converted into the number of turns needed for the adjustment screws holding the surface panels.

References

1. Levy, R., "Optimization of Antenna Structure Design," *Eighth Conference on Electronic Computation*, pp. 114–129, ASCE, Houston, Tex., February 1983.

2. Katow, M S., and Schmele, L. W., "Antenna Structures: Evaluation Techniques of Reflector Distortions," *Space Programs Summary 37–40*, Jet Propulsion Laboratory, Pasadena, Calif., Vol. IV, September 1968, pp. 176–184.
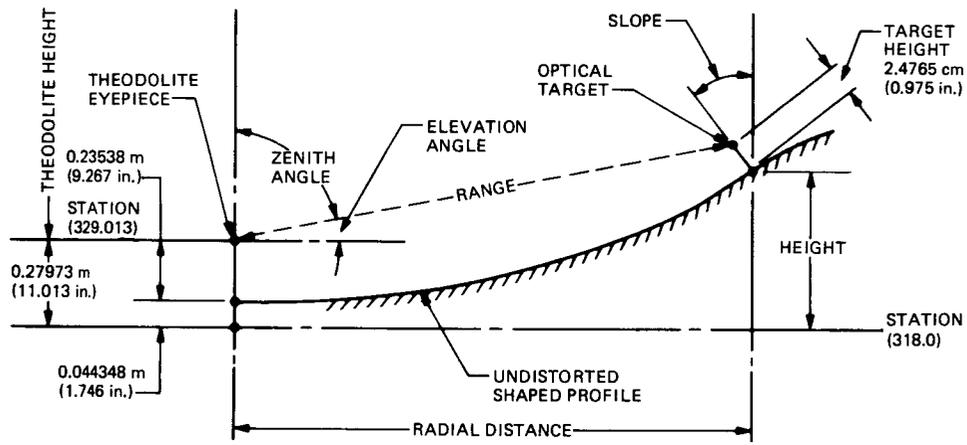
Fig. 1. Shaped main reflector profile

Fig. 2. Target-numbering scheme

G = 1.0 UNIT GRAVITY FORCE

GRAVITY LOADING
VECTORS AT
ZENITH LOOK

SYMMETRIC
(IDEAS) OUTPUT

+0.707 G

0.707 G
(COSINE 45°)

−0.293 G

−0.707 G
(SINE 45°)

45°

0.293

ANTISYMMETRIC
(IDEAS) OUTPUT

1.0 G

ELEVATION 45°
ANGLE

90°
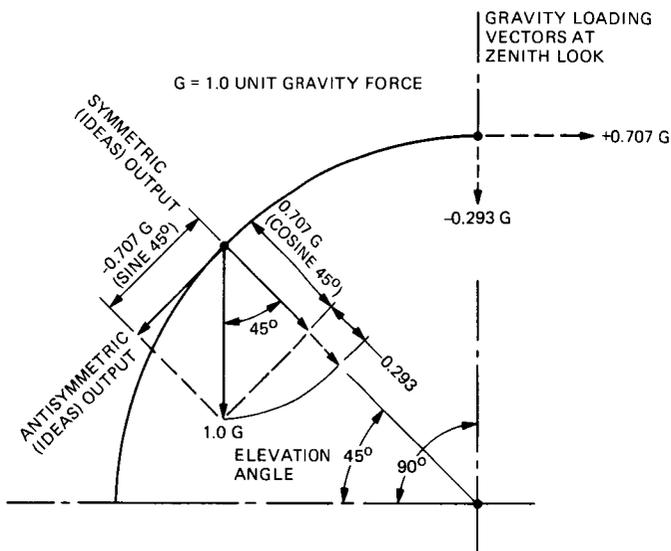
**Fig. 3. 70-m gravity loading components for surface panels
rigged at 45-deg elevation**

+Y

CONTOUR INTERVAL = 0.5 mm
1/2 PATHLENGTH RMS = 0.053
(0.021 in.)

0.

+0.5

+1.5

+1.0

+2.0

−0.5

+1.0

0.

+0.5

−0.5
0

−0.

+X

+1.0

−0.5

−0.5
−1.0

+0.5

−1.5
−2.0

−0.5

+0.5
+0.5

0.

0.
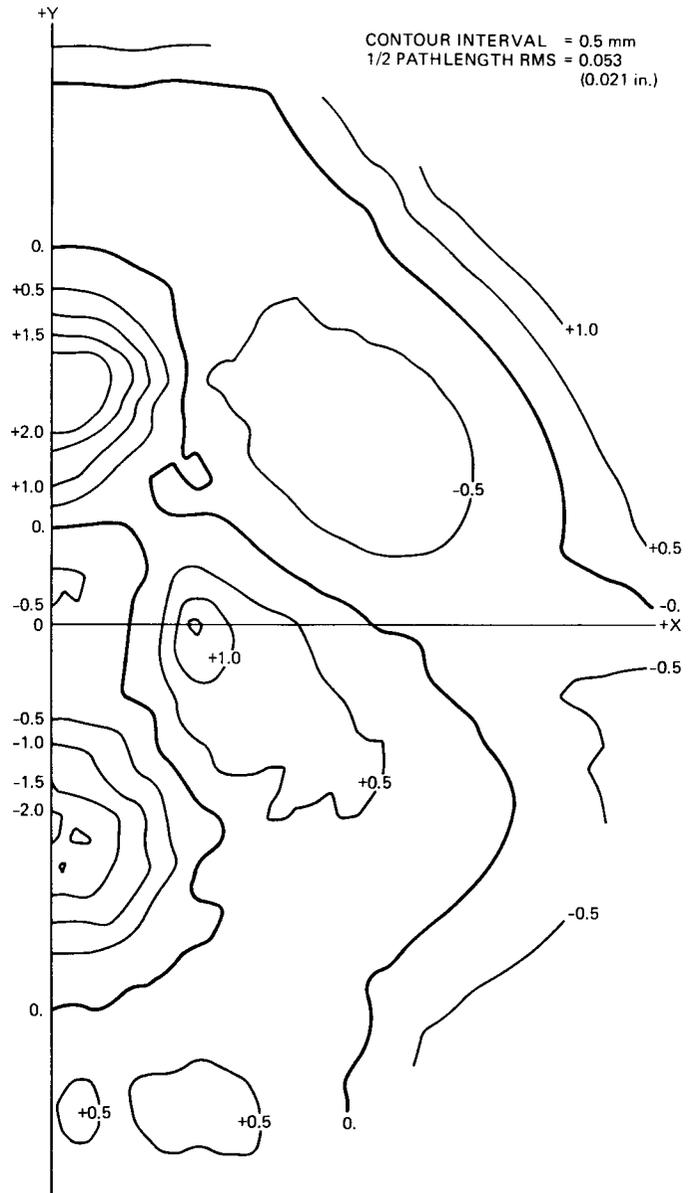
**Fig. 4. 70-m reflector structure best-fit distortion at EL-90
with panels set at 45 EL**
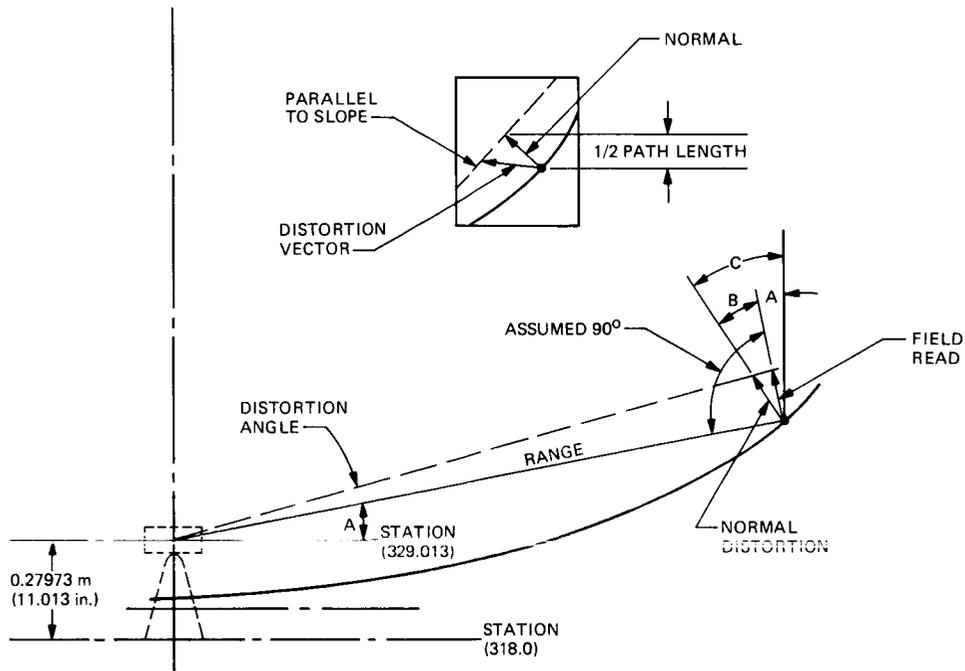
A = THEODOLITE ELEVATION ANGLE
B = C - A
C = SLOPE OR TANGENT ANGLE TO PARABOLOID

$$\text{DISTORTION ANGLE} = \frac{\text{NORMAL} \times \text{COS (B)}}{\text{RANGE}} \text{ rad}$$

Fig. 5. Distortion angle geometry
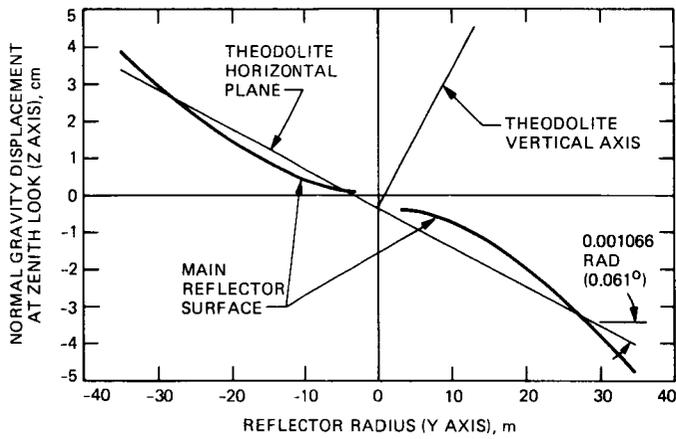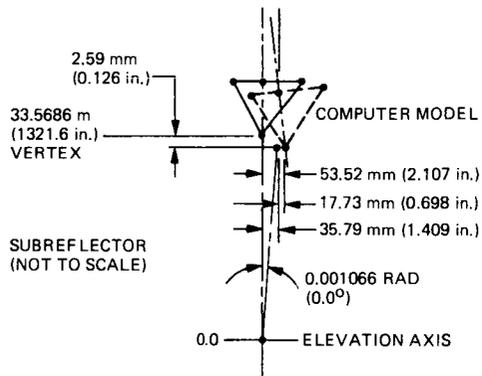
Fig. 6. Gravity loading displacements at zenith look: YZ plane
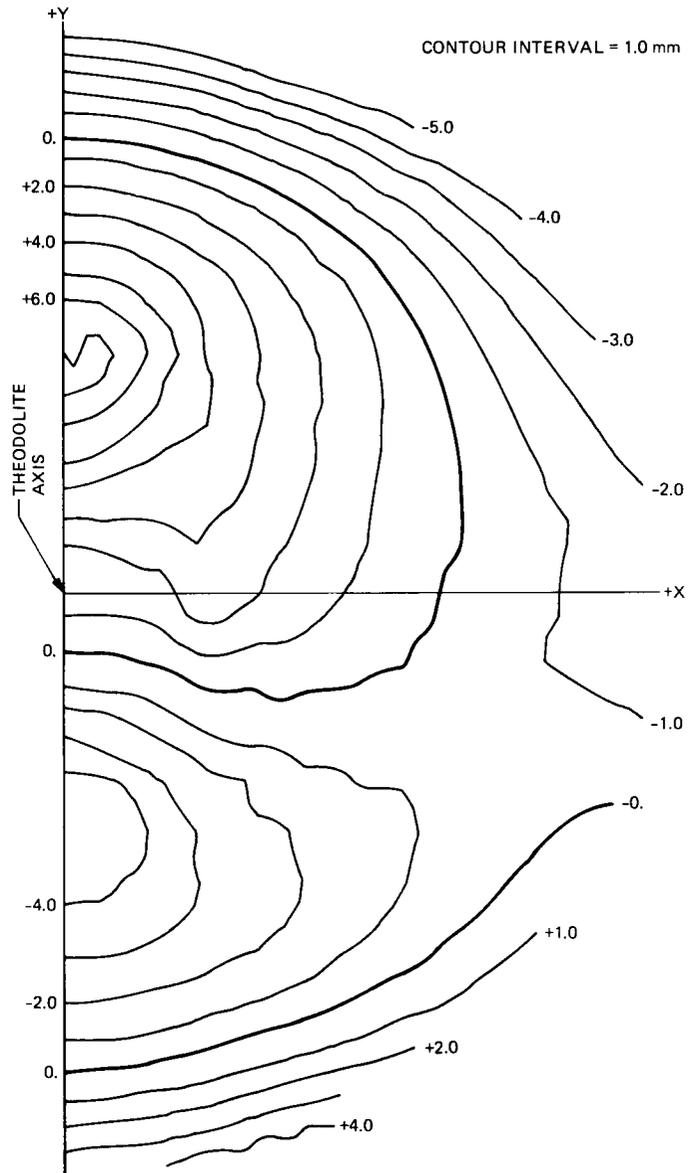
Fig. 7. Distortion contour map referred to the theodolite axis
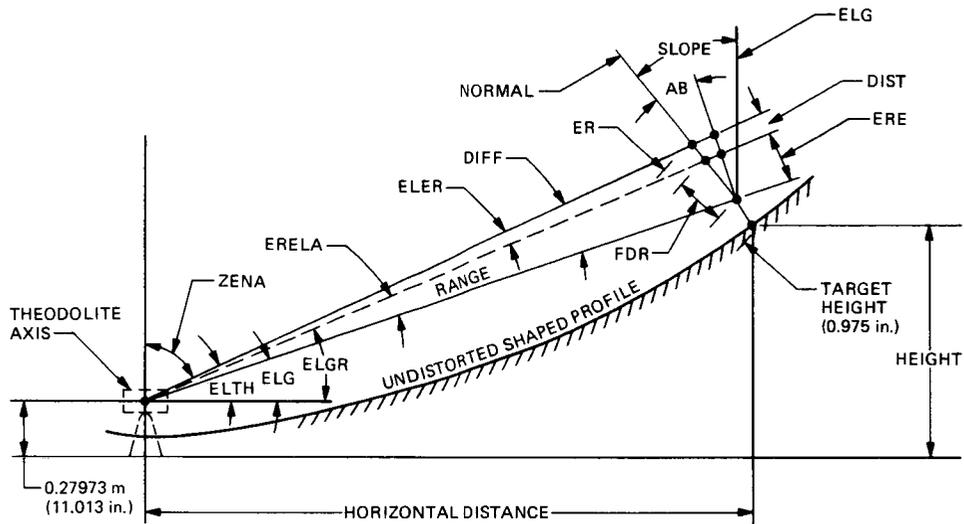
ZENA	MEASURED ZENITH ANGLE (Grad)
ELG	ELEVATION ANGLE TO TARGET ON UNDISTORTED SHAPED PROFILE (Grad)
ELTH	MEASURED ELEVATION ANGLE (Grad)
ERELA	ANGULAR DISTORTION VECTOR (Grad)
ELER	ANGULAR POSITION ERROR OF THE TARGET (Grad)
SLOPE	SLOPE ON UNDISTORTED SHAPED PROFILE (deg)
FDR	NORMAL DISTORTION VECTOR
ER	CORRECTION – CONVERTED TO NO. OF TURNS OF PANEL ADJUSTMENT SCREW (20 THREADS/in.) PANEL ADJUSTMENT SCREWS ARE NORMAL TO SURFACE
ERE	LINEAR DISTORTION VECTOR NORMAL TO LINE OF SIGHT
DIST	TARGET POSITION ERROR NORMAL TO LINE OF SIGHT
DIFF	ELTH – ELG (Grad)
ELGR	CORRECT TARGET SETTING ANGLE (Grad)

Fig. 8. Corrective algorithm notations

A Prototype DSN X/S-Band Feed: Model III Development[1]

D. A. Bathker and H. F. Reilly

Radio Frequency and Microwave Subsystems Section

This article is the seventh in a series documenting development of a prototype X/S-band common aperture Cassegrain feedhorn for DSN use. A Model III combiner has been developed to increase S-band bandwidth to include the Highly Elliptical Earth Orbiter band from 2025 to 2110 MHz, and to provide a 400-kW CW S-band uplink and a possible planetary radar band near 2320 MHz. The combiner uses eight S-band waveguide injection slots arranged in four pairs. The problems of this design geometry associated with rejection filtering and X/S-band matching are discussed.

I. Introduction

This is the seventh in a series of reports (Refs. 1 through 6) documenting development of a common aperture, two-frequency band Cassegrain microwave antenna feed system for specialized DSN use. This report briefly summarizes previous work and discusses problems associated with the most recent configuration, Model III. The problems are documented with the expectation that future efforts of a similar nature will necessarily be undertaken. It is hoped this report will provide future investigators with a valuable background, and increase the efficiency of the probable future development cycle.

---

[1]At inception of the DSN dual frequency common aperture feedhorn development series, it was usual practice to identify the various microwave frequency bands by letter designations, e.g., X- and S-bands. Since then the impreciseness of such letter designators has become apparent, and the use of numerical frequency designators was instituted as the DSN standard. Because of the long series of reports on this development, we retain the X/S designator to maintain the continuity and interrelationship of each report in the series. Wherever cited, S-band is defined to be 2.1 - 2.3 GHz and X-band 8.4 GHz.

II. Previous Work

For many years, DSN 64- and 34-m operational tracking antennas have functioned simultaneously in the 2.1 (transmit), 2.3 (receive), and 8.4-GHz (receive) bands. Proper antenna RF performance parameters are obtained in the two widely separated frequency bands by the reflex-dichroic feed system (Ref. 7). This feed system spatially separates the S- and X-band feed beams and then directs those individual beams to (or from) appropriate single-band, high performance feed-horns. Although it is capable of nearly unlimited CW uplink power and no measurable degradation to S-band receive performance, the reflex-dichroic approach does impact X-band receive noise temperature at the 2-kelvin (0.5-dB G/T) level. For this reason alone, investigations into alternative approaches were indicated to improve the critical X-band performance, even at the cost of a slight impact to S-band performance, if necessary.

Following work which showed generally poor performance of one kind or another from a variety of simple approaches

(such as horn-within-horn), a common aperture feedhorn approach was undertaken. This approach uses a very large feedhorn operating in a beamwidth saturation mode at X-band, such that the radiation pattern at X-band is very similar to that at S-band. Corrugations are used so that the necessary capacitive surface impedance is realized in both bands. The groove depth at S-band is about 0.35 λ_0, while at X-band it is about 1.35 λ_0. This corrugation "harmonic" principle ensures that the horn aperture fields are well tapered in all planes, yielding desirable horn pattern functions at both frequency bands. In fact, an additional band at 0.85 λ_0 groove depth is available (as are others at harmonics above 1.35 λ_0) but is unused. As stated, the horn is large—more than 7 λ_0 in aperture diameter at S-band and over 28 λ_0 at X-band. Yet the saturation phenomenon (caused by phase error in the aperture) functions at X-band to produce −10 dB pattern beamwidths that are nearly the same as at S-band. Little trouble was experienced in the development of the horn portion of this feed system. When separately excited with appropriate single-band throat sections, suitable performance is attained.

The challenge associated with the DSN X/S-band common aperture feed development effort lies in designing the combiner (or separator) portion; that is, extraction (or insertion) of the longer wavelength band at an appropriate horn diameter while providing an uninterrupted propagation path for the shorter wavelength band to exit (or enter) the normal throat region. The basic combiner concept adopted is shown in Fig. 1 of Ref. 2.

Three separate and distinct combiner developments have been made. Model I, described in Refs. 1 through 4, provided approximately 50 MHz of S-band bandwidth, sufficient only to prove the concept. The height of the four circumferential S-band waveguide injection slots was 8.9 mm (0.35 in.). The slots were arranged in a radial line configuration, providing minimal interruption to the X-band axial wave, even given the possibility of imperfect radial line X-band rejection filtering. These reject filters attempt to provide a virtual short-circuit of X-band impedance at a specific location near the corrugated waveguide wall. The specific location is the same as the root, or bottom, of the normal corrugations. In this manner, it was expected that the X-band wave would undergo no discontinuity in the transmission line (the horn).

In Model I, isolation of the X-band wave into the S-band ports was excellent—in excess of 40 dB with consequently good, however not perfect, X-band radiation patterns across a wide band of 7 to 8.5 GHz. Some evidence of EH_{11} moding was present at X-band, as evidenced by small variations in the detailed shape of the various X-band radiation patterns.

Model II, described in Ref. 5, successfully extended the S-band bandwidth to 200 MHz (2100 to 2300 MHz), with slight but acceptable further degradation to X-band patterns. The height of the S-band waveguide injection slots necessary for bandwidth improvement was 12.7 mm (0.50 in.). Off-axis crosspolarization at X-band was maintained at −26 dB in the usual 45-deg pattern planes. Evidence of EH_{11} moding at X-band remained but was acceptable. Model II was designed for as high as possible S-band transmit power. Due to the necessity of a narrow height (3.12-mm or 0.123-in.) impedance matching section of waveguide within the S-band tuner–transformer portions of the combiner, a conservative S-band CW power limitation was specified (5 kW to each of four ports). It is possible, theoretically, that 100 kW (25 kW to each of four ports) could be reliably handled; a test would be required. Testing to 20 kW has been accomplished, but not to 100 kW. It is clear that 400 kW (100 kW to each port) would cause breakdown at normal gas temperature, pressure, and material (air or nitrogen). In order to provide for a higher power application, an alternative Model II tuner–transformer was designed (Ref. 5). This high-power tuner–transformer uses a 12.3-mm (0.484-in.) minimum height waveguide section and should allow 200 kW (50 kW to each of four ports) to be reliably handled. Full 400-kW power is a possibility but would require testing. However, the high-power version tuner–transformer provides only two narrower bands at S-band within which VSWR remains below 1.2:1—about 2090 to 2130 and 2260 to 2310 MHz. The lower power use of the Model II 3.12-mm height tuner–transformer provides continuous 2100- to 2300-MHz matching.

Finally, for the Model II development, Ref. 6 describes a four-function feedcone system that was built to demonstrate the Model II horn combiner and provided to the research and development station, DSS 13. The four functions are 20-kW S-band receive/transmit and X-band receive/transmit, the latter at 20 kW at 7.1 GHz. With requisite thermal design details added, X-band is expected to be capable of 400 kW. In one test of that feedcone (without the reflector antenna, with the feedcone on the ground simply radiating vertically to the sky), 20 kW was simultaneously transmitted at S- and X-bands. It is not known if the presence of reflector surfaces, with presumed non-linear junctions on a micro scale, would produce receive band intermodulation products arising from the dual-band uplinks, but the Model II combiner and horn was found capable of basic two-band transmission without breakdown. For the DSN 34-m high efficiency antennas (initially DSS 15 and 45), the common aperture horn/Model II combiner feed system was selected and successfully implemented. Thus, three stations are fitted with this equipment as of January 1986. DSS 65, the third 34-m high efficiency antenna, is scheduled for completion in 1987 and will similarly use the common aperture horn/Model III combiner.

III. Current Work

Two requirements guided development of a Model III combiner, intended to further increase S-band bandwidth to include the Highly Elliptical Earth Orbiter (HEO) band from 2025 to 2110 MHz, as well as to provide an assured 400-kW CW S-band uplink in both the full operational uplink band (2025 to 2120 MHz) and possible planetary radar band (near 2320 MHz). A key feature of the Model III combiner is the use of eight S-band waveguide injection slots, arranged in four pairs. The height of each slot was maintained at 12.7 mm (0.50 in.) because of the power handling requirement. Figures 1(a) and 1(b) show the Model II and Model III combiners for comparison. In both the Model I and Model II combiners, it was physically possible to arrange the X-band rejection filter structure on both the top and bottom surfaces of the single S-band slot. But in the paired-slot arrangement of the Model III combiner, only single-sided rejection filter geometry is possible due to the thin septum between the slots. Early concerns about obtaining sufficient X-band rejection (undesired coupling of X-band from the X-band horn throat terminal to any of the S-band slots) were allayed by initial testing, which showed sufficiency (approximately 50 dB). However, the combiner junction at that point was not yet matched for S-band, and later testing revealed that the rejection filtering was inadequate in the final configuration, including matching elements.

In all three combiner models, it was necessary to obtain S-band matching by a laborious and time-consuming process. It is a principle of wideband matching in dispersive lines (rectangular waveguide in this instance) that the best results are obtained by placing matching elements as close as possible to the initiating (undesired) reflection. In the case of the basic X/S combiner geometry used here, it was found necessary to actually enter the multi-mode radial line region (Figs. 2(a) and 2(b)) and to experimentally determine placement of matching elements (wide inductive posts in this instance) in order to achieve an acceptable bulk reduction of unmatched reflection. With the combiner roughly matched by experimental means (about 2:1 VSWR or -10 dB reflection) over a wide band, analytic methods are then applied in the 12.7-mm-high single-mode uniform transmission lines. Figure 3 shows a Smith chart presentation of the matched Model II combiner, referenced to full size WR430 waveguide impedance.

As stated above, when the Model III combiner was matched at S-band, the X-band rejection was affected. In some parts of the band, particularly near 7.2 GHz, the rejection was only -25 dB. At first it was expected that such leakage could be cured by additional and distant (from the combiner) simple rejection filtering, such as a (so-called) waffle iron lowpass type. In terms of the leakage level reaching sensitive components (the S-band maser for example), the above is true. But this approach fails to take into account where the reflected leakage finally exits the system, since it is not absorbed within. The conflict facing designers is how to achieve needed S-band wideband matching, a challenge in itself, while maintaining high (more than 40-dB rejection) performance of the X-band rejection filtering with the radial line (which is potentially a multi-mode region).

Radiation patterns were examined in both S- and X-bands of the "leaky" Model III combiner/horn system. As expected, S-band performed without incident, but X-band was found to be seriously degraded by the finite isolation of -25 dB in some parts of the 7- to 8.5-GHz band, particularly near 7.2 GHz (the needed X-band uplink). X-band radiation patterns are variable and generally unpredictable; they are dependent upon what is connected to the S-band lines (uniform WR430 with a 180-deg tee junction or 90-deg hybrid junctions[2]). In a worst-case setup, reflecting the undesired X-band leakage back into the horn in an adverse phase (a possible situation in a full-feed system), the 7.2-GHz radiation pattern is seriously affected and is predicted to cause approximately 1 dB of final reflector antenna system gain loss. Figure 4(a) shows expected normal (high X-band isolation) radiation pattern near 7.19 GHz. Figure 4(b) shows a finite isolation pattern causing 1 dB of final reflector antenna system gain loss. The loss components are illumination (-0.55 dB), cross-polarization (-0.15 dB), and phase (-0.34 dB). Overall, a typical high performance (shaped) dual-reflector antenna system fed with the affected radiation pattern shown in Fig. 4(b) would produce only 57% aperture efficiency (down from a normal 72%). This is considered grossly unacceptable performance for the DSN, and work on this effort has been terminated, at least temporarily. Future work is likely, and the authors hope this summary will provide a valuable starting place.

---

[2]Implementation of these junctions is dependent upon how the S-band polarization network is configured.

References

1. Williams, W. F., "A Prototype DSN X-S Band Feed: DSS 13 First Application Status," *DSN Progress Report 42-44,* pp. 98-103, Jet Propulsion Laboratory, Pasadena, CA, Apr. 15, 1978.

2. Williams, W. F., "A Prototype DSN X-S Band Feed: DSS 13 Application Status (Second Report)," *DSN Progress Report 42-47,* pp. 39-50, Jet Propulsion Laboratory, Pasadena, CA, July 15, 1978.

3. Williams, W. F., et al., "A Prototype DSN X-S Band Feed: DSS 13 Application Status (Third Report)," *DSN Progress Report 42-52,* pp. 51-60, Jet Propulsion Laboratory, CA, July 15, 1979.

4. Williams, W. F., and Withington, J. R., "A Common Aperture S- and X-Band Feed for the Deep Space Network," *Proceedings of the 1979 Antenna Applications Symposium,* University of Illinois, Urbana-Champaign, IL, Sept. 26-28, 1979.

5. Williams, W. F., and Reilly, H., "A Prototype DSN X/S-Band Feed: DSS 13 Application Status (Fourth Report)," *DSN Progress Report 42-60,* pp. 77-88, Jet Propulsion Laboratory, Pasadena, CA, Nov. 15, 1980.

6. Withington, J. R., and Williams, W. F., "A Common-Aperture S- and X-Band Four Function Feedcone," *Proceedings of the 1981 Antenna Applications Symposium,* University of Illinois, Allernton Park, Monticello, IL, Sept. 1981.

7. "Dual Frequency Dichroic Feed Performance," *26th Meeting of Avionics Panel,* AGARD, Munich, Germany, 26-30 November, 1973, AGARD Conference Proceedings No. 139, NATO, June 1974.
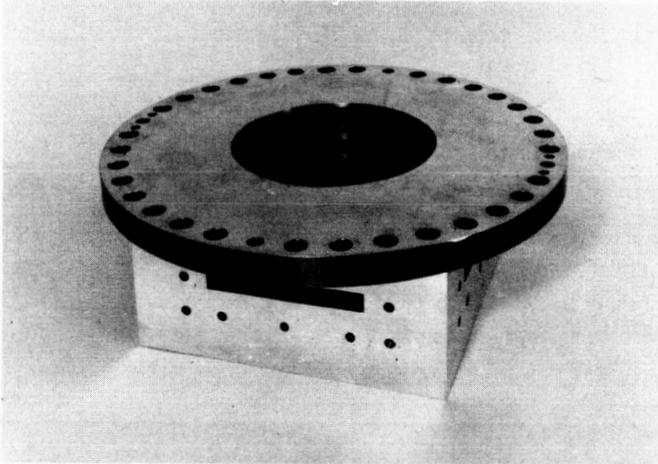
Fig. 1(a). Model II X/S combiner
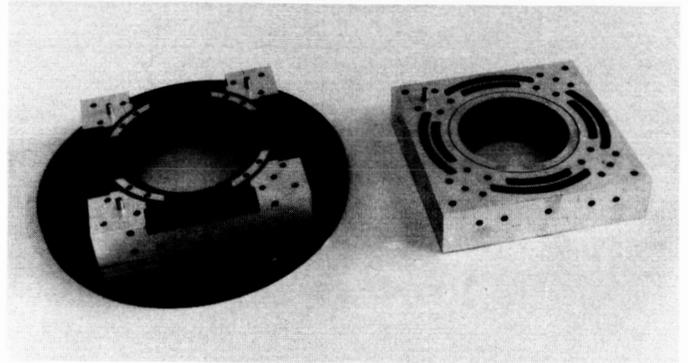
Fig. 2(a). Model II combiner showing radial line region

Fig. 1(b). Model III X/S combiner
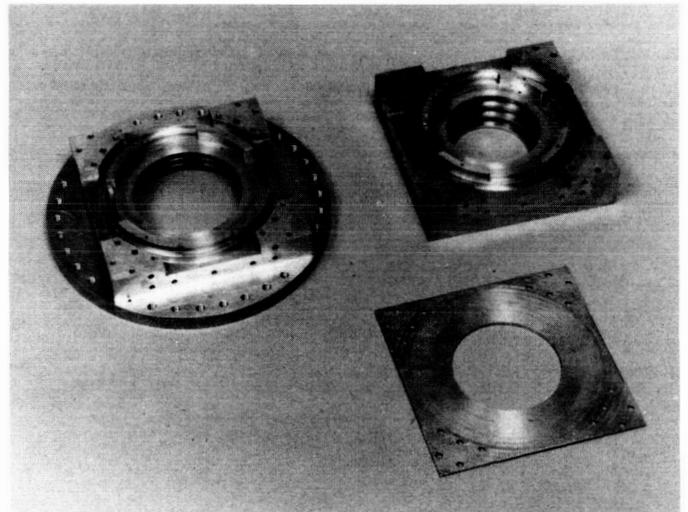
Fig. 2(b). Model III combiner showing radial line region

No. GC 19510

NAME	TITLE	DWG. NO.
SMITH CHART Form 5301-7561-NE	GENERAL RADIO COMPANY, WEST CONCORD, MASSACHUSETTS	DATE

IMPEDANCE OR ADMITTANCE COORDINATES
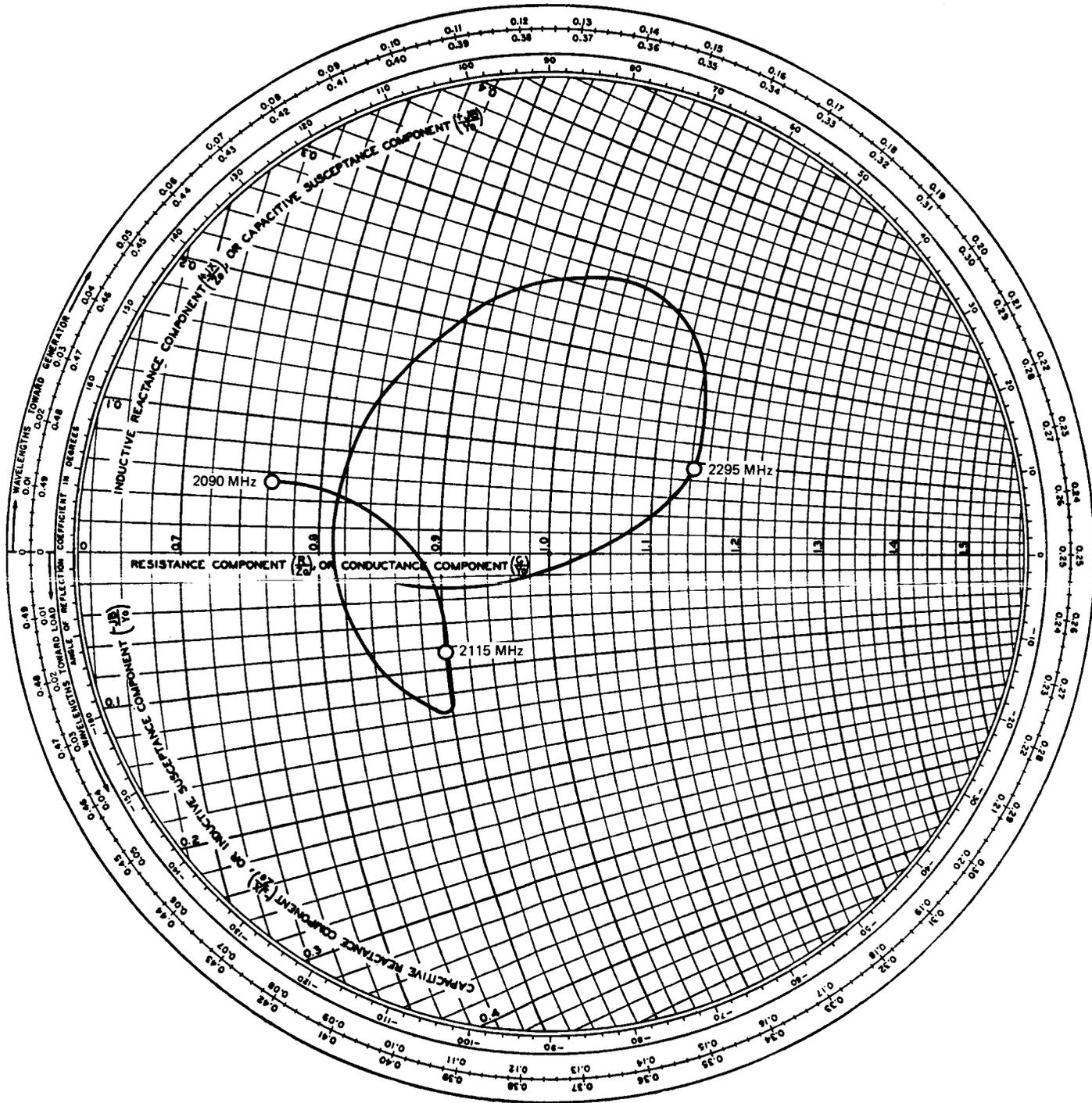
Fig. 3. Smith chart presentation of Model II X/S combiner referenced to WR430 admittance
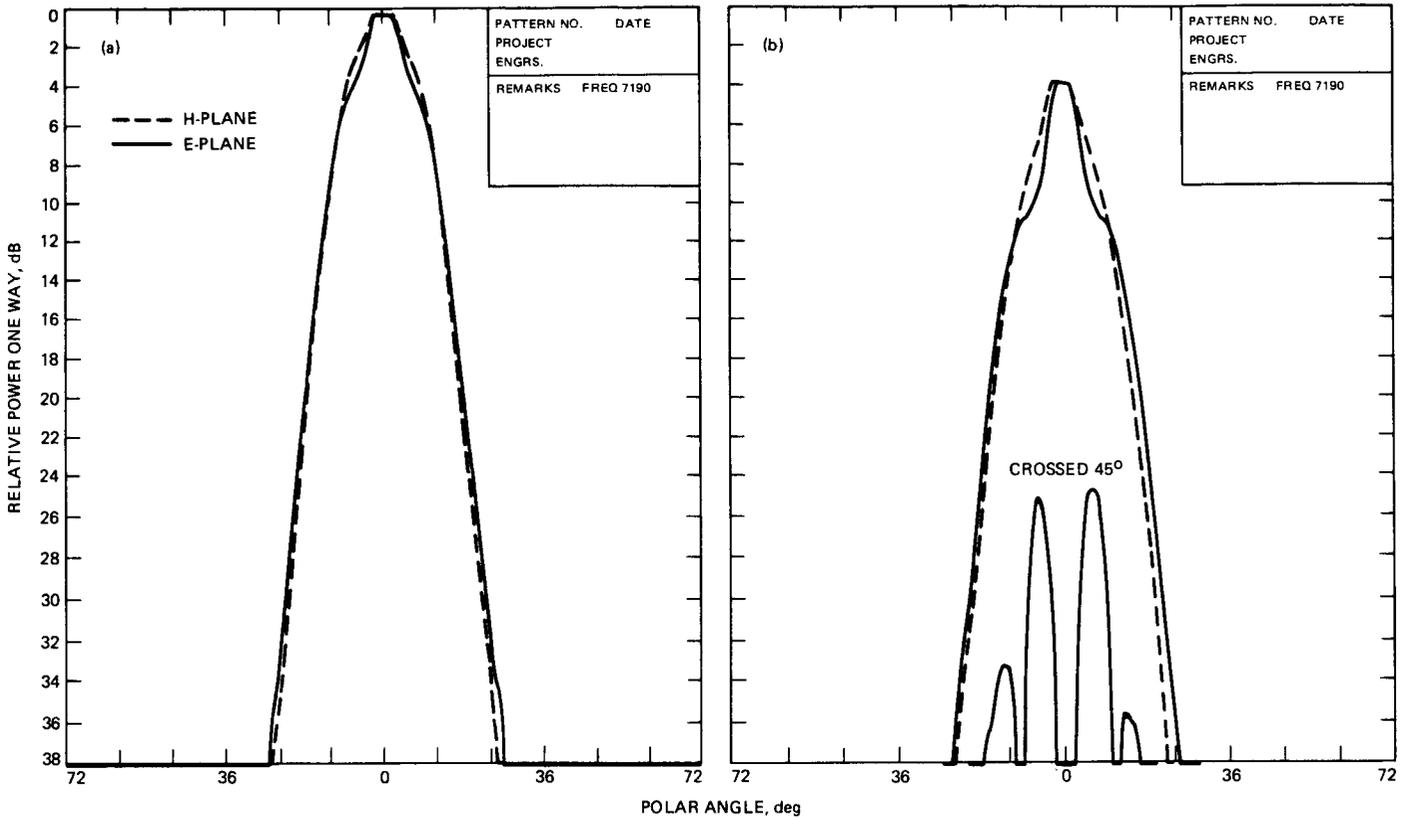
Fig. 4. Radiation patterns of feedhorns in X-band at 7.19 GHz: (a) Model II E- and H-plane patterns, (b) Model III E- and H-plane patterns

Stability Investigation of the Quadripod Structure for the NASA/JPL 70-Meter Antenna

C. T. Chian, J. J. Cucchissi, and R. Levy
Ground Antennas and Facilities Engineering Section

A new, slim-profiled, low-blockage quadripod structure was designed to support the 7.7-m-diameter subreflector for the 70-m antenna. Detailed analyses of quadripod structural stability (in frequency and buckling) are presented. The results indicate that the new quadripod design has an adequate safety margin for buckling, and its lowest natural frequency is above the control system bandwidth. The analytical design frequencies were confirmed by actual field measurements made in Spain in February 1986.

I. Introduction

The upgrade of the three NASA/JPL 64-m diameter antennas will provide a needed increase in Earth-based space communication capability at all three Deep Space Communications Complexes: Goldstone, California (DSS 14); Canberra, Australia (DSS 43); and Madrid, Spain (DSS 63). In addition to the increase of the antenna aperture area from 64 m to 70 m, a number of significant improvements in the quadripod, surface panels, subreflector positioner, and microwave aspects are included in the design. The upgrade objective is to increase the radio-frequency (RF) gain/noise temperature (G/T) by about 1.9 dB at X-band (8.45 GHz).

As part of the upgrade effort, a new, high-precision 7.7-m (25.4-ft)-diameter subreflector and positioning mechanism are needed. Consequently, an entirely new quadripod structure is required to support the subreflector. The new quadripod design particularly emphasizes reduced RF blockage, which is achieved by means of a narrow cross-sectional profile of the legs. The profile adopted provides about 0.32 dB of

gain improvement in comparison with the existing 64-meter design (Ref. 1). This report addresses the stability analyses performed on the new quadripod design to ensure that it has an adequate safety margin for buckling and that the minimum natural frequency is compatible with control system requirements.

After construction, full-scale vibration measurements were performed at the fabricator's plant on the completed and assembled structure with dummy weights to simulate the subreflector and other equipment loads.

II. Design

The quadripod assembly is a tabular space-frame steel structure with four trapezoidally shaped legs connected to another large space frame at the apex, as shown in Figs. 1 and 2. The four legs are supported at the corner points of the rectangular truss system of the main reflector structure as shown in Fig. 2. The final slim profile leg cross-section enve-

lope selected is shown in Fig. 3. Also, the quadripod will be used occasionally for hoisting the cassegrain feed cones, the subreflector, or other heavy equipment which may be removed and reinstalled.

The finite element model of the 70-m quadripod truss structure is a pin-joined frame (3 translational degrees of freedom per node) comprising 156 nodes, 445 axial bars, and 28 membrane plates. The JPL/IDEAS (Iterative Design of Antenna Structures) computer program was used for analysis and design (Ref. 2). The program employs the optimality criterion to minimize the structural weight (objective function) with a constraint placed on the lowest natural frequency. A subsequent analysis of the 70-m model, accounting for bending and torsional stiffness at the joints (6 degrees of freedom per node) using NASTRAN (Ref. 3), showed only a small increase in the torsional natural frequency (Ref. 1). Outrigger braces (Fig. 2) were added thereafter to increase the lowest natural frequency.

Due to the slimness of the quadripod legs, the following requirements had to be considered:

(1) *Dynamic stability*. The original 64-m antenna quadripod, with a minimum natural frequency of 1.22 Hz, presented no control system stability problems. Therefore, it was recommended that the minimum natural frequency of the new 70-m antenna quadripod must be 1.22 Hz. There are several distinct types of vibration modes characteristic of quadripod structures. Torsional modes can be excited at near-zenith antenna elevation by the azimuth drive. Lateral vibration cantilever modes can be excited by the azimuth drive at low elevation angles, and cantilever pitch modes can be excited by the elevation drive at any elevation angle, as shown in Fig. 4.

The torsional mode had the lowest frequency, so this frequency was selected as the primary design constraint. It was also found that this frequency could be significantly increased by adding outrigger braces near the quadripod base. These braces have an important stiffening effect and provide an insignificant increase in blockage. However, the braces are attached to an elastic antenna structure, and the compliance of this structure could reduce the outrigger contribution to the stiffness of the quadripod. Therefore, the consequences of partially effective outrigger braces on natural frequency were also studied.

(2) *Static buckling stability*. The occasional use of the quadripod as a derrick required a check on the possibility of buckling instability. A factor of safety of at least 1.5 was recommended. The smallest eigenvalue

found from a structural buckling analysis is equivalent to this factor of safety. Since the IDEAS program that was used for design and natural frequency analyses does not perform buckling analysis, the new quadripod design was optimized for the frequency requirement using IDEAS and then analyzed for buckling using NASTRAN (NASA Structural Analysis Program). NASTRAN was used to determine the buckling loads of the natural frequency-constrained quadripod design. Two versions of the NASTRAN program were used because of possible different finite element formulations: the NASTRAN-COSMIC (NASA's Computer Software Management and Information Center) (Ref. 4) and the proprietary NASTRAN-MSC version (MacNeal-Schwendler Corporation) (Ref. 3). The two versions were used both for the buckling and natural frequency analyses, and the results obtained were compared.

III. Natural Frequency Analysis and Results

For natural frequency and mode shape analysis, the IDEAS program uses the Simultaneous Iteration method (Ref. 5), which is an iterative extension of Guyan's one-step solution. The NASTRAN programs in this study used the conventional Inverse Power Method.

Table 1 compares the first three natural frequencies of two pin-joined quadripod models; one with outrigger braces and the other without braces. All the values of Table 1 (except mode 1 without outriggers) exceed the goal of 1.22 Hz. Despite their effect on frequency, the braces do not significantly alter the characterization of the mode shapes of interest: the lowest mode is torsional and the next lowest are lateral and pitch cantilever modes. This table makes it evident that the outriggers are effective and approximately double the lowest frequency. Also included in Table 1 are the field measurements made on the assembled quadripod (Ref. 6).

To study the consequences of varying degrees of outrigger fixity caused by partially effective braces, the axial stiffness of the outriggers was parameterized, and the resulting natural frequencies were computed and plotted in Fig. 5. This approach is equivalent to reducing the "stiffness" of the support points. There is a relatively small change in frequency as long as the stiffness is at least 50% of the maximum. This fortuitous condition results from the requirement that the quadripod supports the hoisting loads.

IV. Results of Buckling Analysis

The Rigid Format No. 5 of the NASTRAN program was used to perform the buckling analysis. The results of the

buckling analysis for the quadripod pin-joined model are presented in Table 2. Both the COSMIC and MSC versions of the NASTRAN program were used and compared. Four antenna configurations, each subject to the maximum loads anticipated to be hoisted when employing the quadripod as a derrick, in addition to the quadripod weight, were considered in the quadripod buckling analysis:

(1) Zenith look with outriggers.

(2) Zenith look without outriggers.

(3) Horizon look with outriggers.

(4) Horizon look without outriggers.

Table 2 shows that similar to effects on natural frequency the outriggers tend to at least double the buckling load capability.

V. Finite Element Plate Stiffness Representation

Comparisons of the COSMIC-NASTRAN and MSC-NASTRAN results on the quadripod model in Table 1 show that the plate element CQDMEM2 in the COSMIC version gives a different stiffness matrix representation compared with the CQUAD4 plate element in the MSC version, or with the IDEAS plate element CQDMEM. The lowest natural frequency of the quadripod. for instance. was found to be 1.54 Hz for the COSMIC model, while the MSC and IDEAS models gave

1.30 Hz. Table 3 compares the quadripod natural frequency results by the three computer programs: IDEAS, NASTRAN-COSMIC, and NASTRAN-MSC.

A parametric study was conducted to readjust the moduli of elasticity of the COSMIC CQDMEM2 plate elements to produce results similar to those of the MSC elements. Comparison of the results of the stiffness parameterization study is shown in Table 4 for the quadripod natural frequency analysis and in Table 5 for the quadripod buckling analysis. The plate element used in the NASTRAN-COSMIC employs the constant stress formulation, while the elements used in the NASTRAN-MSC or IDEAS permit a stress variation. As a result, the COSMIC element generates a stiffer structure than the other elements.

VI. Summary

A natural frequency and structural stability study was conducted for the 70-m antenna quadripod. The quadripod was found to be adequate in natural frequency and stable in buckling when the outrigger braces were included. One computer program used in the investigation was found to give an over-estimate of the stiffness. In order to correct the excessive stiffness, a parametric study was conducted to derive empirical coefficients to adjust the plate stiffness for future use of this program. The predicted values of natural frequency were shown to be closely consistent with actual full-scale field tests.

References

1. Cucchissi, J. J., "A New 70-Meter Antenna Quadripod With Reduced RF Blockage," *TDA Progress Report 42-82*, pp. 24-30, Jet Propulsion Laboratory, Pasadena, CA, Aug. 15, 1985.

2. Levy, R., and Chai, K., "Implementation of Natural Frequency Analysis and Optimality Criterion Design," *Computers and Structures*, Vol. 10, pp. 277-282, Pergamon Press Ltd., Great Britain, 1979.

3. McCormick, C. W., Editor, *MSC/NASTRAN User's Manual*, MSR 39, The MacNeal-Schwendler Corporation, 1983.

4. *The NASTRAN Theoretical Manual (level 15.5)*, NASA SP-221 (06), Scientific and Technical Information Office, NASA, Wash. DC, Jan. 1981.

5. Levy, R., *Guyan Reduction Solutions Recycled for Improved Accuracy*, NASA TMX-2378, Vol. 1, pp. 201-220, Sept. 1971.

6. "Vibration Measurements of Quadripod and Apex Structure for 70M Antenna DSS-63," Report No. J-K 5391, Kinemetrics, Inc., Pasadena, CA, May 1986.

Table 1. Comparison of natural frequencies for 70-m quadripod

Mode	With outrigger braces		Without outrigger braces		Predominant mode shape
	Analytical	Field measurement	Analytical	Field measurement	
1	1.302	1.27	0.637	0.70	Torsion
2	1.967	1.76	1.293	1.36	Lateral cantilever
3	2.720	2.62	1.555	1.74	Pitch cantilever

Frequency, Hz

Table 2. The quadripod buckling analysis results

Case	Antenna configuration	NASTRAN COSMIC version	NASTRAN MSC version
1	Zenith look, with outriggers	22.81	12.27
2	Zenith look, without outriggers	9.23	(not run)
3	Horizon look, with outriggers	12.87	3.77
4	Horizon look, without outriggers	4.12	1.66

Table 3. Comparison of lowest natural frequencies of the 70-m quadripod with outrigger braces

Mode	Frequency, Hz		
	Computer program		
	IDEAS	MSC-NASTRAN	COSMIC-NASTRAN
	Plate element		
	CQDMEM	CQUAD4	CQDMEM2
1	1.302	1.304	1.542
2	1.967	1.967	2.089
3	2.720	2.720	3.216

Table 4. Comparison of the plate element stiffness and the lowest quadripod natural frequencies

Program	Plate element	Young's modulus, N/m^2 (psi)	Shear modulus, N/m^2 (psi)	Mode 1 min. freq., Hz	Mode 2 min. freq., Hz
NASTRAN-MSC	CQUAD4	20.0×10^{10} (29.0×10^6)	8.3×10^{10} (12.0×10^6)	1.304	2.720
NASTRAN-COSMIC	CQDMEM2	14.5×10^{10} (21.0×10^6)	6.0×10^{10} (8.7×10^6)	1.312	2.737

Table 5. Comparison of the plate element stiffness and the smallest eigenvalues for the quadripod buckling analysis

Program	Plate element	Young's modulus, N/m^2 (psi)	Shear modulus, N/m^2 (psi)	λ_{min}
(a) Zenith Look Antenna Configuration, with Outrigger Braces:				
NASTRAN-MSC	CQUAD4	20.0×10^{10} (29.0×10^6)	8.3×10^{10} (12.0×10^6)	12.27
NASTRAN-COSMIC	CQDMEM2	10.7×10^{10} (15.5×10^6)	4.4×10^{10} (6.4×10^6)	12.30
(b) Horizon Look Antenna Configuration, with Outrigger Braces:				
NASTRAN-MSC	CQUAD4	20.0×10^{10} (29.0×10^6)	8.3×10^{10} (12.0×10^6)	3.77
NASTRAN-COSMIC	CQDMEM2	5.2×10^{10} (7.5×10^6)	2.2×10^{10} (3.2×10^6)	3.97

REFLECTOR SURFACE

SUBREFLECTOR

REFLECTOR
STRUCTURE
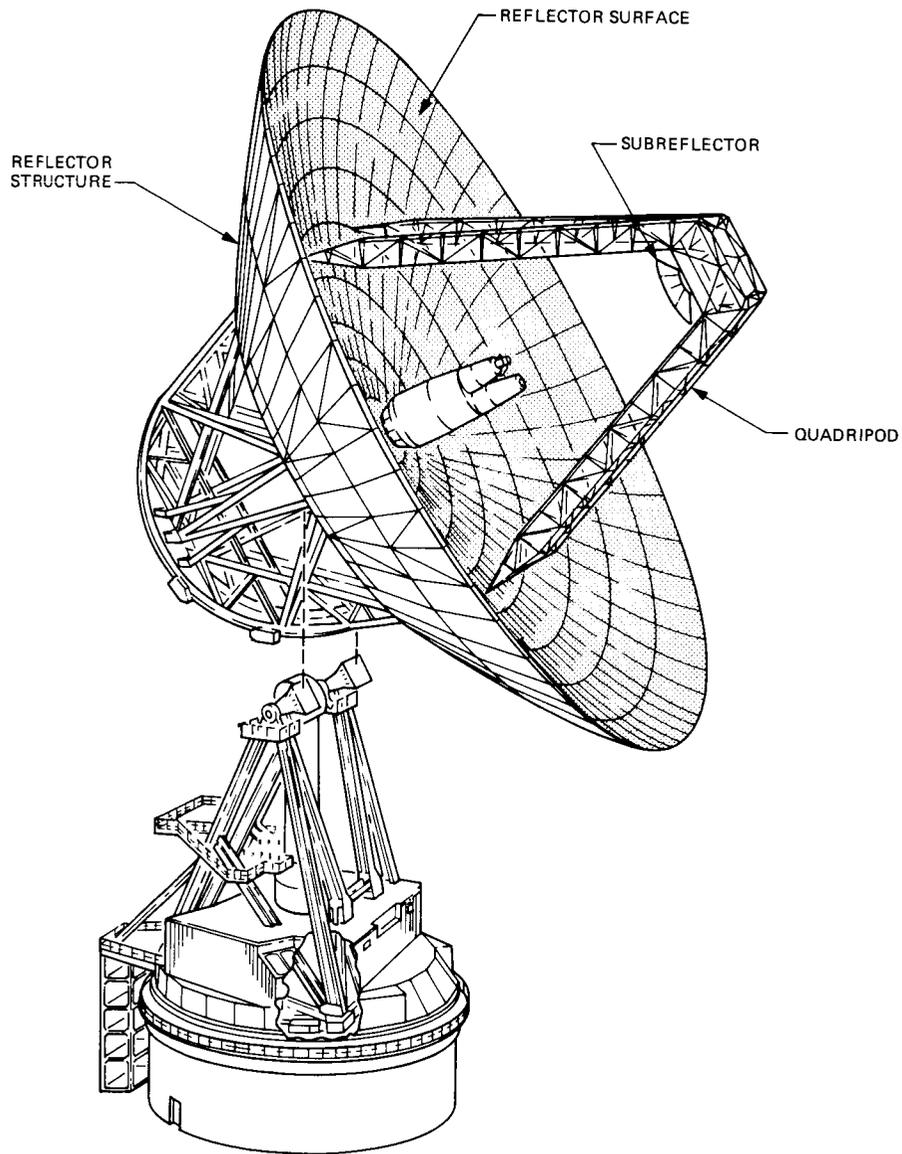
QUADRIPOD

Fig. 1. Antenna quadripod and reflector system

Fig. 3. Quadripod cross-sectional profile
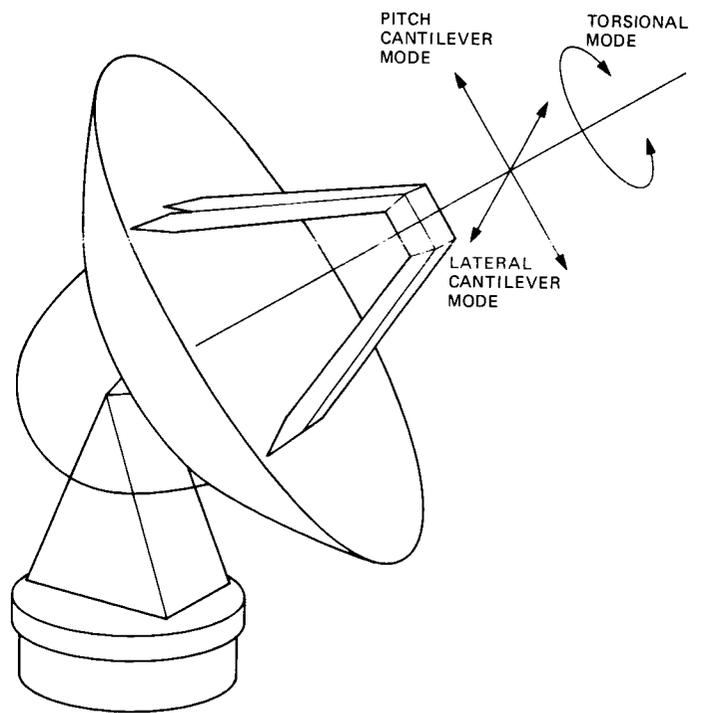
Fig. 4. Quadripod vibration modes

Fig. 2. Plan view of 70-m quadripod with outriggers
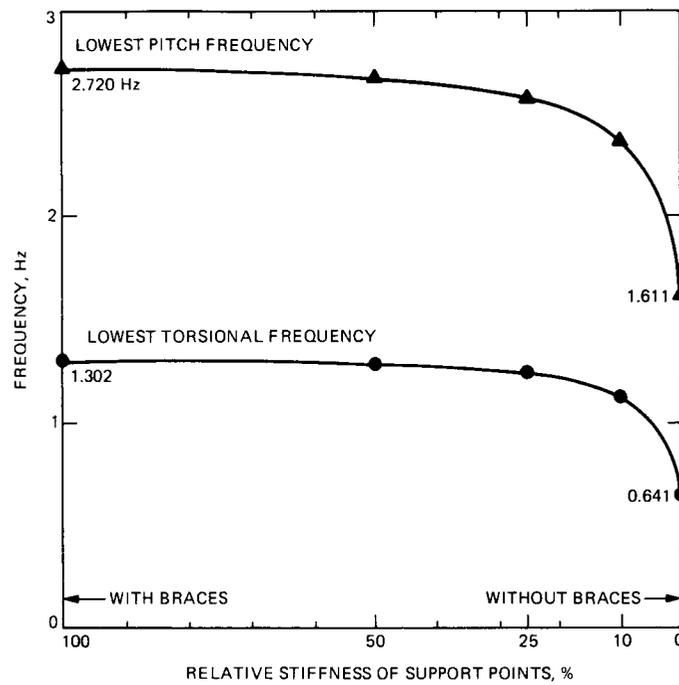
Fig. 5. Effect of outrigger braces on natural frequency

Appendix A

Differential Stiffness Matrices for Geometric Nonlinear Problems

For completeness of the report, the following material, which is extracted from Ref. 3, is included to describe formulation of the differential stiffness matrix that is used in the buckling analysis. The differential stiffness approximation uses terms up to quadratic in the strain-displacement relationship. The linear elastic solution and the differential stiffness solution are the first two iterations in the geometric nonlinear algorithm, which is an iterative technique that utilizes a modified Newton–Raphson method.

The approach requires a "Displaced Element Coordinate System" to be constructed for each element, which follows and rotates with the element as the model deforms. In the displaced element coordinate system, the distortions are small, and linear elastic theory can be used. Element forces in the displaced element coordinate system are computed by simply premultiplying the displacements by the elastic (small motion) stiffness matrix. The incremental stiffness matrix, when expressed in the displaced element coordinate system, is the sum of the elastic and differential stiffness matrices.

The term "differential stiffness" applies to linear terms in the equations of motion of an elastic body that arise from a simultaneous consideration of large, nonlinear motions and the applied loads.

The approach to the theory of differential stiffness is based on Lagrange's equations for the motion of a system with a finite number of degrees of freedom. Consider a system with a finite number of degrees of freedom, q_r; with a set of springs whose potential energy is V; and with a set of loads, P_a, applied to displacements u_a. The equations of motion for the system may be written

$$\frac{\partial V}{\partial q_r} = Q_r, \quad r = 1,2,3,\ldots,n \tag{A-1}$$

where the generalized force Q_r is given by

$$Q_r = \frac{\partial W}{\partial q_r} = \sum_a \frac{\partial u_a}{\partial q_r} P_a \tag{A-2}$$

W is the work done by the external forces. It is assumed in the theory of differential stiffness that the potential energy of differential stiffness is a quadratic function of the degrees of freedom, i.e.,

$$V = \frac{1}{2} \sum_{i,j} a_{ij} q_i q_j \tag{A-3}$$

but that the partial derivatives, $\partial u_a / \partial q_r$, are not necessarily constants.

The Lagrangian discrete element approach can be applied to a general elastic body, if it be imagined that the body is made up of infinitesimal cubes, each of which is joined to its six neighbors by a universal joint at the midpoint of each face. For a given static loading on the body, the stress distribution is computed throughout the body, ignoring differential stiffness effects in the process. This internal stress distribution is taken as the equivalent loading, and is applied to each cube in turn to determine the differential stiffness for the cube.

The work done by the static loads is computed for general motion of the degrees of freedom using Eq. (A-2). The terms in the differential stiffness matrix for the cube are then computed from

$$K_{rs} = -\frac{\partial Q_r}{\partial q_s} = -\frac{\partial^2 W}{\partial q_s \partial q_r} \tag{A-4}$$

The total work done by all components of force on a cube of volume Δv is

$$\Delta W = -\frac{\Delta v}{2} [\omega_x^2 (\sigma_y + \sigma_z) + \omega_y^2 (\sigma_z + \sigma_x) + \omega_z^2 (\sigma_x + \sigma_y)$$

$$-2 \omega_x \omega_y \tau_{xy} - 2 \omega_y \omega_z \tau_{yz} - 2 \omega_z \omega_x \tau_{zx}] \tag{A-5}$$

where ω_x, ω_y, and ω_z are rotations about the x, y, and z axes, respectively. No work is done on the cube during translation because the forces acting on the cube are in equilibrium.

The matrix of differential stiffness coefficients for a cube of volume Δv is written from Eq. (A-4) as

$$\Delta [K^d] = \Delta v \begin{bmatrix} \sigma_y + \sigma_z & -\tau_{xy} & -\tau_{zx} \\ -\tau_{xy} & \sigma_z + \sigma_x & -\tau_{yz} \\ -\tau_{zx} & -\tau_{yz} & \sigma_x + \sigma_y \end{bmatrix} \tag{A-6}$$

The above general result is applied to evaluate the differential stiffness matrices for the quadripod structural elements.

Appendix B

Buckling Analysis Procedure

The formulation of the quadripod linear static response problem by the displacement method is described by the matrix equation

$$[K] \{u\} = \{P\} \qquad (B-1)$$

where $[K]$ is the stiffness matrix, $\{u\}$ is the displacement vector, and $\{P\}$ is the load vector.

The steps for solving a quadripod buckling problem are listed as follows:

(1) Solve the linear static response problem Eq. (B-1) for the quadripod structure in the absence of differential stiffness, and compute the internal forces in elements.

(2) Using the results of Step (1), calculate the differential stiffness matrices for individual elements, and apply the standard reduction procedures (constraints and partitioning) to form the differential stiffness matrix $[K^d]$ in final form.

(3) Replace the load vector $\{P\}$ by $-\lambda [K^d] \{u\}$, and find eigenvalues and eigenvectors for

$$[K + \lambda K^d] \{u\} = 0 \qquad (B-2)$$

The eigenvalues, λ, are the load level factors by which the applied static loading is multiplied to produce buckling:

$$P_{cr} = \lambda P \qquad (B-3)$$

P_{cr} is the critical load for buckling, and P is the applied load.

The eigenvalues, λ_i, and the corresponding eigenvectors, $\{\phi_i\}$, are extracted by the Real Eigenvalue Analysis Module. The criterion for the qaudripod structure to be statically stable (free from buckling) is, therefore:

$$\lambda_i > 1 \qquad (B-4)$$

Array Feed Synthesis for Correction of Reflector Distortion and Vernier Beamsteering

S. J. Blank
Center for Technology
Holon, Israel

W. A. Imbriale
Radio Frequency and Microwave Subsystems Section

This article describes an algorithmic procedure for the synthesis of planar array feeds for paraboloidal reflectors to simultaneously provide electronic correction of systematic reflector surface distortions as well as a vernier electronic beamsteering capability. Simple "rules of thumb" for the optimum choice of planar array feed configuration (i.e., number and type of elements) are derived from a parametric study made using the synthesis procedure described herein. A number of f/D ratios and distortion models were examined that are typical of large paraboloidal reflectors. Numerical results are presented showing that, for the range of distortion models considered, good on-axis gain restoration can be achieved with as few as seven elements. For beamsteering to ±1 beamwidth (BW), 19 elements are required. For arrays with either 7 or 19 elements, the results indicate that the use of high-aperture-efficiency elements (e.g., disk-on-rod and short backfire) in the array yields higher system gain than can be obtained with elements having lower aperture efficiency (e.g., open-ended waveguides). With 37 elements, excellent gain and beamsteering performance to ±1.5 BW are obtained independent of the assumed effective aperture of the array element.

An approximate expression is derived for the focal-plane field distribution of the distorted reflector. Contour plots of the focal-plane fields are also presented for various distortion and beam scan angle cases.

The results obtained show the effectiveness of the array feed approach.

I. Introduction

This article describes an algorithmic procedure for the synthesis of planar array feeds to simultaneously provide electronic correction of paraboloidal reflector antenna distortion as well as a vernier electronic beamsteering capability. This technique would be useful for correcting the gain loss and beam squint errors that are caused by systematic reflector distortions such as those induced by gravity on large ground station antennas. These distortions are large-scale and can be measured and/or predicted with reasonable accuracy by analytic methods. This study is not concerned with random surface errors, which are usually relatively small and must be analyzed statistically.

Rudge and Davies (Ref. 1) discussed the array feed for distortion correction and proposed the use of the Butler matrix as an ingenious means of providing adaptive excitation of the feed array. Their analytic and quantitative results, however, are limited to cylindrical reflectors having one-dimensional distortion profiles requiring only linear array feeds. Amitay and Zucker (Ref. 2) analyzed the use of planar array feeds for aberration correction in spherical reflectors. Their synthesis procedure, however, relies heavily on the circular symmetry of the feed-plane field distribution that exists in this case.

The synthesis algorithm and computer code used in this study are a modification of that developed by Imbriale et al. (Ref. 3). The code includes a numerical search routine that could be used to optimize general performance criteria including gain, sidelobe levels, and beamshape. In this article, however, only gain maximization is considered. Preliminary results showing feasibility of the technique have previously been reported by Blank and Imbriale (Ref. 4). For a given feed configuration, reflector f/D ratio, and distortion profile, the algorithm finds the optimum values of the individual feed excitations to maximize the reflector antenna gain in some desired direction. In doing this, the algorithm uses the Jacobi-Bessel technique (Ref. 5) to calculate the secondary fields of the reflector antenna having an arbitrary (i.e., distorted) shape and an illumination resulting from the displaced sources that comprise the array feed. The optimum choice of the array feed configuration (i.e., the number and type of element radiators) is determined by a parametric study made on a range of f/D ratios and reflector distortion profiles. A qualitative discussion of the choice of these feed parameters is given in the section on numerical results. Questions of system cost, signal-to-noise (S/N) ratio, and hardware availability are not explicitly considered.

It is assumed that the projected aperture of the reflector is sufficiently large (i.e., $D \geqslant 100\lambda$) so that the effects of feed blockage on system gain can be neglected. This study is limited to fixed planar feed arrays of identical elements located in the focal plane.

Contour plots of focal plane field distributions are presented to give graphical insight into the effects of reflector distortion. An analytic expression is also given for the focal plane field distribution based on the scalar diffraction equation. The expression is valid for small magnitudes of reflector distortion and large f/D ratios.

Cross-polarization effects are not considered in this study, although in principle, an array feed could also be used to synthesize some desired polarization distribution, by, for example, varying the orientation of the feed elements or by employing coupling effects. The array feed approach coupled with numerical search opens up the possibility of adaptive closed-loop optimization of system performance.

II. Reflector Feed Configuration

The reflector/feed configuration addressed is the case of single reflector systems[1] having circular apertures of diameter D and maximum subtended angle $\hat{\psi}$ (Fig. 1). The feed consists of either a single circular element (taken as a reference for gain calculations) or a triangular grid array of 7, 19, or 37 elements (Fig. 2). The elements of the array are presumed to be identical and to have circularly symmetric, unidirectional radiation patterns of the form

$$f(\psi) = \cos^q(\psi) \tag{1}$$

The assumption of circularly symmetric element radiation patterns is made for the purpose of simplifying the presentation of results. The technique being described can also be used with more general types of element radiation patterns.

The minimum center-to-center spacing d_e (footnote 2), between the elements of the array, is related to q, the exponent in Eq. (1), by

$$d_e = \sqrt{\frac{q + 0.5}{b}} \tag{2}$$

where the value of b depends on the type of element used (Ref. 7). For circular waveguide elements, $b = 2.07$. For radiators with a higher effective aperture, such as disk-on-rod elements, a representative value of $b = 2.47$ is used.

For given values of b and q, spacing the elements greater than d_e would imply gain loss, since the "open" area thereby left between elements would not be available to capture incident energy. Therefore, in what follows, the element spacings are set equal to d_e, not more nor less. The maximum diameter, d_f, of such a "well-packed" array feed cluster is given by

$$d_f = K \cdot d_e \tag{3}$$

where $K = 1, 2, 3, 7$, corresponding to the number of feed elements $N = 1, 7, 19, 37$, respectively.

---

[1]Cassegrain systems can be considered in terms of equivalent single reflector systems having the appropriate f/D ratio (Ref. 6).

[2]Alternatively, d_e can be considered the minimum element diameter.

III. Calculation of Gain

In what follows, the gain of the reflector antenna $G(\theta_0, \phi_0)$ is taken to be synonymous with its directivity so that

$$G(\theta_0, \phi_0) = \frac{|E(\theta_0, \phi_0)|^2}{P_T/4\pi} \tag{4}$$

where $E(\theta_0, \phi_0)$ is the value of the field of the secondary pattern at the peak of its mainlobe, and P_T, the total power radiated by the antenna as a whole, is equal to that power radiated by the feed array.

Referring to Fig. 1, we have

$$E(\theta_0, \phi_0) = \sum_{i=1}^{N} a_i F_i \exp[jkr_i \sin\theta_0 \cos(\phi_0 - \phi_i')] \tag{5}$$

where

N = number of elements in the array

k = propagation constant

a_i = complex excitation of the ith element in the feed array

r_i, ϕ_i' = polar coordinates of the ith element in the focal plane, i.e., $x_i = r_i \cos\phi_i'$, $y_i = r_i \sin\phi_i'$ (r_i is normalized to wavelength)

F_i = complex vector field received by the ith array element as a result of a plane wave incident on the reflector aperture from the direction (θ_0, ϕ_0), and

$$P_T = \int_0^{2\pi} \int_0^{\pi/2} \left| f(\theta, \phi) \sum_{i=1}^{N} a_i \exp[jkr_i \sin\cos(\phi - \phi_i')] \right|^2$$

$$\times \sin\theta \, d\theta \, d\phi \tag{6}$$

where $f(\theta, \phi)$ is the field pattern of the element used in the array. (All elements are assumed to be identical and mutual coupling effects are neglected.)

By reciprocity, the $\{F_i\}$ can also be viewed as the secondary fields resulting from the illumination of the reflector by the offset elements that comprise the array feed. The $\{F_i\}$ are calculated using the Jacobi-Bessel method (Ref. 5).

The gain of the reflector is maximized in a given direction by setting the feed element excitations, $\{a_i\}$, to be the conju-gate values of the $\{F_i\}$ fields received from that direction. In those cases where the performance criteria are other than simply maximizing gain in a given direction, or when it is not practical from a hardware point of view to take the conjugate values of the fields, then it is necessary to use a numerical search algorithm such as the Rosenbrock method (Ref. 8) to find the feed excitations that optimize performance.

The ideal maximum gain of the reflector is $(\pi D/\lambda)^2$. However, it is perhaps more meaningful to reference the gains achieved with an array feed to that obtained with a single element feed having a pattern giving, say, a -10 dB aperture taper at the maximum subtended angle $\hat{\psi}$. In this case, the value of the exponent for a single element feed, q_1, can be found from the expression

$$\frac{(\cos\hat{\psi})^{q_1}}{(f + D^2/16f)} = 0.316 \tag{7}^3$$

which, after suitable approximations, reduces to

$$q_1 \cong 9.2 \, (f/D)^2 - 0.5 + \frac{1}{64 \, (f/D)^2}$$

$$\cong 9.2 \, (f/D)^2 - 0.5 \text{ for } f/D \geqslant 1 \tag{8a}$$

Together with Eq. (2), this gives

$$d_{e_1} \cong \sqrt{\frac{9.2}{b}} \, (f/D) \tag{8b}$$

For an arbitrary value t of desired aperture field taper,

$$q_1 \cong \frac{-8 \ln(t)}{b} \, (f/D)^2 - 0.5 \tag{8c}$$

which, with Eq. (2), gives

$$d_{e_1} \cong \sqrt{\frac{-8 \ln(t)}{b}} \, (f/D) \tag{8d}$$

IV. The Algorithmic Procedure

The method being described here can be used to optimize general performance criteria such as gain, sidelobe levels, and beamshape. In this study, however, only gain maximization

---

[3] $\hat{\psi} = 2\tan^{-1}(1/4f/D)$; $\cos\hat{\psi} = (f - D^2/16f)/(f + D^2/16f)$.

is considered. For the case of gain maximization, the procedure is as follows:

(1) For a given angle of main-beam pointing observation (θ_0, ϕ_0), reflector f/D ratio and distortion profile, choose the array feed parameters N, d_e, and b.[4]

(2) Calculate the field values $\{F_i\}$ using the Jacobi-Bessel method.

(3) Set the array element excitations $\{a_i\}$ to the conjugate values of $\{F_i\}$ and calculate gain.

For the more general case, a performance function $f(\overline{a})$, $\overline{a} = \{a_i\}$, can be defined as the norm of the difference between the actual secondary pattern $E(\theta, \phi, \overline{a})$, and some desired secondary pattern, $E_D (\theta, \phi)$, i.e.,

$$f(\overline{a}) = \|E(\theta, \phi, \overline{a}) - E_D (\theta, \phi)\| \tag{9}$$

The statement of the problem is then

$$\text{find min } f(\overline{a}) \tag{10}$$

where some numerical search routine, such as the Rosenbrock method, is used to perform the function minimization.

V. Analysis of Focal-Plane Field Distributions for Distorted Parabolic Reflectors

An approximate expression for the focal-plane field distribution is first derived based on the scalar diffraction integral. The expression is valid for reflectors having small magnitudes of distortion and large f/D ratios. It is meant to provide some analytical insight into the effects of reflector distortion. Contour plots of the focal-plane fields, based on data from the Jacobi-Bessel method, are presented in Subsection V.C. for various cases of distortion and beam scan angle.

Using the aperture-field method, the principal component of the electric field distribution, E, in the focal plane of a large parabolic reflector can be related to the corresponding component H of the electric field in the aperture plane by a scalar equation of the form (Ref. 1)

$$E(r, \phi') = \int_0^{2\pi} \int_0^{\hat{\psi}} H(\psi, \zeta) \exp\left[jkr \sin \zeta \cos (\psi - \phi')\right]$$

$$\times \sin \psi \, d\psi \, d\zeta \tag{11}$$

[4]The optimum choice of the feed parameters is discussed in Section VI.

where the integration is taken over the surface of the spherical cap subtended by the angle $\hat{\psi}$.

A. Undistorted Reflector

For the case of a uniform aperture distribution and a perfect reflector having $f/D > 1$, the solution to Eq. (11) has the form

$$E(r, \phi') \cong 2\pi(\hat{u})^2 \frac{J_1(kr\hat{u})}{kr\hat{u}} \tag{12}$$

where $\hat{u} = \sin \hat{\psi}$ and $J_1(x)$ is the Bessel function of first kind, first order.

B. Distorted Reflector

Small profile distortions in the reflector surface may be conveniently considered as effective phase errors, which theoretically may be projected into the aperture plane of the reflector (Ref. 9). The field distribution in the aperture plane is therefore modified by the phase error. The magnitude of the phase error, being proportional to the operating frequency, imposes an upper limit above which the reflector is ineffective as an antenna.

In what follows, models of reflector surface distortion are considered to have the form

$$\Delta z = \epsilon \rho^s \cos (L\zeta) \tag{13}$$

Where ϵ is the maximum deviation in wavelengths, s is a real number, ρ is the normalized distance from the z-axis to a point on the reflector normalized to the reflector radius and L is the integer number of periodic scallops in the distorted reflector. Based on data (Ref. 10) for gravity-induced distortions as a function of pointing angle, typical of large ground-based reflectors, ϵ can vary from 0 to 0.2, s from 0.5 to 2, and L is roughly equal to 3. Random distortion errors are not considered here.

Figure 1, with all distances normalized to the reflector radius, $D/2$, in wavelengths, shows

$$u = \sin \psi = \frac{\rho}{2 f/D + \rho^2/8 f/D} \tag{14}$$

which for $f/D > 1$ reduces to

$$\rho \cong \left(2 \frac{f}{D}\right) u \tag{15}$$

If, as discussed above, the effects of distortion are accounted for by an equivalent phase error in the aperture plane, the resulting phase error distribution, including a factor for reflector curvature as in Ref. 11, can be expressed as

$$\delta = 2\pi (1 + \cos \psi) \Delta z \tag{16}$$

If, however, the effects of reflector curvature are neglected, which is reasonable for $f/D \geqslant 1$, then Eq. (16) becomes

$$\delta \cong 4\pi \Delta z \tag{16a}$$

Substituting Eqs. (13) and (15) into Eq. (16a) results in

$$\delta = 4\pi \epsilon (2f/D)^s u^s \cos (L\zeta) \tag{17}$$

The modified aperture field distribution then becomes

$$H(u, \zeta) = \exp (j\delta) = \exp (j\epsilon' u^s \cos L\zeta) \tag{18}$$

where

$$\epsilon' = 4\pi \epsilon (2f/D)^s$$

Therefore, the focal-plane field distribution, as given by Eq. (11), for the case of a distorted reflector having $f/D \geqslant 1$, becomes

$$E(r, \phi') = \int_0^{2\pi} \int_0^{\hat{u}} \exp [j\epsilon' u^s \cos L\zeta]$$

$$\cdot \exp [jkru \cos (\zeta - \phi')] \, u \, du \, d\zeta \tag{19}$$

This can be evaluated by expanding the two exponential terms in the integrand. In general, since Bessel functions of large order and small argument can be neglected, the procedure is rapidly convergent.

For the case of $s = 2$ and $L = 3$, the following result is obtained:

$$E(r, \phi') \cong 2\pi(\hat{u})^2 \frac{J_1(kr\hat{u})}{kr\hat{u}} + \frac{16\pi \epsilon' \cos (3\phi') (\hat{u})^3}{7 kr} J_4(kr\hat{u}) \tag{20}$$

The first term on the right of Eq. (20) is the focal-plane field of an undistorted reflector. The second term represents a first-order approximation of the field due to distortion.

C. Contour Plots of Focal-Plane Field

Graphical insight into the effects of reflector distortion can be obtained from an examination of contour plots of the focal-plane field distribution. In Figs. 3a through 3e, such plots are shown for the case $f/D = 1.0$ and values of $\epsilon = 0$, 0.12λ, 0.2λ, $s = 1.0$, $L = 3$, and incident plane wave on-axis ($\theta_0 = 0$) and off-axis ($\theta_0 \cong 0.5$ BW).[5] Each of these plots is based on a triangular grid of 37 data points (see Fig. 2).

The deviation from circular symmetry and the spreading out of energy in the focal plane as a function of distortion is clearly shown by these plots. The combination of distortion and off-axis beam scan (Fig. 3e) results in particularly severe spreading of the focal-plane energy. Bearing in mind the dynamic nature of the distortion effects, these plots make it clear that it would be impossible to properly match the focal-plane fields of a distorted reflector with a conventional single-element feed.

VI. Numerical Results — Gain Maximization

In this section, quantitative information is presented about the effectiveness and feasibility of the array-feed approach, as well as optimum array-feed parameter values obtained from gain maximization calculations made using the algorithmic procedure described in Section IV. This is done for a variety of f/D ratios, beam-scan angles θ_0, array-feed parameters (N, d_e, and b), and distortion parameters ϵ and s.

Feasibility of the array-feed approach is largely determined, from a hardware point of view, by the number and type of array-element radiators required to achieve significant gain restoration and vernier beamsteering.

The output of the algorithmic synthesis procedure is the set of array excitations that optimize performance and the resulting performance value (i.e., gain). To maximize gain with a given number of array elements N, the optimum choice of the feed parameters (d_e and b) depends on the need for greater or less granularity and/or greater or less capture area to conjugate-match the array-feed excitation to the focal-plane field distribution.

To see the relationship between gain performance and these parameters, the results of the matrix of gain calculations are presented graphically in Figs. 4a through 4e (on-axis gain loss vs d_e) for values of $N = 7$, 19, 37; $f/D = 0.4$, 1.0, 1.5; $\epsilon = 0.$, 0.12, 0.2; $s = 1$, 2; and $b = 2.07$ and 2.47. In Fig. 5, gain loss is plotted as a function of scan angle, θ, BW, for the optimum

---

[5]Beamwidth, $BW \cong 1.16\lambda/D$ radians (i.e., $\cong 0.66$ deg for $D = 100\lambda$).

choice of d_e (taken from Figs. 4a through 4e) with $N = 7, 19, 37$ ($f/D = 1.0$). In these graphs, gain loss is referenced to the ideal maximum gain of the reflector's circular aperture (e.g., 49.94 dB) for $D = 100\lambda$). It is also useful to have as a reference the gain loss results obtained with a single feed, $N = 1$, for various f/D and distortion parameter values (Table 1).

From these results it is immediately seen that: (a) the lower the f/D value and therefore the more pronounced the reflector curvature (Eq. (16)), the smaller the effects of distortion; (b) gain loss increases with increasing modulus of distortion ϵ; and (c) gain loss is less for larger s (i.e., $s = 2$), since this corresponds to distortion concentrated at the reflector rim, while smaller s (i.e., $s = 1$) corresponds to distortion distributed more widely over the reflector surface, resulting in greater gain loss.

A. Gain Loss as a Function of the Array Parameters N, d_e, b

The results presented in Figs. 4 and 5 show that, as compared to the gain obtained with a single feed, an array feed having the proper choice of parameters can provide significant gain restoration and (for $N = 19, 37$) useful vernier beamsteering. For $N = 7$, Figs. 4a through 4c, it is seen that the value of element diameter that gives maximum on-axis gain loss is approximately that required by a single feed for the given f/D ratio, i.e.

$$\text{optimum } d_{e_{N=7}} \cong d_{e_1} \cong \sqrt{\frac{9.2}{b}} \, (f/D) \qquad (21)$$

where d_{e_1} is given by Eq. (8b). While the amount of gain restoration is significant, there still remains some residual loss due to distortion; i.e., for the case $\epsilon = 0.12, s = 2, b = 2.07$, the gain loss is approximately 0.3 dB greater than that for $N = 1$ with no distortion (Table 1).

Also seen in Figs. 4a, b, and c for $N = 7$, the use of elements having high effective aperture, characterized by a value of $b = 2.47$, gives better on-axis gain (approximately 0.2 dB better for the case $\epsilon = 0.12, s = 2$) than elements with $b = 2.07$. This increase in gain can probably be explained by the fact that the higher value of b allows smaller array element spacing (and therefore better matching granularity) while maintaining a high effective capture area for the feed. The results presented here do not account for coupling effects, and therefore the possible advantages of using high effective aperture elements (e.g., $b = 2.47$) in an array are subject to experimental verification. In what follows, the emphasis is placed on results obtained with elements characterized by the more conservative (i.e., more realizable) value of $b = 2.07$.

As seen in Fig. 5, the optimum seven-element array will not provide continuous low-loss scanning but rather stepped scanning with beams formed on-axis and approximately 1.8 BW off-axis. This is discussed further in Subsection VI.B.

In each of the cases shown in Figs. 4a, b, and c for $N = 7$, it is seen that gain loss has an oscillatory or quasi-periodic character as a function of variable d_e: in addition to the optimum at d_{e_1}, there is a local optimum at approximately $1/3 \ d_{e_1}$. This can be explained heuristically as giving the number and size of identical circular elements that fill the minimum diameter d_{e_1} and that fit on a triangular grid, but that do not overlap the focal plane field null region. This explanation fits well with the results for $N = 19$ and 37, (Figs. 4d and e), where it is seen that the optimum d_e is approximately $1/3 \cdot d_{e_1}$. Therefore

$$\text{optimum } d_{e_{N=19,37}} \cong \frac{1}{3} \sqrt{\frac{9.2}{b}} \, (f/D) \qquad (22)^6$$

From Eqs. (3), (21), and (22), it is seen that the maximum diameter of the optimum 7-element array feed is approximately 30% larger than that of the optimum 37-element array and 80% larger than that of the optimum 19-element array (Fig. 6).

For $N = 19$, the on-axis gain is hardly better than that for $N = 7$. However, for $N = 19$, very good beamsteering is obtained over a range of ± 1 BW (Fig. 5). As is the case for $N = 7$, and also for $N = 19$, the use of an element having high effective aperture characterized by a value of $b = 2.47$ does give better on-axis gain (approximately 0.2 dB better for the case $\epsilon = 0.12, s = 2$) than an element having lower effective aperture ($b = 2.07$, Fig. 4d).

For $N = 37$, the on-axis gain and the beamsteering performance is excellent over a scanning range of ± 1.5 BW (Figs. 4e and 5). In fact, over the range of ± 0.5 BW for $N = 37$, the gain with distortion ($\epsilon = 0.12, s = 2$) is equal to or better than the gain for $N = 1, 7$, or 19 with no distortion ($\epsilon = 0$). With $N = 37$, no appreciable improvement in on-axis gain is obtained by using elements having a value of $b = 2.47$.

For small values of f/D, Eq. (22) indicates that the optimum diameter is less than 0.5λ for arrays with 19 or 37 elements. In most cases, such elements would be impractically small. Therefore, in the context of this study, the use of array feeds with more than seven elements for reflectors with small f/D values (e.g., 0.4) would not be practical.

---

[6] For $N = 19$, the numerical results indicate that the optimum d_e is approximately 10% larger than that given by Eq. (22).

Finally, in Fig. 7, the secondary patterns of a distorted reflector with $f/D = 1$, $\epsilon = 0.12$, and $s = 2$ are shown for feeds with (a) $N = 1$ ($d_e = 2.1$) and (b) $N = 37$ ($d_e = 0.71$). Note that the array feed ($N = 37$) has produced a secondary pattern that is reasonably symmetric despite the asymmetric reflector distortions.

B. Beamsteering

To deal with the question of beamsteering, a simple analysis is made based on the reflector/array feed geometry. For large apertures, the scan angle θ_s is approximately

$$\theta_s \approx \frac{r}{f} BDF \qquad (23)$$

where r is the radial distance of the feed away from the focal point origin, and BDF is the beam deviation factor, which is $\cong 1$ for $f/D \geqslant 1$. The number of beamwidths of scan tor a scan angle of θ_s is given by

$$\frac{\theta_s}{BW} \cong \frac{D}{1.16\lambda} \cdot \frac{r}{f} \cdot BDF \qquad (24)$$

where the secondary pattern beamwidth, $BW \cong 1.16\lambda/D$.

Now, the set of elements that comprise the array generates a set of corresponding beams in the far field. To achieve continuous, low-loss scan, the far-field beam crossovers should not be lower than their 3-dB points. To achieve this, it is necessary that

$$\frac{\theta_s}{BW} \leqslant 1 \qquad (25)$$

For $r = d_e$ (the radial distance to the first ring of elements in the feed array), Eqs. (24) and (25) imply that

$$d_e \leqslant \frac{1.16\lambda}{BDF} (f/D) \qquad (26)$$

The restrictions imposed by Eq. (26) on d_e for continuous, low-loss beamsteering performance may not, in all cases, be consistent with the requirements for optimum on-axis gain correction. For example, with $N = 7$ and $f/D = 1$, the on-axis gain is optimized with $d_e \cong 2.1\lambda$ [Fig. 4b and Eq. (21)]. This value of d_e does not satisfy Eq. (26). By Eq. (23), the scan angle θ_s for an element located at 2.1λ is approximately 1.8 BW and not $\leqslant 1$ BW as required for continuous, low-loss beamsteering. The array feed in this example gives good "stepped" beamsteering but not continuous beamsteering (e.g., $G(0$ BW$) \cong -0.8$ dB and $G(1.8$ BW$) \cong -0.8$ dB but $G(1$ BW$) \cong -7.0$ dB (see Fig. 5). On the other hand, for $N = 37$, the optimum value of d_e to maximize on-axis gain is $\cong 0.71\lambda$ [Fig. 4e or Eq. (22)], which does satisfy Eq. (26), giving excellent, continuous, low-loss beamsteering performance (Fig. 5).

VII. Conclusions

This study has shown that a planar array feed has excellent potential as a solution to paraboloidal reflector distortion problems and beamsteering requirements. The numerical results obtained show that, for the range of distortion models considered, good on axis gain restoration can be achieved with as few as seven elements. For beamsteering to ±1 BW, 19 elements are required. For arrays with either 7 or 19 elements, high effective aperture elements (e.g., $b = 2.47$) give higher system gain than elements having lower effective apertures (e.g., $b = 2.07$). With 37 elements, excellent gain and beamsteering performance to ±1.5 BW are obtained independently of the assumed effective aperture of the array element. Simple "rules of thumb" for the design of the planar array feed configuration have been presented.

References

1. Rudge, A. W., and Davies, D. E. N., "Electronically Controllable Primary Feed for Profile Error Compensation of Large Parabolic Reflectors," *Proc. IEEE (Brit.)*, Vol. 117, No. 2, pp. 351–358, February 1970.

2. Amitay, N., and Zucker, H., "Compensation of Spherical Reflector Aberrations by Planar Array Feeds," *IEEE Trans.*, Vol. AP-20, No. 1, pp. 49–56, January 1972.

3. Imbriale, W. A., Galindo-Israel, V., Rahmat-Samii, Y., and Bruce, R. L., "Improved Wide-Angle Scan Using a Mini-Max Optimization Technique," *Proceedings of the URSI Meeting*, Houston, Texas, May 1983.

4. Blank, S. J., and Imbriale, W. A., "Array Feed Synthesis for Correction of Reflector Distortion and Vernier Beamsteering," *Proceedings of the International Radio Science Union (URSI) Meeting*, Boulder, Colorado, January 1983.

5. Rahmat-Samii, Y., and Galindo-Israel, V., "Shaped Reflector Antenna Analysis using the Jacobi-Bessel Series," *IEEE Trans.*, Vol. AP-28, pp. 425–435, July 1980.

6. Collin, R. E., and Zucker, F. J., *Antenna Theory*, McGraw-Hill, New York, New York, 1969.

7. Rahmat-Samii, Y., Cramer, P., Woo, K., Lee, S. W., "Realizable Feed Element Patterns for Multibeam Reflector Antenna Analysis," *IEEE Trans.*, Vol. AP-29, No. 6, pp. 961–963, November 1981.

8. Wilde, D. J., and Beightler, C. S., *Foundations of Optimization*, Prentice-Hall, Inc., 1967.

9. Silver, S., *Microwave Antenna Theory and Design*, McGraw-Hill, New York, New York, 1949.

10. Katow, M. S., *Structural Technology for Large Radio and Radar Telescope Systems*, pp. 185–280, J. W. Mar and H. Liebowitz, Eds., MIT Press, Cambridge, Massachusetts, 1969.

11. Imbriale, W. A., and Rusch, W. V. T., "Scalar Analysis of Distorted Umbrella Reflector," *IEEE Trans.*, Vol. AP-22, pp. 112–114, January 1974.

Table 1. Gain loss with a single feed, $N = 1$ ($b = 2.07$)

f/D	d_{e_1}, λ	ϵ, λ	Gain Loss, dB	
			$s = 2.0$	$s = 1.0$
0.4		0	-0.84	-0.84
	0.9	0.12	-1,56	-2.07
		0.2	-2.83	-5.28
1.0		0	-0.89	-0.89
	2.2	0.12	-1.95	-2.63
		0.2	-3.79	-6.72
1 5		0	-0.90	-0.90
	3.2	0.12	-2.04	-2.74
		0.2	-4.00	-7.03

d_{e_1} is chosen to be slightly larger than is optimum when there is no distortion ($\epsilon = 0$), to give better gain with distortion.
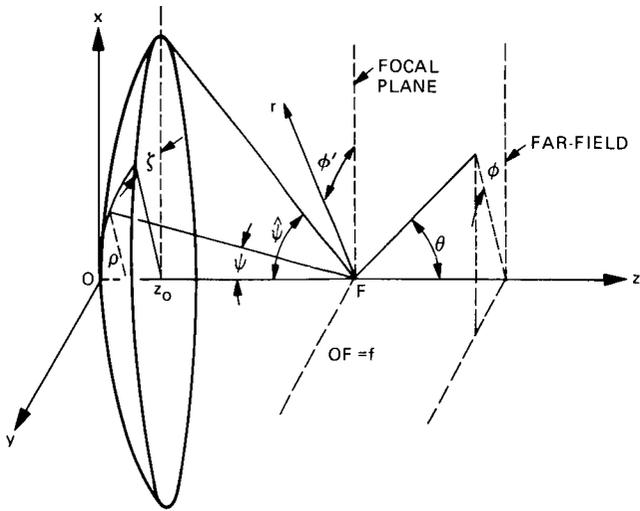
Fig. 1. Geometrical parameters of the parabolic reflector
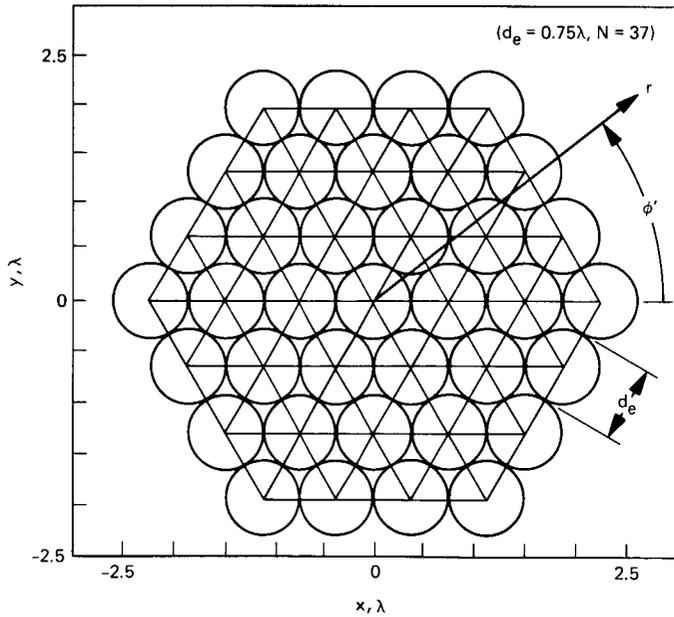
Fig. 2. Triangular grid array of circular elements and focal plane
coordinates r, ϕ'
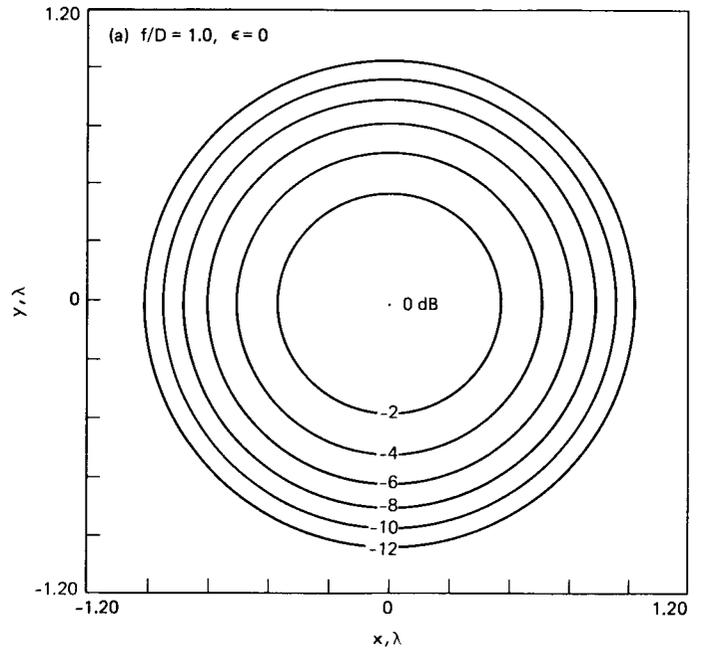
Fig. 3. Contour plots of focal-plane field distribution
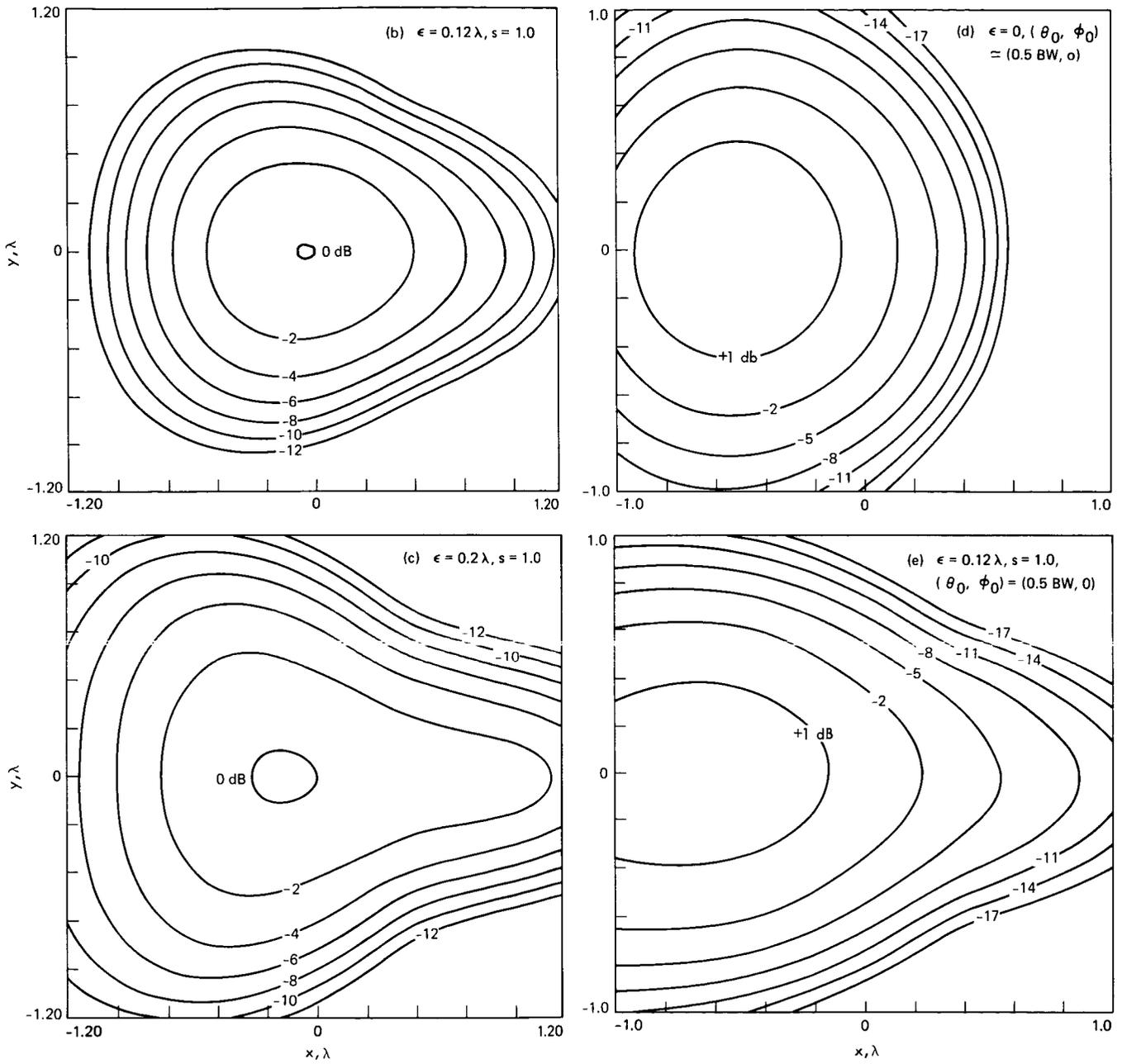
Fig. 3 (contd)
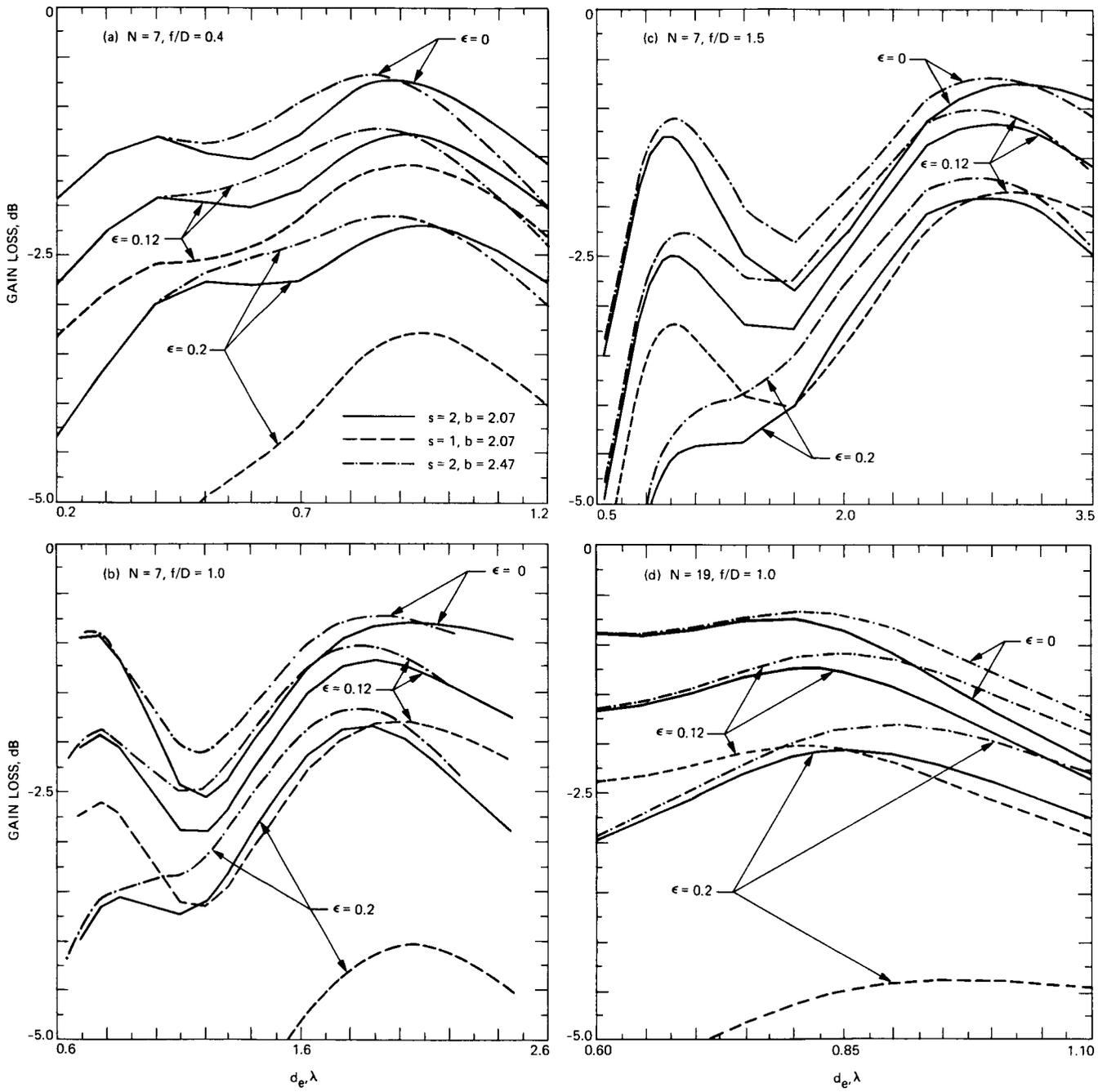
Fig. 4. On-axis gain loss as a function of element spacing
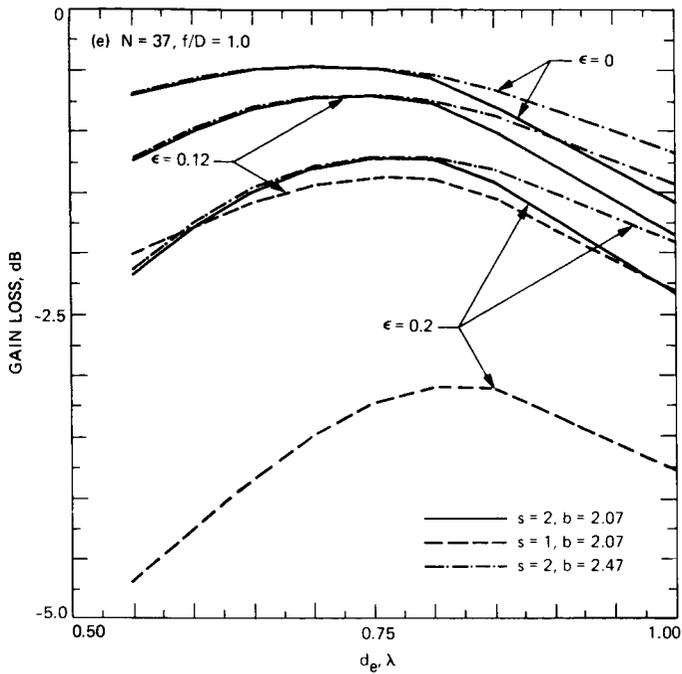
Fig. 4 (contd)
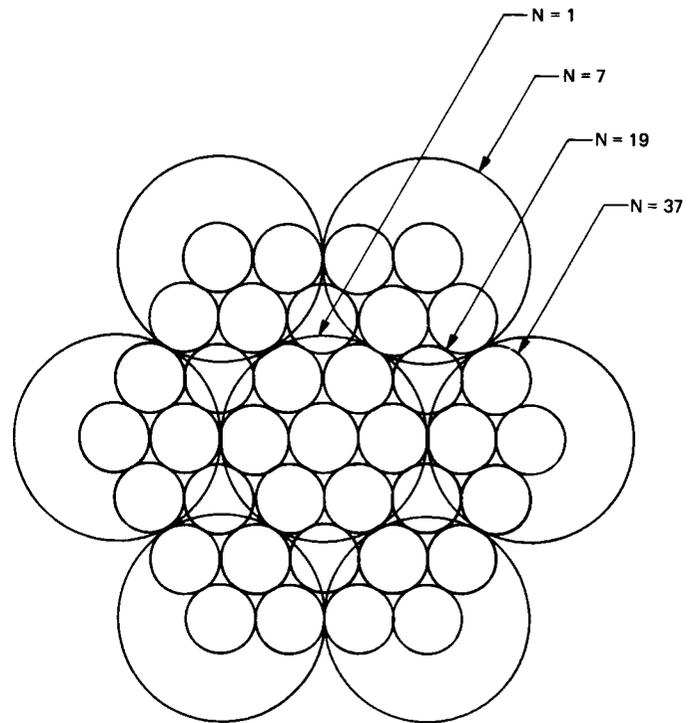
Fig. 6. Optimum array feed configurations for N = 1, 7, 19, and 37 elements
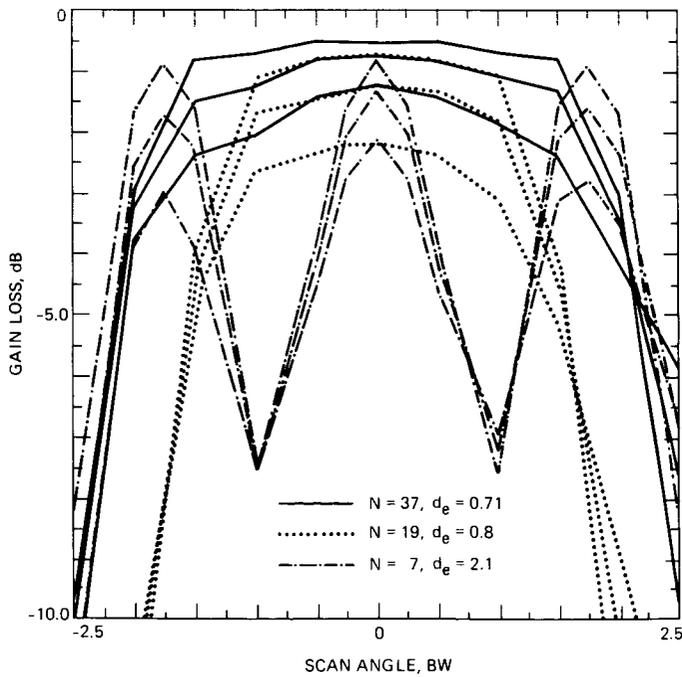
Fig. 5. Gain loss as a function of scan angle (f/D = 1.0, s = 2, ϵ = 0, 0.12, 0.2)
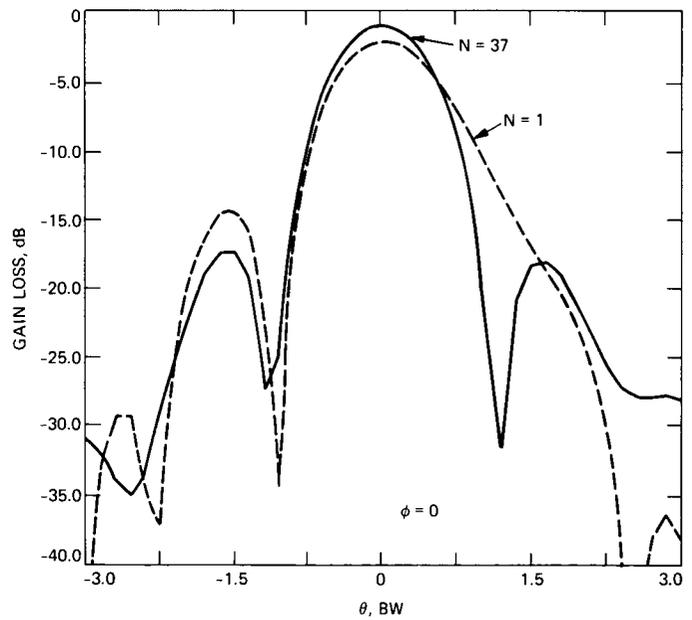
Fig. 7. Secondary patterns of distorted reflector (f/D = 1.0, ϵ = 0.12λ, s = 2.0 for N = 1, d_e = 2.1, and N = 37, d_e = 0.71

1- to 4-K Refrigeration Techniques for Cooling Masers on a Beam Waveguide Antenna

D. L. Johnson

Radio Frequency and Microwave Subsystems Section

The status of technology is reported for various 1- to 4-K commercially available refrigeration systems capable of producing 1.5-K refrigeration to cool masers and superconducting cavity oscillators on the proposed beam waveguide antenna. The design requirements for the refrigeration system and the cryostat are presented. A continuously operating evaporation refrigerator that uses capillary tubing to provide a continuous, self-regulating flow of helium at approximately 1.5 K has been selected as the first refrigerator design for the beam waveguide antenna.

I. Introduction

This past year, JPL installed a 2.3-GHz maser on the beam waveguide antenna in Usuda, Japan, to support the ICE mission and also to measure the performance of the antenna/maser combination. The results were impressive (Ref. 1). Comparing low-noise configurations of the 64-m beam waveguide antenna and a Deep Space Network (DSN) 64-m antenna at 2.295 GHz and at zenith, the measured total system noise temperature of the beam waveguide antenna was 15.0 K, whereas the measured total noise temperature was 16.0 K for the DSN 64-m antenna.

JPL is proposing to NASA that a 34-m beam waveguide antenna be built at Goldstone for research and development purposes. This antenna would support the new 32-GHz communications link. In addition to improved noise temperatures, there would no longer be limitations of size, variable elevation angle, and remote location of the feedcone as design parameters for the maser's refrigeration system. In this new design, the maser and its closed-cycle refrigerator (CCR) would be

situated in a laboratory at the base of the antenna. The advantage of this plan is that the maser/CCR would no longer need to be orientation-independent or as compact; rather, it would be in a stationary, upright position, which would provide easy access to the maser(s) for modifications and maintenance, even during antenna tracking. This arrangement also eases the size and weight restrictions of the refrigeration system and provides a clean laboratory environment for the operators.

In addition to the 32-GHz maser, the beam waveguide antenna could support other masers and High Electron Mobility Transistor (HEMT) amplifiers. Masers require operating temperatures of 4.5 K or less. The HEMT amplifier requires only 12-K refrigeration, but for simplicity and convenience, the HEMT may also be cooled to 4.5 K.

Along with the amplifiers, the new cryostat would be designed to also contain a stabilized superconducting cavity oscillator, such as the one built by Caltech (Ref. 2). The cavity would replace the hydrogen maser as the frequency standard

for the antenna. Initial measurements by Caltech using a ruby maser stabilized by a lead-on-sapphire superconducting cavity at 1 K indicated the stability $\Delta f/f$ to be 3×10^{-14} over 4000 seconds, which extrapolates to a stability of 10^{-15} for 1 second (Ref. 3). In contrast, the present DSN hydrogen maser operates with a stability $\Delta f/f$ of 10^{-15} for 1000 seconds. The theoretical stability of a superconducting stabilized oscillator can be up to three orders of magnitude better than the hydrogen maser, but is dependent on the physical temperature, the temperature stability, and the superconducting material used. With the lead-on-sapphire cavity at 0.6 K, the theoretical stability approaches 10^{-18} for 1000 seconds; at 1.5 K, the stability is reduced to 10^{-15} for 1000 seconds, but a stringent temperature stability of $<10 \mu K$ is also required. As an alternative, an all-sapphire cavity operating at 1.5 K could yield a stability better than 5×10^{-17} for 1000 seconds. This added stability could be traded for a less stringent temperature stability; however, use of an all-sapphire cavity would require much additional development work. The selection of the cavity type for use as the stabilized oscillator on the beam waveguide antenna will depend on the stability goal for the antenna and the temperature and temperature stability achievable by the refrigeration system.

Although the new cavity oscillator requires additional cryogenics to enable it to operate at or below 1.5 K, this oscillator can be put in the same cryogenic package as the masers. A decision must be made whether to cool the maser to this low temperature as well. Since the gain of the maser (in dB) is roughly inversely proportional to the physical temperature of the maser, lowering the maser temperature from 4.5 K to 1.5 K could improve the maser gain by a factor of three (or alternately, increase the maser bandwidth by a factor of three). This improvement in maser performance must be weighed against the added complexity of placing additional components and heat loads in the 1.5-K portion of the refrigeration system.

This report examines the design requirements of the cryostat and the cooling requirements of the maser and cavity oscillator, the availability of 1.5-K refrigeration systems and components in industry, and the feasibility of using such a system in a beam wavebuide antenna.

II. Design Requirements

The 2.3- and 8.5-GHz masers presently used in the DSN are cooled by 1-W closed-cycle refrigerators (CCRs). The 32-GHz maser package will be cooled by a 3-W CCR. The expected heat load for the 32-GHz maser is about 1 W. The superconducting cavity at Caltech has a measured heat load of 50 μW at 1.5 K due to the rf input power, negligible compared to that of the maser. Because of this disparity in cooling requirements,

it is necessary to consider the option of cooling only the cavity oscillator to 1.5 K or less, versus cooling both the cavity and the maser(s) to 1.5 K. If the latter option is chosen, the difference in required temperature stabilities for the devices will also contribute to the complexity of the cryostat. The design requirements for the cryostat are listed below.

(1) Cavity and one or two masers housed in vacuum, cooled by conduction

(2) 1-2 week hold time at 4.2 K in event of power/ mechanical failure

(3) 50 μW of refrigeration at 1.5 K, 300 mW at 4.2 K for the cavity

(4) $<10 \mu K$ stability at 1.5 K for the cavity

(5) 1 W of refrigeration between 1.5 K and 4.2 K for the maser

(6) <10 mK stability for the masers

The masers and the cavity must also be independently accessible for service or maintenance, preferably while the remainder of the electronics in the cryostat remain at the low temperature. Therefore, it would be convenient to be able to hang the electronics from the top of the cryostat. Also, there is a tradeoff between the allowable thermal conductance and the allowable rf losses in the waveguides feeding the maser and cavity in the cryostat.

In selecting a refrigeration system, it is desirable to find a system that is commercially available, with readily available parts and service, and low operator maintenance requirements. The refrigeration system should also require no liquid nitrogen precool in order to simplify cryogenic requirements.

III. 1- to 4-K Refrigeration Systems

Magnetic refrigeration, electrocaloric refrigeration, or evaporative cooling techniques can produce temperatures between 1.0 K and 4.2 K. However, above 1 K, evaporative cooling is most commonly used. The other refrigeration methods are not yet commercially available, but it is of interest to describe these alternate cooling techniques briefly and the progress made in their development.

A. Magnetic Refrigeration

Adiabatic demagnetization, or magnetic cooling, was first suggested in 1925 by two independent researchers, Giauque (Ref. 4) and Debye (Ref. 5), as a means to produce temperatures below those attainable by pumping on a liquid helium bath. The first continuous magnetic refrigerator was built by Collins and Zimmerman (Ref. 6) in 1953, but it used

mechanical heat switches, which produced considerable vibrational heat leak. In 1954, Heer (Ref. 7) built the first continuous magnetic refrigerator with superconducting heat switches. That machine was capable of maintaining temperatures in the range of 0.2 to 1.0 K with a refrigeration capacity of a few microwatts. These refrigerators were made commercially by Arthur D. Little Corporation for a brief period until ^{3}He became commercially available in sufficient quantities for evaporative cooling. In the late 70s and early 80s, efforts were renewed to make a continuous magnetic refrigerator for the 2- to 4-K and the 4- to 20-K ranges. The French (Ref. 8) demonstrated the feasibility of a reciprocating magnetic refrigerator which, operating with cold-and hot-end temperatures of 1.8 K and 4.2 K, could produce 1.2 W of refrigeration for short periods of time. The Japanese have built two magnetic refrigerators that also operate in the 2- to 4.2-K temperature span — a rotational wheel magnetic refrigerator (Ref. 9), and a pulsed coil magnetic refrigerator (Ref. 10) — that have produced cooling powers of 1.5 W at 2.10 K and 0.6 W at 1.80 K, respectively.

B. Electrocaloric Refrigeration

Analogous to the magnetic refrigerator, the electrocaloric refrigerator uses an applied electric field to cause the electric dipoles in a dielectric material to become ordered at a higher temperature. Dielectric materials that have transition temperatures in the 1- to 4-K range do exist, but their small dipole entropies make them incapable of producing significant refrigeration. Shepherd (Ref. 11) reported using an OH-doped KCl dielectric material to produce cooling at 0.3 K from a starting temperature of 1.3 K. The only reported application of electrocaloric refrigeration has been the use of the OH-doped KCl for thermostating crystals below 1 K while they were radiated with short light bursts (Ref. 12). New dielectric materials will have to be developed for this temperature region before electrocaloric refrigeration will become a viable refrigeration method.

C. Evaporative Cooling

Evaporative cooling, or subatmospheric refrigeration, has been used for many years as the means of lowering the bath temperature of the liquid cryogen. The temperature of the liquid is lowered below the normal boiling point by reducing the vapor pressure over the liquid. Temperatures as low as 0.3 K and 0.8 K can be attained with liquid ^{3}He and ^{4}He, respectively, down from their normal boiling temperatures of 3.2 K and 4.2 K at atmospheric pressure. This refrigeration technique is being used in many research programs today because of its ease of operation and the wide range of refrigeration capacity and temperatures achievable. Evaporative cooling is the most suitable at this time for maser cooling on a beam waveguide antenna. Various evaporative cooling methods will be discussed next and the configurations most suitable for use on a beam waveguide antenna will be described.

Temperatures down to 2.3 K have been achieved with the use of an ejector (Ref. 13). An ejector uses the momentum of a high-velocity stream of gas to entrain and accelerate a slower moving gas into which it is directed (Fig. 1). This reduces the pressure in the slower moving gas stream, lowering its temperature. The ejector can be placed outside the cryostat (Ref. 14), as in Fig. 1, driven by the side stream of the main compressor, or it can be installed in the Joule-Thomson (J-T) circuit near the J-T heat exchanger (Refs. 13,15), as shown in Fig. 2. The cold ejector has the benefit of not requiring the room temperature compressor to operate at subatmospheric pressures, thereby avoiding possible air leaks into the gas stream. Staging of ejectors could possibly produce temperatures lower than 2.3 K, but not without significant development work.

Vacuum pumping the vapor above the liquid bath can produce the low temperatures mentioned previously. Pumping the vapor causes the liquid to boil as it tries to maintain a pressure equilibrium between the liquid and the vapor. In so doing, the liquid cools itself, but at the expense of reducing the volume of liquid remaining in a dewar. At 1.5 K, only about half of the original volume of liquid remains, requiring either the use of a larger dewar or more frequent refilling of the dewar, which may be inconvenient for experiments that could last for days. In addition, this process of temperature reduction has low efficiency because it does not use the sensible heat of the helium vapor leaving the dewar. However, the simplicity of this cooling method makes it attractive for research experiments requiring small refrigeration loads where the test duration is on the order of hours.

The vacuum pumping process can be turned into a continuous low-temperature refrigeration system by connecting the low-temperature pot to the outer jacket of 4.2 K liquid helium with a capillary tube (Ref. 16, Fig. 3). The capillary tube provides a continuous flow of 4.2-K liquid to the pot. Helium transfer is only required to top off the 4.2-K jacket, which will not influence the temperature of the pot. The incoming liquid to the pot produces a small heat leak, which may raise the temperature of the pot slightly.

The cryostat used to cool the superconducting cavity at Caltech uses capillary tubing but also incorporates a manual shut-off valve at the entrance to the tubing. This keeps the refrigeration from being continuous and operator-free, but permits lower temperatures to be reached. When the pot is in need of refilling, the valve is opened to the 4.2-K liquid helium jacket. By vacuum pumping and refilling simultaneously, the liquid entering the pot is cooled to a temperature near 1 K.

The valve is closed off when the pot is filled, resulting in only a small fraction of the liquid being boiled off as the system is further cooled to 0.8 K. In the Caltech cryostat, the 0.5-liter volume of liquid in the pot will keep the superconducting cavity at a temperature of 1.0 K for about 5 days before requiring refilling with LHe.

Cryostats with capillary tubing have been used to produce mW of refrigeration at 1 K. However, with the proper sizing of the capillary tubing, it should be possible to attain the 1 W of refrigeration required for the maser. The Caltech cryostat uses an open-cycle process. Liquid helium is transferred periodically from a storage dewar into the 4.2-K jacket of the cryostat. The helium vapor pumped from the pot by the vacuum pump is vented to air. This process can be turned into a closed-cycle process to conserve helium by returning the pumped helium vapor to the compressor of a liquefier for reliquefaction. An important requirement for a DSN system is that the vacuum pump be helium-tight to prevent gas contamination.

If refrigeration on the order of watts or more is required, then maximizing the thermal efficiency of the refrigeration process is of great importance. The thermal efficiency can be improved by incorporating a heat exchanger to recover the refrigeration available in the sensible heat of the exiting helium vapor. Collins (Ref. 17) analyzed and compared the work input required to provide cooling at 1.85 K for refrigeration systems that do and do not incorporate a high-efficiency heat exchanger. His results showed a five-fold improvement in efficiency for a continuous closed-cycle refrigerator using a heat exchanger over an open-cycle method in which 4.2-K liquid is first produced and then the bath temperature is lowered. The heat exchanger may be added either in the gas circuit of the refrigerator (Fig. 4) or as an entirely separate closed-loop gas circuit in heat exchange with the refrigerator circuit (Fig. 5).

A specially designed heat exchanger built into the refrigerator circuit as shown in Fig. 4 was tested by Collins (Ref. 17). The liquefier-refrigerator produces 7-K helium gas[1] which is introduced to the high-pressure supply side of the heat exchanger. High-pressure, 300-K helium gas is also introduced to the heat exchanger to recover all the available refrigeration of the exiting low-pressure helium gas. The exiting gas flows over the finned-tube heat exchanger, cooling the incoming gas. At 7 K, the streams combine and flow through the remainder of the exchanger with the expansion valves. The intermediate expansion valves are used to reduce the supply pressure at the lower temperatures so that the enthalpy change (change in

---

[1] Collins reports minimal improvements when 4.2-K gas is used.

heat content) in the high-pressure stream is more nearly equal to that of the low-pressure stream. This permits recovery of more of the refrigeration capacity of the low-pressure gas, thereby increasing the heat exchanger efficiency.

A heat exchanger of this type is used to cool the linear accelerator at the University of Illinois. Coupled with a CTI Model 1400® liquefier, the university's heat exchanger, without the LN_2 precool, is capable of producing 11 W of refrigeration at 1.85 K. This heat exchanger is contained in a vacuum chamber separated from the liquefier and the accelerator cryostat. The accelerator cryostat is filled with the 1.85-K liquid that has been transported from the heat exchanger through an evacuated transfer line. The transfer line is precooled by the exhaust vapors being pumped from the cryostat. Following the development of this 1.85-K refrigerator, a 300-W, 1.85-K refrigerator was designed and built for Stanford University. The designs are similar, but the Stanford refrigerator requires the LN_2 precool, along with four-stage vacuum pumping of the heat exchanger.

A JPL heat exchanger design using a separate closed-cycle loop in heat exchange with the refrigerator (Fig. 5) was demonstrated in 1975 (Ref. 18). The JPL design used the 1-W CCR as a 4.5-K precooling stage. Room temperature helium gas is supplied to the added heat exchanger loop by the existing CCR compressor. The gas in the loop is cooled at the 70-K, 15-K, and 4.5-K stages of the CCR before passing through a single J-T expansion valve to provide 3-K temperatures. The added heat exchanger loop used the same size heat exchangers as those in the J-T circuit of the unmodified CCR. By regulating the supply pressure to the heat exchanger loop to 3.04×10^5 Pa (3 atm), 200 mW of refrigeration was achieved at 3.1 K. Use of a vacuum pump with a greater pumping speed and heat exchangers with less restrictive low-pressure return paths would have permitted lower temperatures to be reached.

IV. JPL Refrigerator Design

The desire for a 1- to 2-week reserve capacity at 4.2 K requires use of a helium liquefier and an intermediate LHe storage dewar. For an expected heat load of 1 W at 1.5 K, a liquefaction rate of 2 liters per hour is required. Two liquefiers for this liquefaction rate are commercially available. Both have been recently designed and developed for hospital use with nuclear magnetic resonance imaging (MRI). The Model 1200® liquefier, manufactured by Koch Process Systems, Inc., is capable of producing liquid helium at a rate of 5 liters per hour. Its screw compressor can adjust the flow rate to lower the liquefaction rate to 2 liters per hour. A number of these systems are already in operation. The other liquefier is the Model TCFII®, manufactured by Sulzer of Switzerland. This is a 5-liter-per-hour liquefier, and is now undergoing

final pre-marketing tests. It is expected to be commercially available by the end of the year. Although both liquefiers are designed to use LN_2 precooling for optimum performance, both can operate without the LN_2 precool with a slight decrease in liquefaction rate.

For the first engineering design, the JPL refrigeration system should incorporate a cryostat that uses a capillary tube for the continuous flow of helium to the 1.5-K pot and direct vacuum pumping on the 1.5-K bath. A possible configuration for the JPL closed-cycle refrigerator system for the beam waveguide antenna is shown in Fig. 6. All items are commercially available with the exception of the cryostat. The buffer tank is used to store the helium gas when the cryostat and the storage dewar are warm. The compressor and vacuum pump would be set off in a separate room. A vacuum pump having a 150-1/s pumping speed would handle the boil-off from the 1-K pot. The liquefier and the storage dewar can be positioned in close proximity to the cryostat. A 250-liter storage dewar would amply hold a 1-week supply of 4.2-K liquid and would supply the liquid for the cryostat's continued operation in the event of a mechanical failure or electrical power failures. During a mechanical or power failure, the boil-off from the cryostat may be vented to the atmosphere or may be collected in a large bladder for future repurification and reuse. The liquid in the pot would warm to 4.2 K, but the maser's superconducting magnet would remain charged in the superconducting state.

A cryostat that takes into consideration all design requirements for the masers and the superconducting stabilized oscillator will be designed and fabricated at JPL. The design scheme has not yet been determined. Of particular interest is the possibility to warm and remove for maintenance either the maser or the cavity without interfering with the operation of the other. The different temperature stability requirements for the cavity oscillator and the masers may require separate 1.5-K pots, with separate pumping schemes for each. This would permit individual control of the temperature and temperature stability for each pot. The exhaust gas stream will be used to cool outer radiation shields, thus using some of the sensible heat of the venting vapor. This design will determine whether the continuous liquefaction/vacuum pumping refrigeration approach will be sufficient for the 1-W refrigeration requirement.

V. Conclusion

This report has introduced the various methods of attaining temperatures between 1 K and 4.2 K. Magnetic refrigeration, electrocaloric refrigeration, and evaporative cooling techniques are all possibilities. Of these, evaporative cooling is the only commercially available refrigeration method that fills the requirements set forth for cooling the masers and the superconducting stabilized cavity oscillator on a beam waveguide antenna. This method requires a liquefier to provide a supply of 4.2-K liquid to a storage dewar. The stored liquid is then transferred as needed to the cryostat to provide both a 4.2-K bath as a radiation shield and the supply to the 1.5-K pot which cools the masers and cavity. This refrigeration system will be used in a closed-cycle mode, recycling the pumped helium vapor to the refrigerator for reliquefaction.

References

1. Neff, D., "Installation and Operation of a 2.3-GHz Maser on the Usuda, Japan, 64-Meter Antenna," to be published in the *TDA Progress Report*.

2. Thakoor, S., Strayer, D. M., Dick, G. J., and Mercereau, J. E., "A Lead-on-Sapphire Superconducting Cavity of Superior Quality," *J. Appl. Phys.*, Vol. 59, pp. 854–858, 1986.

3. Strayer, D. M., Dick, G. J., and Mercereau, J. E., "Characterization of an All-Cryogenic Oscillator as a Stable Frequency Source," *Proceedings of the 17th Annual Precise Time and Time Interval (PTTI), Applications and Planning Meeting*, pp. 174–188, December 3–5, 1985, U.S. Naval Observatory, 1986.

4. Giauque, W. F., "Paramagnetism and the Third Law of Thermodynamics. Interpretation of the Low-Temperature Specific Heat of Gadolinium Sulfate," *J. Am. Chem. Soc.*, Vol. 49, pp. 1870–1877, 1927.

5. Debye, P., "Einige Bemerkungen zur Magnetisierung bei tiefer Temperatur," *Ann. Phys.*, Vol. 81, pp. 1154–1160, 1926.

6. Collins, S. C., and Zimmerman, F. J., "Cyclic Adiabatic Demagnetization," *Phys. Rev.*, Vol. 90, pp. 991–992, 1953.

7. Heer, C. V., Barnes, C. B. and Daunt, J. G., "The Design and Operation of a Magnetic Refrigerator Maintaining Temperatures Below 1 K," *Rev. Sci. Instrum.*, Vol. 25, pp. 1088–1098, 1954.

8. Lacaze, A. F., Beranger, R., Bon Mardion, G., Claudet, G., and Lacaze, A. A., "Double Acting Reciprocating Magnetic Refrigerator: Recent Improvements," *Advances in Cryogenic Engineering*, Vol. 29, pp. 573–579, Plenum Press, New York, 1984.

9. Hakuraku, Y., and Ogata, H., "A Magnetic Refrigerator for Superfluid Helium Equipped with a Rotating Superconducting Magnet System," *Jap. J. App. Phys.*, Vol. 25, pp. 140–146, 1986.

10. Hakuraku, Y., and Ogata, H., "A Static Magnetic Refrigerator for Superfluid Helium with New Heat Switches and a Superconducting Pulse Coil," *Jap. J. App. Phys.* Vol. 24, pp. 1538–1547, 1985.

11. Shepherd, I. W., "Electrocaloric Properties of OH-Molecules in KCl," *J. Phys. Chem. Solids*, Vol. 28, pp. 2027–2051, 1967.

12. Korrovits, V. K., Luud'ya, G. G., and Mikhkelsoo, V. T., "Thermostating Crystals at Temperatures Below 1 K by Using the Electrocaloric Effect," *Cryogenics*, Vol. 14, pp. 44–45, 1974.

13. Rietdijk, J. A., "The Expansion-Ejector, a New Device for Liquefaction and Refrigeration at 4 K and Lower," *Liquid Helium Technology, Proceedings of the International Institute of Refrigeration*, Boulder, Colorado, pp. 241–249, Pergamon Press, New York, 1966.

14. Rudolf von Rohr, P., and Trepp, C., "Experimental Investigation of an Ejector," *Cryogenics*, Vol. 25, pp. 684–686, 1985.

15. Wu, K. C., Brown, D. P., Schlafke, A. P., and Sondericker, J. H., "Helium Refrigerator with Features for Operation at Supercritical Pressures," *Advances in Cryogenic Engineering*, Vol. 29, pp. 495–501, Plenum Press, New York, 1984.

16. DeLong, L. E., Symko, O. G., and Wheatley, J. C., "Continuously Operating ^{4}He Refrigerator," *Rev. Sci. Instrum.*, Vol. 42, pp. 147–150, 1971.

17. Collins, S. C., Stuart, R. W., and Streeter, M. H., "Closed-Cycle Refrigeration at 1.85°K," *Rev. Sci. Instrum.*, Vol. 38, pp. 1654–1657, 1967.

18. Wiebe, E. R., "Low-Noise Receivers: 3-Kelvin Refrigerator Development for Improved Microwave Maser Performance," *DSN Progress Report 42-28*, pp. 42–44, August 15, 1975.

P = 1.5 x 10^6 Pa
T = 300 K

COMPRESSOR

EJECTOR

HEAT EXCHANGER

EXPANDER

J-T EXPANSION VALVE

P > 1 x 10^5 Pa
T > 4.2 K

P = 0.5 x 10^5 Pa
T = 3.55 K

REFRIGERATOR

P_1, T_1, $\dot{m}_1$

P_3, T_3, $\dot{m}_1 + \dot{m}_2$

P_2, T_2, $\dot{m}_2$
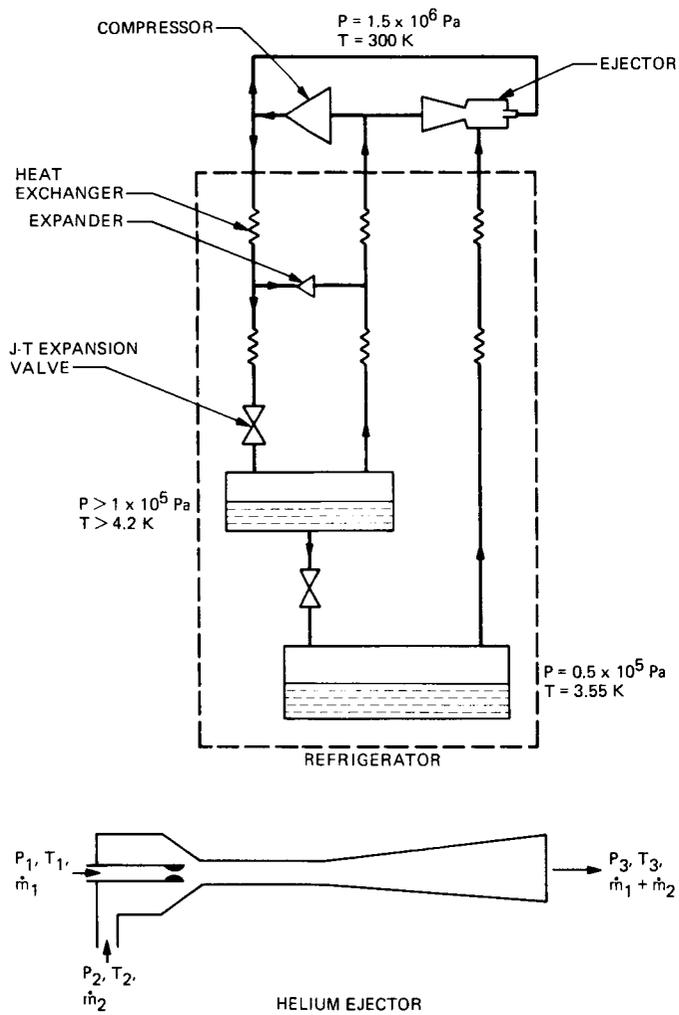
HELIUM EJECTOR

Fig. 1. Helium refrigerator using a room temperature ejector (after Ref. 14)

COMPRESSOR

PRECOOLING STATIONS

$\dot{m}_1$

$\dot{m}_1$

P_1

$\dot{m}_1 + \dot{m}_2$

P_3, T_3

EJECTOR

P_2

$\dot{m}_2$

H.E.

J-T EXPANSION VALVE

P_2, T_2
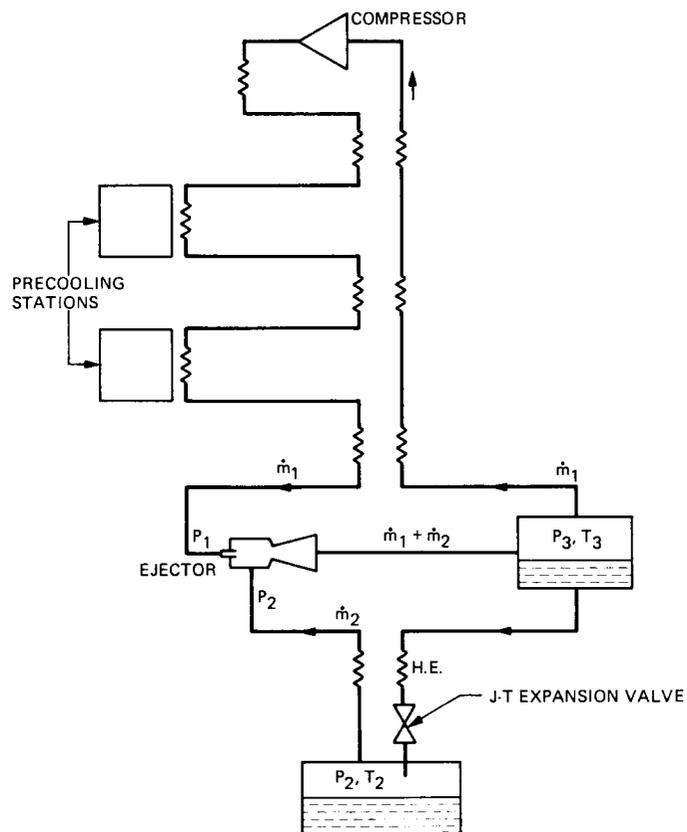
Fig. 2. Diagram of the low-temperature end of a helium refrigerator equipped with a cold ejector (from Ref. 13)
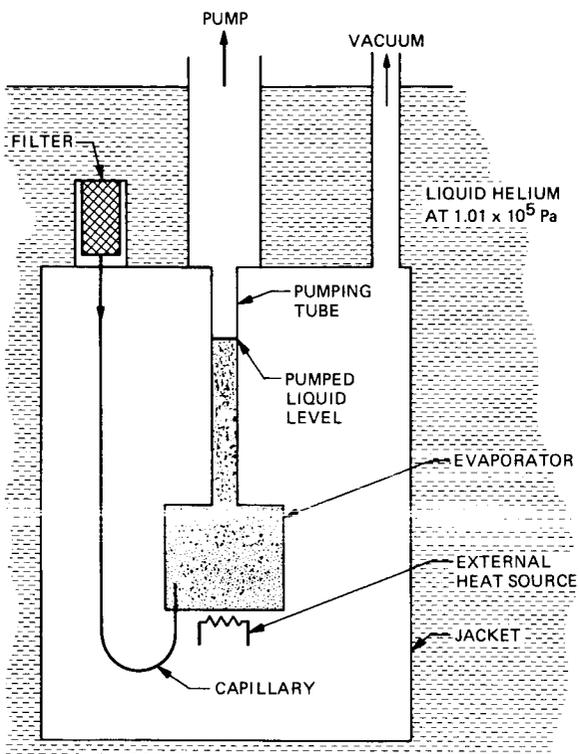
Fig. 3. Schematic of a continuously operating helium refrigerator (after Ref. 16)
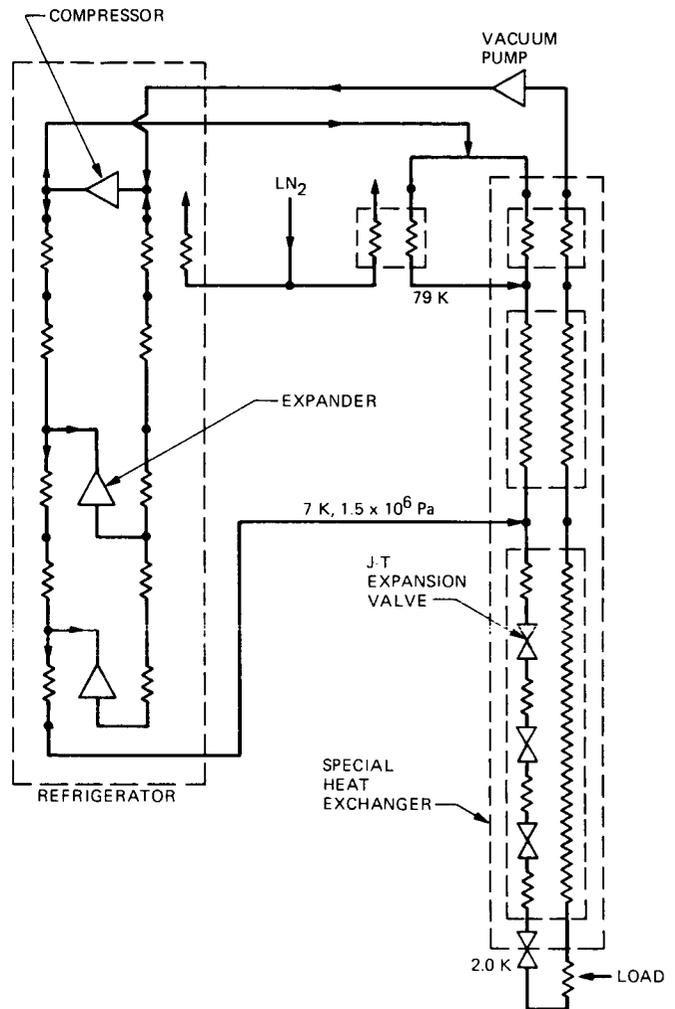
Fig. 4. Test apparatus for a 2-K closed cycle refrigerator (from Ref. 17)
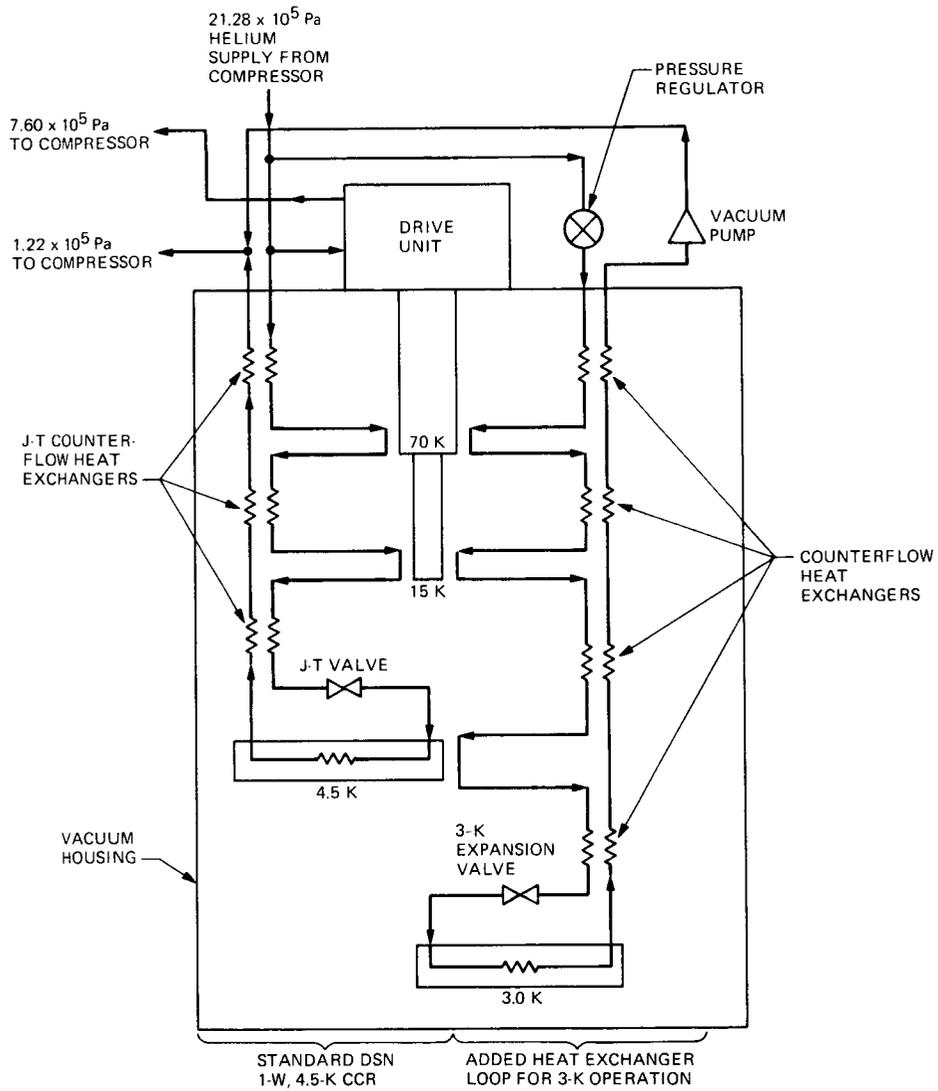
Fig. 5. Schematic of a 3-K closed-cycle refrigerator built at JPL (from Ref. 18)
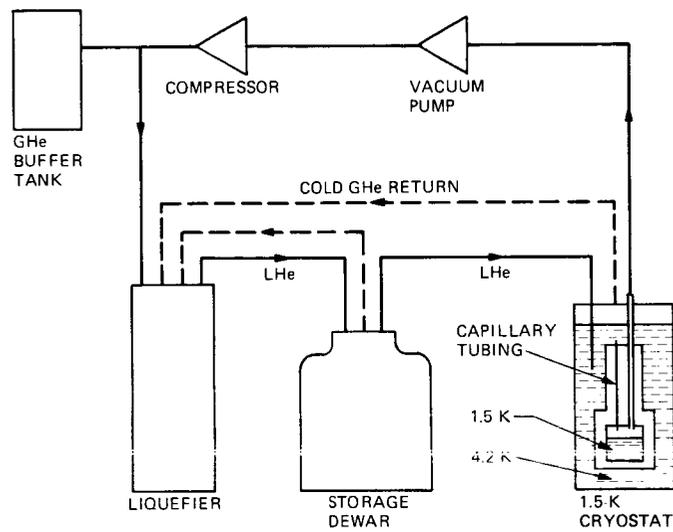
Fig. 6. Block diagram of a 1.5-K refrigerator for a beam waveguide antenna

N87-11031

Single Longitudinal Mode Operation of Semiconductor Laser Arrays with Etalon Control

H. Hemmati
Communication Systems Research Section

A simple method is devised to obtain a single longitudinal output beam from high-power multilongitudinal mode diode laser arrays. Mode control is achieved by simply placing a thin etalon in front of the laser. The three-cavity laser formed by addition of the etalon favors a single longitudinal mode. This technique is applicable to both continuous wave and pulsed laser modes of operation. Experimental results demonstrating the technique along with future work and possible applications are discussed.

I. Introduction

The GaAlAs semiconductor diode array laser and the diode laser pumped Nd:YAG laser are two of the leading candidate sources of radiation for free-space optical communication (Refs. 1 and 2). High electrical-to-optical efficiency along with small size and low mass are some of the features distinguishing these lasers for use in satellites. In general, available diode array lasers have a multilongitudinal mode output. For the heterodyne (coherent) detection technique and various power summing schemes, a single longitudinal mode laser is essential. Also, for such applications as optical pumping of solid-state lasers that have narrow absorption bands, a single mode diode array laser, tuned to a particular absorption band, could be beneficial for more efficient optical excitation.

Most attempts for obtaining single mode array lasers have been directed towards modification of the internal structure of the laser or the use of optical feedback from an external cavity (Ref. 3). When successful, these attempts are limited by low output power from the laser array.

Here a simple method is described to force the multimode output of a laser array into a single longitudinal mode structure. This is accomplished by placing a thin (150- to 250-μm thick) etalon plate at the output port of the diode laser. This etalon (which acts as a resonant reflector) along with the reflective surfaces of the diode laser itself form a three-cavity laser. Resonant operation, where all retroreflections are in phase, results in selection and enhancement of a particular longitudinal mode and simultaneous suppression of the adjacent modes. This method was recently applied to single-element diode lasers with significant (>34 dB) suppression of the adjacent modes (Ref. 4). However, applicability to multimode diode array lasers has not been reported until now. The experimental results presented here pertain to the latter and demonstrate the effectiveness of the method on laser arrays.

II. Experiments

A schematic of the experimental set up is shown in Figs. 1 and 2. The 120-mW output of a 10-element diode gain guided array laser was collimated by a 6.5-mm focal length and 0.615 numerical aperture lens. Three etalon thicknesses were examined: 1-mm, 200-μm, and 150-μm thick solid uncoated etalons. The 150-μm thick plate was just a microscope slide cover glass plate. At such a small thickness, the two faces of the plate are flat and parallel enough to act as an etalon. The 1-mm thick etalon proved ineffective, whereas the Fresnel reflection from either of the two thin ($\leq$200-μm) uncoated glass plates provided sufficient control of the laser mode structure. The etalon was placed on an adjustable mount for fine control. At distances greater than 2 mm from the array (no collimating lens was used), the alignment of the etalon relative to laser facets becomes critical. Just prior to entering the single mode operation, spectral instabilities can be observed due to improper feedback. The addition of a thin uncoated etalon is very power efficient. The overall laser power is reduced by less than 3%.

A typical laser emission spectrum before etalon feedback and after proper etalon alignment is shown in Fig. 3. The measured side mode suppression on this unoptimized setup is about 26 dB. The laser remained single mode for over an hour of examination time. The diode laser was temperature regulated using a thermoelectric cooler. However, single mode operation was maintained with up to 4°C of temperature tuning after which realignment of the etalon was necessary to regain single mode operation. Single mode operation was also maintained when the diode array was pulsed. For pulse rates up to 0.1 MHz, no degradation was observed in the single mode output. Single mode operation is expected to be retained at even higher pulse rates.

The single mode operation method was also examined with an antireflection coated diode laser. Such a diode gain element requires a partially reflecting (typically 50%) output coupler for radiation feedback in order to lase. In this case, the etalon was placed outside the cavity (Fig. 2) where it effectively forces the multilongitudinal mode laser into a single mode. Work is underway to examine this method for the case in which the etalon is placed inside the laser cavity.

Improved efficiency of the diode laser pumped Nd:YAG laser is of major interest. Since the Nd:YAG crystal has a number of narrow absorption bands around the emission wavelength of the laser, a pump laser tuned to one of these bands should result in higher laser power. In preliminary studies where the Nd:YAG crystal was optically pumped with a single mode laser, only 5% improvement in the output power was observed. However, no attempts have been made yet to tune the laser wavelength. Wavelength selection is achievable by diode laser temperature tuning (typically about 3 angstroms per °C), by adjusting the spacing between the laser front facet and the etalon or a combination of both.

For a compact structure and nominal alignment requirement, the etalon may be attached very close to the front facet of the laser, perhaps in place of the window of the protection housing used with many of the commercially available diode lasers. Since the resonant peaks of multiresonant reflectors are sharper and more separated than is the case of a single-element etalon, the effect of using a multiple number of etalons all aligned with the original laser cavity needs to be examined.

III. Conclusion

In conclusion, a simple method is described to obtain single longitudinal mode laser output from multitemporal mode laser arrays with nearly 100% optical efficiency. Work is underway to study the effect of the etalon on the near-field and far-field patterns (the transverse mode structure) of laser arrays, as well as their application to diode array pumping of solid-state lasers.

References

1. Sipes, D. L., "Highly Efficient Nd:YAG Lasers for Free-Space Optical Communications," *TDA Progress Report 42-80*, pp. 31–39, Jet Propulsion Laboratory, Pasadena, Calif., Feb. 15, 1985.

2. Sipes, D. L., "Highly Efficient Nd:YAG Laser End Pumped by a Semiconductor Laser Array," *Appl. Phys. Lett.*, Vol. 47, pp. 74–76, 1985.

3. Yaeli, J., et al, "Array Mode Selection Utilizing an External Cavity Configuration," *Appl. Phys. Lett.*, Vol. 47, pp. 89–91, 1985.

4. Smith, R. J., "Wideband Lasercom Transmitter Performance," *SPIE Optical Technologies for Communication Satellite Applications*, Vol. 616, pp. 100–104, 1986.
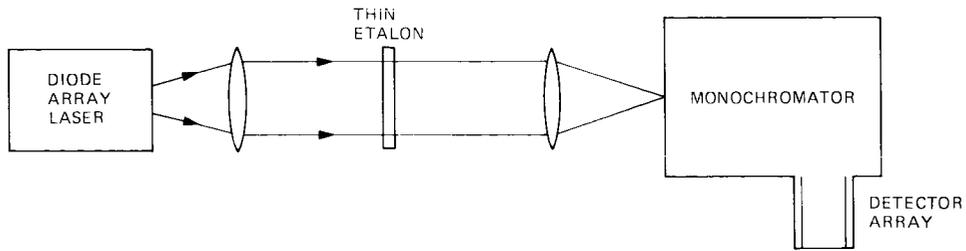
Fig. 1. A Schematic of the experimental setup where a regular ten-element diode
array laser is used
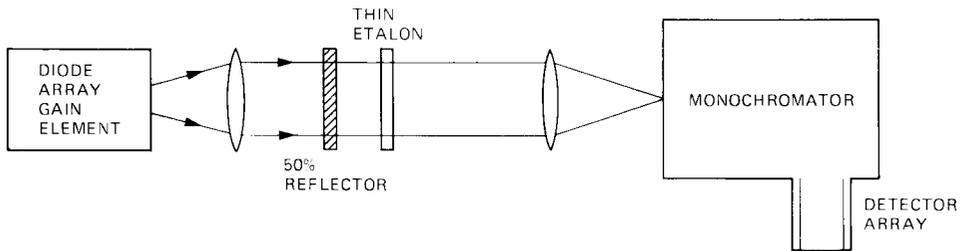
Fig. 2. A schematic of the experimental setup consisting of a front facet antireflection
coated diode array laser (diode array gain element)
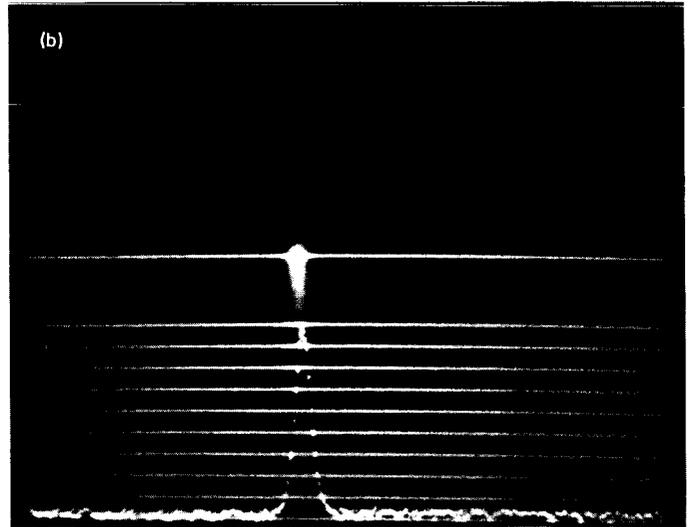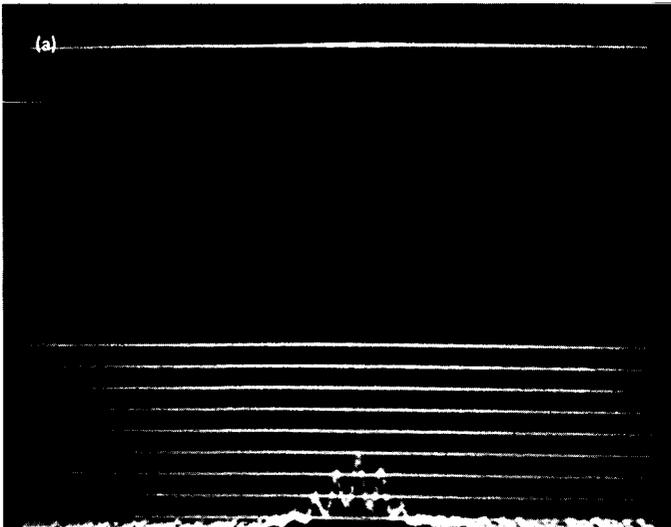
Fig. 3. The spectrum of a ten-element diode array laser as analyzed by a monochromator: (a) before etalon feedback;
(b) after etalon alignment with the array

A Recursive Solution for a Fading Memory Filter Derived From Kalman Filter Theory

J. I. Statman
Communications Systems Research Section

A simple recursive solution for a class of fading memory tracking filters is presented. A fading memory filter provides estimates of filter states based on past measurements, similar to a traditional Kalman filter. Unlike a Kalman filter, an exponentially decaying weight is applied to older measurements, discounting their effect on present state estimates. It is shown that Kalman filters and fading memory filters are closely related solutions to a general least squares estimator problem. Closed form filter transfer functions are derived for a time invariant, steady state, fading memory filter. These can be applied in loop filter implementation of the DSN Advanced Receiver carrier phase locked loop (PLL).

I. Introduction

The problem of estimating system state based on measurements is usually addressed by some form of a least squares estimator (LSE), where a Kalman filter is the common choice. The Kalman filter (Ref. 1) offers a recursive solution for state estimates, as well as for a state estimate covariance matrix. However, most Kalman filter implementations are sensitive to errors in modeling, both in generating a linearized model and in selecting model parameters. The effect of mismodeling is more severe when the system varies with time. Then, as the Kalman filter attempts to fit all past data to a single model, large errors in state estimates occur.

One approach to reducing effects of past data is to use an LSE that applies an exponentially decaying weight to older measurements. This "fading memory" approach, introduced in Ref. 3, overcomes most mismodeling and instability problems associated with standard Kalman filters, at a cost of losing some of the information associated with the discounted past measurements. In addition, the fading memory filter has the advantage that an exponential decay of past data is an intuitively clear concept to most electrical engineers, or anyone who works with basic electrical circuits. Sorensen (Ref. 2) presents a general recursive solution to the linear LSE problem, that with an appropriate selection of parameters can be either a fading memory filter or a standard Kalman filter. Unfortunately, the computational load associated with the fading memory filter implementation of Ref. 2 is excessive compared to what can be achieved.

In this article, the recursive LSE solution proposed in Ref. 2 is simplified for fading memory filters, assuming a particular form of process noise covariance matrix, Q_n. The resulting filter performs well for simulations of real life situa-

tions. The ensuing simplicity of the filter equations, based on a largely intuitive selection of Q_n, may incur some cost in terms of optimality. The simplified equations are then applied to a time invariant system, where filter gains have reached their steady state value, and closed form filter input-to-output transfer functions are derived.

II. System Model and Recursive Solution

The linearized model of a system under consideration is given by Eqs. (1) through (3). Equation (1) describes how the system state vector is propagated from one time point to the next. Equation (2) defines the relationship between system state and present measurement. All differences between state propagation model and behavior of the actual system are represented by the random variable v_{n-1}, while measurement noise is represented by u_n. Usually, $\{v_{n-1}\}$ and $\{u_n\}$ are modeled as zero-mean, white Gaussian random processes, with covariance matrices given by Eq. (3).

$$x_n = \phi_{n,n-1} \, x_{n-1} + v_{n-1} \qquad (1)$$

$$y_n = H_n \, x_n + u_n \qquad (2)$$

$$\left.\begin{array}{c} E[v_{n-1} \, v_{m-1}^T] = Q_{n-1} \, \delta_{m,n} \\[2mm] E[u_n \, u_m^T] = R_n \, \delta_{m,n} \\[2mm] E[v_n \, u_m^T] = 0 \end{array}\right\} \qquad (3)$$

where

$\delta_{m,n}$ The Kronecker delta function

x_n System state vector (at time n)

$\phi_{n,n-1}$ State transition matrix (from time n - 1 to time n)

v_{n-1} Process (or state) noise

y_n Vector of measurements

H_n Measurement transformation matrix

u_n Measurement noise

Q_{n-1} Process noise covariance matrix

R_n Measurement noise covariance matrix

$E(\cdot)$ Statistical expected value

The LSE problem can be stated as follows. Given a set of measurements $\{y_i, \; i = 1 \ldots n\}$ and a weight matrix S_n, find an estimate of the state, $\bar{x}_{n,n}$, that minimizes J_n:

$$J_n = U_n \, S_n^{-1} \, U_n^T \qquad (4)$$

where U_n is the column vector composed of the individual measurement noise vectors $\{u_i, \; i = 1 \ldots n\}$, and S_n is a non-negative definite matrix. In this formulation, the covariance associated with initial state uncertainty is ignored.

The matrix S_n is often defined as a diagonal, or quasi-diagonal matrix, reflecting the stationary nature of the measurement noise processes. When u_n is a scaler, S_n is a diagonal matrix, while when u_n is a k-length vector, S_n is a block diagonal matrix consisting of k by k matrices along the block diagonal, with zeroes elsewhere. Let us first explore the scalar measurement case, with constant measurement noise variance. There are two approaches for selecting the elements of S_n. The first approach is to assign equal weight to all measurements, i.e., S_n is an identity matrix. This approach leads to a standard Kalman filter. The second approach is to degrade older measurements, accounting for less validity of older measurements. In this case, the diagonal elements of S_n, denoted s_i, satisfy:

$$s_i > s_{i-1} \quad i = 2 \ldots n$$

This approach results in a fading memory filter. The two approaches can be easily extended to the cases where each measurement is a k-length vector and measurement noise covariance changes from one time to the next.

As the number of measurements increases, a complete LSE solution (requiring inversion of an nk by nk matrix) becomes computationally unattractive and a recursive form of the algorithm is used. Sorensen proposed Eqs. (5) through (8) as an optimum recursive solution to the LSE problem:

$$\bar{x}_{n,n} = \phi_{n,n-1} \, \bar{x}_{n-1,n-1} + K_n(y_n - H_n \phi_{n,n-1} \, \bar{x}_{n-1,n-1}) \qquad (5)$$

$$P_{n,n-1} = \phi_{n,n-1} \, P_{n-1,n-1} \, \phi_{n,n-1}^T \, e^{c_{n-1}} + Q_{n-1} \qquad (6)$$

$$K_n = P_{n,n-1} \, H_n^T (H_n P_{n,n-1} \, H_n^T + R_n)^{-1} \qquad (7)$$

$$P_{n,n} = P_{n,n-1} - K_n H_n P_{n,n-1} \qquad (8)$$

where

$\bar{x}_{n,n}$ State estimate (at time n)

K_n Filter gain

$P_{n,n}$ State estimate covariance matrix (at time n, including y_n)

$P_{n,n-1}$ State estimate covariance matrix (at time n, but without y_n)

e^{c_n} Decay factor

Typical filter update is as follows. In the nth step, Eq. (6) is evaluated, extrapolating the state estimate covariance matrix in time, and accounting for the process noise covariance matrix. This equation also performs a time decay function, using the multiplier $e^{c_{n-1}}$. Next, filter gain is computed by Eq. (7). Finally, the new measurement is incorporated into the state estimate and the state covariance matrix, using Eqs. (5) and (8), respectively.

When the decay factor is unity, i.e. no degradation of past measurements is used, Eqs. (5) through (8) represent a standard Kalman filter (Ref. 1). On the other hand, when $e^{c_{n-1}} > 1$ for all n, there is actual decay of past measurements, resulting in a fading memory filter.

As seen in the above equations, the computational complexity associated with this implementation of a fading memory filter is at least as high as that of a Kalman filter. However, a significantly simpler filter implementation is derived in Appendix A, for the design value of Q_{n-1} selected according to Eq. (12). The resulting recursion formulas are:

$$\bar{x}_{n,n} = \phi_{n,n-1} \bar{x}_{n-1,n-1} + K_n (y_n - H_n \phi_{n,n-1} \bar{x}_{n-1,n-1}) \quad (9)$$

$$M_n = H_n^T R_n^{-1} H_n + \alpha_n \phi_{n,n-1}^{-T} M_{n-1} \phi_{n,n-1}^{-1} \quad (10)$$

$$K_n = M_n^{-1} H_n^T R_n^{-1} \quad (11)$$

$$Q_{n-1} = \beta_n \phi_{n,n-1} P_{n-1,n-1} \phi_{n,n-1}^T \qquad \beta_n \geq 0 \quad (12)$$

where

M_n Inverse of state estimate covariance matrix

$\alpha_n = \dfrac{1}{e^{c_{n-1}} + \beta_n}$ Filter decay factor, $0 < \alpha_n < 1$

Filter update for the nth measurement consists of updates of M_n, K_n, and $\bar{x}_{n,n}$, in that order. Computations associated with this form of filter update are simpler than corresponding

computations for a Kalman filter. A measure of the simplification is that steady state value of $P_{n,n}$ can be obtained from a linear set of equations, rather than the (quadratic) Riccati equation required for a Kalman filter. The filter decay factor, used in Eq. (10), can be viewed as an exponential:

$$\alpha_n = e^{-\tau_n/T} \quad (13)$$

where T is the filter sample time and τ_n is the filter time constant. With this definition, the filter response to input impulse is somewhat similar to the response of an RC electrical circuit to an impulse, with an exponentially decaying transient. From the above equations it is clear that α_n has a dual role. First, it represents the exponential increase in the state estimate covariance matrix, $e^{c_{n-1}}$. In addition, it also includes the effect of the assumed Q_{n-1}, via a β_n component.

III. Results for Time Invariant, Steady State Filter Gain

Often, the system model assumes that the state transition and measurement matrices and the measurement noise covariance matrix are time invariant, and measurement samples are uniformly spaced in time. It is also assumed that the filter decay factor is constant. In this case, the filter update equations are:

$$\bar{x}_{n,n} = \phi \bar{x}_{n-1,n-1} + K_n (y_n - H \phi \bar{x}_{n-1,n-1}) \quad (14)$$

$$M_n = H^T R^{-1} H + \alpha \phi^{-T} M_{n-1} \phi^{-1} \quad (15)$$

$$K_n = M_n^{-1} H^T R^{-1} \quad (16)$$

In these equations, the unsubscripted ϕ, R, H, and α are the time invariant versions of the corresponding subscripted variables. After a sufficiently long time, the matrices M_n approach a steady state value, M, that depends only on ϕ, H, R, and α. When M_n is approximated by this steady state value, the filter gain defined in Eq. (15) can be precomputed. This results in a significant reduction in the computational load associated with filter updates. Of course there is some loss of flexibility in using constant, steady state, filter gains.

When using steady state filter gain, K, transfer functions from filter input to filter output can be evaluated. The transfer function, in matrix form, is:

$$C(z) = (zI - \phi + KH\phi)^{-1} Kz \quad (17)$$

Note that for an m-input, n-state filter, the matrix $C(z)$ is of dimension n by m.

Analytic steady state tracking filter solutions are often investigated for simple second and third order Kalman filters (Ref. 4). Similar expressions are derived below for fading memory filters. In a typical case, range (or range and velocity) measurements are used in estimating range, velocity, and perhaps acceleration. The resulting ϕ and H, when no velocity measurement is available, are given in Table 1. Without loss of generality, R is assumed to be unity. This can be done since any linear scaling of R causes similar scaling for M_n, but has no impact on the filter gain. Thus, the state update equation is independent of scaling of R.

Table 2 presents the input-to-output transfer function components for these filters, assuming steady state filter gains. It is interesting to notice that the transfer functions have all their poles at $z = -\alpha$, within the unit circle.

Fading memory filters, described by Eqs. (14) - (16), are being investigated for the DSN Advanced Receiver carrier PLL loop filter (Ref. 5), where phase, frequency, and frequency rate correspond to range, velocity, and acceleration. It is expected that these filters, in conjunction with a predictor, will reduce the effect of loop transport lag.

Similar filters were also successfully used in the Mobile Automated Field Instrumentation System (MAFIS) Position Location Demonstration[1,2] and for the High Dynamics GPS Receiver Validation Demonstration (Ref. 6), both performed at JPL.

IV. Conclusions

The fading memory filter and Kalman filter are presented as special cases of a general least squares estimator problem. It is shown that both filters can be implemented by the same set of recursion equations, with an appropriate choice of parameters. A simple recursive solution for a class of fading memory tracking filters is presented. Filter implementation for this class is computationally efficient, and exhibits good stability performance. It is proposed as part of the loop filter for the DSN Advanced Receiver carrier PLL.

[1] Hurd, W. J., *MAFIS Position Location Feasibility Demonstration Final Report* (JPL Internal Document 7011-22), Vol II B.2, March 1982.

[2] Wallis, D. E., private communications.

References

1. Gelb, A., *Applied Optimal Estimation*, Chapter 4, The MIT Press, Cambridge MA, 1974.

2. Sorensen, S. W. and Sacks, J. E., "Recursive Fading Memory Filters," *Information Sciences 3*, pp. 101-119, 1971.

3. Fagin, S. F., "Recursive Linear Regression Theory, Optimal Filter Theory, and Error Analysis of Optimal Systems," *1964 IRE Convention Record*, pp. 216-240, 1964.

4. Ekstrand, B., "Analytic Steady State Solution for a Kalman Tracking Filter," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-19, pp. 815-819, November 1983.

5. Statman, J. I., and Hurd, W. J., "An Estimator-Predictor Approach to the Design of Digital PLL Loop Filter," *TDA Progress Report 42-86*, JPL Propulsion Laboratory, Pasadena, Calif., this issue.

6. Hurd, W. J., Statman, J. I. and Vilnrotter, V. A., "GPS High Dynamic Receiver Tracking Demonstration Results," *Proceedings of the International Telemetering Conference*, pp. 533-545, Las Vega, NV, October 1985.

7. *CRC Standard Mathematical Tables*, 25th Edition, page 32, CRC Press, 1979.

8. Anderson, B. D., and Moore, J. B., *Optimal Filtering*, pp. 135-138, Prentice-Hall, New Jersey, 1979.

Table 1. Matrices for second and third order filters

Order	ϕ	H	R
2	$\begin{Bmatrix} 1 & T \\ 0 & 1 \end{Bmatrix}$	$(1,\,0)$	(1)
3	$\begin{Bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{Bmatrix}$	$(1,\,0,\,0)$	(1)

Table 2. Closed form transfer functions from input range to output parameter

Output parameter	Second order filter	Third order filter
Range	$\dfrac{(1-\alpha^2)\,z\left(z - 2\,\dfrac{\alpha}{1+\alpha}\right)}{(z-\alpha)^2}$	$\dfrac{(1-\alpha)\,z\,((\alpha^2+\alpha+1)\,z^2 - 3\alpha\,(1+\alpha)\,z + 3\alpha^2)}{(z-\alpha)^3}$
Velocity	$\dfrac{(1-\alpha)^2\,z\,(z-1)}{T(z-\alpha)^2}$	$\dfrac{(1-\alpha)\,z\,(z-1)\,((3-3\alpha^2)\,z + 5\alpha^2 - 4\alpha - 1)}{2T\,(z-\alpha)^3}$
Acceleration	—	$\dfrac{(1-\alpha)^3\,z\,(z-1)^2}{T^2\,(z-\alpha)^3}$

Appendix A
Derivation of Simplified Recursive Formulas

This appendix derives simplified fading memory equations, Eqs. (9) through (11), from Sorensen's recursive solution, given in Eqs. (5) through (8). The simplification is accomplished in three steps. First Eqs. (7) and (8) are combined. Then, a specific form of Q_n is assumed, resulting in a simple equation for the update of state estimate covariance matrix. Finally, a new state estimate update equation is derived. The first step uses a matrix inversion lemma (Ref. 7):

$$(B + UV)^{-1} = B^{-1} - B^{-1} U (I + VB^{-1} U)^{-1} VB^{-1} \quad (A-1)$$

Applying Eq. (A-1) to Eq. (8), results in:

$$P_{n,n}^{-1} = P_{n,n-1}^{-1} - P_{n,n-1}^{-1}(-K_n)$$

$$\times (I + H_n P_{n,n-1} P_{n,n-1}^{-1} (-K_n))^{-1}$$

$$\times H_n P_{n,n-1} P_{n,n-1}^{-1} \quad (A-2)$$

or

$$P_{n,n}^{-1} = P_{n,n-1}^{-1} + P_{n,n-1}^{-1} K_n (I - H_n K_n)^{-1} H_n \quad (A-3)$$

but:

$$I - H_n K_n = I - H_n P_{n,n-1} H_n^T$$

$$\times (H_n P_{n,n-1} H_n^T + R_n)^{-1}$$

$$= I - (H_n P_{n,n-1} H_n^T + R_n - R_n) \quad (A-4)$$

$$\times (H_n P_{n,n-1} H_n^T + R_n)^{-1}$$

$$= R_n (H_n P_{n,n-1} H_n^T + R_n)^{-1}$$

combining Eqs. (A-3) and (A-4) results in:

$$P_{n,n}^{-1} = P_{n,n-1}^{-1} + H_n^T (H_n P_{n,n-1} H_n^T + R_n)^{-1}$$

$$\times (H_n P_{n,n-1} H_n^T + R_n) R_n^{-1} H_n$$

$$= P_{n,n-1}^{-1} + H_n^T R_n^{-1} H_n \quad (A-5)$$

Equation (A-5) has a form found commonly in literature (Ref. 4). Next, let us assume that the matrix Q_{n-1} has a special form:

$$Q_{n-1} = \beta_n \phi_{n,n-1} P_{n-1,n-1} \phi_{n,n-1}^T \qquad \beta_n \geqslant 0, \quad (A-6)$$

The rationale for this assumption is discussed at the end of the appendix. With this assumption, Eq. (6) becomes:

$$P_{n,n-1} = (\beta + e^{c_{n-1}}) \phi_{n,n-1} P_{n-1,n-1} \phi_{n,n-1}^T \quad (A-7)$$

Inserting Eq. (A-7) into (A-5), and assuming that $P_{n,n-1}$ and $\phi_{n,n-1}$ are invertible, we get:

$$P_{n,n}^{-1} = H_n^T R_n^{-1} H_n$$

$$+ \frac{1}{\beta + e^{c_{n-1}}} \phi_{n,n-1}^{-T} P_{n-1,n-1}^{-1} \phi_{n,n-1}^{-1} \quad (A-8)$$

Following the notation in the body of the paper, Eq. (A-8) can be represented as:

$$M_n = H_n^T R_n^{-1} H_n + \alpha_n \phi_{n,n-1}^{-T} M_{n-1} \phi_{n,n-1}^{-1} \quad (A-9)$$

where

$$M_n = P_{n,n}^{-1} \qquad \text{Inverse of state estimate covariance matrix}$$

$$\alpha_n = \frac{1}{\beta_n + e^{c_{n-1}}} \qquad \text{Filter decay factor}$$

Note that $e^{c_{n-1}} > 1$ and $\beta_n \geqslant 0$, thus $1 > \alpha_n > 0$. This completes the derivation of the simplified covariance matrix update equation. To complete the proof, the expression from Eq. (15) is now derived for the gain K_n. From Eq. (A-5):

$$M_n = P_{n,n-1}^{-1} + H_n^T R_n^{-1} H_n \quad (A-10)$$

Applying Eq. (A-1) to Eq. (A-10):

$$M_n^{-1} = P_{n,n-1} - P_{n,n-1} H_n^T$$

$$\times (I + R_n^{-1} H_n P_{n,n-1} H_n^T)^{-1} R_n^{-1} H_n^T P_{n,n-1}$$

$$= P_{n,n-1} - P_{n,n-1} H_n^T$$

$$\times (R_n + H_n P_{n,n-1} H_n^T)^{-1} H_n^T P_{n,n-1}$$

$$(A-11)$$

After multiplying both sides of this equation by $H_n^T R_n^{-1}$, and some tedious arithmetic, Eq. (A-11) becomes:

$$M_n^{-1} H_n^T R_n^{-1} = P_{n,n-1} H_n^T (H_n P_{n,n-1} H_n^T + R_n)^{-1} = K_n$$

$$(A-12)$$

This completes the derivation.

The selection of Q_n, as defined by Eq. (A-6), is of particular interest. In most Kalman filter applications, Q_n serves a dual function. First, it represents the modeled process noise, which is its declared objective. Then, it also reduces the risk of numerical instability by establishing a minimal value to the state estimate covariance matrix. The interested reader can evaluate Eqs. (5) through (8) for simple cases, using $Q_n = 0$, and observe that $P_{n,n}$ approaches zero as n goes to infinity. Since, for non-trivial Kalman filters, $P_{n,n}$ must be positive definite, small numerical problems can make the filter unstable. For this reason, Kalman filter designers tend to inflate Q_n above its modeled level.

The fading memory filter does not suffer from a corresponding problem. Even when Q_n is zero, $P_{n,n}$ reaches a finite nonzero value. Conceptually, a Kalman filter $P_{n,n}$ is affected by a larger set of measurements than a fading memory filter $P_{n,n}$ and thus tends to be smaller. Since Q_n is not required for numerical stability, a reasonable choice is for Q_n to be proportional to the state estimate covariance matrix, or to a related quantity, as defined by Eq. (A-6). It can be viewed as an adaptive definition of Q_n. Also, if Q_n is chosen as 0, the decay factor $e^{c_{n-1}}$ defined in this appendix, and the filter decay factor α, are reciprocals of each other.

Anderson and Moore (Ref. 8) suggest that a fading memory filter can be viewed as a Kalman filter with exponential inflation applied to past Q_n and R_n. They also emphasize the inherent stability associated with such a filter. Their approach, though insightful, does not simplify filter mechanization.

In summary, the particular form of Q_n used in this appendix is not derived from an independent statistical model of process noise, thus resulting in a sub-optimal solution. In many real life applications, the benefit associated with numerical stability, reduced sensitivity to mismodeling, and reduced computational load may far outweigh this loss in optimality.

An Estimator–Predictor Approach to PLL Loop Filter Design

J. I. Statman and W. J. Hurd
Communications Systems Research Section

A new approach to the design of digital phase locked loops (DPLLs), using estimation theory concepts in the selection of a loop filter, is presented. The key concept is that the DPLL closed-loop transfer function is decomposed into an estimator and a predictor. The estimator provides recursive estimates of phase, frequency, and higher order derivatives, while the predictor compensates for the transport lag inherent in the loop. This decomposition results in a straightforward loop filter design procedure, enabling use of techniques from optimal and sub-optimal estimation theory. A design example for a particular choice of estimator is presented, followed by analysis of the associated bandwidth, gain margin, and steady state errors caused by unmodeled dynamics. This approach is under consideration for the design of the DSN Advanced Receiver Carrier DPLL.

I. Introduction

A phase locked loop (PLL) is a feedback system that tracks the phase of a received signal. A typical PLL structure is shown in Fig. 1: the difference between the received phase and a voltage controlled oscillator (VCO) phase is filtered and used to steer the VCO, so that the VCO phase tracks the received phase. Designers of PLLs select a loop filter to meet performance requirements, usually specified in terms of bandwidth, gain margin, and dynamic errors. Design methods for analog PLLs are well documented in the literature (Refs. 1 and 2).

It is assumed here that the DPLL is characterized by constant loop update rate. Discussion of DPLLs with variable loop update rate is provided in Refs. 3-6. DPLLs with fixed update rate can be designed either by analogy to continuous domain PLLs or based on an optimality criterion.

Design by analogy to continuous time PLLs, i.e., applying s-plane design rules by translation to the z-plane, suffers from two major disadvantages. The design is based on iteratively placing open-loop poles and zeroes at "well chosen" locations until satisfactory performance is achieved. Such locations are selected based on design experience rather than on a well established set of rules. Also, the design does not account for the transport delay in the digital loop, caused by hardware and loop filter computations. This delay is either "tolerated," i.e., the degradation due to the delay is analyzed and found to be acceptable (Ref. 7), or "compensated" by the equivalent of a lead–lag network (Ref. 8).

Design based on an optimality criterion attempts to optimize a concise performance measure, usually a weighted sum of transient response, gain margin, and noise bandwidth. For some cases (Ref. 9), analytical expressions for the performance measure can be obtained, and closed-form expressions for loop filters defined.

The estimator–predictor approach described in this article offers an alternate method for DPLL design. The DPLL is shown to be equivalent to an estimator followed by a predictor, compensating for the transport lag. The estimator derives an estimate of the state vector, consisting of phase, frequency, and perhaps frequency rate, based on measurements of phase alone. Selection of the specific estimator can then be based on the vast resources available in linear optimal estimation theory. In the next sections, a generic form of the design procedure is presented, followed by a design example using specific estimator, predictor, and for specified transport lags. Then, performance of the designed loops is analyzed. A future study will compare performance of loops designed using this approach to those designed according to alternative criteria.

II. Design Approach

This section discusses the rationale for the new design approach and then presents the resulting design procedure. Time dependent variables, e.g., phase, are represented by their z-plane transforms, and difference equations are replaced by corresponding z-plane transfer functions. Even though the notation does not explicitly account for time variation of difference equation coefficients, such variations can be accommodated by the design procedure.

A. Rationale

The basic purpose of a DPLL is to generate a signal, with phase $\overline{\theta}(z)$, that approximates $\theta(z)$, the phase of a received signal, as shown in Fig. 1. From a linear estimation theory viewpoint, this is a pure and simple estimation problem, and the solution is straightforward. First, a linear state model for $\theta(z)$ is defined, with a corresponding measurement model. Then, statistical models for state (or process) noise and measurement noise are developed. Finally, an appropriate estimator is selected (a fixed-coefficient, recursive least squares estimator is a prime candidate).

There is a slight complication since a non-zero transport delay is associated with DPLL mechanization. The effect of this delay can be compensated by a predictor that extrapolates the state estimate by an appropriate time interval. Thus, the closed-loop DPLL transfer function has the form:

$$H(z) = D(z)C(z)z^{-N} \qquad (1)$$

where

$H(z)$ = closed-loop DPLL transfer function

$D(z)$ = predictor transfer function

$C(z)$ = estimator transfer function

N = DPLL transport lag, in units of loop update time

Equation (1) assumes that the DPLL is updated at uniformly spaced time points, with update interval of T, hence the transport lag is an integer multiple of T. Figure 2 shows a DPLL structure that addresses Eq. (1). The three functional blocks on the right are the elements of $H(z)$: estimator, predictor, and transport lag. In addition, the figure incorporates a phase detector, with gain A, and another summing junction. Note that, when phase detector gain has a nominal value of $A = 1$, the feedback is cancelled, and closed-loop transfer function is $H(z)$, as in Eq. (1). The closed-loop transfer function, for general phase detector gain, is denoted $P(z)$.

The transformation of Fig. 2 into an implementation-oriented structure is presented in Figs. 3 and 4. In Fig. 3, the estimator–predictor–delay combination is replaced by actual components: a digital filter, $S(z)$, and a numerically controlled oscillator, $Q(z)$, where $S(z)$ satisfies:

$$S(z) = \frac{D(z)C(z)}{Q(z)} z^{-N} \qquad (2)$$

Finally, Fig. 4 illustrates a proposed DPLL implementation. Phase detection and "hardware numerically controlled oscillator (NCO)" functions are performed by custom circuitry, while summing junction, $S(z)$, and "software NCO" functions are executed in a computer. Both "hardware NCO" and "software NCO" have identical mathematical representations, as the latter simulates the operation of the former. When comparing Figs. 1 and 4, it is apparent that the function of the loop filter is performed, in Fig. 4, by the inner loop, implemented in the computer-resident portion of the DPLL.

B. Design Procedure

The design can be divided into five distinct steps. The first step is to select a model for the received phase process. This model can include the process dynamics and statistics on process noises and measurement noise. The second step is to model the hardware: NCO, transport lag, and phase detector. As shown later, the transport lag usually can be incorporated as part of the transfer function of the NCO, $Q(z)$.

The next step is selection of an estimator, based on phase model and process noise and measurement noise covariance matrices. Possible realizations of the estimator are some varia-

tions of Kalman filters, or other least squares estimators. The fourth step is determination of a predictor to compensate for transport lag. The last, but definitely not the least, step is to assure that the loop is stable.

III. Design Example

In this section, the design approach presented above is implemented for a particular case.

A. Model of Received Phase

Two separate models are used for the received phase. The first model has two state variables: phase and frequency, while the second model uses three state variables: phase, frequency and frequency rate. Later, these models will be addressed as second- and third-order models, respectively.

B. Model of Phase Detector and NCO

The phase detector inputs are actually time continuous signals. The phase detector measures the average phase difference between the phase of the input process and the NCO phase. This is equivalent to the difference between the phase of the input process and the NCO phase. Thus, $\theta(z)$ represents the input phase, averaged over the measurement interval, and $\bar{\theta}(z)$ represents the NCO phase, averaged over the same interval.

Hurd and Aguirre (Ref. 7) and Simon and Mileant (Ref. 8) discuss a general form for a DPLL NCO. The NCO is updated gT seconds after the phase measurement is made, where T is the loop update interval and $0 \leqslant g \leqslant 1$. The case $g=0$ corresponds to NCO update immediately following the measurement, while $g=1$ is caused by a T-second delay before the NCO update. Such a delay occurs in practical systems while the computer determines the next NCO input. $Q(z)$ is the transfer function between the NCO input and the average NCO phase at the next measurement, including all delays. This transfer function is:

$$Q(z) = T \frac{(1-g)^2 z^2 + (1 + 2g - 2g^2)z + g^2}{2z^2(z-1)} \qquad (3)$$

In particular, for the cases $g=0$ and $g=1$, $Q(z)$ has the simple forms:

$$g = 0: \qquad Q(z) = \frac{T}{2} \frac{z+1}{z-1} z^{-1} \qquad (4)$$

$$g = 1: \qquad Q(z) = \frac{T}{2} \frac{z+1}{z-1} z^{-2} \qquad (5)$$

We observe that in Eqs. (4) and (5), the NCO function can be separated into two parts: integer number of delays, and a kernal, $Q_1(z)$. $Q_1(z)$ is characterized by a numerator and a denominator of the same degree. The cases of $g=0$ and $g=1$ are then examples of $N=1$ and $N=2$ computational delays, respectively. The general expressions for a delay of N update times is:

$$Q(z) = \frac{T}{2} \frac{z+1}{z-1} z^{-N} = Q_1(z)z^{-N} \qquad (6)$$

In particular, cases with $N = 3,4 \ldots$ correspond to pipelined feedback computations.

C. Selection of an Estimator

For this example, we select a least squares, fading memory estimator, with fixed gains (Ref. 10). The estimator computes a state estimate based on a set of measurements, applying an exponentially decaying weight to past data. This "aging function" effectively discards measurements that are older than three or four estimator time constants. The estimator transfer functions for fixed or steady state gain are presented in Table 1, and depend on a single parameter, α, a decay factor:

$$\alpha = e^{-\frac{T}{\tau}} \qquad (7)$$

where τ is the estimator time constant.

D. Selection of a Predictor

The predictor generates a phase predict by extrapolating the current phase estimate using either an estimate of frequency, for a second-order model, or estimates of frequency and frequency rate, for a third-order model. The extrapolation time, NT, is the number of integer delays in the loop multiplied by the update interval. Predictor equations are given in Table 2.

E. Loop Transfer Functions

Recall from Eq. (1) that the transfer function $H(z)$ equals $D(z)C(z) z^{-N}$. Thus, in this example, $H(z)$ depends only on the order of the phase model, the number of delays, and the estimator decay factor, α. The open loop, $G(z)$, closed loop, $P(z)$, and error, $E(z)$, transfer functions for the DPLL are easily expressed in terms of the function $E_1(z) = 1 - H(z)$:

$$G(z) = \frac{\bar{\theta}(z)}{\phi(z)} = \frac{AH(z)}{1 - H(z)} = \frac{A(1 - E_1(z))}{E_1(z)} \qquad (8)$$

$$P(z) = \frac{\bar{\theta}(z)}{\theta(z)} = \frac{G(z)}{1 + G(z)} = \frac{AH(z)}{1 + (A-1)H(z)}$$

$$= \frac{A(1 - E_1(z))}{1 + (A-1)(1 - E_1(z))} \quad (9)$$

$$E(z) = \frac{\phi(z)}{\theta(z)} = 1 - P(z) = \frac{E_1(z)}{1 + (A-1)(1 - E_1(z))} \quad (10)$$

where $\phi(z)$ is the z-transform of the phase out of the phase detector, not including phase detector gain, i.e., $\phi(z) = \theta(z) - \bar{\theta}(z)$. Expressions for $E_1(z)$ are presented in Table 3. A special case of interest occurs when the phase detector gain, A, has its nominal value, $A = 1$. Then:

$$G(z) = \frac{H(z)}{1 - H(z)} \; ; \quad P(z) = H(z) \; ; \quad E(z) = E_1(z)$$
$$(11)$$

F. Realizability and Stability

The DPLL structure of Fig. 4 has three blocks that need to be realized: a "hardware NCO," a "software NCO" and a digital filter $S(z)$. The "hardware NCO" is realizable by definition. The "software NCO" is a mere simulation of the "hardware NCO" function and is thus realizable. This leaves the digital filter $S(z)$. A necessary and sufficient condition for the realizability of this filter is that the order of the numerator is less or equal to that of the denominator (Ref. 11). Recall, from Eqs. (2) and (6) that $S(z) = D(z) C(z)/Q_1(z)$. Since the above condition is satisfied for each component of $C(z)$ (Ref. 10), it is also true for $D(z) C(z)$ (the predictor merely computes a fixed weight sum of components of $C(z)$). The above condition is also satisfied for $1/Q_1(z)$ (see Eq. 6). Thus, $S(z)$ is realizable.

In this section, we restrict the stability discussion to the nominal operating point, $A = 1$, with analysis of the range of stable operating conditions deferred to Section IV. When $A = 1$, the closed loop function is:

$$P(z) = H(z) = 1 - E_1(z) \quad (12)$$

Equation (12), combined with Table 3, shows that $P(z)$ has poles at $z = \alpha$ and at $z = 0$. Since α is real and satisfies $0 < \alpha < 1$, all the poles of $P(z)$ are inside the unit circle and the closed-loop transfer function is stable. Gain margin is computed in the next section.

Even though $P(z)$ is stable, special care must be exercised in the implementation of $S(z)$. Note that $Q_1(z)$ has a zero at $z = -1$; thus, $S(z)$ has a pole at that point. This means that $S(z)$, were it not a part of the feedback loop, would be mar-

ginally stable. $H(z)$ is stable since the pole at $z = -1$, contributed by $S(z)$, is cancelled by the zero of $Q(z)$ at the same location. To maintain numerical stability, this pole of $S(z)$ may be moved slightly into the unit circle. In Section IV, we demonstrate that this shift in pole location has negligible effect on loop performance.

IV. Performance Analysis

In this section, performance of the loops is analyzed from three aspects. First, loop type and errors due to dynamics of the received phase are evaluated, then loop bandwidths are determined, and, finally, gain margins are computed. Discussions of dynamic errors and loop bandwidths is restricted to the nominal operating point, A = 1.

A. Loop Type and Dynamic Tracking Errors

Loop type is defined as the number of integrators in the open loop transfer function, i.e., the number of poles of $G(z)$ at $z = 1$. At the operating point, $G(z)$ is equal to $H(z)/E_1(z)$ forcing the poles of $G(z)$ to be either poles of $H(z)$ or zeroes of $E_1(z)$. However, $H(z)$ has no poles at $z = 1$, and $E_1(z)$ has as many zeros at $z = 1$ as the order of the estimator; thus, second- and third-order phase models correspond to type II and type III loops, respectively. We observe that while loop type is independent of the number of computational delays, order of polynomials in the transfer functions strongly depends on the number of delays. Specifically, the orders of the numerator and denominator of $E_1(z)$ are equal to the loop type plus $N - 1$, where N is the number of delays.

The steady state phase error for constant dynamics is computed using the final value theorem:

$$E_\infty = \lim_{z \to 1} \frac{z-1}{z} E(z) \theta(z) \quad (13)$$

It is well known that a type II loop has zero phase error for phase ramp, constant (non-zero) phase error for phase acceleration, and infinite phase error for phase jerk, or higher dynamics. Similarly, a type III loop has zero, constant, and infinite phase errors for phase acceleration, jerk, and higher dynamics, respectively. Thus, the dynamics of interest are phase acceleration for a type II loop, and phase jerk for type III loop. Table 4 lists the steady state phase errors for the types II and III loops, at the operating point $A = 1$. Figure 5 shows these phase errors as functions of the normalized value of the estimator time constant. The inputs are unit phase acceleration, e.g., rad/s^2, for a type II loop and unit phase jerk for a type III loop, and the errors are in consistent units. As the estimator time constant increases, the phase error increases. In fact, for $T \ll \tau$ and $N = 1$, the normalized phase

errors are τ^2 and τ^3 for type II and type III loops, respectively. Phase errors also increase slightly with increase in the number of delays. This is caused by the increased prediction error.

One must remember that, were a predictor not included, increased delays would have caused catastrophic performance degradation. As an exercise, the interested reader can derive expressions for $E_1(z)$, when no prediction is used, i.e.,

$$D(z) = (1,0) \quad or \quad D(z) = (1,0,0) \tag{14}$$

and prove that both second- and third-order estimators result in type I loops.

B. Bandwidth Results

One-sided bandwidth of the loop is defined as:

$$B_L = \frac{1}{2T} \frac{1}{|P(1)|^2} \frac{1}{2\pi j} \int_{\text{unit}} P(z) P(z^{-1}) \frac{dz}{z} \tag{15}$$

where $P(z)$ is the closed-loop transfer function, and the integral is evaluated along the unit circle. At the desired operating point, $A = 1, P(z) = H(z)$, and the integral can be written as:

$$B_L T = \frac{1}{2} \int_0^1 \left| H(e^{\angle \pi/x}) \right|^2 dx \tag{16}$$

The quantity $B_L T$ is also called normalized bandwidth. Equation (16) uses the fact that in our example, at $z = 1$, $H(z)$ is unity, as can be verified from the definitions in Section II. The integral was numerically evaluated for different estimators and delays. Figure 6 shows the bandwidth as a function of the estimator time constant and N for type II and type III loops. We observe that the introduction of extra delays in the loop increases the noise bandwidth for a fixed time constant, but the effect of the delays is less significant as the estimator time constant increases.

C. Stability and Gain Margin

Let us first discuss the closed-loop root loci for the designed DPLL. Recall the definition of the open loop transfer function in Eq. (8). The open-loop transfer function $G(z)$ has zeros at the zeros of $H(z)$ and poles at the zeros of $E_1(z)$; thus, with increase of the phase detector gain A, the poles of $P(z)$ move from the zeros of $E_1(z)$ to the zeros of $H(z)$.

Figure 7 shows root-locus plots for type II and III loops, for $N = 2$. The transfer function $E_1(z)$, for a type II loop, has

three zeros, two at $z = 1$ and one at $z = -2(1 - \alpha)$. In Fig. 7(a) we observe that at low gain, the poles of $P(z)$ are near these points; then, as the gain increases, they move within the unit circle. Eventually, with higher gains, the poles of $P(z)$ go outside the unit circle and stability is lost. The root locus for a type III loop, illustrated in Fig. 7(b), has a similar form, except that at low gain roots may occur outside the unit circle, near $z = 1$.

As shown previously, the design can be marginally stable due to cancellation of a pole of $S(z)$ by a zero of the NCO. To assure stability, this $S(z)$ pole at $z = -1$ is slightly shifted to the inside of the unit circle. Effect of this shift in pole location is demonstrated in Figs. 7(c) and 7(d). These figures repeat the cases given in Figs. 7(a) and 7(b), with the pole shifted to $z = -0.98$. There is an extra root locus branch from $z = -1.0$ toward $z = -0.98$, with negligible effect on the rest of the root locus. In fact, since the whole root locus branch is inside the unit circle, an arbitrarily small shift in pole location is sufficient to guarantee loop stability. This small shift in pole location has a negligible effect on the three loop parameters of interest: dynamic phase error, gain margin, and noise bandwidth.

Locations of the poles of the closed-loop transfer function at the operating point are of special importance. These poles are commonly selected to be on the real axis to avoid instability and oscillations when the gain A changes slightly. At the operating point, $A = 1$, the closed loop transfer function is $H(z)$. For the estimator–predictor in our example, $H(z)$ has poles at $z = 0$ and two or three co-located poles at $z = \alpha$ for the second- and third-order estimators, respectively. This feature, real closed-loop transfer function poles at the operating point, assures stability.

Gain margin measures the effect of the phase detector gain, A, on the stability of the closed loop. It is defined as the ratio of the critical gain, i.e., gain that forces the poles of the transfer function outside the unit circle, to the nominal unity gain. Figure 8 shows the relationship between upper gain margin and normalized bandwidth. Here, as the loop becomes wider, the gain margin decreases. Similar results are presented in Fig. 9 for the lower gain margin of the type III loop. Note that the lower gain margin for the type II loop is infinity. These figures are useful for quick design evaluation, e.g., finding a favorable trade-off between any two variables.

D. Approximation Formulas

Figures 8 and 9 presented a trade-off between gain margin and normalized loop bandwidth, $B_L T$. Similar results can be obtained for a trade-off between dynamic errors and $B_L T$. Since dynamic phase error, E_∞, is approximately inversely

proportional to the second or third power of the normalized bandwidth, we define phase error coefficients, C_{ϕ_2} and C_{ϕ_3}, so that the steady state phase errors due to unit acceleration or jerk are:

Second-order estimator: $E_\infty = \left(\dfrac{C_{\phi_2}}{B_L}\right)^2$ (17)

Third-order estimator: $E_\infty = \left(\dfrac{C_{\phi_3}}{B_L}\right)^3$ (18)

These coefficients exhibit slower variation with $B_L T$ than do the dynamic phase errors. Figure 10 shows the phase coefficients as functions of normalized bandwidths. The phase coefficients, for narrow normalized bandwidths, are less than 1 for type II loops and slightly over 1 for type III loops, and they increase slowly with increased bandwidths.

Approximation formulas were also derived for dynamic phase error and gain margin as functions of the one-sided bandwidth, B_L. The resulting equations are valid for $B_L T < 0.1$ and $N = 2$, though the dependence of N is weak. Table 5 summarizes the equations.

V. Conclusions

Using an estimation theory approach, design of a DPLL can be accomplished in a systematic procedure rather than a trial-and-error approach. The examples presented here, second- or third-order fixed gain, fading memory estimators, are useful for many applications, and the resulting design curves may thus be directly applied. It is proposed that the design procedure be applied to the DSN Advanced Receiver Carrier DPLL.

C-2

References

1. Gardner, F. M., *Phaselock Techniques*, John Wiley and Sons, 1979.

2. Best, R. E., *Phase Locked Loops, Theory, Design and Applications*, Mcgraw Hill, 1984.

3. Gill, G. S., and Gupta, S. C., "First Order Discrete Phase Locked Loops with Applications to Demodulation of Angle Modulated Carrier," *IEEE Trans. on Communication Technology*, Vol. COM-20, pp. 454-462, June 1972.

4. Weinberg, A., and Liu, B., "Discrete Time Analysis of Nonuniform Sampling First- and Second-Order Digital Phase Locked Loops," *IEEE Trans. on Communication Technology*, Vol. COM-22, pp. 123-137, Feb. 1974.

5. Osborne, H. C., "Stability Analysis of Nth Power Digital Phase Locked Loop — Part I: First Order DPLLs," *IEEE Trans. on Communication Technology*, Vol. COM-28, pp. 1343-1354, Aug. 1980.

6. Osborne, H. C., "Stability Analysis of Nth Power Digital Phase Locked Loop — Part II: Second- and Third-Order DPLL's," *IEEE Trans. on Communication Technology*, Vol. COM-28, pp. 1355-1364, Aug. 1980.

7. Aguirre, S., and Hurd, W. J., "Design and Performance of Sampled Data Loops for Subcarrier and Carrier Tracking," *TDA Progress Report 42-79*, pp. 81-95, Jet Propulsion Laboratory, Pasadena, CA, Nov. 1984.

8. Simon, M. K., and Mileant, A., "Performance of the DSA's Subcarrier Demodulation Digital Loop," *TDA Progress Report 42-80*, pp. 180-194, Jet Propulsion Laboratory, Pasadena, CA, Feb. 1985.

9. Kumar, R., and Hurd, W. J., "A Class of Optimum Digital Phase Locked Loops for the DSN Advanced Receiver," *TDA Progress Report 42-83*, pp. 63-80, Jet Propulsion Laboratory, Pasadena, CA, Nov. 1985.

10. Statman, J., "A Recursive Solution for a Fading Memory Filter Derived From Kalman Filter Theory," TDA Progress Report 42-86 (this issue).

11. Oppenheim, A. V., and Schafer, R. W., *Digital Signal Processing*, Prentice-Hall, N.J., 1975.

Table 1. Closed form transfer functions from input phase to output parameter

Outside Parameter	Second-Order Estimator	Third-Order Estimator
Phase	$\dfrac{(1-\alpha^2)\,z\left(z - 2\,\dfrac{\alpha}{1+\alpha}\right)}{(z-\alpha)^2}$	$\dfrac{(1-\alpha)\,z((\alpha^2+\alpha+1)\,z^2 - 3\alpha(1+\alpha)\,z + 3\alpha^2)}{(z-\alpha)^3}$
Frequency	$\dfrac{(1-\alpha)^2\,z\,(z-1)}{T(z-\alpha)^2}$	$\dfrac{(1-\alpha)\,z\,(z-1)((3-3\alpha^2)\,z + 5\alpha^2 - 4\alpha - 1)}{2T(z-\alpha)^3}$
Frequency rate	—	$\dfrac{(1-\alpha)^3\,z\,(z-1)^2}{T^2(z-\alpha)^3}$

Table 2. Predictor formulas

Predictor	Second-Order Estimator	Third-Order Estimator
$D(z)$	$(1, NT)$	$\left(1, NT,\ \dfrac{(NT)^2}{2}\right)$

Table 3. Formulas for $E_1(z)$

N	Second-Order Estimator	Third-Order Estimator
1	$\dfrac{(z-1)^2}{(z-\alpha)^2}$	$\dfrac{(z-1)^3}{(z-\alpha)^3}$
2	$\dfrac{(z-1)^2}{(z-\alpha)^2}\ \dfrac{z+2(1-\alpha)}{z}$	$\dfrac{(z-1)^3}{(z-\alpha)^3}\ \dfrac{z+3(1-\alpha)}{z}$
3	$\dfrac{(z-1)^2}{(z-\alpha)^2}\ \dfrac{z^2+2(1-\alpha)z+(\alpha^2-4\alpha+3)}{z^2}$	$\dfrac{(z-1)^3}{(z-\alpha)^3}\ \dfrac{z^2+3(1-\alpha)z+(3\alpha^2-9\alpha+6)}{z^2}$

Table 4. Steady state phase errors due to dynamics (phase units)

N	Second-Order Estimator $\left(\theta(t) = \frac{1}{2}t^2\right)$	Third-Order Estimator $\left(\theta(t) = \frac{1}{6}t^3\right)$
1	$\dfrac{T^2}{(1-\alpha)^2}$	$\dfrac{T^3}{(1-\alpha)^3}$
2	$\dfrac{T^2}{(1-\alpha)^2}(3-2\alpha)$	$\dfrac{T^3}{(1-\alpha)^3}(4-3\alpha)$
3	$\dfrac{T^2}{(1-\alpha)^2}(\alpha^2-6\alpha+6)$	$\dfrac{T^3}{(1-\alpha)^3}(3\alpha^2-12\alpha+10)$

Table 5. Approximation formulas

DPLL Parameter	Second-Order Estimator	Third-Order Estimator
Gain margin (g_m)	$\dfrac{0.53}{(B_L T)^{0.9}}$	$\dfrac{0.89}{(B_L T)^{0.75}}$
Dynamic lag (E_∞)	$a\left(\dfrac{0.87}{B_L}\right)^2$	$j\left(\dfrac{1.17}{B_L}\right)^3$

E_∞ = units (e.g., rad)

a = acceleration, units/s^2 (e.g., rad/s^2)

j = jerk, units/s^3 (e.g., rad/s^3)
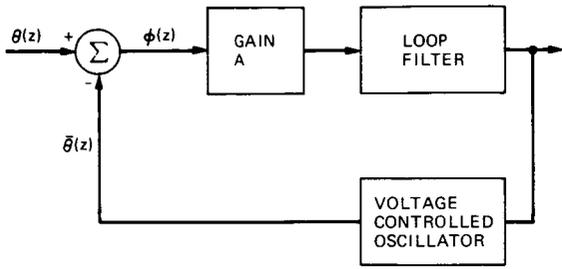
Fig. 1. Traditional phase-locked loop

Fig. 2. DPLL functional block diagram

Fig. 3. DPLL physical block diagram
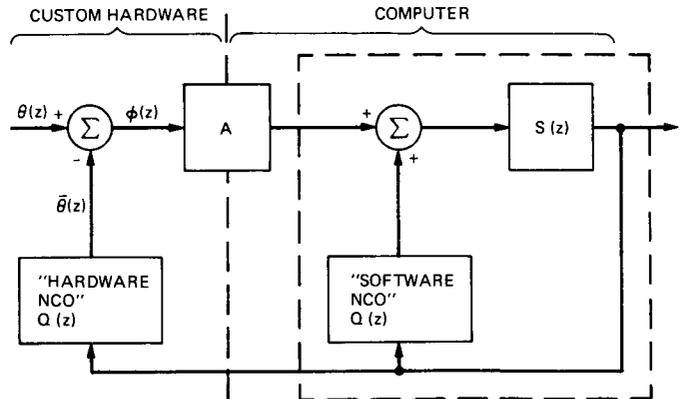
Fig. 4. Proposed DPLL implementation
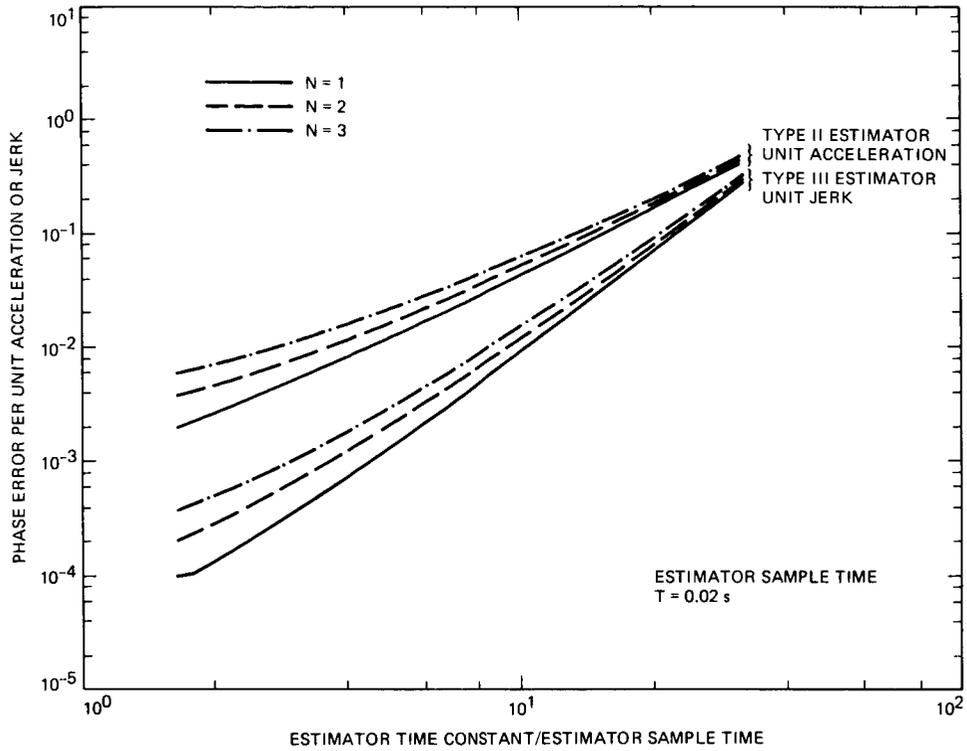
Fig. 5. Dynamic phase error versus estimator parameters
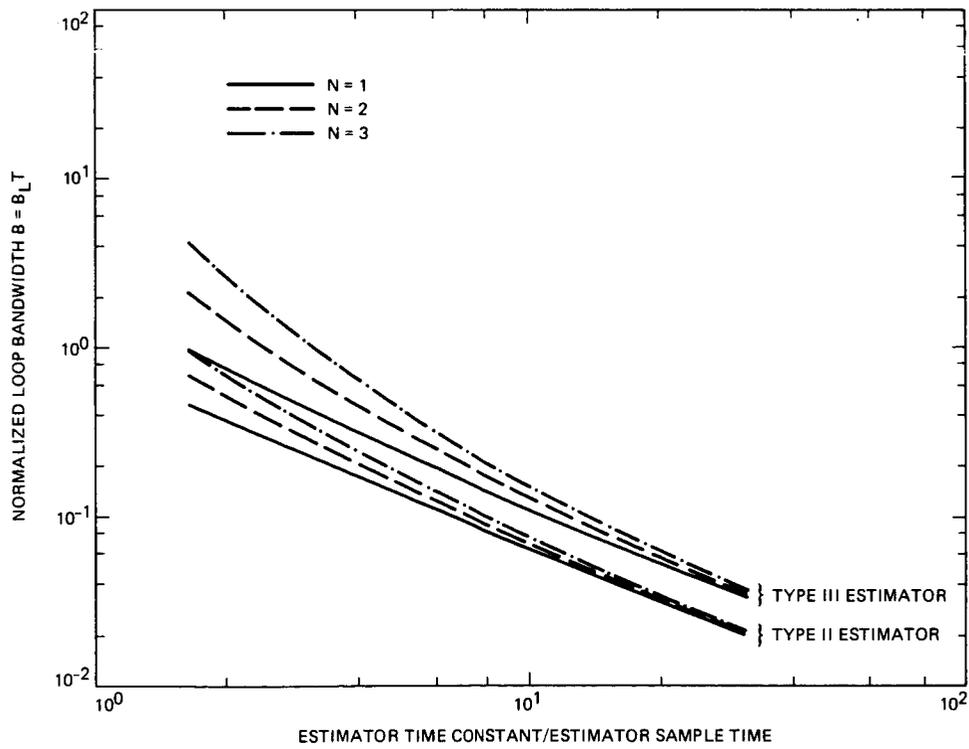
Fig. 6. Loop bandwidth versus estimator parameters
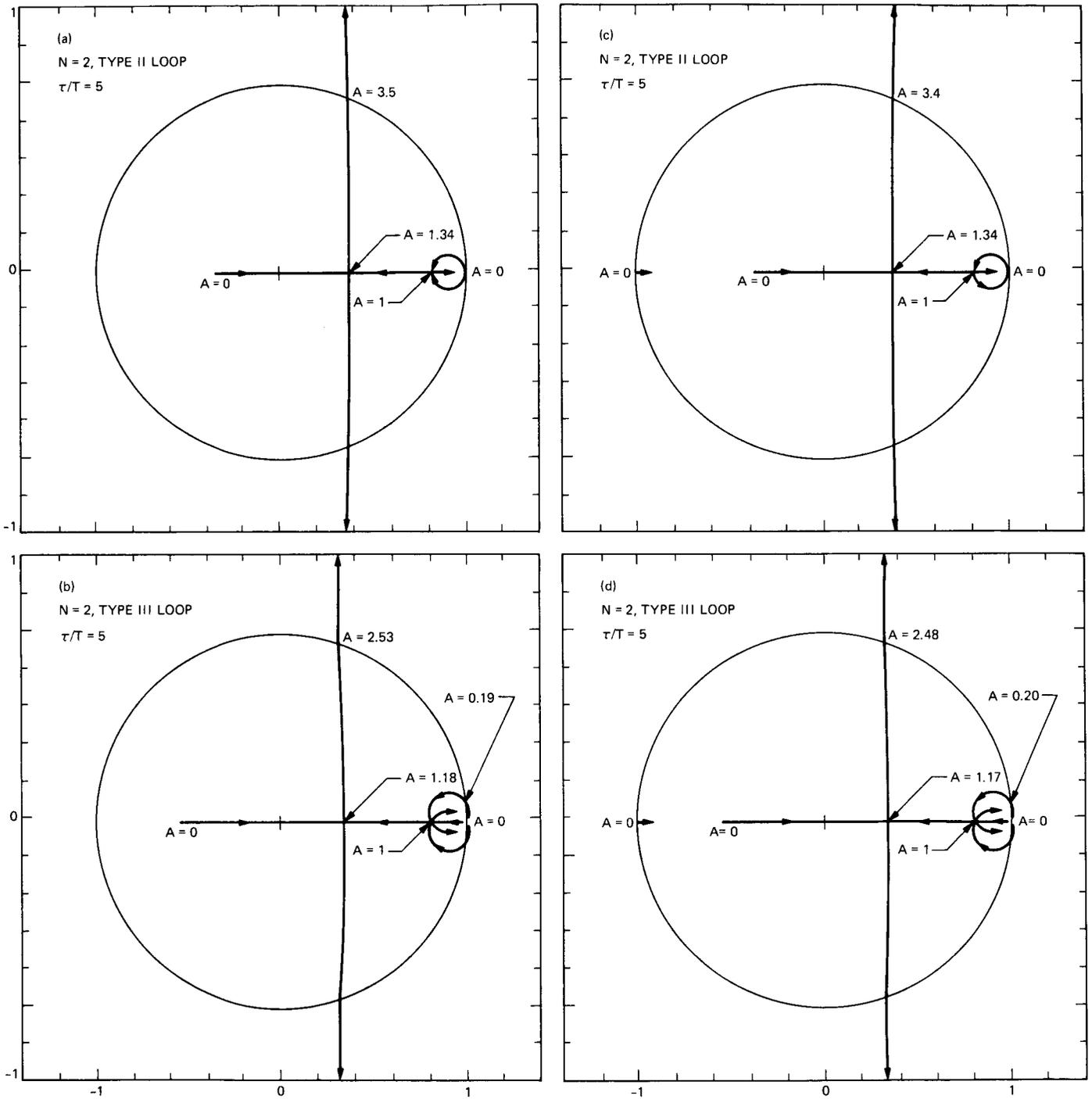
Fig. 7. Examples of root loci: (a) type II loop; (b) type III loop; (c) type II loop with pole at $z = -0.98$**; (d) type III loop with pole at** $z = -0.98$
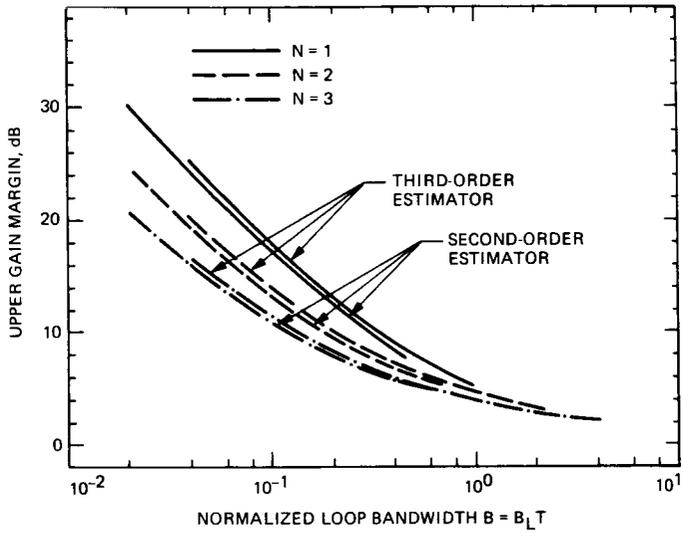
Fig. 8. Upper gain margin versus normalized loop bandwidth

Fig. 10. Phase error coefficients versus normalized loop bandwidth

Fig. 9. Lower gain margin versus normalized loop bandwidth

Choosing Channel Quantization Levels and Viterbi Decoding for Space Diversity Reception Over the Additive White Gaussian Noise Channel

S. Kalson
Communications Systems Research Section

In this article we review previous work in the area of choosing channel quantization levels for a additive white gaussian noise channel composed of one receiver-demodulator, and show how this applies to the Deep Space Network composed of several receiver-demodulators (space diversity reception). Viterbi decoding for the resulting quantized channel is discussed.

I. Introduction

The Deep Space Network receives interplanetary signals at several locations on the Earth. Also, at each complex there are several receiver-demodulators. Thus, we have available a number of receiver-demodulators, each producing a "stream" of sufficient statistics (symbols). For biphase modulation over the unquantized additive white gaussian noise channel (AWGN), it has been shown that maximum likelihood decoding can be realized by first forming a weighted sum of these streams to form a new single stream of statistics, followed by maximum likelihood decoding of this single stream (Ref. 1). This is commonly referred to as symbol stream combining. Each stream of statistics is weighted in proportion to the square root of its received signal energy divided by the noise power spectral density associated with its corresponding receiver-demodulator.

In this article, we are concerned with the quantized AWGN channel, and so the above does not apply. How are the quantization levels for each receiver-demodulator to be chosen and what is the structure of the Viterbi decoder? Choosing quantization levels for the AWGN channel composed of one receiver-demodulator is not new (Ref. 2). This will be reviewed and extended for space diversity reception, resulting in an optimum method for the Deep Space Network. Furthermore, we discuss the Viterbi decoder for space diversity reception.

II. Review of Previous Work

It has been proposed (Refs. 2 and 3) that modulation and demodulation design can be based upon the "R_o criterion," where R_o is the cut-off rate of the channel. That is, modulation and demodulation should be such that the cut-off rate of the resulting channel created by the modulation and demodulation scheme is maximum. The rationale for this is the fact that the probability of bit error P_b for the best convolutional codes is upper bounded by

$$P_b \leqslant \frac{2^{-KR_o/R}}{[1 - 2^{-\epsilon R_o/R}]^2} \tag{1}$$

where $\epsilon > 0$ and $0 \leqslant R \leqslant R_o(1 - \epsilon)$, where R is the code rate in bits/channel-symbols and K is the constraint length (Ref. 4).

In our case, we are interested in choosing the quantization levels such that R_o is maximum.

The cut-off rate in bits/channel-symbols is

$$R_o = -\log_2 \left\{ \min_{q(x)} \sum_y \left[\sum_x \sqrt{p(y|x)}\, q(x) \right]^2 \right\} \quad (2)$$

where x is the transmitted symbol, $q(x)$ is a probability distribution, and $p(y|x)$ is the conditional probability distribution of the received output symbol y given that the symbol x is transmitted. We are interested in biphase modulation, where x is one of two symbols, say $x = \pm\sqrt{E_s}$ (E_s is the received symbol energy). Then for biphase modulation, we have by symmetry that $q(x = +\sqrt{E_s}) = q(x = -\sqrt{E_s}) = 1/2$ is the distribution needed to minimize the bracketed expression in the definition of R_o, Eq. (2). Thus, we have

$$R_o = 1 - \log_2 \left[1 + \sum_y \sqrt{p(y|x = +\sqrt{E_s})\, p(y|x = -\sqrt{E_s})} \right]$$

$$(3)$$

Let us consider the symmetric quantization scheme of Fig. 1, where Q, the size of the output alphabet, is assumed to be even. In this article, we shall only consider the case of Q being even. The case of Q being odd can be handled in essentially the same way. Thus, for symmetric channels, we have $T_i = -T_{-i}$, $i = 1, 2, \ldots, Q/2$. By convention, we shall define $T_o = 0$ and $T_{Q/2} = -T_{-Q/2} = \infty$.

In Fig. 1, we have denoted the output alphabet by $\{a_i, i = -Q/2, \ldots, -1, 1, \ldots, Q/2\}$. Thus, the quantization scheme of Fig. 1 produces the discrete input-binary, output-symmetric channel of Fig. 2 with transition probability distributions $p(a_i|x)$ given by

$$p(a_i|x = \pm\sqrt{E_s})$$

$$= \begin{cases} \dfrac{1}{\sqrt{\pi N_o}} \displaystyle\int_{T_{i-1}}^{T_i} e^{-(z\pm\sqrt{E_s})^2/N_o} dz, \\ \qquad\qquad\qquad i = 1, 2, \ldots, Q/2 \\[2em] \dfrac{1}{\sqrt{\pi N_o}} \displaystyle\int_{-T_{-i}}^{-T_{-i-1}} e^{-(z\pm\sqrt{E_s})^2/N_o} dz, \\ \qquad\qquad\qquad i = -1, -2, \ldots, -Q/2 \end{cases}$$

$$(4a)$$

where N_o is the one sided noise spectral density. In terms of the normalized quantization $\hat{T}_i \equiv T_i/\sqrt{E_s}$ and signal-to-noise ratio $\rho \equiv E_s/N_o$, the above transition probability can be expressed as

$$p(a_i|x = \pm\sqrt{E_s})$$

$$= \begin{cases} \sqrt{\rho/\pi} \displaystyle\int_{\hat{T}_{i-1}}^{\hat{T}_i} e^{-(z\pm 1)^2 \rho} dz \qquad i = 1, 2, \ldots, Q/2 \\[2em] \sqrt{\rho/\pi} \displaystyle\int_{-\hat{T}_{-i}}^{-\hat{T}_{-i-1}} e^{-(z\pm 1)^2 \rho} dz \qquad i = -1, -2, \ldots, -Q/2 \end{cases}$$

$$(4b)$$

Using Eq. (4a) with Eq. (3), and setting $\partial R_o/\partial T_i = 0$, $i = 1, 2, \ldots, Q/2$, we have after some algebra the set of equations:

$$\cosh[2\,\hat{T}_i\rho - 1/2 \ln \Lambda(a_i)] = \cosh[2\,\hat{T}_i\rho - 1/2 \ln \Lambda(a_{i+1})]$$

$$i = 0, 1, \ldots, Q/2 - 1 \quad (5)$$

where $\Lambda(a)$ is the likelihood ratio

$$\Lambda(a) \equiv \frac{p(a|x = \sqrt{E_s})}{p(a|x = -\sqrt{E_s})} \quad (6)$$

Notice that the equation $\cosh(x) = \cosh(y)$ can be satisfied by one of the two conditions: (1) $x = y$, or (2) $x = -y$. Numerical results show that the former condition applied to Eq. (6) cannot be satisfied for any T_i. Thus, we must use the second condition to obtain

$$8\,\hat{T}_i\rho = \ln[\Lambda(a_i)\,\Lambda(a_{i+1})] \qquad i = 0, 1, \ldots, Q/2-1 \quad (7)$$

It can easily be shown that Eq. (7) is equivalent to Eq. (16) of Ref. 2.

Massey (Ref. 2) suggested the following algorithm to solve Eq. (7): Pick some T_1 and use Eq. (7) with $i = 1$ to solve for T_2 (remember that $T_o = 0$). Likewise, use Eq. (7) with $i = 2$ to solve for T_3. Continuing in this way, we finally arrive at a tentative value for $T_{Q/2-1}$. Then, see if this value for $T_{Q/2-1}$ satisfies Eq. (7) for $i = Q/2 - 1$, remembering that $T_{Q/2} = \infty$. If not, then pick a smaller T_1 and try again. If this method still fails to converge, then the tentative guesses for T_1 should be increased instead of decreased.

The fact that the above algorithm may be numerically intensive is not important, since the normalized quantization levels can be determined off-line. Thus, a table look-up can be stored in ROM, giving the normalized quantization levels as a function of signal-to-noise ratio. Notice that the quantization levels are a function of both E_s and N_o. The maximum likelihood estimation of these parameters has been considered in Ref. 5.

III. Space Diversity Reception

For space diversity reception, we have multiple receiver-demodulators, say N of them. If $\mathcal{H}^{(i)}$ is the output space of the ith receiver-demodulator, then the output space $\mathcal{H}$ of the discrete channel composed of N receiver-demodulators is $\mathcal{H}^{(1)} \times \mathcal{H}^{(2)} \times \cdots \times \mathcal{H}^{(N)}$. Let $\mathcal{H}^{(i)}$ be the alphabet $\{a_1^{(i)}, a_2^{(i)}, \ldots, a_Q^{(i)}\}$ (without loss of generality, we assume the number of quantization levels for each receiver-demodulator to be the same). Assuming that the noise of the N receiver-demodulators are uncorrelated, we have that the joint transition probabilities of this discrete channel are

$$p_{\mathcal{H}}(\mathbf{a}|x) = \prod_{i=1}^{N} p_{\mathcal{H}(i)}(a^{(i)}|x) \tag{8}$$

where $p_{\mathcal{H}(i)}(a^{(i)}|x)$ are the marginal transition probabilities of the quantized channel comprised of just the ith receiver-demodulator, and

$$\mathbf{a} \in \mathcal{H}, \quad \mathbf{a} \equiv (a^{(1)}, a^{(2)}, \ldots, a^{(N)}), \quad a^{(i)} \in \mathcal{H}^{(i)} \tag{9}$$

The cutoff rate of the discrete channel composed of N receiver-demodulators is just

$$R_o = 1 - \log_2 \left[1 + \sum_{\mathbf{a} \in \mathcal{H}} \sqrt{p_{\mathcal{H}}(\mathbf{a}|x = \text{``0''}) \, p_{\mathcal{H}}(\mathbf{a}|x = \text{``1''})} \right]$$

$$= 1 - \log_2 \left[1 + \prod_{i=1}^{N} \sum_{a^{(i)} \in \mathcal{H}^{(i)}} \sqrt{p_{\mathcal{H}(i)}(a^{(i)}|x = \sqrt{E_s^{(i)}}) \, p_{\mathcal{H}(i)}(a^{(i)}|x = -\sqrt{E_s^{(i)}})} \right] \tag{10}$$

where $E_s^{(i)}$ is the received symbol energy at the ith receiver-demodulator.

From Eq. (10), we see that choosing the quantization levels to maximize R_o for the channel with N receiver-demodulators is equivalent to choosing the quantization levels of each receiver-demodulator to maximize its corresponding cut-off rate.

IV. Viterbi Decoding

For a branch in the trellis corresponding to the transmitted symbol sequence $(x_1, x_2, \ldots, x_n)$, where we have a rate $1/n$ code, we must calculate the metric

$$\sum_{j=1}^{n} \ln p_{\mathcal{H}}(\mathbf{a}_j|x_j) \tag{11}$$

where $\mathbf{a}_j, j = 1, \ldots, n$ is the received symbol sequence composed of symbols from $\mathcal{H}$. From Eq. (8), we see that the metric of Eq. (11) is just

$$\sum_{j=1}^{n} \sum_{i=1}^{N} \ln p_{\mathcal{H}(i)}(a_j^{(i)}|x_j) = \sum_{i=1}^{N} \sum_{j=1}^{n} \ln p_{\mathcal{H}(i)}(a_j^{(i)}|x_j)$$

$$(12)$$

Thus, we see from Eq. (12) that Viterbi decoding is accomplished by calculating the metrics of each branch in the trellis associated with each receiver-demodulator, and summing these metrics to give the total metric associated with that branch.

V. Remarks

This article reviews how quantization levels should be chosen to maximize the cut-off rate of the discrete channel created by a quantization scheme. For most situations, eight properly chosen quantization levels will give nearly optimum results, (Refs. 2 and 4). However, since a coding gain of only one tenth of a decibel in received bit energy-to-noise ratio is important to the Deep Space Network, more quantization levels may be required.

We have also discussed the Viterbi decoder for space diversity reception over the quantized channel. In a sense, the "combining" of the symbol streams is performed "within" the Viterbi decoder, where the branch metrics are given by the sum of metrics associated with each receiver-demodulator, Eq. (12). It is therefore not optimum to combine the symbols before Viterbi decoding. An example of a suboptimum method is to treat the symbols as real numbers (perhaps the midpoint of the symbol's corresponding quantization zone), and combine them as explained in Ref. 1.

It should be noted that symbol stream combining has been carried out experimentally in Ref. 6, and practical aspects of choosing quantization levels have been discussed in Ref. 7. However, the "R_o criterion" was not considered in these references, nor were the symbols "combined" as discussed here. Does choosing the quantization levels to maximize R_o and performing "true" maximum likelihood (Viterbi) decoding (Eq. 12) lead to any significant coding gain? This question should be investigated further to make sure that the Deep Space Network is making the best use of the channel.

Acknowledgment

References

1. McEliece, R. J., Pollara, F., and Swanson, L., "Symbol Stream Combining in a Convolutionally Coded System," *TDA Progress Report 42-81*, pp. 47–50, Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1985.

2. Massey, J., "Coding and Modulation in Digital Communications," *1974 Int. Zurich Seminar on Digital Communications*, pp. E2 (1), (2), (3), (4), Zurich, Switzerland, 1974.

3. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*. New York: John Wiley & Sons, 1965.

4. Viterbi, A. J., and Omura, J. K., *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.

5. Kalson, S., and Dolinar, S. J., "Maximum Likelihood Estimation of Signal-to-Noise Ratio and Combiner Weight," *TDA Progress Report 42-85*, pp. 32–41, Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1986.

6. Hurd, W. J., et al., "Intercontinental Antenna Arraying by Symbol Stream Combining at ICE Giacobini-Zinner Encounter," *TDA Progress Report 42-84*, pp. 220–228, Jet Propulsion Laboratory, Pasadena, Calif., February 15, 1986.

7. Pollara, F., and Swanson, L., "Effects of Quantization on Symbol Stream Combining in a Convolutionally Coded System," *TDA Progress Report 42-84*, pp. 82–87, Jet Propulsion Laboratory, Pasadena, Calif., February 15, 1986.
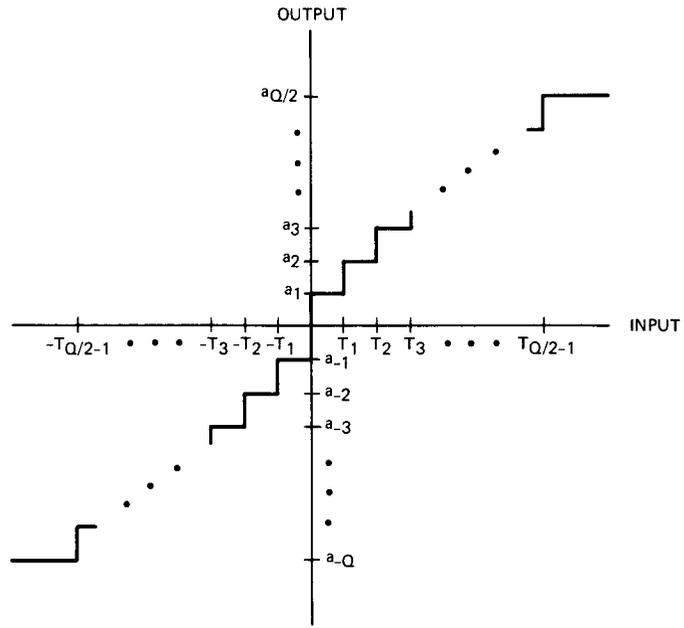
Fig. 1. Quantization scheme
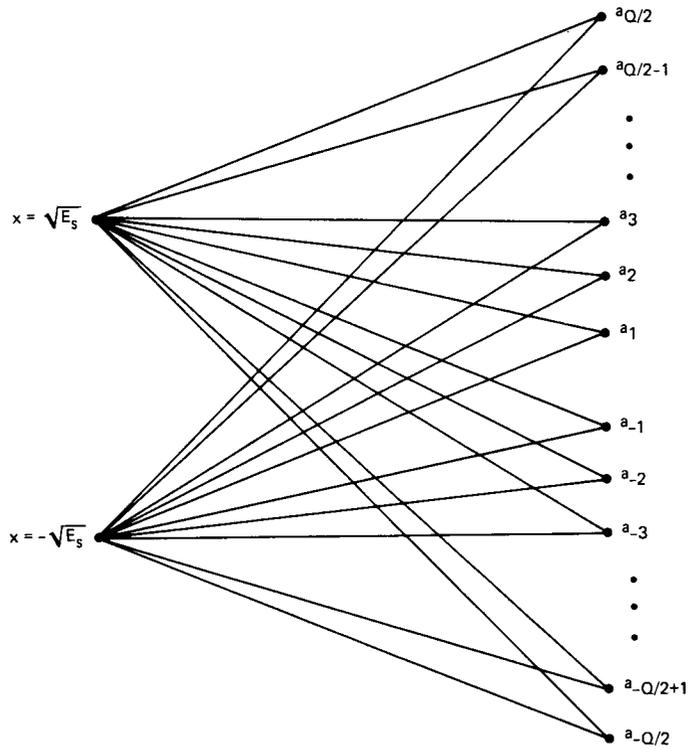
Fig. 2. Discrete channel

Phase Lock Acquisition for Sampled Data PLLs Using the Sweep Technique

S. Aguirre, D. H. Brown, and W. J. Hurd

Communications Systems Research Section

Simulation results of the swept-acquisition performance of residual carrier phase-locked loops (PLLs) are reported. The loops investigated are sampled data counterparts of the continuous time type II and III loops currently in use in Deep Space Network receivers. It was found that sweep rates of $0.2\,B_L^2$ to $0.4\,B_L^2$ Hz/s can be used, depending on the loop parameters and loop signal-to-noise ratio (SNR), where B_L is the one-sided loop noise bandwidth. Type III loops are shown to be not as reliable as type II loops for acquisition using this technique, especially at low SNRs.

I. Introduction

It is well known that phase-locked loops (PLLs) are excellent tracking devices, but they can take an excessive amount of time to acquire when the loops are narrow-band compared to the input frequency offsets.

A popular method of improving the acquisition time is to sweep the center frequency of the local oscillator over the frequency range where the input signal is expected to be (Refs. 1–4). The sweep rate must be held within reasonable bounds; otherwise the PLL may not lock at all. Frazier and Page (Ref. 1) derived an empirical formula based on simulation for the maximum permissible sweep rate for a 0.90 probability of acquisition for noisy signals. Later Gardner (Ref. 4) suggested a more conservative limit for the allowable sweep rates based on practical experience.

These results are valid only for second-order, continuous time PLLs. To our knowledge similar documented information is lacking for third-order continuous time PLLs, or for sampled data versions of these two types of loops.

An Advanced Receiver is now being developed for the Deep Space Network. This receiver uses type II and type III sampled data loops (Ref. 5). The type III is required to achieve low steady state phase error with narrow bandwidths under conditions of high doppler rate, such as at Voyager Neptune encounter. Since the sweep technique is equivalent (as far as the PLL is concerned) to acquiring when frequency ramps are present in the input signal, it is interesting to determine if the type III is reliable for acquisition under these dynamic conditions. This would avoid the need to acquire first with a type II loop and subsequently switch to a type III loop.

We present computer simulations of the swept-acquisition technique. The PLLs involved are those pertinent to the Advanced Receiver, which have been previously discussed in Ref. 5. We present plots of probability of acquisition versus

sweep rate with loop signal-to-noise ratio (SNR) as a parameter for type II and III loops.

We also examine in detail the impact of sampling rate on acquisition and pull-in behavior in general.

II. Description of Simulation

A detailed description of the baseband simulation model was reported in Ref. 6. The existing model only requires the addition of a frequency ramp term (opposite in sign to the frequency offset) to the input phase process. For the simulation runs, the local oscillator was swept 100 times, and the number of times that the loops acquired represented an estimate of the percentage of probability of acquisition. This procedure was repeated for different frequency rates and several combinations of loop parameters, SNR, and loop type.

Phase lock is declared once the phase error drops below 90 deg and stays there for at least ten times the reciprocal of the loop bandwidth. This definition is somewhat arbitrary, but we found it convenient to treat low SNR situations.

The initial frequency offset for every run was arbitrarily chosen as ten times the loop bandwidth, and the initial phase offset was uniformly distributed between $(-\pi, \pi)$. The mean and standard deviation of the acquisition times were recorded.

As in Ref. 6, the simulations were run for $B_L T = 0.02$, where B_L is the loop noise bandwidth, and T is the loop filter update time. This value is typical of the present breadboard implementation.

Other values of $B_L T$ were tested to determine their influence on pull-in behavior with, and without, frequency ramps present.

III. Simulation Results

Figures 1–3 contain the simulation results of probability of acquisition versus normalized sweep rate (where the normalized sweep rate is defined as the sweep rate in Hz/s, divided by B_L^2). In Tables 1 and 2 we include numerical values for the maximum permissible sweep rates for a 0.90 probability of acquisition. The mean times to acquire normalized by the bandwidth of the loop are shown, as well as the corresponding normalized standard deviations.

To compare the various loops, consider a loop SNR of 13 dB. For a 0.90 probability of acquisition Table 1 indicates that a type II loop with $r = 2$ as given in Ref. 5 can be swept at a normalized rate of 0.30. Table 2 indicates that a type II loop

with $r = 4$ can be swept only at 0.25, and Fig. 3 indicates that a type III loop with $r = 4$ and $k = 0.25$ (Ref. 5) can be swept only at 0.17. For type II similar results hold for SNRs down to 7 dB, but type III loops with $r = 4$, $k = 0.25$ do not acquire reliably for loop SNRs below 13 dB. Thus, acquisition time is better for type II loops with $r = 2$ than for the other loops considered.

A type III loop with $r = 2$ was found to be unreliable for acquisition, even for small sweep rates, for all signal to noise ratios considered.

The uncertainty in the plotted points in Figs. 1–3 can be quantified in the following manner. Every simulation run represents a Bernoulli trial, since there are two possible outcomes. One is called a success (if the loop acquires) and the other one a failure (if the loop does not lock).

For a given sweep rate, denote the true probability of acquisition by p, and the probability of failure by q, where $p + q = 1$. The probability of r successes in any order, out of n simulation runs, follows a binomial distribution with mean np and variance npq.

Since n is large (100 in our case), the binomial distribution can be approximated by a Gaussian distribution. With this in mind, we obtain the 95% confidence intervals (following any standard text in statistics)

$$\bar{x} - \frac{1.96\sqrt{\bar{x}(1-\bar{x})}}{10} < p < \bar{x} + \frac{1.96\sqrt{\bar{x}(1-\bar{x})}}{10} \quad (1)$$

where p is the true probability of acquisition for a given sweep rate, and $\bar{x}$ is the estimate of p obtained in the simulations.

These confidence intervals are indicated by bars in Figs. 1–3. In these figures, for the sake of clarity, bars were drawn only for one case of loop SNR. But, of course, similar bars can be drawn for other loop SNRs.

IV. Effects of Sampling Rate

Digital phase-locked loops possess certain distinctive features not encountered in analog phase-locked loops. For example, in theory, an analog PLL with a type II loop has an infinite pull-in range. This is not, in general, the case for digital PLLs. Early evidence supporting this statement may be found in Refs. 8–11. For this reason it is important to examine more carefully the effects introduced by possibly inadequate sampling rates (sampling rate is defined here as the reciprocal of the loop filter update time T).

In Fig. 4 we examine the sensitivity of the sweep technique with respect to $B_L T$. We plot the maximum normalized sweep rates that ensure a 0.90 probability of acquisition as a function of normalized bandwidth. This is done for a type II loop with $r = 2$ and without noise. A maximum normalized bandwidth of 0.08 is tolerable if the sweep technique is employed. Larger values are not allowed because stability problems arise (Ref. 5). (Stable loops can be designed for larger $B_L T$, but these loops are not considered here. See Ref. 7 for example.)

As stated before, Figs. 1–3 were obtained using $B_L T = 0.02$. For this value type II loops require a minimum normalized sweep rate of about 0.05 and type III a minimum of 0.16, otherwise the loops may not lock at all. Smaller values of $B_L T$ require smaller minimum sweep rates (approaching zero for a $B_L T = 0.01$, $r = 4$ type II when noise is not present), which agrees with intuition, since a sampled data loop with small $B_L T$ (<0.01) resembles a continuous time loop.

Computer simulations show that for a given bandwidth and sampling rate, these sampled data loops tolerate only a finite frequency offset. If this limit is exceeded, the loops become unstable and phase lock is never reached. In Fig. 5 we plot this maximum tolerable frequency offset as a function of bandwidth and sampling rate for a type II loop with $r = 4$ and no noise. The most surprising result is that the maximum allowed frequency offset varies almost linearly with the sampling rate.

More detailed analysis is required in the area of pull-in characteristics as a function of bandwidth, sampling rate, and initial phase conditions. This requires somewhat elaborate mathematical tools and graphical aids such as phase planes to study the convergence properties of the nonlinear difference equations describing the digital PLL operation. But this is considered to be outside the scope of our present discussion.

It is interesting to compare our results with empirical formulas reported in Refs. 1, 2, and 4. For a type II continuous time loop with $r = 2$ in a noiseless environment, Frazier (Ref. 1) and Lindsey (Ref. 2) suggest a frequency rate $R < 0.56\ B_L^2$, while Gardner (Ref. 4) suggests a conservative $R < 0.28\ B_L^2$. Our result, $R < 0.38\ B_L^2$, falls between these two values.

V. Effects of Initial Frequency Offsets

In the following discussion we will assume that the sampling rate is adequate such that frequency offsets much larger than $10\ B_L$ can be tolerated, as shown in Fig. 5.

The results presented are for initial frequency offsets of $10\ B_L$. The acquisition times are dominated by the sweep rate, and hence by the maximum sweep rate allowed for a given set of parameters. For type II loops the results can be extended to wider initial frequency offsets without modifying the maximum sweep rate. For type III loops, however, this is not valid for large initial offsets. Heuristically, this is explained as follows. When the initial offset is very large, the phase detector output is almost sinusoidal with nearly zero mean; therefore, the output of the loop filter (which is proportional to the frequency estimate) is approximately described by the integral of a Wiener process, whose variance grows as t^3, where t is time relative to the start of acquisition. For large initial offsets, t can be large and the rate of change of the loop filter output due to accumulated noise can be greater than the sweep rate. Lock may never occur either because the effective sweep rate is too high, or because the noise keeps the loop frequency from ever sweeping through the frequency of the incoming signal.

In conclusion, acquisition is more reliably achieved with type II loops.

Fast Fourier Transform (FFT) techniques can also be used in the acquisition mode to resolve the frequency uncertainty to a pre-determined range, typically a fraction of the loop bandwidth.

A fast acquisition scheme using FFT techniques has been added to the Advanced Receiver breadboard. Initial testing reveals that carrier phase lock can be achieved in 2 s with an initial frequency uncertainty of 350 Hz and a loop bandwidth of 4 Hz.

VI. Laboratory Results

Using a breadboard of the Advanced Receiver, laboratory measurements were taken to verify the results obtained by computer simulations. One hundred experiments were performed for three different cases, each of them with an initial frequency offset equal ten times the loop bandwidth. The bandwidth was set to 10 Hz and the sampling rate to 500 Hz.

Case 1 involved a type II loop with $r = 2$ and an estimated loop SNR of 13.3 dB. The loop was swept at a rate of 30.4 Hz/s, or, equivalently, a normalized rate of 0.304. It was observed that out of 100 trials, the loop attained phase lock 95 times. Extrapolation of the simulation results yields for the same SNR a probability of acquisition close to 0.92.

In case 2, a type II loop with $r = 4$ was employed. This time the estimated SNR was 12.8 dB and the normalized

sweep rate 0.248. The loop acquired with probability 0.88. Extrapolation of the simulation results predict 0.87.

The last case consisted of a type III loop with $r = 4, k = 0.25$ and an estimated loop SNR of 13 dB. The normalized sweep rate was adjusted to 0.208. The measured probability of acquisition was only 0.42, in contrast with 0.82 predicted by the simulations.

It is believed that the discrepancy for type III loops can be attributed to small dc offsets in the output of the loop phase detector. In a type III loop, these dc offsets are accumulated in a double summer resulting in a rapid dc build-up, the net effect of which is to cause the filter output to run off in one direction. This has been observed in the breadboard loop by monitoring the NCO output. Thus, we conclude that in practical applications, undesired dc offsets may preclude the use of type III loops for acquisition.

VII. Empirical Formulas

Empirical formulas were obtained from the plots, relating maximum sweep rates, which gave a 0.90 probability of acquisition, as a function of loop SNR (ρ) and bandwidth. These are summarized as follows (assuming no dc offsets):

Type II, $r = 2$:

$$R < 0.38\, B_L^2 \left(1 - \frac{1}{\sqrt{\rho}}\right) \text{Hz/s}, \quad \rho > 7 \text{ dB} \qquad (2)$$

Type II, $r = 4$:

$$R < 0.32\, B_L^2 \left(1 - \frac{1}{\sqrt{\rho}}\right) \text{Hz/s}, \quad \rho > 7 \text{ dB} \qquad (3)$$

Type III, $r = 4, k = 0.25$:

$$R < 0.27\, B_L^2 \left(1 - \frac{1}{\sqrt{\rho}}\right) \text{Hz/s}, \quad \rho > 13 \text{ dB} \qquad (4)$$

VIII. Conclusions

The simulations show that type II loops are more reliable for acquisition than type III loops when using the swept-acquisition technique. Type II loops can tolerate higher sweep rates and, hence, minimize acquisition time. Type III loops are not recommended for acquisition if this technique is employed, particularly if dc offsets are present.

We also explored in some detail the influence of sampling rate on pull-in behavior, pointing out fundamental differences between analog and digital PLLs.

References

1. Frazier, J. P., and Page, J., "Phase-Lock Frequency Acquisition Study," *IRE Trans. Space Electron. Telemetry*, Vol. SET-8, pp. 210–227, Sept. 1962.

2. Lindsey, W. C., *Synchronization Systems in Communication and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1972.

3. Best, R. E., *Phase-Locked Loops*, McGraw-Hill, Inc., NY, 1984.

4. Gardner, F. M., *Phaselock Techniques*, 2d ed., John Wiley and Sons, NY, 1979.

5. Aguirre, S., and Hurd, W. J., "Design and Performance of Sampled Data Loops for Subcarrier and Carrier Tracking," *TDA Progress Report 42-79*, pp. 81–95, Jet Propulsion Laboratory, Pasadena, CA, Nov. 15, 1984.

6. Aguirre, S., "Acquisition Times of Carrier Tracking Sampled Data Loops," *TDA Progress Report 42-84*, pp. 88–93, Jet Propulsion Laboratory, Pasadena, CA, Feb. 15, 1986.

7. Kumar, R., and Hurd, W. J., "A Class of Optimum Digital Phase Locked Loops for the DSN Advanced Receiver," *TDA Progress Report 42-83*, pp. 63–80, Jet Propulsion Laboratory, Pasadena, CA, Nov. 15, 1985.

8. Weinberg, A., and Liu, B., "Discrete Time Analyses of Nonuniform Sampling First- and Second-Order Digital Phase Lock Loop," *IEEE Trans. Commun.*, Vol. COM-22, pp. 123–137, Feb. 1974.

9. Russo, F., "Graphical Analysis of a Digital Phase-Locked Loop," *IEEE Trans. Aerosp. Electron. Syst.*, Vol AES-15, pp. 88–94, Jan. 1979.

10. Osborne, H. C., "Stability Analysis of an Nth Power Digital Phase-Locked Loop — Part I: First-Order DPLL," *IEEE Trans. Commun.*, Vol. COM-28, pp. 1343–1354, Aug. 1980.

11. Osborne, H. C., "Stability Analysis of an Nth Power Digital Phase-Locked Loop — Part II: Second- and Third-Order DPLL's," *IEEE Trans. Commun.*, Vol. COM-28, pp. 1355–1364, Aug. 1980.

Table 1. Maximum permissible sweep rate for 0.90 probability of acquisition: corresponding normalized acquisition time and standard deviation for initial frequency offset of 10 B_L, type II, $r = 2$

SNR, dB	R_{90}/B_L^2	$B_L \bar{t}_{acq}$	$B_L \sigma_{acq}$
7	0.183	57.5	12.0
13	0.304	32.4	4.2
16	0.350	27.8	2.7
∞	0.380	24.4	0.9

Table 2. Maximum permissible sweep rate for 0.90 probability of acquisition: corresponding normalized acquisition time and standard deviation for initial frequency offset of 10 B_L, type II, $r = 4$

SNR, dB	R_{90}/B_L^2	$B_L \bar{t}_{acq}$	$B_L \sigma_{acq}$
7	0.145	67.6	14.7
13	0.248	41.0	6.4
16	0.280	36.3	5.9
∞	0.320	29.8	3.9
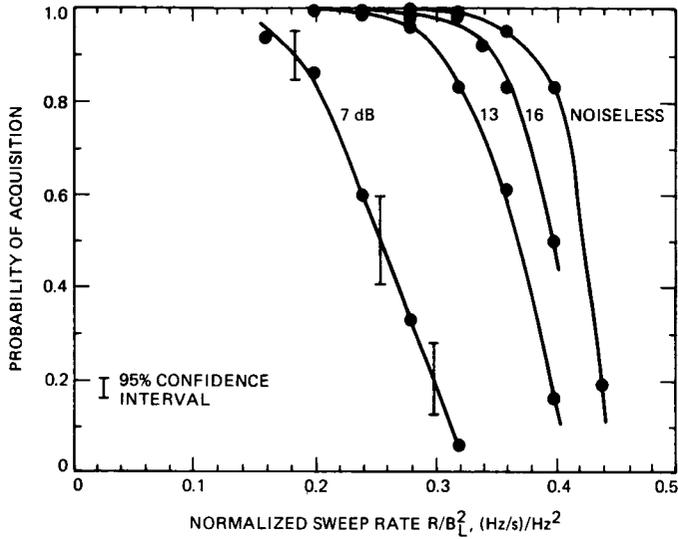
Fig. 1. Probability of acquisition vs normalized sweep rate with loop SNR as a parameter for a type II, $r = 2$ loop (initial frequency offset is ten times the loop bandwidth)
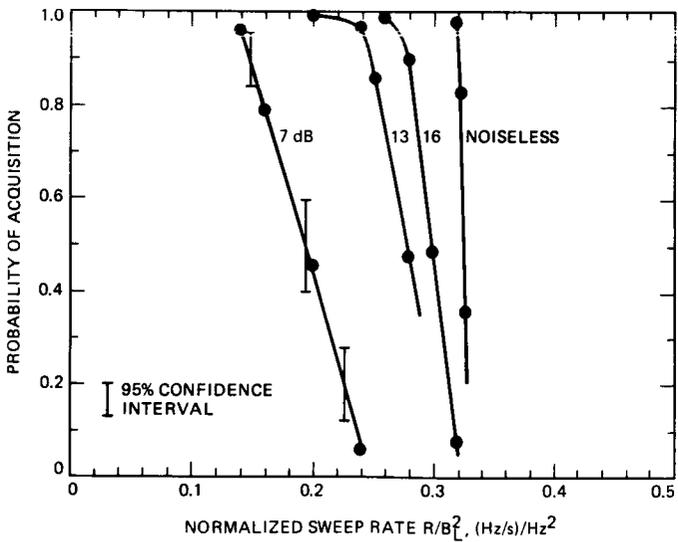
Fig. 2. Probability of acquisition vs normalized sweep rate with loop SNR as a parameter for a type II, $r = 4$ loop (initial frequency offset is ten times the loop bandwidth)
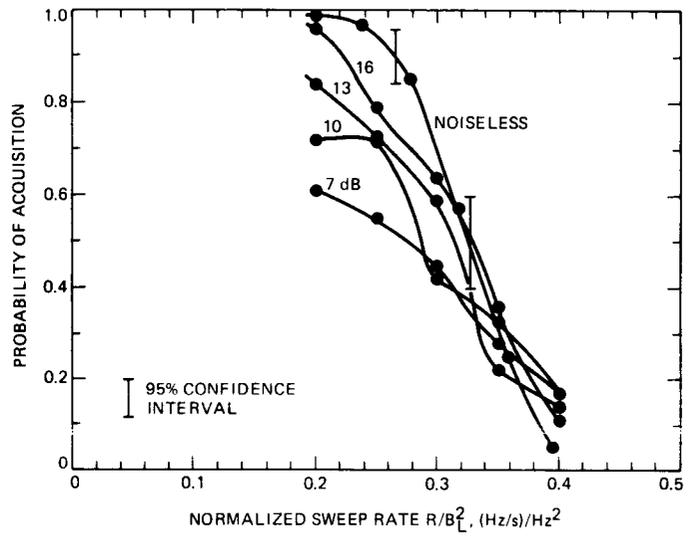
Fig. 3. Probability of acquisition vs normalized sweep rate with loop SNR as a parameter for a type III, $r = 4$, $k = 0.25$ loop (initial frequency offset is ten times the loop bandwidth)
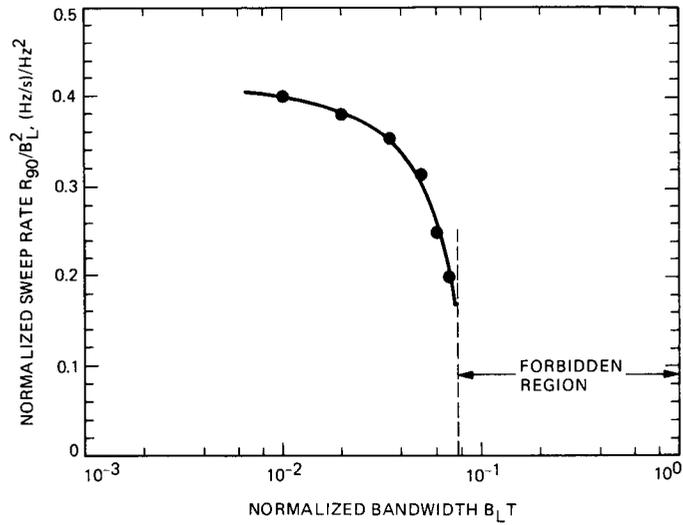
Fig. 4. Sensitivity of the sweep technique to $B_L T$ for sweep rates
that ensure 0.90 probability of acquisition: type II, $r = 2$
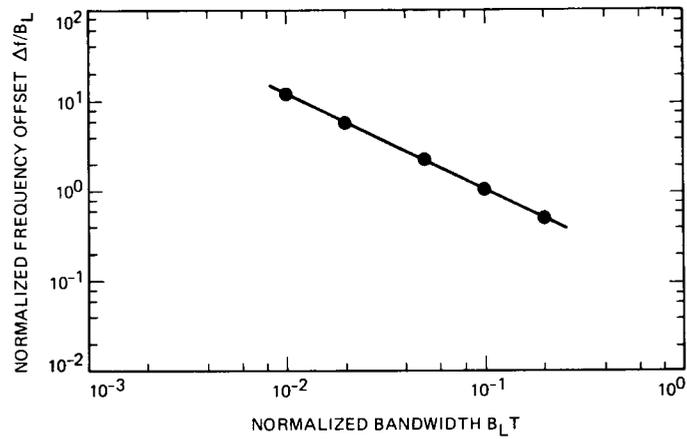
Fig. 5. Maximum initial frequency offset for which phase lock
is guaranteed: type II, $r = 4$, SNR = ∞

A Method to Dramatically Improve Subcarrier Tracking

W. J. Hurd and S. Aguirre

Communications Systems Research Section

A method is presented for achieving a dramatic improvement in phase tracking of square wave subcarriers or other square waves. The method is to set the amplitude of the phase quadrature reference signal to zero except near the zero crossings of the input signal. Without changing the loop bandwidth, the variance of the phase error can be reduced to approximately $W\sigma_0^2$, where σ_0^2 is the phase error variance without windowing, and W is the fraction of cycle in which the reference signal has a nonzero value. Simulation results confirm the analysis and establish minimum W versus SNR. Typically, the window can be made so narrow as to achieve a phase error variance of $1.5\sigma_0^4$.

I. Introduction

In deep space communications, the loss in data signal-to-noise ratio due to phase tracking error is often more severe for subcarrier tracking than for carrier tracking. This is because the subcarriers are often square waves, and the carriers are sinusoidal. The SNR loss varies approximately as the mean square phase error for sinusoids, but only as the rms phase error for square waves.

Subcarrier tracking loss is most significant in low-rate telemetry systems where the subcarrier loop bandwidth cannot be made narrow enough to reduce the rms phase error to a small enough value. For example, the loss in average symbol SNR for the Pioneer 10 spacecraft at a symbol SNR of 0 dB with the narrowest bandwidth Block III or Block IV subcarrier demodulator assembly is 0.4 dB at 16 bps and 0.6 dB at 8 bps. The actual loss in decoder threshold is even greater, just as radio loss is greater for coded than for uncoded systems.

These losses motivated the analysis and simulation of the improved subcarrier tracking method presented here. The method is capable of reducing the loss in the average symbol SNR (SSNR) to under 0.1 dB for the Pioneer example, without reducing the loop bandwidth.

II. Method and Performance

The improvement in subcarrier tracking is achieved by windowing one of the subcarrier channel reference signals as done in a Digital data Transition Tracking Loop (DTTL) bit synchronizer (Ref. 1). A theoretical basis for this method was presented by Layland (Ref. 2), who concluded that, for a first order phase-locked loop and high loop SNR the optimum reference signals needed to track square waves resemble alternating trains of narrow pulses.

Figure 1 shows the windowed quadrature phase referencing waveform and its relationship to the subcarrier and to the standard reference waveform. Let W be the fraction of each cycle of the reference signal which has nonzero value. The reference signal looks like a square wave, multiplied by zero except for the regions within plus or minus $W/4$ of the zero

crossings as illustrated in Fig. 1. The theoretical improvement in loop SNR is approximately a factor of $1/W$, provided that the phase error is small enough that the loop is in the linear region. Based on simulation results, values of W from $1/16$ to $1/64$ appear practical in cases for which the loop SNR would otherwise be low enough to cause significant symbol SNR loss. This means that the loss can be reduced by a factor of 4 to 8. A 0.4 dB loss can be reduced to 0.05 dB, and a 0.6 dB loss can be reduced to under 0.1 dB. These examples are typical of Pioneer 10 at 0 dB SSNR and data rates of 16 bps and 9 bps, respectively.

The implementation of the windowed reference signal is remarkably simple. The waveform can be generated in a read only memory (ROM) whose input address is the phase. The only change from the full square wave case is to zero the reference signal ROM for the appropriate regions of phase. There is no change in the gain of the subcarrier phase detector due to the windowing.

III. Analysis

In this section we develop the equations that describe the operation of the subcarrier loop, shown in Fig. 2. Using linear analysis we compute analytically the variance of the phase tracking error and validate our results with computer simulations. Throughout this article we neglect quantization errors, nonzero data rise time, filtering distortions, etc. We also assume that perfect symbol synchronization is available and has already been established.

A. Phase Detector Model

The ith sample of the digitized subcarrier signal for the DSN Advanced Receiver is assumed to be of the form (Ref. 3)

$$r_i = \sqrt{P_D} \, d_k \, \mathsf{Sin}(\theta_i) \cos(\phi_{ci}) + n_i \qquad (1)$$

where

P_D = average data power (V^2)

d_k = data value of kth binary symbol (± 1 equally probable)

$\mathsf{Sin}(x)$ = sgn $(\sin(x))$

ϕ_{ci} = instantaneous phase carrier estimation error (rad) assumed zero for the rest of the analysis.

n_i = zero mean white Gaussian noise sample with variance σ_n^2

θ_i = instantaneous subcarrier angle (rad)

At this point, it is convenient to introduce the variables

$\widehat{\theta}_i$ = closed loop estimate of θ_i (rad)

$\phi_i = \theta_i - \widehat{\theta}_i$ = instantaneous subcarrier phase estimation error of the ith sample (rad)

To facilitate the analysis, we assume that the subcarrier period T_{sc} is related to the symbol duration T_{sym} by

$$T_{sym} = nT_{sc} \text{ for some integer } n \qquad (2)$$

This circumvents modeling problems associated with "end" effects, which greatly complicate the analysis and lie outside the intended scope of the present discussion. We also postulate that a large number of samples per symbol time are available.

The loop operates as follows (see Fig. 2): The digitized incoming signal r_i is mixed with the reference signals (with no loss of generality we set the multiplier gains equal to one)

$$R_I = \mathsf{Sin}(\widehat{\theta}_i) \qquad (3)$$

$$R_Q = \widetilde{\mathsf{Cos}}(\widehat{\theta}_i) \qquad (4)$$

to produce the signals x_i and y_i respectively. These signals are accumulated over the L samples during a symbol interval. Assuming that the instantaneous phase errors of the samples averaged over any particular symbol interval are equal, then the accumulators have responses

$$X_k = d_k L \sqrt{P_D} \, (1 - |u_k|) + n_{xk}, \qquad k = 0, \ldots, M \quad (5)$$

$$Y_k = d_k L \sqrt{P_D} \, v_k + n_{yk}, \qquad k = 0, \ldots, M \qquad (6)$$

where

$$u_k = \frac{2}{\pi} |\phi_k|, \qquad |\phi_k| \leqslant \pi \qquad (7)$$

and

$$v_k = \begin{cases} u_k, & |\phi_k| \leqslant \pi W/2 \\ \operatorname{sgn}(\phi_k)W, & \pi W/2 \leqslant |\phi_k| \leqslant \pi(1 - W/2) \\ 2\operatorname{sgn}(\phi_k) - u_k, & \pi(1 - W/2) \leqslant |\phi_k| \leqslant \pi \end{cases}$$

$$(8)$$

These last two equations describe the in-phase and quadrature arms of the phase detector respectively. We show them graphically in Fig. 3.

B. Phase Detector S Curve

The S curve is defined as the mean value of the error control signal conditioned on the phase error. The gain slope at the origin of the S curve and the variance of the error control signal are useful in evaluating the closed loop tracking performance.

In the following section we show that the S curve is given by

$$S(\phi_n) = \frac{\pi}{2} (1 - | U_n |) V_n \qquad (9)$$

which for convenience, we have normalized to have unity slope at the origin. This is shown in Fig. 4 for different windows. Notice that for small phase errors, the gain (slope) of the S curve does not change due to the windowing. This has the notable advantage from an implementation point of view of maintaining constant phase detector gain as W is changed. This simplifies implementation in which a wide W is used for acquisition and a narrower W for tracking.

C. Phase Detector Variance

In order to assess the closed loop tracking performance, the variance of the error control signal is needed. Due to the orthogonal nature of the reference signals R_I and R_Q, the noise processes $\{n_{xk}\}$ and $\{n_{yk}\}$ are independent. Strictly speaking, these are cyclostationary processes, but we approximate them by stationary processes. In other words, their first and second order statistics are obtained by time averaging (over just one symbol interval for this case) their ensemble averages. With this in mind, samples of these noise processes have zero mean and variances

$$\text{var } (n_{xk}) = L \sigma_n^2 \qquad (10)$$

$$\text{var } (n_{yk}) = WL \sigma_n^2 \qquad (11)$$

There is also self noise, which is neglected. This self noise is the difference between the actual signal summed over the actual samples, and the mean value which we have used.

The outputs of the in-phase and quadrature arm summers are multiplied together and accumulated subsequently over M symbols to produce the error control voltage that drives the Costas loop. This signal is

$$\epsilon_n = ML^2 P_D (1 - | U_n |) V_n + N_n \qquad (12)$$

with U_n, V_n being the time averages of u_k and v_k over M samples respectively. It can be argued via the central limit theorem that the noise samples N_n are approximately Guassian with zero mean and variance

$$\sigma_N^2 = \sigma_n^2 P_D L^3 M V_n^2$$

$$+ W \sigma_n^2 P_D L^3 M (1 - | U_n |)^2$$

$$+ W \sigma_n^4 L^2 M \qquad (13)$$

D. Linear Tracking in the Presence of Noise

When the loop is in the linear region, U_n and V_n in Eqs. (12) and (13) are close to zero and

$$\sigma_N^2 \approx W \sigma_n^2 P_D L^3 M$$

$$+ W \sigma_n^4 L^2 M \qquad (14)$$

If we assume that the noise samples N_n are stationary, the steady state variance of $\{\phi_n\}$ is given by (Refs. 4 and 5)

$$\sigma_\phi^2 = \frac{1}{2\pi j} \oint_{|z|=1} H(z) H(z^{-1}) z^{-1} \frac{R_N(z)}{A^2} \, dz \qquad (15)$$

where

$$R_N(z) = Z \{E(N_j N_{j+n})\}$$

is the Z transform of the autocorrelation function of the noise process at the input of the loop filter. The function $H(z)$ denotes the closed loop transfer function and A is the gain of the control voltage signal (without noise) evaluated at zero phase error, which is

$$A = \frac{2}{\pi} L^2 M P_D \qquad (17)$$

Since the samples $\{N_n\}$ are uncorrelated, zero mean with variance σ_N^2, then

$$R_N(z) = \sigma_N^2 \qquad (18)$$

so

$$\sigma_\phi^2 = \frac{2B_L T \sigma_N^2}{A^2} \qquad (19)$$

where

$$B_L = \frac{1}{2T} \frac{1}{H^2(1)} \frac{1}{2\pi j} \oint_{|z|=1} H(z)H(z^{-1}) \frac{dz}{z} \qquad (20)$$

is the one-sided noise bandwidth of the loop, and T is the filter update time related to the symbol duration by

$$T = M T_{sym} \qquad (21)$$

Substitution of Eqs. (14) and (17) into Eq. (19) results in

$$\sigma_\phi^2 = \frac{2\left(\frac{\pi}{2}\right)^2 W\sigma_n^4 \left(1 + \frac{P_D L}{\sigma_n^2}\right) B_L T_{sym}}{L^2 P_D^2} \qquad (22)$$

If we assume that the received noise samples n_i are obtained by sampling white noise of one-sided spectral density N_o at a rate $1/T_s$, then

$$\sigma_n^2 = \frac{N_o}{2T_s} \qquad (23)$$

Using this result, and the fact that

$$P_D = \frac{E_s}{T_{sym}} \qquad (24)$$

$$T_{sym} = L T_s \qquad (25)$$

E_s being the symbol energy, then the variance of the tracking error can be put in the form

$$\sigma_\phi^2 = W\left(\frac{\pi}{2}\right)^2 \frac{B_L T_{sym}}{E_s/N_o} \left(1 + \frac{1}{2E_s/N_o}\right) \qquad (26)$$

Notice that windowing improves the loop SNR by a factor of $1/W$. At first glance, it might appear (erroneously though) that arbitrarily small windows can be selected to obtain any desired performance. This is not so, since for very small windows, linear theory is not valid, and the actual tracking variance is much larger than that predicted by Eq. (26). This will be quantified more precisely in the following section.

IV. Simulation Results

Computer simulations were performed to validate the analysis and determine the range of usable values for the windows. To perform the simulations, an equivalent PLL type model is found first for the Costas loop. This resembles a standard PLL, except that the sinusoidal nonlinearity is replaced by the normalized phase detector characteristic given by Eq. (9).

In Fig. 4 we summarize the simulation results and also include results dictated by linear analysis for several window sizes. We do this by computing the variance of the phase error using as parameters loop SNR's when $W = 1$. Thus, when $W = 1$ and loop SNR = 14 dB, the phase error variance is approximately 0.04 rad^2. By just narrowing the window, we can lower the variance by a factor of roughly 16.

The most striking piece of information contained in Fig. 4 is that, given an initial tracking variance, say σ_0^2, this variance can be reduced conservatively to

$$(\sigma_1^2)_{min} = 1.5\, \sigma_0^4 \qquad (27)$$

by selecting the optimum window size whose value is approximately

$$W = 0.5 \text{ to } 1 \times \sigma_0^2 \qquad (28)$$

V. SNR Loss Due to Phase Error and Design Example

First we determine the average loss in symbol SNR (SSNR) due to subcarrier phase error. Equation (5) represents the output of the decision arm in the Costas loop. It is observed that a subcarrier phase estimation error causes the signal voltage term to be degraded by

$$D = 1 - \frac{2}{\pi} |\phi| \qquad (29)$$

The symbol signal to noise ratio (SSNR) is then degraded on the average by the statistical expectation of the square of the above term. If the loop SNR is high, then it is reasonable to assume a Gaussian density function for the phase error. Carrying out the details of the expectation leads to

$$E\{D^2\} = 1 - 2\left(\frac{2}{\pi}\right)^{3/2} \sigma_\phi \qquad (30)$$

where only first order terms were retained.

A design example is considered next: The average loss in SSNR for the Pioneer 10 spacecraft at a coded symbol SNR of 0 dB with a two-sided design point bandwidth of 0.03 Hz and a symbol rate of 33-1/3 sps is about 0.45 dB (see footnote 1). If a narrower window is employed, this degradation can be significantly reduced.

For the parameters previously mentioned, the initial (W = 1) loop SNR is 27.8 dB. From the simulation results,

it appears that a window of size W = 1/256 is feasible. If this smaller time window is used, with the other parameters kept constant, the average loss in SSNR can be reduced to 0.01 dB.

VI. Conclusions

A subcarrier Costas loop capable of tracking square waves with less phase error has been described. By setting the quadrature reference signal to zero at the appropriate phases, the variance of the tracking phase error is reduced by a factor of $1/W$ under linearized conditions. Computer simulations validate the previous statement, and give practical values for W when the loop is not adequately described by linear theory.

[1] "Deep Space Network/Flight Project Interface Design Handbook," JPL internal document 810-5, Rev. D, Vol. I, Module TLM-10, p. 67, Jet Propulsion Laboratory, Pasadena, Calif.

References

1. Hurd, W. J., and Anderson, T. O., "Digital Transition Tracking Symbol Synchronizer for Low SNR Coded Systems," *IEEE Transactions on Communication Technology COM-18*, No. 2, pp. 141–147, April 1970.

2. Layland, J. W., "On Optimal Signals for Phase-Locked Loops," *IEEE Transactions on Communication Technology COM-17*, No. 5, pp. 526–531, October 1969.

3. Sfeir, R., Aguirre, S. and Hurd, W. J., "Coherent Digital Demodulation of a Residual Carrier Signal Using IF Sampling," *TDA Progress Report 42-78*, pp. 135–142, Jet Propulsion Laboratory, Pasadena, Calif., August 15, 1984.

4. Oppenheim, A. V., and Schafer, R. W., *Digital Signal Processing*, Englewood Cliffs, N. J.: Prentice-Hall, 1975.

5. Lindsey, W. C., and Chie, C. M., "A Survey of Digital Phase-Locked Loops," *Proceedings of the IEEE 69*, No. 4, pp. 410–431, April 1981.
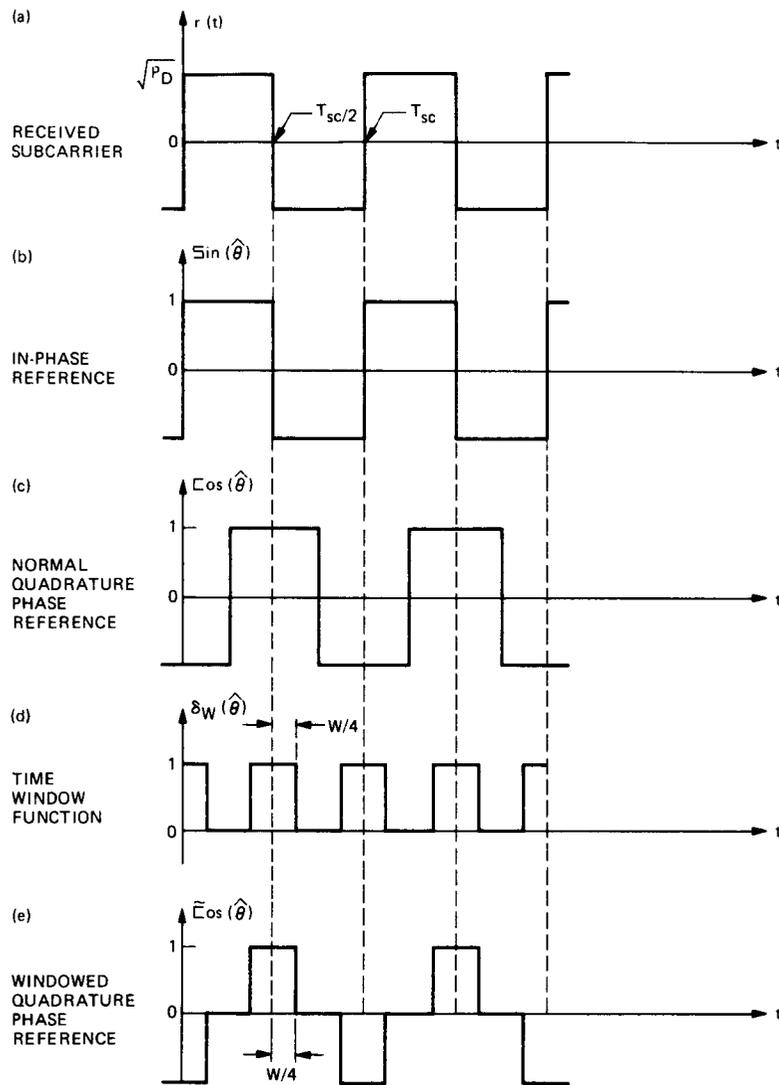
Fig. 1. Loop waveforms in the absence of noise. Continuous and perfect subcarrier synchronization is assumed for pictorial representation.
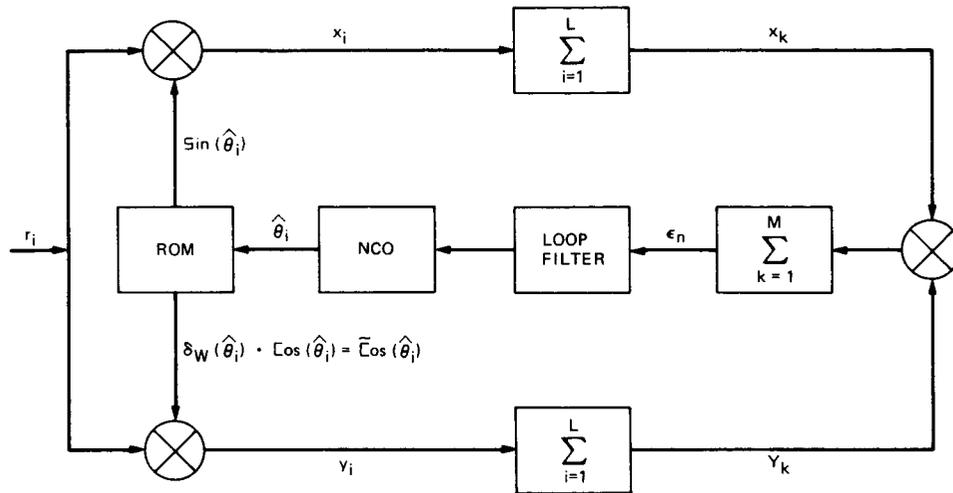
Fig. 2. Block diagram of the all-digital subcarrier Costas loop incorporating a window function in the quadrature reference signal
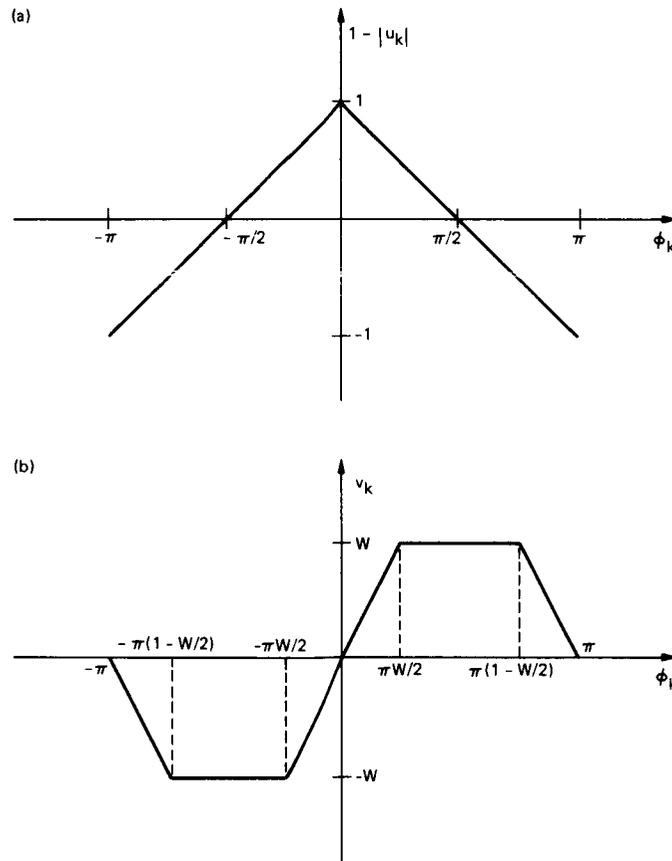
Fig. 3. Phase detector characteristics of in-phase and quadrature arms: (a) in-phase and (b) in-quadrature
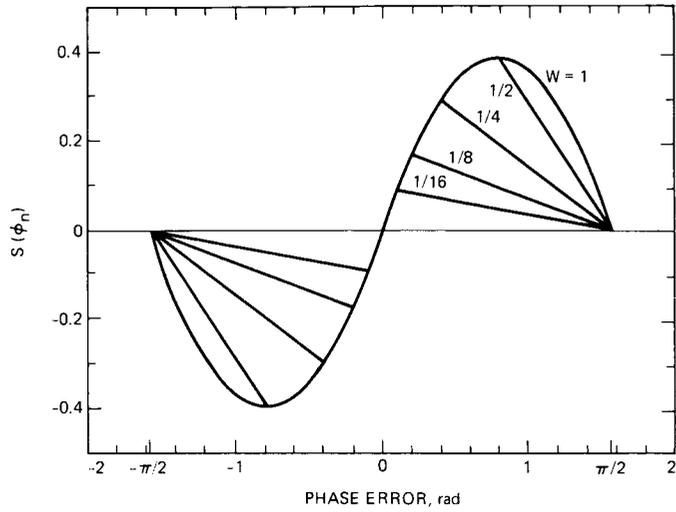
**Fig. 4. Normalized loop phase detector characteristics
for different window sizes**
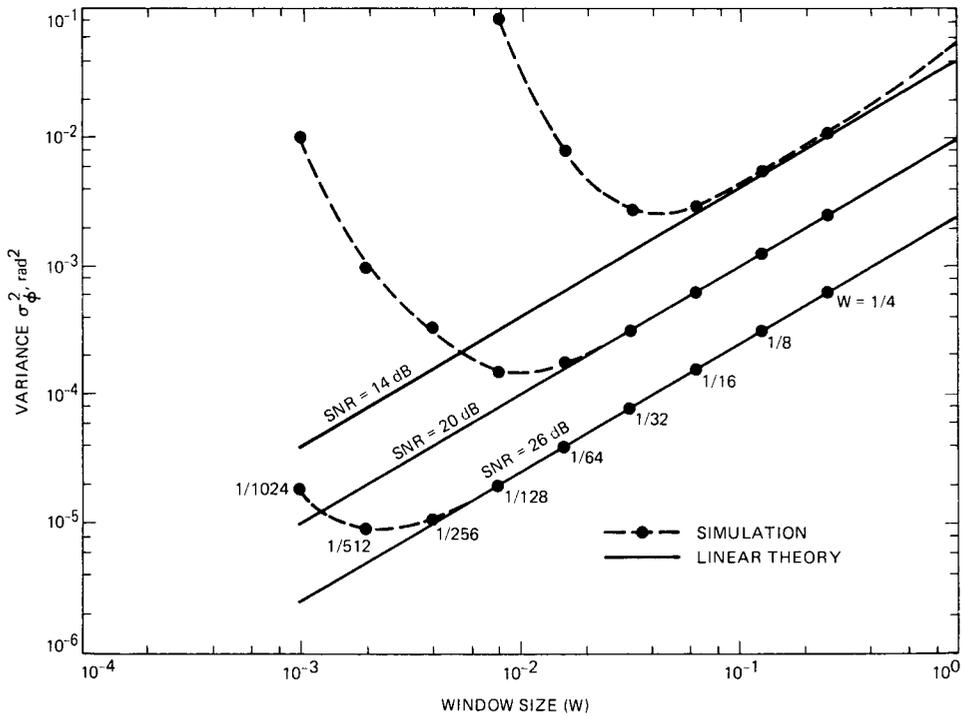
**Fig. 5. Variance of the subcarrier phase estimation error as a function of window size
with loop SNR as a parameter**

A New Method for Frame Synchronization

M. Shahshahani and L. Swanson
Communications Systems Research Section

A new frame synchronization method based on an examination of the shortest bit length containing all error bursts is introduced. It is shown that the new method is more reliable and efficient than the commonly used one based on counting the number of errors for the convolutionally coded channel.

I. Introduction

Data from spacecraft are transmitted to stations on earth in encoded form. The transmission channels are generally noisy, and the encoding enables one to correct some of the errors. This is especially true of convolutional codes because of their high error correcting capability. The encoding/decoding procedure is briefly described in Fig. 1.

The encoded data appear in frames of fixed length and it is essential to identify the beginning of each frame. This is accomplished by the insertion of a marker at the beginning of each one. Correctly identifying the marker is the problem of frame synchronization. This is usually done after the received signal has passed through the Viterbi decoder. Several methods have been proposed for this purpose (see Ref. 1). Here we consider yet another method and make a comparative study of the new frame synchronization technique and the commonly used one based on counting the number of disagreements.

II. The Method

Data are transmitted in frames of N ($\cong 10080$) bits, each of which begins with a marker of $k(= 32)$ bits. Currently, frame synchronization involves counting the number of disagreements with the marker. Since Viterbi decoded data contains errors in bursts, we propose instead to examine the shortest length containing all the bursts. Specifically, choose a positive integer T (called threshold) and examine the k consecutive bits

of data starting at a random point α. If the distance δ between the first and last disagreements with the marker in the sequence $\alpha, \alpha + 1, \ldots, \alpha + k - 1$ is greater than T, then we reject α as the beginning of a marker; otherwise it is retained as a candidate for a marker. In the latter event, we examine the k bits starting at the point $N + \alpha$. If $\delta > T$ for this sequence, we reject the k-bits starting at α (or $N + \alpha$) as a marker and repeat the procedure starting at $\alpha + 1$. Otherwise, we accept the k-bits as the first bit of a marker. In the latter case we continue to test the k-bits starting at $N + 2\alpha, N + 3\alpha, \ldots$ for the marker in the course of decoding data. If for three consecutive trials δ exceeds T then we reject α as the beginning of a marker, and repeat the procedure starting at $\alpha + 1$.

To analyze the performance of this technique we make the customary assumption that the $N - k$ bits of data in each frame is a random sequence of 0's and 1's, so that all sequences are equally likely to occur. Without loss of generality we may assume the marker consists of a sequence of k zeroes which due to the noise in the channel is possibly received erroneously. It has been observed that the errors in the Viterbi decoded data sequence occur in bursts and the lengths of these bursts follow the geometric distribution with parameter p. We recall that a bit sequence is a burst if (a) its first and last bits are incorrect, (b) it does not contain K (constraint length) consecutive correct bits, and (c) it is not contained in any other sequence with properties (a) and (b). Furthermore, the waiting time W between bursts has (shifted) geometric distribution with parameter q, i.e.,

$$P(W = n) = q(1 - q)^{n-K+1} \qquad (1)$$

Numerical values for p and q are given in Ref. 2.

Let M be the event that a marker actually starts at the randomly chosen point α, and O denote the event that α is identified as the beginning of a marker. In the next section we obtain estimates for $P(O|M)$, $P(M|O)$ and $P(O|M')$ where M' is the event complementary to M. These numbers provide a measure of the performance of the method. Note that $P(O'|M)$ and $P(O|M')$ are commonly called probability of miss and probability of false alarm respectively.

III. Probabilistic Estimates

To obtain estimates for the performance of the frame synchronization technique, it is convenient to assume that the probability of occurrence of more than one burst in a k-bit range is negligibly small. Such an assumption is reasonable for high SNR or short marker length. We first quantify this assertion.

Consider the sequence $1, 2, \ldots, k$, and let (a) $I = 1$, (b) $I = j$, $2 \leqslant j \leqslant k$ and (c) $I = \infty$ denote the events (a) 1 is contained in a burst, (b) first burst begins at bit j, and (c) there is no error in the sequence, respectively. Let L denote the length of the first burst where in case $I = 1$, L is measured from bit 1. Notice that if $I = j$ and $L = r$ then the bits $r + j$, $r + j + 1, \ldots, r + j + K - 1$ are correct. Denote by Y the starting point of the second burst. We want to calculate $P(Y \leqslant k)$. We have

$$Y = I + L + W$$

where W is the waiting time between bursts. Therefore

$$P(Y \leqslant k) = \sum_{\substack{j+r \leqslant k-K \\ j \geqslant 1, r \geqslant 1}} P(Y \leqslant k \mid I = j, L = r) \, P(I = j, L = r)$$

$$= \sum_{\substack{j+r \leqslant k-K \\ j \geqslant 1, r \geqslant 1}} P(W \leqslant k - j - r) \, P(I = j) \, P(L = r)$$

$$= \sum_{\substack{j+r \leqslant k-K \\ j \geqslant 1, r \geqslant 1}} \left(\sum_{v=K-1}^{k-j-r} P(W = v) \right) P(I = j) \, P(L = r) \qquad (2)$$

From Eq. (1),

$$\sum_{v=K-1}^{k-j-r} P(W = v) = 1 - (1 - q)^{k-j-r-K+2}$$

Also note

$$
\left.
\begin{aligned}
P(I = 1) &= \frac{B}{B + W} \\[2mm]
P(I = j) &= q(1 - q)^{j-1} \, \theta \\[2mm]
P(L = r) &= p(1 - p)^{r-1}
\end{aligned}
\right\} \qquad (3)
$$

where

$$\theta = \frac{W}{B + W} \quad \text{for} \quad 2 \leqslant j \leqslant k$$

Therefore

$$P(Y \leqslant k) = \sum_{\substack{r \geqslant 1 \\ r \leqslant k-K-1}} \frac{B}{B + W} \, p(1 - p)^{r-1}$$

$$- \sum_{\substack{r \geqslant 1 \\ r \leqslant k-K-1}} \frac{B}{B + W} \, p(1 - p)^{r-1} (1 - q)^{k-r+K+1}$$

$$+ \sum_{\substack{j \geqslant 2, r \geqslant 1 \\ j+r \leqslant k-K}} \theta \, pq \, (1 - p)^{r-1} (1 - q)^{j-1}$$

$$- \sum_{\substack{j \geqslant 2, r \geqslant 1 \\ j+r \; k-K}} \theta \, pq \, (1 - p)^{r-1} (1 - q)^{k-r+K+1}$$

i.e.,

$$P(Y \leqslant k) = \frac{B}{B+W}\left(1 - (1-p)^{k-K-1}\right)$$

$$- \frac{B}{B+W}\frac{p}{p-q}(1-q)^{2+2K}$$

$$\cdot\left((1-q)^{k-K-1} - (1-p)^{k-K-1}\right)$$

$$+ \sum_{j \geqslant 2, r \geqslant 1} \frac{W}{B+W} pq (1-p)^r (1-q)^{j-1}$$

$$- \frac{W}{B+W} pq (1-q)^{k+K} \sum_{r=1}^{k-K}\left(\frac{1-p}{1-q}\right)^{r-1}(k-K-r-1) \tag{4}$$

The numerical values of $P(Y \leqslant k)$ for different values of SNR are given in Table 1. The parameters $p = 1/B$ and $q = 1/(W-5)$ are taken from table C-1 in Ref. 2.

To estimate $P(M \mid O)$ we first note

$$P(M \mid O) = \frac{P(M, O)}{P(O)}$$

$$= \frac{P(O \mid M) P(M)}{P(O \mid M) P(M) + P(O \mid M') P(M')} \tag{5}$$

Let C be the event that the first burst point occurs no sooner than $K - T + 1$. Then it is trivial that

$$P(\widetilde{O} \mid M) \geqslant P(C) + P(1 \leqslant B \leqslant k - T) P(L \leqslant T)$$

where $\widetilde{O}$ denotes the event that $\delta \leqslant T$ for the sequence α, $\alpha + 1, \ldots, \alpha + k - 1$ where α is the randomly chosen starting bit. We assume, without loss of generality, that bit 1 is the beginning of a marker. Now

$$P(C) = \theta \sum_{j=k-T+1}^{\infty} q(1-q)^{j-1}$$

$$= \theta (1-q)^{k-T}$$

and

$$P(1 \leqslant B \leqslant k - T) = 1 - \theta (1-q)^{k-T}$$

(For a justification of using parameter q in evaluation of $P(C)$, see Ref. 3, pp. 12-13.) Hence

$$P(\widetilde{O} \mid M) \geqslant (1-q)^{k-T} + \left(1 - \theta (1-q)^{k-T}\right) \cdot \left(1 - (1-p)^T\right)$$

We denote r.h.s. of the above inequality by β. Since separation $N - k$ between markers is sufficiently large, the error bursts in the sequences beginning at α and $N + \alpha$ are essentially independent. Therefore

$$P(O \mid M) = P(\widetilde{O} \mid M)^2 \geqslant \beta^2 \tag{6}$$

Since for $0 < a < 1, c > 0$ the function

$$q(x) = ax/ax + (1-a)c$$

is increasing, we obtain from Eq. (5)

$$P(M \mid O) \geqslant \frac{\beta^2 P(M)}{\beta^2 P(M) + P(O \mid M') P(M')} \tag{7}$$

To calculate $P(O \mid M')$ we use the acceptability assumption on the marker. Thus if the randomly chosen point α is such that the sequence $\{\alpha, \alpha + 1, \ldots, \alpha + k - 1\}$ overlaps with but is not identical with the marker, then the probability of retaining α as the beginning of a marker is no greater than the case where the marker does not overlap. Now if $\{\alpha, \ldots, \alpha + k - 1\}$ does not overlap with the marker then this probability is bounded by

$$\gamma = \frac{\lambda(k, T)}{2^k}$$

where $\lambda(k, T)$ is one plus the number of binary sequences $a(i), a(i + 1), \ldots, a(i + T - 1)$ such that $a(i) = 1, 1 \leqslant i, i + T - 1 \leqslant k$. We have

$$\lambda(k, T) = 1 + k + (k - 1) + (k - 2) + \ldots + (k - T + 1) 2^{T-2}$$

$$= k + k (2^{T-1} - 1) - (T - 1) 2^T + T 2^{T-1}$$

Substituting

$$P(O \mid M') \leqslant \gamma^2 \tag{8}$$

in Eq. (7) we get

$$P(M \mid O) \geqslant \frac{\beta^2}{\beta^2 + (N - 1) \gamma^2} \tag{9}$$

The above inequality and Eqs. (6) and (8) are the required estimates.

IV. Numerical Calculations

In this section we make a numerical comparison of the quantities $P(M|O)$, $P(O|M')$ and $P(O|M)$ at various SNR values for the new method and one based on counting the number of disagreements. A few remarks are necessary regarding our calculations.

(1) To evaluate $P(O|M)$ and $P(O|M')$ for the old method we use Eq. (5), and calculate $P(O|M)$ and $P(O|M')$ by simulation.

(2) More precisely, we used a random number generator to generate bursts in 32,000 markers at various SNR values, and $P(O|M)$ was calculated accordingly.

(3) For calculation of $P(O|M')$ we assumed the marker is acceptable in the sense of Ref. 1. Thus if the randomly chosen point α is such that the sequence $\{\alpha, \alpha + 1, \ldots, \alpha + k - 1\}$ overlaps with but is not identical with the marker, we still can treat it as one in random data.

(4) Notice that our calculations sometimes give only upper or lower bounds.

(5) The quantity $P(O|M')$ is independent of SNR since we are assuming randomness of data.

(6) To make a meaningful comparison of the two methods we have graphically exhibited $P(O|M)$ and $P(M|O)$ for the same values of $P(O|M')$ (Figs. 2–8). For $P(M|O)$ we have only exhibited the curves for SNR = 1.6 since $P(M|O)$ is very stable relative to the variation of SNR.

(7) Detailed results of our calculations appear in Tables 1–14 and Figs. 2–8 below.

V. Conclusion

A comparison of performance statistics for fixed probability of false detection shows that the new method is significantly more reliable and efficient in detection of the marker. Mathematically, false detection, reliability and efficiency in detection of the marker are measured by $P(O|M')$, $P(M|O)$ and $P(O|M)$, respectively. The reliability of both methods is very stable relative to the variation of the signal to noise ratio.

References

1. Swanson, L., 1982, *A Comparison Frame Synchronization Methods*, JPL Publication 82-100, Jet Propulsion Laboratory, Pasadena, Calif.

2. Miller, R. L., L. J. Deutsch and S. A. Butman, 1981, *On the Error Statistics of Viterbi Decoding and the Performance of Concatenated Codes*, JPL Publication 81-9, Jet Propulsion Laboratory, Pasadena, Calif.

3. Feller, W., 1971. *An Introduction to Probability Theory and its Applications*, vol. 2. New York: John Wiley.

Table 1. Probability of more than one burst

SNR	$P(Y \leqslant 32)$
3	9×10^{-6}
2.5	8×10^{-5}
2.1	4×10^{-4}
1.9	10^{-3}
1.8	10^{-3}
1.7	2×10^{-3}
1.6	3×10^{-3}

Table 2. Comparison of reliability, SNR = 1.6

Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.999+	0.999+
2	↑	0.999
3		0.986
4		0.904
5		0.620
6		0.257
7		0.082
8		0.026
9		--
10		--
11		--
12		--
13		--
14		--
15		--
16	↓	--
17	0.999+	--
18	0.997	--
19	0.991	--
20	0.971	--
21	0.906	--

Table 3. Comparison of reliability, SNR = 1.7

Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.999+	0.999+
2	↑	0.999
3		0.986
4		0.906
5		0.623
6		0.261
7		0.083
8		0.027
9		--
10		--
11		--
12		--
13		--
14		--
15		--
16	↓	--
17	0.999+	--
18	0.998	--
19	0.991	--
20	0.971	--
21	0.906	--

	Table 4. Comparison of reliability, SNR = 1.8	
Threshold (*T*)	New Method $P(M \mid O) \geqslant$	Old Method $P(M \mid O)$
1	0.999+	0.999+
2		0.999
3		0.986
4		0.907
5		0.626
6		0.263
7		0.084
8		0.027
9		--
10		--
11		--
12		--
13		--
14		--
15		--
16		--
17	0.999+	--
18	0.998	--
19	0.991	--
20	0.971	--
21	0.907	--

	Table 5. Comparison of reliability, SNR = 1.9	
Threshold (*T*)	New Method $P(M \mid O) \geqslant$	Old Method $P(M \mid O)$
1	0.999+	0.999+
2		0.999
3		0.987
4		0.908
5		0.628
6		0.264
7		0.082
8		0.027
9		--
10		--
11		--
12		--
13		--
14		--
15		--
16		--
17	0.999+	--
18	0.998	--
19	0.992	--
20	0.971	--
21	0.906	--

Table 6. Comparison of efficiency, SNR = 2.1				Table 7. Comparison of reliability, SNR = 2.5		
Threshold (T)	New Method $P(M\|O) \geqslant$	Old Method $P(M\|O)$		Threshold (T)	New Method $P(M\|O) \geqslant$	Old Method $P(M\|O)$
1	0.999+	0.999+		1	0.999+	0.999+
2	↑	0.999		2	↑	0.999
3		0.987		3		0.987
4		0.909		4		0.910
5		0.632		5		0.635
6		0.267		6		0.269
7		0.085		7		0.086
8		0.027		8		0.027
9		--		9		--
10		--		10		--
11		--		11		--
12		--		12		--
13		--		13		--
14		--		14		--
15		--		15		--
16	↓	--		16	↓	--
17	0.999+	--		17	0.999+	--
18	0.998	--		18	0.998	--
19	0.992	--		19	0.992	--
20	0.971	--		20	0.971	--
21	0.907	--		21	0.908	--

Table 8. Comparison of efficiency, SNR = 1.6		
Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.864	0.878
2	0.875	0.894
3	0.885	0.908
4	0.895	0.920
5	0.903	0.930
6	0.911	0.941
7	0.919	0.949
8	0.926	0.957
9	0.932	--
10	0.938	--
11	0.943	--
12	0.948	--
13	0.953	--
14	0.957	--
15	0.961	--
16	0.964	--
17	0.968	--
18	0.971	--
19	0.973	--
20	0.976	--
21	0.978	--

Table 9. Comparison of efficiency, SNR = 1.7		
Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.884	0.897
2	0.894	0.911
3	0.903	0.922
4	0.911	0.933
5	0.919	0.942
6	0.926	0.950
7	0.932	0.958
8	0.938	0.965
9	0.943	--
10	0.948	--
11	0.953	--
12	0.957	--
13	0.961	--
14	0.965	--
15	0.968	--
16	0.971	--
17	0.974	--
18	0.976	--
19	0.979	--
20	0.981	--
21	0.983	--

Table 10. Comparison of efficiency, SNR = 1.8		
Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.905	0.915
2	0.914	0.927
3	0.921	0.937
4	0.928	0.946
5	0.935	0.955
6	0.941	0.962
7	0.946	0.968
8	0.951	0.974
9	0.956	--
10	0.960	--
11	0.963	--
12	0.967	--
13	0.970	--
14	0.973	--
15	0.976	--
16	0.978	--
17	0.980	--
18	0.982	--
19	0.984	--
20	0.986	--
21	0.987	--

Table 11. Comparison of efficiency, SNR = 1.9		
Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.915	0.929
2	0.923	0.939
3	0.930	0.948
4	0.937	0.956
5	0.943	0.962
6	0.948	0.968
7	0.953	0.974
8	0.958	0.978
9	0.962	--
10	0.966	--
11	0.969	--
12	0.972	--
13	0.975	--
14	0.977	--
15	0.980	--
16	0.982	--
17	0.984	--
18	0.986	--
19	0.987	--
20	0.988	--
21	0.990	--

Table 12. Comparison of efficiency, SNR = 2.1				Table 13. Comparison of efficiency, SNR = 2.5		
Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$		Threshold (T)	New Method $P(M\mid O) \geqslant$	Old Method $P(M\mid O)$
1	0.948	0.955		1	0.977	0.981
2	0.954	0.961		2	0.980	0.986
3	0.958	0.967		3	0.983	0.987
4	0.962	0.973		4	0.985	0.989
5	0.966	0.977		5	0.987	0.990
6	0.969	0.981		6	0.988	0.992
7	0.972	0.984		7	0.990	0.994
8	0.975	0.987		8	0.991	0.996
9	0.978	--		9	0.992	--
10	0.980	--		10	0.993	--
11	0.982	--		11	0.994	--
12	0.984	--		12	0.995	--
13	0.986	--		13	0.995	--
14	0.987	--		14	0.996	--
15	0.989	--		15	0.996	--
16	0.990	--		16	0.997	--
17	0.991	--		17	0.997	--
18	0.992	--		18	0.998	--
19	0.993	--		19	0.998	--
20	0.994	--		20	0.998	--
21	0.994	--		21	0.999	--

Table 14. Probability of false detection

Threshold (T)	New Method $P(O\|M') \leqslant$	Old Method $P(O\|M')$
1	10^{-16}	10^{-16}
2	10^{-16}	10^{-14}
3	10^{-15}	10^{-12}
4	10^{-15}	10^{-10}
5	10^{-14}	10^{-9}
6	10^{-14}	10^{-7}
7	10^{-13}	10^{-6}
8	10^{-12}	10^{-5}
9	10^{-12}	--
10	10^{-11}	--
11	10^{-11}	--
12	10^{-10}	--
13	10^{-10}	--
14	10^{-9}	--
15	10^{-8}	--
16	10^{-8}	--
17	10^{-7}	--
18	10^{-7}	--
19	10^{-6}	--
20	10^{-6}	--
21	10^{-5}	--
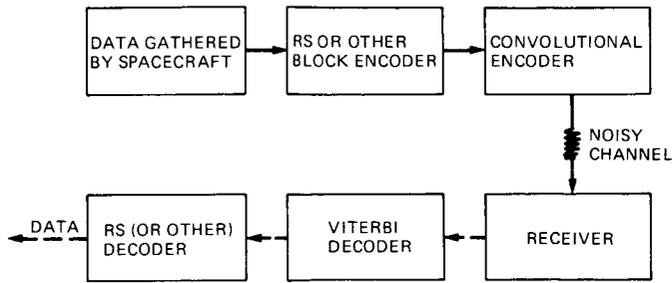
Fig. 1. The encoding-decoding procedure
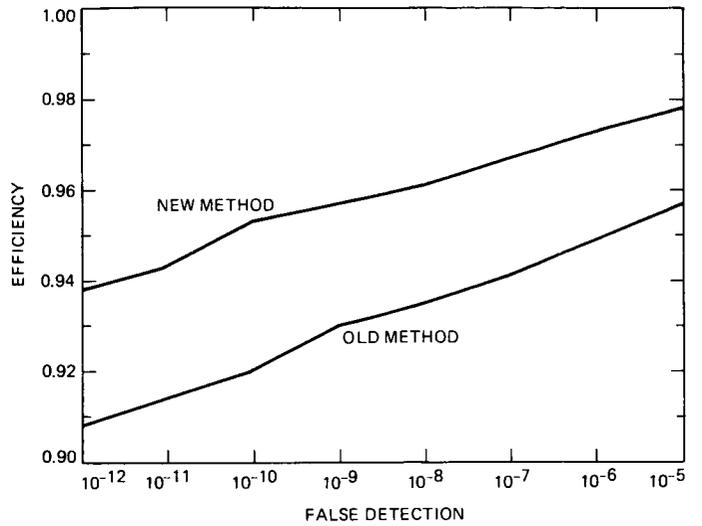
Fig. 3. Graph 2, SNR = 1.6
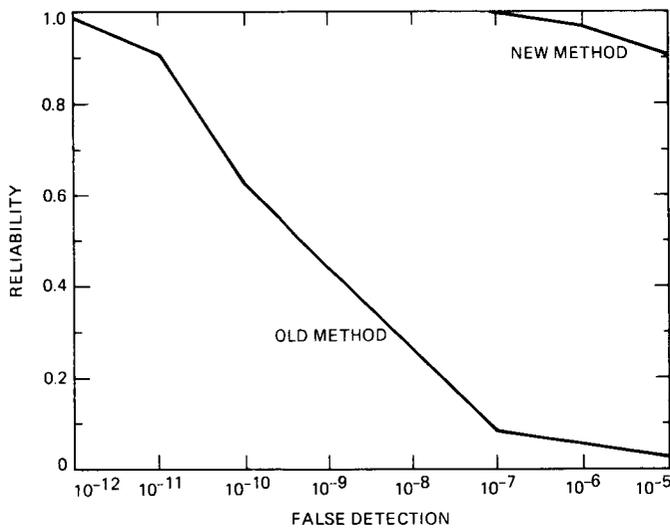
Fig. 2. Graph 1
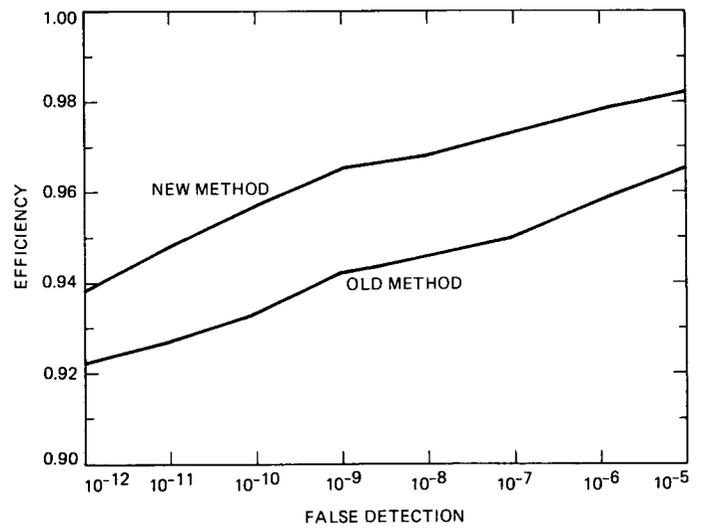
Fig. 4. Graph 3, SNR = 1.7
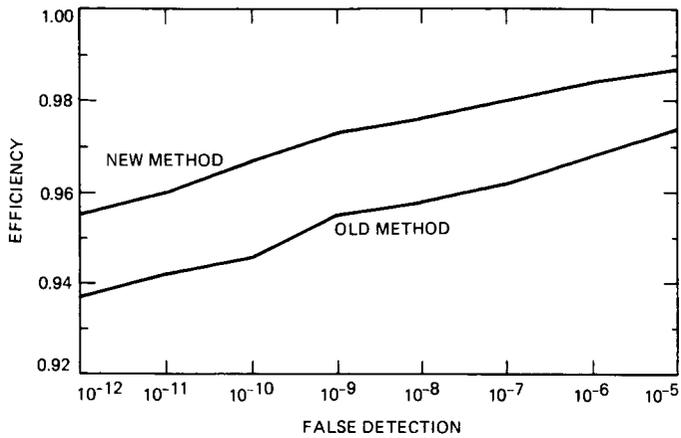
Fig. 5. Graph 4, SNR = 1.8
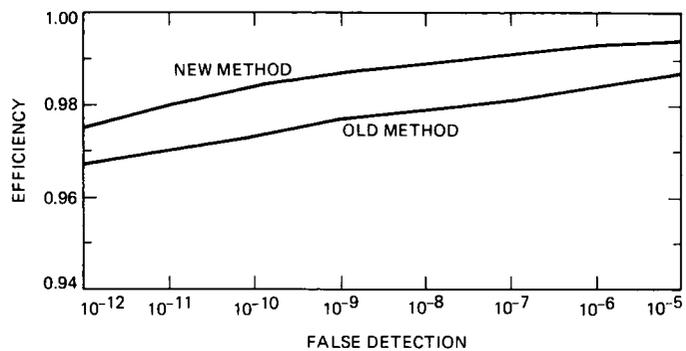
Fig. 7. Graph 6, SNR = 2.1
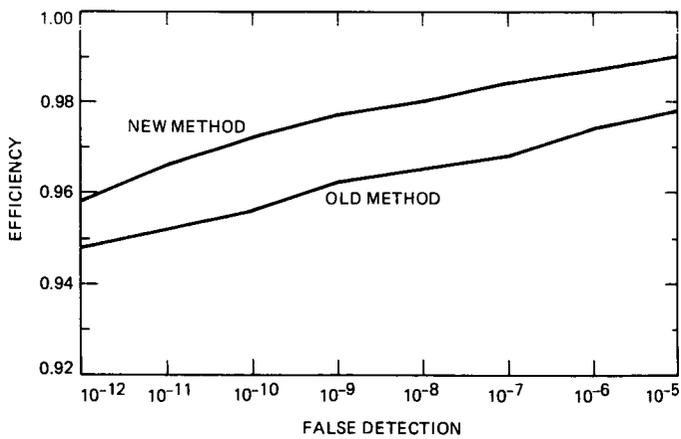
Fig. 6. Graph 5, SNR = 1.9
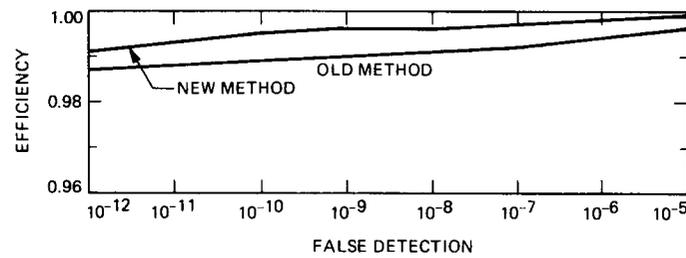
Fig. 8. Graph 7, SNR = 2.5

Cramer-Rao Bounds for Signal-to-Noise Ratio and Combiner Weight Estimation

S. J. Dolinar

Communications Systems Research Section

Cramer-Rao lower bounds on estimator variance are calculated for arbitrary unbiased estimates of signal-to-noise ratio and combiner weight parameters. Estimates are assumed to be based on a discrete set of observables obtained by matched filtering of a biphase modulated signal. The bounds are developed first for a problem model based on one observable per channel symbol period, and then extended to a more general problem in which subperiod observables are also available.

I. Introduction

This article calculates the Cramer-Rao bounds on the performance of arbitrary unbiased estimates of signal-to-noise ratio (SNR) and combiner weight parameters. Estimates are assumed to be based on a discrete set of observables obtained by matched filtering of a biphase modulated signal. Initially, we assume in Section II that exactly one observable or "sample" is available per channel symbol period. Later, in Section III, we consider a more general problem in which multiple observables or "subinterval samples" are obtained per symbol period by filtering over equal-length subintervals of each symbol period.

Estimates of signal-to-noise ratio and combiner weight are of interest in a variety of applications, such as symbol stream combining. In this article, Cramer-Rao bounds are determined directly for these parameters of interest, rather than for the underlying signal mean and noise variance parameters. This approach also reduces the mathematical complexity, because many expressions are separable functions of signal-to-noise ratio and combiner weight. The result is an almost-closed-form solution in which only one easily characterizable function of a single variable (SNR) is not explicitly determined.

Reference 1 provides additional background information on the significance of the parameters being estimated, and on the origin of our probabilistic model for the symbol period observables. Our model for the subinterval observables is a straightforward generalization, and it has been discussed previously as the basis for analyzing so-called "split-symbol" estimators (e.g., see Ref. 2) or "generalized" maximum likelihood estimators.[1]

II. Estimation with Symbol Period Sampling

We first consider estimation based on one sample per symbol period. Under this model, there are N discrete observables

---

[1] Vilnrotter, V. A., "A Generalized Class of Maximum Likelihood Estimators," IOM 331-86.5-82, Jet Propulsion Laboratory, Pasadena, Calif., January 13, 1986 (JPL Internal Document).

x_i, $i = 1, \cdots, N$, which can be represented in the form (see Eq. (7) of Ref. 1):

$$x_i = D_i m + n_i \sigma \qquad (1)$$

where $\{D_i\}$ is a data modulation sequence corresponding to the transmitted channel symbols, and $\{n_i\}$ is a noise sequence. As in Ref. 1, we assume that the $\{n_i\}$ are independent unit normal random variables, and that the $\{D_i\}$ are independent and take on the values +1 and -1 with equal probability. The unknown parameters m and σ represent the magnitudes of the "signal" and "noise" components of the observables $\{x_i\}$.

Our estimation problem is to estimate a signal-to-noise ratio parameter ρ and a combiner weight parameter α, based on the vector of observables $\mathbf{x} = (x_1, \cdots, x_N)$. The parameters ρ and α are defined in terms of the underlying signal and noise parameters m and σ as

$$\left. \begin{array}{l} \rho = \dfrac{m^2}{\sigma^2} \\[4mm] \alpha = \dfrac{m}{\sigma^2} \end{array} \right\} \qquad (2)$$

We note from Eq. (12) of Ref. 1 that the actual signal-to-noise ratio at the receiver is only one-half ρ, but we prefer the definition in Eq. (2) for mathematical convenience.

The log-likelihood function for this problem is taken from Eq. (20) of Ref. 1:

$$\frac{1}{N} \ln p(\mathbf{x}|m, \sigma) = -\frac{1}{2} \ln (2\pi) - \ln \sigma - \frac{\langle x^2 \rangle_N}{2\sigma^2}$$

$$- \frac{m^2}{2\sigma^2} + \left\langle \ln \cosh \frac{mx}{\sigma^2} \right\rangle_N \qquad (3)$$

where the notation $\langle \cdot \rangle_N$ represents a sample average value: for any function F applied to the N samples $x_1, \cdots, x_N$,

$$\langle F(x) \rangle_N \triangleq \frac{1}{N} \sum_{i=1}^{N} F(x_i) \qquad (4)$$

The log-likelihood function may also be written directly in terms of the signal-to-noise ratio and combiner weight parameters,

$$\frac{1}{N} \ln p(\mathbf{x}|\rho, \alpha) = -\frac{1}{2} \ln (2\pi) - \frac{1}{2} \ln \rho + \ln \alpha - \frac{\langle x^2 \rangle_N \alpha^2}{2\rho}$$

$$- \frac{1}{2} \rho + \langle \ln \cosh \alpha x \rangle_N \qquad (5)$$

The Cramer-Rao bound requires computation of the Fisher information matrix $\mathbf{J}$,

$$\mathbf{J} = \begin{bmatrix} -E\left\{ \dfrac{\partial^2 \ln p(\mathbf{x}|\rho, \alpha)}{\partial \rho^2} \right\} & -E\left\{ \dfrac{\partial^2 \ln p(\mathbf{x}|\rho, \alpha)}{\partial \rho \partial \alpha} \right\} \\[6mm] -E\left\{ \dfrac{\partial^2 \ln p(\mathbf{x}|\rho, \alpha)}{\partial \alpha \partial \rho} \right\} & -E\left\{ \dfrac{\partial^2 \ln p(\mathbf{x}|\rho, \alpha)}{\partial \alpha^2} \right\} \end{bmatrix}$$

$$\triangleq \begin{bmatrix} J_{11} & J_{12} \\[3mm] J_{21} & J_{22} \end{bmatrix} \qquad (6)$$

After evaluating the indicated derivatives, we find that

$$\left. \begin{array}{l} \dfrac{1}{N} \rho^2 J_{11} = \dfrac{E\{\langle \alpha^2 x^2 \rangle_N\}}{\rho} - \dfrac{1}{2} \\[5mm] \dfrac{1}{N} \rho\alpha J_{12} = \dfrac{1}{N} \rho\alpha J_{21} = \dfrac{-E\{\langle \alpha^2 x^2 \rangle_N\}}{\rho} \\[5mm] \dfrac{1}{N} \alpha^2 J_{22} = 1 + \dfrac{E\{\langle \alpha^2 x^2 \rangle_N\}}{\rho} - E\{\langle \alpha^2 x^2 \text{ sech}^2 \alpha x \rangle\}_N \end{array} \right\} \qquad (7)$$

The expectations in Eq. (7) may be evaluated by substituting Gaussian random variables $\{u_i\}$ for the non-Gaussian random variables $\{x_i\}$. Defining

$$u_i = \alpha D_i x_i \qquad (8)$$

where α is the unknown combiner weight and $D_i = \pm 1$ is the random data modulation embedded in x_i, we see that u_i is Gaussian with mean and variance both equal to ρ:

$$E\{u_i\} = \alpha m = m^2/\sigma^2 = \rho$$

$$E\{u_i^2\} = \alpha^2(m^2 + \sigma^2) = \rho^2 + \rho \qquad (9)$$

$$\text{var}\{u_i\} = E\{u_i^2\} - [E\{u_i\}]^2 = \rho$$

Furthermore, because $D_i^2 = 1$ and because sech $(\cdot)$ is an even function of its argument,

$$E\{\langle \alpha^2 x^2 \rangle_N\} = E\{u_i^2\} = \rho^2 + \rho$$

$$E\{\langle \alpha^2 x^2 \text{ sech}^2 \alpha x \rangle_N\} = E\{u_i^2 \text{ sech}^2 u_i\} \triangleq E_2(\rho) \qquad (10)$$

The second expectation is written as $E_2(\rho)$, which is not determined in closed form. However, it is important to note that $E_2(\rho)$ is a function of ρ only, because the statistics of u_i are a function of ρ only. Inserting these results into the expressions for J_{ij}, we obtain

$$\frac{1}{N}\rho^2 J_{11} = \rho + \frac{1}{2}$$

$$\frac{1}{N}\rho\alpha J_{12} = \frac{1}{N}\rho\alpha J_{21} = -(\rho + 1) \qquad (11)$$

$$\frac{1}{N}\alpha^2 J_{22} = 2 + \rho - E_2(\rho)$$

In terms of $\mathbf{J}$, the Cramer-Rao bound (Ref. 3) states that for any *unbiased* estimates $\hat{\rho}, \hat{\alpha}$ of the unknown parameters ρ, α,

$$\frac{\text{var}(\hat{\rho})}{\rho^2} \geq \frac{(\mathbf{J}^{-1})_{11}}{\rho^2}$$

$$\frac{\text{var}(\hat{\alpha})}{\alpha^2} \geq \frac{(\mathbf{J}^{-1})_{22}}{\alpha^2} \qquad (12)$$

Calculating $\mathbf{J}^{-1}$ from Eq. (11), we obtain

$$\frac{\text{var}(\hat{\rho})}{\rho^2} \geq \frac{2}{N}\frac{2 + \rho - E_2(\rho)}{\rho - (2\rho + 1)E_2(\rho)}$$

$$\frac{\text{var}(\hat{\alpha})}{\alpha^2} \geq \frac{2}{N}\frac{\rho + \frac{1}{2}}{\rho - (2\rho + 1)E_2(\rho)}$$

general result for arbitrary N, ρ, α (13)

Note that both fractional variance bounds are functions of N and ρ but not α. These are exact expressions so far, but further analysis requires characterization of the function $E_2(\rho)$. This function is easy to evaluate numerically, but first we consider its limiting behavior for large and small ρ. The general case is discussed and plotted at the end of Section III of this article.

A. High SNR Case

For large ρ, it can be shown that the function $E_2(\rho)$ is exponentially small,

$$E_2(\rho) \sim \rho^{-1/2}e^{-\rho/2} \qquad \rho \gg 1 \qquad (14)$$

Thus, the fractional variance bounds can be written

$$\frac{\text{var}(\hat{\rho})}{\rho^2} \gtrsim \frac{2}{N}\left[1 + \frac{2}{\rho}\right]$$

$$\frac{\text{var}(\hat{\alpha})}{\alpha^2} \gtrsim \frac{2}{N}\left[1 + \frac{1}{2\rho}\right] \qquad \rho \gg 1 \qquad (15)$$

Both of these expressions are accurate within terms that are exponentially small in ρ (i.e., there are no $1/\rho^n$ terms for $n > 1$).

B. Low SNR Case

For small ρ, we make use of the Taylor series expansion for $u^2 \text{ sech}^2 u$ around $u = 0$,

$$u^2 \text{ sech}^2 u = u^2 - u^4 + \frac{2}{3}u^6 - \frac{17}{45}u^8 + \cdots \qquad (16)$$

and apply the formula for the moments of a Gaussian random variable (Ref. 4) with mean and variance both equal to ρ,

$$E\{u^2\} = \rho^2 + \rho$$

$$E\{u^4\} = \rho^4 + 6\rho^3 + 3\rho^2$$

$$E\{u^6\} = \rho^6 + 15\rho^5 + 45\rho^4 + 15\rho^3$$

$$E\{u^8\} = \rho^8 + 28\rho^7 + 210\rho^6 + 420\rho^5 + 105\rho^4 \qquad (17)$$

This leads to a Taylor series expansion of the function $E_2(\rho)$ around $\rho = 0$:

$$E_2(\rho) = \rho - 2\rho^2 + 4\rho^3 - \frac{32}{3}\rho^4 + \cdots \qquad (18)$$

When this expression is inserted back into Eq. (13), all of the ρ, ρ^2, and ρ^3 terms in the denominator cancel, i.e.,

$$\rho - (2\rho + 1)E_2(\rho) = \frac{8}{3}\rho^4 + \text{higher order terms} \qquad (19)$$

This deep singularity at $\rho = 0$ causes the bounds on the fractional variance to be very large for low SNR.

$$\left. \begin{array}{l} \dfrac{\text{var}(\widehat{\rho})}{\rho^2} \gtrsim \dfrac{3}{2N\rho^4} \\[3mm] \dfrac{\text{var}(\widehat{\alpha})}{\alpha^2} \gtrsim \dfrac{3}{8N\rho^4} \end{array} \right\} \quad \rho \ll 1 \qquad (20)$$

III. Estimation with Subinterval Sampling

Now we consider the same type of bound for a more general model in which multiple subinterval samples are taken before their mean value has a chance to change sign. We assume M independent and equally spaced subinterval samples X_{ij}, $j = 1, \cdots, M$, for each of the N symbol intervals. The subinterval samples are modeled as

$$X_{ij} = D_i m_0 + n_{ij} \sigma_0 \qquad (21)$$

where $D_i = \pm 1$ is the same data modulation variable defined in Eq. (1), $\{n_{ij}\}$ are independent unit normal random variables, and m_0 and σ_0 denote the subinterval signal and noise parameters. Note that the M subinterval samples X_{ij}, $j = 1, \cdots, M$, are affected by one data modulation variable D_i and M independent noise variables n_{ij}, $j = 1, \cdots, M$.

Our estimation problem is still to estimate the signal-to-noise ratio and combiner weight parameters for the full symbol period. Each block of M subinterval samples X_{ij}, $j = 1, \cdots, M$, sums to form a symbol period sample x_i,

$$x_i = \sum_{j=1}^{M} X_{ij} \qquad (22)$$

This implies that the subinterval signal and noise parameters appearing in Eq. (21) are related to the full symbol period parameters by

$$\left. \begin{array}{l} m_0 = \dfrac{m}{M} = \dfrac{\rho}{M\alpha} \\[3mm] \sigma_0^2 = \dfrac{\sigma^2}{M} = \dfrac{\rho}{M\alpha^2} \end{array} \right\} \qquad (23)$$

The log-likelihood function for the vector of subinterval observables $\mathbf{X} = (X_{11}, \cdots, X_{1M}, \cdots, X_{N1}, \cdots, X_{NM})$ is obtained analogously to Eq. (3) in terms of the subinterval signal and noise parameters as

$$\frac{1}{MN} \ln p(\mathbf{X}|m_0, \sigma_0) = -\frac{1}{2} \ln 2\pi - \ln \sigma_0 - \frac{\langle X^2 \rangle_{MN}}{2\sigma_0^2}$$

$$\qquad (24)$$

$$- \frac{m_0^2}{2\sigma_0^2} + \frac{1}{M} \left\langle \ln \cosh \frac{m_0 x}{\sigma_0^2} \right\rangle_N$$

or, alternatively, in terms of the symbol period signal-to-noise ratio and combiner weight parameters as

$$\frac{1}{MN} \ln p(\mathbf{X}|\rho, \alpha) = -\frac{1}{2} \ln \left(\frac{2\pi}{M}\right) - \frac{1}{2} \ln \rho + \ln \alpha - \frac{\langle X^2 \rangle_{MN} \alpha^2}{2\rho/M}$$

$$- \frac{1}{2} \frac{\rho}{M} + \frac{1}{M} \langle \ln \cosh \alpha x \rangle_N \qquad (25)$$

In Eqs. (24) and (25), $\langle X^2 \rangle_{MN}$ denotes the mean square value of the MN subinterval samples

$$\langle X^2 \rangle_{MN} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij}^2 \qquad (26)$$

and $\langle \ln \cosh \alpha x \rangle_N$ is the same quantity appearing in Eqs. (3) and (5), i.e., an average based on the full symbol period samples x_i,

$$\langle \ln \cosh \alpha x \rangle_N = \frac{1}{N} \sum_{i=1}^{N} \ln \cosh \alpha x_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} \ln \cosh \left(\alpha \sum_{j=1}^{M} X_{ij} \right) \qquad (27)$$

We observe from comparing Eqs. (5) and (25) that the Cramer-Rao bounds for this problem can be obtained trivially from the bounds derived earlier by substituting $ME\{\langle \alpha^2 X^2 \rangle_{MN}\}$ for $E\{\langle \alpha^2 x^2 \rangle_N\}$ and $M^{-1} E\{\langle \alpha^2 x^2 \text{ sech}^2 \alpha x \rangle_N\}$ for $E\{\langle \alpha^2 x^2 \text{ sech}^2 \alpha x \rangle_N\}$. We note that

$$ME\{\langle\alpha^2 X^2\rangle_{MN}\} = M\alpha^2(m_0^2 + \sigma_0^2)$$

$$= \rho\left(1 + \frac{\rho}{M}\right) \tag{28}$$

$$\frac{1}{M}E\{\langle\alpha^2 x^2 \operatorname{sech}^2\alpha x\rangle_N\} = \frac{1}{M}E_2(\rho)$$

where $E_2(\rho)$ is the same function defined earlier. The Fisher information matrix elements can immediately be evaluated from Eqs. (7) and (28):

$$\frac{1}{MN}\rho^2 J_{11} = \frac{\rho}{M} + \frac{1}{2}$$

$$\frac{1}{MN}\rho\alpha J_{12} = \frac{1}{MN}\rho\alpha J_{21} = -\left(\frac{\rho}{M} + 1\right) \tag{29}$$

$$\frac{1}{MN}\alpha^2 J_{22} = 2 + \frac{\rho}{M} - \frac{1}{M}E_2(\rho)$$

Inversion of this matrix produces the bounds on the fractional variance of the estimates $\hat{\rho}, \hat{\alpha}$:

$$\frac{\operatorname{var}(\hat{\rho})}{\rho^2} \geq \frac{2}{N}\frac{2M + \rho - E_2(\rho)}{M\rho - (2\rho + M)E_2(\rho)}$$

$$\left.\begin{array}{l}\text{general result for} \\ \text{arbitrary } N, M, \rho, \alpha\end{array}\right\} \quad (30)$$

$$\frac{\operatorname{var}(\hat{\alpha})}{\alpha^2} \geq \frac{2}{N}\frac{\rho + M/2}{M\rho - (2\rho + M)E_2(\rho)}$$

A. High SNR Case

For high values of ρ, the bounds reduce to

$$\frac{\operatorname{var}(\hat{\rho})}{\rho^2} \gtrsim \frac{2}{MN}\left(1 + \frac{2M}{\rho}\right)$$

$$\left.\right\} \quad \rho \gg 1 \tag{31}$$

$$\frac{\operatorname{var}(\hat{\alpha})}{\alpha^2} \gtrsim \frac{2}{MN}\left(1 + \frac{M}{2\rho}\right)$$

As before, these bounds are accurate within terms that are exponentially decreasing with ρ. We see that the performance bound improves with the total number of samples MN, regardless of whether they are subinterval samples or symbol period samples. However, this conclusion is not correct if the number of *subinterval* samples gets arbitrarily large. If M is increased beyond the value of ρ, the bounds eventually saturate at

$$\frac{\operatorname{var}(\hat{\rho})}{\rho^2} \gtrsim \frac{4}{N\rho}$$

$$\left.\right\} \quad 1 \ll \rho \ll M \tag{32}$$

$$\frac{\operatorname{var}(\hat{\alpha})}{\alpha^2} \gtrsim \frac{1}{N\rho}$$

B. Low SNR Case

For low SNR, the deep singularity at $\rho = 0$ in the denominator of the accuracy bounds is partially relaxed for $M > 1$, i.e.,

$$M\rho - (2\rho + M)E_2(\rho) = 2(M - 1)\rho^2 + \text{higher order terms}$$

$$\text{for } M > 1 \tag{33}$$

The "higher order terms" in Eq. (33) are small with respect to $(M - 1)\rho^2$ as ρ gets small, no matter how large M is. The accuracy bounds are approximately

$$\frac{\operatorname{var}(\hat{\rho})}{\rho^2} \gtrsim \frac{2}{N\rho^2}\frac{M}{M-1}$$

$$\left.\right\} \quad \rho \ll 1 < M \tag{34}$$

$$\frac{\operatorname{var}(\hat{\alpha})}{\alpha^2} \gtrsim \frac{1}{2N\rho^2}\frac{M}{M-1}$$

We see from Eqs. (20) and (34) that the performance bounds for small ρ improve by a large factor $3/(8\rho^2)$ in going from $M = 1$ to $M = 2$, and then by only an additional factor of 2 from $M = 2$ to $M = \infty$.

C. Large Number of Subinterval Samples Case

We have seen that the performance bound saturates at a nonzero limit for both the low SNR and high SNR cases, as the number of subinterval samples M goes to infinity. This saturation value can be calculated from Eq. (30) for all SNR values, in terms of the function $E_2(\rho)$.

$$\frac{\operatorname{var}(\hat{\rho})}{\rho^2} \gtrsim \frac{4}{N[\rho - E_2(\rho)]}$$

$$\left.\right\} \quad M \gg \max(\rho, 1) \tag{35}$$

$$\frac{\operatorname{var}(\hat{\alpha})}{\alpha^2} \gtrsim \frac{1}{N[\rho - E_2(\rho)]}$$

D. General Case

The Cramer-Rao bounds for the general case are plotted in Figs. 1 and 2 for the signal-to-noise ratio and combiner weight estimates, respectively. Each curve shows the lower bound on

the fractional estimator variance times the number of symbol period samples N, as a function of signal-to-noise ratio ρ. Curves are drawn for various numbers of subinterval samples M, including the case $M = 1$, which is equivalent to the case of full symbol period sampling considered in Section II of this article.

The ordinate in these plots may be interpreted as a lower bound on the number of symbol period samples N required to achieve a fractional estimator variance of 100%. If a smaller fractional estimator variance is desired, say ϵ, the bound on the required number of samples is simply increased by the factor $1/\epsilon$.

IV. Conclusions

Figures 1 and 2 present a strong case for taking subinterval samples. In the low SNR region, split-symbol estimators can potentially reduce the number of required samples by orders of magnitude relative to estimates based entirely on full sym-bol period samples. The bulk of this reduction results from splitting the symbol period in half ($M = 2$), and additional improvement is limited to 3 dB as the number of subinterval samples is increased further. In the high SNR region, it pays to keep increasing the number of subinterval samples, but rapidly diminishing returns are encountered when the number of subinterval samples is increased beyond the true value of the signal-to-noise ratio. Of course, if the true SNR is extremely high, the practical limit on the number of worthwhile subin-terval samples may be set by bandwidth constraints rather than SNR constraints.

A caveat must be attached to all of the analysis, and hence the conclusions, in this article. Performance bounds derived here apply only to unbiased estimators. Perhaps the require-ment that the full symbol period estimator be perfectly un-biased is too tight a constraint to impose in the low SNR region. Further work should investigate the possible trade-offs between estimator bias and estimator variance, especially for full symbol period estimators at low SNR.

Acknowledgment

References

1. Kalson, S., and S. J. Dolinar, "Maximum Likelihood Estimation of Signal-to-Noise Ratio and Combiner Weight," TDA Progress Report 42-85, pp. 32-41, Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1986.

2. Simon, M. K., and A. Mileant, "SNR Estimation for the Baseband Assembly," TDA Progress Report 42-85, pp. 118-126, Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1986.

3. Van Trees, H. L., Detection, Estimation, and Modulation Theory, Part I, John Wiley & Sons, New York 1968.

4. Kindal, M. G., and A. Stuart, The Advanced Theory of Statistics, Vol. I, Hafner Publishing Co., New York, 1973.
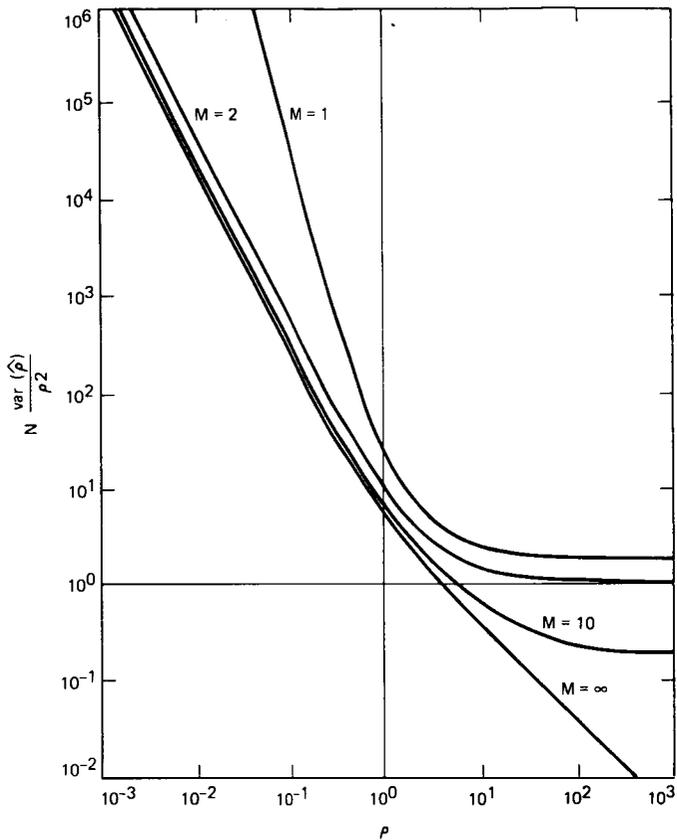
Fig. 1. Cramer-Rao lower bound on unbiased signal-to-noise ratio estimator performance
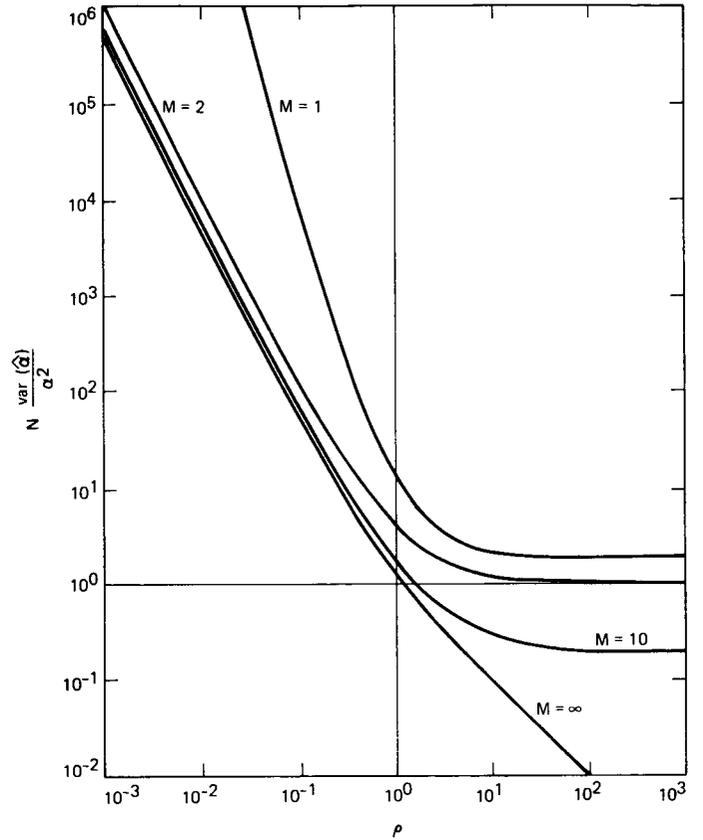
Fig. 2. Cramer-Rao lower bound on unbiased combiner weight estimator performance

Antenna Arraying of Voyager Telemetry Signals by Symbol Stream Combining

W. J. Hurd, J. Rabkin, M. D. Russell, and B. Siev
Communications Systems Research Section

H. W. Cooper
Radio Frequency and Microwave Subsystems Section

T. O. Anderson and P. U. Winter
DSN Data Systems Section

Telemetry signals received from the Voyager 2 spacecraft at Deep Space Stations at Parkes and Canberra, Australia, on February 6, 1986, were combined by the method of symbol stream combining. This second demonstration of symbol stream combining followed the International Cometary Explorer (ICE) demonstration at Giacobini-Zinner encounter in September 1985. The Voyager demonstration was at a symbol rate of 43.2 ksymb/s, compared to 2 ksymb/s for ICE. Recording, playback, and combining at this higher rate were demonstrated. The average symbol signal-to-noise ratio (SNR) of the combined data was 2.84 dB, or 0.23 dB less than the sum of the SNRs of the two input symbol streams. This 0.23 loss from ideal combining was due to use of 4-bit quantization of the input symbol streams and imperfect scaling. A practical implementation with 8-bit quantization could achieve combining losses of under 0.05 dB over a wide dynamic range of input signal levels.

I. Introduction

Use of symbol stream combining for arraying antennas to enhance telemetry performance was first demonstrated in 1985. Signals from the International Cometary Explorer (ICE) spacecraft were combined as the spacecraft encountered the comet Giacobini-Zinner (Ref. 1). The data rate was 2 ksymb/s. This paper describes a second demonstration, using data from Voyager 2 taken February 6, 1986, at a data rate of 43.2 ksymb/s. This higher rate was enabled by changes in both the data acquisition and the combining systems.

The basic system configurations for symbol stream combining and baseband combining are shown in Fig. 1. First, consider real-time systems. In baseband combining, a broadband ground communications link is used to transmit the baseband signal to the local station where the combining is done. The link bandwidth is typically 5 MHz. The combined baseband signals are processed by the telemetry detector (subcarrier demodulator and symbol synchronizer and detector) and the decoder at the local station. In symbol stream combining, the signals at the two stations are processed through telemetry detection. The symbol streams are represented as sequences of

digital numbers. The symbols from the remote station are transmitted over a digital link to the local station where they are combined and decoded. The symbol quantization is typically 4- to 8-bits. For 8-bit quantization and the maximum symbol rate of 43.2 ksymb/s for Voyager at Neptune encounter, the link data rate is 346 kb/s. These data can be transmitted with an order of magnitude less link capacity than required by the baseband system.

Now consider non-real-time combining. For the Parkes baseband array, the baseband signal was low-pass filtered and then sampled at 6 MHz with 4-bit quantization, and recorded with a digital recording rate of 24 Mb/s. The symbol stream data for the present demonstration were quantized to 4 bits and recorded at 173 kb/s, for a two-orders-of-magnitude savings in recording rate over baseband combining.

The reduced data rates for real-time and non-real-time realizations are the main advantages of symbol stream combining over baseband combining. Combining is also easier, since the data rate is lower, and combining can be done in either hardware or software. Combining performance is at least as good for symbol stream combining as for baseband combining, with losses of 0.05 dB ± 0.05 dB possible in an 8-bit quantized symbol stream combining system. Total system losses for the Parkes baseband array are specified as less than 0.4 dB, due to inability to measure performance much better than this operationally.

Other telemetry system losses are approximately the same using symbol stream or baseband combining. Carrier tracking (radio) losses are the same. Subcarrier and symbol tracking losses will be small in a demodulation system designed for low symbol SNRs, such as the DSN Advanced Receiver. Subcarrier losses are kept small at all symbol SNRs of interest by the new technique of Hurd and Aguirre (Ref. 2).

II. Combining Theory

This section presents the basic relationship for telemetry arraying by symbol stream combining. Losses due to quantization of the symbol values are discussed. Then the methods used to estimate SNR are presented. Finally, overall performance is summarized.

A. Problem Formulation

A block diagram showing the mathematics of symbol stream combining is given in Fig. 2. The input detected symbols from Stations 1 and 2 are denoted by X_{1i} and X_{2i} for the ith symbol. These may be quantized by the detection, recording, ground link communications, or combining process, to obtain Y_{1i} and Y_{2i}. These values are multiplied by weights

W_1 and W_2, and the products are summed to obtain the combined symbols, Z_i. The Z_i may be further quantized to form the symbol values S_i, which are processed by the decoder. From now on, we neglect this output quantization, since it is normally a property of the decoder rather than of the combiner. That is, decoders typically quantize the inputs to 3 bits. This dominates the resolution of the combiner.

B. Combining Unquantized Symbol Values

The case of no-input quantization has been well analyzed by Vo (Ref. 3). Suppose that the input symbol mean values for binary symbol values ±1 are $\pm m_1$ and $\pm m_2$, that the variances are σ_1^2 and σ_2^2, and that these parameters are constant. Then the input symbol SNRs are

$$R_i = \frac{m_i^2}{2\sigma_i^2}$$

The optimum combining weights are

$$W_i = \frac{m_i}{\sigma_i^2}$$

and the resulting output SNR, i.e., the symbol SNR of Z_i, is

$$R_0 = R_1 + R_2$$

This is the best that can be achieved, and is the same as the best that can be achieved in baseband combining.

In practice, one does not know the input parameters a priori, so they need to be estimated. Vo (Ref. 3) has studied this problem and has determined the loss in output symbol SNR versus the number of symbols used to estimate the SNRs. The SNR estimation method is the absolute moment method, with unbiasing. For input SNRs of –1 dB and –3 dB, the loss is approximately 0.03 dB if 1000 points are used to estimate the SNRs, and under 0.005 dB if 10,000 points are used. The loss is smaller at higher input SNRs.

C. Effects of Quantization

When there is quantization of the input symbols, the probability distributions of the quantized input symbols and of the combined symbols are no longer Gaussian, conditioned on the transmitted symbol values. Performance, or loss due to quantization, may not be characterized accurately by signal-to-noise ratio. However, SNR is a convenient measure, and our approach here is to determine the loss in SNR due to quantization. Even though our approach does not accurately characterize telemetry system performance, it does accurately char-

acterize loss in symbol SNR, and agreement between theory and experimental results confirms understanding and proper operation of the system. The SNR of the quantized and combined symbols is maximized by using the means and variances of the quantized symbols in calculating the weights. Letting primes denote quantization, the weights are

$$W'_i = \frac{m'_i}{(\sigma'_i)^2}$$

where m'_i and σ'_i are the mean and standard deviation of the station i symbols after quantization. Then the SNRs of the quantized symbols are

$$R'_i = \frac{(m'_i)^2}{2(\sigma'_i)^2}$$

and the best achievable output SNR is

$$R'_0 = R'_1 + R'_2$$

If the quantizer losses (in decibels) are the same for both quantizers, then the loss at the output is equal to the loss in either quantizer. Thus, the maximum of the two quantizer losses is a simple upper bound on loss.

Figure 3 shows the input quantization loss in decibels, i.e., $10 \log_{10} (R_i/R'_i)$, for 4-bit quantization, with various values of m, in units of the quantizer output. The value $m = 2.7$ corresponds to the nominal value if the 8-bit quantized symbols at the output of a DSN Symbol Synchronizer Assembly (SSA) or Demodulator Synchronizer Assembly (DSA) are re-quantized to the appropriate 4 bits. The other values are for gain variations of 3 dB and 6 dB from the nominal. Note that the loss is sometimes negative, i.e., there is a gain in SNR. This occurs when the quantizer is saturating, which reduces variance more than it reduces conditional mean square. This does not indicate a gain in telemetry system performance, but it does indicate an increase in mean square to variance of the symbol values.

We computed the loss in SNR for the parameters of the demonstration, obtaining a calculated loss of 0.13 dB ± 0.1 dB.

1. Decoding of quantized and combined symbols. Prior to the decision to use 4-bit quantization for the demonstration, Pollara and Swanson (Ref. 4) simulated the effect of input symbol quantization, through the decoder, including the effect of 3-bit quantization at the decoder input. The input SNRs and weights were equal. They concluded that the loss due to combining the combiner input quantization is approxi-

mately 0.04 dB ± 0.02 dB at input symbol SNRs of −2 dB to −3 dB. From Fig. 3, there is a gain in SNR of 0.1 dB at −3 dB and of 0.04 dB at −2 dB. This illustrates the inadequacy of the mean square to variance SNR approach insofar as characterizing telemetry performance. On the positive side, it concludes that degradation due to quantization is small.

2. Summary of quantization effects. We conclude that quantization is the dominant effect on symbol SNR, i.e., it is a much larger effect than the loss in SNR due to imperfect weights. Quantization can cause either an increase or a decrease in SNR, or mean square to variance. The degradation in telemetry performance, through the decoder, is approximately 0.04 dB ± 0.02 dB, for typical cases of interest for Voyager. Accurate characterization of the telemetry performance of a combiner with 4-bit quantization would require extensive simulations over a wide range of quantizer scalings and input SNRs.

D. SNR Estimation for Quantized Symbols

Use of 4-bit quantization has a significant effect on the accuracy of the unbiased moment method of SNR and weight estimation. The quantization noise is very large at low input signal levels, and saturation dominates at high input signal levels. In either case, severe biases occur in the estimates of SNR, mean, and standard deviation.

To overcome this problem, Vilnrotter and Rodemich (Ref. 5) developed a new method of estimating the input parameters from the sample absolute first and second moments. This method is to perform a two-dimensional table look-up to determine m and σ from the sample moments. The method accounts for the DC offset inherent in the quantizer code. Preliminary evaluations of the method indicate that the estimates of SNR have zero bias, ±0.02 dB, for input symbol SNRs from −3 dB to 1.5 dB, and for input means of 1.3 to 3.8. The effect of this on the combiner weight bias causes a negligible loss in combiner output SNR. Random variations in the SNR estimates were also evaluated. For the 5700 points used to estimate SNR in the software combiner, the standard deviation of SNR estimation was approximately 0.017 dB for SNRs from −1 dB to 1 dB. This performance is close to that of the unbiased moment method without quantization, so the effect on weight estimation is similar.

The table look-up method estimates the signal mean and noise standard deviation at the input of the quantizers, rather than at the output. In the demonstration, we used these values to compute the weights. The loss due to using this procedure is negligible because the scaling and SNRs of the two symbol streams were so similar that differential biases are negligible. However, an implementation should change the algorithms to use the estimates of the parameters at the quantizer outputs.

E. Performance Prediction Summary

Performance is dominated by quantization of the input symbols. The only other loss is in use of improper combiner weights, but this can be kept to less than 0.01 dB by using sufficient data to estimate the SNRs and the weights. This assumes that the signal parameters are constant or slowly varying. SNR is not an accurate measure of system performance, because of the nonlinearities. Decoder performance should be considered.

For 4-bit quantization, performance losses can be as low as 0.05 dB with proper quantizer scaling and at the SNRs of primary interest, near decoder threshold.

It appears that use of 5-bit quantization would enable losses of less than 0.05 dB over a wide range of scaling. However, this is not a convenient quantization for non-real-time systems, in which the data are typically stored in byte-oriented storage media. Use of 8-bit quantization would make quantization losses negligible — less than 0.01 dB. Time-varying signal level effects might begin to dominate in some situations, since good SNR estimates require 5000 to 10,000 symbols.

Use of 8-bit quantization is probably the best system trade unless ground communications capacity or storage capacity dominates costs. Losses of under 0.05 dB should be easily realizable, with 0.01 dB possible.

F. Accuracy of Performance Evaluation

Although errors in estimating SNRs, and thus combiner weights do not limit combiner performance, the errors in SNR do limit the accuracy of measurement of SNR loss. Since random effects can be reduced by averaging many estimates of input and output SNR, the main effect is estimation bias. We estimate the error in bias of estimations of the input SNRs to be less than 0.02 dB.

To estimate the output SNR, we assumed that the combined symbols are Gaussian-distributed, with the appropriate conditional mean. This is not strictly true, because they are the sums of numbers that are coarsely-quantized and then weighted. Their distribution is very complex and depends on the exact weights. Therefore, for lack of ability to do any better, we estimated the output SNR by applying the unbiased moment method, implicitly making the Gaussian assumption. We estimate that there is less than 0.1 dB bias in the software combining system.

Counting 0.02 dB for input SNR bias, 0.1 dB for output SNR bias, and 0.01 dB for other effects, we estimate our ability to measure performance to be ±0.13 dB.

We could reanalyze the data to obtain more accurate performance measures. One method would be to decode the combined data, reconstruct the input symbol streams, and recompute all SNR estimates using the "known" symbol values. The demonstration data have high SNR compared to decoder threshold, so there would be almost no decoder errors. Therefore, the SNR estimates would be very accurate. DC offsets, believed to be insignificant and neglected in the other SNR estimation methods, could be accounted for. The accuracy would be 0.01 dB or better.

III. Demonstration Description

The Voyager demonstration combined telemetry data received at Parkes and Canberra, Australia, on February 6, 1986. The symbols were recorded on standard 1600-bpi digital computer tapes which were shipped to JPL for combining. Combining was done by two different methods. Using a hardware combiner, the symbols were combined and decoded at 60 ksymb/s, faster than the 43.2-ksymb/s data rate. This demonstrated that data could be combined and decoded in near-real time, i.e., with no data backlog buildup. The data were also combined in software at slower than real time. This was done to achieve more accurate SNR and performance estimates than achieved in the hardware system.

A. Station Configurations

The station configurations are shown in Fig. 4. The receivers were the Parkes Telemetry Receiver at Parkes and a Block IV DSN receiver at DSS 43, Canberra. The receiver baseband outputs were processed by DSAs, which performed subcarrier synchronization and demodulation, symbol synchronization, and symbol detection. The detected symbols were sent to Telemetry Processor Assembly (TPA) Modcomp computers via specially designed interfaces. These are designated SSTIs, which stands for Symbol Synchronizer Assembly (SSA) — Symbol Stream Combiner (SSC) — TPA Interfaces. The symbol streams were recorded onto TPA magnetic tapes.

1. SSTI. The SSTIs were specially designed, fabricated, and installed for this demonstration. They were required in order to record at the 43.2-ksymb/s Voyager rate. They replaced the standard SSA-TPA couplers used in the earlier ICE Giacobini-Zinner demonstration. The SSTIs accept input 8-bit quantized symbol streams from the DSAs, quantize these numbers to 4 bits, and pack two consecutive 4-bit symbols into one 8-bit byte. The byte streams, at one-half the rate of the incoming symbols, are output to the TPAs over DSN 14-line Standard Interfaces. The SSTIs are also used on playback, to transfer the byte streams read out of the Modcomps to the symbol stream combiner.

2. TPA software. The standard Mark IV TPA software was extensively modified to meet the soft-symbol recording requirements and to accommodate the Voyager spacecraft data rate. The system was tested at symbol rates up to 80 ksymb/s, compared to the maximum rate of 43.2 ksymb/s at Voyager Uranus encounter.

The main modifications included removing all Local Monitor and Control (LMC) interfaces, adding operator prompts and directives to the local terminal, removing the commands to the SSA and SSA/TPA coupler, adding the commands to the SSTI and making changes to the magnetic tape handling software. The tape handling changes were made to avoid any unnecessary tape writing delays such as multiple retries on tape errors and time lost switching tapes. The tape handling software rewinds the magnetic tape when the end-of-tape marker is sensed, switches to the alternate tape on the drive, and continues recording without losing data. It prompts the operator to remove the tape just written or read, and to place a new tape on the drive, thereby preparing it for the next tape switch. Approximately every 31 seconds during the recording, the software displays the signal level and the biased SNR estimates. This enables the operator to monitor and control the signal level from the DSA or SSA, and to monitor SNR to assure subcarrier and symbol lock. The data received by the TPA from the SSTI are put into buffers in the standard Original Data Record (ODR) format shown in Fig. TLM-3-13-2 of the JPL Internal Document 820-13 (*Deep Space Network System Requirements. Detailed Interface Design*, Jet Propulsion Laboratory, Pasadena, California, May 1, 1986). Five of these buffers comprise one tape record. Each record has time tags determined by the station clock.

For test purposes, three programs were developed to verify the operation of the recording system. These programs perform pseudo-random noise code verification, count pattern verification, and time-pulse verification.

Playback software was included in the same software package. It reads the soft symbols recorded by the soft-symbol recording software and transmits them to the SSTI. The playback software requires from the terminal operator directives which define the recorded time of the first data and the time at which the transmission is to start. These start times make it possible to line up two sets of symbol streams recorded at two different sites, to within approximately 2 ms. The playback software reads the ODR blocks from the tape, aligns the times, strips off the headers and presents the SSTI with a continuous stream of two-symbol bytes. Approximately every ten seconds during playback, the software displays the current time and the block header time on the block being processed.

B. Installation, Testing, and Operations

The recording and test software and the SSTIs were extensively tested at the JPL Compatibility Test Area, CTA 21, in August 1985. The hardware and software were then shipped to Australia and installed and tested at Parkes and Canberra in October 1985, prior to station configuration freeze for Voyager Uranus encounter. The system was then tested again prior to the first scheduled day of data acquisition. Three days were scheduled for data acquisition. Operations were so smooth on the first day, February 6, that the additional days were cancelled.

C. Near-Real-Time Combining System

The near-real-time combining system was installed and operated in the Compatibility Test Area (CTA-21) at JPL. A block diagram is shown in Fig. 5.

The recorded data are played back using two Modcomp computers. Either TPAs or Area Routing Assembly (ARA) computers can be used if equipped with magnetic tape units and time-code translators. The recorded byte streams, with two symbols per byte, are output from the Modcomps over the DSN Standard Interface to the SSTI, and then to the Symbol Stream Combiner. The SSC has been described in an earlier article (Ref. 6). It (1) unpacks the bytes into symbols, (2) buffers the symbols in two memories, (3) aligns the streams by cross-correlation, (4) measures the power and symbol SNR of each stream, (5) calculates weights, (6) combines the two streams by multiplying each one by the proper weight and summing the two products, (7) measures the output SNR, and (8) displays its status and the three SNRs and records these on a printer and on a floppy disk.

The symbol stream at the combiner output is in the same physical format as symbols at the output of the SSAs or DSAs. These symbols are sent to a Maximum-likelihood Convolutional Decoder (MCD) and a TPA Modcomp for decoding and recording of the decoded data bits. The TPA logs estimates of bit SNR, made in the MCD, onto a printer. During the demonstration, the decoded bits were not recorded due to lack of tape units.

Initial approximate alignment of the data tapes is accomplished by keying the data start time and the playback start time into the Modcomps, via terminals. The computers then read the time tags in the tape data block headers, advance the tapes to the data start times, and then start playback when the CTA 21 station clock reaches the playback start time. Playback rate is controlled by the symbol stream combiner which synchronously clocks two-symbol bytes out of each Modcomp's standard interface. The SSTI and the symbol

stream combiner assure that one byte is read from each Modcomp on each clock pulse.

D. Software Combining System

The combining software developed for the ICE Giacobini-Zinner demonstration was modified and used to process Voyager data. Processing was considerably slower than real time because of the relatively high data rate. The main reason for software processing was to achieve more accurate estimates of input and output symbol SNRs than in the near-real-time system. Better SNR measurements are achieved because the near-real-time system does not use all of the symbol values in estimating SNRs due to software speed limitations.

Only two fundamental modifications were required to the existing processing software. The first was to unpack the symbols, which were two per byte, and to enlarge the data arrays. The second was to modify the symbol SNR estimation algorithm for the input symbols. The new (table look-up) algorithm is described in the previous section.

IV. Demonstration Results

Symbol streams were recorded at Parkes and Canberra, DSS 43, on February 6, 1986, Day-of-Year (DOY) 37. The data were combined three times: (1) by software, (2) by the near-real-time system without decoding, and (3) by the near-real-time system with decoding. Different subsets of the data were processed for each case.

A. Software Combining

Figure 6 shows the results of software combining of data from 2228 to 2318 hours. The figure shows the measured SNRs for the two stations and for the combined symbols, and the theoretically achievable combined symbol SNR, which is the sum of the two input SNRs.

The data were combined in blocks of 5700 symbols, corresponding to the amount of data in one physical tape record. Independent estimates of input and output symbol SNRs and combiner weights were made for each block. The SNR estimates were then averaged over non-overlapping sets of 95 blocks, and each point in Fig. 6 corresponds to one of the 95-block averages, or 541,500 symbols. For all data shown, the average input SNRs are −0.37 dB for Parkes and 0.45 dB for Canberra, with a measurement accuracy estimated in Section II.D as ±0.02 dB. The two input SNRs sum to 3.07 dB ±0.02 dB (by summing the SNRs expressed as ratios, not decibels, and then converting to decibels). The average output

SNR is 2.84 dB ± 0.11 dB, or 0.23 dB ± 0.13 dB less than the sum of the input SNRs.

The theoretical loss in SNR is due to two sources: weight errors and quantization. From Vo (Ref. 3), we estimate the loss due to weight errors as 0.01 ± 0.01 dB. Loss in mean square to variance due to quantization was calculated for the actual signal conditions at each quantizer, and the effect on the combined symbols was calculated as a loss of 0.13 dB ± 0.01 dB. The total theoretical loss is thus 0.14 dB ± 0.02 dB. Our experimental results of 0.23 dB ± 0.13 dB are within the expected tolerance.

B. Near-Real-Time Combining

Portions of the data were combined on seven separate occasions using the near-real-time SSC system, at CTA-21. The combining rate was 60 ksymb/s, compared to the recorded data bit rate of 43.2 ksymb/s. This demonstrated that symbols can be combined at a rate faster than real time, i.e., without building up a backlog of data at the combiner.

Figure 7 shows the results of near-real-time combining conducted on June 6. The approximate data time span was from 2259 to 2311 hours on the data acquisition day, February 6. Unfortunately, the recorded data times are not accurately known during near-real-time combining, because there is no way to get the time tags from the tape recordings to the combiner system.

In the near-real-time combiner, the SNRs are estimated approximately every 3.2 s of playback time, or every 4.4 s of actual data time. Only 4096 symbols are used in each estimate. Each point shown in Fig. 7 is the average of five SNR measurements. Weights are computed from the SNRs, and smoothed using a ten-measurement time-constant low-pass filter. In Fig. 7, the average symbol SNRs are −0.29 dB for Parkes and 0.52 dB for Canberra, which sum to 3.14 dB. The measured output SNR is 2.90 dB, or 0.24 dB below the sum of the input SNRs. This measurement of loss is almost identical to the software combining loss of 0.23 dB, with an estimated tolerance of ±0.13 dB.

Figure 8 shows the results of combining the same data tapes (but not exactly the same data) at a different time, on May 30. This time, the input data streams were reversed with respect to the combiner channels, compared to the data of Fig. 7. Each point in Fig. 8 represents one measurement of SNR, instead of an average of five measurements, as in Fig. 7, because of the lesser amount of data. The average SNR measurement results are −0.21 dB for Parkes and 0.48 dB for Canberra, which sum to 3.16 dB and 3.02 dB for the output

symbols. The measured loss in symbol SNR is 0.14 dB, which is 0.10 dB less than in Fig. 7, and 0.09 dB less than the software result.

Near-real-time combining was done a total of seven times. For the last six times, the measured losses in average SNR were 0.23 dB ± 0.09 dB. The first time the data were combined, the measured loss was 0.42 dB, or 0.19 dB higher than the other times. At first, we thought this higher loss might be due to biases in measuring output SNR, and might be dependent on which symbol stream was input to which processor channel. The last six processing runs used each combination of inputs three times, with no significant differences. Thus, we are uncertain as to the cause of the higher loss the first time the data were processed, but attribute it to configuration or operations error.

In summary, data were successfully combined on seven separate occasions by the near-real-time combiner. Performance was excellent. Except for the first time the data were processed, the loss in SNR was 0.23 dB ± 0.09 dB. This is in close agreement with the software processing, which had an SNR loss of 0.24 dB, with estimated accuracy 0.13 dB.

C. Combining and Decoding in Near-Real Time

Data combined on May 30 simultaneously decoded. The output symbol stream from the combiner was input to an MCD at CTA 21. The combining rate was 60 ksymb/s, versus the recorded symbol rate of 43.2 ksymb/s. Since the MCD and its associated TPA could only accept standard Voyager symbol rates, the bit-rate input to the TPA was 29.9 kb/s, coded (rate 1/2 code).

The estimate of bit SNR made by the MCD was monitored as an indication of decoder performance. The MCD estimates are based on the rate of renormalization of a metric in the decoder. Since the data were coded with a rate 1/2 code, the bit SNR is theoretically twice the symbol SNR. The average estimate was 6.02 dB, with an unspecified tolerance, probably of several tenths of 1 dB. The combined symbol SNR measured by the combiner was 3.02 dB ± 0.13 dB. Despite the measurement tolerances, the measured bit SNR happened to be exactly 3 dB higher than the symbol SNR, as it should be for the rate 1/2 code. This indicated proper decoding.

This demonstration established the ability to combine and decode at a rate faster than real time with the present station equipment, adding only the combiner.

V. Conclusions

Symbol stream combining has been demonstrated to be a practical and low-loss method for antenna arraying of telemetry signals. Combining has been demonstrated at a rate faster than real time, with concurrent decoding, using standard DSN equipment plus the symbol stream combiner.

The main source of loss relative to ideal performance is quantization of the input symbol streams. Simulations by Pollara and Swanson (Ref. 4) indicate that losses in telemetry performance due to quantization are approximately 0.04 dB ± 0.02 dB for 4-bit quantization. More accurate quantizers would reduce this error by approximately a factor of four for each additional bit of quantization. The next significant loss source is error in the combiner weights, but this can be kept under 0.01 dB. This indicates that a system using 8-bit quantization could easily achieve performance within 0.05 dB of ideal, over a wide range of input signal and noise levels.

With coarse (4-bit) symbol quantization, the SNR of the combined symbols, in the sense of the ratio of mean-squared signal level to noise variance, is not a good measure of telemetry system performance. However, it can be accurately predicted and measured, so it is a good indication of proper operation of the SSC. Using software combining of 4-bit quantized symbols, this demonstration achieved a measured loss in symbol SNR of 0.23 dB ± 0.13 dB, which was within the expected tolerance of the calculated loss of 0.14 dB ± 0.02 dB.

For the maximum data rate at Voyager Neptune encounter, symbol stream combining requires two orders of magnitude less data than does baseband combining, in a near-real-time implementation. Performance of the SSC is at least as good as that of baseband combining. This work has been instrumental in developing the current implementation plan to use symbol stream combining for near-real-time combining of the Very Large Array (VLA), in New Mexico, with the Goldstone array, at Neptune encounter.

References

1. Hurd, W. J., Pollara, F., Russell, M. D., Siev, B., and Winter, P. U., "Intercontinental Antenna Arraying by Symbol Stream Combining at ICE Giacobini-Zinner Encounter," *TDA Progress Report 42-84*, pp. 220-228. Jet Propulsion Laboratory, Pasadena, California, Feb. 15, 1986.

2. Hurd, W. J., and Aguirre, S., "A Method to Dramatically Improve Subcarrier Tracking," *TDA Progress Report 42-86*, Jet Propulsion Laboratory, Pasadena, California, Aug. 15, 1986.

3. Vo, Q., "Signal-to-Noise Ratio and Combiner Weight Estimation for Symbol Stream Combining," *TDA Progress Report 42-76*, pp. 86-98. Jet Propulsion Laboratory, Pasadena, California, Feb. 15, 1984.

4. Pollara, F., and Swanson, L., "Effects of Quantization on Symbol Stream Combining in a Convolutionally Coded System," *TDA Progress Report 42-84*, pp. 82-87. Jet Propulsion Laboratory, Pasadena, California, Feb. 15, 1986.

5. Vilnrotter, V. A., and Rodemich, E. R., "Table Look-Up Estimation of SNR for Quantized Symbols." To be published in *TDA Progress Report 42-87*, Jet Propulsion Laboratory, Pasadena, California, Nov. 15, 1986.

6. Hurd, W. J., Reder, L. J., and Russell, M. D., "Symbol-Stream Combiner: Description and Demonstration Plans," *TDA Progress Report 42-78*, pp. 115-121. Jet Propulsion Laboratory, Pasadena, California, Aug. 15, 1984.
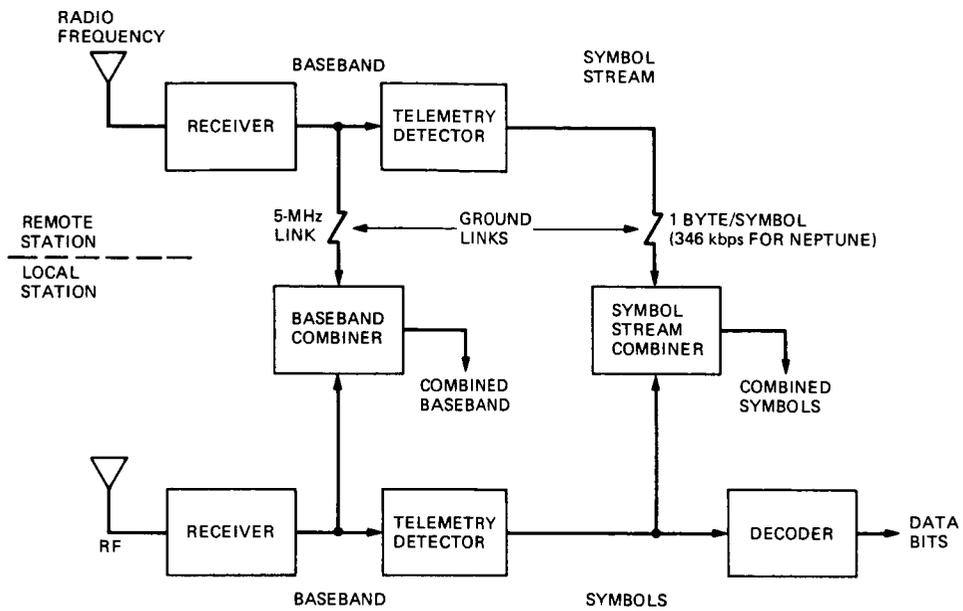
Fig. 1. Symbol stream combining versus baseband combining
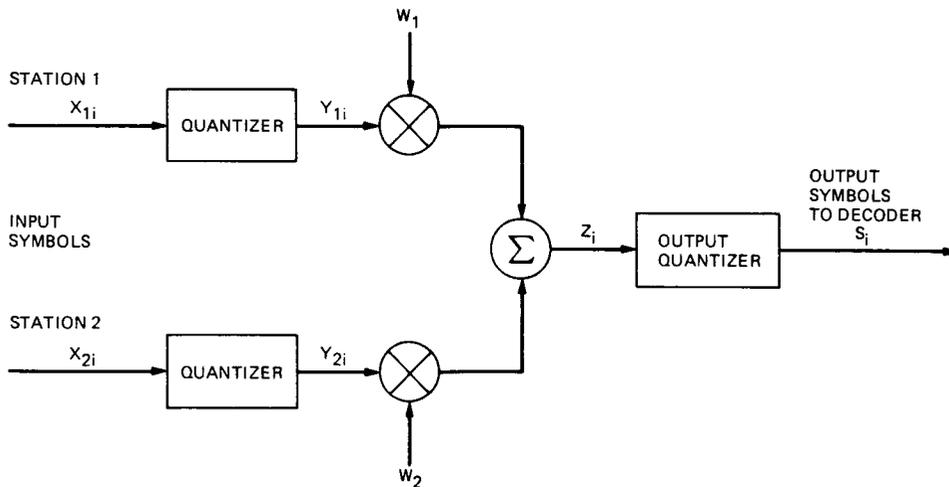
Fig. 2. Mathematics of symbol stream combining
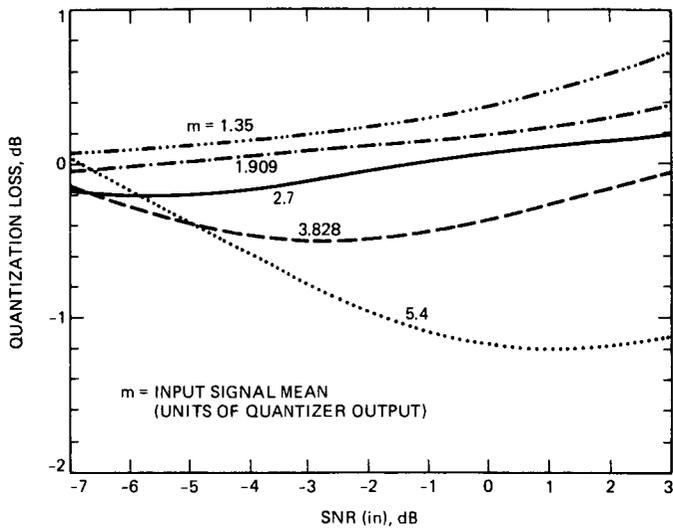
Fig. 3. Loss in SNR for four-bit quantizer
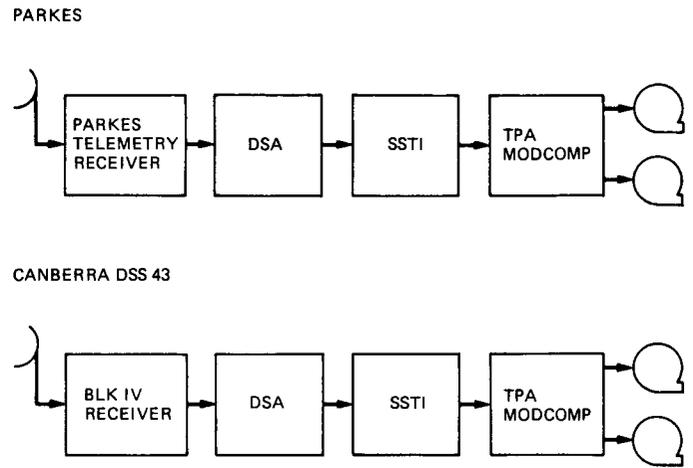
Fig. 4. Station configurations
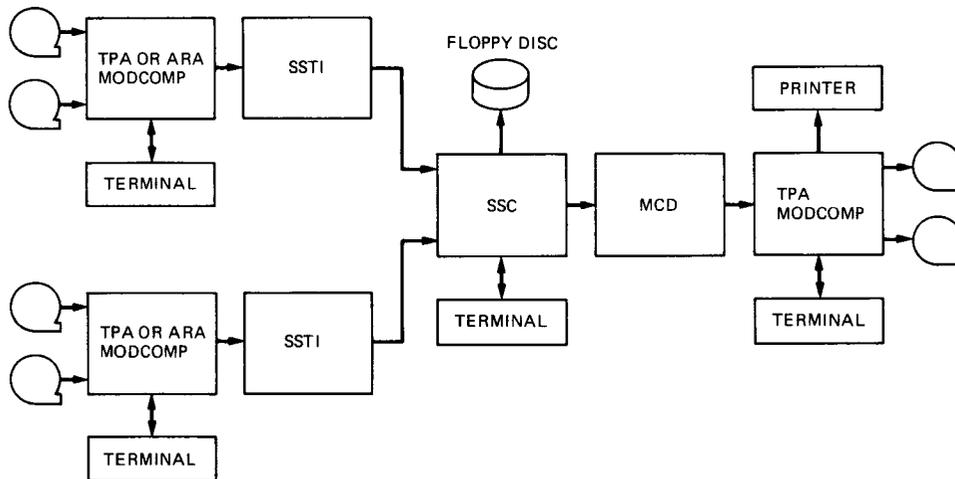
Fig. 5. Near-real-time combining and decoding configuration
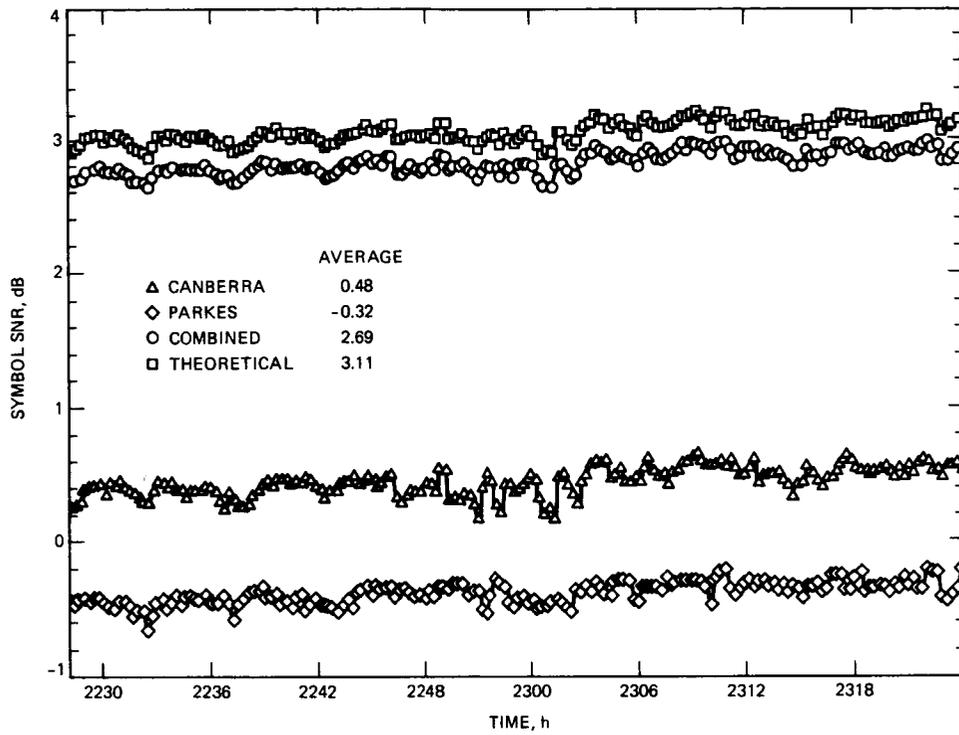
Fig. 6. Results of software combining
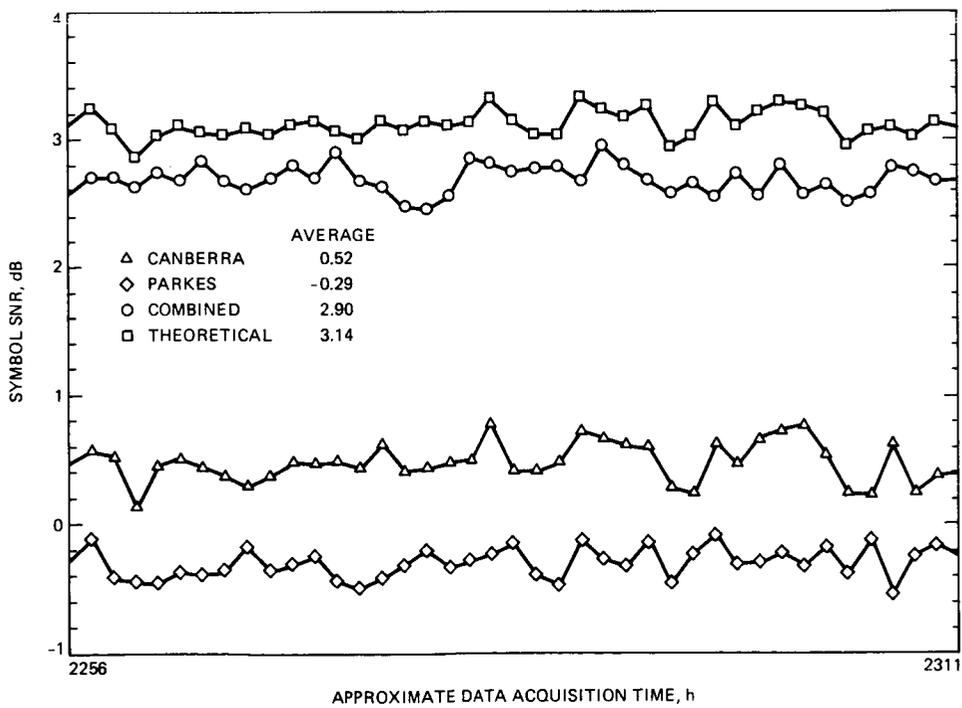
Fig. 7. Results of near-real-time combining on June 6

141
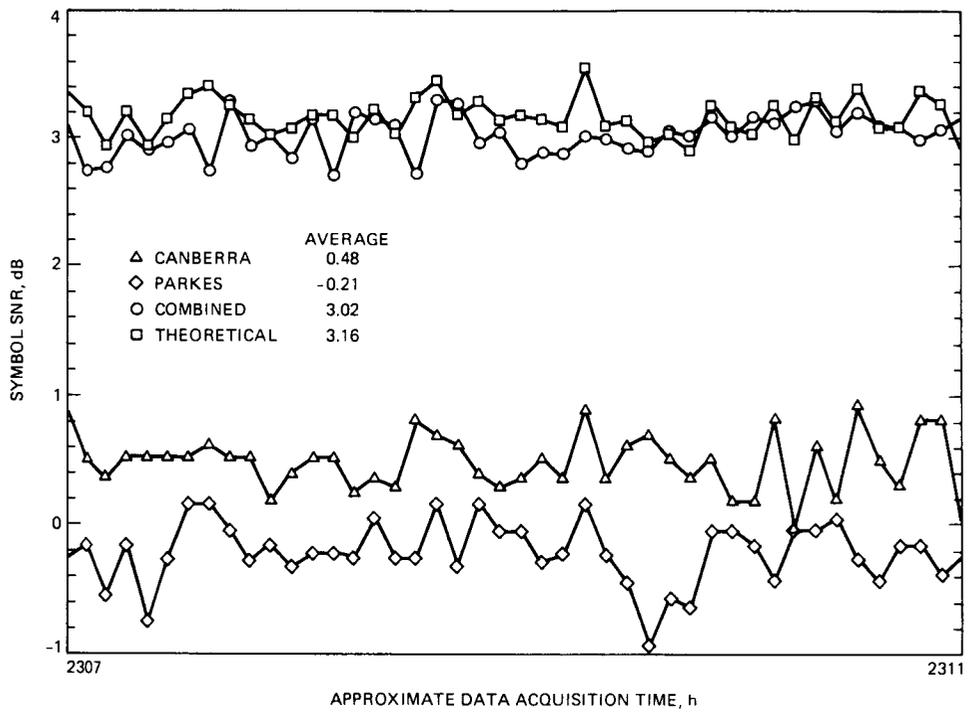
Fig. 8. Results of near-real-time combining on May 30

A VLSI Architecture for Performing Finite Field Arithmetic With Reduced Table Look-Up

I. S. Hsu and T. K. Truong
Communications Systems Research Section

I. S. Reed
Electrical Engineering Department
University of Southern California

A new table look-up method for finding the log and antilog of finite field elements has been developed by N. Glover. In his method, the log and antilog of a field element is found by the use of several smaller tables. The method is based on a use of the Chinese Remainder Theorem. The technique often results in a significant reduction in the memory requirements of the problem. A VLSI architecture is developed for a special case of this new algorithm to perform finite field arithmetic including multiplication, division, and the finding of an inverse element in the finite field.

I. Introduction

A codeword of a cyclic code is a sequence of symbols or characters (Ref. 1). These characters can be represented as the coefficients of a polynomial

$$C(x) = \sum_{i=0}^{m-1} c_i x^i$$

where c_i for $0 \leqslant i \leqslant m - 1$ is an element in a finite field $GF(q)$, and m is the length of the codeword. One such cyclic code is a Reed-Solomon (RS) code with symbols which lie in the Galois field $GF(2^n)$ of order 2^n. If β belongs to $GF(2^n)$, then

$$\beta = \sum_{i=0}^{n-1} a_i \alpha^i$$

where $a_i \in GF(2)$ and α is a root of a primitive irreducible polynomial with degree n over $GF(2)$. It is shown in Refs. 2-4 that the arithmetic used to encode and decode RS codes over $GF(2^n)$ requires the multiplication and division of field elements in $GF(2^n)$. The most straightforward method to perform multiplication and division of two field elements in $GF(2^n)$ is to use table look-up. The same method can also be used to find inverses in the finite field.

To illustrate this procedure, let two field elements be represented in binary. That is, let $x = x_0, x_1, \ldots, x_{n-1}$ and $y = y_0, y_1, \ldots, y_{n-1}$, where $x_i, y_i \in GF(2)$ for $0 \leqslant i \leqslant n - 1$. Next let a "log" table be used to find the exponents i and j in such a manner that $x = \alpha^i$ and $y = \alpha^j$. Binary addressing is used in the table to locate the logarithms i and j of x and y, respectively. After the addition $k = i + j$ mod $(2^n - 1)$ of these exponents, an antilog table is used to find the binary representation of α^k. The exponent k serves as the address of the field element

in the antilog table. For some applications $q = 2^n$ is large and the log and antilog tables may be so large that the consumption of silicon area in the VLSI implementation becomes prohibitive.

Recently, Glover (Ref. 5) developed a new algorithm to reduce the size of the table needed to find the log and antilog of field elements by using several smaller tables. The method is based primarily on a use of smaller look-up tables and the Chinese Remainder Theorem (Ref. 6). To make this new algorithm realizable, a new mapping method based on a special case of the technique in Ref. 5 is found in this article for converting an element in $GF(2^{2n})$ to its counterpart in $(GF(2^n)^2)$ and vice versa, where n is a positive integer. A VLSI architecture is also developed to realize this new algorithm. This VLSI architecture possesses the programmable capability of being able to perform operations such as multiplication, division or the finding of inverses in a finite field.

II. A Reduced Table Look-Up Method for Finding Logs and Antilogs of Elements in a Galois Field

In this section, one of the methods developed by Glover (Ref. 5) to find the log and antilog of finite field elements is described briefly. The details of this method are described in the appendix.

The Log Algorithm. Given $\alpha^m = a + \alpha b$, where α is a primitive element in $GF(2^{2n})$, a, b belong to $GF(2^n)$, and $0 \leq m \leq 2^{2n} - 1$ is the log of element α^m. The procedure to find $m = \log_\alpha (\alpha^m)$ is described as follows:

Step 1: Map α^m to $a + \alpha b$. The particular mapping technique is described in detail in the following section.

Step 2: Compute $x = a^2 + ab + b^2 \beta$ and $y = a/b$, where $\beta = \alpha^{2^{n}+1}$ is a primitive element in $GF(2^n)$.

Step 3: Use the $\log_\beta$ table to find $m_1 = \log_\beta(x)$ and the $\log_\gamma$ table to find $m_2 = \log_\gamma(y)$ for $a \neq 0$, $b \neq 0$, where $\gamma = \alpha^{2^{n-1}}$ is an element in $GF(2^n)$. For $a = 0$, choose $m_2 = 1$. For $b = 0$, choose $m_2 = 0$. Here $\log_\beta(x) = m \bmod (2^n - 1) = m_1$, and $\log_\gamma(y) = m \bmod (2^n + 1) = m_2$.

Step 4: By the Chinese Remainder Theorem (Ref. 6), $m = m_1 \cdot n_2 \cdot M_1 + m_2 \cdot n_1 \cdot M_2$. Here $n = n_1 \cdot n_2 = n_1 \cdot N_1 = n_2 \cdot N_2$, n_i relatively prime and M_i uniquely satisfies (modulo n_i) the congruence $N_i \cdot M_i = 1 \bmod n_i$ for $1 \leq i \leq 2$.

The Antilog Algorithm. Given m, recover $\alpha^m = a + \alpha b$ as follows:

Step 1: Compute $m_1 = m \bmod (2^n - 1)$ and $m_2 = m \bmod (2^n + 1)$.

Step 2: Use the antilog tables to find $\text{antilog}_\beta(m_1) = x = a^2 + ab + b^2\beta$, and $\text{antilog}_\gamma(m_2) = y = a/b$, for $m_2 \neq 0, 1$. For $m_2 = 1$, $a = 0$, and $b = \sqrt{x/\beta} = \text{antilog}_\beta((\log_\beta(x/\beta))/2)$. For $m_2 = 0$, $b = 0$, and $a = \sqrt{x} = \text{antilog}_\beta((\log_\beta(x))/2)$.

Step 3: For $m_2 \neq 0, 1$, use the equation $b = \text{antilog}_\beta((\log_\beta(z))/2)$, where $z = x/(y^2 + y + \beta)$. Then $a = b \cdot y$.

Step 4: Map $a + \alpha b$ to α^m. This inverse mapping is described in detail in the following section.

To illustrate the above procedures, two examples are given for the finite field $GF(2^8)$.

Example 1: Given $a + \alpha b \in GF(2^8)$, where $a = (0\ 1\ 1\ 0)$ and $b = (1\ 1\ 1\ 0) \in GF(2^4)$. Then, find m such that $\alpha^m = a + \alpha b$. By the log algorithm, $x = a^2 + ab + b^2\beta = (1\ 1\ 1\ 0)$ and $y = a/b = (1\ 1\ 1\ 1)$. Now use Tables A-1 and A-3 to find m_1 and m_2, respectively. The results are $m_1 = 7$ and $m_2 = 8$. For this example, $n_1 = 15$ and $n_2 = 17$. Thus, $M_1 = 8$ and $M_2 = 8$ are the smallest numbers such that $17M_1 \equiv 1 \bmod 15$ and $15M_2 \equiv 1 \bmod 17$, respectively. Hence, $n_2 M_1 = 136$ and $n_1 M_2 = 120$. By Eq. (10), Ref. 8, $m \equiv (136 \cdot m_1 + 120 \cdot m_2) \bmod (2^8 - 1) = 127$.

Example 2: Given $m = 127$, find $\alpha^{127} = a + \alpha b \in GF(2^8)$. Using the ANTILOG algorithm, $m_1 \equiv m \bmod (2^n - 1) = 7$ and $m_2 \equiv m \bmod (2^n + 1) = 8$. Then use Tables A-2 and A-4 to find x and y, respectively. The results are $x = (1\ 1\ 1\ 0)$ and $y = (1\ 1\ 1\ 1)$. Thus, $z = x/(y^2 + y + \beta) = (0\ 0\ 1\ 1)$. By Eq. (9), Ref. 8, $b = \text{antilog}_\beta((\log_\beta(z))/2) = \text{antilog}_\beta(7) = (1\ 1\ 1\ 0)$. Thus, $a = b \cdot y = (0\ 1\ 1\ 0)$. Therefore, $\alpha^{127} = (0\ 1\ 1\ 0) + \alpha(1\ 1\ 1\ 0)$.

III. A Method for Mapping Elements of $GF(2^{2n})$ Onto $GF((2^n)^2)$ and Vice Versa

To perform Step 4 of the Antilog Algorithm in Section II, a method is developed in this section for the required mapping. This is accomplished by first considering the mapping of an element α^m in $GF(2^{2n})$ to its counterpart $a + \alpha b$, where $a, b \in GF(2^n)$. This mapping procedure is best described by an example. The extension to other finite fields $GF(2^{2n})$ can be obtained in a similar fashion.

First, Theorem 1 in Ref. 5 is repeated here. A proof of this theorem is given here also in order to make the algorithm more understandable and self-contained.

Theorem 1 (Ref. 5): Let β be a primitive element in $GF(2^n)$ such that the polynomial $p(x) = x^2 + x + \beta$ is irreducible in this field. Also let $\alpha \in GF(2^{2n})$ for $\alpha \neq 0$ where $GF(2^{2n})$ is the quadratic extension field of $GF(2^n)$. If α is a root of $p(x)$, i.e., $p(\alpha) = 0$, and $2^n + 1$ a prime, then α is a primitive element in $GF(2^{2n})$.

Proof: It is shown in Ref. 7 (page 34) that if α is a root of $p(x)$, its conjugate $\bar{\alpha} = \alpha^{2n}$ is also a root of $p(x)$. Thus,

$$(x + \alpha) \cdot (x + \bar{\alpha}) = x^2 + (\alpha + \bar{\alpha})x + \alpha \cdot \bar{\alpha}$$

$$= x^2 + x + \beta$$

where $\alpha + \bar{\alpha} = 1$ and $\alpha \cdot \bar{\alpha} = \beta$. Hence,

$$\alpha^{2^n+1} = \beta \tag{1}$$

Now $2^{2n} - 1 = (2^n + 1) \cdot (2^n - 1)$ and $2^n + 1$ is a prime by hypothesis. Let γ be any number such that $\gamma | (2^{2n} - 1)$ and $\gamma \neq 2^n + 1$. Then $\gamma | (2^n - 1)$ so that by Eq. (1)

$$\alpha^{(2^{2n}-1)/\gamma} = (\alpha^{2^n+1})^{(2^n-1)/\gamma} = \beta^{(2^n-1)/\gamma} \tag{2}$$

Since β is primitive over $GF(2^n)$, $2^n - 1$ is the least integer such that $\beta^{(2^n-1)} = 1$. Hence $\beta^{(2^n-1)/\gamma} \neq 1$ unless $\gamma = 1$.

On the other hand, if $\gamma = 2^n + 1$, then

$$\alpha^{(2^{2n}-1)/\gamma} = \alpha^{(2^n+1)(2^n-1)/(2^n+1)} = \alpha^{2^n-1}$$

Since α is a root of $p(x) = x^2 + x + \beta$, by Eq. (1), one has

$$p(\alpha) = \alpha^2 + \alpha + \beta$$

$$= \alpha^2 + \alpha + \alpha^{2^n+1}$$

$$= 0$$

Thus, $\alpha^{2^n-1} = \alpha^{-1} + 1 \neq 1$ for otherwise $\alpha^{-1} = 0$, which is impossible since $\alpha \neq 0$.

Thus, by Eq. (2) and the above, if $\gamma | 2^{2n} - 1$, then $\alpha^{(2^{2n}-1)/\gamma} \neq 1$ unless $\gamma = 1$. Therefore, the order of α is $2^{2n} - 1$ and α is primitive in $GF(2^{2n})$.

Q.E.D.

To illustrate the consequences of the above theorem, let $n = 4$ and let α be the root of polynomial

$$p(x) = x^2 + x + \beta$$

where β is a primitive element in $GF(2^4)$. For this case by Eq. (1)

$$\beta = \alpha^{2^4+1} = \alpha^{17} \tag{3}$$

By the above theorem, since 17 is a prime, α is a primitive element in the extension field $GF(2^8)$.

Since α is the root of $p(x)$, $p(\alpha) = \alpha^2 + \alpha + \alpha^{17} = 0$. Hence, α satisfies the reduced equation

$$\alpha^{16} = \alpha + 1 \tag{4}$$

Now the root of the irreducible primitive polynomial $f(x)$ over $GF(2)$ which generates the finite field $GF(2^8)$ must also satisfy Eq. (3). Of the many irreducible polynomials which generate $GF(2^8)$, consider the special irreducible polynomial

$$f(x) = x^8 + x^6 + x^5 + x^3 + 1 \tag{5}$$

Let α be any root of $f(x)$ so that α satisfies

$$\alpha^8 + \alpha^6 + \alpha^5 + \alpha^3 + 1 = 0$$

or alternatively

$$\alpha^8 = \alpha^6 + \alpha^5 + \alpha^3 + 1 \tag{6}$$

It is now shown that α also satisfies Eq. (4). Squaring both sides of Eq. (6) one obtains,

$$\alpha^{16} = \alpha^{12} + \alpha^{10} + \alpha^6 + 1$$

$$= (\alpha^7 + \alpha^3 + \alpha^2 + \alpha + 1) + (\alpha^7 + \alpha^6 + \alpha^3 + \alpha^2 + 1) + \alpha^6 + 1$$

$$= \alpha + 1$$

Hence the irreducible polynomial $f(x)$ in Eq. (5) is so chosen that a root α of $f(x)$ also is a root of the quadratic generator polynomial $p(x) = x^2 + x + \beta$ of finite field $GF(2^8)$ over the smaller field $GF(2^4)$, where $\beta = \alpha^{17}$.

In order to simplify operations by table look-up, one would like to represent an element α^m in $GF(2^8)$ by $\alpha^m = a + \alpha b$, where $a, b \in GF(2^4)$, the smaller field. To find the

mapping which makes this representation possible, one must find the generator polynomial $g(x)$ of $GF(2^4)$ which β satisfies. By Eq. (3), $\beta = \alpha^{17}$, where β is a primitive element in $GF(2^4)$. Hence $g(x)$, the generator polynomial of field $GF(2^4)$ over $GF(2)$, must satisfy the equation

$$g(\beta) = g(\alpha^{17}) = 0$$

Try the irreducible polynomial

$$g(x) = x^4 + x^3 + 1$$

over $GF(2)$. Substituting $x = \alpha^{17}$ into the above $g(x)$ yields

$$g(\alpha^{17}) = (\alpha^{17})^4 + (\alpha^{17})^3 + 1 \qquad (7)$$

However, by Eq. (4),

$$\alpha^{16} = \alpha + 1$$

which implies

$$\alpha^{17} = \alpha^2 + \alpha$$

Substituting the above equation into Eq. (7) yields

$$g(\alpha^{17}) = (\alpha^2 + \alpha)^4 + (\alpha^2 + \alpha)^3 + 1$$

$$= (\alpha^8 + \alpha^4) + (\alpha^6 + \alpha^4 + \alpha^5 + \alpha^3) + 1$$

$$= 0$$

Hence $g(x) = x^4 + x^3 + 1$ is the correct generator polynomial of the finite field $GF(2^4)$.

Given an element α^m in $GF(2^8)$, α^m can be expressed in standard basis form as

$$\alpha^m = C_0\alpha^0 + C_1\alpha^1 + C_2\alpha^2 + C_3\alpha^3 + C_4\alpha^4$$

$$+ C_5\alpha^5 + C_6\alpha^6 + C_7\alpha^7 \qquad (8)$$

where $C_i \in GF(2)$ for $0 \leqslant i \leqslant 7$. α^m also can be represented as $\alpha^m = a + \alpha b$, where $a,b \in GF(2^4)$.

Now $\beta = a^{17}$ is the primitive element found above for $GF(2^4)$. Hence a,b in $GF(2^4)$ can be represented over $GF(2)$ in standard basis form as

$$a = a_0\beta^0 + a_1\beta^1 + a_2\beta^2 + a_3\beta^3 \qquad (9a)$$

and

$$b = b_0\beta^0 + b_1\beta^1 + b_2\beta^2 + b_3\beta^3 \qquad (9b)$$

where $a_i, b_i \in GF(2)$, for $0 \leqslant i \leqslant 3$.

Now substitute representations (8) and (9) into equation $\alpha^m = a + \alpha b$ to obtain

$$C_0\alpha^0 + C_1\alpha^1 + C_2\alpha^2 + C_3\alpha^3 + C_4\alpha^4 + C_5\alpha^5 + C_6\alpha^6 + C_7\alpha^7$$

$$= a_0\beta^0 + a_1\beta^1 + a_2\beta^2 + a_3\beta^3 + (b_0\beta^0 + b_1\beta^1 + b_2\beta^2$$

$$+ b_3\beta^3) \cdot \alpha$$

Since $\beta = \alpha^{17}$, the right side of the above equality can be represented in terms of powers of α as

$$= a_0 + a_1\alpha^{17} + a_2\alpha^{34} + a_3\alpha^{51} + b_0\alpha + b_1\alpha^{18} + b_2\alpha^{35}$$

$$+ b_3\alpha^{52}$$

$$= a_0 + (a_1 + b_0)\alpha + (a_1 + a_2 + b_1)\alpha^2 + (a_3 + b_1 + b_2)\alpha^3$$

$$+ (a_2 + a_3 + b_3)\alpha^4 + (a_3 + b_2 + b_3)\alpha^5$$

$$+ (a_3 + b_3)\alpha^6 + b_3\alpha^7$$

If the coefficients of the corresponding powers of α are equated and expressed in matrix form, one obtains

$$\mathbf{C} = \mathbf{M} \cdot \mathbf{A} \qquad (10)$$

where

$$\mathbf{C} = \begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \end{bmatrix}$$

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$A = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

The inverse matrix M^{-1} of M is obtained readily and is given by

$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{11}$$

Hence, one obtains

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \\ C_6 \\ C_7 \end{bmatrix}$$

or $A = M^{-1}C$. Equation (11) is the mapping of elements in $GF(2^8)$ to the corresponding elements in $GF((2^4)^2)$. It is the inverse mapping of the mapping from $GF((2^4)^2)$ to $GF(2^8)$, given in Eq. (10). Equation (11) illustrates for $n = 4$, the mapping needed in Step 4 of the Antilog Algorithm in Section II.

IV. VLSI Architecture for Performing Multiplication, Division and Finding the Inverse of Finite Field Elements

In this section, a VLSI architecture is developed to perform multiplication, division and the finding of an inverse element in the finite field $GF(2^8)$ using the new algorithm described in the last two sections. This chip is designed to be programmable so that multiplication, division or the finding of an inverse in $GF(2^8)$ can be performed.

Figure 1 depicts the overall diagram of this chip. In Fig. 1, IN_1 and IN_2 are the two inputs to this chip. They are field elements in $GF(2^8)$. If the operation is to find the inverse element, IN_1 equals zero and IN_2 is the element to be inverted. Here it is assumed that these field elements are represented in the standard basis. That is,

$$IN_i = \sum_{j=0}^{7} a_j \alpha^j \qquad (i = 1, 2)$$

Input C_1 is the control signal used to control which operation is to be performed. If C_1 equals one, multiplication is performed. If C_1 equals zero, then either division or the

inverse operation is performed, according to whether IN_1 is zero or not. If IN_1 equals zero, then the inverse operation is performed, otherwise the division operation is carried out.

OUTPUT is the data output pin for the calculated result. Figure 2 shows the block diagram of this chip. As shown in the figure, the input field elements α^m and α^n are first converted to their corresponding elements $a + \alpha b$ and $a' + \alpha b'$ in $GF((2^4)^2)$. The needed mapping matrix is found in Eq. (11) of Section III.

Next, the values of x, y, x' and y' shown in Fig. 2 are calculated by the following equations:

$$x = a^2 + ab + b^2$$

$$y = a \cdot b^{-1}$$

$$x' = a'^2 + a'b' + b'^2$$

$$y' = a'b'^{-1}$$

The arithmetic operations such as the square and the inverse, needed above, are performed also by the table look-up method since they can be performed first in the much smaller field of $GF(2^4)$. The look-up table for $GF(2^4)$ involves only a small area in silicon and as a consequence is easier to realize in VLSI.

After the values of x, y, x', y' are obtained, their corresponding log values m_1, m_2, m_1', m_2' are obtained through the table look-up method as well. Again, this is readily realizable since these values are in the smaller field $GF(2^4)$. The log values of α^m and α^n can be obtained by equations

$$m = m_1 \cdot n_2 \cdot M_1 + m_2 \cdot n_1 \cdot M_2$$

$$n = m_1' \cdot n_2' \cdot M_1' + m_2' \cdot n_1' \cdot M_2'$$

respectively.

In Fig. 2, either $m + n$ or $m - n$ is performed in accordance with the control signal C_1. If C_1 equals one, addition is performed as needed in the multiplication operation. Otherwise a subtraction is performed as needed for division or the inverse operations. If the operation is to find the inverse element, then one of the inputs is zero, and its corresponding log value is also zero. Therefore, the operation $m - n$ corresponds to a negation of the logarithm of the nonzero input, i.e., $-n$ is obtained. This is the logarithm value of the inverse element.

The number p, in Fig. 2, results from either $(m + n)$ mod 255 or $(m - n)$ mod 255. These results are obtained through the modulo 255 circuit. The result of this computation is then fed to two modulus circuits to obtain the results modulus m_1, and m_2. m_1 and m_2 are the values of p modulo $2^4 - 1$ and $2^4 + 1$, respectively. Two antilog tables are then used to find the values x and y as described in Section II, Step 2 of the Log Algorithm. Next, $z = x/(y^2 + y + \beta)$ is calculated and a $\log_\beta$ table is used to find the logarithm of z. This value is then divided by two to obtain the value "A" as shown in Fig. 2. At this stage, element y is delayed for a certain number of clock cycles for the purpose of synchronization. An antilog table is then used to obtain element b. If b is combined with y, obtained previously, then element a can be obtained by performing the following operation:

$$a = by$$

Finally, the inverse mapping circuit described in the previous section is used to obtain the corresponding element of $a + \alpha b$, i.e., α^p, in $GF(2^8)$.

Because all operations performed in this chip are in the smaller finite field $GF(2^4)$, operations in $GF(2^4)$ are performed by the table look-up method occupying nearly half the silicon area as one would expect for $GF(2^8)$. It is obvious that the tables in $GF(2^4)$ can be reduced to smaller tables in $GF(2^2)$ by the same technique. The increased overhead associated with this reduction appears to be larger than the benefits that this further reduction might obtain for this case.

V. Conclusion

A VLSI architecture for performing finite field arithmetic is described in this article. A chip is designed to be programmable so that multiplication, division or the finding of an inverse in $GF(2^8)$ can be performed. The algorithm used is due to Glover (Ref. 5), where a new table look-up method for finding the log and antilog of finite field elements is developed. In that method, the log and antilog of a finite field element are found by the use of several smaller tables. The method is based on a use of the Chinese Remainder Theorem. The technique often results in a significant reduction in the memory requirements of the problem.

A method for mapping elements of $GF(2^{2n})$ onto $GF((2^n)^2)$ and vice versa is also developed in this article. An example of mapping elements of $GF(2^8)$ onto $GF((2^4)^2)$ is given for the purpose of illustration. The extension to other finite fields $GF(2^{2n})$ can be obtained in a similar fashion.

References

1. MacWilliams, F. J., and Sloane, N. J. A., *The Theory of Error-Correcting Codes*, North-Holland Publishing Company, NY, 1981.

2. Berlekamp, E. R., *Algebraic Coding Theory*, McGraw-Hill, NY, 1958.

3. Reed, I. S., Truong, T. K., and Miller, R. L., "Simplified Algorithm for Correcting Both Errors and Erasures of Reed-Solomon Codes," *Proc. IEE*, Vol. 126, No. 10, Oct. 1979.

4. Winograd, S., "On Computing the Discrete Fourier Transform," *Proc. Natl. Acad. Scie.*, USA, Vol. 73, pp. 1005-1006, 1976.

5. Glover, N., *Practical Error Correction Design for Engineers*, Data Systems Technology Corp., Broomfield, CO, 1982.

6. Ore, O., *Number Theory and Its History*, McGraw-Hill, NY, 1948.

7. Lin, S., and Costello, D. J., Jr., *Error Control Coding*, Prentice-Hall, NJ, p. 34, 1983.

8. Glover, N., Reed, I. S., Truong, T. K., and Huang, J. T., "Methods to Reduce the Table Look-Up Requirements for Finding Logs or Antilogs of Elements in a Power Galois Field," submitted to *IEEE Trans. on Computers*.
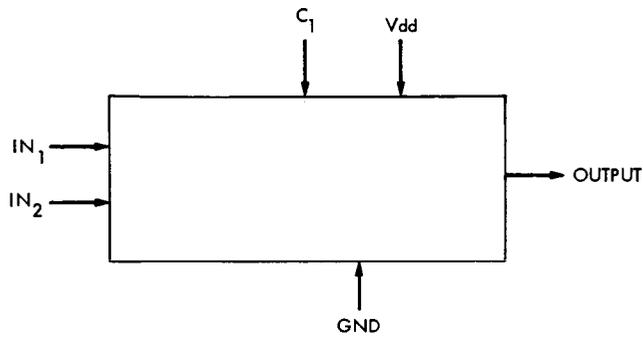
Fig. 1. Overall diagram of the VLSI chip for performing multi-plication/division or finding inverse in a finite field
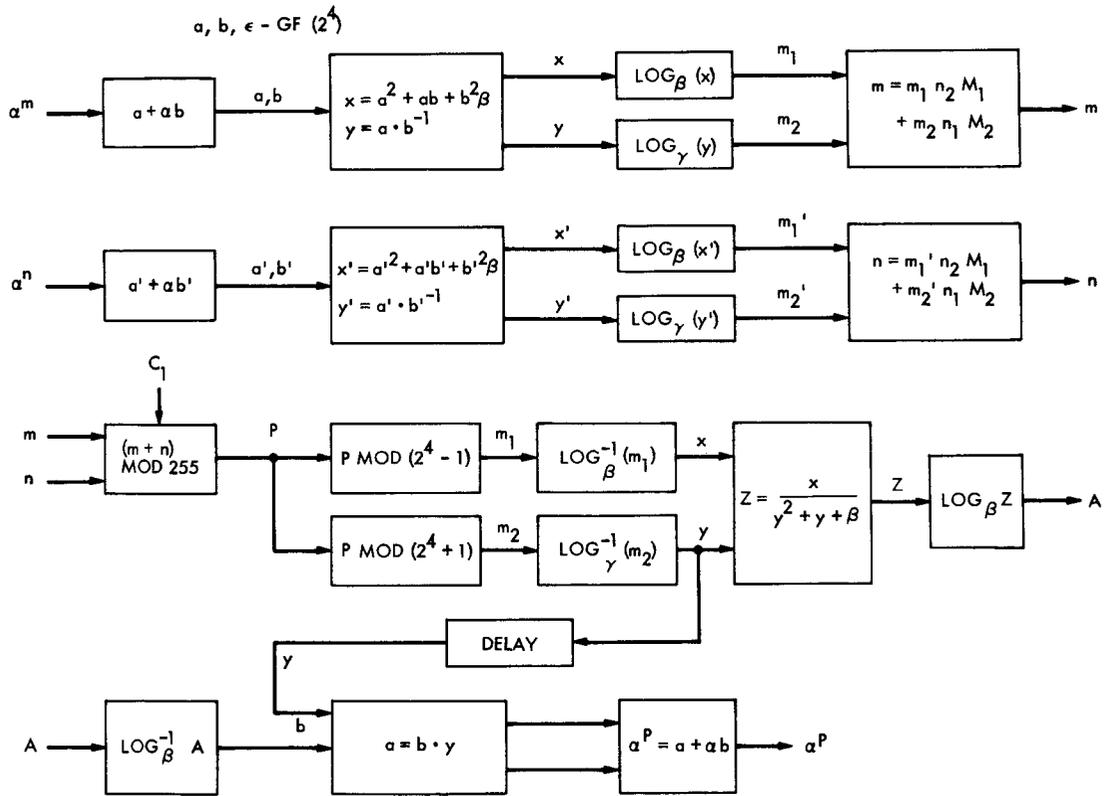
Fig. 2. Block diagram of the VLSI chip for performing multiplication/division or inverse of finite field elements

Appendix

Algorithm for Finding Log and Antilog of Finite Field Elements

In this appendix, the algorithm used to find the log and antilog of finite field elements is described in detail. A special case of Theorem A-1 and A-2, given next, was originally found by Glover (see Ref. 5).

Definition A-1: For $\alpha \in GF(2^{2n})$ and $a + \alpha b \in GF(2^{2n})$, where $a, b \in GF(2^n)$, define the norm of $a + \alpha b$ to be

$$\|a + \alpha b\| = (a + \alpha b) \cdot \overline{(a + \alpha b)}$$

where the bar denotes element conjugation in $GF(2^{2n})$.

Theorem A-1: Let β be a primitive element in $GF(2^n)$ such that the quadratic polynomial $x^2 + x + \beta$ is irreducible over $GF(2^n)$. Suppose also that $2^n + 1$ is a prime integer. Next let α be the root of this polynomial in the extension field $GF(2^{2n}) = \{a + \alpha b | a, b \in GF(2^n)\}$ of $GF(2^n)$. Suppose $\alpha^m = a + \alpha b \in GF(2^{2n})$. Then $m_1 = \log_\beta (a^2 + ab + b^2\beta) \equiv m \bmod (2^n - 1)$, i.e., m_1 is least integer such that $\beta^{m_1} = \|\alpha^{m_1}\| = \|a + \alpha b\| = a^2 + ab + b^2\beta$.

Proof: Since $x^2 + x + \beta$ is irreducible over $GF(2^n)$, it has roots α and

$$\overline{\alpha} = \alpha^{2^n} \tag{A-1}$$

in the extension field $GF(2^{2n})$. By Theorem 1, α is primitive in $GF(2^{2n})$. By definition A-1 and Theorem 1, one has the following:

$$\|a + \alpha b\| = (a + \alpha b) \cdot \overline{(a + \alpha b)} = (a + \alpha b) \cdot (a + \overline{\alpha} b)$$

$$= a^2 + ab + b^2\beta \tag{A-2}$$

If $c + \alpha d$ is any other element in $GF(2^{2n})$ and $c, d \in GF(2^n)$, then

$$\overline{(a + \alpha b) \cdot (c + \alpha d)} = (a + \alpha b)^{2^n} \cdot (c + \alpha d)^{2^n}$$

$$= \overline{(a + \alpha b)} \cdot \overline{(c + \alpha d)}$$

Thus, by the definition of the norm, one has

$$\|(a + \alpha b)(c + \alpha d)\| = (a + \alpha b)(c + \alpha d)\overline{(a + \alpha b)}\,\overline{(c + \alpha d)}$$

$$= \|a + \alpha b\| \cdot \|c + \alpha d\| \tag{A-3}$$

Observe next by Eq. (A-1) that $\|\alpha\| = \alpha \cdot \overline{\alpha} = \beta$ so that the theorem is true for $m = 1$.

For purpose of induction assume that

$$\|\alpha^k\| = \beta^k \tag{A-4}$$

for all k such that $1 \leqslant k \leqslant m$. Then, by (A-3), for $k = m + 1$, $\|\alpha^{m+1}\| = \|\alpha^m\| \cdot \|\alpha\| = \beta^{m+1}$. Hence the induction is complete and (A-4) is true for all k.

Now represent α^m by $a + \alpha b$ for some $a, b \in GF(2^n)$. Then, by (A-2) and (A-4), $\|\alpha^m\| = \beta^m$. Since β has order $2^n - 1$, the theorem must be true. Q.E.D.

By Theorem A-1 one can construct a $\log_\beta$ table of $2^n - 1$ elements by storing the value $m_1 \equiv m \bmod (2^n - 1)$, where $0 \leqslant m_1 < 2^n - 1$, at location $a^2 + ab + b^2\beta$ such that $\alpha^m = a + \alpha b$. Then, with a, b known, one can find m_1 using the $\log_\beta$ table. Similarly, the antilog_β table is constructed by storing the binary representation of $a^2 + ab + b^2\beta$ at location m_1 such that $\alpha^m = a + \alpha b$ and $\text{antilog}_\beta (m_1) = a^2 + ab + b^2\beta = x$.

Next, the construction of tables of $2^n + 1$ elements is shown.

Theorem A-2: Let $\gamma = \alpha^{2^n - 1} \in GF(2^{2n})$, where α is a primitive element of $GF(2^{2n})$. Suppose $\alpha^m = a + \alpha b \in GF(2^{2n})$ for some $a, b \in GF(2^n)$. Then,

$$m_2 = \log_\gamma \frac{(a + \overline{\alpha} b)}{(a + \alpha b)} \equiv m \bmod (2^n + 1) \tag{A-5}$$

i.e., m_2 is least integer such that $\gamma^{m_2} = \overline{\alpha^m}/\alpha^m$.

Proof: Since α is a primitive element in $GF(2^{2n})$ and $\gamma = \alpha^{2^{n-1}}$, the order of γ is $2^n + 1$. By the definition of the norm, one has

$$\|\alpha\| = \alpha \cdot \overline{\alpha} = \alpha \cdot \alpha^{2^n} = \gamma \cdot \alpha^2 \tag{A-6}$$

For purposes of induction assume that

$$\|\alpha^k\| = \gamma^k \alpha^{2k} \tag{A-7}$$

for $1 \leqslant k \leqslant m$. Then, by (A-3), $\|\alpha^{m+1}\| = \|\alpha^m\| \cdot \|\alpha\| = (\gamma^m \cdot \alpha^{2m})(\gamma\alpha^2) = \gamma^{m+1} \cdot \alpha^{2(m+1)}$. Hence the induction is complete and (A-7) is true for all k.

Representing α'^m by $a + \alpha b$ for some $a, b \in GF(2^n)$, it follows from (A-6) that

$$\|a + \alpha b\| = \gamma^m \cdot (a + \alpha b)^2 \qquad \text{(A-8)}$$

Multiplying both sides of (A-8) by $(a + \overline{\alpha} b)^2$ yields $\|a + \alpha b\| \cdot (a + \overline{\alpha} b)^2 = \gamma^m (a + \alpha b)^2 (a + \overline{\alpha} b)^2 = \gamma^m \|a + \alpha b\|^2$.

Therefore, from the definition of the norm,

$$\gamma^m = \frac{(a + \overline{\alpha} b)^2}{\|a + \alpha b\|} = \frac{a + \overline{\alpha} b}{a + \alpha b} \qquad \text{(A-9)}$$

Since the order of γ is $2^n + 1$, the theorem must follow. Q.E.D.

Using the results of Theorem A-2, let

$$f(a/b) = \gamma^m = ((a/b) + \overline{\alpha})/((a/b) + \alpha)$$

$$= (a + \overline{\alpha} b)/(a + \alpha b) \qquad \text{(A-10)}$$

To construct the $\log_\gamma$ table, notice that when $a = 0, f(a/b) = \gamma^m = \alpha^{2^n-1} = \gamma$ and $m = 1$. For $m_2 \equiv m \bmod (2^n + 1)$, one has $m_2 = 1$ when $a = 0$. When $b = 0, f(a/b) = (a + 0)/(a + 0) = 1$. Thus, $m = 0$ and $m_2 = 0$. The remaining part of the $\log_\gamma$ table can then be constructed by storing the values $m_2 \equiv m \bmod (2^n - 1)$ at locations "a/b" for $\alpha'^m = a + \alpha b$, where $2 \leqslant m_2 \leqslant 2^n$. The antilog$_\gamma$ table is constructed by storing the binary representation of $a/b \in \{\beta^1, \beta^2, \ldots, \beta^{2^n-1}\}$ at the corresponding locations, $i = m_2$, for $2 \leqslant i \leqslant 2^n$. Thus,

$$\text{Antilog}_\gamma(m_2) = a/b = y \qquad \text{(A-11)}$$

From (A-10) and (A-11) the two simultaneous equations needed to solve for a and b in the expression $\alpha'^m = a + \alpha b$ are given as follows:

Let

$$a^2 + ab + b^2 \beta = x \qquad \text{(A-12a)}$$

and

$$\frac{a}{b} = y \qquad \text{(A-12b)}$$

Relations (A-12) yield the following solution:

$$b = \sqrt{\frac{x}{y^2 + y + \beta}}$$

and

$$a = b \cdot y$$

For $b \in GF(2^n)$ it is verified readily that

$$b = \text{antilog}_\beta \left(\frac{\log_\beta z}{2} \right)$$

where $z = x/(y^2 + y + \beta)$.

It is desired now to find the logarithm with base α of $\alpha'^m = a + \alpha b \in GF(2^{2n})$, where $a, b \in GF(2^n)$, and $\alpha \in GF(2^{2n})$ is primitive. This can be found from the powers m_1 and m_2, by using the tables of the $2^n - 1$ powers of β and the $2^n + 1$ powers of γ, respectively, as follows:

Let

$$2^{2n} - 1 = (2^n + 1) \cdot (2^n - 1) = n_1 \cdot n_2 = n_1 \cdot N_1 = n_2 \cdot N_2$$

where $N_1 = n_2$ and $N_2 = n_1$. Then, by the Chinese Remainder Theorem, $m = \log_\alpha(\alpha'^m)$ is given by

$$m = m_1 \cdot n_2 \cdot M_1 + m_2 \cdot n_1 \cdot M_2 \bmod (2^{2n} - 1)$$

where M_1 and M_2 are the smallest integers such that $n_2 M_1 \equiv 1 \bmod n_1$ and $n_1 M_2 \equiv 1 \bmod n_2$, respectively.

Table A-1. Log_β table for finding $m_1 \equiv m \bmod 2^n - 1$ from known $\alpha^m = a + \alpha b$ where $2^n - 1 = 15$

Location $x = a^2 + ab + b^2\beta$				Content $m_1 \equiv \log_\beta x \bmod 15$
β_3	β_2	β_1	β_0	
0	0	0	1	3
0	0	1	0	2
0	0	1	1	14
0	1	0	0	1
0	1	0	1	10
0	1	1	0	13
0	1	1	1	8
1	0	0	0	0
1	0	0	1	4
1	0	1	0	9
1	0	1	1	11
1	1	0	0	12
1	1	0	1	5
1	1	1	0	7
1	1	1	1	6

Table A-2. Log_β table for finding $x = a^2 + ab + b^2\beta$ from known $m_1 = m \bmod(2^n - 1)$ where $2^n - 1 = 15$

Location m_1	Content $x = \mathrm{antilog}_\beta(m_1)$			
	β_3	β_2	β_1	β_0
0	0	0	0	0
1	0	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	0	0	1
5	1	1	0	1
6	1	1	1	1
7	1	1	1	0
8	0	1	1	1
9	1	0	1	0
10	0	1	0	1
11	1	0	1	1
12	1	1	0	0
13	0	1	1	0
14	0	0	1	1
15	1	0	0	0

Table A-3. Log_γ table for finding $m_2 \equiv m \bmod (2^n - 1)$ where $2^n + 1 = 17$ from known $\alpha^m = a + \alpha b$

Location				Content
\multicolumn{4}{c|}{$y = a/b$}	$m_2 = \log_\gamma (y) \bmod 17$
β_3	β_2	β_1	β_0	
0	0	0	1	14
0	0	1	0	12
0	0	1	1	6
0	1	0	0	2
0	1	0	1	10
0	1	1	0	4
0	1	1	1	9
1	0	0	0	16
1	0	0	1	3
1	0	1	0	5
1	0	1	1	11
1	1	0	0	15
1	1	0	1	7
1	1	1	0	13
1	1	1	1	8

Table A-4. Antilog_γ table for finding $y = a/b$ from known $m_2 \equiv m \bmod (2^n + 1)$ where $2^n + 1 = 17$

Location	Content			
m_2	$y = \text{antilog}_\gamma (m_2)$
	β_3	β_2	β_1	β_0
2	0	1	0	0
3	1	0	0	1
4	0	1	1	0
5	1	0	1	0
6	0	0	1	1
7	1	1	0	1
8	1	1	1	1
9	0	1	1	1
10	0	1	0	1
11	1	0	1	1
12	0	0	1	0
13	1	1	1	0
14	0	0	0	1
15	1	1	0	0
16	1	0	0	0

A New VLSI Complex Integer Multiplier Which Uses a Quadratic-Polynomial Residue System With Fermat Numbers

T. K. Truong and I. S. Hsu
Communicatioins Systems Research Section

J. J. Chang
Spacecraft Data Systems Section

H. C. Shyu and I. S. Reed
Department of Electrical Engineering
University of Southern California

A quadratic-polynomial Fermat residue number system (QFNS) has been used to compute complex integer multiplications. The advantage of such a QFNS is that a complex integer multiplication requires only two integer multiplications. In this article, a new type Fermat number multiplier is developed which eliminates the initialization condition of the previous method. It is shown that the new complex multiplier can be implemented on a single VLSI chip. Such a chip is designed and fabricated in CMOS-pw technology.

I. Introduction

The era of very large scale integrated (VLSI) circuits has arrived. VLSI systems have the characteristic of being compact, high speed and of low power consumption. Therefore, a large system can be integrated into a VLSI chip. Many systems which were realized in discrete components can be improved dramatically by taking advantage of VLSI technology.

Since the available area on a chip is limited, a residue number system can be introduced in order to reduce the computing complexity. The ring of integers modulo the Fermat number $F_n = 2^{2^n} + 1$ has some special simplifying characteristics for residue number systems. Recently, Leibowitz (Ref. 2) developed a binary arithmetic for implementing the Fermat number transform (FNT). In this development a special representation of binary numbers, the diminished-1

representation, was introduced. Arithmetic operations using this representation were developed also in Ref. 2. The FNT has been widely discussed in many papers (Refs. 3–5). Recently, the authors (Ref. 6) developed a VLSI architecture for a modified Leibowitz multiplier of integers modulo a Fermat number. This bit-modulo multiplier uses only addition and cyclic shifts. With this architecture a single chip integer multiplier was designed, fabricated, and proved to work well.

Recently, the authors (Ref. 1) also developed an algorithm to compute the DFT using the residue Fermat number systems. In this algorithm, a complex multiplier was developed which used the direct sum of two copies of the residue ring of integers modulo F_n. The advantage of this approach is that the operations for computing the complex multiplier need only two integer multiplications in Z_{F_n}, the residue ring of integers modulo F_n. Hence the number of multiplications required for

computing a systolic array of the DFT can be reduced substantially over previous approaches. The basic unit of such a complex multiplier is the integer multiplier over Z_{F_n}. In this article, the integer multiplier in Ref. 6 is modified. This new multiplier uses the diminished-1 representation for both input numbers. The second improvement is that the new design does not require the computation of an initial value. Hence considerable computation time and hardware can be saved. Thus, the new integer multiplier is easier to connect to a quadratic-polynomial residue Fermat number system for computing complex multiplications. As a consequence the new complex integer multiplier unit is readily implemented on a single VLSI chip.

II. The Integer Modulo F_n Multiplier

Integer multiplication with small dynamic range is often implemented by look-up tables. When the dynamic range is large, however, this method is undesirable (Ref. 13). Hence an arithmetic algorithmic solution for implementing the multiplier is needed. In Leibowitz's paper (Refs. 1 and 2), general multiplication has the form

$$(A \cdot B - 1) = (A - 1) \cdot (B - 1) + (A + B - 1) - 1 \qquad (1)$$

in the diminished-1 representation.

In the algorithm developed in Ref. 2 the binary multiplication of $(A - 1) \cdot (B - 1)$ was computed first. Then the term $(A + B - 1)$ was added to the result in the diminished-1 representation. In Ref. 6, a new multiplier was designed which uses the diminished-1 representation of numbers. The product of two integers, A and B, was obtained as

$$(A \cdot B - 1) = (A - 1) \cdot B + \bar{D} + 1 \qquad (2)$$

where $\bar{D}$ is an initial value calculated from the number of ones in B, and where $(x - 1)$ denotes the diminished-1 representation of the number x. The main disadvantage of the multiplier in Ref. 6 is that the initial value $\bar{D}$ must be calculated before the process starts. The use of different representation of numbers, i.e., the diminished-1 representation of A along with the ordinary binary representation of B, leads to some confusion.

A new modulo F_n multiplier is now derived. Let $F_n = 2^n + 1$ be a Fermat number where n is a positive integer. Also let $+$ and $\cdot$ be the symbols for addition and multiplication and $\oplus$ and $*$ be the symbols for the diminished-1 addition and multiplication. Also let $\boxed{\Sigma}$ denote the diminished-1 summation (i.e., consecutive diminished-1 additions). Then the following theorem holds.

Theorem 1: Let A and B be two elements of the finite residue ring of integers modulo F_n, i.e., Z_{F_n}. Then the diminished-1 multiplication can be calculated within Z_{F_n} as follows:

$$(A - 1) * (B - 1) = (A \cdot B - 1) = \sum_{i=0}^{2^n} (b_i 2^i A - 1) + (A - 1)$$

$$\qquad (3)$$

where

$$(B - 1) = \sum_{i=0}^{2^n} b_i 2^i, \; b_i = 0 \text{ or } 1$$

is the binary representation of $B - 1$.

Proof: From Ref. 2, it is known that

$$(A \cdot B - 1) = (A - 1) \cdot (B - 1) + (B - 1) + (A - 1)$$

But

$$B - 1 = \sum_{i=0}^{2^n} b_i 2^i$$

Thus,

$$(A \cdot B - 1) = \sum_{i=0}^{2^n} (b_i 2^i A - 1) - \sum_{i=0}^{2^n} b_i 2^i + \sum_{i=0}^{2^n} 1$$

$$+ (A - 1) + \sum_{i=0}^{2^n} b_i 2^i$$

$$= \sum_{i=0}^{2^n} (b_i 2^i A - 1) + 2^n + 1 + (A - 1)$$

Hence,

$$(A \cdot B - 1) = \sum_{i=0}^{2^n} b_i 2^i (A - 1) + (A - 1) \qquad (4)$$

Note that the diminished-1 addition can be performed by an adder of the type used in Ref. 6.

The recursive architecture for computing developed in Ref. 6 can be modified to compute Eq. (4). To see this, let the initial value be $C_0 = A - 1$. Then the multiplication algorithm in Eq. (4) can be put into the following recursive form:

$$C_{i+1} = C_i + (b_i 2^i A - 1) + 1 = C_i \oplus (b_i 2^i A - 1) \quad (5a)$$

if one successively computes C_{i+1} in Eq. (5a) for $0 \leqslant i \leqslant 2^n$, the required result is obtained as follows:

$$C = A \cdot B - 1 = C_{2^n+1} = C_{2^n} \oplus (b_{2^n} 2^{2^n} \cdot A - 1) \quad (5b)$$

In Eq. (5a), one observes that if $b_i = 1$, then $C_{i+1} = C_i \oplus (2^i A - 1)$ and if $b_i = 0$, then $C_{i+1} = C_i - 1 + 1 = C_i$. In other words, $(2^i \cdot A - 1)$ is added into C_i for $b_i = 1$ and no operation is needed for $b_i = 0$.

This new algorithm does not need a calculation of the initial value and to transform $B - 1$ to B. Now let the new algorithm be illustrated by an example for $F_2 = 2^4 + 1$. The same structure clearly extends to more general multiply algorithms over Z_{F_n}.

Consider an example in F_2. The elements in $GF(2^4 + 1)$ with their decimal equivalents in a normal binary representation along with their values in the diminished-1 representation are shown in Table 1.

Example 1: Let $A - 1 = 0\ 1\ 0\ 1\ 0$ and $B - 1 = 0\ 0\ 1\ 0\ 1$. Compute $C = (A\ B - 1) = 0\ 1\ 0\ 1\ 0 \times 0\ 0\ 1\ 0\ 1$ modulo $2^4 + 1$.

To compute C, let $C_0 = A - 1 = 0\ 1\ 0\ 1\ 0$ and $B - 1 = 0\ 0\ 1\ 0\ 1 = b_4 b_3 b_2 b_1 b_0$. The sequence of computation for $0\ 1\ 0\ 1\ 0 \times 0\ 0\ 1\ 0\ 1$ is then as follows:

```
  0 1 0 1 0    C_0 = A - 1 = 0 1 0 1 0
+ 0 1 0 1 0    b_0 2^0 A - 1 = 0 1 0 1 0,    2A - 1 = 0 0 1 0 0
  1 0 1 0 0
+         0
  0 0 1 0 0    C_1 = C_0 ⊕ (b_0 2^0 A - 1)
+ 1 0 0 0 0    b_1 2A - 1 = -1,             2^2 A - 1 = 0 1 0 0 1
  1 0 1 0 0
+         0
  0 0 1 0 0    C_2 = C_1 - 1 + 1 = C_1
+ 0 1 0 0 1    b_2 2^2 · A - 1 = 0 1 0 0 1,  2^3 A - 1 = 0 0 0 1 0
  0 1 1 0 1
+         1
  0 1 1 1 0    C_3 = C_2 ⊕ (b_2 2^2 A - 1)
+ 1 0 0 0 0    b_3 2^3 A - 1 = -1,          2^4 A - 1 = 0 0 1 0 1
  1 1 1 1 0
+         0
  0 1 1 1 0    C_4 = C_3 - 1 + 1 = C_3
+ 1 0 0 0 0    b_4 2^4 · A - 1 = -1,        2^5 A - 1 = 0 1 0 1 1
  1 1 1 1 0
+         0
  0 1 1 1 0    C_5 = C_4 - 1 + 1 = C_4
```

Thus, $C = C_5 = 0\ 1\ 1\ 1\ 0$ is the desired result of $0\ 1\ 0\ 1\ 0$ times $0\ 0\ 1\ 0\ 1$, module $2^4 + 1$ in diminished-1 notation.

In Example 1, one observes that no operation is needed for $b_i = 0$, that is, $C_{i+1} = C_i$ for $b_i = 0$. This example can be simplified as follows:

```
  0 1 0 1 0    C_0 = A - 1 = 0 1 0 1 0
+ 0 1 0 1 0    b_0 2^0 A - 1 = 0 1 0 1 0
  1 0 1 0 0
+         0
  0 0 1 0 0    C_1 = C_0 ⊕ (b_0 2^0 A - 1) = C_2
+ 0 1 0 0 1    b_2 2^2 A - 1
  0 1 1 0 1
+         1
  0 1 1 1 0    C_3 = C_2 ⊕ (b_2 2^2 A - 1) = C_4 = C_5 = C
```

Example 1 shows that diminished additions require the addition of the complement of an end-around carry to its sum. It was shown (Ref. 6) that a considerable speed improvement can be obtained by performing this operation simultaneously with the summation. A modified algorithm with this simultaneous addition is given for the previous example as follows:

Example 2:

```
  1 1 0 1 0    C̄_0 = 1 0 0 0 0 + (A - 1)
  0 1 0 1 0    b_0 2^0 A - 1
+         0
  1 0 1 0 0    C̄_1 = C̄_0 ⊕ (b_0 2^0 A - 1)
  0 1 0 0 1    b_2 2^2 A - 1 = 0 1 0 0 1
+         0
  0 1 1 0 1    C̄_3 = C_2 + (b_2 2^2 A - 1), C_2 = C_1 = C̄_1 + 1
  0 0 0 0 0
+         1
  0 1 1 1 0    C_3 = C̄_3 + 1 = C_4 = C_5 = C
```

III. A VLSI Structure for Implementing an Integer Multiplication Modulo F_n

In Fig. 1, A, B, C, and D are 5-bit, 6-bit, 5-bit, and 6-bit registers, respectively. Initially registers A, B, C, and D contain the multiplicand $A - 1$, the multiplier $B - 1$, $2^4 + (B - 1)$, and $2^5 - 1$, respectively. The content in register D is used to add $0\ 0\ 0\ 0\ 0$ into $\bar{C}_5 = C_4 + (b_4 \cdot 2^4 \cdot A - 1)$. That is, $C_5 = \bar{C}_5 + 1$. The content in register B is used to control whether $C_{i+1} = C_i$ for $b_i = 0$ or $b_i 2^i + 1$ is added into register C for $b_i = 1$. At the very same moment $\bar{C}_{i+1} = (\bar{C}_i + 1) + (b_i 2^i A - 1) = C_i + (b_i 2^i A - 1)$ is computed and loaded into the register C for

$0 \leqslant i \leqslant 4$. The diminished-1 multiplication of $(A - 1)$ by 2 is performed by a left cyclic shift of the four least significant bits of register A with the A_3 bit circulated into the first significant bit and complemented. Also at the same time registers B and D are shifted right by one bit. These operations are continued repetitively until the MSB of registers B and D are shifted out. The desired final result 0 1 1 1 0 is obtained in register C after six iterations.

The layout of the structure in Fig. 2 has been completed by the use of the CAESER design tool (Ref. 7). The final layout of the multiplication chip is shown in Fig. 2. Both logic and circuit level simulations were performed using the logic simulator "Esim" (Ref. 8) and circuit simulator "Spice" (Ref. 9). A timing analysis was done using the timing simulator "Crystal" (Ref. 10). The VLSI chip is being fabricated. The operating frequency is estimated at around 5 MHz with 3 μm CMOS technology. The total number of transistors in this chip is about 480. The chip for multiplication modulo F_2 in Ref. 6 requires 1000 transistors. Thus, this new multiplication algorithm requires only 50% of the transistors than the one in Ref. 6. The area of this chip with pads is estimated to be about 0.28 cm $\times$ 0.28 cm (110 mil $\times$ 110 mil).

IV. The Complex Modulo F_n Multiplier

Let $a + ib$ and $c + id$ be two complex numbers where a, b, c, and d are integers and $i^2 = -1$. The general complex multiplication of $(a + ib)$ and $(c + id)$ is $(a + ib) \cdot (c + id) = (ac - bd) + i(bc + ad)$ which needs four integer multiplications and two integer additions. An algorithm which can perform a complex multiplication by only two integer multiplications is introduced in Refs. 11 and 12. A special case of this algorithm for Fermat number F_n is introduced in Ref. 1. It is shown in Ref. 1 that $Z_{F_n}(i)$ is isomorphic to the direct sum of two copies of Z_{F_n}, i.e., $S_{F_n} = Z_{F_n} + Z_{F_n}$, integers modulo F_n. In Ref. 1, let s be the solution of $x^2 \equiv -1 \bmod F_n$. For $a + ib$ $Z_{F_n}(i)$, the following mapping,

$$f : a + ib \rightarrow ((a + sb)_{F_n}, (a - sb)_{F_n}) = (\alpha, \overline{\alpha}) \qquad (6)$$

is an isomorphism of $Z_{F_n}(i)$ onto S_{F_n} where addition and multiplication in S_{F_n} are defined by

$$(\alpha, \overline{\alpha}) + (\beta, \overline{\beta}) = (\alpha + \beta, \overline{\alpha} + \overline{\beta}) \qquad (7a)$$

and

$$(\alpha, \overline{\alpha}) \cdot (\beta, \overline{\beta}) = (\alpha \cdot \beta, \overline{\alpha} \cdot \overline{\beta}) \qquad (7b)$$

In the set of F_n, s can be found as $s = \pm 2^{2^{n-1}}$. Thus the forward mapping from $a + ib$ to $(\alpha, \overline{\alpha})$ requires cyclic shifts

and additions only. The inverse mapping of f, f^{-1}, is also simple. From Ref. 1,

$$a \equiv -2^{2^{n-1}}(\alpha + \overline{\alpha}) \bmod F_n \qquad (8a)$$

and

$$b \equiv -2^{2^{n-1}-1}(\alpha + \overline{\alpha}) \bmod F_n \qquad (8b)$$

Note that in Ref. 2 the negative number is the complement of the 2^n least significant bits of its positive counterpart. Hence the inverse mapping requires cyclic shifts, complements, and additions only. This complex integer multiplier algorithm is illustrated in the following example.

Example 3: Compute $(0\ 0\ 1\ 1\ 0 + i\ 0\ 0\ 0\ 1\ 0) \cdot (0\ 0\ 1\ 0\ 0 + i\ 0\ 0\ 0\ 1\ 1)$ module $F_2 = 2^4 + 1$.

In residue rings of $F_2 = 17$, one obtains $s = \pm 2^2$. By Eq. (6), the forward mapping of $(0\ 0\ 1\ 1\ 0 + i\ 0\ 0\ 0\ 1\ 0)$ is

$$(0\ 0\ 1\ 1\ 0 + i\ 0\ 0\ 0\ 1\ 0) \rightarrow ((0\ 0\ 1\ 1\ 0 + 2^2 (0\ 0\ 0\ 1\ 0)),$$

$$(0\ 0\ 1\ 1\ 0 - 2^2 (0\ 0\ 0\ 1\ 0))$$

$$= ((0\ 0\ 1\ 1\ 0 + 0\ 1\ 0\ 1\ 1), (0\ 0\ 1\ 1\ 0 - 0\ 1\ 0\ 1\ 1)$$

$$= (0\ 0\ 0\ 0\ 1, (0\ 0\ 1\ 1\ 0 + 0\ 0\ 1\ 0\ 0))$$

$$= (0\ 0\ 0\ 0\ 1, 0\ 1\ 0\ 1\ 1)$$

Similarly

$$(0\ 0\ 1\ 0\ 0 + i\ 0\ 0\ 0\ 1\ 1) \rightarrow (0\ 0\ 0\ 1\ 1, 0\ 0\ 1\ 0\ 1)$$

Thus, the multiplication over Z_{F_2} is

$$(0\ 0\ 1\ 1\ 0 + i\ 0\ 0\ 0\ 1\ 0) \cdot (0\ 0\ 1\ 0\ 0 + i\ 0\ 0\ 0\ 1\ 1)$$

$$= (0\ 0\ 0\ 0\ 1, 0\ 1\ 0\ 1\ 1) \cdot (0\ 0\ 0\ 1\ 1, 0\ 0\ 1\ 0\ 1)$$

$$= (0\ 0\ 0\ 0\ 1 \cdot 0\ 0\ 0\ 1\ 1, 0\ 1\ 0\ 1\ 1 \cdot 0\ 0\ 1\ 0\ 1)$$

Using the integer multiplier modulo F_2 developed in the previous section, one obtains $(0\ 0\ 1\ 1\ 0 + i\ 0\ 0\ 0\ 1\ 0) \cdot (0\ 0\ 1\ 0\ 0 + i\ 0\ 0\ 0\ 1\ 1) = (0\ 0\ 1\ 1\ 1, 0\ 0\ 0\ 1\ 1)$. By Eqs. (8a) and (8b), the inverse mapping of $(0\ 0\ 1\ 1\ 1, 0\ 0\ 0\ 1\ 1)$ are

$$e = -2^3 (0\ 0\ 1\ 1\ 1 + 0\ 0\ 0\ 1\ 1) = -2^3 (0\ 1\ 0\ 1\ 1)$$

$$= -0\ 1\ 0\ 1\ 0 = 0\ 0\ 1\ 0\ 1$$

and

$$f = -2(0\,0\,1\,1\,1 - 0\,0\,0\,1\,1) = -2(0\,0\,1\,1\,1 + 0\,1\,1\,0\,0)$$

$$= -2(0\,0\,0\,1\,1) = -0\,0\,1\,1\,1 = 0\,1\,0\,0\,0$$

Thus, $0\,0\,1\,0\,1 + i\,0\,1\,0\,0\,0$ is the desired result of $(0\,0\,1\,1\,0 + i\,0\,0\,0\,1\,0) \cdot (0\,0\,1\,0\,0 + i\,0\,0\,0\,1\,1)$ modulo $2^4 + 1$ in diminished-1 notation.

V. A VLSI Structure for Implementing a Complex Integer Multiplication Modulo F_n

In most digital signal processing applications, the multipliers are usually known. Thus, this multiplier can be precomputed to be $(\beta, \bar{\beta})$. Figure 3 shows the architecture for implementing Example 3. The same structure clearly extends to more general multiply algorithms over Z_{F_n}.

The layout of this multiplication chip is shown in Fig. 4. The logic, circuit, and timing simulations are performed. The chip of a complex integer multiplication circuit for $F_2 = 17$ is being fabricated. The operating frequency is around 5 MHz with 3 μm CMOS technology. The total number of transistors in this chip is about 2300. The area of the chip with pads is estimated to be about 0.41 cm $\times$ 0.41 cm (162 mil $\times$ 162 mil).

VI. Conclusion

A new Fermat number integer multiplier is described in this article. This new Fermat integer multiplier does not need the initialization procedure as the one developed previously. Both the area and operating speed of the chip are greatly enhanced with this modification. In Ref. 1, a quadratic-polynomial Fermat number system was used to compute complex integer multiplications. The advantage of such a system is that a complex integer multiplication requires only two integer multiplications while the conventional complex integer multiplier needs four integer multipliers.

In this article, the new Fermat number integer multiplier is used as an integer multiplier in the quadratic-polynomial Fermat number system to compute complex number multiplications. A VLSI architecture for this new complex integer multiplier is developed and it is demonstrated in this article that this new complex integer multiplier, which uses the Fermat number F_2, can be implemented on a single VLSI chip. Such a chip would be designed and fabricated in CMOS-pw technology.

References

1. Truong, T. K., Chang, J. J., Hsu, I. S., Pei, D. Y., and Reed, I. S., "Techniques for Computing the DFT Using the Residue Fermat Number System and VLSI," *GOMAC-85 Advance Program*, Orlando, FL, Nov. 5-7, 1985.

2. Leibowitz, L. M., "A Simplified Binary Arithmetic for the Fermat Number," *IEEE Trans. Acoustic, Speech and Signal Processing*, Vol. ASSP-24, No. 5, pp. 356-359, Oct. 1976.

3. Rader, C. M., "Discrete Convolution via Mersenne Transforms," *IEEE Trans. Comput.*, Vol. C-21, No. 12, pp. 1269-1273, Dec. 1972.

4. Agarwal, R. C., and Burns, C. S., "Fast Convolution Using Fermat Number Transform with Application to Digital Filtering," *IEEE Trans. Acoustic, Speech and Signal Processing*, Vol. ASSP-22, No. 2, pp. 89-97, Apr 1974.

5. Reed, I. S., Truong, T. K., and Welch, L. R., "The Fast Decoding of Reed-Solomon Code Using Fermat Number Transforms," *IEEE Trans. Information Theory*, Vol. IT-24, No. 4, pp. 497-499, July 1978.

6. Chang, J. J., Truong, T. K., Reed, I. S., and Hsu, I. S., "VLSI Architecture for the Multiplication of Integers Modulo a Fermat Number," *IEEE Trans. on Acoustic Speech, and Signal Processing*, Vol. ASSP-35, No. 6, pp. 1599-1602, Dec. 1985.

7. Ousterhout, J., *Editing VLSI Circuits with Caesar*, Computer Science Division, Electrical Engineering and Computer Sciences, University of California, Berkeley, Apr. 21, 1982.

8. Tevman, C., *Esim-Event Driven Switch Level Simulator*, Dept. of Electrical Engineering, MIT, 1977.

9. Negal, L. W., and Pederson, D. O., *SPICE-Simulation Program with Integrated Circuit Emphasis*, Memorandum No. ERL-M382, Electronics Research Laboratory, 1976.

10. Ousterhout, J., *Using Crystal for Timing Analysis*, Computer Science Division, Electrical Engineering and Computer Sciences, University of California, Berkeley, Mar. 1983.

11. Despain, A. M., Peterson, A. M., Rothaus, O. S., and Wold, E., "Fast Fourier Transform Processors Using Complex Residue Arithmetic," to appear in *The Journal of Parallel and Distributed Processing*.

12. Cozzens, J. H., and Finkelstein, L. A., "Computing the Discrete Fourier Transform Using Residue Number Systems in a Ring of Algebraic Integers," *IEEE Trans. on Information Theory*, Vol. IT-31, No. 5, pp. 580-588, Sept. 1985.

13. Shyu, H. C., Truong, T. K., and Reed, I. S., "A Complex Integer Multiplier Using the Quadratic Polynomial Residue Number System with Numbers of Form $2^{2n} + 1$," *International Workshop on Systolic Array*, University of Oxford, England, July 2-4, 1986.

Table 1. The correspondence among decimal numbers, their values in the normal binary representation, and in the diminished-1 representation

Decimal number	Normal binary representation	Diminished-1 representation
0	0 0 0 0 0	1
1	0 0 0 0 1	2
2	0 0 0 1 0	3
3	0 0 0 1 1	4
4	0 0 1 0 0	5
5	0 0 1 0 1	6
6	0 0 1 1 0	7
7	0 0 1 1 1	8
8	0 1 0 0 0	9 (−8)
9 (−8)	0 1 0 0 1	10 (−7)
10 (−7)	0 1 0 1 0	11 (−6)
11 (−6)	0 1 0 1 1	12 (−5)
12 (−5)	0 1 1 0 0	13 (−4)
13 (−4)	0 1 1 0 1	14 (−3)
14 (−3)	0 1 1 1 0	15 (−2)
15 (−2)	0 1 1 1 1	16 (−1)
16 (−1)	1 0 0 0 0	0
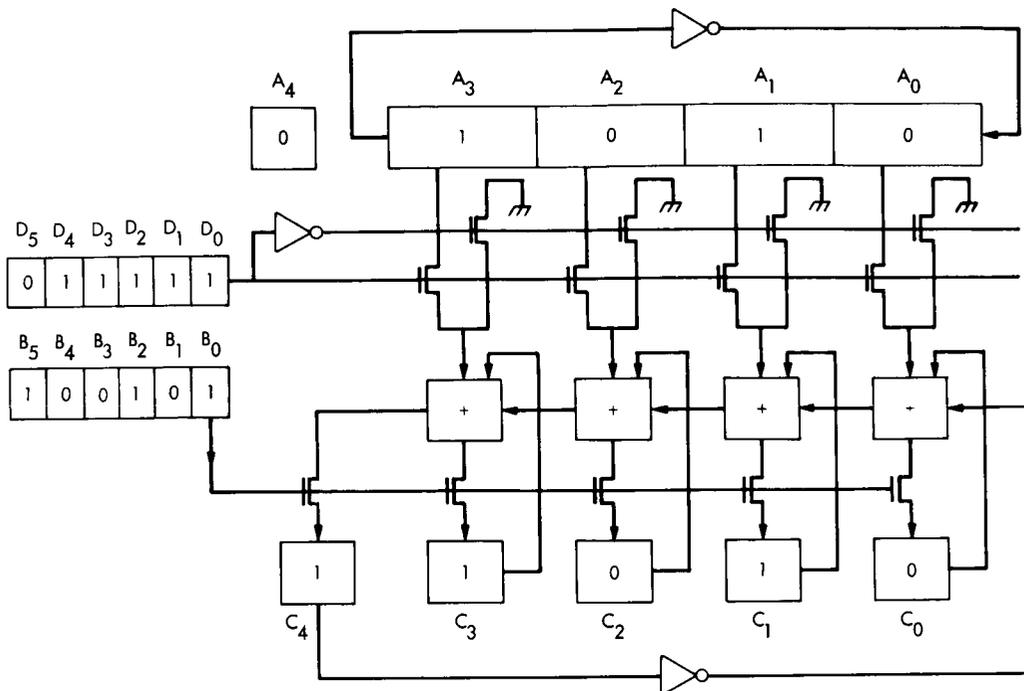
Fig. 1. The pipeline architecture for the implementation of multiplication modulo the Fermat number $2^4 + 1$ using diminished-1 number presentation
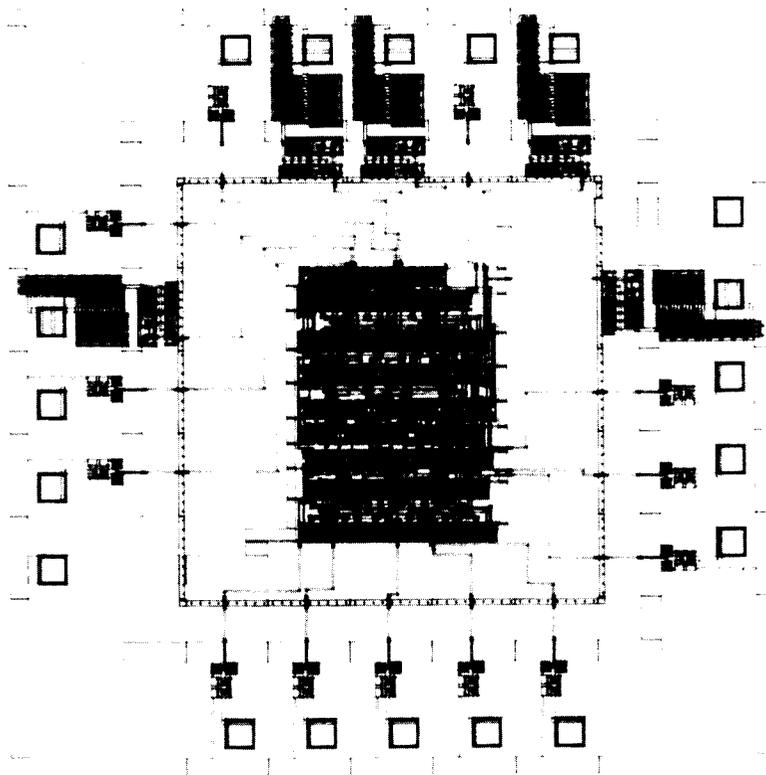
Fig. 2. VLSI layout of an integer multiplication circuit for $F_2 = 2^4 + 1$
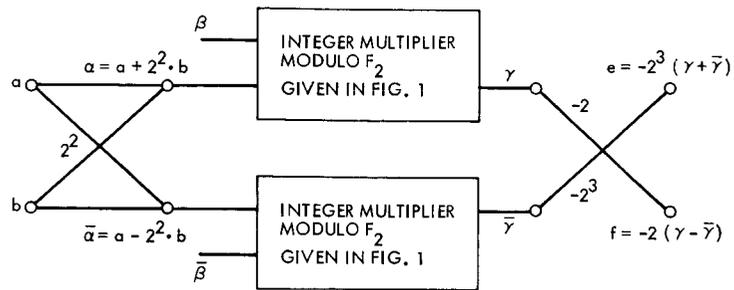
Fig. 3. The architecture for implementation of complex integer multiplication modulo $F_2 = 17$ using diminished-1 number representation
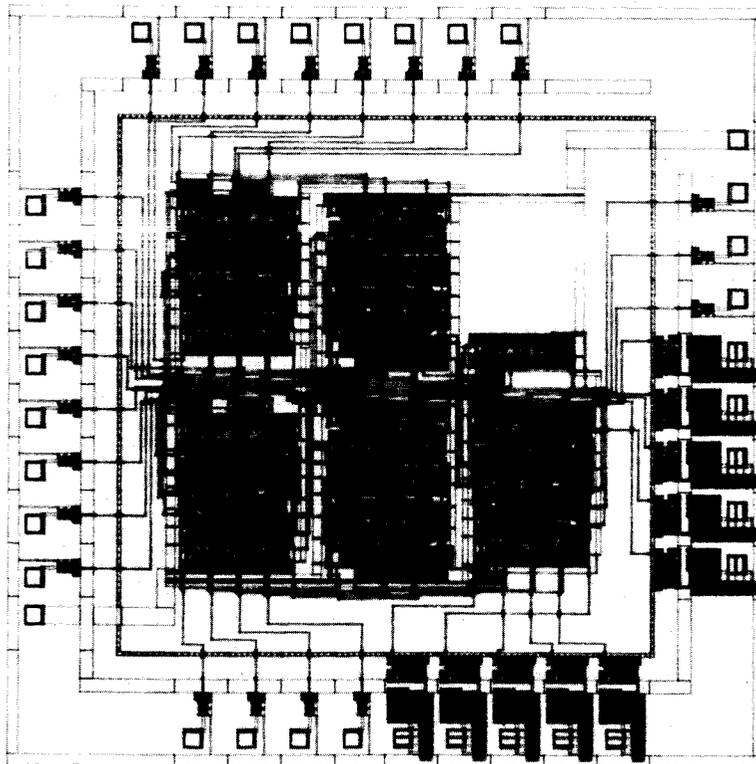
Fig. 4. VLSI layout of a complex integer multiplication circuit for $F_2 = 17$

Unattended Deep Space Station Tracking Station Development: Monitor and Control Technology

C. Foster

Radio Frequency and Microwave Subsystems Section

The major developments leading to successful demonstration of fully unattended operation of a DSN station (DSS 13) are reviewed. Unattended operation was demonstrated by reliably tracking, commanding, and delivering telemetry from the Pioneer 8 spacecraft. Transfer of automated monitor and control technology to DSN implementation is summarized, along with related accomplishments.

I. Introduction

In 1975, the Deep Space Network began to investigate operating an unattended tracking station as a means of lowering cost, adding to personnel safety, providing improved tracking time, and reducing DSN equipment maintenance. The DSS 13 station was used as the testbed for developing a remotely controlled, unattended station. The unattended operation configuration at DSS 13 included a 26-m antenna, a high-power transmitter, a receiver-exciter, a subcarrier demodulator, a command system, a waveguide switching system, and a data path to JPL (Fig. 1). The project was stretched over several years due to funding constraints and the need to develop complex subsystem transducers.

Several demonstrations were scheduled to validate the design concepts well in advance of final program completion. These demonstrations were carried out as each major subsystem was added to the testbed. Fully unattended receiver capability was demonstrated for 6 months in 1979 to provide controlled lifecycle cost data (Ref. 1). Unattended operation of the high-power transmitter was demonstrated for 2 months in 1981 (Ref. 2). Fully unattended station operation was demonstrated by tracking Pioneer 8 in 1984.

II. Control Architecture

A functional requirements document was developed describing software standards and design. This document set forth distributed control as the system design philosophy. Each major class of equipment (microwave, receiver-exciter, subcarrier demodulator, command, and transmitter) had its own controller, which interfaced to the station controller. Each subsystem controller is capable of operating, diagnosing, and monitoring its assigned subsystem from a local terminal. This philosophy of distributed division of control allowed a viable fallback position for semiautomatic operation of the station if the station controller failed, thereby providing uninterrupted use of DSS 13 for other projects. One or two engineers for each subsystem were responsible for developing software that could configure, operate, and maintain safe operation of their subsystem. An added advantage to this independent design philosophy was that concepts could be tested separately and individual subsystem implementations could be late with little impact on the overall station.

The basic unattended station architecture consists of a remote satellite controller at JPL's NOCC, and a station controller at DSS 13 that communicates with the independent

subsystem controllers (Fig. 2). All controllers use an 8080 CPU-based microprocessor developed and assembled at JPL. The principal monitor and control functions are distributed among the subsystem controllers as described below.

(1) NOCC controller. The NOCC controller serves as an automatic scheduler, message router, and operator interface. This controller provides an interface to the operator, DSN predicts files, scheduling files, and the DSS 13 station controller. The NOCC controller routes directives for configuration, predicts, command, and operation. It receives and automatically logs all configuration and performance data.

(2) Station controller. The station controller functions as the supervisory computer, responsible for receiving and transmitting subsystem control directives and for coordination of the overall tracking system. The station controller can be operated from either the NOCC controller or by a local operator. Based on the operational scenarios supplied by the station controller, each subsystem controller performs the detailed tasks required to calibrate, safely operate, and read/write its assigned functions. For example, the station controller has control of the acquisition process after receiving a directive to operate from the NOCC or via local terminal. The station controller continuously monitors the necessary conditions for proper operation. If a subassembly fails, it is possible to maintain the communications channel via the station controller by implementing a prearranged reconfiguration of a backup assembly.

(3) Antenna controller. The antenna controller monitors the antenna servo and mechanical subsystem and local weather conditions. It controls antenna pointing via high-level operator directives. The antenna controller also contains a separate maintenance program that allows maintenance personnel to fully characterize the electrical and mechanical functions of the antenna during scheduled off-line maintenance (Ref. 3).

(4) Receiver-exciter controller. The receiver-exciter controller maintains the receiver-exciter configuration and tunes the uplink and downlink signal. Acquisition of the spacecraft downlink is accomplished automatically in a manner similar to the actions performed by a skilled operator.

(5) Microwave controller. The microwave controller manages and monitors high-power transmitter paths to ensure the safety of personnel and equipment. The controller also monitors and controls the transmitting and receiving signal paths.

(6) Subcarrier demodulator controller. The subcarrier demodulator controller monitors the performance and controls the configuration to provide automatic acquisition. It provides a graphic output of the correlation meter to the operator for use in manual acquisition when the predicts supplied are not within the demodulator's acquisition bandwidth.

(7) High-power transmitter controller. The high-power transmitter controller controls and monitors all safety functions, power calibration, and operation of the high-power transmitter subsystem (Ref. 2).

(8) Traveling wave maser (TWM) and closed-cycle refrigerator (CCR) controller. The TWM/CCR controller provides continuous monitoring of CCR performance data. This data was automatically logged 24 hours a day over a 3-month period and was used to supply engineering with performance numbers previously unavailable because of manpower limitations. This data will be used to characterize CCR performance and eventually to prevent data loss due to unexpected warmup of the TWM (Ref. 4).

III. Pioneer 8 Demonstration and Related Design Tasks

Before a planned tracking demonstration using Pioneer 8 could take place, two new design tasks arose. An improved acquisition detector was required to reach the receiver threshold (Pioneer 8 was 5 dB below the current auto acquisition detector threshold). Also, a multibus version of the DSN command modulator was required because the existing modulator was needed to support operations at Mil 71.

A fast Fourier transform (FFT) automatic acquisition module was designed into the unattended station configuration for the Pioneer 8 demonstration. The FFT module performs a parallel search of the RF spectrum for the desired signal and estimates the difference in frequency between the spacecraft signal and the frequency to which the receiver is tuned. The spacecraft downlink signal is nearly sinusoidal, and based on previous experience and theory of optimum receiver sinusoidal signals of unknown phase (Ref. 5), it was decided that the best approach would be to use power spectrum analysis techniques. The idea was to produce the spectrum, scan it for peaks, and estimate the difference between the frequency of the peak and the frequency to which the receiver was currently tuned. The implementation is shown in Fig. 3. The 10-MHz IF signal plus noise is mixed with a 10-MHz reference signal. The two quadrature components produced are then digitized and processed through a complex-to-complex FFT routine. The resulting Fourier transform is

squared to produce a power spectrum, averaged to statistically reduce the variance of the signal, and inspected by the computer for significant special peaks. The FFT acquisition module runs continuously and asynchronously without control inputs. The spectrum estimated SNR and the estimated tuning correction are passed to the receiver controller, where they are used to automatically retune to the spacecraft frequency and to acquire the signal.

The command modulator assembly (CMA) was designed to fit on a standard multibus board and reside in the station controller chassis (Fig. 4). The CMA was tailored for the Pioneer 8 demonstration, but was designed so that it could be easily modified to support the tracking of other spacecraft. The CMA provides frequency shift key (FSK) modulation synchronous with the subcarrier. It also provides configuration verification and transmitted bit confirmation. If any data is not verified, the command sequence is aborted, the operator is notified that the command was not sent, and the probable cause of failure is displayed.

By 1984, DSS 13 was successfully operating in the unattended station configuration — routinely tracking several spacecraft downlinks and remotely operating the high-power transmitter on a daily basis. The next logical step was to take on a full tracking assignment, including uplink tuning and command of a spacecraft. When the Pioneer 8 spacecraft became available, a set of demonstration objectives was defined in order to evaluate whether the DSS 13 unattended tracking configuration could safely and reliably perform assigned spacecraft tracking missions. The evaluation was based on tracking Pioneer 8 twice a week and measuring the station's ability to acquire the spacecraft signal at the scheduled time, tune the uplink, send and verify commands, return the spacecraft to its previous state, and collect telemetry data at JPL. The data was sent to the Project Office at NASA/Ames for validation. If station personnel were required in any

capacity, the track pass was deemed a failure. The Pioneer 8 demonstration took place for a 12-month period and resulted in 87 tracks. The results are shown in Fig. 5. As can be seen, the success and reliability improved over time. At no time after successful uplink acquisition did a command fail to be sent. Also, the spacecraft was never left in a state different than what was planned at the start of the track.

IV. Accomplishments

(1) The unattended station development effort has provided the DSN with new options for automated and remotely controlled tracking stations.

(2) High-level computer languages were adopted, tested, and transferred to implementation.

(3) The high-power transmitter and waveguide switch subsystems were transferred to DSN implementation with little or no change.

(4) New tools were developed and tested that increased productivity in the development and troubleshooting of complex real-time control systems that have distributed processors embedded in the control feedback.

(5) New digitally programmed oscillators were developed and transferred directly to operations in time to solve tuning problems that occurred with the Voyager spacecraft (POCA, DCO).

(6) The philosophy of containing all safety decisions at the subsystem level and out of the hands of the operator eliminated the threat of injury to personnel and equipment during tracking.

(7) Not one piece of equipment was ever damaged due to the automation, and equipment failure rates went down because handling of equipment was reduced.

References

1. "Deep Space Station (DSS 13) Automation Demonstration," *TDA Progress Report 42-57*, pp. 103–119, Jet Propulsion Laboratory, Pasadena, Calif., June 15, 1980.

2. "High-Power Transmitter Automation," *TDA Progress Report 42-57*, pp. 71–81, Jet Propulsion Laboratory, Pasadena, Calif., June 15, 1980.

3. "DSS 13 Microprocessor Antenna Controller," *TDA Progress Report 42-77*, pp. 64–74, Jet Propulsion Laboratory, Pasadena, Calif., May 15, 1984.

4. "Traveling-Wave Maser, Closed Cycle Refrigerator Automation," *TDA Progress Report 42-65*, pp. 99–103, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 15, 1981.

5. Viterbi, A. J., *Principles of Coherent Communication*, McGraw Hill, NY, NY, 1966.
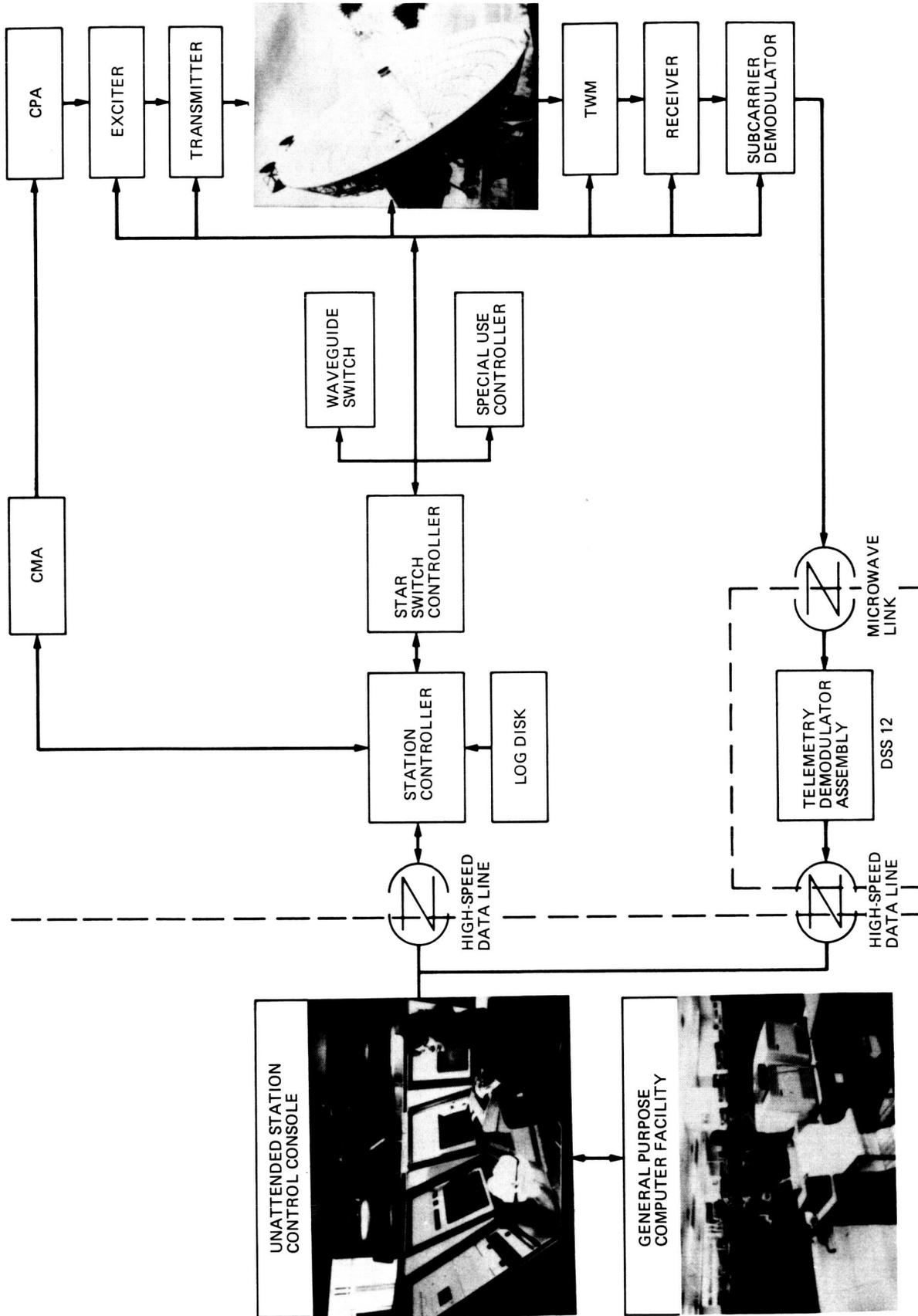
Fig. 1. DSS 13 unattended station system block diagram
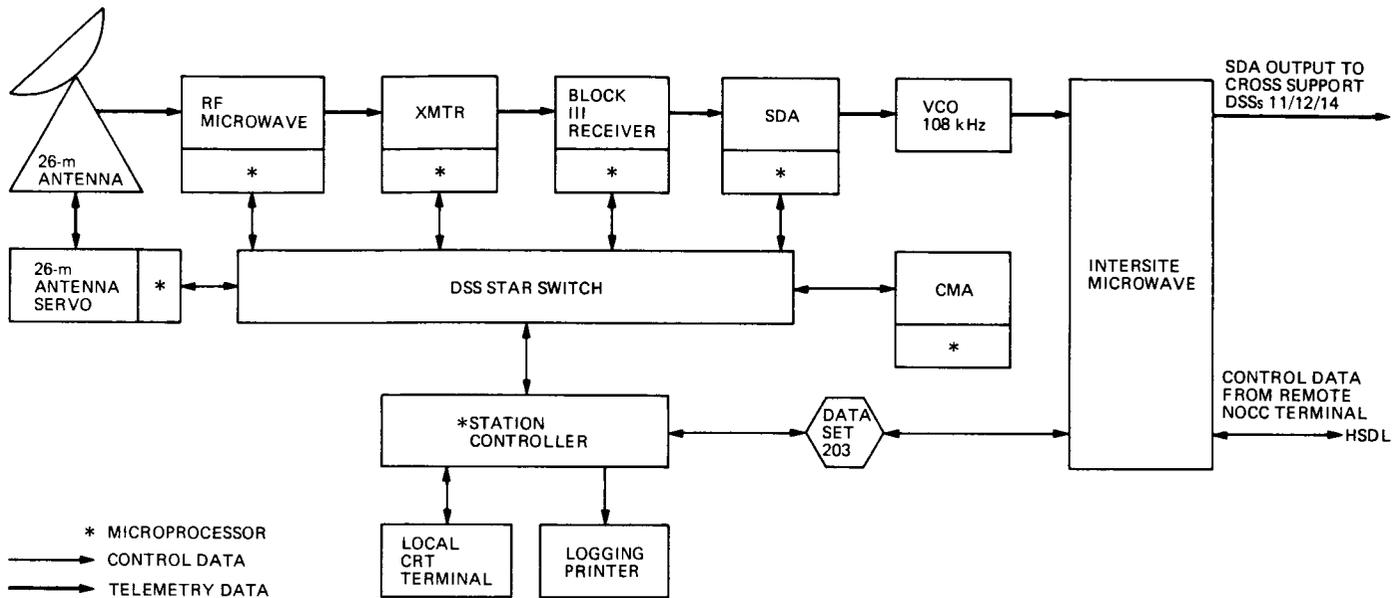
Fig. 2. DSS 13 unattended station functional block diagram
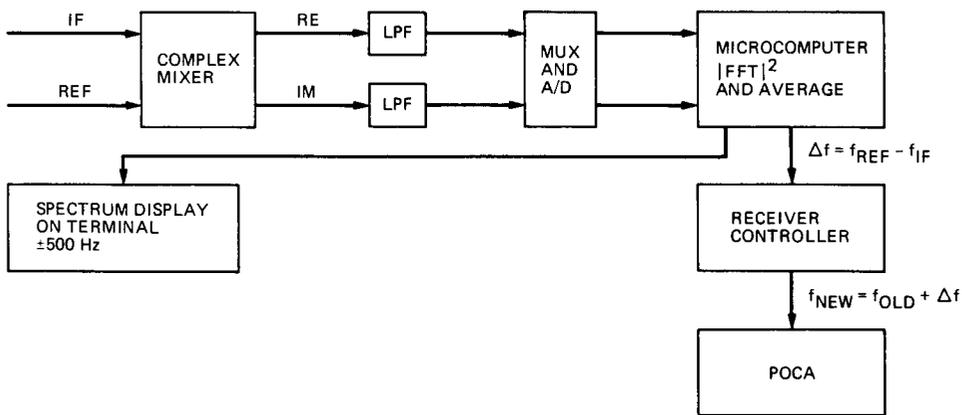
Fig. 3. FFT acquisition aid functional block diagram
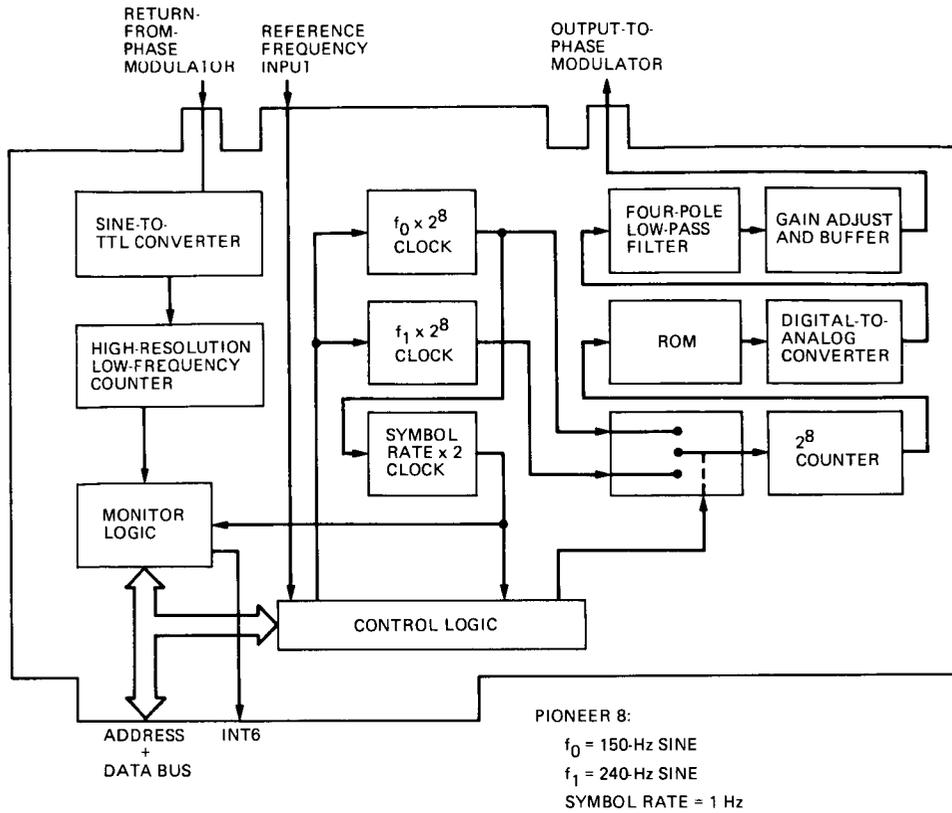
Fig. 4. Command modulator assembly functional block diagram

PIONEER 8:

f_0 = 150-Hz SINE
f_1 = 240-Hz SINE
SYMBOL RATE = 1 Hz
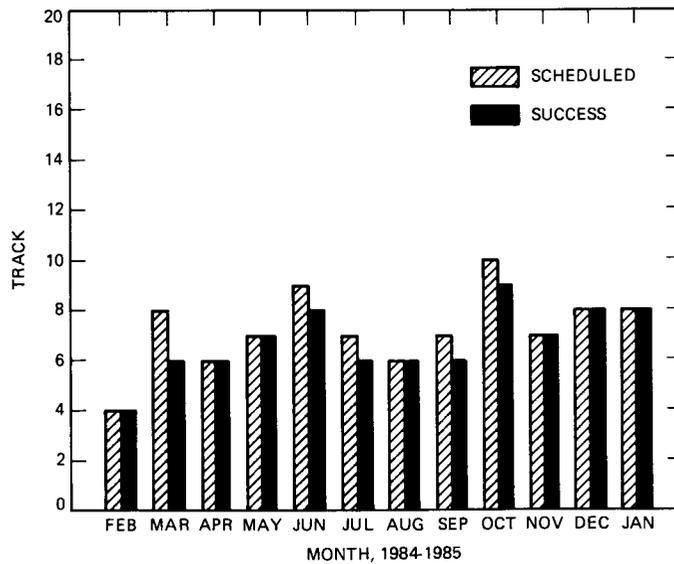
Fig. 5. DSS 13 Pioneer 8 tracking demonstration schedule and success ratio

C-3

N87-11043

A System for the Functional Testing and Simulation of Custom and Semicustom VLSI Chips

E. M. Olson and L. J. Deutsch

Communications Systems Research Section

This article describes a system for the functional testing and simulation of custom and semicustom very large scale integrated (VLSI) chips that are designed using the integrated UNIX-based computer-aided design (CAD) system. The testing and simulation system consists of two parts. One of these is a special purpose hardware device that is capable of controlling the digital inputs and outputs on a custom chip. This device, the Digital Microcircuit Functionality Tester (DMFT) system, can be operated by itself or in conjunction with the VAX host computer on the CAD system. The DMFT is integrated into a microprobe station so that these signals can be injected or read from nodes inside the chip, as well as at the pins. The second part of the system is a software package that is installed on the VAX. This software package, "logic," includes a full-screen editor for developing chip test sequences and drivers for both the DMFT and the "esim" logic simulator.

I. Introduction

The design and testing of custom and semicustom VLSI chips for the Deep Space Network (DSN) Advanced Systems Program at JPL is accomplished on an integrated UNIX-based CAD system. This system, as described in [1], comprises various software tools and hardware peripheral devices that facilitate the design, simulation, verification, and testing of very complex microcircuits in a number of different solid-state technologies. This article is devoted to the portion of the system that is used to perform functional testing of the microcircuits after their fabrication.

One of the most difficult tasks in the production of custom or semicustom digital microcircuits is verifying that the com-

pleted part performs the function for which it was designed. This is, in general, a long and laborious process that involves the use of many varied pieces of test equipment. The following is a description of how this process was performed before the present system was completed.

Chips would first be examined under a microscope to weed out those that were obviously bad. After this process, the remaining chips would be tested for functionality. In order to prove that a chip is performing the desired function, it is often necessary to submit the chip to many thousands of bits of data while monitoring the chip's outputs. This has been done in the past by using digital word generators for input and oscilloscopes or logic analyzers for output. Even the best word gen-

erators, however, are limited in the number of bits they can store. A typical generator can produce signals that are up to 1024 bits in length. This number is inadequate for most of the chips that are being designed in this effort. A Reed-Solomon encoder chip [2], for example, requires at least 2500 bits of non-periodic data for even the simplest test. An alternative is to build special hardware "jigs" for each chip that produce the desired bit sequences. This takes a lot of time and can be very expensive. In either case, there is no good way for the user to define the sequences that are needed. Finally, the output signals, as seen on the oscilloscope or logic analyzer, are not easily stored for future use. These problems have been essentially alleviated by the development of the Digital Microcircuit Functionality Tester (DMFT) and the "logic" software package.

The DMFT has been developed as part of the JPL Advanced Systems Program specifically for incorporation into the integrated UNIX-based CAD system. The tester was built internally to meet specific budget demands that weren't addressable by testers of similar functionality and speed that were currently on the market. Furthermore, since the cost of developing the tester was relatively small, the VLSI design engineers in this program were afforded a chance to develop requirements for future testers from first-hand experience and evaluation of this prototype tester. Since the tester was inexpensive, the cost of developing this first-hand experience is considerably less than it would cost to rent or purchase a tester from the current market. The DMFT has also allowed these design engineers to begin developing standards and requirements for testers to be incorporated into the system in the future, while at the same time addressing the pressing need for test capability at the present.

The DMFT consists of special purpose hardware and firmware for the generation and monitoring of digital signals to and from a VLSI chip. It is capable of generating eight independent sequences of up to 4096 bits at a time. The DMFT also generates several clock signals that are synchronized to these data. It can subject these to the chip being tested (sometimes called the device under test or DUT) at selectable clock rates of up to 20 MHz. This maximum frequency is sufficient at present because all the design work that is currently going on involves MOSFET [3] devices and fabrication techniques that cannot operate any faster than this. The DMFT can monitor eight chip outputs for up to 4096 bits at a time. The firmware in the DMFT supports a terminal interface so that the user may define and edit the test vectors, run tests, and display the output. In addition, the firmware includes the support of a host computer interface with asynchronous protocol so that the DMFT may be completely controlled by the VAX computer.

This special purpose hardware and firmware is built into a specially designed microprobe station. The station consists of a high-power stereoscopic microscope for examining the device under test, a set of four probes that can be positioned on points inside the actual chip, and an image-shearing device for measuring the thickness of layers on the chip. The DMFT is mounted so that the chip being tested is positioned under the microscope. Signals to and from the chip can be routed through the chip's pins and/or through any of the probes. This means that the DMFT system may be used for interactive debugging of bad designs by tracing signals back through the chip to find problem areas.

The logic software package runs on the UNIX-based CAD system. It was written here at JPL so that users of the CAD system could easily control the DMFT from any remote site. The package is fairly comprehensive and includes a full-screen editor for defining, displaying, and modifying large sets of test vectors. The format of the editor is similar to that of many commercial logic analyzers and is, therefore, easy for the designers to use. The sequences may be stored on a disk and read back later. In this way, the designer can build a large database of test vectors for a design. The logic program can use these vectors to run the DMFT and display the results in real time. The results can also be stored on the disk or printed on a line printer. Logic can also use the vectors to simulate a chip before fabrication. This is done by automatically invoking the "esim" logic simulator [1, 4]. By using this capability, a designer does not have to learn two separate systems for simulation and testing. Also, the designer can use the same test vectors for testing as were used in simulation. The results can be compared automatically using the UNIX compare utilities to verify the functionality of the finished chip.

The remainder of this report describes the tester system in detail. Section II describes the testing philosophy behind the design. Section III contains an overview of the DMFT hardware. Sections IV and V describe the DMFT hardware in detail. Section VI describes the logic software package. Section VII is devoted to an assessment of the current system and plans for future generation testing systems.

II. The Testing Philosophy

Integrated circuit testing is a very broad field and incorporates many disciplines and technologies. In order to fully verify that an integrated circuit is qualified for its application, it must not only pass electrical specifications, but also mechanical and cosmetic specifications. Currently the Advanced Systems Program VLSI design effort is directed to research and development of prototype LSI and VLSI architectures and their applications. As such, this effort is currently not con-

cerned with mechanical testing, and is concerned with cosmetic testing only for the purpose of rejecting obviously bad dies. Also, the output of effort to date has exclusively been digital MOSFET designs. Thus, the testing effort is further limited primarily to digital qualification at speeds under 20 MHz. Although parametric testing (i.e., testing that considers the analog performance of devices) is a very important step in the qualifying of digital circuits, the tester described herein does not have any capabilities to directly measure currents, voltages, or any other analog qualities of an integrated circuit — it has been designed only to test for logical functionality.

Logical testing of digital integrated circuits can be over-simply characterized as the systematic presentation of stimulus vectors to the device under test, while systematically recording response vectors from the device. Vectors consist of sets of bits either presented to (stimulus) or read from (response) the device. The voltage-to-logic-level correspondence is preset to some range for all the bits in a vector (for example 0 to 0.8 volts on a pin for a logical "0" and 2.4 to 5.0 volts for a logical "1"). Usually each bit in a test vector corresponds physically to the logic level that is presented to (or recorded from) a particular pin on the device. The vector itself can be thought of as a copy of the logical status of some particular pins on the device at some point in time or during an interval of time.

III. The DMFT — An Overview

The DMFT hardware and firmware are housed in an aluminum chassis that mounts under a microprobe station platform (see Figure 1). The entire electronics of the tester is contained in this chassis, including three power supplies, front panel control logic, and a standard multibus (IEEE standard number 796) circuit board on which most of the circuitry itself is mounted. A 64-pin zero insertion force socket and a signal patch panel are also mounted on the multibus board. The socket and the patch panel protrude through an opening in the probe platform allowing access to the device being tested with the microscope and probing equipment, while maintaining the bulk of the tester's electronics out of sight. The test setup is controlled partly by front panel controls and partly by setup through the tester firmware via either a host computer or a standard RS-232 (CCITT recommended standard number V.24) terminal (this later mode of operation is called stand-alone).

Figure 2 shows the layout of the front panel controls and indicators. Their functions are described as follows:

Power-On Indicator — This indicator is lit when the DMFT system tester power is on

Slow Clock Indicator — This indicator is lit when the slow internal clock range is selected; if it is not lit, then the fast internal clock is selected

Test Indicator — This indicator is lit while a test is actually in progress

Internal/External Clock Switch — This switch is used to select either the internal DMFT clock or an external clock for testing

External Clock In — This is a connector for external clock input

Frequency Adjust — This is a Vernier dial that is used to set the internal clock frequency

Clock Out — This is a connector that provides the selected clock signal as an output

DUT Voltage Adjust — This is a Vernier dial that is used to adjust the power supply voltage from 1.5 to 7 volts for the device under test

DUT Current Meter — This is a meter that monitors the electrical current being drawn by the device under test

DUT Voltage Meter — This is a meter that monitors the voltage applied to the device under test

Main Power Switch — This is the main power switch for the DMFT

DUT Power Switch — This switch is used to turn power, to the device under test, on and off

Reset Button — This push-button resets the entire tester and restarts tester firmware.

A block diagram of the DMFT is shown in Fig. 3. It is clear from the figure that the concept is fairly simple. As mentioned above, the entire tester electronics including a 64-pin zero insertion force socket and a signal patch panel, fit on one multibus board. The tester hardware itself can be divided into two sections. The first section consists of the MUX, DEMUX, Vector Memory, CNTR ADDR GEN, and CLOCK GENERATOR sections of the block diagram and will be called the "Vector Buffer Section." The second section of the tester consists of the controlling microprocessor and its associated support circuitry, and also includes the firmware contained in the microprocessor's memory. It will be called the "Control Section." These two portions will be discussed in the following two sections.

An abbreviated list of the important DMFT hardware specifications appears in Table 1.

IV. The Vector Buffer Section

The Vector Buffer Section (VBS) of the tester is really the core of the tester and is that part of the hardware that actually presents and records stimulus and response vectors during a test. The tester's VBS and the test itself are entirely synchronized to the Clock Generator's master clock. The Vector Memory is designed to be able to transmit and receive vectors at a rate of 20 MHz (one vector output and one received every 50 nS). The memory is actually implemented on 4K-deep-by-8-bits-wide, 45 nS, static random access memory (SRAM) chips and are arranged as 4 groups of 4K-deep-by-8-bits-wide chips. The idea behind the organization and support circuitry is to maintain two groups of two parallel memories, one group to drive vectors to the device under test and one group to receive vectors from it. The two parallel memories within each group are cycled 180 degrees apart, allowing the DMFT to receive and return vectors at twice the rate of which one individual memory is capable. This organization and clocking scheme is effected in the following manner.

Memories C and D each contain test vectors to be sent to the device being tested. These vectors are placed in the two memories in such a way that memory C contains the data for odd-numbered time slots and memory D contains the data for even-numbered time slots. The Control Section handles this ordering such that it is transparent to the user. Memories A and B each contain vectors returned from the device being tested as the results of prior test vectors. Those resultant vectors are also ordered in memories A and B the same way the odering takes place in C and D. Again, upon reading the resultant vectors the multiplexed organization is transparent to the user by manipulation through the Control Section. Stimulus vectors from buffers C and D are synchronously multiplexed to the device under test while response vectors are synchronously multiplexed to buffers A and B.

Control of the addressing of the Vector Memories is effected by the Address Generators A and B. Initial addresses for the memories are loaded into presettable counters by the Control Section, commanded by the user. The initial address determines the starting point for a test or a vector transfer from the Vector Memories to the Control Section and on to the outside world. The ending address for a test is also loaded into the Counter Address Generator block by the Control Section, and the loaded value determines the length of a test. When the ending address has been reached, the end of test interrupt is asserted and the VBS is returned to the Control Section.

All clocks in the vector buffer section are derived in some way from the output of the internal voltage-controlled oscillator (VCO) or the external clock. The internal VCO is selectable for a slow or fast range through the Control Section. Since everything in the VBS is synchronized to the master clock, the use of the internal VCO or an external clock, as the master, allows continuously variable test rates of up to 20 MHz. The clock rate for the internal VCO is a few kHz to about 20 MHz. Additionally, since the entire VBS is treated as an asynchronous input/output device by the Control Section, the tests run completely asynchronously to the Control Section and the microprocessor is not required to do real-time work for the performance of tests. The actual frequency of the internal clock can be set by the front panel Frequency Adjust and monitored through the Clock Out connector.

V. The Control Section

The Control Section consists of an Intel-Z80A 4-MHz microprocessor, a serial input/output controller (SIO), a counter timer circuit (CTC), some address decoding logic, some interrupt arbitration logic, some buffers and latches, and the Z80's core memory. The CTC is programmed to interrupt the microprocessor every second for user timing features. The CTC also provides baud rate clocking to the SIO for both SIO ports. Interrupt arbitration is handled internally by the SIO (highest priority) and CTC (second priority) and externally for the End-Of-Test (EOT) signal by the interrupt control logic. The SIO and CTC are programmed to present their own interrupt vectors with their interrupts. The EOT interrupt vector is presented at the appropriate time in conjunction with the end of a test.

The primary function of the Control Section is to act as the interface between the Vector Buffer Section, which actually performs tests, and the user who desires a test. The user can be either the host computer or someone at a terminal connected directly to the tester (stand-alone mode). The use of the microprocessor in the Control Section, along with the SIO controller allows a convenient interface to the tester through an RS-232 standard interface. The Control Section hardware and firmware combine to allow the user to generate and input test vectors, and to view the stimulus and response vectors after a test, in any one of three formats in the stand-alone mode. Also in the stand-alone mode, the user has complete control over the tester through a simple monitor program that, among other things, allows the user to enter and execute his own firmware for special configurations or testing. The Control Section hardware and firmware also handles all high-speed data transfers and handshaking between the host computer and the tester. Finally, the Control Section handles all the interpretation of commands for test setup including the time interlacing of vectors as mentioned above in the vector buffer discussion.

In addition to the visible functions the Control Section performs, there are several other functions that are handled by the Control Section that are not as readily apparent. For example, on power up or reset, the vector memories are tested and bad memory chips are called out by part number, permitting quick diagnosis and replacement of bad parts. The Control Section also keeps track of the length and duration of tests and insures that commands aren't accidentally executed during a test run that might interfere with the test results. The Control Section buffers all commands internally allowing the "repeat-last-command" function and correction on the fly. Special commands available to the user through the stand-alone terminal permit non-periodic and variable length periodic stimulus vectors to be generated quickly with a single command. The tester is capable of detecting and flagging erroneous data transfers or requests from the host computer, adding further confidence to test results. The host computer interface is defined entirely in the American Standard Code for Information Transmission (ASCII), allowing the tester to be used with virtually any host computer that can implement the command protocol over an RS–232 interface.

VI. The Logic Software Package

The logic software package was written to run on the UNIX-based CAD system and to allow the user to create, edit, store, and retrieve test vectors and to run both the esim logic simulator and the DMFT. The program is interactive and it may be run from any user terminal on the CAD system. This means that both simulation and testing of a chip design may be performed from the user's office. The user must be communicating to the host VAX computer on an ANSI standard (American National Standards Institute documents X3.4 and X3.64) terminal. All of the terminals that are currently produced by the Digital Equipment Corporation (such as the VT100 and VT200 series terminals) meet this standard as well as most personal computer terminal emulation programs. In fact, logic has been run on many different terminals including Apple and IBM personal computers. A photograph of the logic program screen in use appears in Fig. 4.

The logic program has three modes of operation. These are called the "display," "edit," and "test" modes and they are each described below. Each of these modes has a set of single letter commands associated with it. The appropriate command menu for the current mode is always displayed in the upper portion of the terminal screen. Most of the remainder of the screen is used to display a 16-line by 64-bit window on the test vectors. The program has the capacity to handle 80 lines of 4096 bits — and this can be easily changed by recompiling the code. The bottom line of the screen is the "command" line. It is used by the user to enter various parameters during the operation of the program and to display error messages and

warnings. The very top part of the display contains various pieces of useful information including the name of the file being edited, the current bit, the number of inputs and outputs in the current set of test vectors, and the current mode.

The display mode comprises a mechanism for scanning through the test vectors that are currently loaded into memory. It also contains commands for writing vectors to the host disk and reading vectors from the disk and for obtaining a hardcopy printout of the current vectors (also in logic analyzer format). The screen contains a cursor that points to the current signal (line) and bit. The commands in the display mode are actually available to the user in all three modes. The display menu only appears in the display mode.

The second mode in the logic program is the "edit" mode. The edit mode contains a set of commands that are used to create test vectors and to modify them. Each signal, or line, in the display can be made either an "input" or "output" signal. Input signals are stimuli to the simulator and tester while output signals are responses. Figure 4 shows the edit mode screen. Commands are included for a comprehensive set of editing functions including generating arbitrary periodic sequences, moving and copying signals, and even producing random signals.

The third and final mode in the logic program, the "test" mode, is used to run the simulator and the tester. When the user enters the test mode, a consistency check is run on the current vector database to insure that the vectors represent a valid test. Among the things that are checked are multiple signals with the same name, signals with undefined bits in them (i.e., the user failed to enter a one or zero at some location), and that the signals are all of the same length. In each of these cases, logic lets the user decide if the problem should be fixed automatically. For example, if a particular signal has a greater bit length than the rest, logic can truncate that signal if desired. Once the consistency check has completed without errors, the test mode is brought into operation. Either the esim simulator or the DMFT may be run with the existing data from this mode. Additional data, however, are required to run either of these and the test mode offers the user the ability to merge the necessary data at this point.

If the user requests a run of the esim simulator, the logic program will ask the user for the name of a file set that contains the netlist information of the design to be simulated. Logic then checks this file set to be certain that all the signal names that have been defined in the current test vector set actually appear in the design. If they do not, this information is made available to the user. If they are, then esim is run and the results are displayed on the screen. Esim output consists of 1's, 0's, and X's. The X's represent unresolved states. Logic

will display the X's and allow them to be saved, read, or edited just as the other logical states.

If the user wants to run the DMFT, then information is required that tells the logic program which signals are wired to each channel (of the eight input and eight output channels) on the DMFT. This information can be created by invoking the "edit pin file" command or by reading it in from an existing file. Before the DMFT is actually run, a consistency check on the pin list and the signal set is performed to make sure that no more than eight inputs and eight outputs have been requested and that inputs are wired to input channels and outputs to output channels. Assuming that all is fine, then the tester is run. The results appear on the screen just as in the case of esim, with the exception that X's cannot occur.

If the user wants to run another test (either with esim or the tester), the outputs can be reset in the test mode, also.

VII. Conclusions and Future Work

The system for testing and simulation that has been described in this report has been operational since February of 1985. In the short time since the system has been in use, it has proven to be of essential importance to the microcircuit projects in the Advanced Systems Program. It was used to test the Reed-Solomon encoder chips in only two weeks time. These chips included two test chips and three versions of the complete system chip. The chips all worked as expected — but this would have been next to impossible to determine without the DMFT. One other chip, part of a Hopfield neural memory system [5] being developed on a Director's Discretionary Fund grant, was designed, simulated

using the logic program, and tested using the DMFT and logic, bringing to fruition the design and test methodology described in this report. The design and simulation of the chip took only two weeks and the testing was completed in just two days.

Some limitations of the system have already been discovered. Many engineers had suggestions for certain commands to be added to the logic program. These included the random signal generator and the undo command (which undoes the last edit operation). These have already been added to the program. Also, the need for external clock input protection circuitry for the DMFT was discovered and this too was added.

Among the development activities planned for the system are expanding the DMFT's capability in a second-generation device that will have many more input and output channels and automatic crossbar switching to replace the patch panel. Also, the logic program could be expanded to allow other simulators (such as "rsim" [6] or JPL in-house programs such as "ULYSES") to take advantage of the user-friendly interface.

The DMFT hardware is well documented. The documentation includes complete logic diagrams, a hardware manual, and a user manual. The DMFT can be duplicated for about $1400 including JPL engineering costs. The design is available to other parties at JPL who may wish to use it. The logic software package is written entirely in the "C" programming language, but it relies heavily on the UNIX operating system and the ANSI terminal standard. Those functions that are operating-system dependent are separated into a single C source module just as those for ANSI standard are. This means that the program might be rewritten to run on either a different operating system or terminal type by changing these files only. Logic, like the tester, is available to other parties at JPL.

Acknowledgments

References

1. Deutsch, L. J., "An Integrated UNIX-Based CAD System for the Design and Testing of Custom VLSI Chips," *TDA Progress Report 42-81*, Jet Propulsion Laboratory, Pasadena, California, May 15, 1985, pp. 51-62.

2. Truong, T. K., Deutsch, L. J., Reed, I. S., Hsu, I. S., Wang, K., and Yeh, C. S., "The VLSI Design of a Reed-Solomon Encoder Using Berlekamp's Bit Serial Algorithm," Proceedings of the Third Caltech Conference on VLSI, California Institute of Technology, Pasadena, California, 1983, pp. 303-330.

3. Mead, C. and Conway, L., *Introduction to VLSI Systems*, Addison-Wesley Publishing Company, Menlo Park, California, 1980, pp. 1-37.

4. Baker, C. and Terman, C., "Tools for Verifying Integrated Circuit Designs," *Lambda*, Fourth Quarter, 1980.

5. Hopfield, J. J., "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," Proceedings of the National Academy of Sciences, Vol. 79, April 1982, pp. 2554-2558.

6. Terman, C. J., "RSIM — A Logic-Level Timing Simulator," Proceedings of the IEEE Conference on Computer Design: VLSI in Computers, IEEE Computer Society Press, Silver Spring, Maryland, 1983.

Table 1. DMFT Specifications

Power requirements	120 VAC rms at 60 Hz and 2 Amps
Terminal connection	RS-232 standard, 9600 baud
Host computer connection	RS-232 standard, 9600 baud
Stimulus vector	8 user-definable logic bits wide, TTL logic levels
Response vector	8 bits wide, TTL logic levels
Power to device under test	1.5 to 7 Vdc at up to 100 mAmps
Minimum time between stimulus vectors or response vectors	50 nanoseconds
Maximum time between vectors	Infinite (dc)
Resolution of vector separation time	Infinite, continuously variable
Maximum number of vectors per test	8190 (4095 stimulus, 4095 response)
Minimum number of vectors per test	4 (2 stimulus, 2 response)
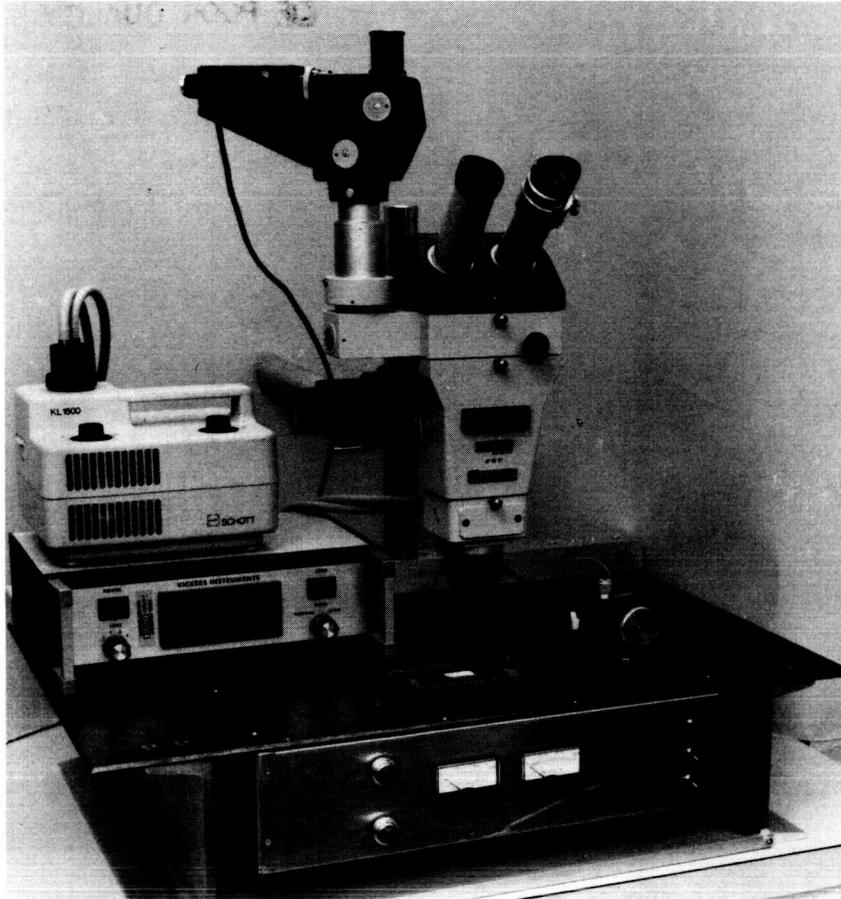
Fig. 1. The Digital Microcircuit Functionality Tester (DMFT) and microprobe station
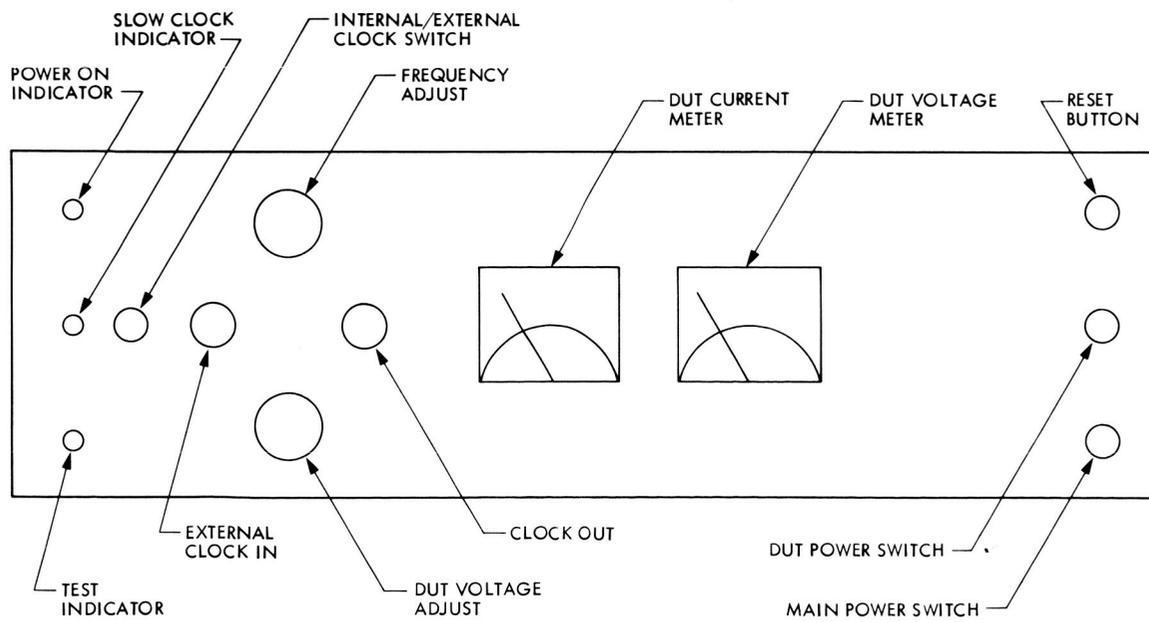
Fig. 2. Front panel controls and indicators on the DMFT
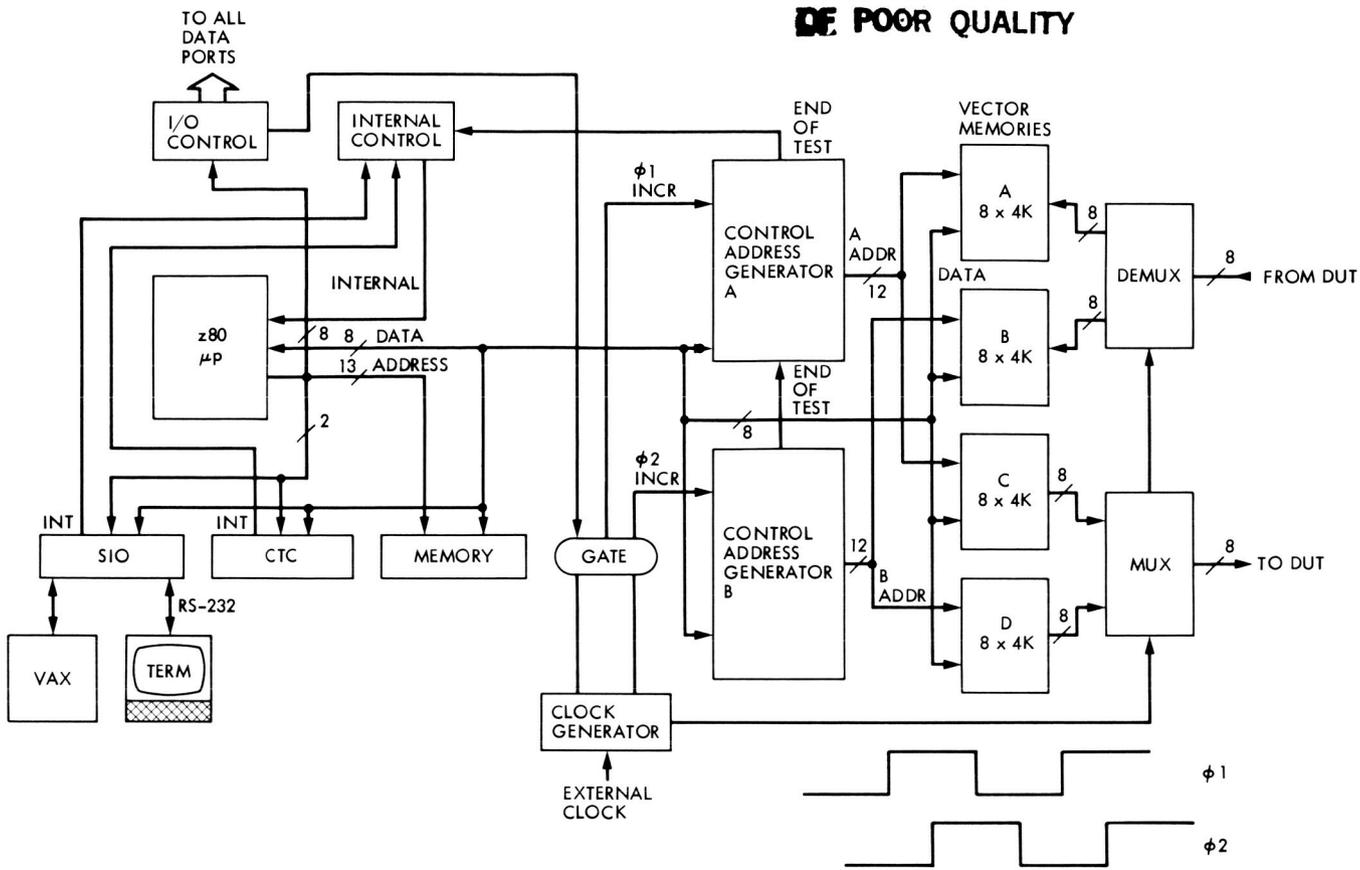
Fig. 3. System block diagram for the DMFT
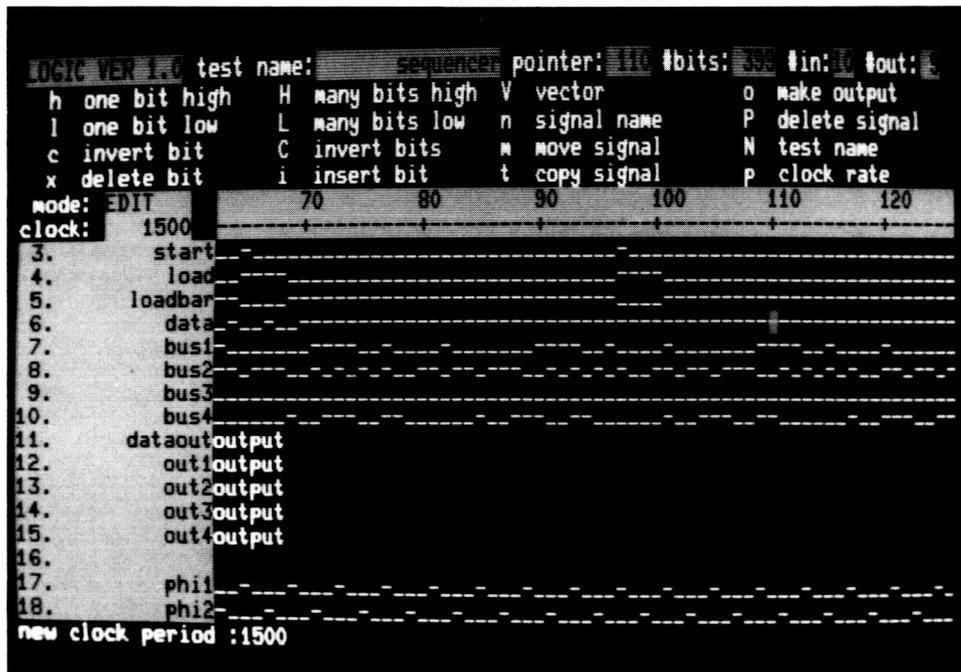
Fig. 4. The logic program display screen

Prototyping and Implementing Flight Qualifiable Semicustom CMOS *P*-Well Bulk Integrated Circuits in the JPL Environment

E. M. Olson

Communications Systems Research Section

Presently, there are many difficulties associated with implementing application specific custom or semi-custom (standard cell based) integrated circuits (ICs) into JPL flight projects. One of the primary difficulties is developing prototype semi-custom integrated circuits for use and evaluation in engineering prototype flight hardware. The prototype semi-custom ICs must be extremely cost-effective and yet still representative of flight qualifiable versions of the design. A second difficulty is encountered in the transport of the design from engineering prototype quality to flight quality. Normally, flight quality integrated circuits have stringent quality standards, must be radiation resistant and should consume minimal power. It is often not necessary or cost effective, however, to impose such stringent quality standards on engineering models developed for systems analysis in controlled lab environments. This article presents work originally initiated for ground based applications that also addresses these two problems. Furthermore, this article suggests a method that has been shown successful in prototyping flight quality semi-custom ICs through the Metal Oxide Semiconductor Implementation Service (MOSIS®) program run by the University of Southern California's Information Sciences Institute. The method presented has been used successfully to design and fabricate through the MOSIS three different semi-custom prototype CMOS p-well chips. The three designs make use of the work presented here and were designed consistent with design techniques and structures that are flight qualifiable, allowing one hour transfer of the design from engineering model status to flight qualifiable foundry-ready status through methods outlined in this article. The design techniques presented here that permit the flight qualifiable prototyping arose as a natural extension of other purely ground-based work that will also be described.

I. Introduction

Recently, during the course of developing a standard cell library for use in ground based applications for the Deep Space Network (DSN), a new method and set of standard cells was also developed, almost with no additional work, that addressed many of the flight quality prototyping issues that are critical to the use of custom integrated circuits in JPL flight projects. The general process of bringing a new standard cell family online for use in ground application microcircuit design at JPL is not unlike that process necessary to develop the capability to prototype flight quality custom ICs for flight projects. As one looks to the future of space investigation, it is clear that the widespread, successful use of

custom integrated circuits (ICs) in NASA flight projects and ground signal processors will hinge on the ability of NASA personnel to maintain full design cognizance of those circuits.

Although custom ICs have already been used in major flight projects at JPL, they have been used only on a very limited basis. Such limited use allows the electrical and physical environment to be precisely defined so that it is possible to make effective use of non-JPL personnel for custom IC design. The future economics of unmanned space exploration, however, point toward the use of multi-mission hardware and equipment.[1] The need to design multi-mission hardware in turn points toward an increasing use of custom and semi-custom integrated circuits, whose replication cost in time, money *and* mission risk is extremely low (Ref. 1). If whole flight subsystems are then designed around several custom integrated circuits, it will not be possible technically nor economically to maintain control and cognizance over flight hardware without involving JPL custom microcircuit designers directly.

Let us diverge for the next few paragraphs and investigate what might be needed to enhance JPL's ability to use custom microcircuits in flight projects. This short investigation will help us understand the parallels of ground based and flight quality prototyping IC development. In order to allow JPL personnel to effectively use custom and semi-custom ICs in flight quality hardware, three things are necessary:

(1) Engineering prototype chips must be available quickly, reliably and for extremely low cost in small quantities, and must reflect the flight quality chip speed performance, architecture, and testability.

(2) Transfer of the design from engineering prototype to flight quality foundry ready must be done without redesign or exposure of the design to operations that may affect chip size, pin-out, power consumption, floor plan or architecture. Nevertheless, the flight quality design must fully address all issues of reliability and radiation resistance.

(3) The flight quality design must be consistent with designs and foundries that have a proven success record and have demonstrated qualified custom ICs in the past.

Given that set of ground rules, if one were interested in real prototyping capability, it would then be necessary to choose a particular integrated circuit technology to further

focus the development of a method for prototyping flight quality custom ICs. For technical reasons that are well outlined in the literature (Ref. 2), and for its availability and history of use in qualified systems, one would probably choose bulk CMOS p-well for the design of custom integrated circuits in the near future. Bulk CMOS p-well offers extremely low power operation with wide noise margins and immunity to power supply fluctuation. Even more important for the economically strained flight project reality, bulk CMOS p-well is much cheaper to produce than other technologies that have the same or lower reliability and radiation resistance.

To further focus a hypothetical effort on developing technology to permit rapid and effective use of custom ICs in JPL flight projects, one would also consider the JPL design environment. In order to be effective, prototype flight quality custom ICs would have to be designed quickly, by designers with a firm knowledge of the target system's functional and timing constraints and some simple rules of reliable, testable and radiation resistant architectures. Those designers could be expected to have little detailed IC design expertise or detailed knowledge of silicon structures and device physics. That typical designer profile would point the way to implementing custom integrated circuits with standard cells rather than fully hand crafted structures. Standard cells provide nearly the density of hand crafting and yet free the designer from detailed and esoteric analysis that is necessary to properly build digital logic gate structures in silicon for space flight and prototype custom ICs. Further gains on the density of hand crafting could be had by hand placing standard cells in designs that must be dense, as opposed to allowing placement of cells by automatic computer tools.

Up to now, we have investigated in general terms what might happen if there were an effort to develop a comprehensive capability to design and implement custom microcircuits in JPL flight projects. We can, however, continue the conjecture in more detail. The directives outlined above indicate the following possible approach to achieving the ability to quickly and accurately obtain engineering prototype custom ICs for flight quality prototype hardware:

(1) Search for a CMOS p-well bulk standard cell family that has an outstanding flight quality performance history when fabricated through the supporting foundry.

(2) If such a standard cell family is found, determine what modifications would be necessary to fabricate the cells through the inexpensive, readily available MOSIS foundry service, with the understanding that the MOSIS design rules must be met in order to get typical MOSIS yield and performance. It is assumed that the MOSIS performance levels are adequate for

[1]Jordan, J. J., Getaway Special Project Manager, Physical Science Laboratory, Private Communication, June 1984.

engineering prototypes. Also determine if interconnection structures designed to the MOSIS design rules would meet the design rules of the original foundry. If so, then proceed.

(3) If the individual standard cells from the selected family can be modified under the constraints outlined below then do so; verify (Ref. 3) the results by fabricating through the MOSIS.

 (a) Modified cells shall be of the same bounding size, port placement and power bussing as the original cells. The periphery and connectivity of the modified cells will be identical to the original cells.

 (b) Modified cells shall have the same transistor structures and drive capability (size) as the original cells in order to maintain similar performance between original and modified cells.

 (c) Modified cells must not contain any structures that would result in MOSIS or original foundry design rule violations with any possible adjacent structures.

 (d) No modifications to the original cells can be made.

(4) If the verification-through-MOSIS results from Step 3 are positive, then develop a method for designing with modified cells and then substituting original cells without exposure of the design to normal modification and design operations that could introduce human error.

(5) Verify with computer aided analysis that designs done to MOSIS design rules, with original cells subsequently substituted for modified cells, do in fact meet the design rules of the original foundry.

(6) Verify the design sequence by fabricating the same design through the MOSIS *and* the original foundry, using appropriate cells for each, and compare performance.

(7) If performance comparisons are favorable then release, with appropriate legal agreements, two sets of standard cells of identical function and port placement and of similar performance. One set of modified cells is suitable for fabrication through the MOSIS, and the original untouched library of cells can be fabricated through the originating, high reliability, flight qualifiable foundry. Also release the means for substituting original cells for modified cells without affecting the design floor plan or architecture.

We have now investigated to a detailed planning level what would be involved in developing prototyping capability for flight project custom microcircuits at JPL. Although the above approach may seem special, perhaps peculiar only to the specific interests of flight projects, the approach outlined above has only minor differences with the work that is necessary to bring any typical vendor's standard cell family to JPL for use in ground applications. In fact, in the course of bringing on-line a standard cell family for ground based applications, all items in the list above have been completed up to, but not including, Step 6.

A standard cell family was found that looked suitable for use in DSN applications and it was decided to proceed with the necessary modifications to make the cell library more widely useful at JPL, thus spreading future maintenance costs over more users. It typically takes 2 to 3 years to design and verify an entire cell library from scratch, so it was determined that a suitable alternative would be to modify an existing cell library for use within the existing JPL design framework. Those modifications typically amount to changes to the mask layer names, changes to mask geometry to meet the MOSIS design rules and then subsequent verification through the MOSIS.

The standard cell family chosen to bring to JPL was Sandia National Laboratories' $4/3\mu$ CMOS p-well Bulk radiation hardened cell library (Ref. 4). The library was chosen for its completeness, its ease of use, easily available documentation and the fact that minimal modifications were needed in order to use the cell library with the MOSIS for ground based applications. Furthermore, since Sandia is also a government agency, there were no legal difficulties or licensing fees needed to get the cell library information from Sandia. Sandia's excellent reputation and historical record of producing flight quality and mil-spec semi-custom ICs for military and spaceflight applications was also intriguing, but not of primary concern at the time. Of course, it was realized that those structures in the Sandia cells that were necessary for flight quality results would not and could not be fabricated through MOSIS. That fact was irrelevant, though, because the new modified cell library was targeted only for ground applications.

As work began in modifying and bringing the Sandia cell library on-line for JPL ground based use and MOSIS fabrication, little consideration was given to maintaining any compatibility with the original Sandia cells. It soon became apparent, though, that if slightly stricter modification guidelines were used, namely, those numbered 3 through 7 in the approach outlined above, not only would a useful cell library for ground based work be obtained, as was originally desired, but also, a set of standard cells that permitted quick prototyping of flight quality custom ICs could be had. It was clear that the items 1 and 2 in the approach outlined above

had already been met simply by our choice of the original Sandia cell family.

The more stringent guidelines were, in fact, adhered to, and the same cell library can now be used for ground based and flight prototyping applications. The rest of this article will outline a summary of the problems, methods and analysis of reconstructing Sandia cells to MOSIS standards guided by the constraints outlined above, concentrating on the more interesting flight prototyping characteristics that came out of this work. The article will also address the problems and methods of implementing modified Sandia cells, hereafter referred to as MOSIS cells,[2] in designs. Results and limited analysis of MOSIS fabrication verification will also be presented.

II. Design Rule Comparison

Table 1 lists important differences between the MOSIS design rules for 3μ bulk CMOS p-well (Ref. 5) and Sandia's $4/3\mu$ bulk CMOS p-well radiation hardened process (Ref. 6). The most important differences that affect the modification of the cells are first level metal minimum width and contact overlap requirements. Although the MOSIS requires much more contact overlap than Sandia, some help in cell internal modification space availability is lent by the 1μ smaller minimum metal width for the MOSIS. One can see by the poly and metal contact and poly and metal spacing and width rules that interconnection structures satisfying the MOSIS design rules will also satisfy Sandia's design rules. That fact is of critical importance for avoiding two design iterations in the advancement from prototype to flight quality.

Table 2 lists the layers that are used by the MOSIS (Ref. 5) and Sandia (Ref. 6) in their respective CMOS p-well processes. This table shows radical differences between the specification of diffusion masks. The diffusion mask specification differences have a profound effect on many of the modifications and performance of the MOSIS cell based designs and will be discussed later. Two additional layers that are present in the Sandia line that are not available through the MOSIS are the P+ guard ring and the V+ threshold adjust implant. The guard ring and the threshold adjust are included in Sandia's line to combat latch-up and body-effect threshold degradation in radiation environments and thus are unnecessary in the prototype MOSIS cells.

The only layers available through the MOSIS that are not available through Sandia are the second level metal and con-

[2]Not to be confused with the MOSIS Charger cell library released by MOSIS. In the context of this article, the term "MOSIS cells" refers to Sandia cells modified for fabrication through MOSIS.

tact cut to first level metal. The use of second metal interconnections in MOSIS designs can offer a speed improvement of up to a factor of 3 over designs using polysilicon for interconnections. Such a speed improvement is not representative of Sandia fabrication since second level metal is not available on Sandia's $4/3\mu$ line. Therefore, interconnections using second metal are not permitted in prototype designs. If second metal were used in prototype designs, transfer of the design to Sandia for fabrication would require redesign of the interconnections, which violates one of the ground rules of the prototyping endeavor. Second metal must be used, however, in the modified Sandia bonding pad cells at the bonding pads. If second metal is not present at the bonding pads, the MOSIS bonding operation will attempt to wire bond to the insulating material between metal one and metal two and an unsatisfactory connection will result.

III. Modification Synopsis

Figure 1 shows an unmodified Sandia cell with Sandia layers. Table 3 shows the correspondence between the layer names and the plot names. The n channel field effect transistors (nFETs) are contained in the guard ring at the bottom and the p channel FETs are along the top. The power bus is also at the top and ground is at the bottom. The polysilicon gates are 3μ long and the gate width direction runs vertically. Figure 2 shows the same cell modified for fabrication through the MOSIS. Layers have been deleted and modified as necessary to meet the MOSIS design rules both internally and to possible adjacent structures. The modified cell still maintains the original functionality, periphery physical characteristics and approximate transistor sizes as the cell shown in Fig. 1.

Since Sandia makes use of a deep P+ guard ring surrounding the p-well, it is permitted in the Sandia cell to cross the p-well boundary with thin oxide. The width of the Sandia cell nFETs is then controlled by the height of the $n+$ implant mask. (Note that Sandia actually implants source and drain regions in order to avoid the high temperature processing required for diffusion. High temperature processing undermines the radiation resistant characteristics built in prior processing steps.) Since the MOSIS supports neither a guard ring nor an explicit definition of the $n+$ diffusion mask, the MOSIS version of the cell must be modified to reflect that fact.

First, the thin oxide edge must be brought inside the p-well boundary in the nFET locale. Since all thin oxide in MOSIS designs is either diffused $p+$ or $n+$, a shorting forward biased condition between the grounded p-well and the biased substrate outside the p-well would exist if the thin oxide were permitted to cross the p-well and subsequently allow $n+$ diffusion under its area.

Second, the width of the nFETs must be controlled by the thin oxide mask rather than by the unavailable n+ diffusion mask. Figure 3 shows the modifications in the nFET locale from the Sandia cell on the left to the modified MOSIS compatible cell on the right.

One may also note by comparing Figs. 1 and 2 that the MOSIS cell has substantially more contact overlap of the contacting layers than the original Sandia cell. This is necessary to meet the MOSIS design rules. Although intuition might lead one to believe that the oversized looking metal lines in the MOSIS cell would increase node capacitance significantly, the typical increase in capacitance due to increased metal area around the contacts is not substantial enough to be of interest. SPICE (Ref. 7) simulations of the MOSIS and Sandia cells show that the electrical circuit is essentially unaffected by the slight differences in node capacitance between the Sandia cell and its MOSIS counterpart.

Cells were modified using Caesar for geometrical manipulation and Lyra for design rule checking to MOSIS specifications (Ref. 8). It was necessary in special situations to push some conservative minimums by $1/2\mu$ in order that the modified cell maintain the same bounding dimensions as the original Sandia cell. The two design rules that were violated approximately 3% of the time were the first metal to first metal spacing (was made 3.5μ instead of 4μ) and the contact to channel spacing (was made 2.5μ instead of 3μ). Since the MOSIS design rules are a conservative superset of design rules for many foundries, it was estimated that pushing the design rule limits in such a manner would have minimal effect on yield. To date, none of the 72 chips returned from the MOSIS have shown any indication of poor performance or failure due to these design rule violations.

IV. Method of Transferring From Prototype to Flight Quality

Given that now two sets of outwardly identical cells exist, it is relatively simple to transfer a design consisting of interconnections and MOSIS cells to one consisting of the same interconnections and Sandia cells. The transfer assumes that the MOSIS prototype design has accounted for architecture and timing requirements appropriate to flight quality designs.[3] Again, a constraint is placed on the interconnections to not employ second level metal.

The use of the Caltech Intermediate Format (CIF) as the design database provides a quick means through symbol calls

_____

[3]A survey and analysis of specific architecture and timing characteristics desirable for achieving flight qualification is beyond the scope of this article.

to effect the desired cell switch (Ref. 9). The entire design is completed with MOSIS cells and then transferred with the cell calls only (not the cell symbol definitions) to another environment where the cell symbols are defined as Sandia cells. Care is taken by the design tool maintainer to insure that CIF layer names are consistent throughout Sandia and MOSIS representations. A typical design sequence might proceed as follows on a UNIX® system using the Berkeley Computer Aided Design tool set:

(1) The designer constructs the design with his choice of tools and the MOSIS cells. The designer is only constrained at this point to produce a physical representation of the design in CIF that is constructed using MOSIS cells and design rules. Such construction might be accomplished by using Caesar or Magic (Ref. 10) to construct the design in a UNIX directory that only contained MOSIS cells with Caesar's or Magic's path then set to that working directory.

(2) The design is design rule checked and sent in CIF format to the MOSIS for fabrication (Ref. 11).

(3) The design is transferred into a new UNIX directory that contains all of the information that existed in the original directory except for the representations of the MOSIS cells. This is done with a UNIX copy command. The Caesar or Magic path is set to the new directory and the original Sandia cells are copied from a master directory into the new directory. It is critical that the MOSIS cells have the identical call names as their Sandia counterparts.

(4) The design is write change protected and brought up for read only editing by Caesar or Magic in the new directory. The tools automatically invoke the Sandia cells where the MOSIS cells were previously. The design is written out with a new name and is immediately write protected.

(5) The design is design rule checked and sent to Sandia or a suitable second source in CIF format, or in the more industry accepted GDSII format, for mask making and fabrication.

V. MOSIS Verification and Performance Results

To date, three different designs have been fabricated through the MOSIS. The first consisted of large sections of shift registers and test cells. The fabricated chips were tested and found functional but with higher operating current and slower speed than expected, based on simulations. A problem in substrate biasing and p-well isolation within the cells was inferred and modifications were made to adjust diffusion

masks to provide better connectivity and correct the biasing problems.

Those modifications proved helpful to the second design, which was actually very similar to the first. The second design exhibited normal speed and only slightly increased power consumption over normal. Further modifications to diffusion masks were made and a third design, much different than the first two, was completed with the modified cells and sent to the MOSIS for fabrication.

The third MOSIS design exhibited much higher speeds than worst case simulations. This was partly due to careful hand calculations of buffering and loading in the circuit architecture and partly due to the success of the modifications from the first two designs. A third factor relating to the polysilicon may have also been a factor and is discussed below. The third design was fully static and functioned at clock rates from DC to 22 MHz. Microprobing of individual gates indicated loaded gate delays on the order of 3 to 5 nanoseconds. Unfortunately, there was still a problem with excessive power dissipation.

The higher than expected speed exhibited by the third design, which was an array of quasi-synchronous counters, may be attributed in part to the $n+$ diffused polysilicon used for interconnections in the MOSIS fabrication. Such polysilicon would not be expected in a typical Sandia fabrication. Since the designer only specifies the $p+$ diffusion mask to the MOSIS and the MOSIS generates the $n+$ diffusion mask as the complement of the $p+$ mask, most of the polysilicon interconnections in the MOSIS design are exposed to the $n+$ diffusion operation. Since diffused polysilicon has a much lower resistivity than implanted or *in situ* doped polysilicon (Ref. 12), circuit interconnection RC time constants are somewhat smaller in designs with diffused polysilicon interconnections. Sandia generates $n+$ and $p+$ regions with implanting techniques rather than diffusion techniques and furthermore does not allow implanting of most of the polysilicon used for interconnections. More investigation of the exact differences between the resistivities of Sandia's typical polysilicon and the MOSIS typical polysilicon is needed.

Unfortunately, differences in polysilicon interconnection resistivity cannot come close to accounting for the high speeds. Other factors also contributing to high operational speed of the third design include maximum use of metal for interconnections, carefully analyzed buffering of critical signals (mentioned above), careful hand placement of cells and manual floor planning that minimized interconnection length. The P-well doping concentrations may also be a major factor in the high observed speed.

The increased power consumption in the third MOSIS design may be at least in part attributed to the inability of the designer to explicitly specify the $n+$ diffusion mask to the MOSIS. Since the designer cannot explicitly specify the $n+$ diffusion mask, the designer has very little control over the pad protection diode junction characteristics. Initial investigation[4] indicates that high reverse currents may be present between supply rails in pad protection diodes (see Figure 4). The Sandia cell protection diodes, in fact CMOS diode structures in general, do not lend themselves well to obtaining good performance without the ability to specify both the $n+$ and $p+$ diffusion masks explicitly. Further testing is necessary to verify that the excess currents are, in fact, flowing through the pad protection diodes.

The third MOSIS design also contained many individual cells with probe pads for individual testing. Three different sizes of inverter logic and buffer cells were tested for electrically induced latch-up susceptibility, with favorable results. The inverters were powered at 3 volts and driven to a -2 and +5 volt logic signal input without latching up. Next the inverters were powered at 5 volts and subject to a -5 to +10 volt logic signal at 1 MHz. Again there was no latch-up. Finally the inverters were powered at 6 volts and driven with a -5 to +10 volt logic signal at 1 MHz that had an additional 3 volt transient spike induced on both the rising and falling edges. The inverters still did not latch up. There was no purposeful on-chip signal clamping or special probe pad structures used in conjunction with these tests. The tests indicated excellent localized latch-up resistance for the MOSIS cells. A test was not devised to observe chip-wide latch-up susceptibility.

In summary, designs fabricated through the MOSIS with the modified Sandia cells have exhibited very good performance and robustness. Although the MOSIS designs have not been tested at 10 volts (a typical operating voltage for a Sandia chip) because they were fabricated by a 5 volt line. There are no expected problems concerning supply voltage in transferring from MOSIS to Sandia.

One possibly serious problem that may arise, though, is in attempting to operate with high supply voltages in the other direction. That is, attempting to operate MOSIS prototype designs at the same supply voltage that Sandia designs will operate at comfortably. The MOSIS bases most of its fabrication parameters on up to 6 volt operating supply, while Sandia typically fabricates for up to 10 volts and beyond. The problem would appear when operating a prototype MOSIS design in hardware running at typical flight system supply voltages of 10 volts and greater. For flight systems employing 5 volt supply voltages, no prototyping problems are expected.

---

[4]Shafer, B. D., Sandia National Laboratories. Private Communication, May 1986.

VI. Conclusions and Future Work

A method and set of standard cells has been developed that permits not only quick ground application microcircuit development, but also easy and extremely high confidence prototyping of flight quality standard cell based custom microcircuits. The flight prototyping method has been tested up to actual fabrication by the flight quality foundry. Despite any differences in the electrical performance of the prototype designs compared to typical flight quality performance, a high degree of confidence in the design can be obtained for very low cost before the decision is made to include a custom microcircuit in a flight system. Also, it is possible to evaluate architecture and timing constraints with real hardware in the lab by prototyping potential flight quality custom microcircuits through MOSIS. The job of prototyping and obtaining flight qualifiable chips has been reduced to a single design iteration with a simple "human-free" substitution method for going from prototype to flight quality foundry-ready.

But not to be overshadowed by flight applications, a reasonable general purpose standard cell library for routine use with the MOSIS has also been generated. Routine use of this new cell library has the advantage of possibly permitting a quick development of flight quality chips for designs that may not necessarily have been originally intended to fly.

Future work will include more definitive analysis and characterization of electrical parametric operating differences between MOSIS and Sandia fabrications. Work is presently proceeding, under other funding, on an actual flight project prototype design for Mariner Mark II. The prototype chip is a digital filter containing adders, multipliers and other functions and is built entirely from modified Sandia cells. The prototype is in fabrication with the MOSIS at the time of this printing. It is hoped that the circumstances will present an opportunity to fabricate the prototype through a flight quality foundry after substitution of the flight quality cells. Such an opportunity could further verify the work and methods introduced here.

Future work will also include analysis of Sandia's new 2μ CMOS p-well, double metal cell library. Initial surveys[5] indicate that the new radiation hardened 2μ library will be capable of direct fabrication through MOSIS without modification. Another good ground based cell library would be had and radiation hardened performance would then become purely a product of the fabrication line rather than both the cell structure and the line.

[5]Bair, R. E., Barnard, W. J., Shafer, B. D. (of Sandia National Laboratories), Olson, E. M. (of JPL) and Steelman, J. E. (of New Mexico State University), Technical Discussions at Sandia National Laboratories, Albuquerque, New Mexico, 1985.

Acknowledgments

References

1. Millman, J., *Microelectronics* (Columbia University, McGraw-Hill Book Company, 1979), pp. 92-93.

2. Weste, N. and Eshraghian, K., *Principles of CMOS VLSI Design* (Bell Laboratories, Addison-Wesley Publishers, 1985), 531 pp.

3. Olson, E. M. and Deutsch, L. J., "A System for the Functional Testing and Simulation of Custom and Semicustom VLSI Chips," *TDA Progress Report 42-86*, Jet Propulsion Laboratory, Pasadena, California, 1986.

4. Flores, R. W., *ELA Standard Cell Users' Guide 4/3 Micron Design Rules Radiation-Hardened Silicon Gate CMOS*, Sandia Report Number SAND83-0184 (Revised), Sandia National Laboratories, Albuquerque, New Mexico, February 1985.

5. MOSIS Project, *3 Micron CMOS P-Well Design Rules*, Version 1.1, USC/ISI, Marina del Rey, California, April 1985.

6. Babicz, S. J., *Design Information for the 4/3 Silicon-Gate CMOS Technology*, Sandia Report Number SAND85-0052, Sandia National Laboratories, Albuquerque, New Mexico, June 1985.

7. Negal, L. W. and Pederson, D. O., *SPICE—Simulation Program with Integrated Circuit Emphasis*, Memorandum Number ERL-M382. Electronics Research Laboratory, University of California, Berkeley, California, April 1973.

8. Ousterhout, J. K., *Editing VLSI Circuits with Caesar*, Documentation of 1983 VLSI Tools, Computer Science Division, University of California, Berkeley, California, 1983.

9. Hon, R. W. and Sequin, C. H., "A CIF Primer," *A Guide to LSI Implementation*, Second Edition, Xerox Palo Alto Research Center, Palo Alto, California, January 1980, pp. 79-123.

10. Scott, W. S., Mayo, R. N., Hamachi, G. and Ousterhout, J. K., *1986 VLSI Tools: Still More Works by the Original Artists*, Report No. UCB/CSD 86/272, Computer Science Division, University of California, Berkeley, California, December 1985.

11. MOSIS Project, *The Mosis System (what it is and how to use it)*, Report Number ISI/TM-84-128, USC/ISI, Marina del Rey, California, 1984.

12. Sze, S. M., *VLSI Technology* (Bell Laboratories, McGraw-Hill Book Company, 1983), pp. 103-106.

Table 1. Important design rules for cell modifications

Design Rule	Sandia Rule, μ	MOSIS Rule, μ
Min. metal line	4	3
Min. metal space	4	4
Min. poly line	3	3
Min. poly space	4	4
Min contact to gate	2.5	3
Metal overlap of contact	0.5	2

Table 2. Layer sets for Sandia and MOSIS

Layer	Sandia	MOSIS
Polysilicon	Yes	Yes
Thin Oxide	Yes	Yes
Threshold Adjust	Yes	No Access
Metal One	Yes	Yes
Metal Two	No	Yes
Contact	Yes	Yes
Via (m1-m2)	No	Yes
P-Well	Yes	Yes
P+	Yes, implanted	Yes, diffused
N+	Yes, implanted	No, diffused, logical not of P+
P+ Guard Ring	Yes	No

Table 3. List of layer names and plotted names for Figs. 1, 2 and 3

Layer Name	Plot Name
Polysilicon	CP
Thin Oxide	CA
Threshold Adjust	CX
Metal One	CM
P-Well	CW
P+	CS
Contact	CC
N+	CT
P+ Guard Ring	CG

LAYERS
PLOTTED

CP

CA

CX

CM

CW

CS

CC

CT

CG
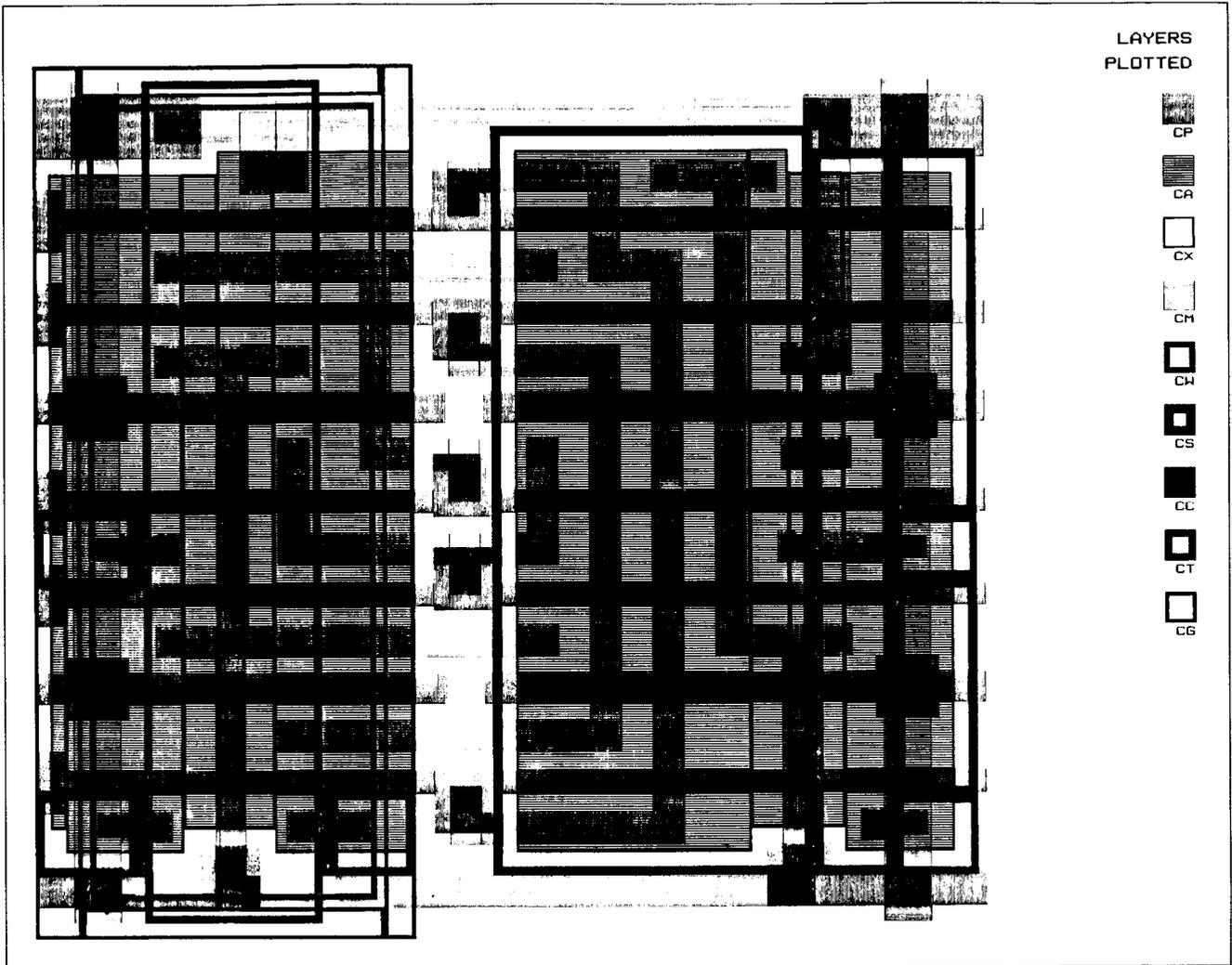
Fig. 1. An original Sandia cell showing the guard ring and overlap of *p*-well edge by thin-oxide
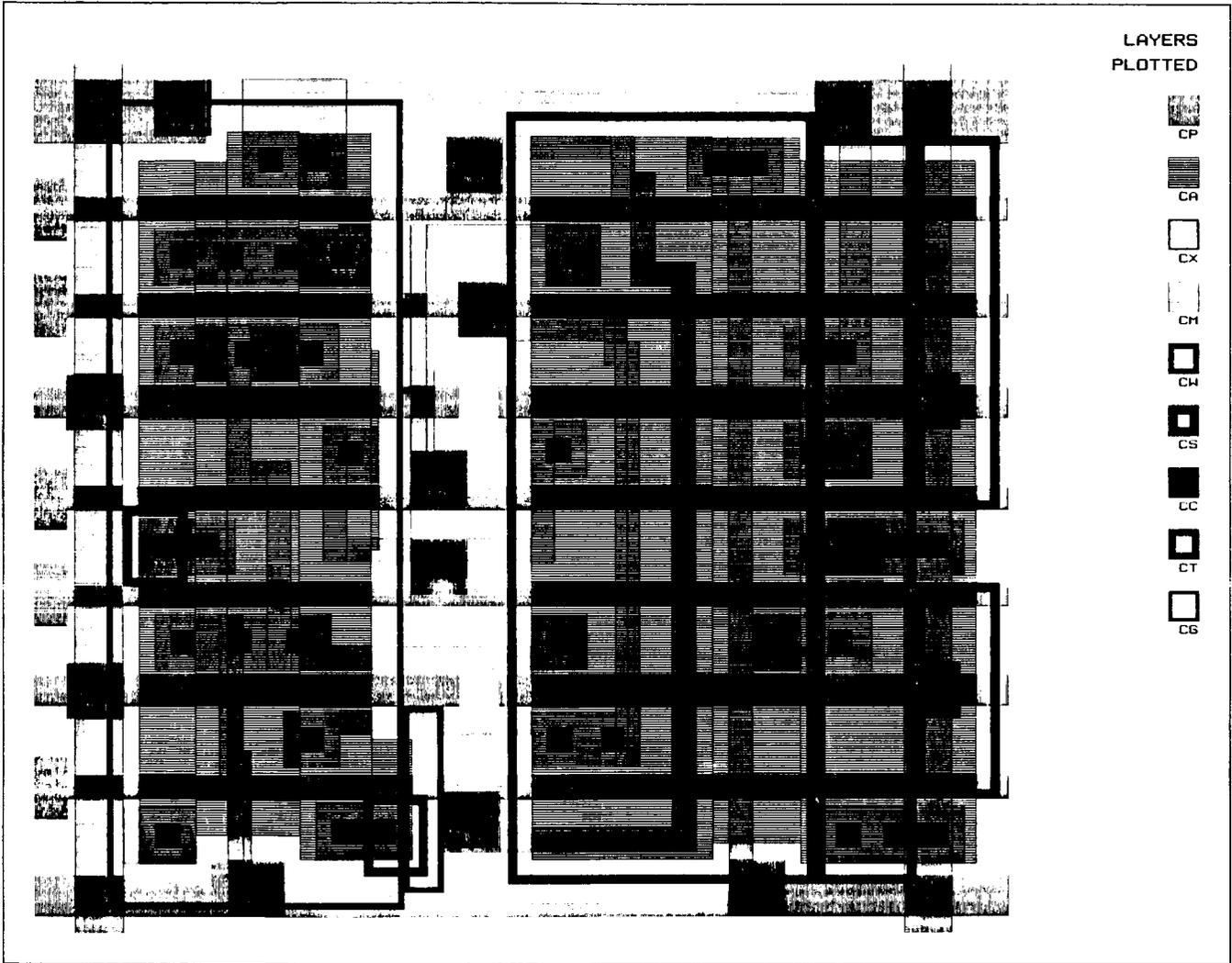
Fig. 2. A modified Sandia cell suitable for fabrication through the MOSIS

LAYERS
PLOTTED

CP

CA

CX

CM

CW

CS

CC

CT

CG
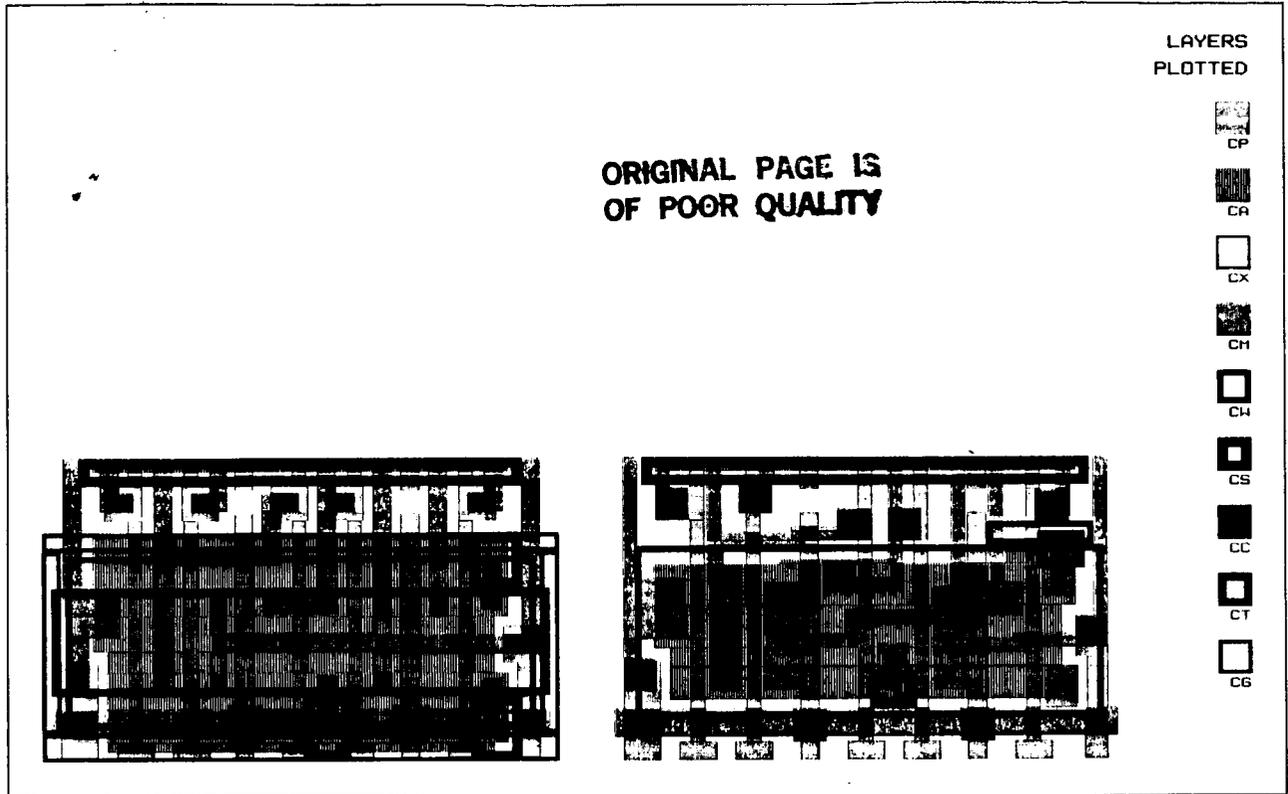
Fig. 3. A closeup comparison of the nFET regions of an original Sandia cell (left) and a MOSIS or modified Sandia cell (right). The tops of the cells have been chopped off for illustrative purposes only. Note that the modified cell's nFET gate width is slightly larger than the original cell's in order to accommodate some necessary contacts.
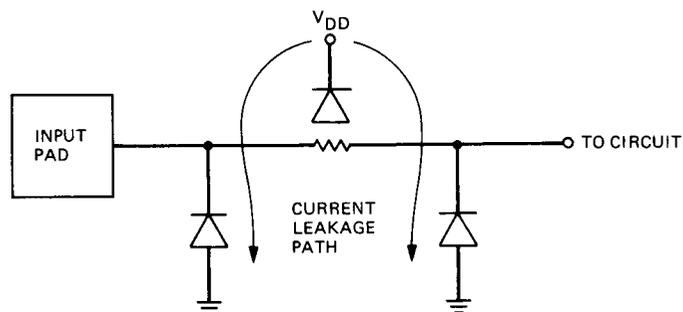
Fig. 4. Schematic of input pad circuitry showing possible path of high power supply currents due to input protection diode leakage

Ulysses, a Functional Description and Simulation Software System

T. W. Griswold and D. F. Hendry

Microelectronics Technology Section

Current design tools for digital circuits and systems are not well-integrated among the behavioral, gate, and transistor levels of design. Ulysses is a prototype software system that consists of a description language, a description compiler, and a simulator that make no distinction among these levels. The language is uniform over the entire range of logical descriptions, the description is hierarchical with no fundamental restrictions on depth or mixing of levels, and the simulator is fully integrated with the description. The structure of the language, compiler, and simulator are described in terms of their relationships to the abstractions of physical systems that are made in order to create logical descriptions and models of behavior.

I. Introduction

Design and implementation of a custom microcircuit involves several levels of design and several technical disciplines. Currently available commercial design tools do not provide good interfaces among these levels and disciplines. In particular, the distinction between "logical" and "behavioral" levels of description and simulation is sharp: The logical level is handled by pre-defined and pre-compiled internal models of logic modules, such as gates and flip-flops; the behavioral level is handled by user-written procedures in some language (such as PASCAL or C) that are compiled and linked to the system. The user does not have direct access to the internal data structures. As a result, the mixing of descriptions and simulations at the logical and behavioral levels is limited.

The Ulysses system is a set of computer programs that provides a uniform language for description and simulation of digital systems from the transistor-switch level through the gate level to the functional-block level. It is fully hierarchical, and descriptions and simulations at all levels may be mixed without restriction. A basic concept is that all logical, functional, and behavioral definitions are made in the same way, namely by means of truth tables created by the user. The user thus has full access to all internal data structures, and has full control of the system (he also has full responsibility for the correctness of all descriptions).

Ulysses is, in effect, an algebra of logical behavior. It provides a set of operators, function types, and data types that are used to form descriptions of logical systems of arbitrary size and complexity, and to evaluate those descriptions (i.e., to simulate the behavior of physical systems). The limit on the size of the logic system that can be described and simulated is set by the size and speed of the host machine, not by any internal limitation imposed by Ulysses itself.

Ulysses was initiated as part of the Silicon Structures Project (SSP) of the Computer Science Department of the California Institute of Technology. It was developed to its present state at JPL, as part of a development program for VLSI design tools. It is written in MINT (Ref. 1), which is a language that was designed for portability (i.e., machine independence). MINT is defined in terms of a "virtual machine," which contains a set of primitive functions that provide the actual interface between the MINT language and the host-machine hardware and operating system. Successful porting of the virtual machine to a new host machine guarantees that any program written in MINT (e.g., Ulysses) will run correctly on the new machine. (The MINT system includes a diagnostic program for verification of correct operation of the virtual-machine primitives.) MINT has been ported at various times and by various organizations to a number of machines, including Apple II, IBM PC-XT and PC-AT under MS-DOS, 68000-based machines running UNIX-like operating systems, VAX 11/780 running VMS, and Univac 1100. The basic system at JPL used for development of MINT and Ulysses is VAX 11/780 under VMS, with the virtual machine written in C.

II. Logical Representation of Physical Systems

A. Levels of Abstraction

Description and simulation of digital circuits at the logic level, that is, with signal values low and high, is an approximation to physical reality. The approximation is obtained by a sequence of abstractions. The first is from physical structures to lumped circuit elements. Structures created by doped semiconductor regions, insulating layers, and conducting layers are represented by device models: diodes, transistors, resistors, capacitors, wires, and contacts. In addition to the devices created explicitly for the desired circuit structure, second-order devices, such as series resistances, stray capacitances, and parasitic diodes and transistors coupled through the substrate, must be included in the circuit representation for good accuracy. The popular electrical simulation program SPICE is designed to help extract these models from a geometrical layout and to analyze the electrical circuit constructed from them.

The second abstraction is from the electrical representation to a switch-level representation. Transistors are replaced by switches that are either conducting or non-conducting, depending on their gate voltages. At this point, the circuit is still electrical, and it can be analyzed by SPICE if the switches are represented by relays. Signals are represented by amperes and volts.

The third abstraction replaces wire voltages with logic levels: If the voltage is above a certain threshold, the logic value is high; if the voltage is below another threshold, the logic value is low. This switch-level logical representation is the one that is used most of the time in design and manual checking of digital transistor circuits. The principle is simple: When a transistor is off, it has a high resistance; when it is on, it has a low resistance. An N-channel transistor is on when its gate is high; a P-channel transistor is on when its gate is low. When an electrical value is needed, such as the ratio of current to load capacitance for calculation of voltage slew rate, the electrical representation is immediately available; the circuit diagrams are essentially the same.

The fourth abstraction is from switch level to gate level: The switch-level circuit is divided into blocks, and the blocks are replaced by logical modules that are represented by truth tables and propagation-delay values. The delay values are obtained by calculations and summations of internal voltage slew rates at the switch or electrical levels, or by measurement.

In all of the abstractions described above, the description of the circuit is directly related to the way in which it is built out of transistors and other components. Higher-level abstractions depart from the structural description level, and take on a flavor of behavior: They describe what the circuit does, rather than how it is structured. An ALU, for instance, can be constructed in a number of ways and still perform the same computations. Complex circuits are generally defined and developed at the behavioral level in terms of data objects, modules with specified computational functions, and interconnections. They are then expanded hierarchically down to the gate and transistor levels in specific implementations.

It would obviously be of great value to have a single language for as many of these levels of abstraction as possible. Ulysses was designed to cover the range from behavioral descriptions to transistor-switch diagrams, that is, the entire range in which logical description in terms of high and low signal values is applicable. It is based on a relatively small set of operators that handle data structures of arbitrary complexity.

B. Logic Description and Simulation

The key to digital-description systems is the way in which electrical behavior is abstracted to form a logical description. In general, node voltages are represented as the logical values true and false. In positive logic, true is denoted variously by T, 1, or H (high), and false by F, 0, or L (low). High means that the node voltage is greater than some threshold value, and low means that it is lower than some other threshold value. Voltages in an intermediate state, below the upper threshold and above the lower threshold, are treated as part of a transitional state of vanishingly small duration. The transi-

tional state is represented as an "edge," either rising (R), or falling (F). Representation of edges is important in descriptions of synchronous digital circuits, because edge-triggered flip-flops are used to control system timing.

Logic modules (gates, flip-flops, microprocessors, and memories) are abstracted to functions with input and output argument lists. The input arguments correspond to signals connected to input ports, and the output arguments correspond to signals connected to output ports. A function drives its output signals to values determined by the values of its input signals and the behavior of the function, just as an algebraic function does. Unlike ordinary algebraic functions, however, logical functions that describe circuit behavior must incorporate the notion of propagation delay: Output value changes are delayed in time with respect to input value changes. The logic function must return the delay time for each output signal, together with its new value.

Simulation of the behavior of such a system is carried out by assigning logical values to the inputs to the system, and executing the functions as dictated by the connectivity of the system. The signal values at the outputs of the system describe its behavior. The connectivity, in effect, is described by the input/output structure of the functions: The outputs of one function are the inputs of another, corresponding to the way in which the wires of the hardware implementation are connected between the functional modules. Wires and module ports that are connected together form a circuit node; a logical signal value is assigned to represent the voltage of the node.

A logical abstraction to a two-level (binary) representation cannot deal properly with systems in which a node voltage is in the intermediate state for any length of time, that is, the voltage lies between the lower and upper thresholds for a long time. Such a condition may arise, for example, when two modules try to drive a signal at the same time. If one is driving it high and the other driving it low, the resultant voltage can have any value; the state is not known. The concept of "strength" is introduced to deal with such situations: A signal of greater strength always wins in a contest with a signal of lesser strength. This approach is used in wired-or and tri-state constructs, in which a pull-up resistor (strength = weak) keeps a node high unless one or more drivers (strength = strong) is turned on to pull the node low. When two drivers of equal strength try to pull the node in opposite directions, the result is represented by U or X, for undefined. The notion of an open circuit, or high impedance, is closely related to wired-or and tri-state structures: A function is either driving a node or it is turned off. The open-circuit condition may be represented either by a high-impedance value Z, or by zero strength.

Modules represented by truth tables have definite input and output ports. They are unidirectional, and their output values can be looked up in a table when their input values are known. The input and output signals must be known at compile time, that is, they must be determined by the connectivity of the circuit. Bidirectional elements, such as resistors, capacitors, and pass transistors, do not behave this way in full generality. Input and output signals are determined at run time by the signal values, not at compile time by the connectivity. In many cases, however, such elements are connected in such a way that signals flow in only one direction. They are effectively unidirectional. A pass transistor, for instance, that is connected between the output of one unidirectional module and the input of another has definite input and output ports. It is unidirectional in that particular connection. Elements that cannot be treated as unidirectional cannot be handled properly in the gate-level abstraction; they should be inside some block, where they can be dealt with at the electrical or switch level.

III. The Ulysses Description and Simulation System

A. Description Language

Unlike most description and simulation systems, Ulysses makes no fundamental distinction between descriptions of behavior and implementation structure. Its central principle is that the "level of complexity" is related to the complexity of the functions and the data structures, not to the complexity of the descriptive language itself. An algebraic notation is quite independent of functional simplicity or complexity. Accordingly, Ulysses provides a relatively small number of language "constructs" that deal with functions of unrestricted complexity. The primary constructs are shown in Table 1.

There are no built-in primitives. The user defines all functions, and he therefore has complete control of them at all levels of description. He may use functions from a library or any other source, or he may build his own at any time and use them immediately as components in his circuit. Functions from all sources may be intermixed freely.

The descriptive part of Ulysses consists of a language and a compiler. The language provides constructs for definition of functions, logical signals, and connections. A "scope" mechanism provides a means for controlling the visibility of object names, which is essential for hierarchical descriptions. The compiler generates data structures that are used by the simulation part of Ulysses to exercise the circuit. There are two ways to use the system for description. In the first, which might be called bottom-up, a logic schematic can be transcribed, module by module and wire by wire, into a Ulysses

description. The modules are represented by functions, and the wires are represented by signals. Groups of modules and wires that are repeated can be collected together into single functions that are given names and treated as units. A hierarchical description can be composed to any depth in this way. In the second, which might be called top-down, a behavioral description is expressed in terms of functions and signals, and is expanded into a hierarchy of functions and signals. Since the user controls all functional definitions, the top-down and bottom-up descriptions may be used in any combination, to any depth of hierarchy, and at any level of complexity.

A Ulysses description is written as an ASCII text file, using the constructs listed in Table 1:

Signals (SIG) are variables that represent circuit nodes. A signal has a name and a logical value of low or high, L and H. In order to be able to handle edge-sensitive functions, the domain of values in Ulysses includes rising and falling edges, R and F. An undefined value, U, signifies an unknown signal value (which may or may not be an error value), and high impedance, Z, means there is an open circuit. Signals are declared with the SIG construct: SIG A1 CLOCK Q.

Replication (REPL) is used with signal declarations to create signal vectors, or arrays of any number of dimensions. Vector and array components are specified by index values. For example, REPL[7...0] SIG BUS creates an 8-bit signal vector. BUS[2] is a scalar signal, and BUS[2 3] is a 2-bit signal vector. Ulysses has a full set of operations for composition, decomposition, and manipulation of signal vectors and arrays.

Case tables (CASETABLE) are definitions of primitive logic functions in truth-table form. They are the only form of functional definition in Ulysses. The user defines all functionality with them; there are no built-in definitions.

A case table consists of one or more rows of input signal values, output signal values, and output signal-delay values. There may be any number of input columns and output columns. For input values, a "don't care" notation (X) is provided, which can be used to condense tables. When a table is referenced, the values of the input arguments are compared with the values stored in the input columns, starting with the first row and proceeding through the table. When a match is found, the output values and delays for that row are returned. The delays represent propagation delays of signals in the module that the function describes. If no match is found for any user-written row, the values U and zero delay are returned for all output columns. A single delay value may be given for the table as a whole, or delays may be specified separately for each row and each output column. The latter method allows the user to define, for example, different delays for low-to-high and high-to-low transitions.

Devices with storage are handled by a feedback mechanism, in which the "old" value of an output signal is included in the argument list when the case table is referenced; the case table returns the "new" value. The extension of the input-argument list is done inside a template that references the case table (see below).

A case table for a JK flip-flop illustrates the essential features. It has an asynchronous clear input, and it is negative-edge triggered. The F/B (feedback) statement in the header is omitted if no storage is involved, that is, for combinational-logic functions.

CASETABLE C_JKFF ::						*/	Its name is C_JKFF
DOMAIN L H U Z F R						*/	Names of signal values
I/O (4 2)						*/	Four inputs, two outputs
F/B 2						*/	Both outputs are fed back
DELAYS 20						*/	Delay value for all transitions
PORTS CLR CLK	J	K	Q	QB	*/	Inputs CLR CLK J K, outputs Q QB
L X	X	X	L	H	*/	Clear active low
H F	L	L	Q	QB	*/	No change: copy old outputs to new
H F	L	H	L	H	*/	Synchronous load Q low
H F	H	L	H	L	*/	Synchronous load Q high
H F	H	H	QB	Q	*/	Toggle: copy old Q to QB, old QB to Q
H F	U	X	U	U	*/	Catch glitches — return U values
H F	X	U	U	U		
X X	X	X	Q	QB	*/	If none of the above, do nothing

ENDCASES

This case table is typical of user-created functional definitions. It may or may not be a correct representation of a different JK flip-flop in another application; that judgement is up to the user. The case table may be edited to implement the desired behavior.

Case tables can represent transistors, in a limited sense. A transistor is treated as a switch with a control port (gate or base) and a switched data path. The limitation comes from the fact that a MOS transistor is like a resistor, in that it is bidirectional: The current can flow in either direction. This behavior cannot be represented by a function with specific input and output signals. If the transistor is connected in such a way that the input and output ports never change, it can be represented as a case table. Pullup and pulldown transistors meet this requirement and can be represented in case tables, because the terminal that is connected to the supply terminal (power or ground) is the source, the other terminal is the drain (output), and the gate voltage (input signal) is referenced to a specific node (the source node). Pass transistors, on the other hand, are inherently bidirectional, because the direction of current flow is determined at run time by the signal values, rather than at compile time by the connections. If the nature of the circuit is such, however, that the direction of signal flow (which is not necessarily the same as the direction of current flow) is fixed, the pass transistor can often be represented in a case table.

Case tables handle only scalar signals directly; signal vectors and arrays are handled by case table references in templates.

Templates (TEMPL) represent circuit modules. A template has a name, a list of input signals (dummy arguments), signal definitions, any number of definitions of actions (DEFs), and a list of output signals. Connectivity is defined by the relationships among input and output signals of functions. A template may reference case tables, other templates, or itself. The DEF construct defines functional relationships. Its form is

DEF output-signal list = function_name (input-signal list)

As described later, the basic action of the Ulysses simulator is to drive DEFs whenever one or more of their input signals change value. Driving a DEF schedules all of its output signals to assume new values at later times, as defined by the signals' delay values.

There is no notion of a sequence of events in a template; all actions are simultaneous. Consequently, the order in which the DEF statements are written has no effect on the meaning of the template.

The template example below illustrates several features of the construction. The C_JKFF case table listed above is the primitive behavior definition; the feedback of outputs to inputs is done in the template. The internal signals are vectors with four components.

```
TEMPL JKFF = << PARS          */ Dummy argument names
(SIG CLEAR CLOCK J K )
  REPL[3...0] J K             */ J and K are 4-bit signals
  REPL[3...0] SIG Q QB        */ Internal 4-bit signals
  DEF Q QB @ C_JKFF           */ Outputs fed back to inputs
  ( CLEAR CLOCK J K
  Q QB )
  Q QB >>                     */ Return two 4-bit signals
```

The line REPL[3...0] SIG Q QB contains the SIG keyword, so it creates two internal signal vectors of four components each. The names Q and QB are private to the template. The line REPL[3...0] J K does not contain the SIG keyword, so it does not create signals; it tells the compiler that the J and K inputs are 4-vectors. The output signals are listed in the final line. In this example they are the two 4-vectors that are the outputs of the JK flip-flops. The case table is referenced once for each component, which generates, in effect, four distinct copies of the flip-flop. The CLEAR and CLOCK inputs are signal scalars. Ulysses fans them out to provide these inputs to all four copies of the flip-flop.

In the particular case of references to case tables with feedback, the template construction listed above is mandatory, because it is the only way in which the case table can be provided with the six inputs (I/O plus F/B) that it expects. Higher-level references to the template JKFF supply only the four "normal" input values (CLEAR, CLOCK, J, and K). The feedback mechanism that implements storage is hidden inside this template. In effect, storage is implemented as a case table wrapped in a template.

Templates may reference other templates, which provides the means of constructing a hierarchical description of a digital system. In some other template, for instance, there might be the line

DEF A B = JKFF(RESET PHI1 P Q)

This DEF references the template JKFF defined above. A and B must both be present, and they must be defined as 4-vectors, because JKFF returns two 4-vector values. Another template might reference this one, that template might be referenced by another, and so on upwards in the hierarchy.

Scopes (SC.) provide the means of controlling the visibility of names in the hierarchy. A scope may be thought of as a black box with a name. All structure inside the box is invisible from the outside. The only communication with it is through the inputs and outputs. A scope may be "entered" by giving its name, which is equivalent to removing the cover of the box: The internal names become visible. It is "exited" by

the END statement, which is equivalent to replacing the cover.

Templates are scopes. Their internal names are private, that is, invisible in the scope of the referencing function, unless the template has been entered and not yet exited.

Instantiations of circuit blocks are created as scopes. They consist of signal definitions, declarations of inner scopes, scope names and END statements for entering and exiting scopes, and DEF statements that define the functionality and connectivity.

The scope mechanism provides the means for identifying particular instances of modules that are defined as templates and used repeatedly in the system. The internal names of the signals in a particular template are the same in all instances of the template, but the names for a particular instance are made unique by concatenating the template name with the names of the scopes that were entered in sequence to get at the signal. For example, the names Q and QB in the template JKFF are private to the template. A particular instance of the template, which corresponds to a physical part, might be extracted in a netlist with the name COUNTER.INPUT.Q, where '.' signifies concatenation of names, and COUNTER and INPUT are scope names.

B. Compiled Data Structures

Compilation of the description of a digital circuit generates the data structures shown in Table 2.

Each signal (scalar or vector component) has a number assigned to it. The number is its index into the array of signal records. For each signal, there is a dependency list that contains the numbers of all signals that are driven by functions that are driven by the signal. In other words, it is a list of all signals that are affected by changes in that one particular signal's value. There is also a pointer to the function that drives the signal, that is, to the function named in the DEF statement that defines how the signal is driven. (A signal can be driven by only one DEF statement.)

At the address of the signal's driving function there are two pieces of information: the address of a list of the names of its arguments, and the address of a list of elementary simulator actions that evaluate the function.

The argument list contains the signal numbers of the input and output arguments and the address of the compiled case table (i.e., primitive function) that is to be executed. (The description compiler expands the hierarchy of template references down to the primitive-function level in order to generate the argument lists.)

The run-time function is usually the implementation of a single DEF statement at the primitive-function level, that is, the execution of a single case table reference. There are four elementary actions:

(1) Get the value of an object in the argument list and push it on the operand stack.

(2) Get the address of an object in the argument list and push it on the operand stack (the address of a signal is its number).

(3) Execute the case table whose address is on the stack.

(4) Schedule the signal whose address, new logic value and delay value are on the stack.

The construction is more general than this, however; any Ulysses function can be executed, using the operand stack for communication of arguments. RAM and ROM behaviors have been implemented compactly as general functions. While they can be implemented with case tables and templates, the amount of host-machine memory needed for a case table description of a large RAM or ROM is excessive, and the user has to wrestle needlessly with the details of addressing and read-write control if he is not actually building a RAM, but only using its behavior in the description of a digital system.

The way in which these structures are used is as follows: The signals that are dependent on a particular signal are in its dependency list. When a signal changes value, its dependency list is scanned in order to mark the dependent signals for processing. A dependent signal is processed by executing the function that drives it. The pointer to the driving function gives its location and the location of its arguments. The execution of a case table results in new value and delay pairs for all of its output signals. The simulator uses the delay values to schedule the output signals to change value at the time "now" plus delay.

C. The Ulysses Simulator

The notion of time is implemented as a linearly increasing quantity, digitized into "slots." Each slot has a number. There is a slot counter, whose value is the current slot, or "now." When the slot count is N, the time is within the range corresponding to N times the width of a slot in some time unit (e.g., one nanosecond). In each slot, a signal has a value selected from the domain L(ow), H(igh), F(alling), R(ising), U(ndefined), and Z (high impedance). This treatment of time is similar to the way in which logic analyzers deal with it.

Associated with each slot is an event list, which contains the names and new values of all signals that have been scheduled to change value in that slot. The simulator has an event counter.

Every value change generates an event, for which the event counter is incremented. When the slot counter advances to a new slot, all of the events in the slot's list are processed. The event counter is decremented as each list element is processed by evaluating the functions that drive the signals in its dependency list. Each evaluation results in a new value and delay pair for each of the function's output signals. Each new value is an event, and it is scheduled in (i.e., appended to the event list of) the slot whose number is current-slot plus delay.

The simulator is initialized with the slot counter and event counter set to zero, and all event lists empty. All internal signal values except those connected to power or ground are set to U and the system is driven once, in imitation of a power-up sequence. The simulator is started by forcing one or more input signals to particular logic values in particular slots. Each forcing action is an event: The signal name and value are scheduled in the specified slot, and the event counter is incremented. After all forcing events have been specified, the RUN command is issued. The simulator processes the forcing events, which generate new events, which are processed and which generate new events, et cetera. The event counter follows the progress of the computation, being incremented for new events, and decremented as events are processed. The simulator runs until the event count returns to zero. The state of the system at stop time is preserved, that is, the values of all signals are remembered. A new cycle is started with a new set of forcing events, followed by the RUN command.

The simulator does not know or care what the individual functions are. It simply applies them to signals, as dictated by the event lists. Simple gate-level functions are treated in the same way as functions representing large, complex modules. It is a mixed-level simulator that can exercise circuits at all levels and in any combination of levels, from transistors (when they meet the requirements of case table representation) through logic gates to functional blocks of arbitrary complexity.

IV. Future Development of Ulysses

There are a number of extensions and improvements that are needed for a full capability. Some are fundamental, some are cosmetic. Only the key fundamental extensions will be described; cosmetic improvements include improved error diagnostics and reporting, and an interface to a graphical means of generating circuit descriptions.

A. Netlist Generation and Acceptance

The compiled data structures described above define the circuit completely at the logical level. They provide the infor-

mation needed by the simulator to analyze the behavior of the circuit. They are in a form, however, that is meaningful only to Ulysses. In order to use the circuit description in some external context, such as a schematic-capture editor in an engineering workstation, it is necessary to generate a netlist for it; in order to replicate in Ulysses a circuit description generated externally, it is necessary to accept a netlist from the external source. A netlist is a list of nodes. Each list element contains the name of the node and a list of module-terminal names connected to it. Associated with the netlist is a description of each module: terminals, logical behavior, delays, strengths, etc. Netlists and module descriptions are generally written in ASCII format. Beyond this, there is no generally accepted standard format for composing and combining the lists of nodes and module ports. Generation and acceptance of netlists from one system to another is essentially a problem of format translation. One of the main problems is to work out the correspondence of names in the description hierarchy (in Ulysses, scope names) and names in the netlist. Lacking a general netlist format, it is necessary to write a separate translator for each external system. A standard format, called NIF (Netlist Intermediate Form) has been defined for Ulysses. Development of a netlist capability for Ulysses would have two stages:

(1) Write translators directly for specific systems, such as the Mentor CAE workstations that JPL uses, in order to work out the concepts and details and to debug the translation process.

(2) Develop the NIF concept, write the translators between Ulysses and NIF once, and write the translators between NIF and external systems as needed.

The advantage of inserting NIF into the process is that it separates the details of the Ulysses description from the details of the translation process.

There is a more general advantage to a standard netlist format: In the general case, with N different netlist formats, and with separate translators needed for generation and acceptance, a total of $N(N-1)$ translation programs must be written for a complete data-interchange capability. With a standard format, two translation programs must be written for each external system: one into NIF, and one out of it. The total translation-program count becomes $2N$, which is significantly smaller than $N(N-1)$, for large N. The disadvantage of having to use a two-step process for translation (external — NIF — external) is relatively minor. Development of the standard netlist format concept is advantageous, independently of the Ulysses system. The NIF concept can be developed, pending the arrival of a generally accepted format, such as has been proposed in the EDIF (Electronic Data Interchange Format) documentation that is currently being reviewed by

the electronic industry, or a format associated with the VHDL (VHSIC Hardware Description Language) concept sponsored by the U.S. Department of Defense.

B. Multiple DEFs

At present, Ulysses allows only one DEF statement for each output signal value. This restriction prevents the user from defining conditional executions in the full behavioral sense; he is forced to construct them as actual circuit elements. For example, an ALU executes one of a number of logical operations, such as ADD, SUBTRACT, AND, or OR, depending on the value of an index signal vector. If the vector is three bits wide as in the 2901 bit-slice processor, one of eight logical operations is selected. An extension has been mapped out that generalizes the DEF construct to include a logical condition based on signal values: If the condition is true, then execute the DEF; else do nothing. This construct will be quite general in nature. It is the analog in the rule-based Ulysses system of the IF-THEN-ELSE construction in procedural languages. The present Ulysses system of a single DEF per output signal is a special case: The condition is always true.

C. Removal of Edges From the Signal Domain

Edge values (rise and fall) were added to the signal domain in the early stages of development, in order to have a means of treating edge-triggered components. They do not properly belong there; they are properties of the components, not of the wire voltages. The case table construct can be modified to test for edges in those inputs that are edge-sensitive. When this is accomplished, the case table definition will be shorter and cleaner, because all of the edge cases in inverters and gates may be deleted. The user will have the option of using them as he sees fit; the case table will know how to deal with them when they are encountered.

D. Removal of Strength From the Signal Domain

The development of Ulysses lags behind that of more conventional simulators in this regard: Drive strength is expressed as signal values (e.g., H = strong high, and h = weak high), rather than as a separate attribute associated with node capacitance and module output currents. An extension has been mapped out tentatively: Node voltage is expressed as an n-bit digitized value. If $n = 1$, the representation is binary, or boolean, as it is at present. If $n > 1$, the representation is intermediate between boolean and real-variable. The concept of strength may then be included as an abstraction of Ohm's law, in that it can deal with currents, voltages, and admittances. The basic principle is that the evolution of node voltage may be tracked by calculating its derivative and extrapo-

lating to the upper or lower threshold voltage; a delay time may be extracted from this extrapolation. In the simplest cases, the derivative of the node voltage is determined by drive currents and node capacitances:

$$I = C \cdot dV/dt,$$

$$dt = (C/I) \cdot dV.$$

The quantity dt is the predicted time delay before reaching a threshold voltage. The quantity dV is the difference between the current voltage value and the threshold voltage. The quantity C/I is related to signal strength, as shown in the following dimensional argument:

$$I = g \cdot V, \qquad g = \text{conductance}$$

$$C/I = C/(gV) = (C/g) \cdot (1/V).$$

The quantity C/g has the dimensions of time. Strength may therefore be defined as its reciprocal, with the property that high strength corresponds to short time. In other words, a strong signal leads to short delay values. The reason for defining strength in terms of conductance rather than current is that, at least in a simple constant-conductance model, the conductance is a fixed circuit parameter.

It is anticipated that development of a clean method of scheduling and rescheduling events will be one of the most difficult tasks in implementing this approach.

V. Summary and Conclusion

Ulysses is a system that provides a description language for digital circuits of arbitrary complexity, and a simulator for predicting their behavior and performance. The description language is uniform and consistent over the range of descriptions from arbitrarily high behavioral levels down to the detailed transistor-switch level. At the transistor-switch level, the correspondence of the circuit to an actual electrical and physical implementation is direct: Each transistor, wire, and contact has an electrical and physical counterpart.

The fundamental usefulness of this system is that the user can follow the hierarchical structure from the top behavioral level down to the transistor level selectively, in as much or as little detail as required. Since there are no discontinuities in the language over this entire range, the integrity of the description is guaranteed. (It should be pointed out that Ulysses knows only what the designer tells it about the circuit. If the

design or the circuit parameters contain errors, Ulysses will propagate those errors faithfully.)

The uniformity of description from the switch level upward can potentially be used to control the transistor implementation from the architectural level. Such control would be valuable in applications in which transistor-level factors are important at the system level: power dissipation, sensitivity to single-event upsets, testability, and fault tolerance. Current methods of investigation are based on building the hardware and testing it; modelling such behavior in software is less expensive and more flexible. A major development problem, in general, is to make the modelling sufficiently accurate to provide useful results.

Reference

1. Godfrey, M. D., D. F. Hendry, H. J. Hermans, and R. K. Hessenberg, "Machine-Independent Organic Software Tools (MINT)," Academic Press, Orlando, Florida, Revised 2nd Edition, 1985.

Table 1. Description constructs[a]

Name	Meaning
SIG	Signal, a variable representing a circuit node.
REPL	Replication of signals to form signal vectors.
CASETABLE	Truth table. Defines the behavior of primitive functions.
DELAYS	CASETABLE timing. Represents propagation delay.
TEMPL	Template. Defines interconnections of modules, and enables building new functions.
DEF	Defines action: DEF outputs = function (inputs).
SC.	Scope. Defines visibility of names in the hierarchy.

[a]Ulysses keywords are capitalized.

Table 2. Compiled data structures

Structure	Contents
Signal record	Dependency list Pointer to its driving function
Driving function	Address of the argument list Address of the run-time function
Argument list	List of input-signal names Address of an executable case table image List of output-signal names
Run-time function	List of actions: Get input values Execute the case table Schedule output values

Improved Performance of a Digital Phase-Locked Loop Combined with a Frequency/Frequency-Rate Estimator

A. Mileant and M. Simon
Telecommunications Systems Section

When a digital phase-locked loop with a long loop update time tracks a signal with high doppler, the demodulation losses due to frequency mismatch can become very significant. One way of reducing these doppler-related losses is to compensate for the doppler effect using some kind of frequency-rate estimator. The performance of the fixed-window least-squares estimator and the Kalman filter is investigated; several doppler compensating techniques are proposed. It is shown that the variance of the frequency estimator can be made as small as desired, and with this, the doppler effect can be effectively compensated. The remaining demodulation losses due to phase jitter in the loop can be less than 0.1 dB.

I. Introduction

Figure 1 depicts the major components of a digital phase-locked loop (DPLL) used in tracking low signal-to-noise ratio (SNR) signals. The integrate-and-dump circuit integrates the phase error at the output of the phase detector producing the so-called error signal. At the loop update instants, this error signal is sampled, filtered, and used to set the frequency of the digitally controlled oscillator (DCO) to its new value. So, the continuously changing frequency of the tracked signal is followed by the DCO in a staircase fashion.

In selecting the optimum loop update time, T, one encounters two conflicting effects. On the one hand, the phase jitter in the loop decreases in proportion to $1/T$. On the other hand, the static phase error due to frequency mismatch increases in proportion to T. During high doppler rates, the demodulation losses due to the static phase error may be very significant. Figure 10[1] illustrates our point.

At the Deep Space Network (DSN) stations, the static phase error of an analog PLL is reduced by ramping the voltage-controlled oscillator (VCO) using a predicted trajectory file and the Programmed Oscillator Control Assembly (POCA). A similar technique of ramping the DCO's frequency to reduce the phase error due to doppler rate in a DPLL is considered in this analysis. However, no predicted trajectory files to do this DCO ramping will be assumed in this article. Instead, techniques of estimating the frequency of the tracked signal using linear estimators is investigated.

---

[1]From Simon, M., and A. Mileant, "DSA's Subcarrier Demodulation Losses," IOM 3395-85-55. Jet Propulsion Laboratory, Pasadena, Calif., March 1985. (JPL Internal document.)

In Sections II and III, the equations and performance of two linear estimators are derived, namely, of the "fixed-window" least-squares estimator and the Kalman filter. In Section V, several possible implementations of the combined estimator/DPLL demodulator are compared. It is shown that the incorporation of an estimator in the carrier/subcarrier demodulation process can virtually eliminate the doppler-related demodulation losses. Finally, in Appendix A, the transfer functions for the estimators are given for future reference.

II. Fixed-Window Least-Squares Estimator

Because of the Doppler effect, the instantaneous frequency $x(t)$ tracked by a DPLL is a time-varying function, which in the Taylor series expansion is of the form

$$x(t) = x_1 + x_2 (t - T_r) + x_3 (t - T_r)^2/2 + \cdots \tag{1}$$

where T_r is some arbitrary reference time. In this analysis, it will be assumed that only the first *two* parameters, i.e., the frequency at time T_r, x_1, and the frequency rate x_2, have significant value and need to be estimated from the available data.

At the loop update instants t_k, the loop produces $\hat{x}(k)$, which is the estimate of $x(k)$, at the DCO's output.[2] Since the DCO does not have any offsets or frequency drifts, there is a one-to-one relation between the DCO's input, $y(k)$, and its output, $x(k)$; i.e., $x(k) = cy(k)$, where c is a constant denoting the DCO gain. Without loss of generality, from now on we will assume that $c = 1$, so that $x(k) = y(k)$. Using M frequency samples, we want to estimate x_1, the frequency at time T_r, and x_2, the frequency rate in the time interval $T_r \leqslant t \leqslant MT + T_r$.

At time instants $t_k = kT$ we obtain the DCO's frequency sample, $y(k)$, which is assumed to be of the following mathematical form (see Fig. 2):

$$y(k) = x_1 + x_2 (t_k - T_r) + v(k) \tag{2}$$

Here $v(k)$ is the noise due to the phase jitter in the DPLL. It is modeled as a Gaussian random variable with zero mean and variance σ_y^2. T is the loop update time and $k = 1, 2, \cdots M$.

Given M noisy frequency samples, $y(k)$, x_1 and x_2 can be estimated using the following least-squares algorithm (Ref. 1):

$$x = (H^T H)^{-1} H^T y \tag{3}$$

---

[2] For convenience, we shall denote sampled values of a process $x(t)$ by $x(k)$ rather than $x(t_k)$.

where x is a two-dimensional estimation vector defined as

$$\hat{x} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} \tag{4}$$

y is an M-dimensional data vector

$$y = \begin{bmatrix} y(1) \\ y(2) \\ \cdot \\ \cdot \\ \cdot \\ y(M-1) \\ y(M) \end{bmatrix} \tag{5}$$

and H is an $M \times 2$ observation matrix, which by inspection of Eq. (2), is

$$H^T = \begin{bmatrix} 1 & 1 & \dots\dots\dots\dots 1 \\ (T - T_r) & (2T - T_r) & \dots\dots (MT - T_r) \end{bmatrix} \tag{6}$$

Using Eq. (6), it can be shown that

$$(H^T H)^{-1} = \frac{12}{T^2 M(M-1)(M+1)}$$
$$\times \begin{bmatrix} \left[T_r^2 - T_r T(M+1) + \dfrac{T^2(M+1)(2M+1)}{6}\right] & T_r - \dfrac{T(M+1)}{2} \\ T_r - \dfrac{T(M+1)}{2} & 1 \end{bmatrix}$$

$$\tag{7}$$

The above equation simplifies considerably when the reference time T_r is set to zero (equivalent to reducing the Taylor series of Eq. (1) to a Maclaurin series). In this case, the above equation becomes

$$(H^T H)^{-1} = \frac{2}{M(M-1)} \begin{bmatrix} (2M+1) & \dfrac{-3}{T} \\ \dfrac{-3}{T} & \dfrac{6}{T^2(M+1)} \end{bmatrix} \tag{8}$$

With $T_r = 0$ and inserting Eqs. (5), (6), and (8) into Eq. (3), $\hat{x}_1$ and $\hat{x}_2$ are computed as follows

$$
\mathbf{x} = \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} = \frac{2}{M(M-1)}
$$

$$
\times \begin{bmatrix} (2M+1) & \dfrac{-3}{T} \\[2ex] \dfrac{-3}{T} & \dfrac{6}{T^2(M+1)} \end{bmatrix} \begin{bmatrix} \displaystyle\sum_{k=1}^{M} y(k) \\[3ex] T \displaystyle\sum_{k=1}^{M} ky(k) \end{bmatrix} \qquad (9)
$$

It can be shown that both $\hat{x}_1$ and $\hat{x}_2$ are *unbiased* estimates of x_1 and x_2, respectively.

The covariance matrix of the error in the estimator, according to Ref. 1, because of the independence of the noise samples $v(k)$, is

$$
E[\mathbf{e}\mathbf{e}^T] \triangleq \mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \sigma_y^2 (\mathbf{H}^T\mathbf{H})^{-1} \qquad (10)
$$

where the error vector $\mathbf{e}$ is defined as

$$
\mathbf{e} \triangleq \begin{bmatrix} \hat{x}_1 - x_1 \\ \hat{x}_2 - x_2 \end{bmatrix} \qquad (11)
$$

$(\mathbf{H}^T\mathbf{H})^{-1}$ is given by Eq. (7) or Eq. (8) and σ_y^2 is the variance of the DCO's frequency samples.

Observing Eq. (8), we see that for large M, the error variance for x_1 is proportional to $1/M$ whereas for x_2 the error variance is proportional to $1/M^3$. This implies that the uncertainty in the estimate of x_2 diminishes much more rapidly than for x_1.

Using the estimates $\hat{x}_1$ and $\hat{x}_2$, the frequency estimate at time t will be

$$
\hat{x}(t) = \hat{x}_1 + (t - T_r)\hat{x}_2 \qquad (12)
$$

with expected value

$$
E[x(t)] = x_1 + (t - T_r)x_2 \qquad (13)
$$

and variance

$$
\text{var}[\hat{x}(t)] \triangleq \sigma_e^2(t)
$$

$$
= p_{11} + 2(t - T_r)p_{12} + (t - T_r)^2 p_{22}
$$

$$
= \sigma_y^2 \frac{2}{M(M-1)} \left[(2M+1) - \frac{6}{T}(t-T_r) + \frac{6(t-T_r)^2}{T^2(M+1)} \right] \qquad (14)
$$

Note that the variance of the estimated frequency, $\sigma_e^2(t)$, is a parabola with minimum value at time

$$
t_{min} = T(M+1)/2 + T_r \qquad (15)
$$

which lies in the *midpoint* of the data stream. The above equations (12) through (14) are true for prediction $(t > MT + T_r)$ as well as for smoothing $(t < MT + T_r)$.

Let γ be defined as the ratio of the variance of the frequency estimator to the variance of the frequency samples, i.e.,

$$
\gamma \triangleq \frac{\sigma_e^2(t)}{\sigma_y^2} \qquad (16)
$$

and again without loss of generality let $T_r = 0$. Then it can be shown that

$$
\frac{1}{M} \leqslant \gamma \leqslant \frac{4M+2}{M(M-1)} \triangleq \gamma_{max} \qquad (17)
$$

for $T(M+1)/2 \leqslant t \leqslant T(M+1)$.

The above equation says that our estimator will have minimum variance σ_y^2/M when smoothing is performed in the middle of the data stream. On the other hand, if we want to use the estimator as a *predictor* T seconds ahead of the most recent data point, then the variance of the estimator will be $\gamma_{max} \sigma_y^2$, where γ_{max} will be equal to the upper bound of Eq. (17). To improve demodulation, we want the variance of the predictor to be less then the variance of the samples, i.e., we want $\gamma_{max} < 1$. This sets the lower bound on the number of samples needed for computation of $x(t)$, namely, $M \geqslant 6$. Of course, the larger M is, the lower will be the error variance with the penalty of bigger computational burden.

In the proposed implementation of the least-squares algorithm, each time a new frequency sample becomes available,

the oldest sample is discarded. This can be accomplished with a shift register as shown in Fig. 3. In this implementation, in each loop update period the tracked frequency and frequency rate are estimated from the M last frequency samples. Hence the name "fixed-window" estimator.

III. Kalman Filter Estimator

In this section, the performance of a second-order Kalman filter for estimating the frequency and the frequency rate is investigated. The Kalman filter belongs to the class of recursive linear estimators. We begin by making several definitions. Let $x(k)$ be the two-dimensional state vector defined in Eq. (4). Then, in accordance with our previous discussion, the state-space equations describing the evolution of our system from time t_k to time t_{k+1} will be (see Fig. 4)

$$x(k + 1) = Fx(k) + w(k) \tag{18}$$

$$y(k) = Hx(k) + v(k) \tag{19}$$

where F is a 2×2 state transition matrix given by

$$F = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \tag{20}$$

and H is a 1×2 observation matrix given by

$$H = [1 \ 0] \tag{21}$$

and is different from H defined previously. $y(k)$ is again the frequency noisy sample. $w(k)$ is modeled as a stationary white noise process with covariance matrix

$$Q = E [w(k)w(k)^T] = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \tag{22}$$

which takes into account the unmodeled disturbances in the system e.g., the higher order terms in Eq. (1). $v(k)$ is the same as before. Note that F, H, and Q are assumed to be time-invariant.

The operation of the Kalman filter is given in terms of the error covariance matrix P and the gain matrix K. Referring to the time diagram of Fig. 5, the equations describing the operation of the filter at the instant when the measurement $y(k)$ becomes available are (Ref. 2):

State Estimate Update

$$\hat{x}_f(k) = x_p(k) + K(k) [y(k) - Hx_p(k)] \tag{23}$$

Error Covariance Update

$$P_f(k) = [I - K(k)H] P_p(k) \tag{24}$$

Kalman Gain Update

$$K(k) = P_p(k)H^T [HP_p(k)H^T - \sigma_y^2]^{-1}$$

$$= \begin{bmatrix} k_1(k) \\ k_2(k) \end{bmatrix} \tag{25}$$

where the subscript p stands for predicted or extrapolated values and the subscript f denotes filtered or updated values. The corresponding extrapolation or prediction equations are

State Estimate Extrapolation

$$x_p(k) = Fx_f(k - 1) \tag{26}$$

Error Covariance Extrapolation

$$P_p(k) = FP_f(k - 1) F^T + Q \tag{27}$$

To assess the performance of the Kalman filter, we need to know the steady-state value of the error covariance matrix P in terms of Q and σ_y^2. It turns out that the closed form solution of the steady-state matrix P (and K) is quite difficult to derive given an arbitrary Q matrix, even for a second-order system. We will use the results of Ref. 3 where the steady-state covariance and gain matrices are derived for Q of the following form

$$Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} = \sigma_q^2 T^2 \begin{bmatrix} T^2/3 & T/2 \\ T/2 & 1 \end{bmatrix} \tag{28}$$

In our analysis we will treat σ_q^2 as a constant which we can set to some desired value. Rewriting Eqs. (12) through (18) of Ref. 3, the elements of the steady-state filtered covariance matrix are obtained, namely,

$$p_{11} = \sigma_y^2 \left[2\alpha + 2\sqrt{\alpha}\sqrt{\alpha + r^2} - 8\sqrt{\alpha + r^2} - 8/3\right]/r^2$$

$$p_{12} = p_{21} = \sigma_y^2 \, 4\left[\sqrt{\alpha + r^2} - \sqrt{\alpha}\right]/(Tr^2) \tag{29}$$

$$p_{22} = 8\sigma_y^2 \left[\sqrt{\alpha} - 1\right]/(Tr)^2$$

where

$$r^2 \triangleq 16\sigma_y^2/(\sigma_q^2 T^4) \tag{30}$$

and

$$\alpha = 4/3 + 2\sqrt{r^2 + 1/3} \tag{31}$$

When the Kalman filter is used to estimate $\hat{x}(t)$ in the time interval $t_k \leqslant t \leqslant t_k + T$ (for 'real-time' demodulation), the variance of this estimator will be (analogous to Eq. (14))

$$\sigma_e^2(t) = p_{11} + 2tp_{12} + t^2 p_{22}, \quad 0 \leqslant t \leqslant T \tag{32}$$

Inserting Eq. (29) into Eq. (32) with $t = T$ and simplifying, we obtain the maximum value of the variance of our estimated frequency, namely,

$$\sigma_e^2 \Big|_{max} \triangleq \sigma_e^2 = \sigma_y^2 \left[2\alpha + 2\sqrt{\alpha^2 + \alpha r^2} - 16/3\right]/r^2 \tag{33}$$

Defining again γ as the ratio of the input to output variances of the estimator, i.e., $\gamma = \sigma_e^2/\sigma_y^2$, the curve of γ versus r^2 is obtained and shown in Fig. 6. From that figure we see that $\gamma < 1$ for $r^2 > 300$. Since r^2 is a scaled version of the ratio of the measurement variance, σ_y^2 to σ_q^2, the upper bound for choosing the latter should be (from Eq. (30))

$$\sigma_q^2 < 16\sigma_y^2/(300T^4) \tag{34}$$

The lower bound for σ_q^2 is determined by the unaccounted dynamics of the tracked signal such as the frequency acceleration. In estimating $x(k)$, the matrix Q has the effect of "washing out" the old data: the larger Q is, the less effect the old measurements will have against new ones. Conversely, small Q makes the estimate insensitive to new measurements. This is undesirable when significant signal dynamics are expected. The Q matrix also determines the size of the error in the estimator (see Eq. (27): the larger Q is, the larger P will be. The optimum set of values for Q has to be determined possibly by simulation.

Figure 6 also shows curves of the elements of the steady-state gain matrix, k_1 and k_2 vs r^2. These gains are related to p_{11} and p_{12} as follows:

$$k_1 = p_{11}/\sigma_y^2$$

$$k_2 = p_{12}/\sigma_y^2 \tag{35}$$

where p_{11} and p_{12} are given by Eq. (29).

IV. Some Comparisons Between the Two Estimators

Both estimators, the least-squares and the Kalman filter, belong to the class of minimum-variance, unbiased, linear estimators. The first one represents a "batch-type" approach; the second, a recursive.

On each cycle (loop update period), the least-squares algorithm requires approximately $2M$ summations and $M + 3$ multiplications, while the Kalman filter requires approximately 19 summations and 16 multiplications.

The comparison of the error variance for the two estimators is not that obvious. In the least-squares algorithm, the variance depends on M the number of samples considered. In the Kalman filter, the variance depends on our specification of the Q matrix, which can be arbitrarily selected depending on how fast we want the filter to follow the new data.

The Kalman filter seems to be a more elegant approach to the estimation problem. However, the least-squares algorithm does not have the instability and divergence problems of the P matrix of the Kalman filter.

V. Signal Demodulation Using Frequency Rate Compensation

There are many ways in which frequency estimators can be combined with a DPLL or can become part of a DPLL in order to improve the demodulation process of a doppler distorted signal. We will compare here three implementations which we shall call the "Parallel" the "Serial" and the "Single Loop" Estimator-Added Demodulators.

The concept of a *Parallel Estimator-Added Demodulator* is represented in Fig. 7. It consists of M DPLLs and $(M - 1)$ delays. Each subsequent loop tracks the signal delayed by T seconds relative to the previous loop. At the loop update instants, the noisy frequency samples of the M DCOs are fed

into the computing device that estimates $\hat{x}_1$ and $\hat{x}_2$ and from them $\hat{x}(t)$. Finally, $\hat{x}(t)$ drives the DCO, which performs the actual demodulation. In this implementation, demodulation is accomplished with an 'open-loop receiver'. Fine tuning of the phase can be done with an epoch-tracking loop, which is shown with dotted lines. This "parallel" demodulation scheme works only with the least-squares algorithm discussed in Section II. This implementation requires the minimum number of components (DPLLs and delays) when the demodulated signal comes from the middle delay where the variance of the estimator has its minimum value. Four DPLLs with three delays should be sufficient to give an $\hat{x}(t)$ with small variance.

The same demodulation results can be obtained in a more economical way with the *Serial Estimator-Added Demodulator*, which is shown in Fig. 8. Here the noisy frequency samples, $y(k)$, are obtained from a single DPLL. These samples can be fed either into the least-squares estimator or the Kalman filter where the estimated frequency $x(t)$ is computed. The final demodulation process is identical to the one described previously, i.e., it can be performed with an open-loop receiver or with the aid of an epoch-tracking loop.

Finally, in Fig. 9, the *DPLL with a Frequency Rate Compensator* concept is depicted. At loop update instants, the $y(k)$ samples are fed into either the least-squares estimator or the Kalman filter where the frequency rate $x_2(k)$ is estimated. Then $\hat{x}_2(k)$ is used to ramp the DCO between the loop update instants. This compensates for the doppler effect and reduces the loop phase error due to frequency mismatch. Since, in this implementation, the frequency rate estimator becomes part of the DPLL, the loop filter has to compensate for the poles of the estimator. Appendix A gives the transfer functions for the least-squares estimator and the Kalman filter and Figs. 11 and 12 show in a block diagram the interaction of the DPLL components with a second order estimator. However, this is done here only for future reference. A detailed analysis of a DPLL enhanced with a frequency-rate estimator will be treated in a subsequent article.

Comparing the above three schemes, we observe that the "parallel" implementation is the least economical from the standpoint of the number of components. From the standpoint of reducing the phase error in the loop, they all appear to be equal.

VI. Conclusion

When a DPLL with a long loop update time tracks a signal with high doppler, the losses due to frequency mismatch can become very significant. One way of reducing these doppler-related losses is to compensate for the frequency rate using some kind of estimator. It was shown that the variance of the estimator can be made as small as desired. In other words, the doppler effect can be effectively compensated. The remaining demodulation losses due to phase jitter in the loop will be less than 0.1 dB, as is illustrated in Fig. 10.

In Sections II and III, the performances of the fixed-window least-squares estimator and the Kalman filter are investigated. They both belong to the class of minimum-variance linear estimators. The least-squares is a batch-type algorithm, whereas the Kalman filter uses a recursive algorithm. Appendix A gives the transfer functions of these estimators for future reference. In Section V, several possible doppler compensating techniques are proposed.

Acknowledgment

References

1. Sorensen, H. W., *Parameter Estimation*, Marcel Dekker, Inc., N.Y., 1980.

2. Gelb, A., *Applied Optimal Estimation*, The MIT Press, Cambridge, Mass., 1984.

3. Ekstrand, B., "Analytical Steady State Solution for a Kalman Filter," *IEEE Trans. on Aerospace and Electronics Systems*, Vol. 19, Nov. 1983, pp. 815-819.

4. Simon, M., and A. Mileant, "Digital Filters for Digital Phase-Locked Loops," *TDA Progress Report 42-81*, January-March 1985, Jet Propulsion Laboratory, Pasadena, Calif., pp. 81-92.
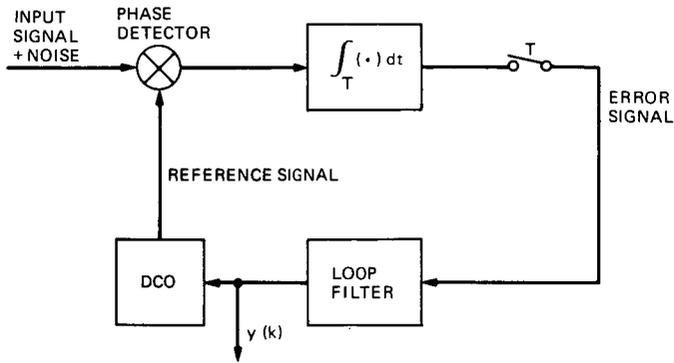
Fig. 1. Digital phase-locked loop block diagram

Fig. 2. Model of the frequency process

Fig. 3. Fixed-window least-squares estimator block diagram
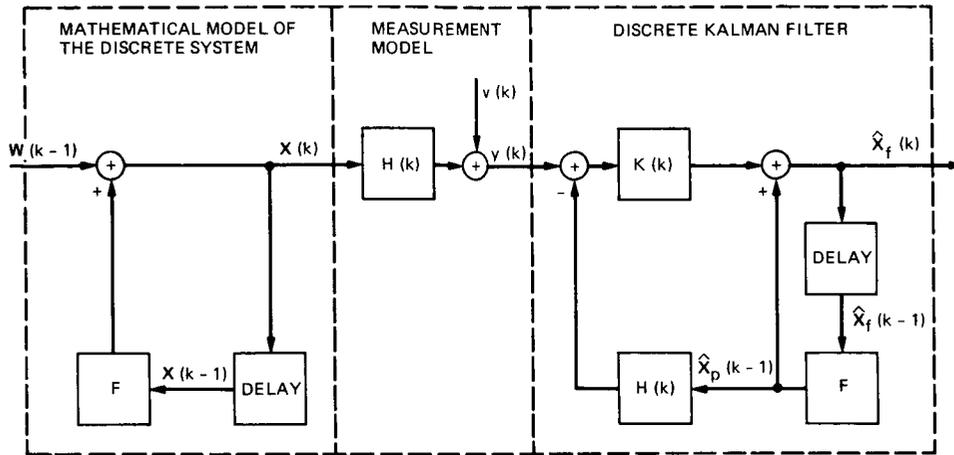
Fig. 4. Kalman filter timing diagram
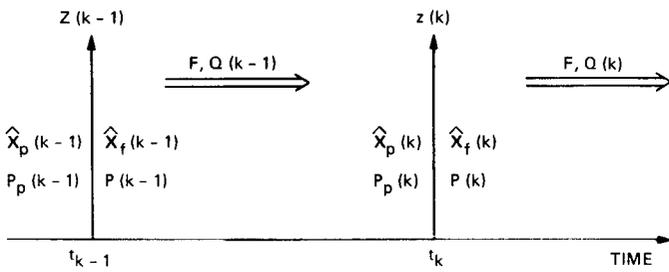
Fig. 5. System model of the discrete Kalman filter
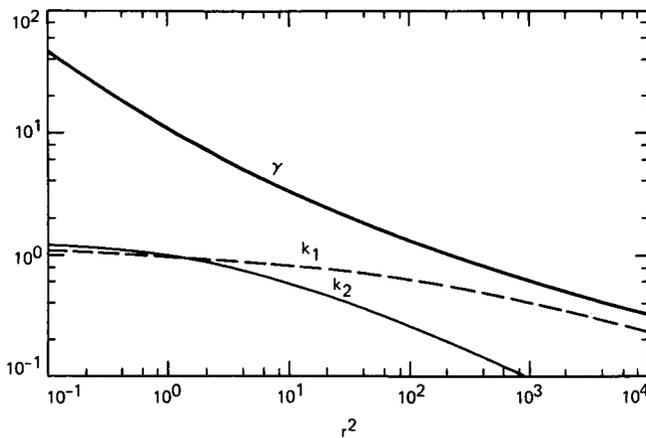
Fig. 6. Steady-state normalized variance of the Kalman filter estimated frequency and gains vs r^2
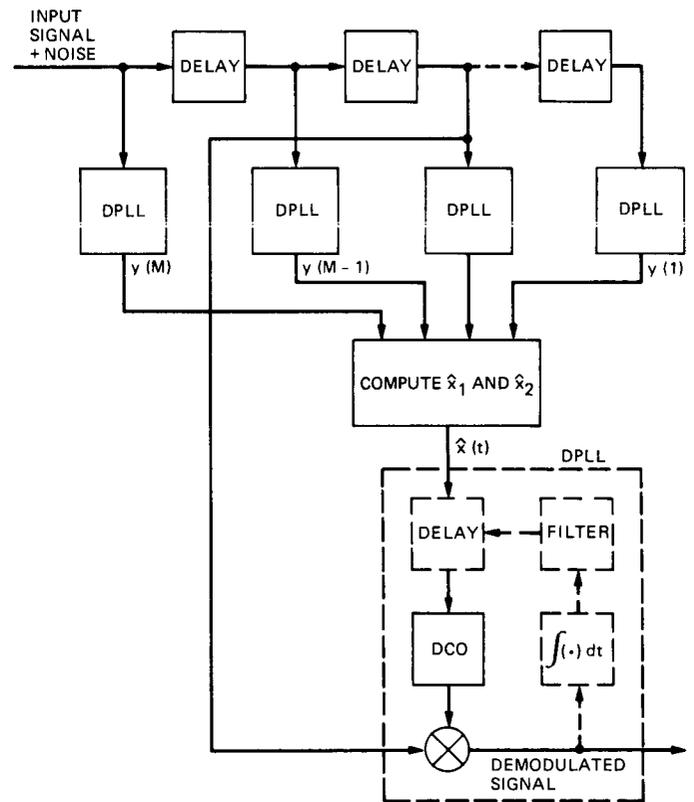
Fig. 7. Demodulation with several DPLLs — "parallel implementation"
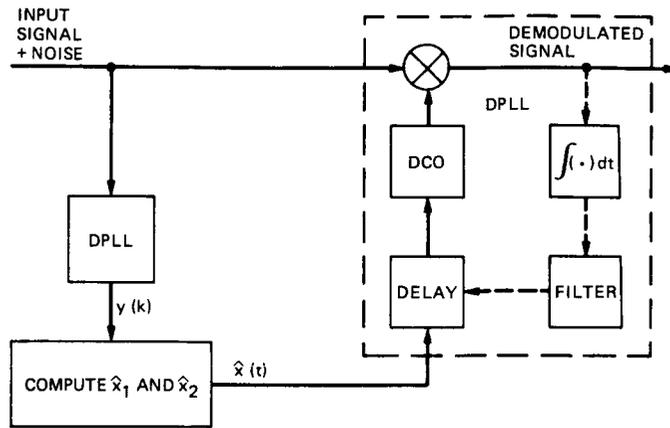
Fig. 8. Demodulation with two DPLLs—"serial implementation"
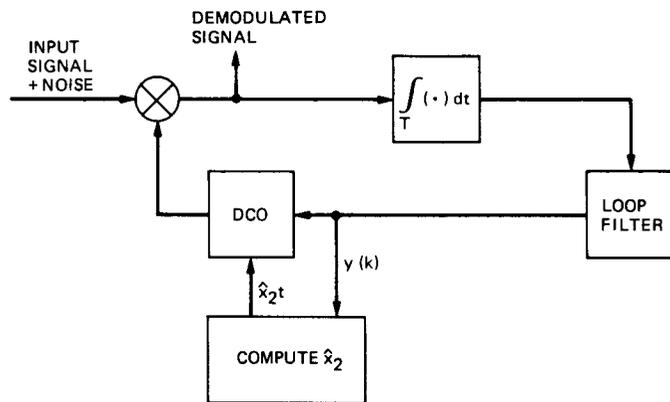
Fig. 9. DPLL with frequency rate compensation
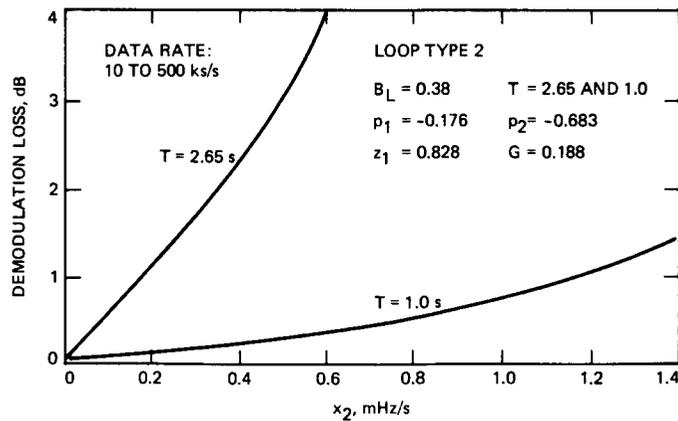
Fig. 10. Subcarrier demodulation losses vs. frequency rate (x_2)
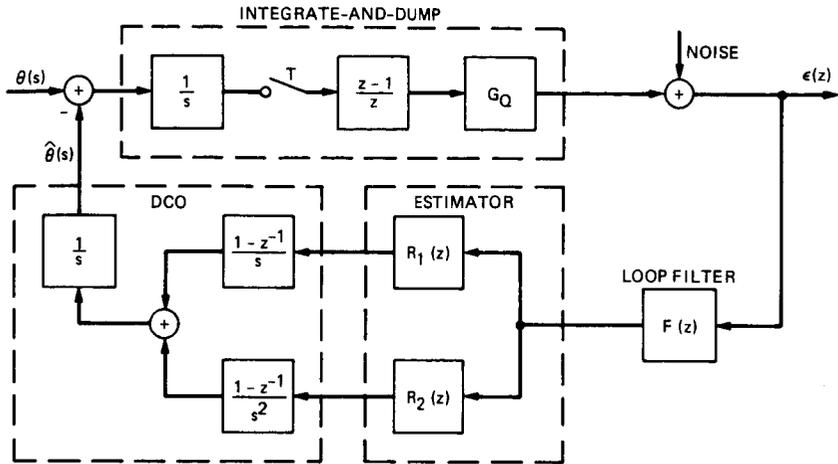
Fig. 11. Hybrid s/z diagram of a DPLL with a second-order estimator
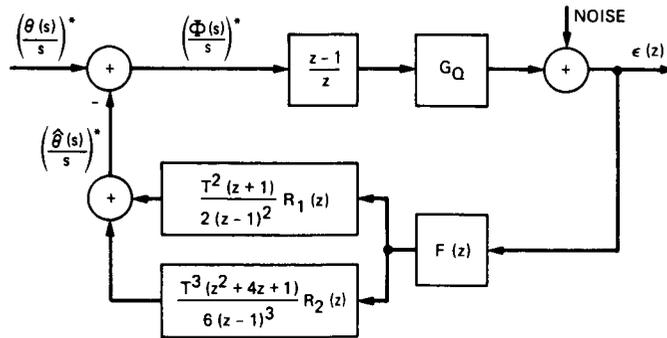
Fig. 12. Equivalent z-domain loop diagram

Appendix A

Transfer Functions for the Fixed-Window Least Squares Estimator and the Kalman Filter

The *fixed-window least squares* algorithm requires the sums

$$S_1 = \sum_{i=1}^{M} y(k) \quad \text{and} \quad S_2 = T \sum_{i=1}^{M} ky(k) \qquad \text{(A1)}$$

Defining z^{-1} as the delay operator, i.e., $y(k)z^{-1} = y(k-1)$, we can write the above sums as follows

$$S_1 = y(M) \sum_{i=0}^{M-1} z^{-i} \quad \text{and} \quad S_2 = Ty(M) \sum_{i=0}^{M-1} (M-i)z^{-i} \qquad \text{(A2)}$$

Performing the above summations, we get

$$S_1 = y(M) \frac{z^M - 1}{z^{M-1}(z-1)}$$

and

$$S_2 = Ty(M) \frac{Mz^{M+1} - z^M(M+1) + 1}{z^{M-1}(z-1)^2} \qquad \text{(A3)}$$

Taking the z-transform of the above expressions we obtain

$$S_1(z) = Y(z) \frac{z^M - 1}{z^{M-1}(z-1)}$$

and

$$S_2(z) = TY(z) \frac{Mz^{M+1} - z^M(M+1) + 1}{z^{M-1}(z-1)^2} \qquad \text{(A4)}$$

Combining Eq. (9) with the above equations, we finally obtain the desired transfer function for the fixed-window least squares estimator, namely,

$$R_1(z) \triangleq \frac{\hat{X}_1(z)}{Y(z)} = \frac{2}{M(M-1)}$$

$$\times \left[\frac{(2M+1)(z^M-1)}{z^{M-1}(z-1)} - \frac{3(Mz^{M+1} - z^M(M+1)+1)}{z^{M-1}(z-1)^2} \right]$$

and

$$R_2(z) \triangleq \frac{\hat{X}_2(z)}{Y(z)} = \frac{2}{M(M-1)}$$

$$\times \left[-\frac{3(z^M-1)}{Tz^{M-1}(z-1)} + \frac{6(Mz^{M+1} - z^M(M+1)+1)}{T(M+1)z^{M-1}(z-1)^2} \right]$$

$$\text{(A5)}$$

To obtain the transfer function of the *Kalman filter,* we insert Eqs. (26), (20), and (21) into Eq. (23) and obtain

$$\begin{bmatrix} \hat{x}_1(k) \\ \hat{x}_2(k) \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_1(k-1) \\ \hat{x}_2(k-1) \end{bmatrix}$$

$$+ \left\{ y(k) - \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{x}_1(k-1) \\ \hat{x}_2(k-1) \end{bmatrix} \right\} \begin{bmatrix} k_1(k) \\ k_2(k) \end{bmatrix}$$

$$\text{(A6)}$$

In steady-state, $k_1(k) = k_1$ and $k_2(k) = k_2$. Taking now the z-transform of the above equation and recombining terms, we obtain

$$\hat{X}_1(z)(1 - z^{-1}(1-k_1)) = \hat{X}_2(z)Tz^{-1}(1-k_1) + k_1Y(z)$$

$$\text{(A7)}$$

$$\hat{X}_2(z)(1 - z^{-1}(1-k_2T)) = -\hat{X}_1(z)k_2z^{-1} + k_2Y(z)$$

Writing the above equation in matrix form and performing matrix inversion, we finally obtain the transfer function for the Kalman filter, namely,

214

$$R_1(z) = \frac{\hat{X}_1(z)}{Y(z)}$$

$$= \frac{z(zk_1 + k_2T - k_1)}{z^2 + z(-2 + k_1 + k_2T) + (1 - k_1)}$$

$$R_2(z) = \frac{\hat{X}_2(z)}{Y(z)}$$

$$= \frac{zk_2(z - 1)}{z^2 + z(-2 + k_1 + k_2T) + (1 - k_1)} \qquad \text{(A8)}$$

The above transfer functions are needed when the least-squares estimator or the Kalman filter becomes part of a digital phase-locked loop (DPLL). For future reference, Fig. 11 shows the building blocks of a DPLL with the frequency and frequency-rate estimator incorporated in the loop. This block diagram is in the hybrid s/z transform domain. Figure 12 shows the corresponding z-domain block diagram, which is obtained from Fig. 11 using techniques of Ref. 4. In Figs. 11 and 12, zero computation time was assumed. Using techniques of Ref. 4, the loop filter $F(z)$ can be designed so that the estimator-enhanced DPLL will have optimum stability and bandwidth characteristics. All of these will be the subject of a future analysis.

VLA Feedhorn for Voyager Encounter of Neptune

F. Manshadi, D. A. Bathker, and H. W. Marlin

Radio Frequency and Microwave Subsystems Section

A high gain, low noise corrugated feedhorn was designed and developed by JPL for use in Very Large Array (VLA) antennas, near Soccoro, New Mexico. The new feedhorn will enable the VLA to support the Voyager encounter of Neptune in August of 1989. This will significantly enhance the receiving capability of the United States for that historic event.

I. Introduction

The original mission of the two Voyager spacecraft was to fly by Jupiter and Saturn. However, after they successfully completed their objectives at those two planets, Voyager 2 was rerouted to pass by Uranus and Neptune. Figure 1 shows the trajectories of the Voyager spacecraft from launch in 1977 through 1989. As can be seen from the figure, the Voyager 2 signal suffers a significant decline at the times of the encounter with Uranus and Neptune due to large increases in distance. During the Uranus encounter in January of 1986, this decline in signal power was compensated for by on-board control software modifications and increasing the receiving antenna aperture by arraying existing DSN ground stations as well as the Parkes, Australia (CSIRO) radio telescope. However, the 3.5-dB decline in the signal from Uranus to Neptune indicates additional improvement is needed in receiving capability. To accomplish this, further arraying of the DSN antennas to other large aperture systems such as the VLA in New Mexico is planned.

The VLA is owned and operated by the National Radio Astronomy Observatory (NRAO). It consists of 28 dual-shaped Cassegrain reflector antennas, each 25 m in diameter. On the basis of simple area alone, the VLA equals over four 64-m-diameter reflectors. Noise and other differences may change the four factor somewhat, but the conclusion that VLA represents a significant asset to the Voyager Neptune encounter tracking is inescapable.

A major modification plan was drawn for the VLA antennas to make them capable of efficiently receiving the Voyager 8.4 GHz (X-band) signal. Part of these modifications were new 8.0 to 8.8 GHz high performance feedhorns for all VLA antennas. The feedhorns were to be designed and developed by JPL and to be installed and owned by NRAO.

The following sections describe the design criteria, physical dimensions, and theoretical and measured electrical characteristics of these feedhorns.

II. Design Factors

Functionally, the feedhorn must efficiently couple the 25-m dual-shaped reflector system previously designed by NRAO to a specified location designated as the low noise amplifier reference (input) flange. The amplifier flange is a circular waveguide carrying either right- or left-circularly polarized waves, located about 1.9 m (75 in.) below the Casse-

grainian focus. The refrigerated amplifier is to be a dual unit with a cryostat-integral polarizer-orthomode. By efficiently couple, we mean maximizing the antenna gain to the system noise temperature quotient. In 1989, Voyager will appear low (30-deg elevation maximum) in the New Mexico sky, and the noise level of the amplifiers was not well known in 1984 when the feedhorn design was done. Accordingly, a noise budget was constructed as follows. Although early model Gallium Arsenide Field Effect Transistor (GASFET) amplifiers were available with 30-K amplifier noise levels (leading to approximately 50-K overall system noise levels) an approach based on expected future high electron mobility transistor (HEMT) development was elected. Present HEMT projections indicate amplifier noise will likely be 20 K, leading to approximately 40-K overall system levels at an elevation angle of 30 deg. The feedhorn design was therefore optimized for a system of approximately 40 K. The noise budget is shown in Table 1.

The design proceeded by computational trial-and-error using the horn aperture variable. The previously determined dual-shaped surface overall geometry and profiles of the VLA 25-m element were not available as design variables. Ordinarily, JPL low noise reflector antenna design is carried out by synthesizing the overall geometry and profiles for a previously proven horn design. Invariably, the JPL synthesis technique results in too much spillover of the subreflector scattered radiation pattern (viewed in reciprocal transmission), therefore collecting ground noise in reception via spillover. JPL then synthesizes an additional annular ring around the main reflector periphery as a noise shield, balancing minimum size against the noise screening benefit. In the VLA case, this was not possible, and the feedhorn aperture was the only design variable of any importance. Of perhaps third order impact was the horn flare angle. Since the design factor of a 1.9-m separation of the Cassegrainian focus from amplifier flange needed to be met, and since nearly all this length was needed for the horn, it was decided to use essentially the total available length as the tapered portion of the horn in order to minimize the dissipation noise of the small-diameter (25.9 mm or 1.02 in.) connecting circular waveguide. Thus, the horn design consisted of iteratively varying aperture size, flare angle, and phase center location to fit the available axial space, to place the horn phase center at the subreflector forward focus, and provide maximum G/T for the approximately 40-K system noise level.

The software used to accomplish this design consists of the JPL developed HYBRID HORN, RUSCH SCATTERING, and EFFICIENCY programs (Refs. 1, 2, 3). A simplified symmetric analysis provided by the Rusch scattering software is considered sufficiently accurate in this instance, even though the VLA feeds are disposed asymmetrically (similar to the NASA-JPL 64-m tricone arrangement).

The horn satisfying the above (for a 40-K total system) is 334.3 mm (13.16 in.) in diameter with a flare angle of 4.28 deg. The midband (8.4 GHz) zenith spillover is computed as 2.3 K, indicative of the higher (than usual for JPL) system noise. That is, from experience, usual JPL systems are sized to produce only 0.7-K zenith spillover noise. Such sizing optimizes G/T for a lower noise system of about 16 to 18 K.

Figure 2 shows the selected horn radiation pattern, and Figure 3 shows the VLA-shaped subreflector scatter pattern, both at 8.4 GHz. Table 2 shows performance factors for the horn alone. Table 3 shows performance factors related to the scatter pattern. Both Tables 2 and 3 cover approximately 10% bandwidth, of interest to NRAO for purposes in addition to Voyager tracking.

For a single narrow band system of about 5%, a smooth walled dual-mode-type feedhorn would have been adequate. The NRAO desired a usable bandwidth of 8.0 to 8.8 GHz. Considerable analytical study of the band-edge performance roll-off of a dual-mode horn was accomplished in order to potentially reduce fabrication costs. It was found the bandwidth requirement of approximately 10% could not be adequately met with a smooth walled dual-mode horn because of poor efficiency and high noise. For this reason, a conventional corrugated horn was necessary, despite increased manufacturing costs. Table 4 shows the theoretical band-edge performance degradation using a smooth walled dual-mode horn carefully tailored to require no throat end phasing section. When bandwidth is of no concern, common practice with dual-mode horns is to size the aperture, select a flare angle, and then bring the two waveguide modes (TE_{11} and TM_{11}) into proper aperture phase relationship by use of a throat-end phasing section. The high dispersion (different velocities) of the two modes in a small waveguide enables a convenient differential phase shifter at one frequency. Over a wideband, however, a small diameter differential phase shifter is counterproductive. The analytical study of a smooth walled dual-mode horn conducted as part of this work therefore carefully arranged the aperture size and flare angle such that no small phasing section was necessary—the horn alone provided the necessary phase shifting, thereby giving best bandwidth performance. However, as seen in Table 4, about 1.2-dB gain at 8.0 GHz and 1.0-dB gain at 8.8 GHz are lost due to degradation in the smooth walled dual-mode horn radiation pattern caused by mode asynchronism away from design frequency. We conclude that the corrugated horn is necessary for the 8.0- to 8.8-GHz application. Despite elimination of the throat end high dispersion phasing section, the rather large (+27.5 dBi) horn still limits the bandwidth of a smooth walled dual-mode horn to approximately 5% for approximately a 10% (-0.5 dB) performance roll-off.

III. Start-up Configuration

Prior to analytical investigation of smooth walled horn performance and prior to final design of the selected corrugated horn, it was necessary to conduct certain systems-level testing at VLA using the Voyager 8.4 GHz (X-band) signal. Testing involved questions centering on signal summing from at least two VLA elements. In order to accomplish this, two microwave horns were needed on a short timescale. JPL had previously developed a large aperture X-band horn with 406.8-mm (16.02-in.) aperture and 6.25-deg flare angle. Although too large in diameter and not long enough for the VLA application, modifications were made by application of JPL software. First, a discrete section of the horn was removed, yielding an aperture near 315 mm (12.38 in.), which was not optimum but acceptable. Second, the remaining horn was juggled fore and aft, balancing the accruing defocusing loss from the misplaced horn phase center to subreflector forward (Cassegrain) focus against the noise increase due to the circular waveguide needed between the horn throat and amplifier location. The final design for temporary service was calculated to be within 0.3 dB of optimum (about 82.6% efficiency on the same basis as Table 3). Horn number one was available from an earlier program and the DSN Advanced Systems Program supplied resources to construct the second. Thus, an early start-up configuration was provided for systems-level testing in 1985.

IV. Feedhorn Physical Configuration

The VLA feedhorn assembly consists of seven sections as shown in Figure 4. Section 1 is a cosine taper, providing a well-matched transition from the VLA amplifier's standard circular waveguide (1.02 in. in diameter) to the JPL horn standard circular waveguide (1.369 in. in diameter) as shown in Figure 5. This taper was designed by an existing JPL program. The tight tolerance on the outside diameter of this section was required by the VLA for mounting purposes. Section 2 (the input section) is a corrugated waveguide used for matching the smooth wall waveguide to the corrugated feedhorn. This part is designed using the method discussed in Section 5.5 of Ref. 4. Sections 3 through 6 make up the main body of the horn and are corrugated uniformly with a constant flare angle of 4.283 deg. Figure 6 shows Section 3 of the horn assembly with detailed dimensions of the corrugations. Section 7 is the radome window and its accessories, as shown in Figure 7. JPL feedhorn windows are ordinarily made from thin Kapton® (trademark of the Dupont Corporation) sheets. However, due to occurrence of heavy hailstorms at the VLA site, a half-wavelength-thick Teflon window was designed for this feed. The window is epoxied to an aluminum frame and then clamped to the horn aperture flange with a rubber O-ring in between. The clamp is used to facilitate the removal or assembly of the window for servicing the feed. The rubber O-ring seals the horn from dust and moisture. By VLA request, two bleed holes are provided: one in the cosine taper and the other near the horn aperture. The bleed holes are used to purge moisture out of the horn by running either dry air or nitrogen through it.

All aluminum parts are fabricated from high quality 6061-T6 alloy. Details of the feedhorn assembly and individual parts are shown in JPL Drawings 9488078, 138372, 138373, and 9488080 (JPL internal documents). Figure 8 shows one complete feedhorn before it was shipped to the VLA site.

V. Measurements

The dissipative loss of the VLA feedhorn was measured by a maser low noise amplifier installed on a rooftop. The sky temperature was measured first using a known feedhorn and then by the VLA feed. From this measurement, the dissipative loss of the VLA horn, without any radome, was found to be 0.023 dB at 8.4 GHz. The additional loss due to the Teflon window was approximately 0.006 dB. The return loss of the horn is measured by an HP 8510 automatic network analyzer. For the feed without the Teflon window, the best return loss is obtained near the upper edge of the 8.0- to 8.8-GHz band (see Figure 9). Therefore, the Teflon window was designed approximately 0.005 in. thicker than the half wavelength at 8.4 GHz (0.488 in.). This adjustment caused the best return loss to be between 8.4 and 8.5 GHz, which is the band of interest for DSN applications. Figure 10 shows the return loss of the horn with an 0.488-in.-thick Teflon window to be better than 34.6 dB (equivalent to VSWR of lower than 1.04:1) between 8.4 to 8.5 GHz.

The far-field pattern of the feed was measured in the JPL 60-ft anechoic chamber. The E- and H-plane patterns, with and without the Teflon radome, are shown in Figs. 11 and 13. The 45-deg-plane co-polar and cross-polar patterns for the same configurations are shown in Figs. 12 and 14. These patterns exhibit low sidelobes and cross-polarization levels and are very close to the theoretically predicted patterns of Fig. 2. Moreover, it can be seen that the Teflon window does not change the far-field pattern of the feed significantly. The slight degradation of 45-deg cross-polarization pattern (worst cross-polar level on axis), for the horn with the Teflon radome, is due to radiation from the non-circularly symmetric clamp (see Figs. 7 and 8).

VI. Summary

A low noise, high performance feedhorn has been designed and developed by JPL for the VLA. The feed shows high performance, very good match, and low loss over the frequency band of interest. The measured data has excellent agreement with the theoretical data.

Acknowledgment

References

1. Moyer, T. D., *Mathematical Formulation of Double-Precision Orbit Determination Program (DPOP)*, TR 32-1527, Vol. XIII, Jet Propulsion Laboratory, Pasadena, Calif., Feb. 1973.

2. Potter, P. D., Antenna feedhorn software upgrade, *The Deep Space Network Progress Report 42-51*, 75-84, Jet Propulsion Laboratory, Pasadena, Calif., June 1979.

3. Ludwig, A., *Computer Programs for Antenna Feed System Design and Analysis*, TR 32-979, Jet Propulsion Laboratory, Pasadena, Calif., April 1967.

4. Clarricoats, P. J. B., and A. D. Olver, *Corrugated Horns for Microwave Antennas* (IEE Electromagnetic Wave Series 18), London, Peter Peregrinus, Ltd., 1984.

Table 1. JPL/VLA 8.4-GHz Feedhorn Design System Noise Budget

Parameter	Noise, K
Amplifier	20
30-deg elevation atmosphere	6
Galactic	3
Quadripod scatter (30°)	5
Horn and other dissipation	3
Spillover (30°)	(1/2 of zenith value; value to be determined in G/T optimization)
Total	Approx. 40

Table 2. Theoretical Horn Selected Performance at 9-deg Half Angle

Frequency, GHz	Taper, dB	Phase center, behind aperture, mm (in.)	Beam efficiency, %	Directivity, dBi
8.0	–15.8	–263 (–10.37)	0.956	+27.13
8.4	–16.65	–315 (–12.42)	0.962	+27.53
8.8	–18.07	–372 (–14.64)	0.966	+27.92

Table 3. Theoretical Scatter Pattern VLA Subreflector With JPL Horn

Frequency, GHz	Overall efficiency, %	Zenith noise, K
8.0	89.2	3.05
8.4	89.6	2.27
8.8	89.4	1.66

Note: Overall efficiency includes illumination amplitude and phase, cross-polarization, spillover, and central blockage. Not included are surface tolerance, quadripod blockage, and dissipation factors.

Table 4. Smooth Wall Dual-Mode Horn VLA Subreflector Scatter Patterns Performance Summary

Efficiency factors	Frequency, GHz				
	8.0	8.2	8.4	8.6	8.8
Forward spillover	0.793	0.862	0.937	0.926	0.863
Rear spillover	0.983	0.983	0.988	0.997	0.989
Noise (rear) at zenith, K	3.3	3.62	2.62	2.05	2.30
Illumination	0.980	0.986	0.980	0.966	0.947
X-polarization	0.983	0.999	0.998	0.980	0.946
Phase	0.913	0.959	0.996	0.984	0.935
Efficiency	0.674	0.786	0.887	0.838	0.700
Total gain degradation, dB	–1.2	–0.5		–0.2	–1.0

Table 5. Amplitude and Phase Coefficients for Dual-Mode Horn Mode Generator*

Frequency, GHz	TE_{11} Mode		TM_{11} Mode	
	Amplitude	Phase	Amplitude	Phase, deg
8.0	1.0	0.0	0.401	67.5
8.2	1.0	0.0	0.653	51.3
8.4	1.0	0.0	0.766	42.2
8.6	1.0	0.0	0.831	35.8
8.8	1.0	0.0	0.871	30.8

*34.77/46.02-mm (1.369/1.812-in.) step discontinuity.
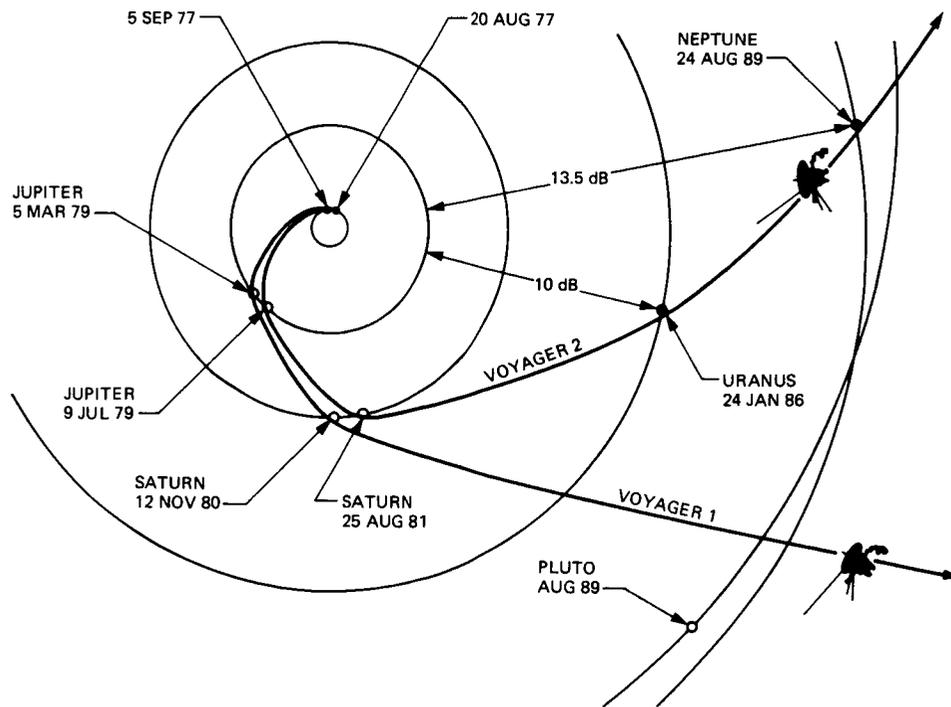
Fig. 1. Voyager heliocentric trajectory
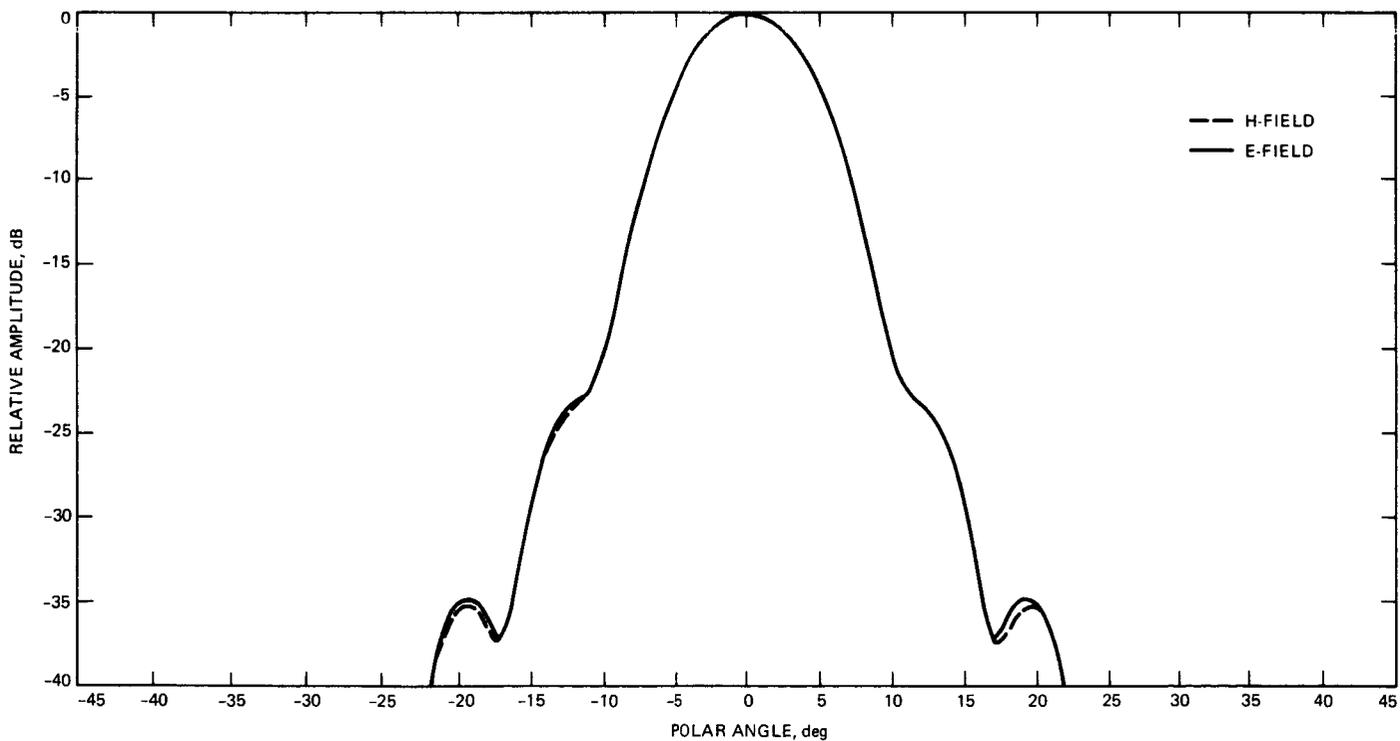
Fig. 2. VLA feedhorn theoretical pattern at 8.4 GHz
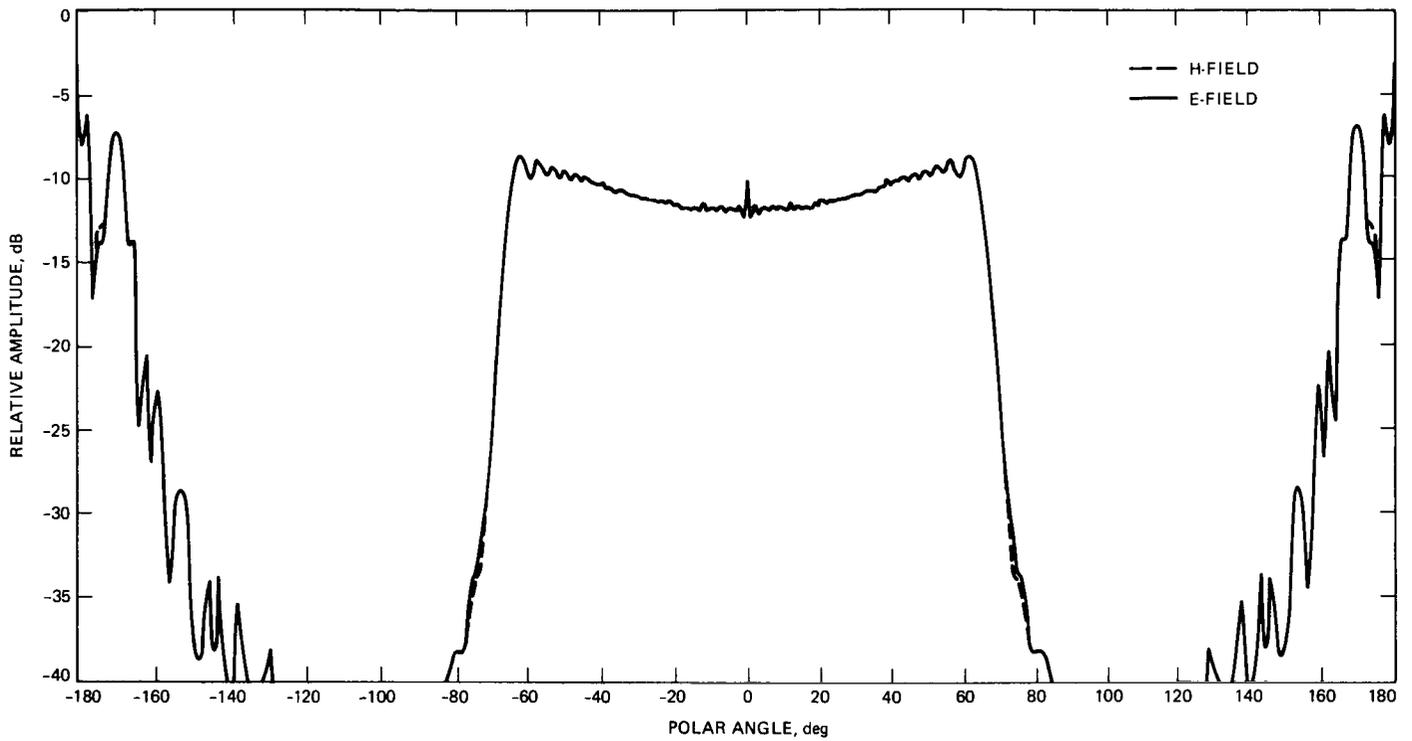
Fig. 3. VLA shaped subreflector scatter pattern

Fig. 4. VLA feedhorn assembly

Fig. 5. VLA feedhorn cosine taper

Fig. 7. Teflon window and accessories

Fig. 6. Section 3 of the feedhorn assembly

Fig. 8. VLA feedhorn assembly

223
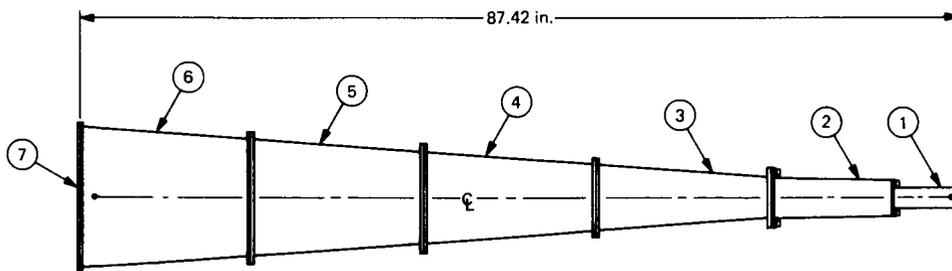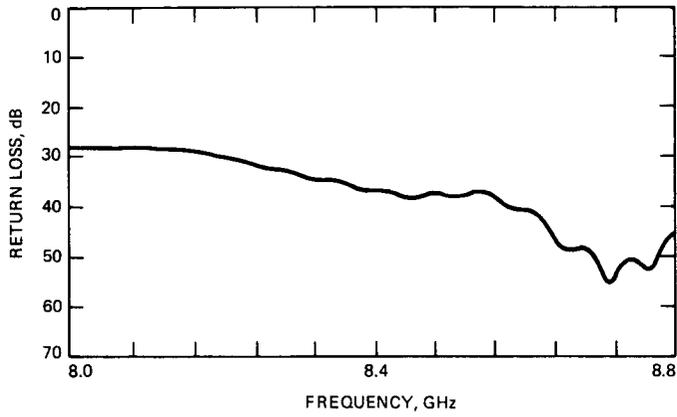
Fig. 9. Return loss of VLA feedhorn without window
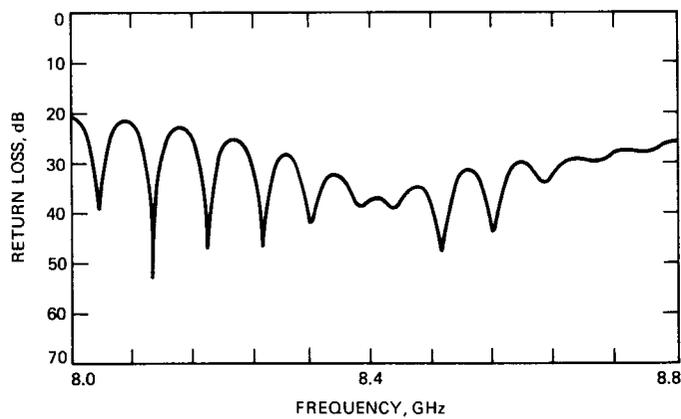
Fig. 10. Return loss of VLA feedhorn with 0.488-in.-thick window
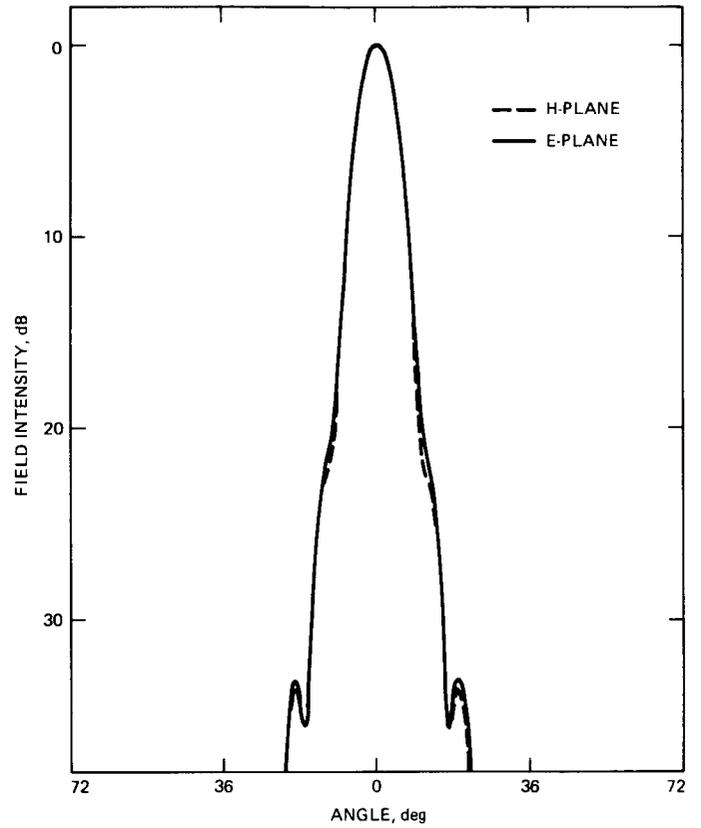
Fig. 11. E-plane and H-plane patterns of the VLA feedhorn without radome at 8.4 GHz
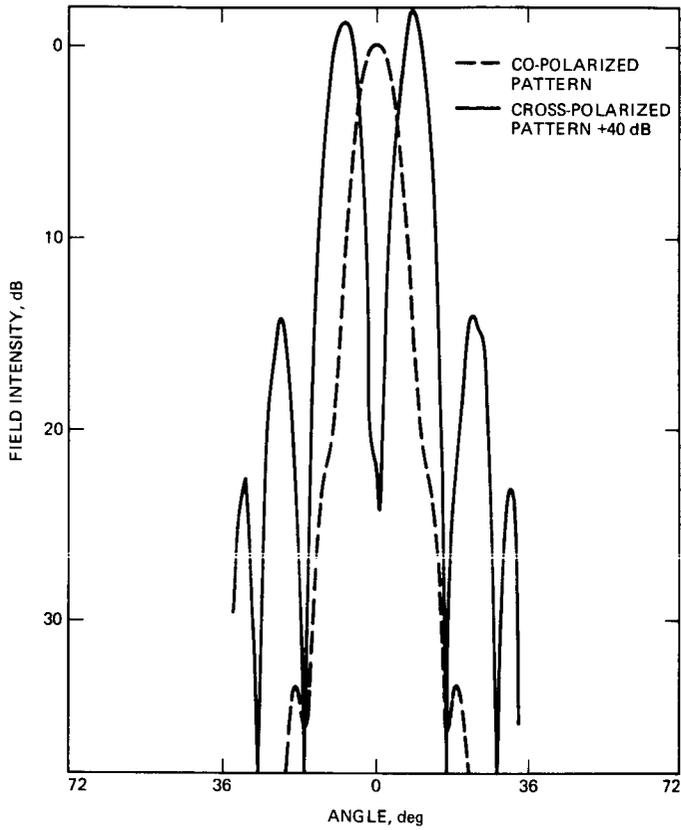
Fig. 12. The 45-deg plane co-polar and cross-polar patterns of the VLA feedhorn without radome at 8.4 GHz
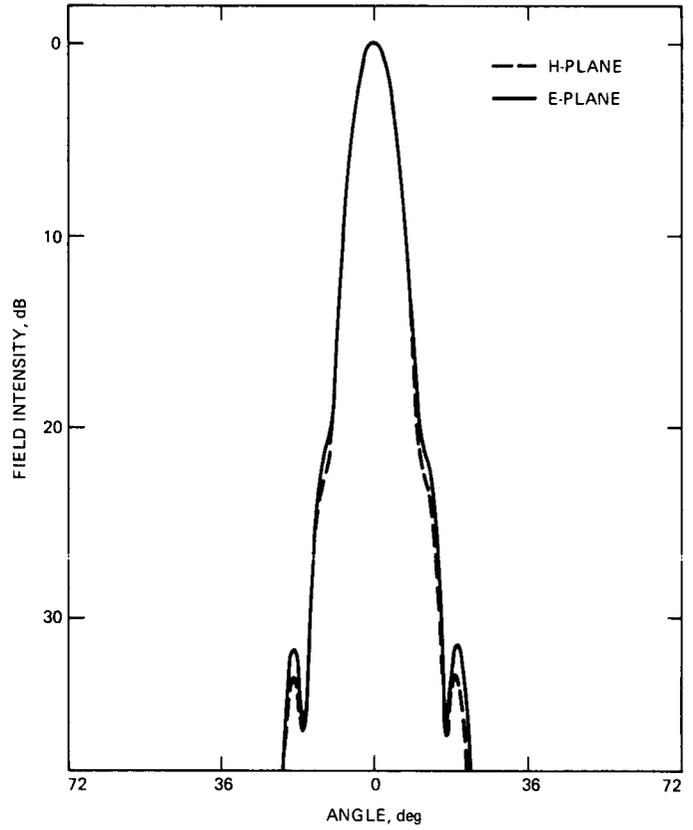
Fig. 13. E-plane and H-plane patterns of the VLA feedhorn with 0.488-in.-thick Teflon radome at 8.4 GHz
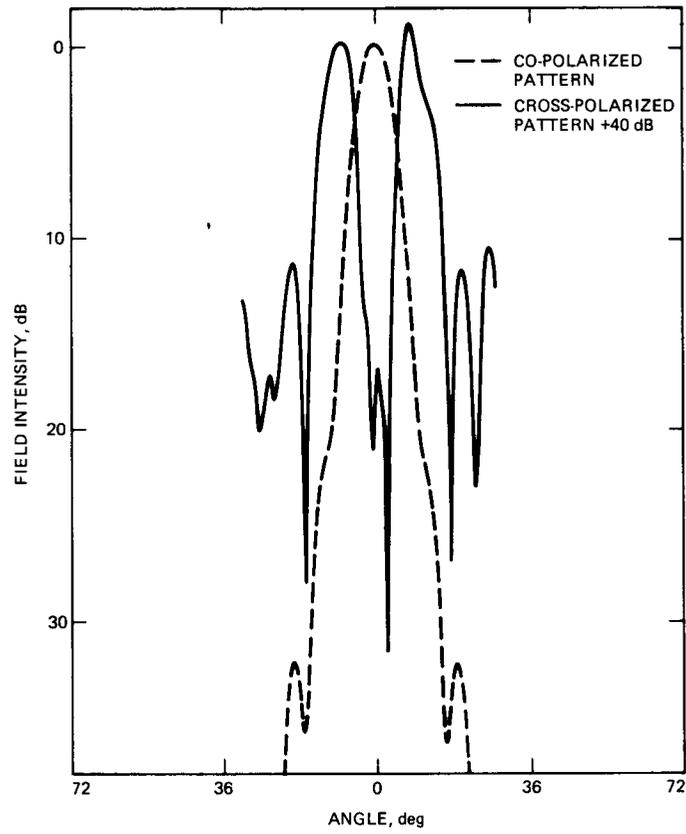
Fig. 14. The 45-deg plane co-polar and cross-polar patterns of the
VLA feedhorn with 0.488-in.-thick Teflon radome at 8.4 GHz

X-Band Preamplifier Filter

F. Manshadi

Radio Frequency and Microwave Subsystems Section

This report describes a low-loss bandstop filter designed and developed for the Deep Space Network's 34-meter high-efficiency antennas. The filter is used for protection of the X-band traveling wave masers from the 20-kW transmitter signal. A combination of empirical and theoretical techniques was employed as well as computer simulations to verify the design before fabrication.

I. Introduction

The X-band preamplifier filter (XPF) is a bandstop filter used for protection of the traveling wave masers of the DSN 34-meter high-efficiency antennas from the high-power (20-kW) transmitter signal. The requirements are as follows:

Rejection = 70 dB from 7.145 to 7.235 GHz

VSWR = 1.05:1 from 8.4 to 8.5 GHz

Insertion loss = 0.05 dB (0.03 dB goal) from 8.4 to 8.5 GHz.

During design and measurements, consideration was also given to the entire 8.2- to 8.6-GHz band in order to permit wideband VLBI without significant degradation compared to the present non-filtered (no transmitter) systems at DSS 15 and DSS 45.

The theory of passive microwave bandstop filters is well known and documented in numerous papers. It has also been compiled in a book by Matthaei et al. (Ref. 1). These filters are traditionally designed based on lumped-element prototype filters such as the one shown in Fig. 1. The component values are determined by considering either maximally flat or equi-ripple characteristics for the filter. At microwave frequencies and for waveguide filters, the resonant circuits are most often realized by cavities coupled to the top or the side wall of a waveguide through some irises. Over narrow frequency bands, these cavities behave very similarly to lumped-element resonant circuits such as the one shown in Fig. 1. The number of resonant circuits used is a function of the stopband rejection requirements. In practice, because of the physical size of waveguide cavities as well as for simplicity, either parallel or series resonant circuits are used and the adjacent cavities are spaced a multiple of quarter wavelength apart.

For the design requirement of 70-dB rejection over the stopband, figures 4.03-4 through 4.03-10 of Ref. 1 suggest that a four-cavity filter would be marginal. Therefore, it was decided to use five cavities. Additionally, since for our application the passband is well away from the stopband, designing the filter based on maximally flat or equi-ripple criteria would not be advantageous. Therefore, for ease of fabrication, all cavities were assumed to be identical. Finally, for better tuning of the passband as well as the stopband, the cavities were coupled to the side wall as shown in Fig. 2.

II. Design Procedure

For the filter of Fig. 2., values of L_c, D_I, and D_c are to be determined. In order to model the cavity waveguide junction, the return loss and transmission loss of a one-cavity filter was measured for several values of D_I and L_c. Typical curves for each value of D_I and L_c are shown in Figs. 3(a) and 3(b). By use of a simple computer code, these curves were fitted to the response of a series inductance and capacitance and the equivalent lumped inductance and capacitance for each cavity/waveguide junction was computed. These values were later used for simulation of multi-cavity filters. Additionally, a series of curves were obtained showing the variation of the resonant frequency as a function of D_I and L_c (Fig. 4). For the XPF, the stopband is centered at 7.190 GHz; however, the filter was designed for 7.240 GHz to allow for a 50-MHz tuning capability in the stopband. Figure 4 shows that an infinite number of D_I and L_c pairs yield the same resonant frequency. But, in selection of the optimum pair, it should be noted that the stopband transmission loss as well as the passband insertion loss is dependent on D_I. The larger D_I is, the larger are the tramsmission and insertion losses. Therefore, the optimum L_c and D_I pair is the one with the smallest value for D_I that will provide the required transmission loss over the stopband. This would guarantee the smallest possible insertion loss for the filter. The optimum pair of D_I and L_c was selected by use of the equivalent inductance and capacitance of cavity waveguide junction in a computer program that simulated the characteristics of a five-cavity filter. This computer program found the optimum value for cavity spacing D_c (corresponding to highest transmission loss) for each pair of D_I and L_c by varying D_c about three quarters of the guide wavelength, at 7.240 GHz. The optimum values for the filter were computed to be L_c = 0.82 in., D_I = 0.78 in., and D_c = 1.58 in. The computed stopband characteristic of the filter, after it is tuned to 7.190 GHz, is shown in Fig. 5. This figure shows a minimum transmission loss of approximately 72 dB from 7.145 to 7.235 GHz.

III. Test Results

The filter was fabricated from the standard WR125 stock with a tuning screw in each cavity (for the stopband) as well as waveguide/cavity junctions (for passband VSWR) as shown in Fig. 6. The filter was tested by the HP 8510 automatic network analyzer; its passband and stopband characteristics are shown in Figs. 7 through 9. The stopband transmission loss is better than what was computed, which is due to the fact that in the computer simulation the resistive loss of each cavity/waveguide junction was not taken into account. All other characteristics are better than the specification.

Acknowledgment

Reference

1. Matthaei, G., Young, L., and Jones, E. M. T., *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Artech House, Delham, MA, 1964.
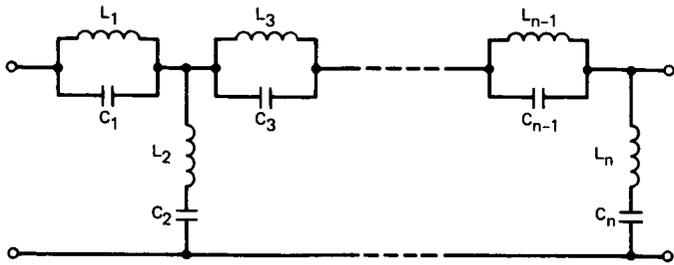
Fig. 1. A bandstop prototype filter
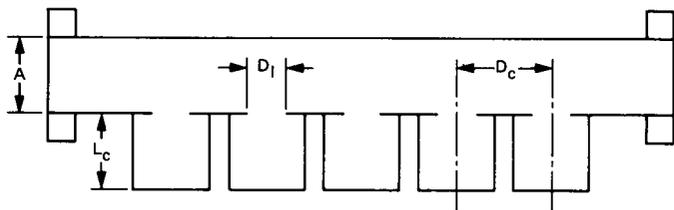
Fig. 2. Five-cavity bandstop filter in waveguide
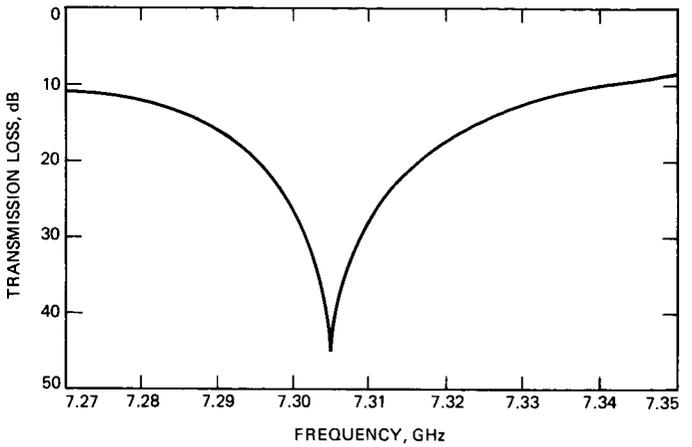
Fig. 3(b). Transmission loss for one-cavity bandstop filter
(D_I = 0.7", L_C × 0.85")
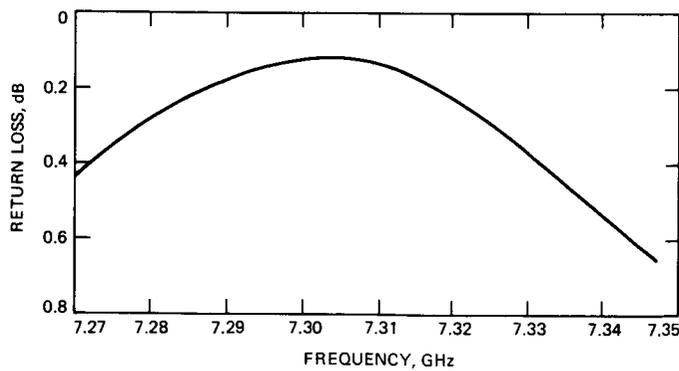
Fig. 3(a). Return loss for one-cavity bandstop filter
(D_I = 0.7", L_c = 0.85")
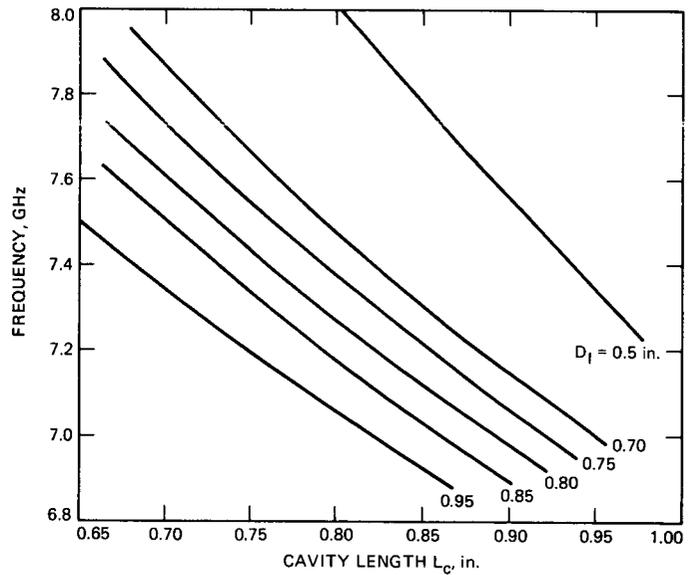
Fig. 4. Resonant frequency of one cavity coupled to waveguide
through an iris of size D_I vs cavity length L_C
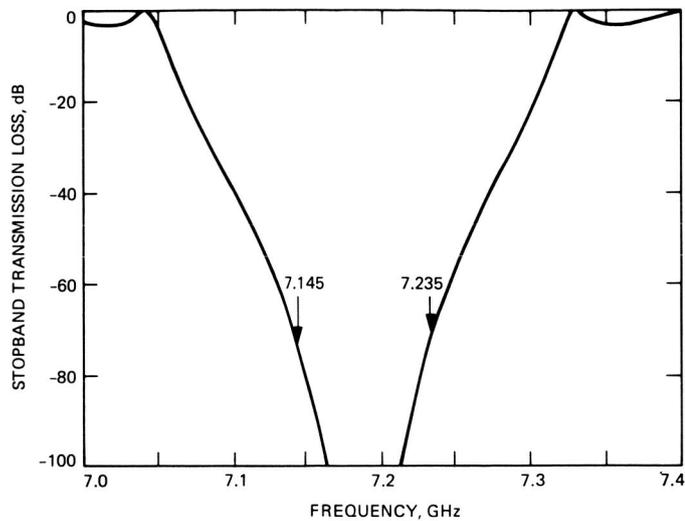
Fig. 5. Theoretical stopband transmission loss of the X-band preamplifier filter vs frequency
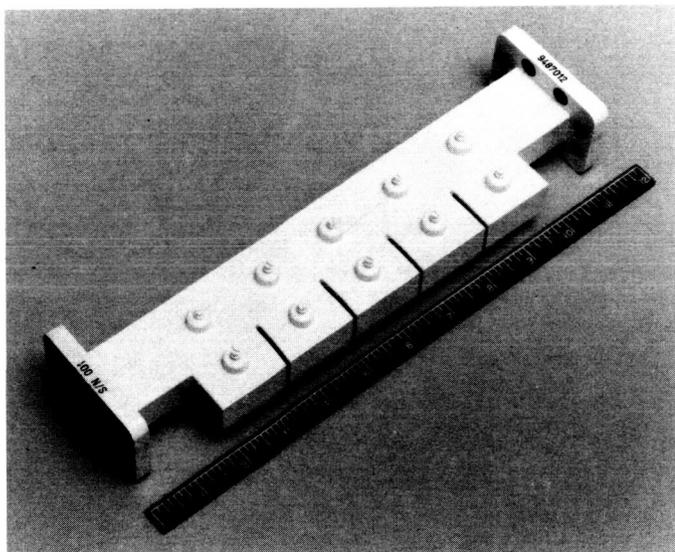
Fig. 6. X-band preamplifier filter: production unit
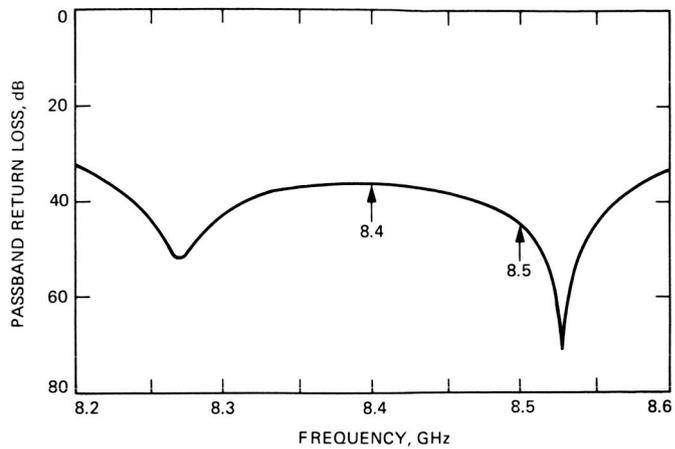
Fig. 7. Passband return loss of the X-band preamplifier filter vs frequency
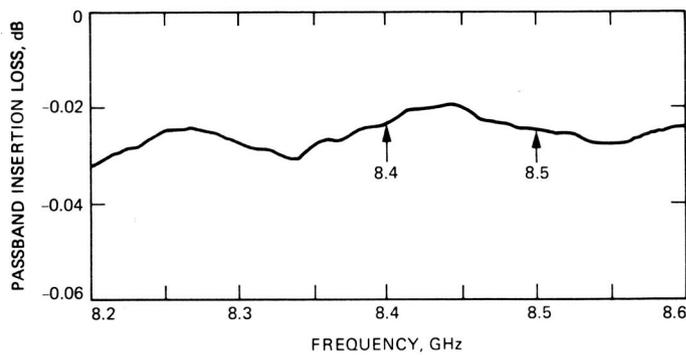
Fig. 8. Passband insertion loss of the X-band preamplifier filter vs frequency
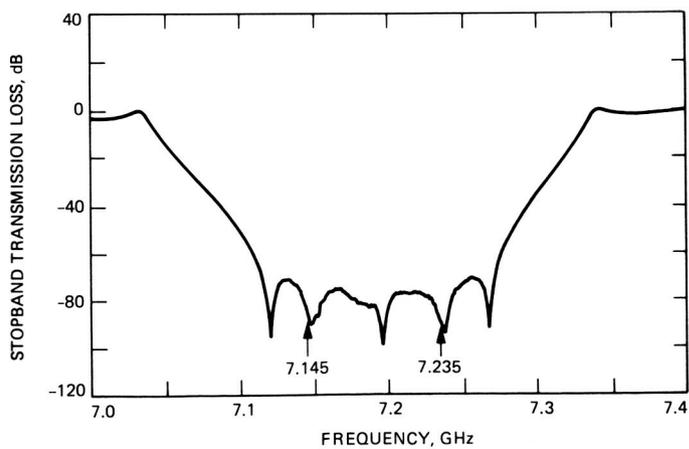
Fig. 9. Stopband transmission loss of the X-band preamplifier filter vs frequency

Physical Optics Analysis of a Four-Reflector Antenna Part 2

A. G. Cha

Radio Frequency and Microwave Subsystems Section

The results of a rigorous analysis of the DSN 70-m antenna S-band (2.295-GHz) RF performance are presented. Previous estimation of 1.6 dB S-band gain improvement of the 70-m antenna over the 64-m antenna has been revised to 1.5 dB by this analysis. The S-band RCP beam position offset relative to X-band (8.45-GHz) beam position is predicted to be 0.0045° (0.04 beamwidth). The effective S-band gain loss resulting from non-coincidence with the X-band beam is predicted to be 0.02 dB. Therefore, this is no longer a concern for the 64-m to 70-m upgrade project.

I. Introduction

The initial configuration of the DSN 70-m antennas, shown in Fig. 1, will retain the present 64-m S/X band (2.295/8.45 GHz) reflex dichroic feed system, but will replace the hyperboloid subreflector and parabolic main reflector by an asymmetric shaped subreflector and a symmetric shaped main reflector. In Part 1 of this article (Ref. 1), the 64-m antenna four-reflector S-band system was analyzed using a rigorous physical optics (PO) based procedure; theoretical predictions were found to be in excellent agreement with known 64-m antenna experimental information. The same analysis procedure is now applied to the 70-m shaped antenna four-reflector S-band system.

In Part 1 of this article, the NASA/JPL 64-m antenna RF performance was analyzed at S-band (2.295 GHz) using a rigorous physical optics approach. Excellent agreement with all known 64-m experimental characteristics was achieved, establishing the analysis procedure as a valid tool for predicting RF performance in arbitrary multiple reflector antenna systems. In this article, the same approach is extended to

analyzing the S-band performance for the upcoming 70-m antenna conversion, with particular interest in determining the S-band beam pointing direction. The uncertainty in S-band performance associated with the uncertainty of S-band beam position had prompted this study.

II. Comparisons of Diffraction and Polarization Characteristics for 64-m Classical Cassegrain vs 70-m Dual-Shaped Subreflectors

In Figs. 2 through 4, the 2.295-GHz diffraction patterns of the ellipsoid subreflector, the dichroic plate, the 64-m hyperboloid, and the 70-m asymmetric subreflector are shown. The ellipsoid E- and H-plane diffraction patterns have a small but observable asymmetry compared to the incident corrugated horn patterns. (The corrugated horn patterns are not shown but exhibit perfect E- and H-plane symmetry.) The ellipsoid diffraction pattern asymmetry is a manifestation of the presence of higher order azimuthal Fourier modes induced by

the asymmetric geometry of the horn and ellipsoid. This moding effect is the primary factor causing S-band (2.295-GHz) beam pointing direction of the 64-m and 70-m antennas to be polarization dependent (and to be offset from the X-band (8.45-GHz) beam pointing direction). The dichroic diffraction pattern is substantially the same as the ellipsoid diffraction pattern, as one would expect from a planar reflector. The hyperboloid diffraction patterns shown in Fig. 3 correspond to the 64-m subreflector, less its flange. As noted in Part 1, the flange is not expected to have any significant effect in beam pointing direction prediction. The hyperboloid diffraction pattern exhibits a tapered illumination from main reflector axis ($\theta = 0°$) to optical edge ($\theta = 60°$). By contrast, the 70-m antenna shaped subreflector diffraction patterns shown in Fig. 4 show the expected inverse illumination taper, which is needed to realize nearly uniform aperture illumination for high aperture efficiency. The center dip in the pattern (approximately, polar angle = $0°$ to $10°$ in Fig. 4) is due to a "vertex plate" design, which is estimated to improve 70-m S- and X-band gain by more than 0.1 dB.

Each diffraction pattern has been analyzed by a Fourier series expansion analysis to determine its azimuthal Fourier modes (m-modes) power content. This is shown in Fig. 5 and Table 1. It is well known that only the $m = 1$ mode contributes to antenna boresight gain, while the $m \neq 1$ modes lead to gain loss and cross-polarization radiation. The reference for viewing the mode content of each diffraction pattern is the corrugated horn pattern illuminating the ellipsoid. For most practical purposes, nearly 100% of the power radiated by the corrugated horn is in the $m = 1$ mode. After reflection from the ellipsoid, it is seen that about 2% of the power is converted into the $m = 0$ and $m = 2$ modes. The dichroic reflector diffraction pattern and mode power contents are essentially the same as those of the ellipsoid. Both the 64-m antenna hyperboloid and the 70-m antenna shaped subreflector are illuminated by the dichroic diffraction pattern. Although the two subreflector diffraction patterns, Figs. 3 and 4, look drastically different, the amount of power in the undesirable higher order m modes is approximately the same (3%) in both cases.

The computed 70-m S-band (2.295-GHz) far-field radiation pattern is shown in Tables 2 and 3 in two perpendicular planes (defined as $\phi = 180°$ and $270°$). Similar to the 64-m antenna case presented in Part 1, beam position offsets in these two planes correspond respectively to the 2.54-cm (1-in.) shim of the dichroic position and the depolarization effect of the $m \neq 1$ modes on circularly polarized waves (RCP in the present case). Figure 6 shows the predicted 70-m antenna S-band beam position relative to the known 64-m antenna X- (8.45-GHz) and S-band positions. The beam position offset in the $\phi = 270°$ plane is approximately the beam separation

between the X- and S-band beams. This is computed to be $0.0045°$ (equal to 4% of the 70-m antenna S-band beamwidth). Assuming the antenna is boresighted by the X-band beam, an S-band equivalent pointing loss of 0.02 dB is to be expected.

The 70-m antenna S-band (2.295-GHz) beam position offset calculated above is considerably smaller than that observed in or calculated for the 64-m antenna (0.0086° or 6.1% beamwidth observed, 0.0095° or 6.8% beamwidth calculated). Although this comes somewhat as a surprise, the results appear reasonable from the following consideration. Figure 7 shows an aperture with a linear phase gradient. The aperture phase is $(-)\Phi_{AP}$ at $x = -D/2$, and increases linearly to $+\Phi_{AP}$ at $x = +D/2$. The beam offset angle θ from aperture normal is then

$$\theta = \frac{\lambda \Phi_{AP}}{\pi D} \qquad (1)$$

where λ is the wavelength.

From Eq. (1), one can attribute the smaller beam offset partly to the larger diameter of the 70-m antenna. It is further expected that the aperture edge phase deviation Φ_{AP} would also be smaller in the 70-m case. The aperture edge phase deviation Φ_{AP} is approximately

$$\Phi_{AP} = \tan^{-1} \left| \frac{E_c}{E_p} \right| \qquad (2)$$

where E_c and E_p are fields arising from the $m \neq 1$ modes and $m = 1$ mode near the aperture edge, respectively. From Table 1, we see the higher order mode energy in both cases is mostly in the $m = 2$ component. For a qualitative argument, we assume E_c to be $E_{m=2}$, and also that E_c is approximately the same for both antennas. However, E_p should be larger at the aperture edge for the uniformly illuminated 70-m antenna than for the 64-m antenna with highly tapered illumination. We can therefore expect Φ_{AP} to be smaller in the 70-m case, further reducing the offset angle θ from boresight, Eq. (1).

III. 70-m Antenna S-Band RF Performance

The 70-m antenna S-band (2.295-GHz) theoretical performance was previously estimated based on a two-reflector, tricone feed geometry system.[1] The two-reflector analysis

[1]Cha, A. G., *Physical Optics Analysis of NASA/JPL Deep Space Network 70-m Antennas* (JPL Internal Document D-1853), Jet Propulsion Laboratory, Pasadena, CA, Nov. 1984.

results were modified based on the best information then available as to what the effects of the two additional subreflectors (the ellipsoid and the dichroic plate) would be in the overall four-reflector system. The baseline 64-m antenna S-band performance used in deriving the performance improvement of the 70-m over the 64-m antenna was based on Ref. 2. The estimated S-band performance improvement was 1.6 dB, from area enlargement and RF optics improvement. It was recognized that there were two significant approximations involved in the above derivation of the 1.6-dB performance improvement. First, the 70-m analysis was not a real four-reflector analysis. Second, the 64-m and 70-m performance values were not derived on the same computational bases and could therefore involve bias errors when the two sets of performance data are used to predict the performance improvement. The present four-reflector S-band analysis of the 64-m and 70-m antennas eliminates both of the above approximations and their associated errors. Based on the four-reflector analysis, S-band gain values for 70-m and 64-m antennas are 63.97 dB and 62.45 dB, respectively. These values include losses due to spillovers, non-uniform amplitude and phase in aperture illumination, $m \neq 1$ modes, crosspolarization-equivalent pointing loss from S-band beam position shift, and central blockage. The revised 70-m over 64-m performance values are shown in Fig. 8. The theoretical S-band performance improvement due to area and RF optics is seen to be 1.5 dB, which is 0.1 dB below the earlier estimate but still 0.1 dB over project requirements. There is no revision on X-band (8.45-GHz) performance improvement, which is shown for reference. In addition, there should be some gain improvement due to reduced quadripod blockage. (This is shown as a 0.0 to +0.3 dB uncertainty in Fig. 8.)

References

1. Cha, A. G., "Physical Optics Analysis of a Four-Reflector Antenna: Part 1," *TDA Progress Report 42-84*, pp. 94–100, Jet Propulsion Laboratory, Pasadena, CA, Feb. 15, 1986.

2. Bathker, D. A., "Dual-Frequency Dichroic Feed Performance," presented at 26th Meeting of Avionics Panel, AGARD, Munich, Germany, November 26–30, 1973, *AGARD Conference Proceedings #139*, June 1974, NATO.

Table 1. Azimuthal Fourier mode power content of subreflector diffraction patterns

	Mode Power as Percent of Total Power			
m	22-dB Horn	Ellipsoid or Dichroic	64-m Hyperboloid	70-m Shaped Subreflector
0	~0.0	0.9	0.6	1.1
1	~100.0	98.1	97.1	97.0
2	~0.0	0.9	2.3	1.8
3	~0.0	<0.1	<0.1	<0.1

Table 2. Beam offset caused by 2.54-cm (1-in.) shim of dichroic position[a]

Theta	E Theta		E Phi		Axial Ratio	Ellipse Tilt Angle	RCP Gain, dB
	Volts	Phase	Volts	Phase			
0.00000	176.583868	87.147	176.457674	2.850	0.006	1.99780	63.953
0.00100	176.871906	87.088	176.736952	2.912	0.007	0.22150	63.967
0.00200	177.124393	87.029	176.981068	2.973	0.007	178.63421	63.979
0.00300	177.341164	86.970	177.189932	3.034	0.007	177.19939	63.989
0.00400	177.522308	86.912	177.363615	3.095	0.008	175.89006	63.998
0.00500	177.667601	86.854	177.501917	3.156	0.008	174.68122	64.005
0.00600	177.777067	86,796	177.604851	3.217	0.009	173.55564	64.010
0.00700	177.850657	86,738	177.672361	3.278	0.009	172.49714	64.014
0.00800[b]	177.888355	86.680	177.704437	3.338	0.009	171.49430	64.015
0.00900[b]	177.890163	86.622	177.701069	3.398	0.010	170.53761	64.015
0.01000	177.856033	86.565	177.662327	3.459	0.010	169.61446	64.014
0.01100	177.786058	86.508	177.588160	3.519	0.010	168.72233	64.010
0.01200	177.680252	86.450	177.478649	3.579	0.011	167.85401	64.005
0.01300	177.538679	86.393	177.333836	3.639	0.011	167.00534	63.998
0.01400	177.361317	86.339	177.153761	3.696	0.011	166.16812	63.989
0.01500	177.148457	86.282	176.938639	3.756	0.012	165.34344	63.979
0.01600	176.900034	86.225	176.688459	3.816	0.012	164.52626	63.966
0.01700	176.616140	86.169	176.403273	3.875	0.012	163.71426	63.952
0.01800	176.296972	86.113	176.083372	3.934	0.013	162.90046	63.937
0.01900	175.942684	86.056	175.728769	3.994	0.013	162.09123	63.919
0.02000	175.553347	86.001	175.339638	4.053	0.013	161.27848	63.900

[a]Plane $\phi = 180°$ (direction of 0.0114° beam offset, see Fig. 6).
[b]Beam position for peak gain.

Table 3. Offset of right circularly polarized beam from antenna boresight [a]

Theta	E Theta		E Phi		Axial Ratio	Ellipse Tilt Angle	RCP Gain, dB
	Volts	Phase	Volts	Phase			
0.00000	176.457674	2.850	176.583868	92.853	0.006	1.99475	63.953
0.00100	176.593958	2.909	176.726660	92.914	0.007	3.23790	63.960
0.00200	176.694437	2.968	176.834131	92.975	0.007	4.35443	63.965
0.00300	176.759111	3.026	176.906185	93.035	0.007	5.35418	63.968
0.00400[b]	176.787991	3.084	176.942875	93.095	0.008	6.24177	63.970
0.00500[b]	176.781010	3.141	176.944080	93.154	0.008	7.02990	63.970
0.00600	176.738192	3.198	176.909853	93.214	0.009	7.72686	63.968
0.00700	176.659531	3.255	176.840151	93.273	0.009	8.34092	63.964
0.00800	176.545029	3.311	176.735043	93.331	0.010	8.87782	63.959
0.00900	176.394760	3.368	176.594500	93.390	0.010	9.35170	63.952
0.01000	176.208757	3.423	176.418629	93.448	0.011	9.76480	63.943
0.01100	175.987118	3.479	176.207468	93.505	0.012	10.12679	63.932
0.01200	175.729912	3.534	175.961098	93.563	0.012	10.44147	63.920
0.01300	175.437237	3.589	175.679642	93.620	0.013	10.71383	63.905
0.01400	175.109095	3.641	175.363056	93.675	0.014	10.94936	63.889
0.01500	174.745895	3.695	175.011717	93.731	0.014	11.15447	63.872
0.01600	174.347519	3.749	174.625599	93.787	0.015	11.32791	63.852
0.01700	173.914185	3.802	174.204796	93.843	0.016	11.47777	63.831
0.01800	173.446131	3.856	173.749630	93.899	0.017	11.60288	63.808
0.01900	172.943472	3.909	173.260151	93.954	0.017	11.70823	63.783
0.02000	172.406422	3.961	172.736588	94.009	0.018	11.79514	63.756

[a]Plane ϕ = 270° (direction of 0.0086° beam offset, see Fig. 6).
[b]Beam position for peak gain.
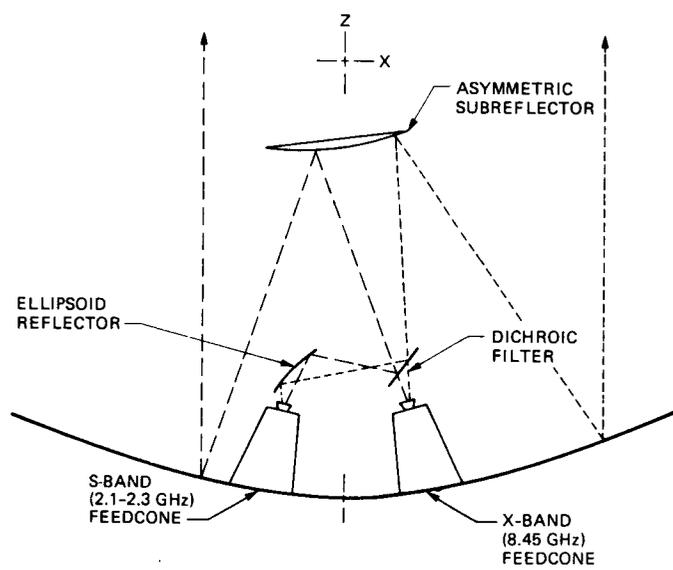
Fig. 1. 70-m antenna four-reflector geometry
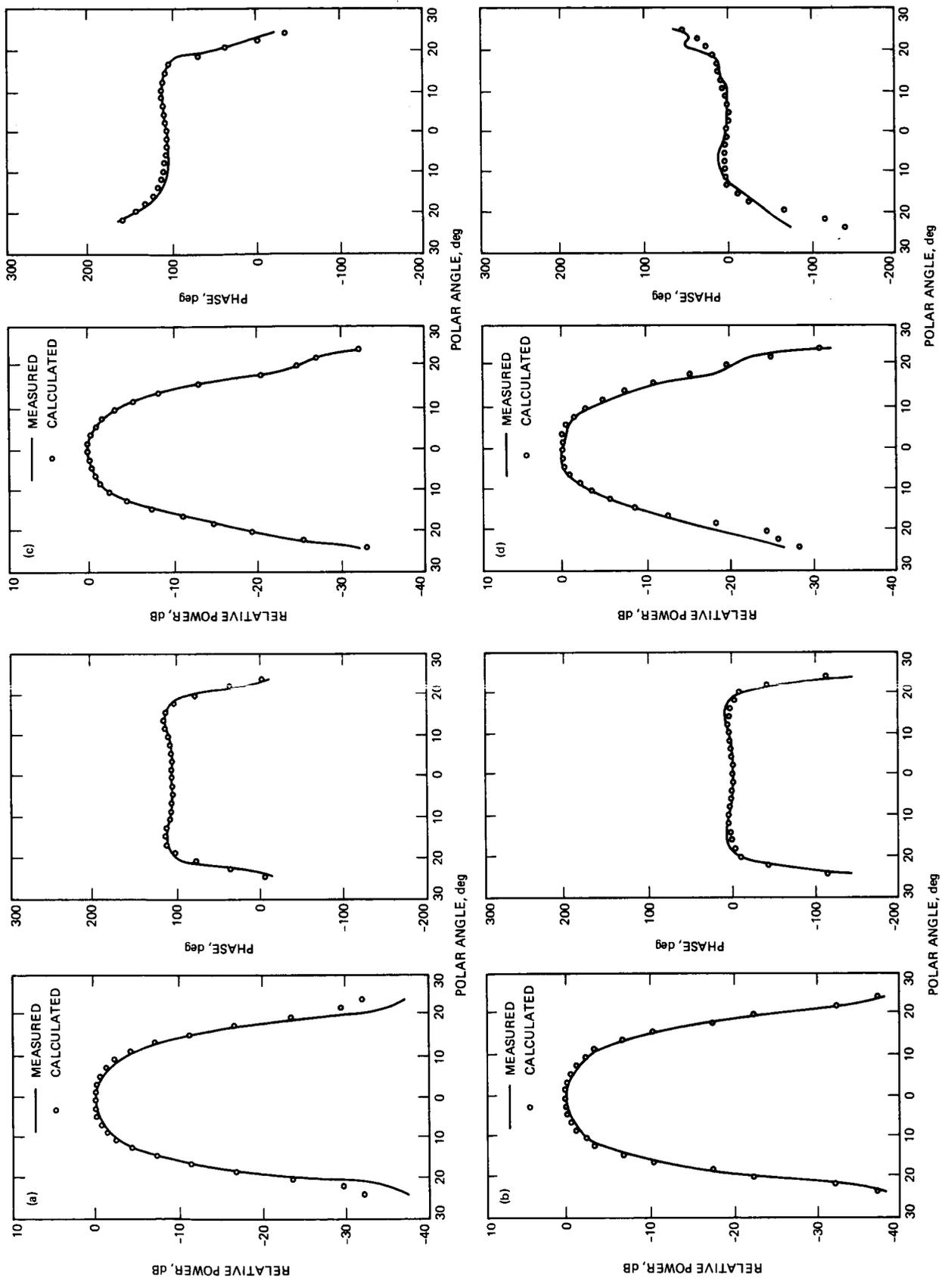
Fig. 2. Ellipsoid and dichroic diffraction patterns: (a) ellipsoid E-plane; (b) dichroic E-plane; (c) ellipsoid H-plane; (d) dichroic H-plane
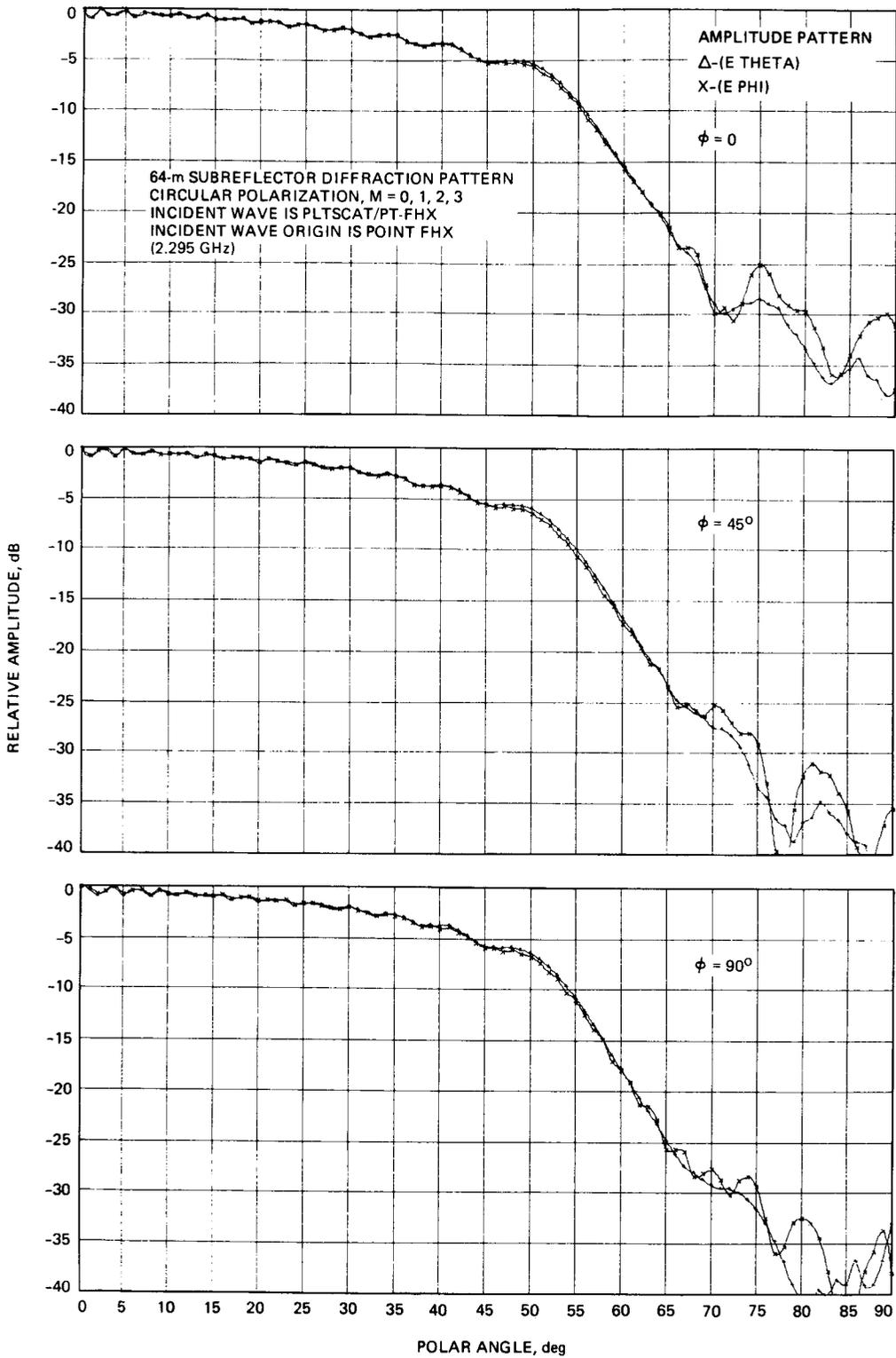
Fig. 3. 64-m hyperboloid diffraction patterns
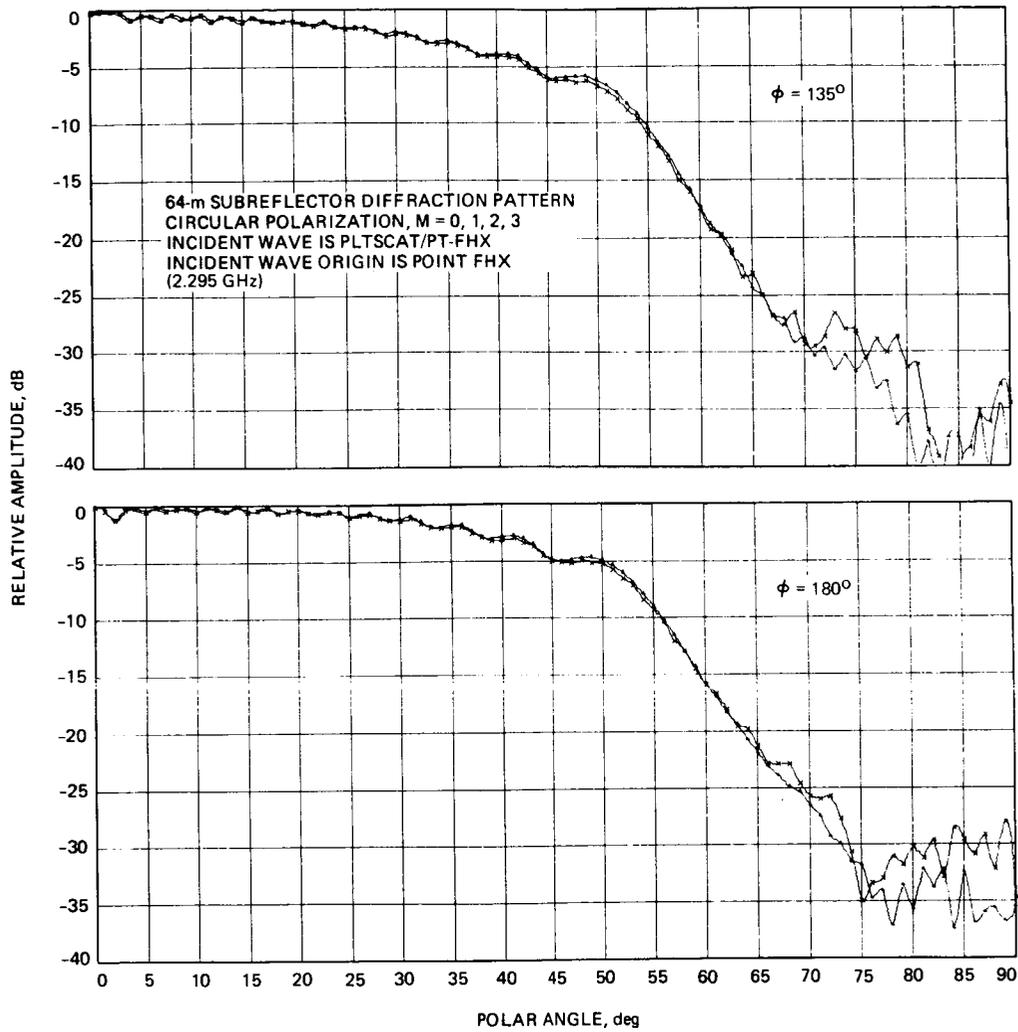
Fig. 3 (contd)

Fig. 4. 70-m shaped subreflector diffraction patterns

70-m SUBREFLECTOR DIFFRACTION PATTERN
CIRCULAR POLARIZATION, M = 0, 1, 2, 3
INCIDENT WAVE IS PLTSCAT/PT-FHX
INCIDENT WAVE ORIGIN IS POINT FHX
(2.295 GHz)

$\phi = 135^o$

$\phi = 180^o$
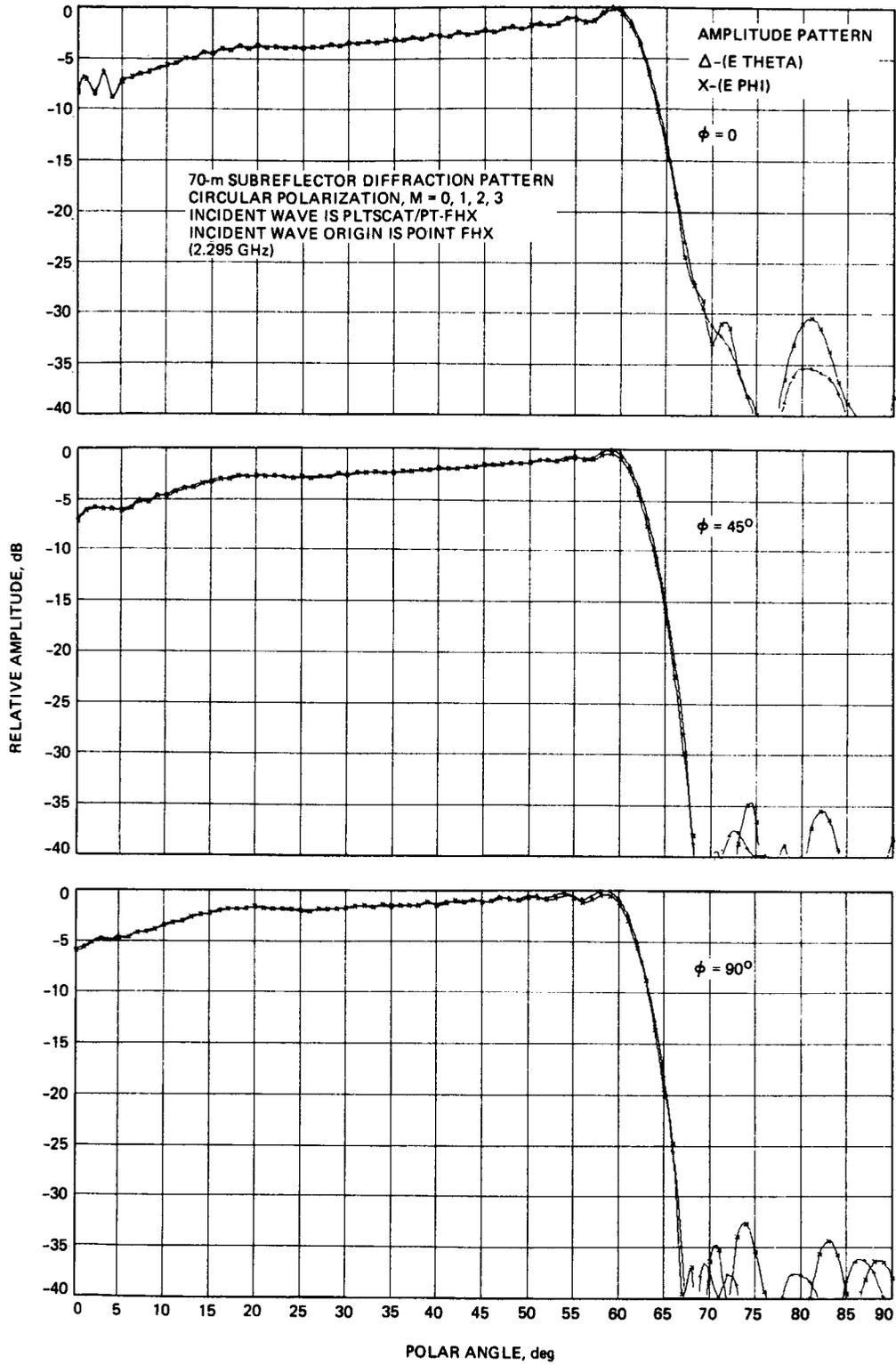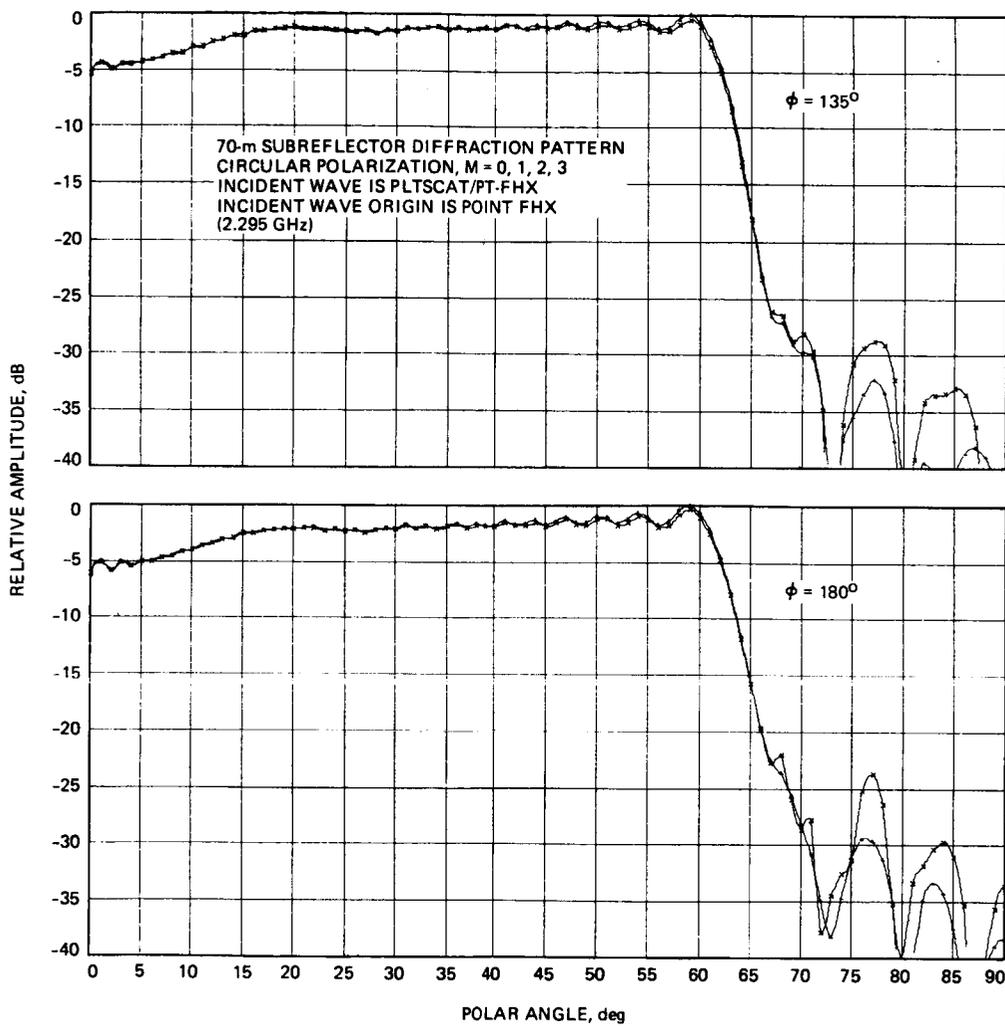
Fig. 4 (contd)
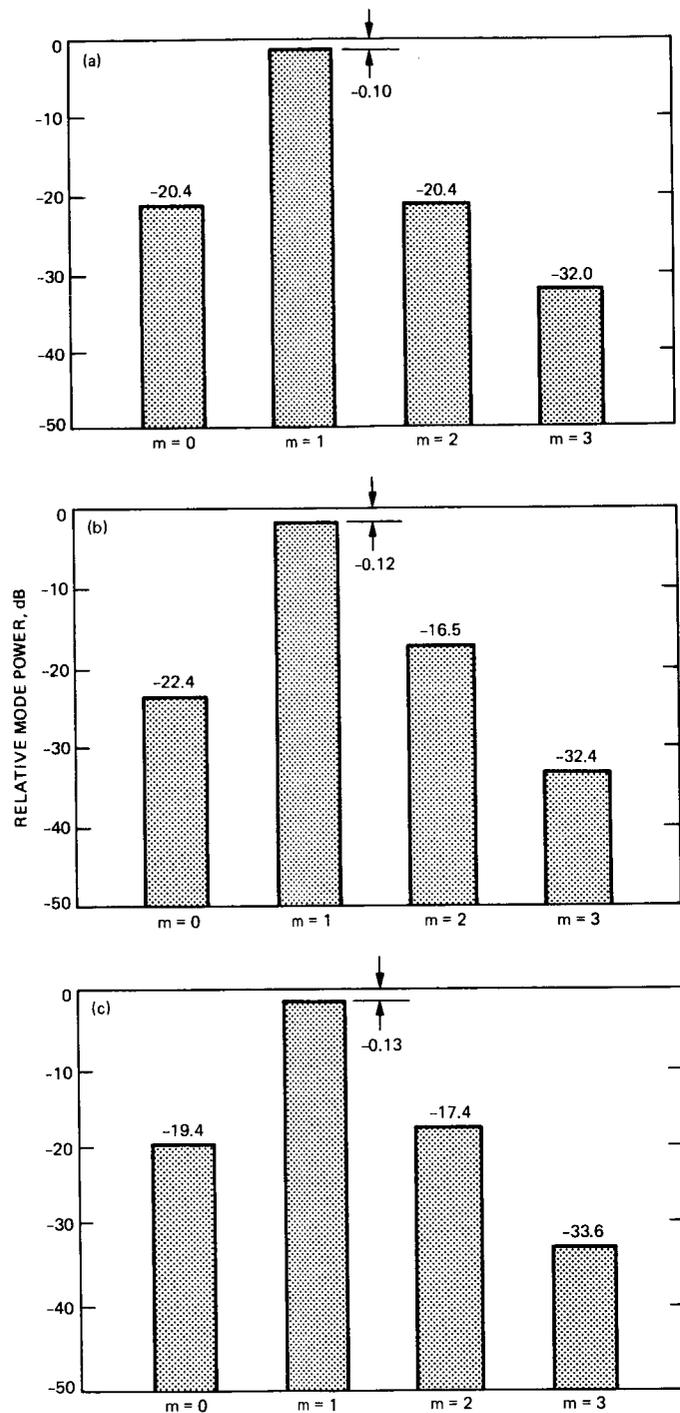
Fig. 5. Azimuthal Fourier mode power content: (a) ellipsoid diffraction pattern; (b) 64-m hyperboloid diffraction pattern; (c) 70-m shaped subreflector diffraction pattern
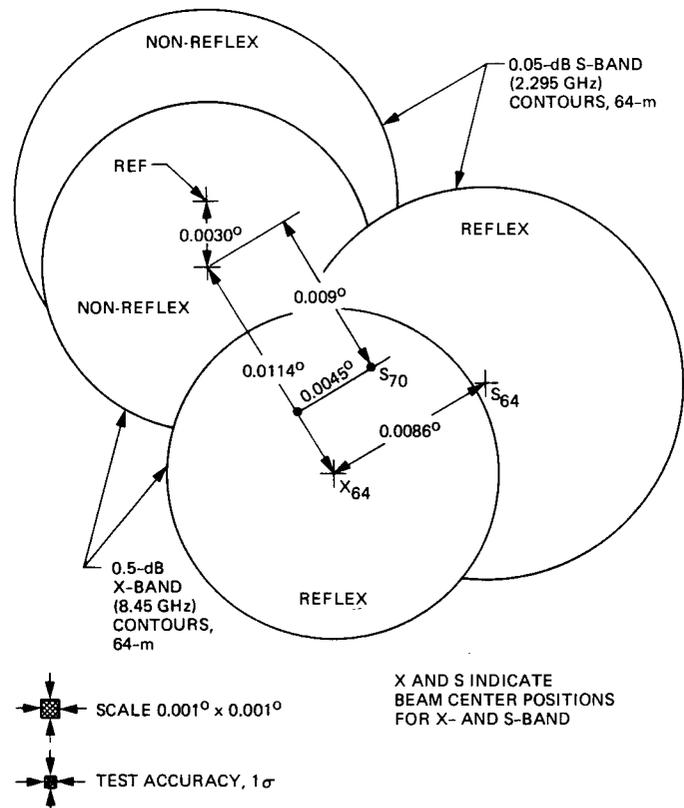
Fig. 6. Predicted S- and X-band beam positions for 70-m antenna
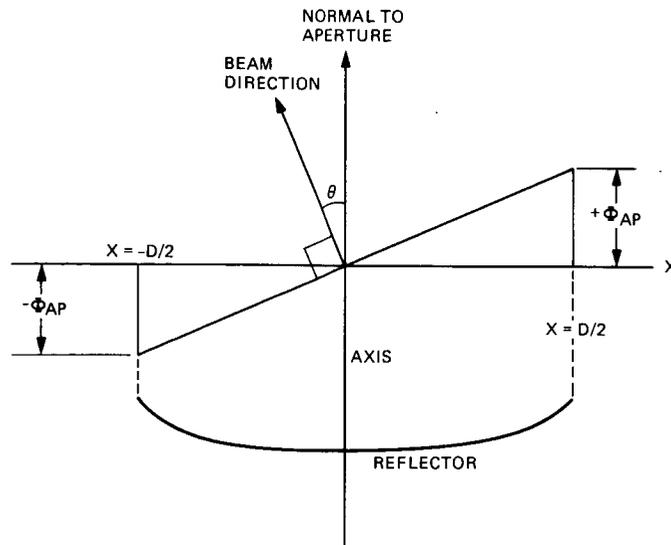
Fig. 7. Beam offset due to linear aperture phase deviation

= PREDICTED VALUE (ACCEPTS 1/2
OF QUADRIPOD POTENTIAL)

= UNCERTAINTY
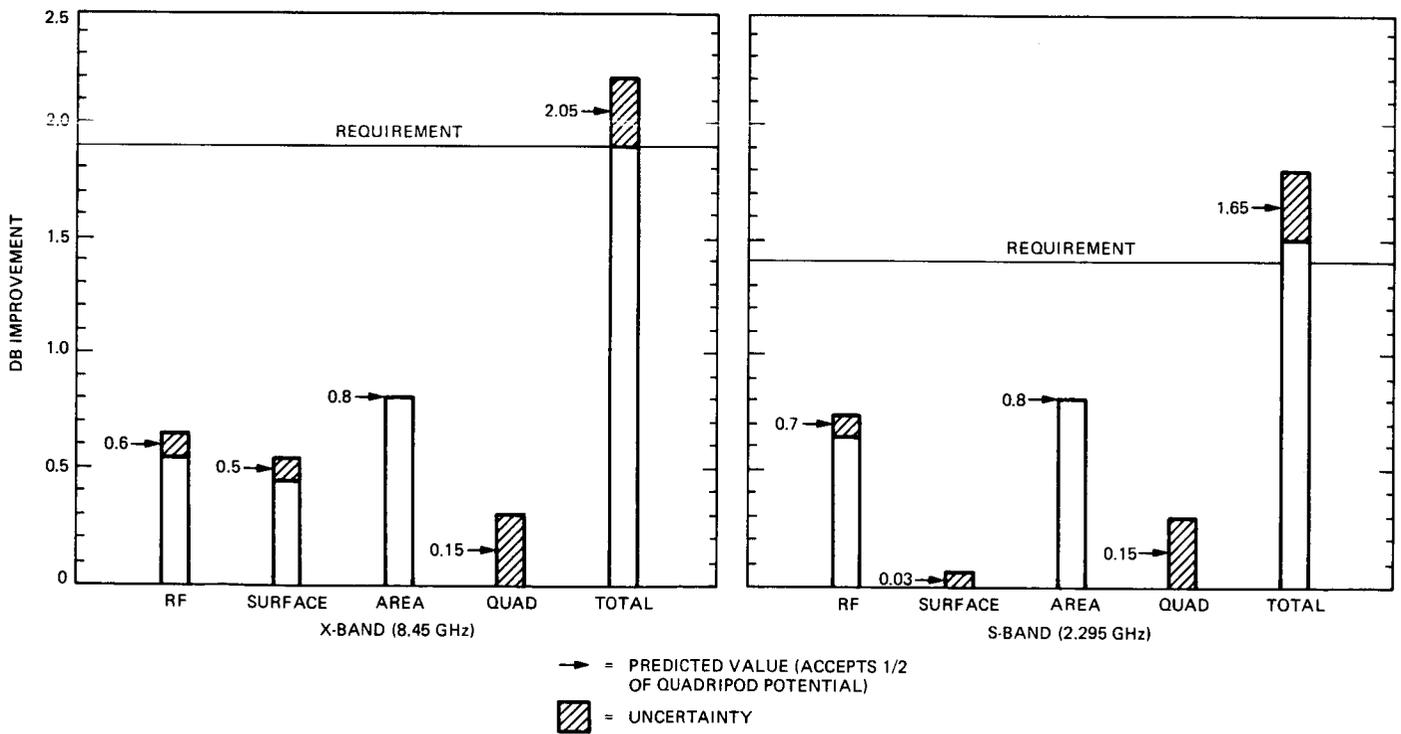
Fig. 8. 70-m antenna reflex feed performance

Finite Wordlength Implementation of a Megachannel Digital Spectrum Analyzer

E. H. Satorius, M. J. Grimm, G. A. Zimmerman, and H. C. Wilck
Communications Systems Research Section

This article presents the results of an extensive system analysis of the megachannel spectrum analyzer currently being developed for use in various applications of the Deep Space Network. The intent of this analysis is to quantify the effects of digital quantization errors on system performance. The results of this analysis provide useful guidelines for choosing various system design parameters to enhance system performance.

I. Introduction

Development of a million channel, Fast Fourier Transform (FFT)-based spectrum analyzer is currently under way at JPL for use in various applications of the Deep Space Network. The system is being designed to provide contiguous output spectra at a real-time throughput rate of 40 MHz. Although the basic system architecture is reasonably well established as described in Ref. 1, i.e., based on a Radix-2 decimation-in-frequency (DIF) FFT algorithm, a number of issues relating to system performance tradeoffs remain to be addressed prior to hardware development.

To this end, we present in this article the results of an extensive computer-aided systems analysis aimed at quantifying the effects of digital quantization noise on system performance. In general, there are three specific sources of quantization error in the implementation and operation of a digital spectrum analyzer: input data quantization; computational errors resulting from finite precision arithmetic operations; and coefficient quantization errors. All of these errors degrade system performance in terms of dynamic range and sensitivity. The results of our analysis not only serve to quan-

tify these degradations, but also provide guidelines for choosing various system design parameters to enhance system performance.

In the remainder of this article, we first briefly describe the general system architecture in Section II. Then, in Section III, we discuss front-end design considerations focusing on image rejection capability. Finally, Section IV contains the results of an extensive quantization error analysis of the basic Radix-2 DIF architecture. The intent of this analysis is a determination of system dynamic range and sensitivity as a function of various input data and wordlength parameters.

II. System Overview

A block diagram of the digital spectrum analyzer system is depicted in Fig. 1 including an IF stage, postprocessing and preprocessing stages, and the FFT processor. As summarized in Ref. 1, the digital spectrum analyzer (FFT processor) is based on the Radix-2 DIF algorithm and will consist of 20 Radix-2 butterfly stages to provide up to 2^{20} output spectral points. Initially (Ref. 1), it was envisioned that 22-bit floating

point arithmetic would be utilized exclusively throughout the FFT processor. However, this is currently being re-evaluated for two reasons. First, due to the excessive memory requirements for the first butterfly stages, it becomes important to minimize wordlengths and simplify the arithmetic in the first stages to meet a 1-rack system packaging goal (this is especially important for a 40-MHz system). Based on the results of the analysis presented in Section IV, it is seen that 16-bit fixed point arithmetic can be used to implement the first butterfly stages (6–8 stages) without sacrificing system performance. Consequently, a hybrid fixed/floating point architecture is being considered as a viable candidate for system implementation. In Section IV, results are presented which quantify hybrid system performance as a function of the number of front-end fixed point stages.

A second reason for re-evaluating a 22-bit floating point architecture also stems from the quantization error analysis as well as the recent advent of fast (10 MHz and faster) 32-bit IEEE floating point chips for performing multiplication and addition operations. In particular, as discussed in Section IV, the dynamic range constraint imposed by a 22-bit floating point architecture is less than that imposed by an 8-bit input quantizer. Thus, dynamic range can not be extended for a 22-bit floating point-based architecture by simply increasing the input quantizer resolution. Such an extension, however, is possible with a 32-bit floating point system. Thus, with the advent of the new chips, 32-bit floating point arithmetic is being considered for hardware implementation.

In addition to re-evaluating the FFT arithmetic, two other critical system design areas are being re-examined prior to system development: (1) system bandwidth, and (2) input IF signal conditioning. A 20-MHz system bandwidth was the original design goal. However, by reducing the FFT front-end size requirements as discussed above, it is possible to package a 40-MHz spectrum analyzer system within one rack. This is highly desirable from the viewpoint of ultimately developing a super wideband (approximately 300 MHz wide), high-resolution system. This can be done with little risk by simply replicating the prototype narrowband system. However, it is certainly more feasible to replicate a 40-MHz system 8 times, for example, than it is to replicate a 20-MHz system 16 times to achieve the same final bandwidth.

Finally, reconsideration of the IF input signal conditioner is also being carried out. In particular, as discussed in Ref. 1, complex mixing was initially considered for translating the input IF signal to baseband. However, only 30-dB image rejection can be realistically provided by complex base-banding due to inherent phase and amplitude mismatching between the in-phase and quadrature channels. Alternatively, a single real channel, sampled at twice the signal bandwidth, can be used to

provide the same spectral information. Furthermore, as discussed in Section III, the real baseband approach is not limited to 30-dB image rejection. The only requirement for this approach is an AD converter which operates at twice the input signal bandwidth. Such a requirement can be met, for example, with existing 8-bit AD converters that operate in excess of the 80-MHz sample rate required for a 40-MHz wide system. Thus, a real baseband, 40-MHz wide system is feasible and is currently being evaluated for system implementation.

III. Front End Design Considerations

With reference to Fig. 1, the system front end prior to digital spectrum analysis consists of the IF conditioning stage as well as input buffer memory and preprocessing stages. Signal basebanding and AD conversion are performed at the IF stage, whereas data reformatting, buffering, and windowing are typically performed in the remaining front-end stages. Critical areas in the design of the system front end include dynamic range constraints imposed by the AD converter as well as the generation of spurious noise components (images) by the basebanding processor. The former design area is discussed below in Section IV, whereas the latter is discussed in this section.

In particular, consider a complex baseband front end such as is used in existing digital spectrum analyzer systems at JPL (Refs. 1 and 2). In this case, a single-channel analog input with signal bandwidth B is converted into two in-phase and quadrature channels each with bandwidth $B/2$. Each channel is sampled at B samples/second and is treated as a complex signal throughout the rest of the system. This approach requires two AD converters each operating at the signal IF bandwidth as opposed to twice the IF bandwidth as would be required for real basebanding.

One limitation with this approach, however, is the generation of images as depicted in Fig. 2. These images are a consequence of phase and amplitude mismatching between the in-phase and quadrature local oscillators. In practice, it is difficult to exceed a 30-dB image rejection ratio (IRR). Consequently, strong RFI components which exceed the desired signal by 25–30 dB contaminate twice the RFI band.

An alternative approach is to convert the analog input into a single baseband channel with bandwidth B. This channel is sampled at $2B$ samples/second and the resulting samples are then split into even and odd pairs in the preprocessor stage. The even–odd sample pairs are treated as complex data by the FFT processor and the resulting output is then recombined in a special, "real adjust" FFT stage to produce the spectrum (Ref. 3, pp. 167–169). Since only one baseband channel is

involved, images created by amplitude and phase mismatch are eliminated.

The only limit to image rejection with this approach is numerical precision, which is most critical in the real adjust FFT stage. This stage effectively uses coherent digital subtraction to cancel images as depicted in Fig. 3. The complex signal, X_{RFI}, denotes an RFI spectral component and the two noise components, n_1 and n_2, arise from digital multiplication and addition operations in previous FFT stages. These components represent the limiting factors on image rejection.

Assuming that only the last FFT stage contributes significantly to these noise components, then the variance of n_1 and n_2 can be approximated by (Ref. 8):

$$\sigma_{n1}^2 = \sigma_{n2}^2 \cong \frac{2^{-2(b-1)}}{3} A_{RFI}^2$$

where A_{RFI} denotes the magnitude of X_{RFI} and b is the number of bits (including sign) used to represent the mantissa of the digital words in the FFT. Thus, the digital IRR is approximately:

$$2 \frac{\sigma_{n1}^2}{A_{RFI}^2} \cong \frac{2}{3} 2^{-2(b-1)}$$

For a 16-bit mantissa, image rejection is approximately 90 dB as opposed to 30 dB provided by complex basebanding. This is also borne out by computer simulation experiments. Based on these results, a real baseband IF stage is being seriously considered for system implementation.

IV. Digital Quantization Error Analysis

The intent of this analysis is to quantify the effects of quantization errors on system sensitivity and dynamic range. In this section, we describe the basic system model used in our analysis (Section IV-A). We then present some approximate analytical results (Section IV-B), which provide a basis for at least understanding trends in system performance over a wide class of input signal conditions. Finally, in Section IV-C, we present a summary of results stemming from an extensive simulation study.

A. System Model

In performing this analysis, we have utilized the system model depicted in Fig. 4. The additive noise components, n_Q and n_{FFT}, correspond to input AD quantization noise (n_Q) and roundoff noise (n_{FFT}) which originates from the FFT butterfly computations. The factor, 1/2, arises from

scaling the input data to the FFT down one bit from the MSB to prevent overflows which might arise within the first butterfly stage (subsequent stages also perform scaling by 1/2).

In quantifying system dynamic range and sensitivity, we have used the following simplified complex input signal model:

$$x(t) = I(t) + s(t) + n(t)$$

where

$$I(t) = A_I e^{2\pi i f_I t} \equiv \text{interference}$$

$$s(t) = A_s e^{2\pi i f_s t} \equiv \text{desired signal}$$

$$n(t) \equiv \text{system noise}$$

The system noise component is assumed to be zero-mean, complex Gaussian noise with total power: $\langle n^2 \rangle = \sigma_0^2 = \alpha k T_s F_s$, where α is a system gain constant (which in practice will vary across the IF bandwidth), k is Boltzman's constant, T_s is the system temperature, and F_s is the sampling rate. It is also assumed that the interference and signal frequencies, f_I and f_s, are at cardinal frequencies of the FFT, and thus the effects of spectral leakage, which is really a separate issue, are not addressed here. (Reference 4 contains an extensive summary of window effects on spectral leakage.)

Even though this is a simplified input signal model, it permits us to develop several important parameters characterizing system performance. First, the FFT input signal-to-noise ratio prior to quantization is given by:

$$SNR_{in} = A_s^2 / \sigma_0^2$$

In the absence of either FFT computational noise or input quantization errors, the output SNR from the FFT is given by:

$$SNR_o = N \, SNR_{in}$$

where $N = 2^L$ is the length of the FFT. With quantization errors, the actual output SNR, SNR_o', can be expressed as:

$$SNR_o' = N \frac{\sigma_s^2}{\sigma_0^2 + N(\sigma_{QN}'^2 + 4 \sigma_{FFT}^2)}$$

where $\sigma_{QN}'^2$ denotes the variance of the AD quantization noise at the FFT output, and σ_{FFT}^2 denotes the variance of the FFT computational noise. (Approximate expressions for $\sigma_{QN}'^2$ and σ_{FFT}^2 will be given below in Section IV-B.) The ratio

of SNR_o to SNR_o', which is a good measure of degradation in system sensitivity due to quantization errors, is given by:

$$\rho \equiv \frac{SNR_o}{SNR_o'} = 1 + \frac{N}{\sigma_0^2}\ (\sigma_{QN}'^2 + 4\sigma_{FFT}^2)$$

Another useful parameter characterizing system performance is the interference-to-input system noise ratio (INR) which is defined as:

$$INR \equiv A_I^2/\sigma_0^2$$

A related parameter is the interference-to-desired signal ratio (ISR):

$$ISR \equiv A_I^2/A_s^2$$

Both of these parameters can be used as measures of system dynamic range. For instance, the INR corresponding to a degradation in system sensitivity of 1 dB ($\rho = 1.26$) can be defined as the input system dynamic range. Similarly, the ISR corresponding to a 1-dB degradation can be defined as the system output or "two-tone" dynamic range.

Finally, we utilize the output noise standard deviation-to-mean ratio, σ:

$$\sigma = \frac{\sqrt{VAR\,(X_n^2)}}{\langle X_n^2 \rangle}$$

where X_n^2 denotes the magnitude squared of the complex FFT coefficient in a noise bin and $VAR(\cdot)$ denotes variance. In the absence of quantization errors (assuming no averaging over multiple bins or transforms):

$$\sigma = 1$$

As discussed in Section IV-B, quantization errors not only increase the average noise level in the output spectrum, thereby increasing ρ, but also increase the noise fluctuations due to the generation of narrowband noise "spurs." Furthermore, these spurs can not generally be reduced by incoherently averaging transforms. They can only be reduced by increasing the numerical precision used in the FFT arithmetic. Consequently, depending on the FFT wordlengths, these spurs may represent the limiting factor on system performance.

B. Input Quantization/Computational Noise Approximations

Regardless of the simplicity of our input signal model, an exact quantization error analysis for this class of inputs

would be extremely complex due to the nonlinear interactions between the various noise components and would provide little insight into system performance. The problem is further complicated by the complex nature of the hybrid fixed/floating point FFT processor under consideration. This processor consists of P fixed point stages followed by $L - P$ floating point stages. Further, the floating point stages utilize floating point addition operations but may only use fixed point multiplication operations due to the fixed point representation of the twiddle factors (such a "quasi" floating point stage model has been used to evaluate existing candidate FFT hardware architectures).

Clearly, an exact analysis of such a complicated system would be very difficult. Consequently, we consider here a simplified analytical statistical model for the quantization errors which arises from a large body of existing literature (Refs. 5-10).[1] This model provides simple, approximate analytical expressions for the various system performance parameters described above in Section IV-A. These expressions are useful for at least predicting trends in system performance as evidenced by the simulation results presented in Section IV-C.

We first discuss the input quantization noise. In particular, it has been shown (Refs. 5-8 and Footnote 1) that for an input signal class consisting of sinewaves plus Gaussian noise, the spectral density of the quantization noise consists of various harmonics and intermodulation products related to the input sinewaves as well as a white (flat) spectral noise component. The amplitudes of the sinewave-related components are a function of both the number of quantizer bits and the levels of the input sinewaves. Typically these components are diminished by the presence of the additive system noise at the input to the quantizer (Footnote 1). Consequently, we assume that the dominant quantization noise component is the uncorrelated, spectrally flat noise component with variance given by (Ref. 8):

$$\sigma_{QN}^2 = q_I^2/12$$

where

$$q_I = 2^{-(BI-1)}$$

and BI is the number of bits (including sign) used in the AD converter.

[1] Also see: Martin, D. R., and Secor, D. J., *High Speed Analog-to-Digital Converters in Communications Systems*, TRW internal report, Nov. 1981.

By virtue of this uncorrelated assumption, the variance of the quantization noise at the output of the FFT, σ'^2_{QN}, is reduced by a factor of $1/N$ from the input:

$$\sigma'^2_{QN} = \sigma^2_{QN}/N$$

Thus, the signal-to-noise ratio, ρ, defined in Section IV-A can be expressed in terms of σ^2_{QN} as follows:

$$\rho = 1 + \frac{1}{\sigma^2_0} (\sigma^2_{QN} + 4N \sigma^2_{FFT})$$

Alternatively, as a function of the interference-to-noise ratio, we have:

$$\rho = 1 + \frac{INR}{A^2_I} (\sigma^2_{QN} + 4N \sigma^2_{FFT})$$

As is seen, the levels of both quantization and FFT computational noise increase as INR increases for fixed A_I. Furthermore, for a given value of σ^2_{FFT}, the effects of computational noise become worse as the FFT length, N, increases whereas input quantization effects are independent of N. Consequently, for shorter transforms the limiting factor on system performance tends to be input quantization noise (regardless of the implementation) whereas roundoff noise starts to dominate as the transform size is increased depending on the implementation. These trends are demonstrated in the simulation results presented in Section IV-C.

The variance of the FFT computational noise, σ^2_{FFT}, has been derived for both fixed and floating point arithmetic (Refs. 8–10) although none of these derivations exactly model the combined effects of quantization/computation/coefficient noise. Nevertheless, we have used these results to derive approximate expressions which characterize system performance trends. Specifically, for a hybrid fixed/floating point architecture, we split the computational noise into two components: one arising from the first P fixed point stages and the other arising from the last L-P floating point stages, i.e.,

$$\sigma^2_{FFT} = \sigma^2_{FXP} + \sigma^2_{FLP}$$

where σ^2_{FXP} and σ^2_{FLP} denote the fixed and floating point noise variances, respectively. Utilizing methods summarized in Ref. 8, we have:

$$\sigma^2_{FXP} \cong \frac{4}{3} 2^{-2(BF-1)} \left(\frac{1}{2}\right)^{L-P}$$

where BF denotes the total number of bits (including sign) used to represent the fixed point FFT words.

Expressions for the floating point noise variance for a wide class of inputs (but assuming floating point multiplication operations) have been derived in Ref. 9. There it is shown that the floating point noise variance can be approximately split into two components: one related to the input system noise and the other related to the large level interference, i.e.,

$$\sigma^2_{FLP}(k) \cong \frac{2}{3} 2^{-2(BM-1)}(L - P)\sigma^2_0 /4N + \sigma^2_{sp}(k)$$

where k denotes the FFT spectral bin number, BM is the number of bits (including sign) in the mantissa of the FFT floating point words, and σ^2_{sp} is the variance of the interference-related noise spur components. The system noise-related component is negligible compared to the other quantization noise components (being scaled by σ^2_0). The spur noise components are narrowband with spectral centers related to the interference frequency.

Based on the analysis presented in Ref. 9, it can be shown that the strongest noise spur occurs at either $k_0 + N/2$ if $k_0 < N/2$ (k_0 = interference spectral bin number) or at $k_0 - N/2$. The two next largest noise spurs (3 dB down) are separated by $N/4$ bins from the maximum spur, etc. The maximum spur level depends on the interference level and numerical precision via:

$$\sigma^2_{sp} \leqslant \frac{1}{3} 2^{-2(BM-1)}(A_I/2)^2$$

Using the above analytical expressions, we can now establish dynamic range constraints imposed by the various quantization noise components. To do this, we first expand ρ in terms of all the quantization noise variances given above, i.e.,

$$\rho = 1 + \frac{INR}{A^2_I} \sigma^2_{QN} + 4N\, INR \frac{\sigma^2_{FXP}}{A^2_I} + 4N\, INR \frac{\sigma^2_{sp}(k)}{A^2_I}$$

$$\equiv 1 + INR\, (E_1 + E_2 + E_3)$$

As discussed in Section IV-A, system input dynamic range can be defined as the INR which corresponds to a 1-dB degradation in system sensitivity, i.e., $\rho=1.26$. Thus the input dynamic range constraint imposed by any quantization noise source is given by:

$$INR_i = 0.26/E_i, \quad i = 1,2,3$$

Additionally, we can compute system output dynamic range constraints from:

$$ISR_i = INR_i \, (N/SNR_o)$$

A summary of dynamic range constraints is presented in Table 1 corresponding to 8-bit input quantization, 16-bit fixed point stages, and both 22-bit (16-bit mantissa, 6-bit exponent) and 32-bit (24-bit mantissa, 8-bit exponent) floating point stages. Also, we have fixed $SNR_o = 10$ (ideal output SNR) and $A_I = 0.75$. As is seen, AD quantization dominates all noise sources (smallest dynamic range constraint) for $N = 2^{14}$. As N increases to $N = 2^{20}$, AD quantization noise continues to dominate fixed point noise from the first butterfly stages. However, for 22-bit floating point stages, spur noise establishes the overall system dynamic range constraint for larger transform sizes ($N = 2^{18}$ or 2^{20}). For 32-bit floating point stages, AD quantization noise remains the dominant noise source for all transform sizes up to 2^{20}. These trends are observed in the simulation results presented in the next section.

C. Simulation Results

We have carried out extensive computer simulations of various hybrid arithmetic FFT systems with transform sizes up to 2^{18}. In addition, we have also obtained some limited hardware simulation results utilizing a 2^{14} point, 22-bit floating point FFT spectrum analyzer. In general, all of these simulation results support the analytical results presented in Section IV-B. As an example, consider the 32-bit floating point simulation results presented in Fig. 5. Here we display a segment (2000 bins) from a 2^{18} point power spectrum (no averaging) containing both interference and desired signal components. Two such segments are presented corresponding to 8-bit input quantization (Fig. 5(b)) and no input quantization, i.e., 32-bit floating point input words (Fig. 5(a)). For both cases, $INR = 60$ dB, $A_I = 0.75$, and $SNR_o = 10$ dB. Referring to Section IV-B (Table 1), we would expect that for this case input quantization will be the limiting factor on system performance inasmuch as FFT computational noise is relatively negligible. This is indeed the case as evidenced by the much higher noise floor (approximately 10 dB higher) in Fig. 5(b). (Note also the addition of noise spurs in Fig. 5(b) evidently due to input quantization effects.)

For hybrid 16-bit fixed point/22-bit floating point systems, the effects of computational noise can be clearly observed in the computer simulated data. For instance, consider the simulation results presented in Fig. 6 corresponding to a hybrid FFT architecture consisting of six 16-bit fixed point stages followed by twelve 22-bit floating point stages. Here we present a 2000-bin spectral segment resulting from inco-

herently averaging five successive transforms (no frequency averaging). This particular segment was chosen to contain the desired signal component as well as the dominant floating point noise spur (based on the analytical model presented in Section IV-B). Two such segments are presented corresponding to $INR = 40$ dB with 8-bit input quantization (Fig. 6(a)) and $INR = 60$ dB with 12-bit input quantization (Fig. 6(b)). For both cases, $A_I = 0.75$ and $SNR_o = 10$ dB. Based on the discussion in Section IV-B (Table 1), we would expect that even at $INR = 40$ dB, floating point noise spurs will represent the limiting factor on system performance. This is indicated in Fig. 6(a), although unambiguous signal detection at least over the 2000 bins displayed is still possible. However, the situation degrades significantly in Fig. 6(b) where noise spurs larger than the signal spectral level would prevent unambiguous signal detection.

A summary of further 2^{18} point FFT computer simulation experiments is presented in Table 2 corresponding to the hybrid 16-bit fixed point/22-bit floating point architecture again with $A_I = 0.75$ and $SNR_o = 10$ dB. The system parameters tabulated in Table 2, SNR'_o and σ, were obtained by frequency averaging over one transform. Specifically, SNR'_o was computed by averaging over 64 redundant signals uniformly spaced throughout the 2^{18} output spectral points and does not correspond to the worse-case output SNR at the noise spurs. Indeed, degradation in system performance is only evidenced by the computed values of σ which were obtained by averaging over all the 2^{18}-64-1 noise bins (the interference bin is excluded from computing σ). For $INR = 60$ dB and 12-bit input quantization, $SNR'_o \cong 9.3$ dB, indicating satisfactory system performance but $\sigma \cong 2.4$–3.0 dB revealing the influence of the noise spurs. Note that, for this case, increasing the number of fixed point butterfly stages from 4 to 6 leads to a 0.6-dB increase in σ, possibly due to interaction between fixed point and floating point noise components. A more ideal behavior is indicated for $INR = 40$ dB and 8-bit quantization; however, even in this case low level noise spurs are present in the output spectrum as noted above (Fig. 6(a)).

Finally, we present in Fig. 7 an output spectrum resulting from a hardware simulation of a 2^{14} point, 22-bit floating point FFT spectrum analyzer currently under development. This spectrum resulted from first generating a simulated complex input data set (on a MASSCOMP computer) consisting of a sinewave (amplitude = 0.375 at spectral bin 250) plus a low level of Gaussian noise with standard noise deviation approximately equal to the quantization level. This data was quantized to 15 bits and memory-mapped to the FFT hardware which consists of 14 butterfly stages and a data unscrambler stage to perform the bit reversal operation. The resulting spectral output (16,384 bins) was then decimated

by only keeping the largest spectral level in every 4 bins and then plotted.

As is seen in Fig. 7, numerous spectral noise peaks are present at or above the maximum levels predicted by either the general computer FFT simulator program or analysis. It should be noted, however, that a separate, special-purpose computer program written to exactly model this FFT hardware system did agree precisely with the hardware results. Clearly, these results indicate the presence of computational noise spurs resulting from the 22-bit floating point arithmetic although more detailed experiments utilizing multiple transform averages will be performed to further evaluate this system.

V. Conclusions

As a result of this systems analysis, several important conclusions emerge. First, spurious images due to the front-end baseband processor can be significantly reduced by utilizing real basebanding in combination with an additional real-adjust stage at the end of the FFT processor. Second, for 8-bit input quantization, which will most likely be used in implementing the prototype megachannel spectrum analyzer, system input dynamic range will be at most 45 dB and possibly less depending on whether 22-bit or 32-bit floating point arithmetic is used. Certainly the results of this analysis would support a 32-bit implementation. Third, 16-bit fixed point arithmetic can be used to implement the first butterfly stages without sacrificing system performance, provided the number of such stages is 8 or less. If the number of fixed point stages exceeds 8, then the resulting roundoff noise from these stages will begin to dominate the 8-bit input quantization noise. Finally, although the effects of window leakage on system performance were not addressed in this study, it is an important issue in designing a very large FFT. For instance, an input $INR = 40$ dB corresponds to a spectral dynamic range of 100 dB when $N = 2^{20}$. Thus, window sidelobes should be at least 100–110 dB down near the desired signals or else spectral leakage from the large level interference will prevent signal detection. Clearly, the particular window used for the megachannel system must be carefully chosen, i.e., using the results of previous detailed analysis on FFT window design (e.g., as summarized in Ref. 4).

References

1. Quirk, M. P., Wilck, H. C., and Grimm, M. J., "A Wide-Band, High-Resolution Spectrum Analyzer," *TDA Progress Report 42-83*, pp. 180-190, Jet Propulsion Laboratory, Pasadena, CA, 1985.

2. Morris, G. A., and Wilck, H. C., "JPL 2^{20} Channel 300 MHz Bandwidth Digital Spectrum Analyzer," *1978 IEEE Int. Conf. Acoust., Speech, and Signal Processing*, April 10-12, 1978, Tulsa, OK, pp. 808-811.

3. Brigham, E. O., *The Fast Fourier Transform*, Prentice-Hall, 1974.

4. Harris, F. J., "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, Vol. 66, pp. 51-83, Jan. 1978.

5. Bennett, W. R., "Spectra of Quantized Signals," *Bell Sys. Tech. J.*, Vol. 27, pp 446-472, July 1948.

6. Widrow, B., "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems," *Trans. AIEE*, Vol. 79, pp. 555-568, Jan. 1961.

7. Hurd, W. J., "Correlation Function of Quantized Sine Wave Plus Gaussian Noise," *IEEE Trans. Infor. Theory*, Vol. 13, pp. 65-68, Jan. 1967.

8. Oppenheim, A. V., and Schafer, R. W., *Digital Signal Processing*, Chapter 9, Prentice-Hall, 1975.

9. Kaneko, T., and Liu, B., "Accumulation of Round-Off Error in Fast Fourier Transforms," *J. Assoc. Comput. Machin.*, Vol. 17, pp. 637-654, Oct. 1970.

10. Weinstein, C. J., "Roundoff Noise in Floating Point Fast Fourier Transform Computation," *IEEE Trans. Audio, Electroacoust.*, Vol. 17, pp. 209-215, Sept. 1969.

Table 1. Quantization noise contributions to system dynamic range

Noise Source	N	INR, dB	ISR, dB
AD Quantization,	2^{14}	45	77
σ^2_{QN} (8-bit ADC)	2^{18}	45	89
	2^{20}	45	95
Fixed Point	2^{14}	62 (P=4)	94 (P= 4)
Roundoff, σ^2_{FXP}	2^{18}	62 (P=4)	106 (P=4)
(16 bit)	2^{20}	56 (P=6)	106 (P=6)
	2^{20}	50 (P=8)	100 (P=8)
Floating Point	2^{14}	47	79
Roundoff, $(\sigma^2_{sp})_{max}$	2^{18}	35	79
(22 bit)	2^{20}	29	79
Floating Point	2^{14}	95	127
Roundoff, $(\sigma^2_{sp})_{max}$	2^{18}	83	127
(32 bit)	2^{20}	77	127

Table 2. Summary of hybrid 16-bit fixed point/22-bit floating point computer simulation results for $N = 2^{18}$

Number of AD Bits	INR, dB	P	SNR'_o, dB	σ, dB
8	40	6	9.0	0
12	60	4	9.3	2.4
12	60	6	9.4	3.0
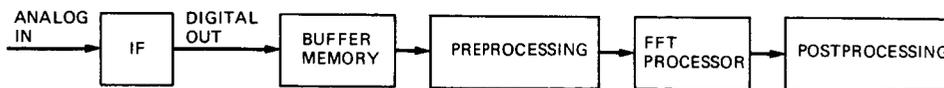
Fig. 1. System block diagram
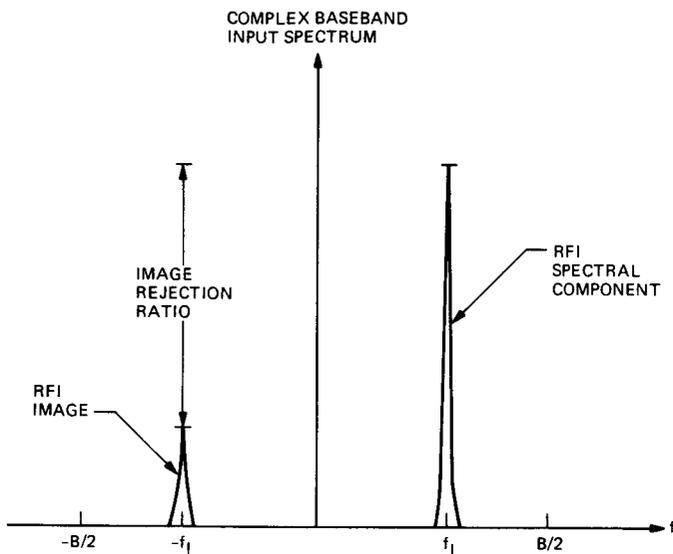
Fig. 2. Image generated by complex basebanding
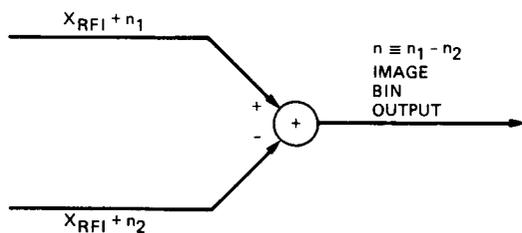
Fig. 3. Model of image cancellation in the real adjust stage
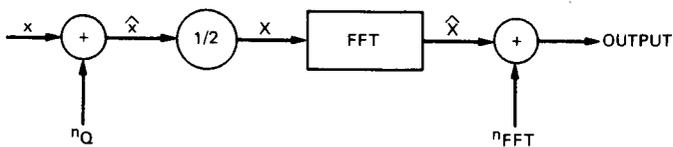
Fig. 4. Basic system model
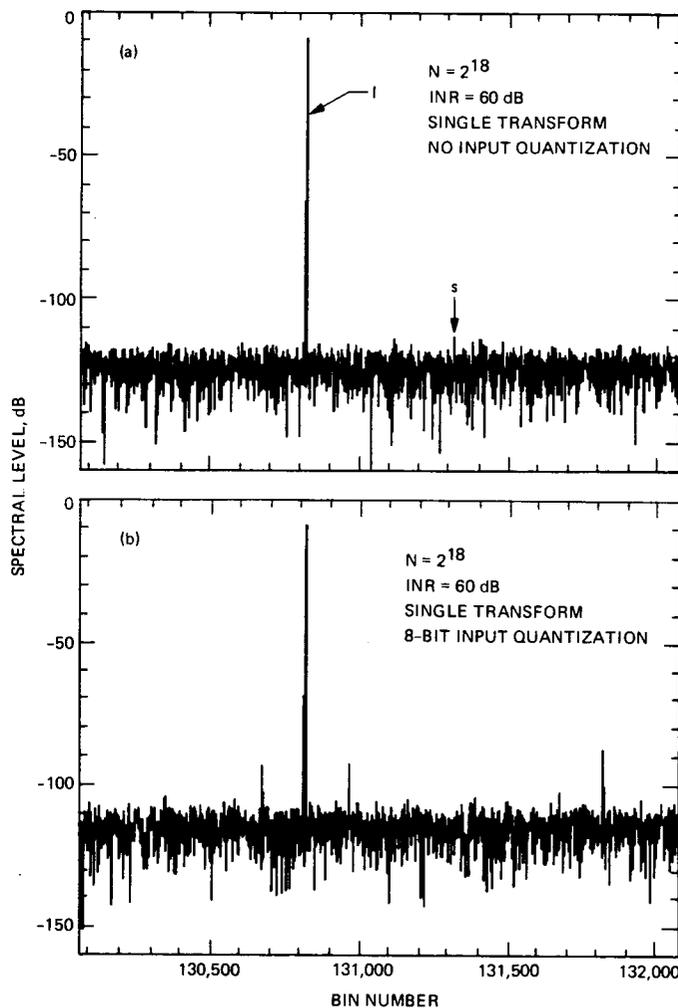
Fig. 5. Computer simulation results for 32-bit floating point FFT arithmetic
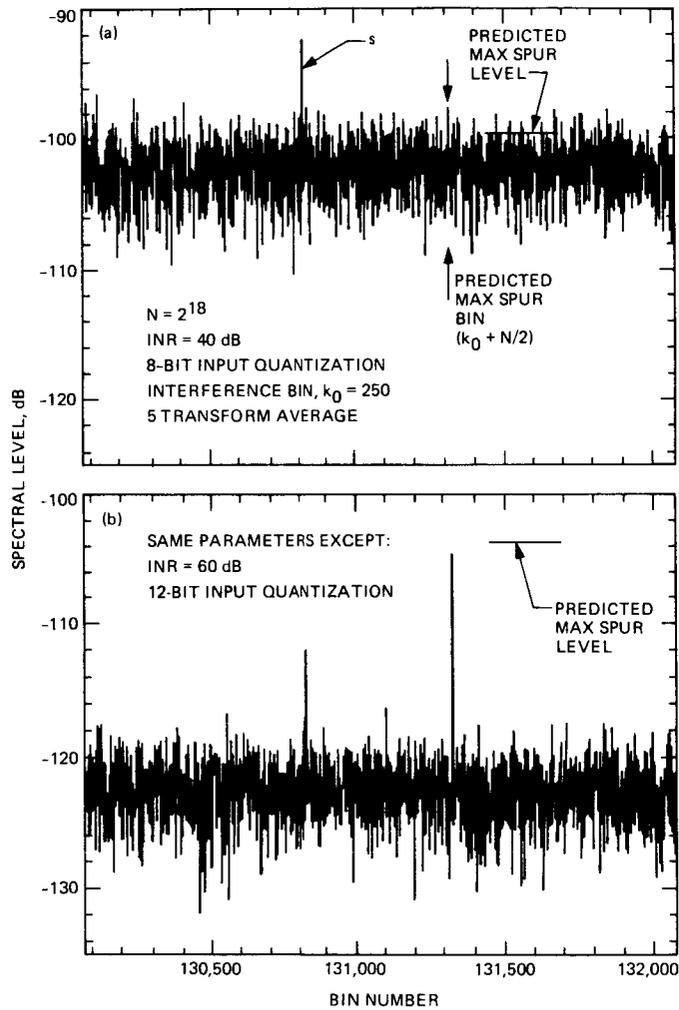
Fig. 6. Computer simulation results for hybrid 16-bit fixed point/22-bit floating point FFT arithmetic with 6 front end fixed point stages
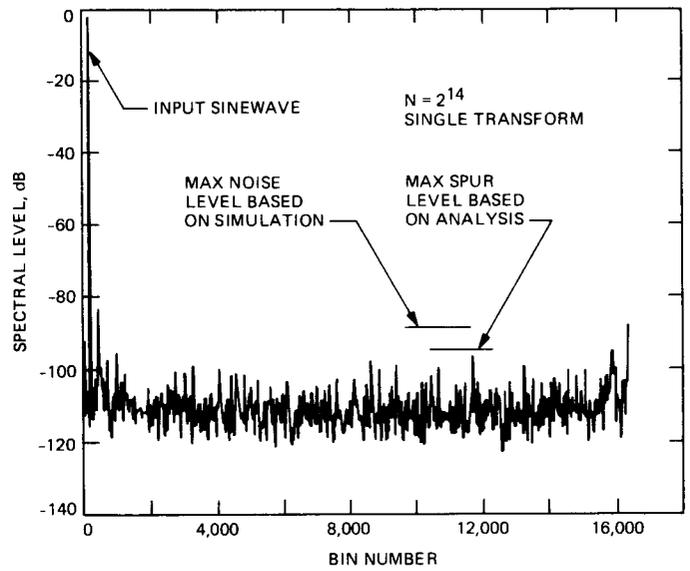
Fig. 7. Hardware simulation for 22-bit floating point system

Deep Space Network Mark IVA Description

R. J. Wallace and R. W. Burt
TDA Engineering Office

This article describes the general system configuration for the Mark IVA Deep Space Network. The arrangement and complement of antennas at the communications complexes and subsystem equipment at the Signal Processing Centers are described. A description of the Network Operations Control Center is also presented.

I. Introduction

The Mark IVA Deep Space Network (DSN) Implementation Project was initiated in 1980 to implement the most complex change to be made to the DSN since its inception. The objectives of the Mark IVA Project were to reduce operation costs, to improve reliability and maintainability, and to increase telemetry reception capability for support of the Voyager Uranus encounter and subsequent planetary missions. These objectives were accomplished by replacing separated, individual antenna control rooms with a centralized signal processing center at each of the three Deep Space Communications Complexes (DSCCs), adding new control and data system computers and local area networks, and building new 34-meter antennas. At the same time, 26-meter antennas from NASA's Ground Spaceflight Tracking and Data Network (GSTDN) were added to the DSN as part of the Network Consolidation Project (Ref. 1).

The Mark IVA DSN has provided outstanding support for all required critical mission activities. These missions include Venus Balloon, International Comet Explorer (ICE), Halley Pathfinder, GIOTTO, and Voyager Uranus Encounter. A significant increase in data return to the Voyager Project was possible because of the techniques developed to array the new and existing antennas for improved telemetry reception.

II. Mark IVA Network Configuration

The new network configuration is shown in Fig. 1. It consists of four antennas, a Signal Processing Center (SPC), and an Earth Orbiter Link at the DSCCs located at Goldstone, California, and Canberra, Australia. There are three antennas at the Madrid, Spain, DSCC; the 34-meter High Efficiency (HEF) antenna is to be added in 1987. Communication facilities to connect these complexes to the Network Operations Control Center are included in the DSN.

At the overseas complexes, the antennas are colocated within 1 or 2 kilometers in order to ease operations and maintenance and to enhance the capability of arraying multiple antennas. At Goldstone, the 34-meter Standard antenna is located at the Echo site which is 12 miles from the SPC at the Mars site. Also, at Goldstone, the 26-meter antenna with the Earth Orbiter Link is located 7 miles from the SPC. Each antenna within the complex has some locally mounted equipment: antenna drive and control equipment, low-noise amplifiers, receiver front ends, and transmitters. The balance of the Deep Space Communications antenna-associated equipment — antenna pointing, microwave instrumentation, transmitter control, receiver, and metric data — are located at the SPC. The SPC also includes the assignable link equipment: telemetry, command, radio metric, radio science, and Very

Long Baseline Interferometry (VLBI) processing equipment. The Complex Global equipment including the Frequency and Timing Subsystem (FTS), Test Support, and Communications are also located in the SPC. The 26-meter link equipment for Earth Orbiter communication is not located in the SPC at any of the DSCCs.

III. Antennas

Each complex is to have four antennas, configured as follows:

(1) *64-Meter*: This pre-Mark IVA antenna is configured for S-band transmission and reception and X-band reception. This is the prime antenna for deep space communications and for radio metric data, radio science data, and VLBI applications.

(2) *34-Meter Standard*: This pre-Mark IVA antenna at each complex is configured for transmission of S-band and reception of S-band and X-band. It is used for deep space communications and for radio metric data. It is also used to support high earth orbiters.

(3) *34-Meter High Efficiency*: New 34-meter High Efficiency antennas were implemented at the Australian and Goldstone complexes. (DSS 65 in Madrid will be added in 1987.) They are configured for both X-band and S-band reception and are used primarily for deep space down-link support. An X-band transmission and radio science VLBI capability is planned for addition in 1987.

(4) *26-Meter*: The 26-meter antennas were relocated in Spain and Australia to the proximity of the SPCs. The Goldstone 26-meter was not relocated. These antennas are used for S-band transmission and reception in support of Earth Orbiter and High Earth Orbiter communications and for radio metric data.

IV. Signal Processing Center

A simplified block diagram of a Signal Processing Center is shown in Fig. 2. This shows the general subsystem complement. Detailed configurations and connections are described in other articles. The SPC is configured to support operation of each antenna individually, or to array any combination of the antennas.

As indicated, antenna control, receiver/exciter, and radio metric tracking subsystem equipment is associated with each antenna as appropriate. Only transmitting antennas have exciters and tracking equipment. For the 64- and 34-meter

antennas, the Telemetry and Command Subsystems are each organized into four groups while the Monitor and Control Subsystem is organized into three groups. Each group can be independently assigned; the groups are thus assigned to form up to three "links." Each "link" has the equipment necessary to support one spacecraft mission, with receiver, antenna, tracking, command, and telemetry equipment. The link can handle a single antenna or an array. Each link is controlled by a single operator stationed at the Link Monitor and Control Console. The link assignments, including the antenna and associated equipment, are performed by the Complex Monitor and Control according to an established schedule. Other subsystems provide test support, technical facilities, frequency and timing, maintenance, and radio science support. Details for the SPC System configurations are presented in other articles.

V. Earth Orbiter Link

The control and processing equipment for Earth Orbiter and High Earth Orbiter support is not integrated into the SPC. In general, the equipment is the same as that being used in the GSTDN. There are two notable exceptions: only a single spacecraft link can be supported and the telemetry processing equipment has been updated. Additionally, there is a cross-support connection that enables a spacecraft link through the 26-meter front end to be processed in the SPC. The reverse is also true in that a 34-meter Standard antenna front end can be connected to the Earth Orbiter link processing equipment.

The Earth Orbiter link support includes separate communications circuits to the Goddard Space Flight Center Network Control and Project Operations Control Centers.

VI. DSN Network Operational Control Center

Data received at the control center is forwarded either to local JPL or remote mission operations centers as shown in Fig. 3. In parallel, the data is also routed to the Network Operations Control Center (NOCC) where data monitoring and network control functions are performed. Monitoring functions are performed in the real-time monitor processors, which were retained from Mark III. The software in these computers was upgraded to be compatible with the Mark IVA design. In addition, a processor was added to perform the Network support functions, including providing predicts, sequence of events, and schedule data to the SPCs. The controllers in the NOCC are provided with displays allowing them to monitor and control network activities including Earth Orbiter link activities.

VII. Communications

The communications between the SPCs and the Central Communications Terminal (CCT) at JPL were modified as part of the Mark IVA Project. This modification permits the cost-effective utilization of available circuits while still providing the flexibility to protect critical data. The communications processor at the SPC is capable of multiplexing data from all of the SPC links into a single circuit or of transmitting critical data in a separate complex circuit.

Reference

1. Yeater, M. L., Herman, D. T., and Sanner, G. E., "Networks Consolidation Program," in *The Telecommunications and Data Acquisition Progress Report 42-59*, July and August 1980, p. 107–120, Jet Propulsion Laboratory, Pasadena, Calif., Oct. 15, 1980.
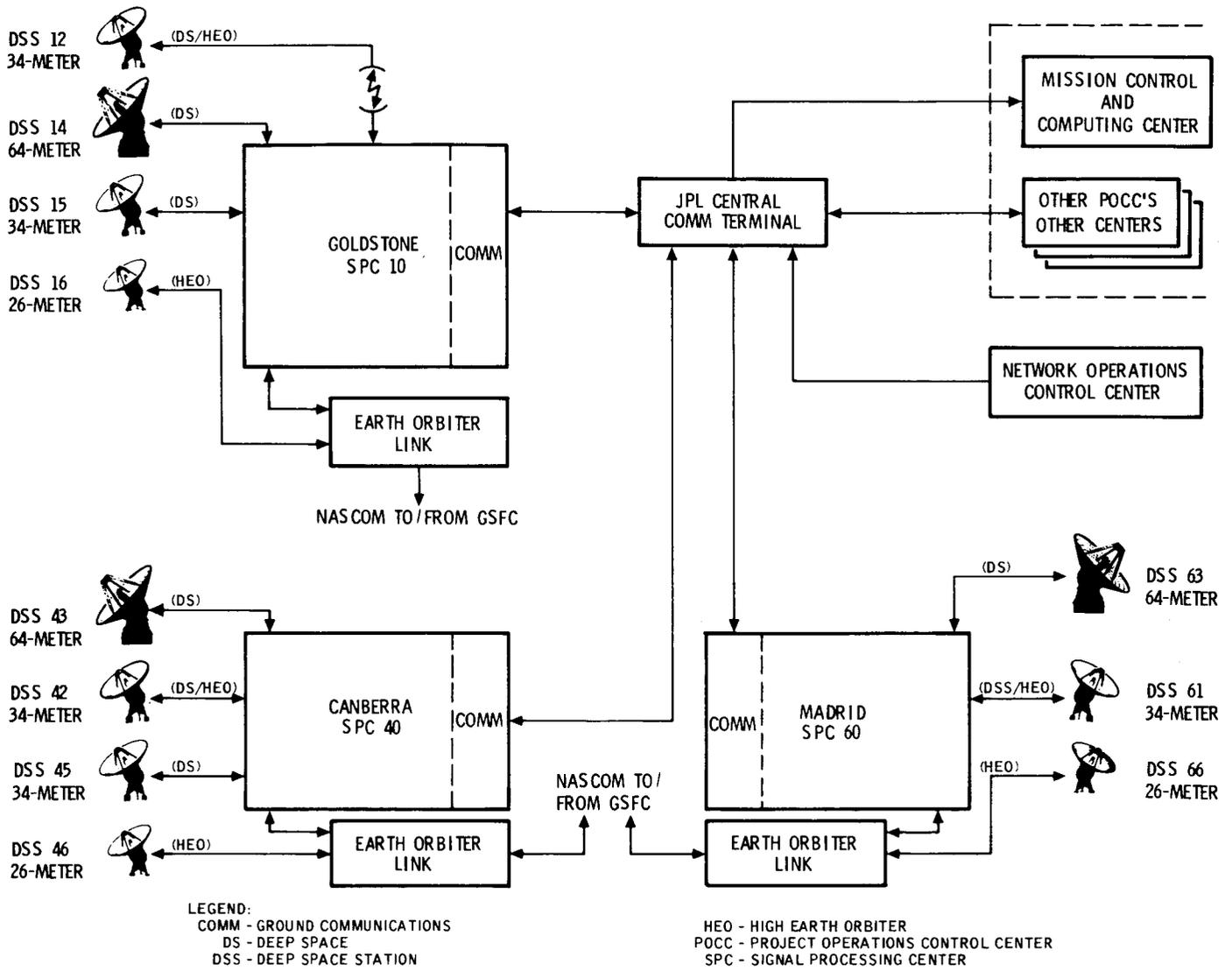
Fig. 1. Deep space network Mark IVA configuration

DSS 12 34-METER

DSS 14 64-METER

DSS 15 34-METER

DSS 16 26-METER

(DS/HEO)

(DS)

(DS)

(HEO)

GOLDSTONE SPC 10 | COMM

EARTH ORBITER LINK

NASCOM TO/FROM GSFC

JPL CENTRAL COMM TERMINAL

MISSION CONTROL AND COMPUTING CENTER

OTHER POCC'S OTHER CENTERS

NETWORK OPERATIONS CONTROL CENTER

DSS 43 64-METER

DSS 42 34-METER

DSS 45 34-METER

DSS 46 26-METER

(DS)

(DS/HEO)

(DS)

(HEO)

CANBERRA SPC 40 | COMM

EARTH ORBITER LINK

NASCOM TO/FROM GSFC

COMM | MADRID SPC 60

EARTH ORBITER LINK

DSS 63 64-METER

DSS 61 34-METER

DSS 66 26-METER

(DS)

(DSS/HEO)

(HEO)

LEGEND:
COMM – GROUND COMMUNICATIONS
DS – DEEP SPACE
DSS – DEEP SPACE STATION

HEO – HIGH EARTH ORBITER
POCC – PROJECT OPERATIONS CONTROL CENTER
SPC – SIGNAL PROCESSING CENTER

258

C-24

Fig. 2. Mark IVA Signal Processing Center, simplified block diagram
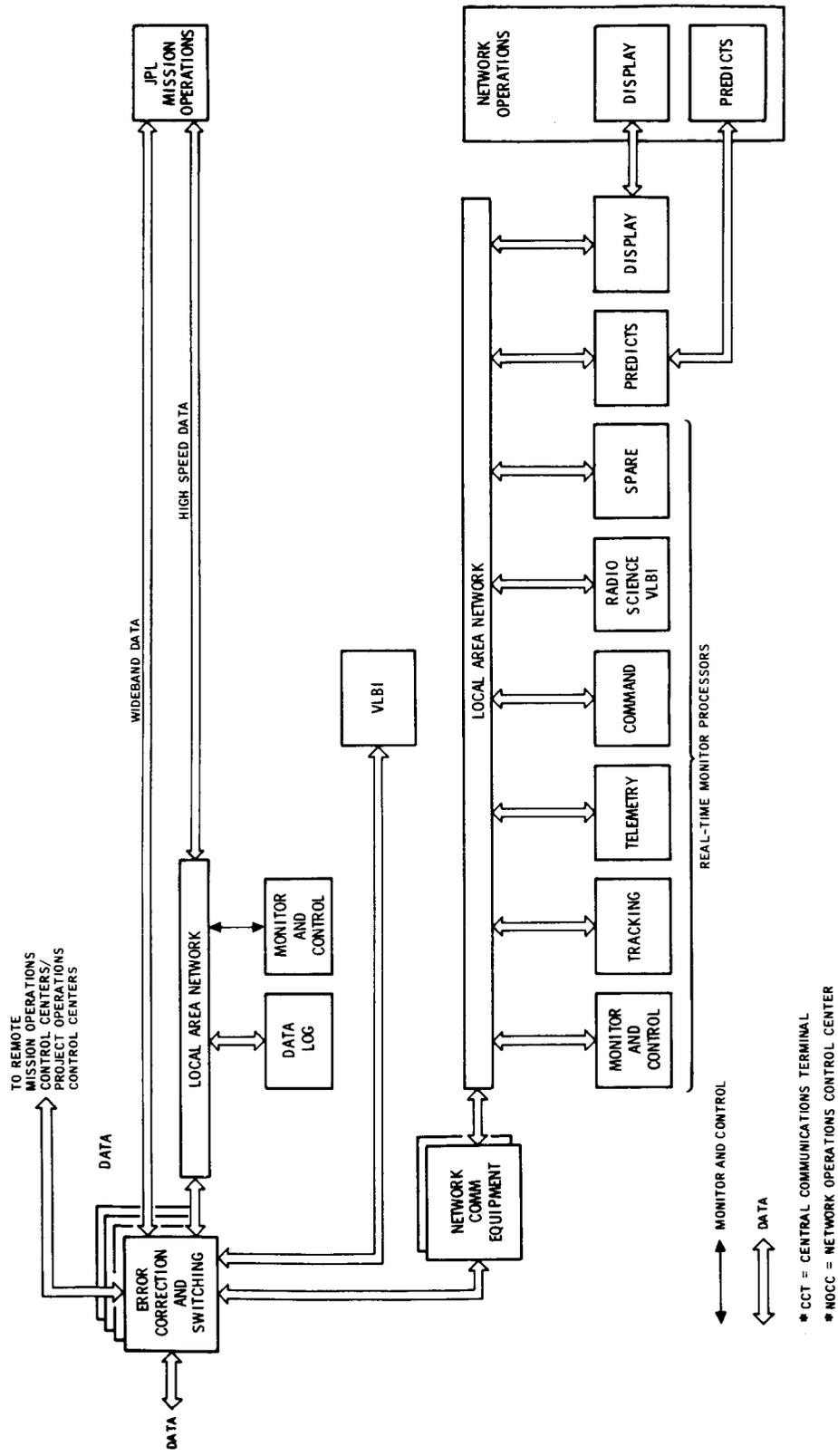
Fig. 3. Mark IVA NOCC, simplified block diagram

HEO Multimission Navigation Concept

J. Ellis

Navigation Systems Section

As part of the Network Consolidation Program, the 26-meter Tracking and Communication Subnet was transferred to JPL. Along with this transfer JPL assumed responsibility for tracking and navigation support for Earth orbiter missions normally tracked by the 26-meter sites. The High Earth Orbiter (HEO) Multimission Navigation Facility was formed as a component of the DSN Tracking System for the purpose of supporting Earth orbiter missions and certain classes of deep space missions. This facility has been implemented on a dedicated VAX 11/780 minicomputer within the Network Operations Control Center (NOCC). The primary function of the system is to process radio metric data and estimate the orbit of a spacecraft in near-Earth or deep space environment. The system is capable of processing radio metric data in near-real time and providing the quick turnaround required for Earth orbiter operations. It is also capable of generating precision spacecraft ephemeris for use by the NOCC Support Subsystem and external agencies. This article discusses the implementation and functional operation of the Multimission Navigation Subsystem and describes the support that has been provided for an array of missions.

I. Introduction

As part of the Network Consolidation Program, the 26-meter Ground Spaceflight Tracking and Data Network (GSTDN) was transferred from Goddard Space Flight Center (GSFC) to JPL as an element of the MARK IVA DSN. Along with this transfer, JPL assumed responsibility for tracking and navigation for Earth orbiter missions normally tracked by the GSTDN sites. This includes all high Earth orbiter missions and Earth orbiters which are not compatible with the Tracking and Data Relay Satellite System (TDRSS). The High Earth Orbiter (HEO) Multimission Navigation Facility was formed as a component of the DSN Tracking System for the explicit purpose of supporting Earth orbiter missions and certain classes of deep space missions. The latter typically includes non-JPL deep space missions that are tracked by the DSN.

The development of this facility was influenced by a DSN Advanced Systems program in which a limited prototype navigation system was developed for a VAX 11/780 minicomputer. This research effort demonstrated the feasibility of providing low cost and efficient multimission navigation support using a dedicated minicomputer. It also provided the opportunity to introduce new technical capabilities to enhance the navigation operations process. This has included automatic loading of data files, electronic transfer of files, interactive graphic display and data editing, and the capability of providing the quick turnaround needed for the near-real-time navigation support for critical Earth orbiter mission events.

Based on the results of the Advanced Systems demonstration, the HEO multimission capability (NAV) was imple-

mented within the Network Operations Control Center (NOCC) using the backup Network Support Controller VAX 11/780 (NSC1). The primary VAX 11/780 (NSC2) is used for the generation and transmission of station predicts, the generation of scheduling information, and the preparation of mission sequence of events. Data are transferred between NSC1 and NSC2 via the DECNET. In the event of a failure of either machine, the Advanced Systems VAX 11/780 serves as backup to NAV.

The implementation of the NAV subsystem was scheduled to meet August 1984 AMPTE (Active Magnetospheric Particle Tracer Explorers) mission requirements, which was the first high Earth orbiter mission supported by the MARK IVA system. The early completion of the deep space development phase enabled the system to be operational for the ICE (International Comet Explorer) mission support starting in March 1984. To date, the NAV facility has supported 13 missions for 7 different space agencies. This support has included navigation for launch, Earth orbiter, geosynchronous orbit transfer, deep space cruise, and planetary and comet encounter phases.

II. Functional Description

Figure 1 shows the NAV subsystem functional interfaces and data flow. The HEO facility is responsible for performing the following functions:

(1) Prelaunch navigation systems analysis. This includes defining tracking data requirements and establishing expected orbit determination accuracies.

(2) Radio metric data analysis.

(3) Operational orbit determination.

(4) Generation of precision spacecraft ephemerides for use by the NOCC Support Subsystem (NSS) and for transmission to external agencies.

(5) Spacecraft maneuver analysis.

(6) Generation of DSN navigation deliverables for flight projects and external agencies. These include processed tracking data files, spacecraft state vectors and spacecraft ephemerides, and other trajectory-related products.

As shown in Fig. 1, the NAV system receives the following data from external sources:

(1) Radio metric data transferred from the DSN subnets via the Ground Communication Facility (GCF) Digital Communication Subsystem.

(2) Correlated VLBI data from the NOCC Radio Science/ VLBI Processor Subsystem (NRV).

(3) Timing, polar motion, and media calibrations from the Tracking and Systems Analytical Calibrations Group (TSAC).

(4) Spacecraft state vectors generated by external agencies, typically transferred via the high-speed data lines in an Intercenter Vector format (ICV).

NAV delivers the following products for external transmission:

(1) Spacecraft ephemeris files for antenna pointing and frequency predictions.

(2) Validated radio metric data in the form of an Orbit Data File (ODF) for transfer to external agencies.

(3) State vector estimates generated by the NAV subsystem in the form of an ICV.

(4) Project-related trajectory products generated from the spacecraft ephemeris.

III. NAV Subsystem Organization

The NAV subsystem consists of the Radio Metric Data Conditioning (RMDC), Orbit Determination (OD), and Trajectory Analysis (TRAJ) subsystems. Figure 2 shows the organization of the NAV subsystem.

A. NAV Subsystem Characteristics

1. Radio metric data conditioning subsystem. The primary function of the RMDC is to receive the radio metric data and prepare an orbit data file for use by either JPL NAV or for transmission to external agencies. The RMDC receives radio metric data consisting of doppler, range, angles, VLBI, and other ancillary information (e.g., calibrations, reference frequencies, validation codes, mode indicators) required to process these data. The data are transmitted in real-time via an electrical interface and automatically loaded onto NSC1. The data are edited — this consists of reordering the incoming data, identifying invalid data, deleting unwanted data types, compressing doppler and narrowband VLBI, and applying calibrations based on station and spacecraft characteristics. The primary output is a validated tracking data file (ODF) in a format suitable for use by NAV or in a format for transmission via the high-speed data lines to external agencies. During critical periods, the RMDC must be capable of processing a one-hour span of data within 5 minutes; for non-critical periods RMDC must edit a one-week data file within 60 minutes.

2. Orbit determination subsystem (OD). The primary function of the orbit determination segment is to estimate

the spacecraft state along with other relevant parameters using the processed ODF provided by the RMDC. The OD component consists primarily of the Orbit Determination Program (ODP), which includes a complete set of algorithms necessary to model the motion of a planetary orbiter or a deep space probe and to model the radio metric observations. The program is capable of estimating the spacecraft state, spacecraft dynamic parameters (i.e., maneuvers, solar pressure constants, gas leaks), astrodynamic and geophysical parameters, and observational parameters. It has the capability of processing the data using either a gaussian least squares filter or a batch sequential filter and smoother algorithm which accounts for the effects of correlated process noise. Estimates and their statistics can be time-mapped and displayed in a variety of coordinate systems with respect to different reference frames. The ODP also displays the pre- and post-fit data residuals and provides an assessment of the quality of the radio metric data.

3. Trajectory analysis subsystem. The primary function of the Trajectory Analysis Segment is to generate a precise spacecraft ephemeris using either an estimated state provided by the ODP or a state vector estimate transmitted by an external agency. This spacecraft ephemeris is used as the source for the trajectory files and Probe Ephemeris Tapes (PET) transmitted to the NOCC Support Subsystem (NSS) for station antenna pointing and frequency predicts. It is also used to generate specific trajectory-related data required by the project, and to generate state vectors in the form of an Intercenter Vector (ICV) for transmission via the high-speed data lines to external agencies. TRAJ will also accept as input an ICV transmitted by an external source and generate a spacecraft ephemeris based on this state.

B. Auxiliary Capabilities

In addition to the primary functions, the NAV system includes the following capabilities:

(1) Simulation of radio metric data for use in covariance analysis studies, operations readiness testing and general operations test, and training activities. This capability has been used to provide external users with simulated data for compatibility testing.

(2) Generation of station view periods.

(3) Preparation of NAV-related mission support timelines and sequence of events.

(4) Procedures to monitor the consistency of orbit estimates.

(5) Procedures for the near-real-time assessment of maneuvers and determination of spacecraft spin rates.

C. Software Implementation

The HEO NAV subsystem consists principally of software modules inherited from the UNIVAC-based navigation system. The UNIVAC version of the software was converted to be compatible with the VAX. Additional models were added to meet the specific NAV mission requirements. These included Earth atmospheric models, an Earth tide model, and the ability to process GSTDN range data. The Earth orbiter navigation phase of the software was tested and certified by comparing results with those independently determined by the GSFC Orbit Determination Program. The deep space modules were certified using the test case library that had been established to certify the UNIVAC version.

IV. Mission Set

The multimission NAV system has provided navigation for launch, Earth orbiter, deep space cruise, planetary flyby, and comet encounter mission phases. All missions that encountered the comet Halley in March 1986 were supported by the NAV subsystem. The missions that have been supported by this facility from FY84 to mid-FY86 and the level of support are described in Table 1. Future missions (mid-FY86 through FY88) are given in Table 2. This future mission set consists primarily of geosynchronous orbiter missions which will require HEO NAV support during the geosynchronous orbit transfer phase.

Table 1. Missions supported: 1984 — 1986

Mission	Agency	Significant Dates	NAV Role	Interface
International Comet Explorer (ICE)	GSFC	Launched as ISEE-3: 08/12/78 Renamed ICE: 12/22/83 Encounter: 09/11/85	Orbit determination in parallel with GSFC during lunar swingby Dec 83 and then prime during interplanetary cruise, comet Giacobini-Zinner encounter and for subsequent solar wind observations upstream of comet Halley during 1986.	*From JPL*: Intercenter State Vectors (ICV) initially via Telex. Now via high-speed data line (HSDL) electrical interference.
Active Magnetospheric Particle Explorer (AMPTE)/ Charge Composition Explorer (CCE)	GSFC	Launched: 08/16/84 Projected Lifetime: 4 yr	Provide OD support and generate trajectory products throughout mission lifetime.	*From JPL*: ICV (TRK 2-17) sent over HSDLs. *From JPL*: Flight path information sent via tape to Applied Physics Lab (TRK 2-22).
AMPTE Ion Release Module (IRM)	German Space Operations Center (GSOC)	Launched: 08/16/84 Projected Lifetime: 1 yr	Provide OD solutions during first 10 days after launch for compatibility testing with GSOC. Generate and transmit orbit data files for DSN tracking data acquired during periods of cannister release. GSOC will provide ICV for Predix generation.	*From JPL*: ICV (TRK 2-17) for first 10 days for solution comparison via HSDL. *From GSOC*: ICV (TRK 2-17) for data acquisition during cannister release via HSDL. *From JPL*: Orbit data files (TRK 2-18) via HSDL.
(MST5) Sakigake	Institute of Space and Astronautical Science, Japan (ISAS)	Launched: 01/07/85 Encounter: 03/11/86	Assist ISAS with validation of their OD software by providing Voyager tracking data for testing, comparing OD solutions and providing analysis and software consulting service; includes plans for workshop at JPL. Provide OD solutions and tracking data for the first 10 days after launch. Provide OD consulting for 6 months following launch.	*From JPL*: ICV via telex. *From JPL*: ODF tape (TRK 2-18) covering 10-day period.

Table 1 (contd)

Mission	Agency	Significant Dates	NAV Role	Interface
VEGA Venus Balloon Experiment	Centre National D'Etudes Spatiales (CNES)	Launched VEGA1: 12/15/84 Encounter: 06/11/85	Determine VEGA orbits during heliocentric cruise and Venus flyby phases using L-band (1.667 GHz) delta DOR, one-way doppler combined with geocentric information provided by IKI.	*From CNES*: VEGA orbital state estimates and spacecraft dynamic parameters.
	Space Research Institute (IKI), USSR	Launched VEGA2: 12/21/84 Encounter: 06/15/85		*From CNES*: L-band (1.667 GHz) oscillator characteristics. *From JPL*: VLBI-based OD solutions.
			Provide processed delta DOR observations to support IKI post-Venus flyby maneuver.	*From JPL*: Processed delta DOR observations.
Giotto	European Space Agency (ESA)	Launched: 07/02/85 Encounter: 03/14/86	Assist ESA with OD software development and validation by comparing solutions using test cases based on Voyager tracking data; conduct joint workshops at JPL and ESOC.	*From JPL*: ODF files (TRK 2-18) via HSDL. *From ESOC*: ICV (TRK 2-17) via HSDL. *From JPL*: ICV (TRK 2-17) via HSDL.
			Provide orbit data files during cruise and encounter phases. Determine flight path and compare solutions for designated rehearsal periods during cruise.	*From ESOC*: State vector solutions, covariances and dynamic parameter estimates via Telex.
Pathfinder	European Space Agency (ESA)	Launched VEGA1: 12/15/84 Encounter: 03/06/86	Provide VLBI-determined orbit solutions for the heliocentric cruise phase (post-Venus encounter) and the comet Halley encounter phase.	*From ESOC*: Orbit state vectors; transponder characteristics.
	Space Research Institute (IKI), USSR	Launched VEGA2: 12/21/84 Encounter: 03/09/86		*From JPL*: VLBI-determined orbit estimates.
			Provide processed delta DOR observations.	*From JPL*: Processed delta DOR observations.
(Planet-A) Suisei	ISAS	Launched: 08/18/85 Encounter: 03/08/86	Provide OD solutions to support USUDA station predicts and tracking data covering the first 4 mo following launch. Provide consulting service for 6-mo period after launch.	*From JPL*: ICV via Telex. *From JPL*: ODF tape (TRK 2-18) for first 10 days.

Table 1 (contd)

Mission	Agency	Significant Dates	NAV Role	Interface
Broadcast Satellite-2B (BS-2B)	National Space Development Agency, Japan (NASDA)	Launched: 02/12/86 Support Period: Launch to L+30 days	Orbit determination during orbit transfer phase from L to L+(Apogee Kick Motor Firing) AKM + 30h with solutions at L+6.5 h, L+10 h, L+29 h, L+61 h; contingency support to L+1 mo; covariance studies to assess the performance of tracking from a single site for L to L+5 h.	*From NASDA*: ICV (TRK 2-17) via HSDL. *From JPL*: ICV (TRK 2-17) via HSDL. *From JPL*: ODF files (TRK 2-20) via HSDL GSTDN data only.
Extended Missions (ISEE-1, ISEE-2, DE-1, NIMBUS)	GSFC		Provide antenna predicts for 26-m subnet using ICV provided by GSFC.	ICV from GSFC.

Table 2. Future missions: 1986 — 1988

Mission	Agency	Date
GOES-H	NASA	10/86
TELECOM-1C	CNES	12/86
TV-SAT	GSOC	03/87
TDF-1	CNES	07/87
ETS-V	NASDA	08/87
TELEX-X	CNES	12/87
DFS-1	GSOC	02/88
CS-3A	NASDA	02/88
TV-SAT2	GSOC	05/88
TDF-2	CNES	07/88
CS-3B	NASDA	08/88
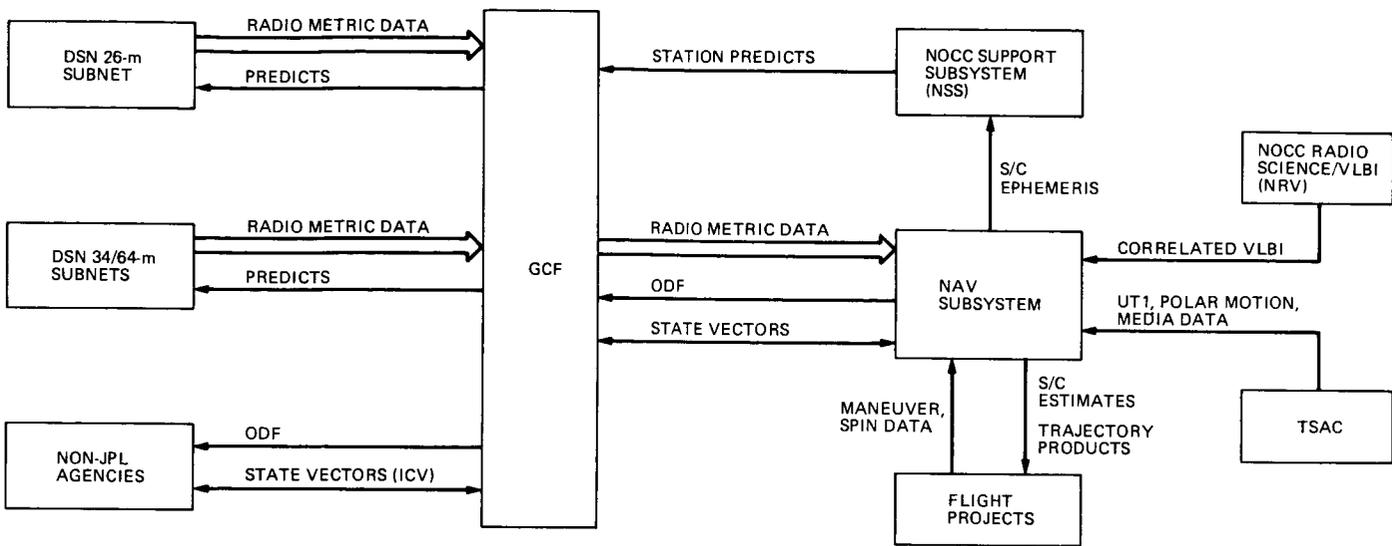
Fig. 1. NOCC Navigation Subsystem functional data flow
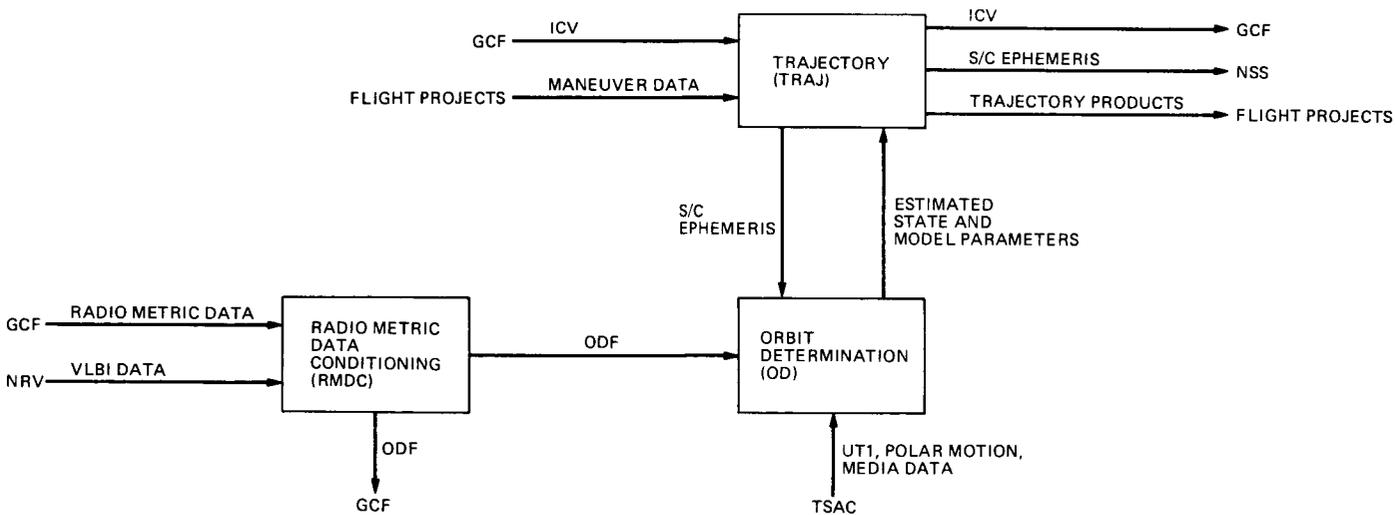
Fig. 2. NOCC Navigation Subsystem organization

ICE Navigation Support

L. Efron, R. J. Muellerschoen, and R. I. Premkumar

Navigation Systems Section

The first in-situ measurements at a comet occurred on 11 September 1985 when the International Cometary Explorer (ICE) passed through the tail of Comet Giacobini-Zinner approximately 7870 km downstream of the nucleus. Encounter took place 7 years after the spacecraft's original launch on 12 August 1978 as the International Sun Earth Explorer 3 (ISEE-3), part of a three-spacecraft project to study the interaction between the solar wind and the Earth's magnetosphere. Transfer to an interplanetary trajectory (and the name change) was performed via a 119-km-altitude, gravity-assist, lunar swingby on 22 December 1983. Navigation support during interplanetary cruise and comet encounter was provided by orbit determination utilizing radio metric data from the DSN 64-meter antennas in Goldstone, California and Madrid, Spain. Orbit solutions yielding predictions of 50-km geocentric delivery accuracy in the target aim plane were achieved during interplanetary cruise and at comet encounter using 6-to-12-week data arcs between periodic attitude-change maneuvers. One-sigma two-way range and range rate residuals were consistently 40 meters and 0.2 mm/s or better, respectively. Non-gravitational forces affected the comet's motion during late August and early September 1985 and caused a 2300-km shift in the orbit of the comet relative to the spacecraft. This necessitated a final ICE orbit trim maneuver 3 days prior to encounter. Near-real-time assessment of two-way 2-GHz (S-band) Doppler pseudo-residuals during the June and July 1985 trajectory change maneuvers aided in calibration of the spacecraft's thrusters in preparation for this final critical maneuver. Post-flight analysis indicates tail centerline passage was achieved within 10 seconds of the predicted time and geocentric position uncertainty at encounter was less than 40 km.

I. Introduction

A. Historical Overview

ICE was initially launched on 12 August 1978 as one of three spacecraft participating in a study of the interaction between the solar wind and the Earth's magnetosphere. Then known as ISEE-3, its instruments made measurements of particles and fields in the solar wind upstream of the Earth. Four years were spent in a six-month-period "halo" orbit in the vicinity of the Sun-Earth L_1 libration point located approximately 1.5×10^6 km (0.01 AU) from the Earth in the direction of the Sun along the Sun-Earth line (Refs. 1 and 2). The gravitational and centrifugal forces on an object on station at a libration point are in balance. Transfer-orbit

and halo-phase geometries are illustrated in Figs. 1 and 2. Occasional station-keeping was required to maintain the halo orbit about the libration point. Although inherently unstable, this configuration eliminated problems of the intense solar radio noise background, which would have plagued tracking a spacecraft located exactly at the L_1 libration point (Fig. 3).

On 10 June 1982, a propulsive ΔV maneuver of 4.5 m/s was used to nudge ISEE-3 out of its halo orbit and onto a geocentric flight path which included excursions into the Earth's geotail. Geotail mission phase geometry is illustrated in Figs. 4 and 5. During this phase of the mission, the flight path carried ISEE-3 into the vicinity of the Moon on four occasions. Lunar swingby dates are noted at points S_1, S_2, S_3, and S_4. Lunar swingby distances varied from 11.1 to 14.2 lunar radii. Propulsive ΔV maneuvers in conjunction with lunar gravity assists during these distant flybys were used to control the motion of the line of apsides, thereby permitting ISEE-3 to spend considerable time deep in the Earth's geotail. At one point, ISEE-3 spent over 3 months in the geotail at distances in excess of 200 Earth radii (RE). The maximum geotail distance was 236.6 RE on 30 June 1983.

All the geotail phase trajectory acrobatics between June 1982 and June 1983 were merely the prelude for the final two orbits of the Earth, which culminated with a propulsive trajectory correction maneuver (TCM) in November 1983. This targeted ISEE-3 for a close trailing-edge flyby of the Moon in late December 1983 (Fig. 6). A total of 15 propulsive correction maneuvers were carried out between 10 June 1982 and 23 November 1983. Four of these were planned and 11 were subsequent small trims required to correct for maneuver execution errors. Analysis and design of the entire sequence of maneuvers is discussed in Refs. 3, 4, and 5.

Targeting for the final lunar swingby was designed to place the spacecraft on a trajectory which would encounter Comet Giacobini-Zinner (G-Z) six days after its 5 September 1985 perihelion passage. As an added bonus, this same trajectory afforded ICE the opportunity to make measurements in the solar wind upstream of Halley's Comet late in March of 1986 (Fig. 7). This would occur several weeks after Halley was to play host to an armada of five visiting European and Japanese spacecraft (Russia's Vega 1 and Vega 2, ESA's Giotto, and Japan's Sakigake and Suisei).

From initial launch, the ISEE-3 mission was supported by the Ground Spaceflight Tracking and Data Network (GSTDN) with orbit determination (OD) performed at the Goddard Space Flight Center (GSFC). At the comet encounter distance of 70.5×10^6 km (0.47 AU), the spacecraft was well beyond the range of GSTDN. This necessitated the total transfer of support from GSTDN to the Deep Space Network (DSN) in January 1984. During December 1983, GSFC and the Jet Propulsion Laboratory (JPL) formally carried out parallel tracking and OD support. Responsibility for maneuver analysis and design remained at GSFC for the duration of the mission.

B. Spacecraft Description

The ICE spacecraft is depicted in Fig. 8. The cylindrical, drum-shaped spacecraft is spin stabilized, with its spin axis oriented perpendicular to the ecliptic plane. The tower, extending in the direction of the north ecliptic pole, supports the medium-gain-phased array 2-GHz (S-band) antenna, which has a flat, disk-like pattern. Table 1 provides a compilation of navigation-related spacecraft characteristics.

Orbit and attitude control are provided by a system of hydrazine engines on pods mounted on the outside and bottom of the main cylinder. The system consists of 12 canted nozzles to provide radial, axial, and spin rate thrusting control. Design goals were to allow propulsive ΔV maneuvers in any direction without perturbing the spin axis orientation. Solar radiation pressure forces acting on the spacecraft did, however, result in torques which slowly altered the spacecraft's attitude. Periodic attitude change maneuvers were necessary to keep the Earth centered in the antenna beam.

C. Navigation Challenge

Navigation support for ICE has been unique in several aspects. First and foremost was the large uncertainty in the predicted September 1985 position of the target at the time of interplanetary transfer orbit insertion in December 1983. From covariance studies, uncertainty in Comet G-Z's time of perihelion passage on 5 September 1985 was perhaps 0.1 day. This represented an anticipated along-track error of approximately 3×10^5 km, and complicated the targeting for the critical low-altitude gravity-assist lunar swingby. This critical event occurred several months before the first ground-based optical observation even confirmed the return of the comet to the inner solar system. Secondly, covariance analysis to demonstrate the capability of a limited subnet of the DSN to meet the ICE mission navigation requirements made assumptions in regard to anticipated tracking data quality. This was done prior to any DSN experience with the generation of radio metric data in conjunction with a transponder of the type aboard the probe. Finally, with the NASA decision that responsibility for orbit determination and maneuver design be split between two agencies, there was a need for verification of dynamics modeling and trajectory propagation compatibility between JPL and GSFC.

The premission navigation covariance analysis and DSN navigation support of the lunar swingby and early interplane-

tary cruise are described in detail in Refs. 6, 7, and 8. Ephemeris improvement efforts at JPL (led by D. K. Yeomans) utilizing ground-based optical observations gathered by the Astrometry Network of the International Halley Watch (IHW) are described in Refs. 8 and 9.

II. Metamorphosis to Cometary Explorer

The period from July to November 1983 was used by the DSN and JPL to demonstrate orbit solution compatibility in a mission operations support environment. From 1 December 1983 through 3 January 1984, the DSN and JPL provided backup tracking and orbit determination support to GSFC. This meant that complete orbit solution turnarounds were required as often as every two days at both centers. GSFC processed the GSTDN data, while JPL solutions utilized DSN-generated data. Mission planning was thereby able to make use of two independent orbit solution results for flight path analysis in preparation for the critical 22 December 1983 close lunar swingby.

To avoid the necessity of having to model maneuvers, orbit solutions were attempted between trajectory correction or attitude change events while ISEE-3 was in high Earth orbit. Trajectory correction maneuvers calculated using either JPL or GSFC orbit solutions were essentially the same. DSN two-way radio metric tracking data from two 34-m antenna sites yielded unbiased 1-sigma residuals that were consistently 0.2 mm/s for 60-s count-time Doppler and 30 to 50 m for range.

Targeting for the critical close lunar swingby began with a trim maneuver on 10 November 1983. A small execution error resulted in prediction of a lunar impact. On 23 November 1983, this situation was corrected by a small trim maneuver at the apoapsis of the spacecraft's final orbit of the Earth. From there, it began a month-long plunge toward the Moon (Fig. 6). The subsequent interplanetary trajectory of ICE relative to the orbits of both Comet G-Z and Halley's Comet is illustrated in Fig. 7. Spacecraft heliocentric inertial velocity at the lunar swingby was on the order of 30 km/s. A velocity vector change of 1 m/s at Earth-Moon departure was capable of producing up to a 50,000-km change in the Comet G-Z closest approach distance. The parameter of most importance in determining the lunar aim point was the final flyby altitude. Maneuver analysis indicated that a closest approach 119 km above the surface was desired. Trim and attitude change maneuvers were planned for 8 December and, if necessary, as many as two later opportunities for trim maneuvers prior to perilune were included in the mission sequence of events. Orbit solutions were performed three times a week. A minimum of 4 or 5 days' data arc was required to obtain solutions with reasonable uncertainties. On 1 December 1983, DSS-63

(the 64-meter antenna in Spain) became available to occasionally provide tracking support. Having three stations prior to encounter greatly enhanced the tracking geometry. The sequence of orbit solutions following the 23 November trim maneuver suggested that the actual trajectory was so close to nominal that only the planned 8 December attitude adjustment was performed. Perilune passage occurred on 22 December 1983 at 18:45:15.2 (ET) as ISEE-3 raced past the trailing limb of the Moon at an altitude of 119.4 km. At that moment, the spacecraft's identify formally underwent the change from ISEE-3 to ICE.

The JPL navigation team's use of DSN-generated radio metric Doppler and range data predicted the radius and time of closest approach to the Moon to better than 1 km and 1 s, respectively, during mission operations support. Post-perilune solutions have further reduced uncertainties in these swingby parameters to less than 100 m and 0.1 s. Discussion, in more detail, of the close lunar swingby is provided in Reference 8.

III. Cruise Orbit Determination

After 6 January 1984, tracking and data acquisition support of ICE was limited to the DSN 64-meter subnet. However, during the first year of interplanetary cruise, only one antenna, DSS-63, was available. Beginning in mid-December 1984, occasional passes were provided by DSS-14, at Goldstone. In mid-January 1985, DSS-63 went down for modifications, leaving only DSS-14 to provide support. After June 1985, both stations supported the mission through the comet encounter.

Daily DSN two-way tracking coverage was provided from lunar swingby until 10 January 1984. Afterwards, coverage was nominally 2 passes a week. One pass provided range-rate data coverage above a 15-degree elevation angle, horizon to horizon. Approximately 10 range points were collected over a subinterval of 30 minutes. The second pass was shorter in duration, limited to elevation angles above 30°, and typically only collected Doppler data. More frequent passes were provided for 10 days after attitude change maneuvers, which occurred at approximately 3-month intervals. Orbit solutions were limited to the data arcs between maneuvers. Data weights were 100 m for range and 1 mm/s for range rate. Residuals throughout the interplanetary phase continued to be unbiased and 50 m and 0.2 mm/s, 1 sigma, for range and range rate, respectively.

Comet encounter conditions were determined relative to the then-current Comet G-Z ephemeris. Cruise orbit solutions consistently predicted a comet miss distance of approximately 62×10^3 km on the sunward side of the comet. This would be outside the anticipated cometary coma. Trajectory correc-

tion was put off until June 1985 to allow comet ephemeris improvement based on post-recovery optical observations.

IV. Comet Giacobini-Zinner Recovery

Comet G-Z, with an orbital period of 6.5 years, was first described in 1900 by Michel Giacobini at the Nice Observatory in France. The comet was rediscovered in 1913 by Ernst Zinner at the Remeis Observatory in Bamberg, Germany. Its orbit is inclined at 32° to the ecliptic plane and varies from perihelion near the orbit of the Earth to aphelion between the orbits of Jupiter and Saturn. Early 1984 marked the eleventh opportunity for recorded Earth-based observation of the comet. The ICE encounter with Comet G-Z was planned at the nodal crossing of the ecliptic at a heliocentric distance of 1.03 AU six days after Comet G-Z's perihelion passage.

To effect an early telescopic recovery of the comet in 1984, an orbit and search ephemeris was computed based on astrometric observations from the three previous Comet G-Z apparitions in 1965, 1972, and 1979. The ephemeris was then distributed to observatories capable of observing objects with the extremely faint predicted magnitude of approximately 23.

An observing team using the 4-m aperture telescope at Kitt Peak National Observatory near Tucson, Arizona successfully recovered the comet on 3 April 1984. Confirmation (pre-recovery) images were then quickly reported from plates recorded 28 January 1984 at the European Southern Observatory in La Silla, Chile and for 28 March 1984 (again at Kitt Peak). When the recovery observations were included in an orbital solution update, the required correction to the predicted perihelion passage time was only +0.01 day.

By means of precise ephemerides from the Astrometry Network of the IHW, Comets Halley and Giacobini-Zinner were both recovered very close to their predicted positions. Both the successful recovery of Comet Halley on 16 October 1982 at Mt. Palomar and the Comet G-Z recovery on 3 April 1984 at Kitt Peak were made when the comets were far fainter than magnitude 23, making these two cometary recoveries the faintest on record. Conditions remained favorable for ground-based observations through August 1984 and again beginning early 1985. A total of 1031 observations were processed. They represented 72 observatories in 21 countries.

V. Comet Encounter

A. Homing In

With the availability of updated comet orbit solutions incorporating post-recovery observation data, the ICE/G-Z radius of closest approach remained about 62.5×10^3 km on

the sunward side of the nucleus. As confidence in the comet's orbit increased in the spring of 1985, conditions became favorable for a TCM to move the ICE aim point into the tail. By May 1985, the predicted 11 September position uncertainties were on the order of 1000 km for Comet G-Z and 50 km for ICE. On 5 June 1985, a ΔV of 39 m/s altered the trajectory. This moved the encounter to a point about 17.8×10^3 km from the comet on the anti-Sun side of the nucleus.

By the first week in July, formal uncertainty in the predicted 11 September ICE aim point (based on a 4-week data arc) was about 100 km, but solution stability and small residuals provided the confidence for a go-ahead on 9 July with a 1.25 m/s ΔV orbit trim maneuver. The purpose was to reduce the flyby distance to 10,000 km and get the trajectory on course for passage through the center of an aberrated tail. Tail aberration results from the comet's motion through the solar wind, which is flowing radially outward from a rotating Sun. The tail was modeled to lag the Sun-comet line by 5.4° and lie 0.1° above the comet orbit plane.

Post-trim orbit solutions indicated aberrated tail passage would be at a nucleus distance of about 9880 km and less than 100 km off the tail centerline. With perhaps as much as a 1000-km uncertainty in Comet G-Z's predicted position and only an intelligent guess for the tail aberration angle, the ICE Flight Dynamics Director, R. W. Farquhar, kept a watchful eye on Comet G-Z's ephemeris updates.

B. The Comet Burps

Observations which became available the last week of August indicated that non-gravitational forces were having an effect on Comet G-Z's motion. These forces may have been the result of explosive gas jetting due to the boiling off of volatiles within the icy nucleus as the comet approached its 5 September 1985 perihelion.

Orbit solutions for ICE from 20 August onward provided geocentric position uncertainty in the target aim plane (B-plane) under 50 km (Fig. 9). These statistics included conservative consideration of station location errors and a 10-percent error in the coefficient of reflectivity. However, updated comet ephemerides began to indicate a change in the predicted comet relative aim point. The encounter point on 11 September seemed to jump 2300 km to a point 7830 km downstream of the nucleus and over 600 km ahead of the aberrated tail axis. This effect is illustrated in Fig. 10.

C. The Final Trim

A decision now had to be made in regard to the final trajectory correction maneuver, planned for 8 September. Two mission requirements controlled trajectory design: that ICE

intercept the tail centerline 10,000 km downstream from the nucleus, and that encounter occur at 11:00 UTC. The scientific rationale for the first requirement is discussed in Ref. 9. The second requirement placed encounter near the time of ICE meridian passage at the 305-m Arecibo Radio Astronomy Observatory in Puerto Rico (Ref. 10). The project decided to leave the encounter distance unchanged and risk encountering the comet within 8,000 km of the nucleus. A small propulsive ΔV of 2.34 m/s was implemented to retarget for tail centerline passage. ICE was now on course for its historic encounter deep within the comet's visible coma at a point hidden from the view of Earth-based observers.

D. Bullseye Confirmed

The final ICE orbit solutions before 8 September 1983 coupled with the nominal a priori final trim ΔV predicted tail center closest approach on 11 September at approximately 11:02:24 UTC. This agrees closely with the latest orbit solutions in conjunction with the final definitive Comet G-Z ephemeris based on observations through 26 September 1985, which yielded an encounter time of 11:02:22 UTC at an encounter distance of 7870 km.

If modeled as a 1-km radius sphere with a density equal to water ice, the mass of Comet G-Z is less than 10^{-12} that of the Earth. Hence, the ICE flyby caused no detectable Doppler shift in the two-way 2-GHz (S-band) radio metric tracking data. The principal signature for determination of the actual closest approach would come from magnetometer detection of spacecraft passage through the comet's neutral (or plasma) sheet. This is illustrated in Fig. 11 and discussed in Ref. 9. The X-Y plane lies in the plane of the comet's orbit.

The vector helium magnetometer aboard ICE reported entrance and exit from the magnetotail at 10:59:40 UTC and 11:07:40 UTC, respectively (Ref. 11). This implies a centerline closest approach of 11:03:40 UTC, assuming symmetry and no wagging of the tail. However, Fig. 12 (Ref. 11) illustrates the magnetic field observations obtained during the comet tail traverse. Centerline passage is indicated by the sign reversal of the X-component of the magnetic field at approximately 11:02:30 UTC. The 70-second difference between tail centerline crossings defined by tail entry and exit times and plasma sheet passage can be explained by motion of the tail during the traverse. Such motion is supported by the magnetometer and other instrument measurements during the encounter.

Plasma sheet orientation is dictated by the external interplanetary magnetic field. The ICE experiment measurements now indicate spacecraft passage through the sheet at an incidence angle of 30°. The plasma sheet is believed to participate with any motion of the ion tail. Hence, there may have been some degree of luck in the close agreement of the time of plasma sheet centerline passage determined from orbit solutions for ICE and for Comet G-Z with that determined from magnetometer measurements.

E. Postmortem

At encounter, the spacecraft's velocity relative to the comet was 21 km/s. The 8-minute transit time through the tail implies a magnetotail diameter on the order of 10^4 km. Within the magnetotail was a plasma sheet approximately 10^3 km in thickness. The tail itself was enclosed in an extended coma at least 10^5 k in radius.

Figures 13 and 14 illustrate the post-encounter reconstruction of Comet G-Z's tail passage geometry. Tail centerline passage is indicated at about 11:02:25 UTC, several seconds after closest approach. This differs by 5 seconds from the centerline passage time cited above on the basis of a magnetic field reversed detection and 3 seconds from the final pre-encounter predictions.

Analysis of ICE's orbit solutions after 8 September provides confidence that the spacecraft geocentric position uncertainty at encounter was less than 40 km. In comparison, the geocentric Comet G-Z position uncertainty at the same time was about 400 km. Discussions with investigators associated with the magnetometer give confidence that the time of tail-centerline plasma-sheet passage, based on instrument measurements, and the joint spacecraft/comet orbit determinations agree to better than 10 seconds.

VI. Maneuver Support

The 2-GHz (S-band) two-way range and range rate residuals from ICE orbit solutions throughout the mission were comparable to those of other deep space missions supported by the DSN. Residuals from the 12-week data arc used for the definitive encounter orbit solution (9 September through 9 November 1985) were typical of all data arcs during the interplanetary phase of the mission. These are displayed in Figs. 15 and 16. The standard deviation of range rate residuals was consistently 0.2 mm/s for 600-second Doppler count times. Maximum range rate residual amplitudes were always less than 0.7 mm/s.

The ICE orbit inclination to the ecliptic is approximately 0.5° and the spacecraft spin axis is maintained nearly normal to the ecliptic. Therefore, out-of-plane and in-plane orbit

corrections were provided by combinations of continuous axial thruster burns and pulsed radial thruster burns, respectively. Pulse sector widths could be varied in the radial burn mode. Individual radial burn pulses imparted a velocity change of 7 mm/s.

Real-time monitoring of the range rate pseudo-residuals (observed minus calculated predicted measurements derived from an earlier orbit solution) allowed resolution of the line-of-sight component of maneuver-induced velocity changes to the sub-mm/s level. Near-real-time evaluation by this technique of the TCM performed on 5 June 1985 indicated "hot" burns, which agreed closely with the later non-real-time maneuver evaluation at the Goddard Flight Dynamics Facility (FDF). During the 9 July 1985 trim maneuver, the number of pulses in the final radial mode burn was altered based on the observed line-of-sight velocity changes. Non-real-time maneuver evaluation again confirmed the validity of the technique. The June and July maneuver evaluations both indicated "hot" burns. Enough data was now in hand to perform a recalibration of the thrusters in preparation for the final 8 September trim maneuver.

Real-time monitoring of the pseudo-residuals during the final trim indicated a maneuver execution error at approximately a 0.7-percent level. A change of 2 in the planned total of 287 pulses for the final radial burn was indicated, but not implemented in light of both the effort required to alter the command sequence and the uncertainty in the comet ephemeris. More discussion of the JPL/GSFC interaction in regard to the June, July, and September orbit adjustment maneuvers is contained in Section 3 of Ref. 12.

VII. What's Next?

A 1.2-m/s radial ΔV implemented on 27 February 1986 and a 38.5-m/s axial ΔV executed on 7 April 1986 are part of the long-range plan to allow ICE to return to the vicinity of the Earth in the year 2014. The orbit of ICE from April 1986 through August 2014 is illustrated in Fig. 17 relative to a fixed Sun-Earth line. Numerical integration of the current ICE orbit indicates approach to both the Earth and Moon at about 5×10^5 km on 10 August 2014. When ICE is in the vicinity of solar opposition during 1998-1999, the spacecraft will appear to be within 2.5° of the Sun (10 solar radii) during a 14-month period from September 1998 to October 1999. If link margins are adequate and the s/c remains healthy during this time frame, the opportunity may exist for radio science investigations of the solar corona.

ICE may perform another mission if the spacecraft survives until 2 days short of its 36th birthday. Through 7 April 1986, the ICE spacecraft had performed a total of 106 separate propulsive trajectory, attitude, and spin-rate change maneuvers. Of the original 89 kg of hydrazine onboard at launch, 29 kg remain available. A propulsive ΔV in conjunction with another close lunar swingby could achieve recapture into Earth orbit. On the assumption that passage through the Comet G-Z coma and tail subjected ICE's surface to a bombardment of dust and ions, it is conceivable that NASA may have a candidate for a low-cost sample-return mission.

References

1. Farquhar, R. W., Muhonen, D. P., and Richardson, D. L., "Mission Design for a Halo Orbiter of the Earth," *Journal of Spacecraft and Rockets*, Vol. 14, No. 3, pp. 170–177, March 1977.

2. Farquhar, R. W., Muhonen, D. P., Newman, C. R., and Heuberger, H. S., "Trajectories and Orbital Maneuvers for the First Libration Point Satellite," *Journal of Guidance and Control*, Vol. 3, No. 6, pp. 549–554, November-December 1980.

3. Farquhar, R. W., and Dunham, D. W., "A New Trajectory Concept for Exploring the Earth's Geomagnetic Tail," *Journal of Guidance and Control*, Vol. 4, No. 2, pp. 192–196, March-April 1981.

4. Muhonen, D. P., Davis, S. A., and Dunham, D. W., "Alternative Gravity-Assist Sequences for the ISEE-3 Escape Trajectory," *Journal of Astronautical Sciences*, Vol. 33, No. 3, pp. 255–273, July-September 1985.

5. Farquhar, R. W., Muhonen, D. P., and Church, L. C., "Trajectories and Orbital Maneuvers for the ISEE-3/ICE Comet Mission," *Journal of the Astronautical Sciences*, Vol. 33, No. 3, pp. 235–254, July-September 1985.

6. Joyce, J. B., Leszkiewicz, S. J., and Schanzle, A. F., "Trajectory Determination Support and Analysis for ISEE-3 from Halo Orbit to Escape from the Earth/Moon System," AIAA Paper No. 84-1980, August 1984.

7. Efron, L., Ellis, J., Yeomans, D. K., Chodas, P. W., and Premkumar, R. I., "ISEE-3/ICE Navigation Analysis," AIAA Paper No. 84-1981, August 1984.

8. Efron, L., Yeomans, D. K., and Schanzle, A. F., "ISEE-3/ICE Navigation Analysis," *Journal of the Astronautical Sciences*, Vol. 33, No. 3, pp. 301–323, July-September 1985.

9. Yeomans, D. K., and Brandt, J. C., "The Comet Giacobini-Zinner Handbook," Jet Propulsion Laboratory, Pasadena, Calif., JPL 400-254, March 1985.

10. Fanelli, N. A., and Morris, D. G., "ICE Encounter Operations," *TDA Progress Report 42-84*, pp. 176–185, October-December 1985, Jet Propulsion Laboratory, Pasadena, Calif.

11. Slavin, J. A., Smith, E. J., Tsurutani, B. T., Siscoe, G. L., Jones, D. E., and Mendis, D. A., "Giacobini-Zinner Magnetotail: ICE Magnetic Field Observations," *Geophysical Research Letters*, Vol. 13, No. 28, p. 3, 1986.

12. Roberts, C. E., "International Cometary Explorer (ICE) Maneuvers 82 through 99," Goddard Space Flight Center Report CSC/TM-85/6117, prepared by Computer Sciences Corporation, December 1985.

Table 1. Spacecraft Characteristics

Feature	Description/value/resolution
Bus shape	16-sided cylindrical drum
Bus height	1.58 m
Bus diameter	1.77 m
Mass	479 kg at launch
	Experiments = 104 kg
	Hydrazine = 89 kg = 430 m/s ΔV
Antennas	
Radial (4)	92 m end to end
Axial (2)	14 m end to end
Solar array	2 bands around cylinder
	182 W at 82 V (at launch)
Spin rate (Sun) sensor (2)	$\pm 1.0°$ (nominal rate = 19.75 $\pm$ 0.2 rpm)
Attitude sensor	$\pm 0.25°$ ($0.1°$ with 4-week data arc)
Hydrazine thrusters (12)	1.1 N
Transponders (2)	2-GHz (S-band) for tracking, telemetry, and command
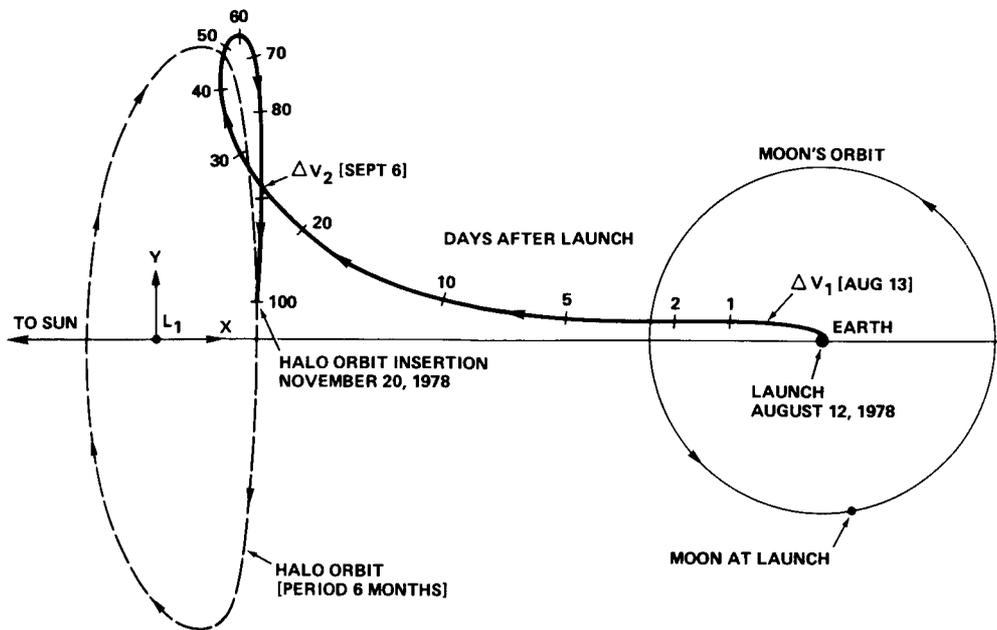
Fig. 1. Transfer trajectory to halo orbit relative to fixed Sun—Earth line
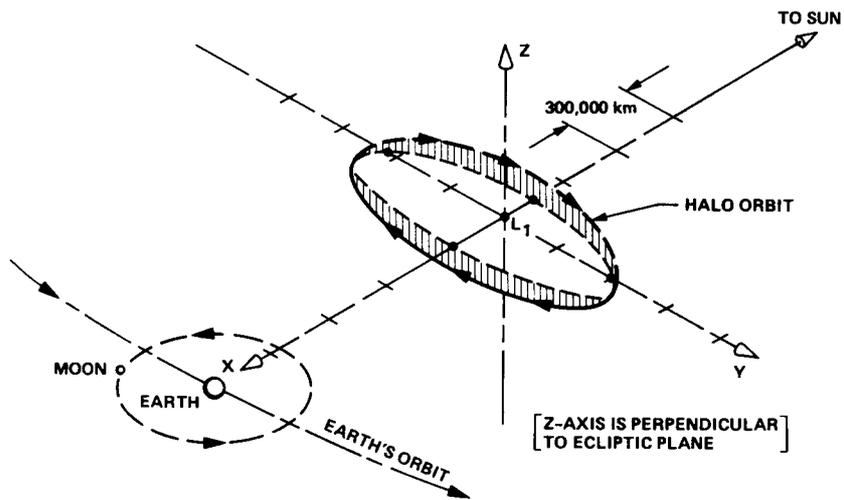
Fig. 2. Halo orbit around the Sun—Earth L$_1$, libration point
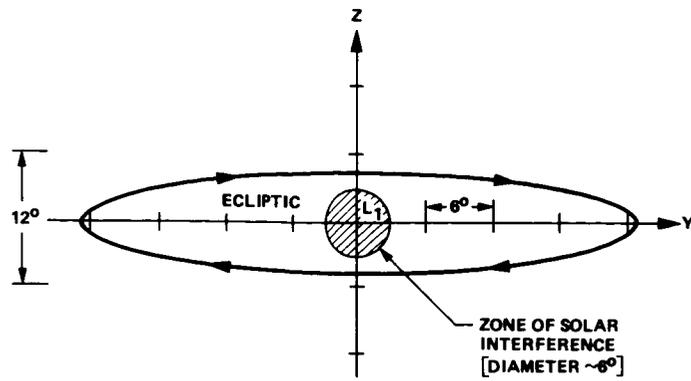
Fig. 3. Halo orbit as seen from Earth
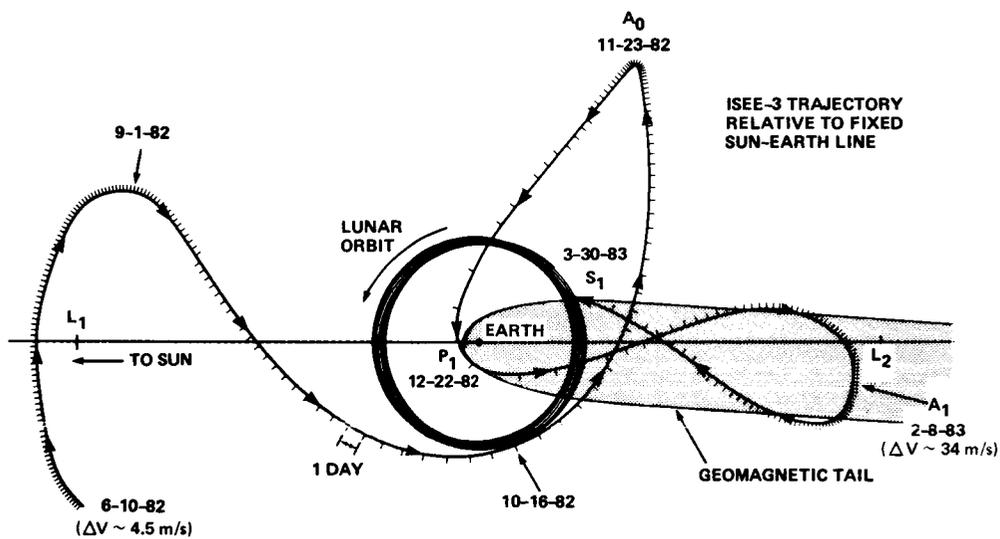
Fig. 4. Transfer from L_1 halo orbit to geomagnetic tail relative to fixed Sun—Earth line
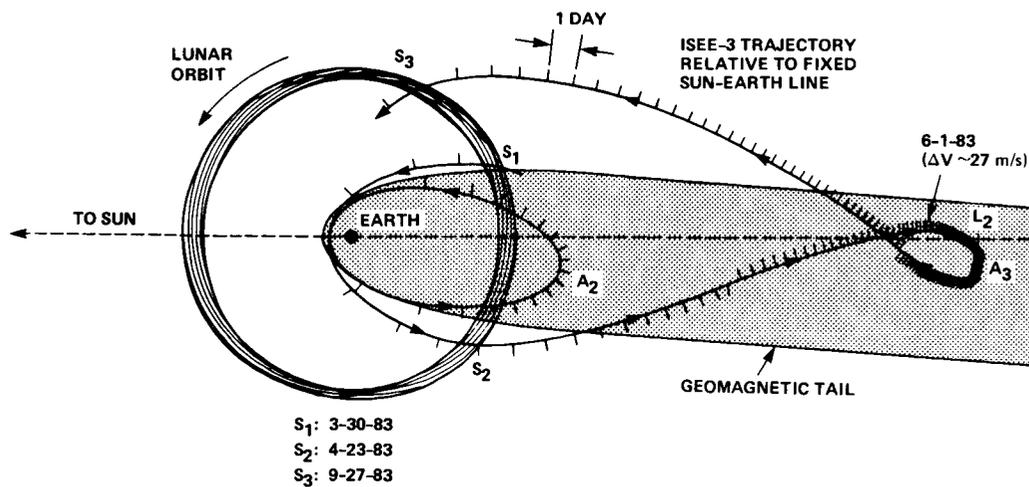
Fig. 5. Five-month geotail excursion relative to fixed Sun—Earth line
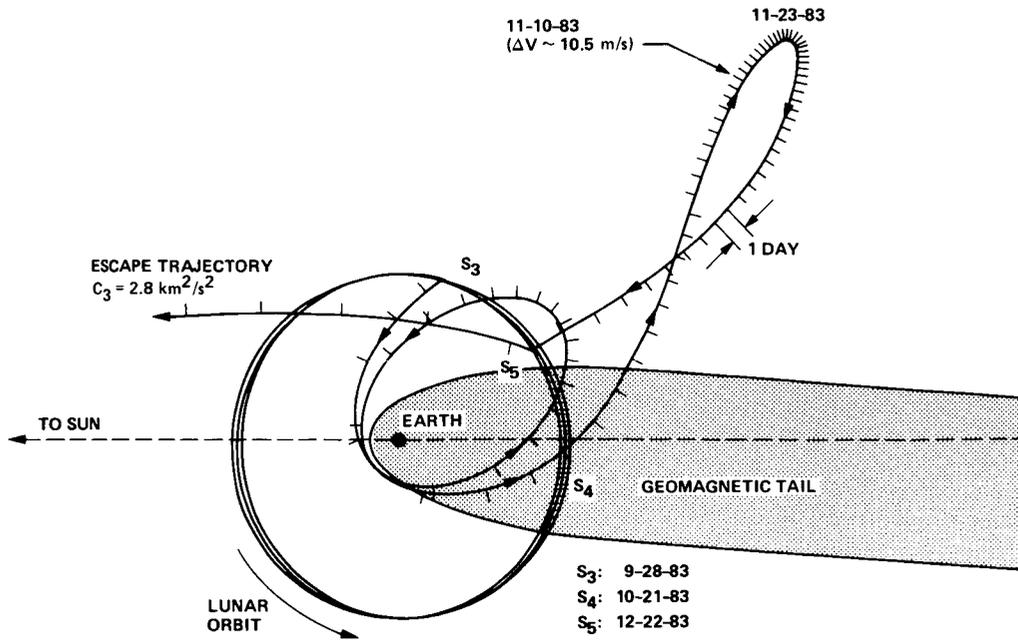
Fig. 6. Escape trajectory relative to fixed Sun−Earth line

11-10-83
($\Delta V \sim$ 10.5 m/s)

11-23-83

ESCAPE TRAJECTORY
C_3 = 2.8 km^2/s^2

S_3

TO SUN

EARTH

S_5

1 DAY

S_4

GEOMAGNETIC TAIL

LUNAR ORBIT

S_3: 9-28-83
S_4: 10-21-83
S_5: 12-22-83
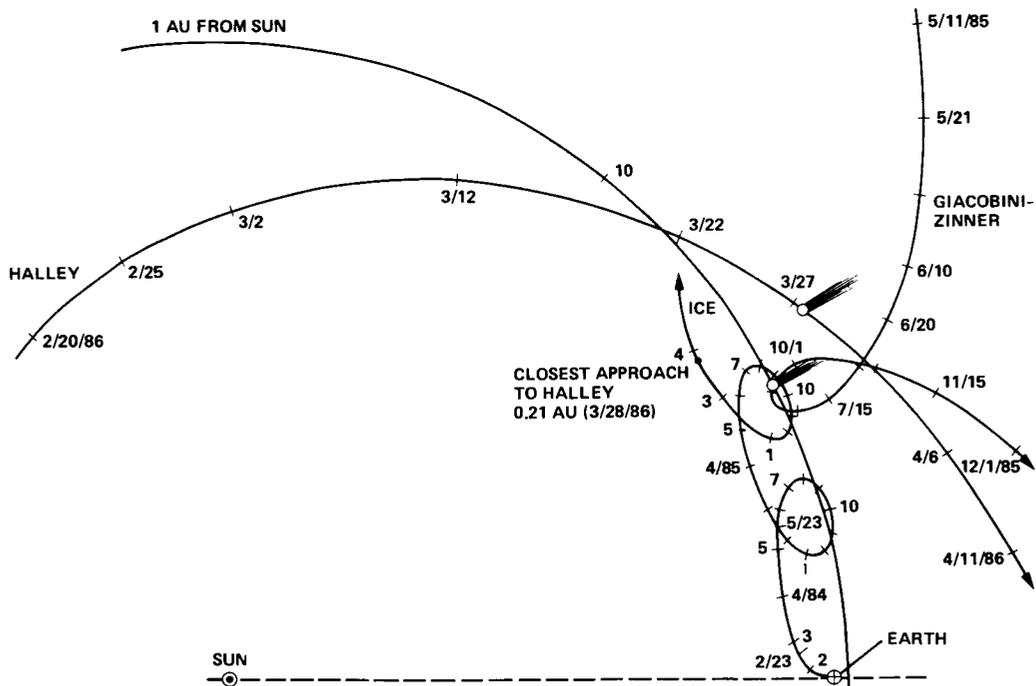
Fig. 7. ICE, Halley's Comet, and Comet G-Z trajectories relative to fixed Sun−Earth line

1 AU FROM SUN

5/11/85

5/21

10

3/12

3/22

GIACOBINI-
ZINNER

HALLEY

3/2

3/27

6/10

2/25

ICE

6/20

2/20/86

10/1

11/15

CLOSEST APPROACH
TO HALLEY
0.21 AU (3/28/86)

7

4

3

10

7/15

5

1

4/85

7

4/6

12/1/85

5/23

10

4/11/86

5

4/84

3

EARTH

SUN

2/23

2

TOTAL S/C WEIGHT: 479 kg

| EXPERIMENTS: | 104 kg |
| HYDRAZINE: | 89 kg |

DIMENSIONS (MAIN BODY)

| DIAMETER: | 1.77 m |
| HEIGHT: | 1.58 m |

3D RADIO MAPPING ANTENNA
LENGTH (TIP TO TIP)

AXIAL: 14 m
RADIAL: 92 m

SHORT ELECTRIC
ANTENNA

SEARCH COIL

AXIAL ΔV
AND ATTITUDE
CONTROL
THRUSTER

RADIAL
THRUSTERS

AXIAL ΔV AND ATTITUDE
CONTROL THRUSTER

SPIN AND DESPIN
THRUSTERS

MEDIUM GAIN
2-GHz (S-BAND)
ANTENNA

X-RAY
TELESCOPE

EXPERIMENT
BAYS

MAGNETOMETER

SOLAR
ARRAY

SPIN STABILIZED AT 19.75 RPM

Fig. 8. ICE spacecraft

| SOLUTION | DATA ARC | | | SEMIMAJOR |
	FROM	TO	DAYS	AXIS (km)
1	7/10	8/13	35	78
2	7/10	8/20	42	45
3	7/10	8/27	49	38
4	7/10	9/03	56	37

150 km

150 km

T

R

Fig. 9. ICE pre-encounter target aim plane error ellipses

COMET NUCLEUS

TO SUN

5.4°

3000 km

6000 km

9000 km

ABERRATED TAIL CENTERLINE

G-Z 31

G-Z 29

G-Z 26

G-Z 24

G-Z 16, 18, 20

G-Z EPHEM	LAST OBSERVATION
16	29 MAY 85
18	24 JUNE 85
20	21 JULY 85
24	26 AUG 85
26	2 SEPT 85
29	5 SEPT 85
31	26 SEPT 85

INDICATES COMET ORBIT PLANE PASSAGE
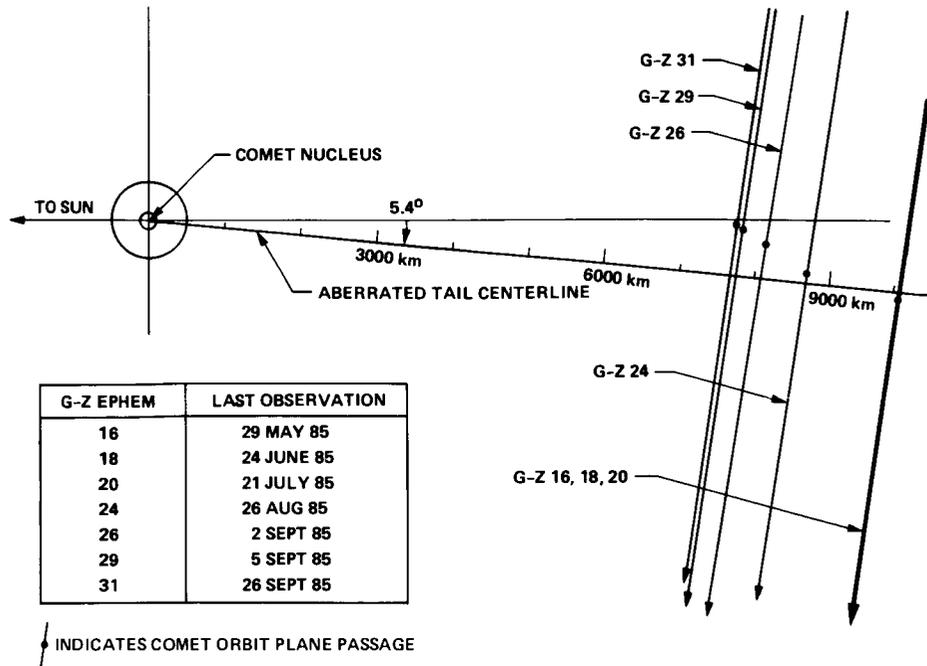
Fig. 10. Projection of the ICE trajectory before 8 September final trim onto the Comet
G-Z orbit plane as a function of comet ephemeris

SOLAR WIND

Z

X

Y

VISIBLE TAIL

MAGNETIC FIELD LINES
ORGANIZED INTO 2
LOBES OF N (NORTH)
AND S (SOUTH)
POLARITY

BOW SHOCK

ELECTRIC CURRENT
FLOWS IN PLANE
(PLASMA SHEET)
SEPARATING LOBES
OF OPPOSITE
MAGNETIC POLARITY
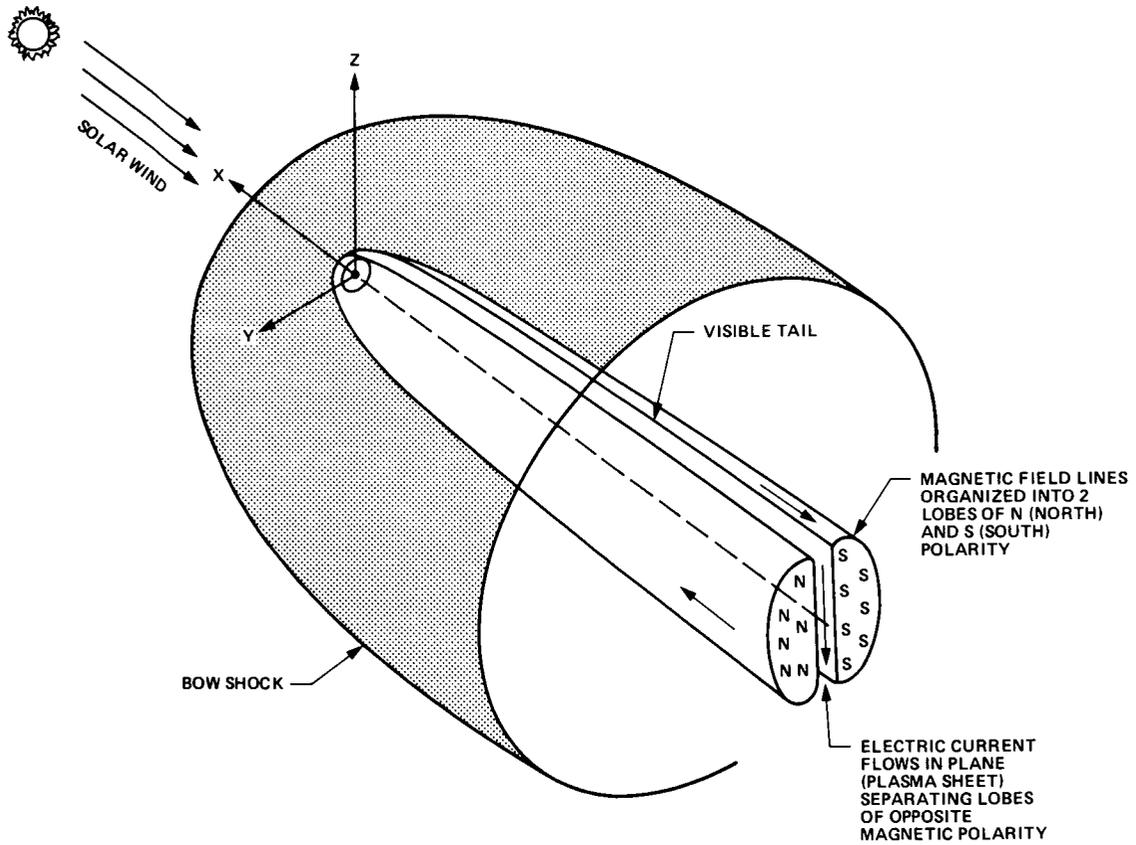
Fig. 11. Comet with visible ion tail composed of magnetic lobes separated by plasma sheet (from Ref. 9)
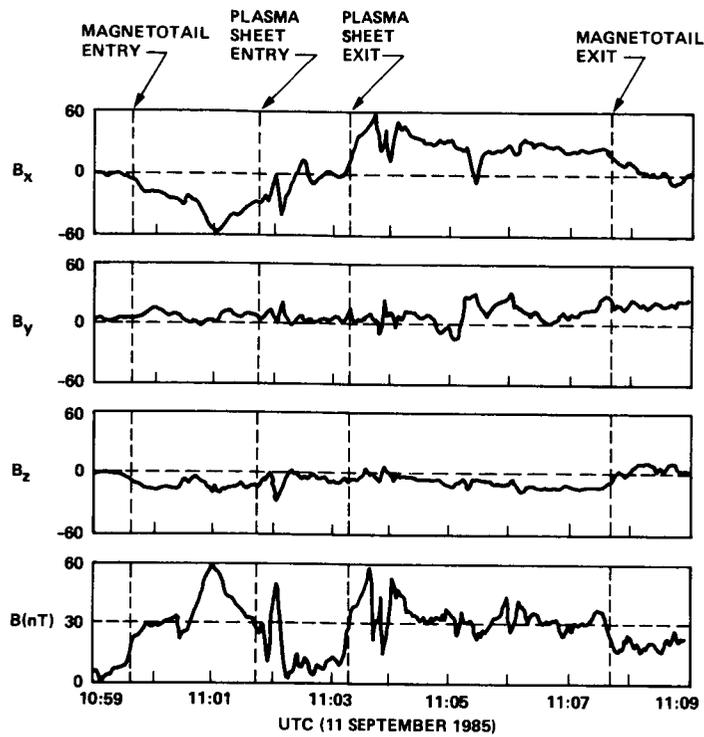
Fig. 12. ICE magnetometer magnetic field observations during the Comet G-Z magnetotail traverse

TIME TICKS ARE AT 10-second INTERVALS
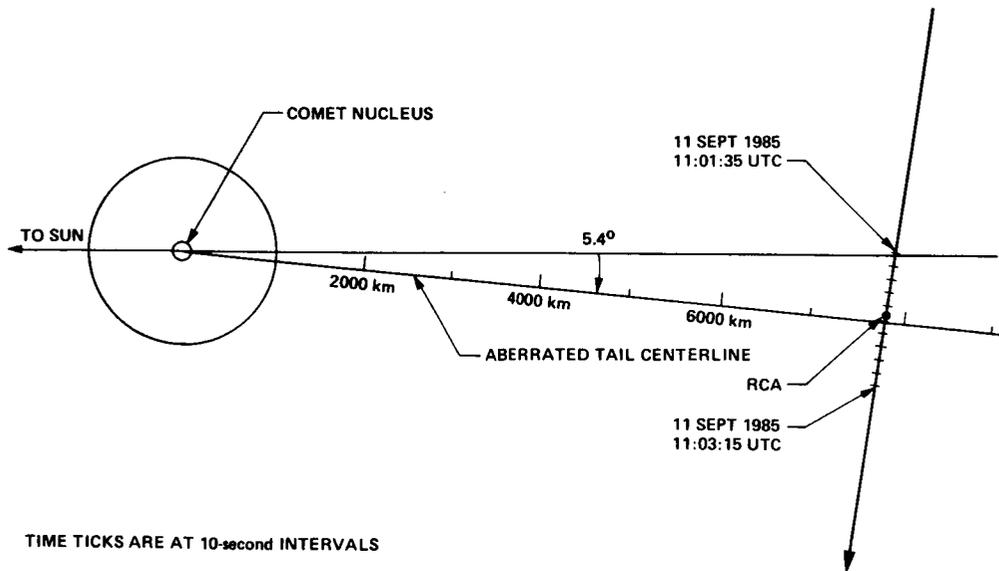
Fig. 13. Projection of ICE encounter trajectory onto the Comet G-Z orbit plane relative to a fixed Sun–Comet line

NOTE:
AT THE ICE ENCOUNTER DISTANCE, THE TAIL CENTERLINE IS
ASSUMED TO LIE ≈14 km ABOVE THE COMET ORBIT PLANE

TIME TICKS ARE AT 10-second INTERVALS
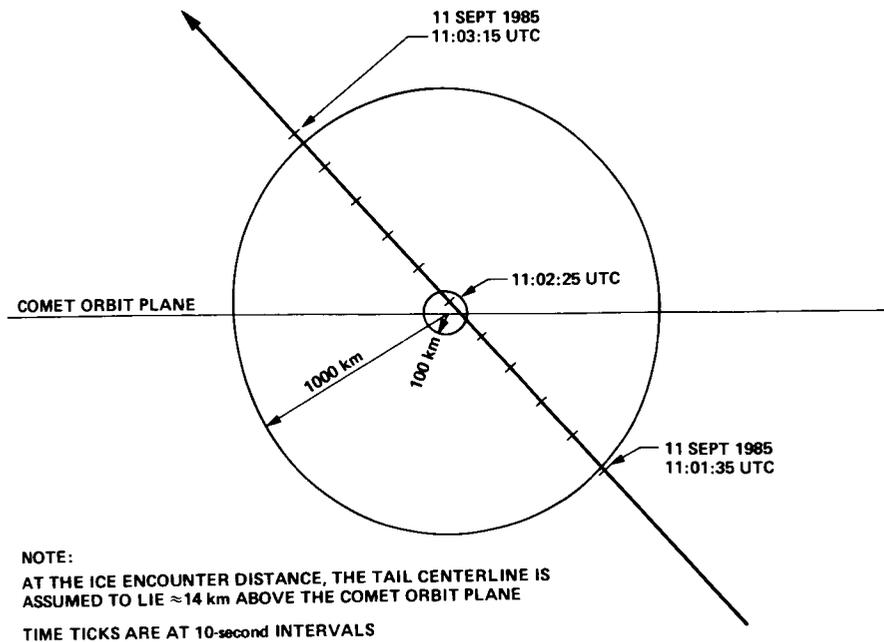
Fig. 14. Projection of ICE encounter trajectory as viewed along the aberrated
tail centerline looking toward the nucleus
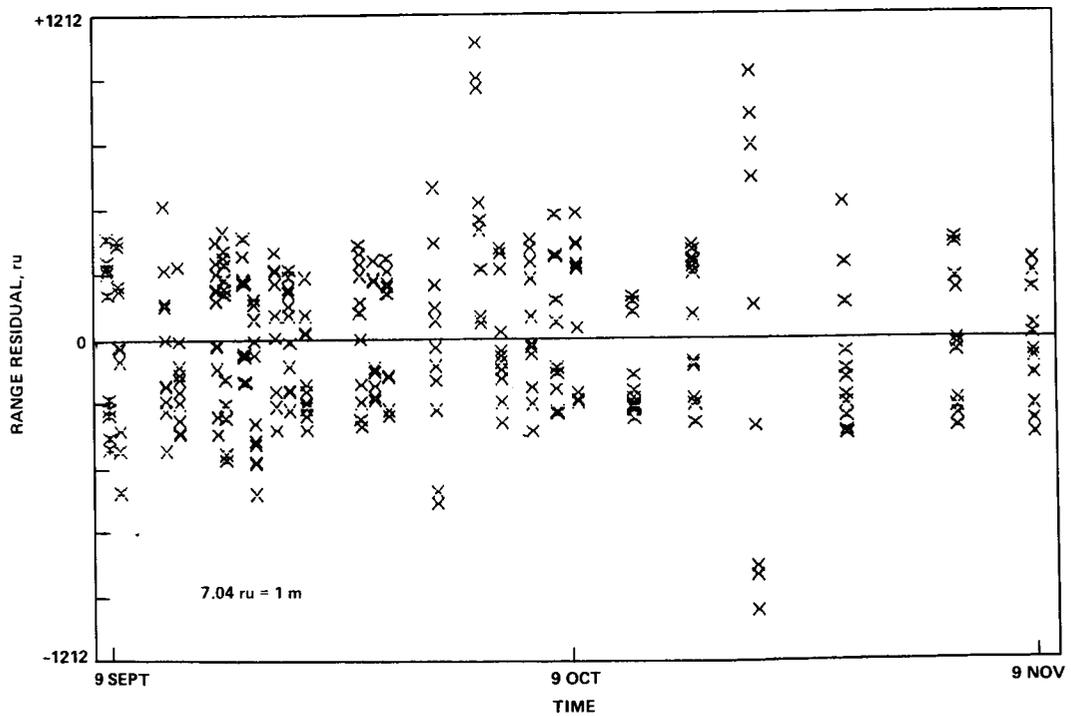
Fig. 15. 2-GHz (S-band) two-way range residuals from comet encounter solution data arc
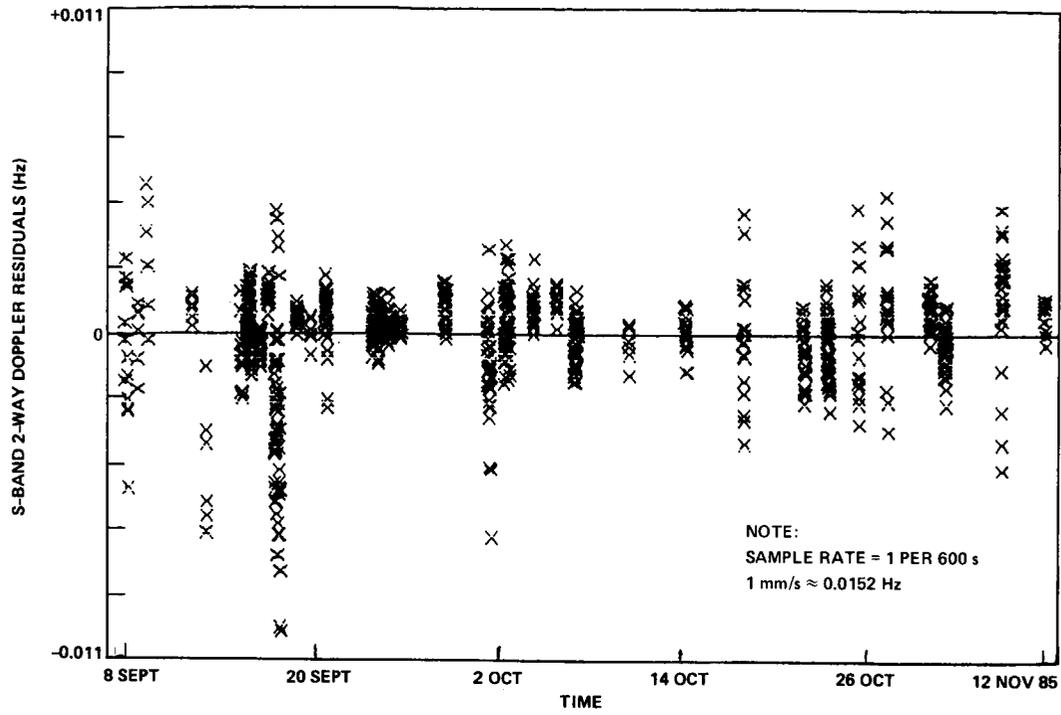
Fig. 16. 2-GHz (S-band) two-way range rate residuals from comet encounter solution data arc
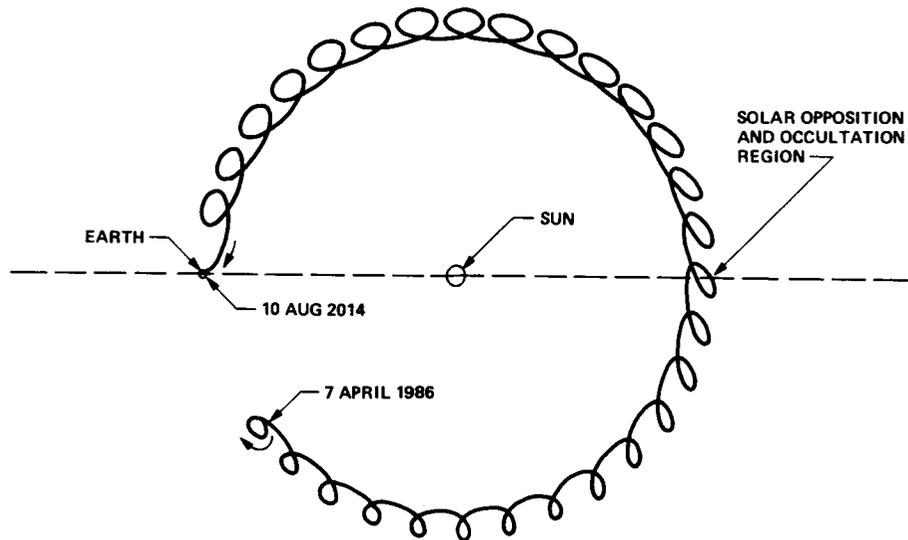
Fig. 17. ICE trajectory April 1986 to August 2014 relative to a fixed Sun–Earth line

TDA Progress Report 42-86

Objectives and First Results of the NASA SETI Sky Survey Field Tests at Goldstone

S. Gulkis, M. J. Klein, and E. T. Olsen
Space Physics and Astrophysics Section

R. B. Crow and R. M. Gosline
Radio Frequency and Microwave Subsystems Section

G. S. Downs and M. P. Quirk
Communications Systems Research Section

A. Lokshin
Automated Systems Section

J. Solomon
Imaging Processing Applications and Development Section

Field tests of SETI prototype hardware and software began in March 1985 at Goldstone. With emphasis on the Sky Survey component of the NASA SETI search strategy, the article describes the survey characteristics, the detection strategy, and preliminary results of system tests.

I. Introduction

The long-range plan (Ref. 1) for the NASA SETI program calls for a systematic search for microwave signals of extraterrestrial intelligent origin. The search will be conducted with existing radio telescopes, state-of-the-art low-noise receiving systems, and special purpose data acquisition and analysis systems. Two complementary search strategies have been adopted (Ref. 2): a Sky Survey of the entire celestial sphere in the 1-10-GHz frequency range and a high sensitivity Targeted Search in the 1-3-GHz range.

The program is currently in a research and development phase that is being jointly carried out by the NASA Ames Research Center (ARC), Moffett Field, CA, and JPL. The two centers are developing detailed plans to begin the full-scale search in the early 1990s.

The near-term plans are to develop prototype instrumentation, search algorithms and observing procedures, and to test the system concepts under observatory conditions. The initial field tests, using prototype hardware and software,

began at DSS 13 in March 1985. An aerial view of the site is shown in Fig. 1. This article reports the objectives and preliminary results of the test activities that relate primarily to the Sky Survey component of the NASA search plan.

II. Field Test Instrumentation

The field test instrumentation consisted of a 26-m antenna, low noise receiving systems in the 2200–2290-MHz and 8400–8500-MHz bands, a 65K channel Fast Fourier Transform (FFT) spectrum analyzer, and a prototype spectrum analyzer system designed specifically for SETI. The prototype system consists of a 72K channel digital spectrum analyzer, a VAX 11/750 computer and an interactive graphics terminal. The prototype analyzer, known as the Multichannel Spectrum Analyzer (MCSA), was designed and built at Stanford University (Ref. 3). The FFT spectrum analyzer is a multi-user facility designed and built at JPL (Refs. 4 and 5) and operated by the Deep Space Network. It is housed in a movable van, which is currently parked near the DSS 13 control building for the SETI field tests.

The MCSA and the FFT analyzers were connected to a common receiver output, and both systems were used to detect signals of various strengths from the Voyager 2 and the Pioneer 10, 11, and 12 spacecraft. Modifications were made to the MCSA to implement and test baseline and threshold algorithms in the processors. The MSCA system was also used to test signal detection algorithms that will be used for the Targeted Search. The results of these tests will be reported elsewhere.

A portable spectral surveillance system was installed at DSS 13 for the field tests. This system was designed and built for SETI to enable the SETI team to survey the RFI environment over the frequency range 1–10 GHz. The system consists of a 1-m paraboloidal antenna, an RF amplifier module that can be sequentially tuned over the 1–10-GHz band, a Tektronic swept-frequency spectrum analyzer, a programmable Tektronic controller, and a floppy disk system. The system is designed to operate unattended for several days at a time. It is also portable so that it can be moved to other observatory sites that are candidates for SETI observations.

III. Sky Survey Approach

The emphasis for the Sky Survey is complete sky coverage and expanded frequency coverage relative to the Targeted Search (Refs. 6 and 7). This component of the search strategy will be capable of detecting a class of strong signals, whose location may be unpredictable. We estimate that more than a million solar-like stars lie within a thousand light years of the

Earth, and that their density per unit steradian is approximately uniform. Since we do not know how strong a signal might be, it may be true that there is no preferred direction within this volume of stars. A survey of the entire sky insures that all potential life site directions are observed.

To complete a microwave survey of the sky over a wide range of frequencies in a reasonable amount of time, one must scan the antenna at a rate that is considerably faster than, e.g., the sidereal rate. However, rapid scans result in less sensitivity because a signal remains in the antenna beam for only a short time. We have adopted a reasonable survey time (approximately 5–7 years) and frequency coverage (1–10 GHz) and have designed our strategy to achieve maximum sensitivity and spatial uniformity. The characteristics of the Sky Survey are listed in Table 1.

To meet these objectives and requirements, the system design and the observing procedures are selected through an iterative process involving the scan strategy, the signal detection strategy, the limitations of the antenna drive system, and the signal processing algorithms that will recognize and ignore radio frequency interference (Ref. 8). Potential strategies and signal detection concepts are being tested as part of the field test activity.

A. Sky Pixelation

Recognizing that SETI will not be the sole user of the antenna, we decided to subdivide the celestial sphere into elements that can be observed within a few hours and be easily incorporated into a complete sky map with minimum overlap. Figure 2 is a schematic representation of the scan strategy that is currently being tested. The sky is divided into 10- by 60-deg pixels along lines of declination and right ascension. The third-dimension in the figure represents the frequency bins of the spectrum analyzer and Sky Survey processor system. The pixels visible to the observatory are mapped by scanning the antenna beam in the boustrophedonic pattern depicted schematically on the frequency element in the foreground. Alternative scan patterns will be considered for parts of the sky where antenna drive rates become excessively high, e.g., near the celestial poles (for equatorially mounted antennas) and near the zenith (for azimuth–elevation systems).

The field test activity includes tests of the scan strategy to evaluate the effect of variations of system temperature with antenna motion, the survey time that is lost during changes in scan direction, and the advantage of comparing signals from adjacent scans. The adjacent scan comparison is an important element in the Sky Survey processor strategy described below.

B. Sky Survey Detection Strategy

In the operational system, the Sky Survey processor will identify and record potential SETI "events" in each of 10-million frequency bins derived from a wideband spectrum analyzer. For the field tests, software modifications in the 65,000-channel FFT analyzer will support quasi-real-time tests of the Sky Survey detection algorithms.

The detection strategy being tested is depicted in Fig. 3. The antenna beam is swept back and forth across the sky pixel. Consecutive scans are separated slightly less than the half power beamwidth (HPBW) as indicated by the two cross-hatched circles. The power of each of the complex points from the spectrum analyzer is accumulated, baselined, and passed through a digital convolutional filter that is programmed to match the antenna response as it sweeps across a point source. The data in each frequency bin are processed and stored in memory just long enough to enable the processor to compare data from adjacent scans. These processing steps are schematically shown by the two panels in the lower left corner of Fig. 3.

Signals in each frequency bin that surpass a preset threshold and that appear at the same location on the sky on at least two adjacent scans are identified as a "SETI event." Signals that pass threshold but fail the SETI event tests (e.g., repeatable, fixed on the sky and exhibiting "reasonable" Doppler frequency drifts) are either saved in the Radio Astronomy data file or designated as RFI and discarded (lower right hand panel in the figure).

Signals that pass the SETI event test are time tagged, the location on the sky is computed, and the relevant information is recorded in a disk file for further analysis. Confirmation observations of events will then be carried out after the pixel observation is completed. Events passing this confirmation will be re-observed at later times and, perhaps, by a different radio astronomy observatory.

IV. Sky Survey Field Tests

The principal objectives of the field test activity are to provide verification of hardware designs, to test and further develop the signal detection strategy described above, to develop automated search procedures, and to characterize the RFI environment at the site over the 1–10-GHz frequency range. In the process, a limited observing program will be conducted from time to time in the 2200–2290-MHz and 8400–8500-MHz frequency bands using the extant receiving systems at DSS 13. These observations are intended to provide end-to-end system tests of the survey hardware, software and search strategy. Results of these tests, which will undoubtedly include RFI, will be compared with theoretical performance characteristics based on Gaussian noise statistics. The observational experience with the Sky Survey system will also stimulate ideas on ways to use the SETI system and/or the data for Radio Astronomy experiments.

The system configuration currently in use for the Sky Survey field tests at DSS 13 is shown in Fig. 4.

A. Pioneers 11 and 12

Among the first of the SETI field tests at DSS 13 was the detection of narrow-band signals from some of NASA's spacecraft. Signals received by the 26-m antenna from Voyager and Pioneer spacecraft were processed by the prototype MCSA and by the DSN FFT analyzer. In Fig. 5 we show the output from the FFT analyzer at three spectral resolutions for Pioneer's 11 and 12.

The upper left panel shows the suppressed carrier and the upper and lower sidebands of the signal from Pioneer 11 (PN11). At its distance from Earth of more than 18 Astronomical Units (AU), the signal is clearly the weaker of the two spacecraft. The resolution is 305 Hz, which is the normal resolution of the FFT running at full speed.

The signal in the upper right panel is from the Pioneer 12 (PN12) spacecraft in orbit around the planet Venus. The resolution of the FFT for this test was 256 Hz. Note that the carrier does not appear at this resolution because it is attenuated by a 300-Hz high-pass filter in the analyzer. Being so much closer to Earth, the PN12 signal is more than 20 dB stronger than that of PN11. The signal levels in the spectra from both spacecraft were calibrated by injecting broadband noise from a thermally controlled test diode. (Note that a narrowband test signal, which also appears in the figure, was used for a different purpose.) The comparison demonstrates the dynamic range of the system.

The lower right panel shows the upper sideband of the PN12 signal at higher resolution. Once again, note that approximately ten resolution elements at the center are lost due to the 300-Hz high-pass filter. This test illustrates the operation of the FFT analyzer at a typical resolution for the Sky Survey.

With the software modifications that were implemented for SETI, the FFT spectrum analyzer is now capable of running at any resolution between 305 Hz and 1 Hz. The PN12 signal was observed at the 1-Hz resolution but is not shown here.

B. Baseline Tests

The design of the Sky Survey processor includes a baseline module that will remove frequency-dependent variations in

the system gain by multiplying each frequency bin by the inverse of a baseline value. After each accumulation cycle of the processor, the baseline will be updated by applying an exponentially decaying filter to each of the frequency bins. The baseline algorithm is:

$$S_k(n) = P_k(n)/B_{k-1}(n) \qquad (1)$$

where

$S_k(n)$ = baseline adjusted signal in the nth frequency bin from the kth accumulation

$P_k(n)$ = accumulated signal in the nth frequency bin

$B_k(n)$ = (1-alpha)$B_{k-1}(n)$ + alpha $P_k(n)$ $\qquad (2)$

Alpha is the exponential factor with a time constant equal to the time required for the beam to sweep through approximately five HPBWs. A range of values for this time constant is being tested to determine the optimum value under field test conditions.

An illustration of this exponentially updated baseline technique is shown in Fig. 6. The effectiveness of the baseline updates can be seen as the accumulation time steps progress from the top spectrum (with no baseline update) to the bottom spectrum (where the baseline has been updated 11 times). For this example, a sinusoidal variation with amplitude 3 dB was imposed on the system gain while the gain in the first 100 frequency bins was increased by 1.5 dB above the average gain of the remaining channels. We note that the effects of these gain variations have been removed from the baseline adjusted spectra by the ninth accumulation time step.

C. Antenna Control Tests

The SETI requirement to drive the antenna at a relatively high angular rate (0.2 deg/s) calls for several tests of the antenna monitor and control subsystems. Antenna dynamics and servo control time constants must be understood in order to avoid the excitation of the normal modes of oscillation of the structure as the antenna is decelerated and accelerated into and out of the turnaround segments of each scan. Not only is it desirable to optimize the turnaround to avoid wasting time, it is extremely important to prevent excessive wear to the antenna drive systems. The accuracy of the antenna pointing system during rapid scanning must be determined.

Software has been written to enter sky mapping patterns in the standard right ascension and declination coordinates. The DSS 13 Station Controller converts these to time steps in azimuth–elevation coordinates. The monitor and control subsystem was also modified to deliver time-tagged antenna position data to a floppy disk recorder; the station was not equipped to monitor and record the position data while the antenna is driven at Sky Survey speeds.

Having overcome some unforeseen problems with the implementation of these tasks, we are about to begin the first of these antenna scanning tests.

D. RFI Survey

The impact of radio frequency interference on the Sky Survey is expected to be significant. One of the objectives of the field tests is to study RFI environment at the DSS 13 site to determine how to minimize the problem with the observational strategy and in the Sky Survey processor. We have begun to conduct a survey of RFI in the 1–10-GHz frequency range at the DSS 13 site. The objectives of the survey are to determine the signal strength and time history of RFI as a function of frequency. With this information we hope to identify ways to work around the problem, for example, by scheduling our SETI observations at troublesome frequencies on weekends, or at certain times of the day, or to avoid certain directions, etc.

The survey is being carried out with the SETI RFI Spectral Surveillance System (SETI RSSS) shown in Fig. 7. The one-meter parabolic antenna shown in the photo is currently located atop the DSN's FFT spectrum analyzer van. The system can be programmed to calibrate the system at any time and scan a specified number of frequency ranges between 1 and 10 GHz at specified azimuths and report signals that exceed a requested threshold. An example of the report from a typical frequency interval is shown in Fig. 8, where five signal components were detected above threshold (−143 dBm) approximately 240 kHz below the center frequency (1722.8 MHz).

The system is being operated to collect preliminary data that will be used to evaluate the SETI relational data base management software.

V. Conclusion

The SETI field test activities reported here are scheduled to continue for at least two years. The test results and observational experience will be used to finalize the design of the SETI Sky Survey processing system and to optimize the observational strategy and procedures in time to begin a full-scale Microwave Observing Program in 1990.

References

1. "SETI Science Working Group Report," *NASA Technical Paper 2244*, F. Drake, H. H. Wolfe, and C. L. Seeger (eds.), NASA, Wash. DC, Oct. 1983.

2. Gulkis, S., Olsen, E. T., and Tarter, J. C., "A Bimodal Search Strategy for SETI," in *Strategies for the Search for Life in the Universe*, M. D. Papagiannis (ed.), pp. 96–105, D. Reidel Publishing Co., 1980.

3. Peterson, A. M., Chen, K. S., and Linscott, I. R., "The Multichannel Spectrum Analyzer," *Proceedings of IAU Symposium 112, June 1984*, pp. 373-383, D. Reidel Publishing Co., 1985.

4. Morris, G. A., and Wilck, H. C., "JPL 2^{20} Channel 300 MHz Bandwidth Digital Spectrum Analyzer," *Proce. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, Oklahoma, pp. 808–811, 1978.

5. Emerson, R. F., "Biplexed Pipelined FFT," *The Deep Space Network Progress Report 42-34*, pp. 54–59, Jet Propulsion Laboratory, Pasadena, CA, 1976.

6. Klein, M. J., and Gulkis, S., "SETI: The Microwave Search Problem and the NASA Sky Survey Approach," *Proceedings of IAU Symposium 112, June 1984*, M. Papagiannis (ed.), pp. 397–403, D. Reidel Publishing Co., 1985.

7. Gulkis, S., "Optimum Search Strategy for Randomly Distributed CW Transmitters," *Proceedings of IAU Symposium 112, June 1984*, M. Papagiannis (ed.), pp. 411–417, D. Reidel Publishing Co., 1985.

8. Olsen, E. T., Lokshin, A., and Gulkis, S., "An Analysis of the Elements of an All Sky Survey," *Proceedings of IAU Symposium 112, June 1984*, M. Papagiannis (ed.), pp. 405–410, D. Reidel Publishing Co., 1985.

Table 1. Characteristics of the Sky Survey for a 34-m parabolic antenna

Spatial coverage	All sky
Frequency coverage	1–10 GHz & higher spot bands
Survey duration	5–7 yr (8 h/day)
Dwell time on source	0.5–10 s
Frequency resolution	10–30 Hz
Sensitivity:	$\sim 10^{-23}$ W/m
Spatial uniformity	12% peak-to-peak
Variation with frequency	$\sim f^{1/2}$
Polarization	Simultaneous dual circular
Signal type	Continuous wave (CW)
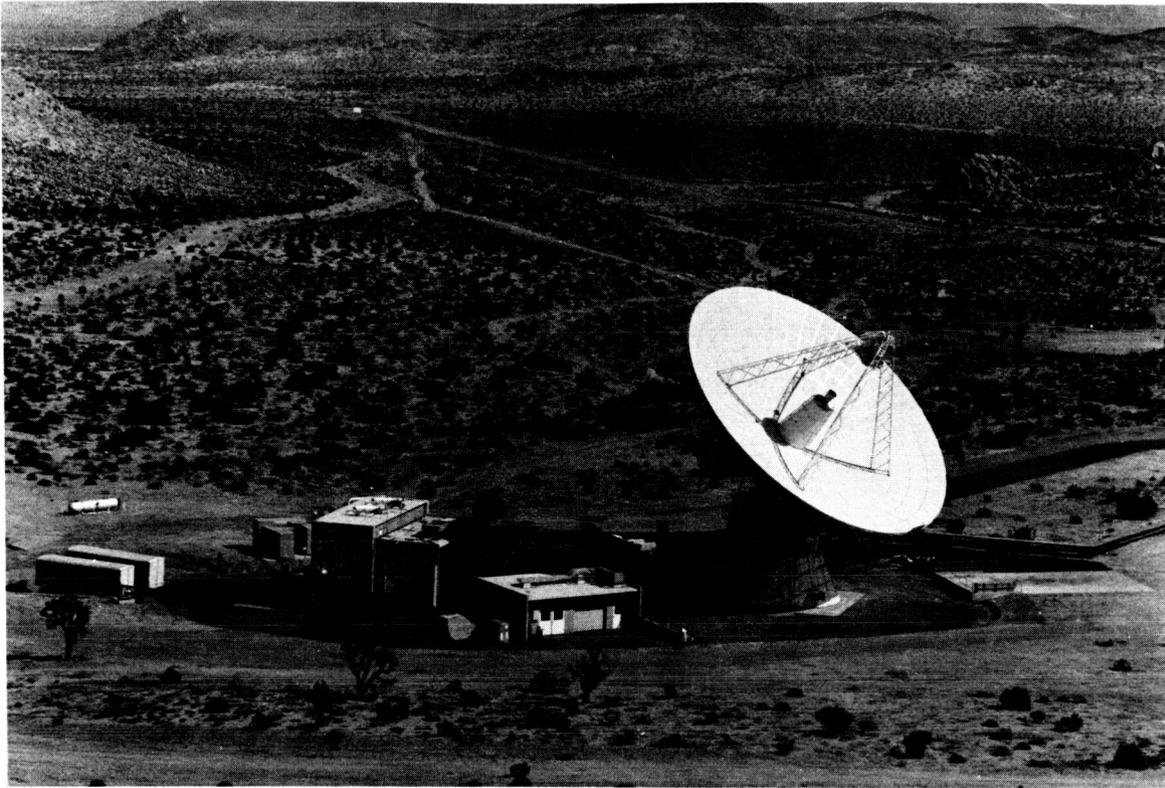
Fig. 1. SETI Sky Survey field tests at DSS 13, 26-m antenna
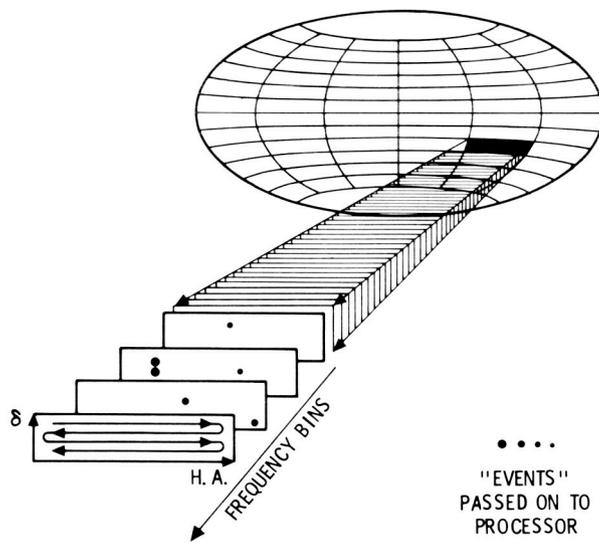
Fig. 2. 60-deg × 10-deg sky pixelation schematic
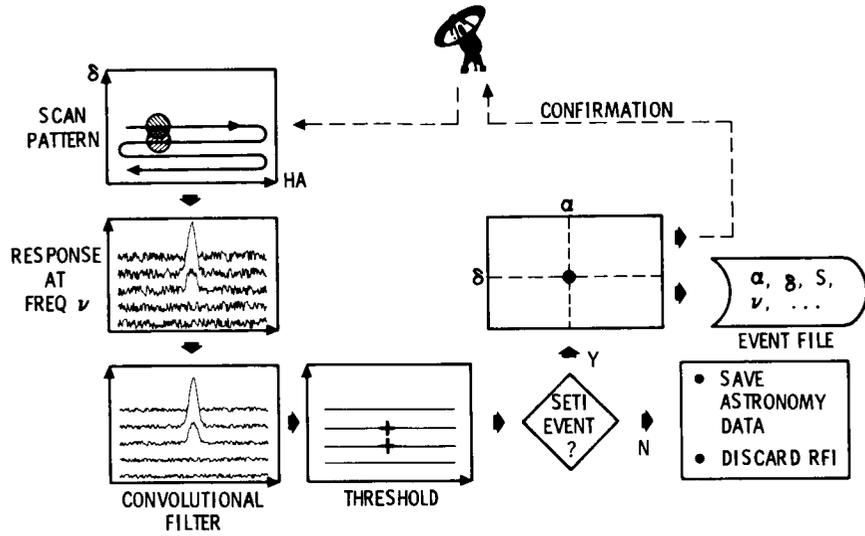
Fig. 3. Sky survey detection strategy
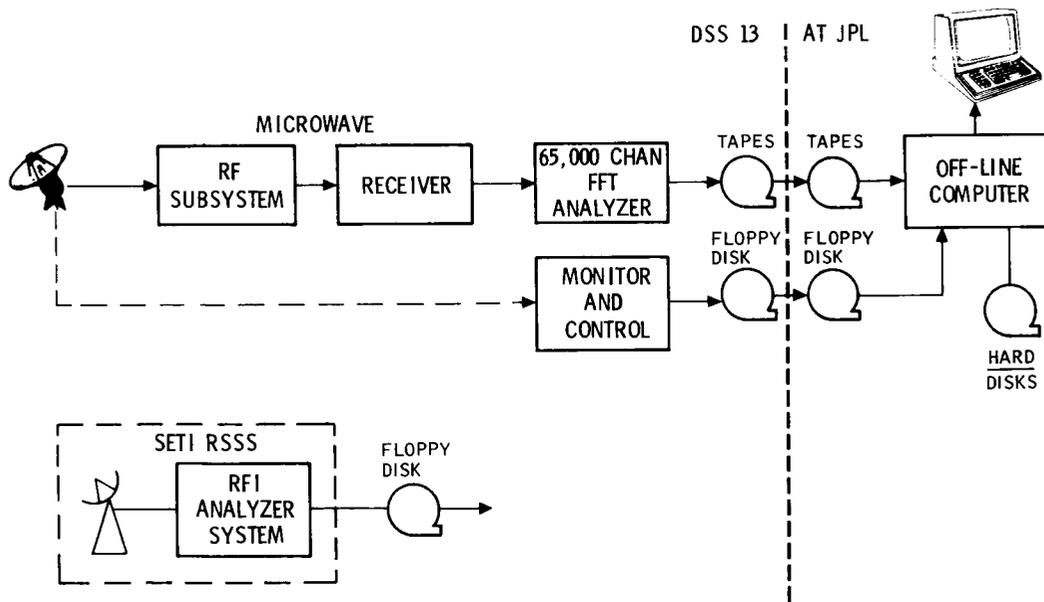
Fig. 4. System configuration for Sky Survey field tests (1985)

Fig. 5. DSS 13 verification tests using spacecraft signals
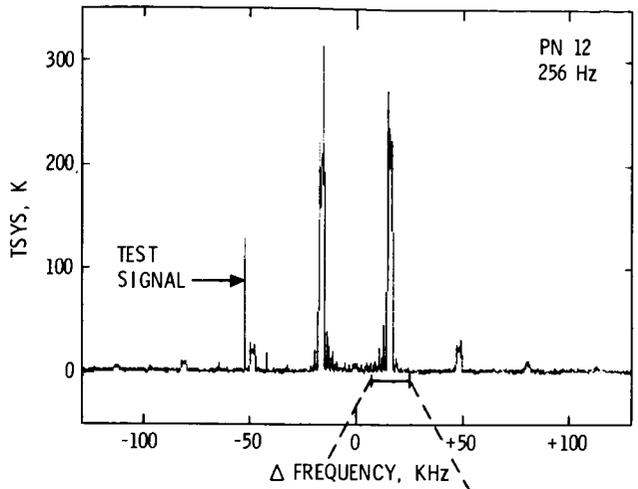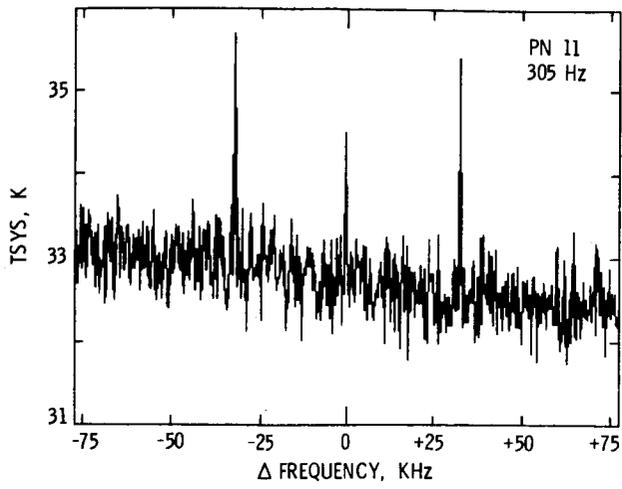
SYSTEM:
- DSS 13 26-METER ANTENNA
- DSN FFT SPECTRUM ANALYZER
- SETI CONFIGURATION AT 2295 MHz

SPACECRAFT DISTANCES (MARCH 1985)
- PIONEER 11......280 x 10^7 KM
- PIONEER 12.........6 x 10^7 KM
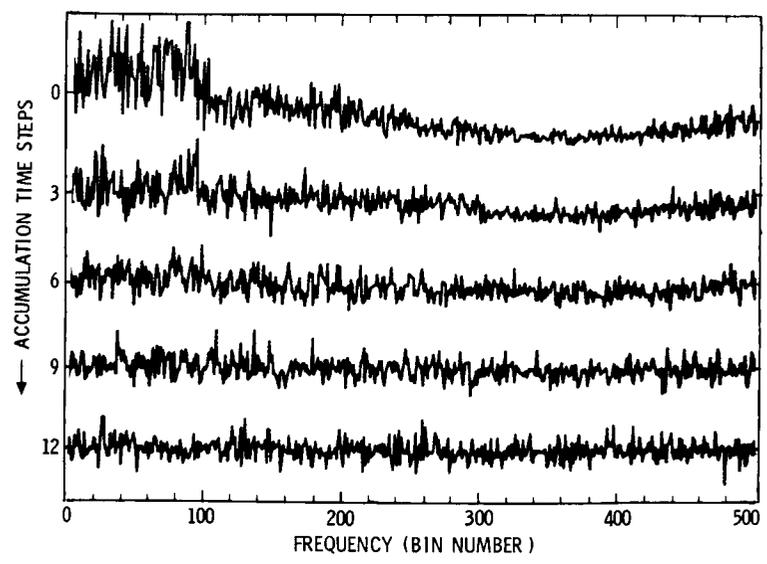
Fig. 6. Exponential baseline update
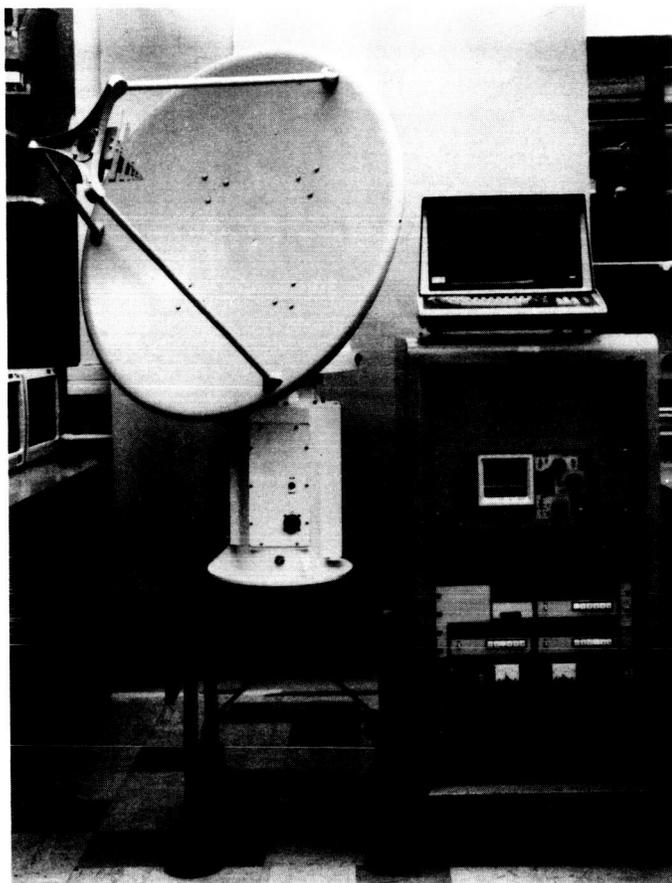
Fig. 7. SETI RFI spectral surveillance system
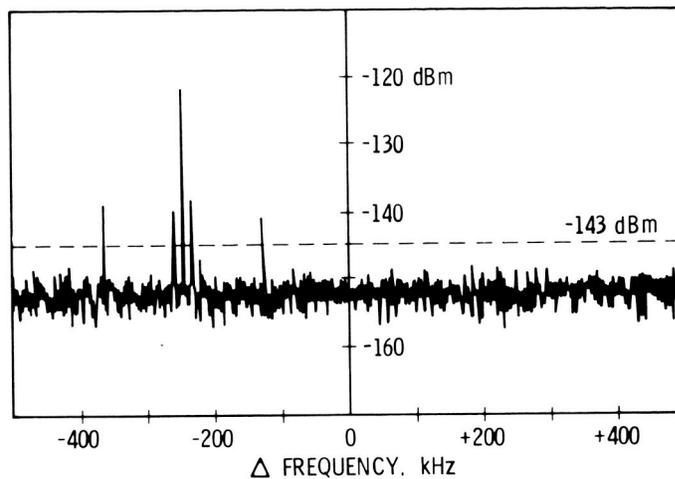
DATE....................03 / 05 / 85
TIME....................14 : 56 : 11
AZIMUTH..............60
CENTER FREQ........1.7228 GHz
RESOLUTION.........1 KHz

```
          REPORT
TASK = 1    FRAME = 5
   5 HITS  DETECTED

X = 137        Y = 153
X = 242        Y = 139
X = 256        Y = 196
X = 269        Y = 146
X = 374        Y = 148
```

Fig. 8. SETI radio spectrum surveillance system report

Fast Magnetic Tape Utilities for VAX/VMS Computers

N. E. Olson, R. F. Jurgens, and J. L. Robinett

Communications Systems Research Section

A package of subroutines for VAX/VMS computers has been developed to simplify the use of the QIO (Queued Input/Output) interface with magnetic tape drives. Routines are provided to perform all of the most common tape functions.

I. Introduction

As part of the standard software for their VAX/VMS computers, the Digital Equipment Corporation provides the SYS$QIO and SYS$QIOW routines to enable a program to request low-level operations on many I/O devices. Unfortunately, because of the general nature of these routines, the parameters they demand are often much more complex than would be necessary for simple operations such as rewinding a magnetic tape drive. In addition, there are several confusing deviations from the standard methods of argument binding. In order to decrease the frequency of programming errors and improve programmer productivity, we have written MTU-TILS, a package of magnetic tape utilities which provide most of the flexibility of the QIO interface for tape operations with a maximum of simplicity. These routines use only the standard FORTRAN and Pascal methods of argument binding and have comparatively simple argument lists.

Table 1 contains a summary of the major routines. In addition to the source code for the routines, the package includes a help file for use with the VAX/VMS on-line help facility, a command procedure to build and install the software, and a FORTRAN include-file and a Pascal environment file to define the structures used by the package.

The routines can be divided into two groups: one which provides basic tape operations, and one which allows certain operations to be performed on a "list of tape drives." This last group is meant for use with datasets which are too large to be stored on a single magnetic tape. It allows the programmer to define a sequence of tape drives which will be written or read in cyclic order, the next drive in the sequence being automatically engaged when the current drive encounters the end of its tape.

II. Usage

The actual caller interface for each routine is listed in the appendix. Below is a description of the sample FORTRAN program shown in Fig. 1. It uses the MTUTILS package to write a record to tape. The file "MTUTILS. ICL" is the include-file which must be included in all FORTRAN routines that use the MTUTILS package. Note the variable called DSCR in the variable declaration section. It is a descriptor block filled by the INIT_UNIT routine and used by the other routines to store information about the tape unit. The first thing the program does is enable the diagnostics. These are informational and error messages which, if requested, are written to the logical file name FOR$PRINT. After getting the user's

input, the program calls the TRANSLATE_DENSITY routine to translate the density in bytes per inch into the density code required by INIT_UNIT. As with most of the routines, an error flag is returned indicating whether its function was successfully performed. If this flag is set, the main program can call several status-checking routines to obtain more detailed information about the error.

Next, the program calls INIT_UNIT, which performs several functions such as mounting and allocating the drive and assigning it a channel, as well as filling the descriptor. Now that the descriptor has been filled, it can be passed to the other routines which perform the following functions: read a record, write a record, write an End-Of-File mark, rewind the tape drive, skip files (forward or reverse), skip records (forward or reverse), search for the End-Of-Volume mark, or release the unit. (For a list of tape units, as opposed to a single unit, the following exceptions apply: rewind cannot be done, and neither files nor records can be skipped in the reverse direction.) Also provided are routines that will make asynchronous requests to read a record, write a record or rewind a tape drive. After making such a request, the program can proceed to do some computation, and when it needs to know that the request has been satisfied, it can call the WAIT-READY routine, which waits until this occurs.

Finally, the sample program writes the record to tape, prints out a message to that effect, and calls RELEASE_UNIT to perform clean-up functions such as dismounting and deallocating the unit, with the option specified to automatically unload the tape as well.

The program does not take measures to recover from errors, but the MTUTILS package provides several facilities to do so. In addition to requesting diagnostic messages, the program can obtain status flags and other data concerning the tape drive. For example, in the sample program, the length of the last record read or written is retrieved using the GET_LENGTH function. In case of more obscure error conditions, the program can get the actual system status word from the most recent system service call by using the GET_STATUS function. The most commonly checked flags are included as parameters to the relevant routines (e.g., WRITE_TAPE automatically returns an end-of-tape flag). The help file which comes with the package contains complete details about all of the routines.

III. Performance

The read and write routines have been timed with a TU-78 tape drive. The following formulas give a good estimate of the time needed to read or write a record of L bytes on a lightly loaded VAX for the particular tape drive on which the routines were timed (these times, of course, will be too optimistic for a heavily loaded VAX). TU-78 tape drives can operate at a density of either 1600 or 6250 bytes per inch (bpi). At 1600 bpi, the time in milliseconds per record is given by:

$$T = 0.00537L + 13.3$$

At 6250 bpi, the time is given by:

$$T = 0.00138L + 10.7$$

These figures indicate a burst rate of about 116 inches per second (ips) for our tape drive. The nominal rate for TU-78 drives is 125 ips.

Table 1. Summary of the major routines[a]

Operation[b]	Versions Included			
	For single tape drives		For a list of tape drives	
	Synchronous	Asynchronous	Synchronous	Asynchronous
Initialize unit	X		X	
Release unit	X		X	
Read record	X	X	X	X
Write record	X	X	X	X
Write EOF mark	X		X	
Rewind	X	X		
Skip files forward	X		X	
Skip files reverse	X			
Skip records forward	X		X	
Skip records reverse	X			
Search for EOV mark	X		X	
Wait for completion (of asynchronous request)	X		X	

[a]See the appendix for a complete list including status tests, etc.
[b]EOF means End Of File; EOV means End Of Volume.

```
********************************************************************
*   This program writes a record of zeroes as the first record on a   *
*   tape.  The user specifies the record length, the name of the      *
*   tape drive and the tape density.                                  *
*                                                                     *
*   If a tape error occurs, an error message is printed and the       *
*   program halts.                                                    *
*                                                                     *
*   This program is not bullet-proof, e.g. the record length is not   *
*   tested to be within bounds.  This allows the user to force some   *
*   error messages out of the MTUTILS diagnostic facility.            *
********************************************************************

        INCLUDE 'MTUTILS.ICL'          !This program uses the MTUTILS package

        PARAMETER (MAX_RECORD_LEN = 65535)

        LOGICAL ERROR_FLAG, EOT_FLAG   !EOT = End Of Tape
        INTEGER DENSITY
        INTEGER FORMAT
        DATA FORMAT /0/                !Flag to use the default tape format
        INTEGER LENGTH
        BYTE BUFFER(MAX_RECORD_LEN)
        DATA BUFFER /65535*0/          !Initialize BUFFER to all zeroes

        CHARACTER*10 TAPE              !Name of the desired tape drive
        RECORD /TAPE_DSCR/ DSCR        !Declare a tape descriptor


C  Enable printing of diagnostic and error messages
C  (DIAGS is a constant defined in the INCLUDE file)

        CALL SET_DIAGNOSTICS(DIAGS)

C  Get the user's specifications

        WRITE(*,*), 'Enter the length of the record to be written (in bytes):'
        READ(*,*), LENGTH

        WRITE(*,*), 'Enter the name of the desired tape unit:'
        READ(*,10), TAPE
10      FORMAT(A10)

        WRITE(*,*), 'Enter the desired tape density (bpi):'
        READ(*,*), DENSITY

C  Translate the density in bytes per inch into a tape density code

        CALL TRANSLATE_DENSITY (DENSITY,ERROR_FLAG)
        IF (ERROR_FLAG) STOP    !Diagnostics have already been printed

C  Initialize the unit

        CALL INIT_UNIT (TAPE,DENSITY,FORMAT,DSCR,ERROR_FLAG)
        IF (ERROR_FLAG) STOP    !Diagnostics have already been printed

C  Write the record

        CALL WRITE_TAPE (DSCR,LENGTH,BUFFER,RETRY,EOT_FLAG,ERROR_FLAG)
        IF (ERROR_FLAG) STOP    !Diagnostics have already been printed

C  Print out the length of the last record written

        WRITE(*,*), 'A record of length ', GET_LENGTH(DSCR)
        WRITE(*,*), 'has been written.'
        IF (EOT_FLAG) WRITE(*,*), 'The end of the tape was reached.'

C  Release the unit and unload the tape.

        CALL RELEASE_UNIT (DSCR,UNLOAD,ERROR_FLAG)
        IF (ERROR_FLAG) STOP    !Diagnostics have already been printed

        END
```

Fig. 1. Sample program

Appendix

Summary of FORTRAN Routines

A summary of the FORTRAN caller interfaces to each of the routines in the package is provided in this appendix.

```
Routines with plural names (e.g. INIT_UNITS and READ_TAPES) and

those whose names begin with "LIST_" are for a list of tape units.

Others apply to single tape units.  Routines whose names begin

with a "Q" perform asynchronous requests.

After the list of routines is a description of their parameters.


Initialization and release:

 CALL TRANSLATE_DENSITY(DENSITY, ERROR_FLAG)

 CALL SET_DIAGNOSTICS(DIAG_FLAG)

 CALL INIT_UNIT(NAME, DENSITY, FORMAT, DSCR, ERROR_FLAG)

 CALL INIT_UNITS(NUMUNITS, NAMES, DENSITY, FORMAT, DLIST, ERROR_FLAG)

 CALL RELEASE_UNIT(DSCR, UNLOAD_FLAG, ERROR_FLAG)

 CALL RELEASE_UNITS(DLIST, UNLOAD_FLAG, ERROR_FLAG)


Reads and Writes:

 CALL QREAD_TAPE(DSCR, BYTE_COUNT, BUFFER, RETRY_FLAG, ERROR_FLAG)

 CALL QREAD_TAPES(DLIST, BYTE_COUNT, BUFFER, SWAPNUM, RETRY_FLAG, ERROR_FLAG)

 CALL READ_TAPE(DSCR, BYTE_COUNT, BUFFER, RETRY_FLAG, EOV_FLAG, ERROR_FLAG)

 CALL READ_TAPES(DLIST, BYTE_COUNT, BUFFER, SWAPNUM, RETRY_FLAG, EOV_FLAG, ERROR_FLAG)

 CALL QWRITE_TAPE(DSCR, BYTE_COUNT, BUFFER, RETRY_FLAG, ERROR_FLAG)

 CALL QWRITE_TAPES(DLIST, BYTE_COUNT, BUFFER, SWAPNUM, RETRY_FLAG, ERROR_FLAG)

 CALL WRITE_TAPE(DSCR, BYTE_COUNT, BUFFER, RETRY_FLAG, EOT_FLAG, ERROR_FLAG)

 CALL WRITE_TAPES(DLIST, BYTE_COUNT, BUFFER, SWAPNUM, RETRY_FLAG, ERROR_FLAG)

 CALL WRITE_EOF(DSCR, ERROR_FLAG)

 CALL LIST_WRITE_EOF(DLIST, SWAPNUM, ERROR_FLAG)
```

```
Skipping and Searching:

 CALL QREWIND(DSCR,UNLOAD_FLAG,ERROR_FLAG)

 CALL REWIND(DSCR,UNLOAD_FLAG,ERROR_FLAG)

 CALL SKIP_FILES(DSCR,COUNT,ERROR_FLAG)

 CALL LIST_SKIP_FILES(DLIST,COUNT,SWAPNUM,ERROR_FLAG)

 CALL SKIP_RECORDS(DSCR,COUNT,ERROR_FLAG)

 CALL LIST_SKIP_RECORDS(DLIST,COUNT,SWAPNUM,ERROR_FLAG)

 CALL SEARCH_EOV(DSCR,EOT_FLAG,ERROR_FLAG)

 CALL LIST_SEARCH_EOV(DLIST,EOT_FLAG,SWAPNUM,ERROR_FLAG)


Synchronizers:

 CALL WAITREADY(DSCR,ERROR_FLAG)

 CALL LIST_WAITREADY(DLIST,SWAPNUM,ERROR_FLAG)


Routines to extract extra information and error conditions:

 LENGTH = GET_LENGTH(DSCR)        LENGTH = LIST_GET_LENGTH(DLIST)

 STATUS = GET_STATUS(DSCR)        STATUS = LIST_GET_STATUS(DLIST)

 BOT_FLAG = TEST_BOT(DSCR)        BOT_FLAG = LIST_TEST_BOT(DLIST)

 EOF_FLAG = TEST_EOF(DSCR)        EOF_FLAG = LIST_TEST_EOF(DLIST)

 EOT_FLAG = TEST_EOT(DSCR)        EOT_FLAG = LIST_TEST_EOT(DLIST)

 EOV_FLAG = TEST_EOV(DSCR)        EOV_FLAG = LIST_TEST_EOV(DLIST)

 HWL_FLAG = TEST_HWL(DSCR)        HWL_FLAG = LIST_TEST_HWL(DLIST)

 LOST_FLAG = TEST_LOST(DSCR)      LOST_FLAG = LIST_TEST_LOST(DLIST)

 PAR_ERR_FLAG = TEST_PAR_ERR(DSCR)

 PAR_ERR_FLAG = LIST_TEST_PAR_ERR(DLIST)

 CURRENT_DRIVE = GET_UNIT_NUM(DLIST)

 CALL ERROR_CHECK(DSCR,BOT_FLAG,EOF_FLAG,EOT_FLAG,HWL_FLAG,
                  LOST_FLAG,PAR_ERR_FLAG)

 CALL LIST_ERROR_CHECK(DLIST,BOT_FLAG,EOF_FLAG,EOT_FLAG,HWL_FLAG,
                       LOST_FLAG,PAR_ERR_FLAG)
```

Below is a table explaining the parameters used in this package.

Note: Non-standard types such as "word" and "TAPE_LIST" are defined in the include file (for FORTRAN) or the environment file (for Pascal)

Parameter	FORTRAN type	Pascal type	Description
BOT_FLAG	LOGICAL	boolean	Beginning Of Tape
BUFFER	BYTE or INTEGER	integer	Data record buffer
BYTE_COUNT	INTEGER*2	word	Length of record
COUNT	INTEGER*2	word	Number of records or files
DENSITY	INTEGER	integer	Tape density
DIAG_FLAG	LOGICAL	boolean	Controls diagnostic messages
DLIST	RECORD/TAPE_LIST/	TapeList	List of descriptors
DSCR	RECORD/TAPE_DSCR/	TapeDscr	Tape descriptor
EOF_FLAG	LOGICAL	boolean	End Of File
EOT_FLAG	LOGICAL	boolean	End Of Tape
EOV_FLAG	LOGICAL	boolean	End Of Volume
ERROR_FLAG	LOGICAL	boolean	Error
FORMAT	INTEGER	integer	Tape format
HWL_FLAG	LOGICAL	boolean	Hardware Write Lock
LOST_FLAG	LOGICAL	boolean	Tape position Lost
NAME	CHARACTER*L	packed array of char	Logical name of tape unit
NAMES	CHARACTER*L(N)	array of Name	List of logical unit names
NUMUNITS	INTEGER	integer	Length of name-list
PAR_ERR_FLAG	LOGICAL	boolean	Parity Error
RETRY_FLAG	LOGICAL	boolean	Controls automatic error-retry
SWAPNUM	INTEGER	integer	If non-zero, tape to swap*
UNLOAD_FLAG	LOGICAL	boolean	Controls unloading of tape

* Number of the tape unit in a tape list which has reached EOT and needs to
  be swapped for the next tape in the volume set. This value is returned
  by the list-of-tape routines so that the calling program can alert the
  operator when tape swaps are necessary.

Demonstration of the Fiducial Concept Using Data From the March 1985 GPS Field Test

J. M. Davidson, C. L. Thornton, S. A. Stephens, S.-C. Wu,
S. M. Lichten, J. S. Border, and O. J. Sovers
Tracking Systems and Applications Section

T. H. Dixon
Geology and Planetology Section

B. G. Williams
Navigation Systems Section

The first field test of NASA's GPS Geodetic Program took place in March of 1985. The principal objective of this test was the demonstration of the feasibility of the fiducial station approach to precise GPS-based geodesy and orbit determination. Other objectives included an assessment of the performance of the several GPS receiver types involved in these field tests and the testing of the GIPSY Software for GPS data analysis. In this article, we describe the GIPSY (GPS Inferred Positioning SYstem) software system and examine baseline solutions for consistency with independent measurements made using very long baseline interferometry.

I. The Fiducial Network Concept

The accuracy of GPS-based baseline measurements depends in large part on the accuracies with which the GPS satellite orbits have been determined. Currently available post-fit ephemerides define a self-consistent frame of reference at the level of approximately 1 part per million (ppm). Since expected values of crustal motion are on the order of 1–10 cm/yr, or less, a geodetic measurement program relying on post-fit ephemerides, and requiring temporal resolution of less than a decade, would be sensitive to such motions over distances of no more than several tens of km. On the other hand, a reference system based on very long baseline interferometry (VLBI) or satellite laser ranging (SLR) measurements is potentially self-consistent at a level of better than 0.01 ppm. Hence, a GPS-based geodetic network tied to the VLBI reference system could be utilized for geodetic measurements on a worldwide basis.

A means of establishing this frame tie and improving the GPS ephemeris has been developed at JPL and other institutions, and is referred to as the fiducial network approach (Ref. 1 and a private communication with J. L. Fanselow and J. B. Thomas, 1983). In this approach, three or more receivers are placed at sites, called fiducial sites, whose positions have been well-established by an independent technique, such as

VLBI. Additional receivers, known as mobile receivers, are placed at sites of geodetic interest. During a GPS measurement, the fiducial receivers record data jointly with the mobile receivers, enabling simultaneous determination of accurate GPS satellite orbits and geodetic baselines. It should be noted that by locating the fiducial receivers at VLBI stations, the GPS orbit and baseline solutions are inherently expressed in the coordinate frame of the quasi-stellar radio sources. This has the aesthetic appeal of unifying the GPS and VLBI results in an inertial, or absolute, frame of reference, enabling simultaneous display, direct comparison and simultaneous use in geophysical interpretation. A schematic illustration of the fiducial network method for precision GPS-based geodesy is shown in Fig. 1.

Covariance analysis has shown that orbit accuracies on the order of 0.1 ppm (or about 3 m) are attainable using currently available systems, involving the NOAA/NGS POLARIS VLBI sites as fiducial stations and carrier phase as the data type (Ref. 2). GPS orbit estimation using data from the March Test has verified this projection (Ref. 3). Anticipated improvements in the next generation systems, including improved models of orbital dynamics, widespread availability of 0.5 cm accurate water vapor radiometers, improved knowledge of fiducial station locations, the availability of a fiducial station in the Northern part of South America (for Caribbean geodesy), and the availability of 1–2 cm accurate absolute range as a data type will lead to orbit and baseline accuracies of order 0.01 ppm by 1989 (Ref. 4).

II. The March 1985 Field Tests

The March 1985 Field Test took place between March 28 and April 4, 1985. Receivers were placed at the NOAA/NGS POLARIS VLBI stations (Ref. 5) in Westford, MA, Richmond, FL, and Ft. Davis, TX; and at the Mojave, Owens Valley and Hat Creek VLBI stations in California. Additional GPS receivers located near Mammoth Lakes, CA, Austin, TX, Dahlgren, VA and Point Mugu, CA (Fig. 2) also recorded data during the test. These sites were equipped with a combination of GPS receiver types, including SERIES-X, TI-4100 and Air Force Geophysical Laboratory (AFGL) dual-frequency receivers. Three of the sites in California were equipped with WVRs. In all, seventeen different institutions made substantial contributions to this test (Ref. 1). A more detailed description of the March 1985 test can be found in Refs. 1 and 4.

III. The GIPSY Software System

The GIPSY (for GPS Inferred Positioning System) software for GPS data analysis was developed between January 1985 and the present at JPL and is still undergoing significant modi-

fication to improve observable models and the user interface. It consists of approximately 70,000 lines of code, approximately one third of which was adapted from preexisting software systems, such as the VLBI data analysis code MASTERFIT (Ref. 6) and the satellite orbital dynamics modeling code PATH VARY[1] (see also Ref. 9). The remaining two-thirds consist of newly written code, a good portion of which is also used in the OASIS (Orbit Analysis SImulation System) error analysis software (Ref. 10). It is written in FORTRAN and resides on a VAX 11/785, running under the VMS Version 4.3 operating system.

The GIPSY software system consists of a series of eight principal modules. These perform various operations, including translation of data to the VAX, interactive editing, calibration and compression, computation of an a priori ephemeris and theoretical observables, generation of differenced data types, least-squares parameter estimation and postprocessing. This software is capable of combining data from all receiver types and of processing different data types (e.g., carrier phase and pseudo-range) simultaneously. The modeling and calibrations implemented in these modules are sufficient to support 0.01 ppm accuracy and better.

Estimated parameters may include satellite position and velocity at epoch, station position and velocity, station and satellite clocks, range biases (for carrier phase data), two solar radiation pressure coefficients and a y-bias factor for each satellite, zenith tropospheric delay at each station, earth orientation parameters, geocenter coordinates, solid earth tide parameters, corrections to precession and nutation, and general relativity gamma. (It should be noted that in normal data processing, only a subset of these parameters are estimated.) Options for clock modeling include constant, linear, quadratic and stochastic clocks for undifferenced and single differenced observables. Clocks may also be removed altogether using double differencing.

Typical throughput times are approximately 15 min of VAX central processing unit time per station day of GPS data (assuming observations recorded at 6-min intervals), excluding the time required for meteorology data reduction. A schematic flow chart illustration of the GIPSY software system is shown in Fig. 3.

IV. Analysis and Results

In this initial investigation of the fiducial approach, a subset of the March test data was selected as representative of the

---

[1]S.-C. Wu, et al., 1986, OASIS Programmer's Guide, Version 1.0, JPL Internal Document D-3140, Jet Propulsion Laboratory, Pasadena, Calif.

overall network. This subset consisted of the carrier phase data from April 3, 1985 from the AFGL dual-frequency GPS receivers located at the POLARIS VLBI sites at Westford, MA, Richmond, FL and Ft. Davis, TX and from the SERIES-X receivers located at the Mojave and Owens Valley VLBI sites in California. GPS satellites included in this subset were GPS 1, GPS 3, GPS 4, GPS 6, GPS 8 and GPS 9 (launch sequence numbers).

Tropospheric calibration was done using water vapor radiometer data for the stations in California and surface meteorology data as input to atmospheric models for all other stations. Ionospheric calibration was achieved using an appropriate linear combination of the phase observables from the separate L-band channels (1.57542 GHz and 1.22760 GHz) based on an assumed inverse frequency-squared dispersive relationship. Differences in measurement epochs for the two receiver types were reconciled by "compressing" the Series-X data, which have observation epochs occurring at 14-s intervals, to match the epochs of the AFGL receiver data, which occur at 6-min intervals. Compression was achieved by fitting the difference between the data and a computed observable to a second order polynomial over the 6-min interval.

Estimated parameters included satellite positions and velocities, station positions, station and spacecraft clocks and range biases. Fiducial station locations were given a priori constraints of 1 cm in each coordinate; mobile stations were given constraints of 2 km. Clocks were treated as stochastic bias parameters, having time constants of zero, which has the same effect as removing clocks by explicitly double-differencing. A priori station locations were derived from several sources, including both VLBI (Ref. 7) and local survey (Ref. 8) data and a combination of these with correction to the geocenter, using SLR data (C. C. Counselman and R. W. King, 1985, private communication). A priori ephemerides were those of the U.S. Naval Surface Weapons Center (NSWC), rotated and scaled to correct for known differences in orientation and GM for the coordinate systems in current use at NSWC and JPL.

Data weights were adjusted to make chi-square per degree of freedom equal to unity.

Two separate parameter estimations were made using this data set. The strategy employed in this initial effort was to designate a single mobile receiver and look at mobile-fiducial baselines. In the first, the GPS receiver at Mojave was designated as the "mobile" receiver; in the second, the receiver at Ft. Davis was taken to be the mobile receiver. Baseline length results from these two cases, excluding results for the fiducial-fiducial baselines, were compared to the independent results based on VLBI measurements (Refs. 7, 8 and C. C. Counselman and R. W. King, 1985, private communication). This comparison showed agreement between GPS and VLBI lengths at the level of 0.13 ppm. In a third parameter estimation, all GPS receivers were treated as mobile receivers, with no estimation of satellite orbit parameters. The agreement in baseline length between the GPS- and VLBI-based estimations for this third case was at the level of 0.72 ppm, a degradation of over a factor of five from the case in which fiducial stations were used for orbit estimation. A summary of these results is given in Fig. 4.

V. Conclusion and Summary

The GIPSY software system for GPS data analysis has been developed at the Jet Propulsion Laboratory. This software was used to process a portion of the data from the March 1985 precision baseline tests and to investigate the feasibility of the fiducial station approach to precision GPS geodesy. It was found that the use of fiducial stations for GPS orbit determination improved the RMS agreement in baseline length between collocated GPS and VLBI antennas by over a factor of 5, to a level of agreement of 0.13 ppm, a level of accuracy which is sufficient for regional geodesy (baselines up to 1000 km). Anticipated system improvements should further improve GPS-based geodetic measurements to a level of approximately 0.01 ppm, which will enable the utilization of GPS-based geodetic systems on a worldwide basis.

References

1. Davidson, J. M., Thornton, C. L., Vegos, C. J., Young, L. E. and Yunck, T. P., 1985, "The March 1985 Demonstration of the Fiducial Network Concept: A Preliminary Report," *Proceedings First International Conference on Precise Positioning With the Global Positioning System*, 603–612, April 15–19, 1985, Rockville, Maryland.

2. Bertiger, W. I., Wu, S.-C., Border, J. S., Lichten, S. M., Williams, B. G. and Wu, J.-T., 1986, "High Precision GPS Orbit Determination Using March 1985 Demonstration Data," *Proceedings AIAA 24th Aerospace Sciences Meeting*, January 6–9, 1986, Reno, Nevada (in press).

3. Wu, S.-C., Lichten, S. M., Williams, B. G., Border, J. S., Bertiger, W. I., Wu, J.-T. and Yunck, T. P., 1986, "Precise Orbit Determination of GPS and LANDSAT-5," *Proceedings Fourth International Geodetic Symposium on Satellite Positioning*, April 28–May 2, 1986, Austin, Texas (in press).

4. Melbourne, W. G., Dixon, T. H. and Thornton, C. L., "NASA GPS-Based Geodetic Program in Mexico and the Caribbean," *Proceedings Fourth International Geodetic Symposium on Satellite Positioning*, April 28–May 2, 1986 (in press).

5. Carter, W. E., Robertson, D. S., Pettey, J. E., Tapley, B. D., Schutz, B. E., Eanes, R. J., and Lefeng, M., 1984, "Variations in the Rotation of the Earth," *Science 224*, 957–961.

6. Fanselow, J. L., and Sovers, O. J., 1985, *Observation Model and Parameter Partials for the JPL VLBI Parameter Estimation Software "MASTERFIT-V2.0,"* JPL Publication 83-39, Rev. 1, Jet Propulsion Laboratory, Pasadena, Calif.

7. Ryan, J. W. and Ma, C., 1985, "Crustal Dynamics Projece Data Analysis Fixed Station Geodetic Results," NASA Technical Memorandum 86229, Washington, D.C., October 1985.

8. NASA, "Crustal Dynamics Project: Catalogue of Site Information," NASA/GSFC Document X-601-85-6, Greenbelt, MD, March, 1985.

9. Wu, S. C., and Thornton, C. L., 1985, "OASIS - A New GPS Covariance and Simulation Analysis Software System," *Proceedings First International Conference on Precise Positioning with the Global Positioning System*, pp. 337–346. April 15–19, 1985, Rockville, Maryland.
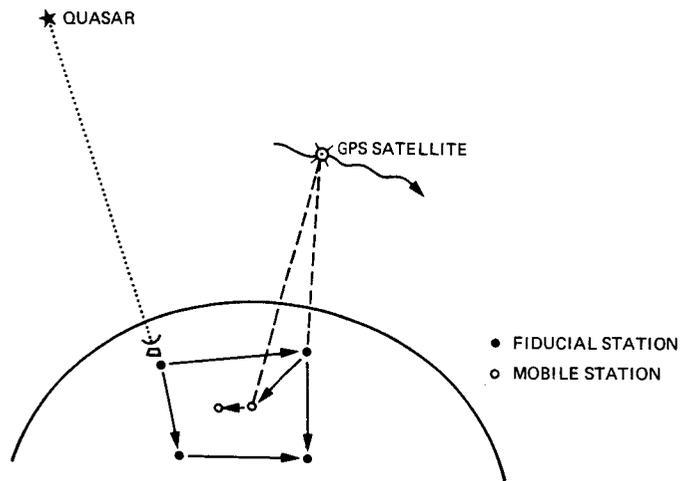
Fig. 1. A schematic illustration of the fiducial network method for precision GPS geodesy is shown
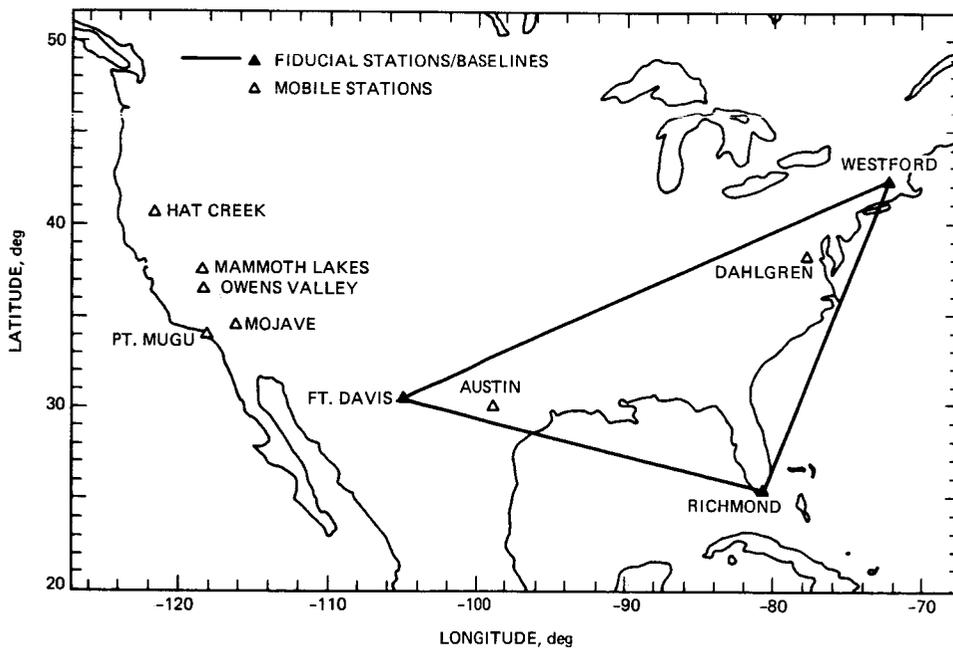
Fig. 2. The locations of the fiducial and mobile GPS receivers involved in the March 1985 demonstration of the fiducial network concept are shown
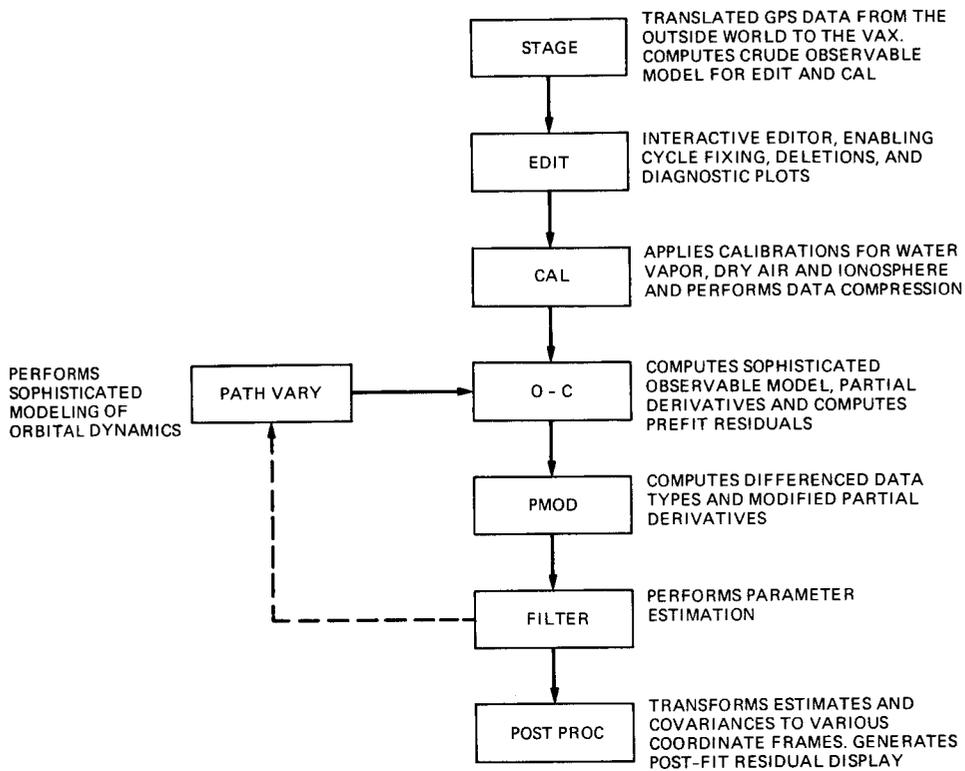
Fig. 3. A schematic flow chart of the GIPSY software system is shown
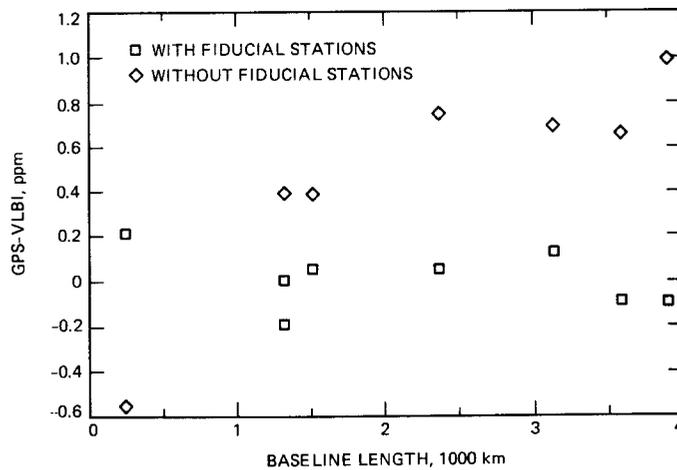
Fig. 4. A comparison of GPS- and VLBI-based geodetic measurements is shown

Referees

The following people have refereed articles for *The Telecommunications and Data Acquisition Progress Report*. By attesting to the technical and archival value of the articles, they have helped to maintain the excellence of this publication during the past year.

D. A. Bathker	E. C. Gatz	S. M. Lichten	J. I. Statman
J. S. Border	G. L. Goltz	W. K. Marshall	C. T. Stelzried
R. C. Clauss	R. Hartop	W. L. Martin	L. Swanson
R. C. Coffin	L. Howard	J. P. McDanell	S. A. Townes
R. B. Crow	T. Howe	S. Petty	R. N. Treuhaft
F. Davarian	I.-S. Hsu	F. Pollara	V. A. Vilnrotter
L. J. Deutsch	R. Jurgens	W. Rafferty	R. J. Wallace
D. Divsalar	J. Katz	D. L. Robinson	C. C. Wang
G. S. Downs	M. J. Klein	E. H. Satorius	W. M. Whitney
R. Emerson	R. Kumar	H. M. Shao	J. H. Wilcher
T. M. Eubanks	C. R. Lahmeyer	W. L. Sjogren	R. Winkelstein
R. H. Evans	F. L. Lansing	M. A. Slade	L. E. Young
B. W. Falin	J. W. Layland	S. D. Slobin	T. P. Yunck
N. Fanelli	J. R. Lesh	P. H. Stanton	

1. Report No. PR 42-86	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
The Telecommunications and Data Acquisition Progress Report 42-86	August 15, 1986
	6. Performing Organization Code

7. Author(s) E. C. Posner (ed.)	8. Performing Organization Report No. PR 42-86

9. Performing Organization Name and Address	10. Work Unit No.
JET PROPULSION LABORATORY California Institute of Technology 4800 Oak Grove Drive Pasadena, California 91109	11. Contract or Grant No. NAS7-918
	13. Type of Report and Period Covered

12. Sponsoring Agency Name and Address	TDA Progress Report April--June 1986
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

Mother #29085 (MG)

16. Abstract

This publication, one of a series formerly titled The Deep Space Network Progress Report, documents DSN progress in flight project support, tracking and data acquisition research and technology, network engineering, hardware and software implementation, and operations. In addition, developments in Earth-based radio technology as applied to geodynamics, astrophysics and the radio search for extraterrestrial intelligence are reported.

17. Key Words (Selected by Author(s))	18. Distribution Statement
Astronautics Communications Mathematical and Computer Sciences Space Sciences	Unclassified--Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	vii + 307	

JPL 0184 R 9/83